Introduction à l'économétrie

Le modèle de régression linéaire

Support de cours destiné aux étudiants de licence et master SHS/MIASHS/MASE

version du 29/10/2013, 20:34

Université Charles-de-Gaulle Lille 3
UFR MSES

O. Torrès

Ce document contient des animations ; pour les visualiser, il est nécessaire d'utiliser Adobe Reader

Introduction : présentation du cours

Ce cours est une introduction aux méthodes et modèles de base de l'économétrie. Cette dernière s'entendra ici comme une branche de la statistique mathématique (ou inférentielle) dans laquelle

1. les modèles statistiques utilisés sont constitués à partir d'une adaptation d'un modèle économique théorique ou peuvent avoir une interprétation qui relève du raisonnement économique
2. les données utilisées pour l'inférence statistique proviennent de l'observation du fonctionnement de l'économie

On peut résumer la définition proposée de l'économétrie en assimilant cette dernière à la statistique appliquée à des situations pouvant être décrites par la science économique.

Sur le plan de la statistique, cette définition amène plusieurs remarques.

1. Du fait de cette connexion avec la science économique, les variables pour lesquelles les modèles statistiques [1] de l'économétrie (qu'on appellera simplement modèles économétriques par la suite) sont construits sont également des variables que l'on retrouve dans les modèles économiques. Ces derniers décrivent typiquement les relations qui existent entre plusieurs variables économiques. Par conséquent, les modèles économétriques sont destinés à représenter des relations qui sont supposées exister entre les variables tout en permettant de les interpréter. Ces modèles mettront ainsi en évidence des paramètres qui expriment des relations entre variables et en caractérisent la forme.
2. L'inférence statistique qui sera menée dans le contexte du modèle économétrique portera essentiellement sur ces paramètres ; ceux-ci seront donc les paramètres d'intérêt (voir les points 6 et 7 à la page 236) du modèle économétrique.
3. De par la nature même des modèles économétriques (voir le point 1 ci-dessus), les méthodes d'inférence qui seront mises en œuvre pour étudier ces paramètres seront quasi-exclusivement *multivariées*.

Ce cours peut être considéré comme un cours de statistique, et dans lequel on présentera des modèles et des méthodes d'inférence de base couramment utilisés en économétrie. Bien que le contenu de ce cours soit orienté par la pratique statistique dans le domaine des modèles économiques, les méthodes statistiques qui seront présentées peuvent bien entendu s'appliquer à des contextes autres (les premières applications du modèle de base qui sera présenté dans le cours sont d'ailleurs apparues dans des domaines bien distincts de l'économie).

1. On rappelle qu'un modèle statistique — et donc un modèle économétrique — est un ensemble d'hypothèses probabilitstes sous lesquelles il sera notamment possible de dériver les propriétés des diverses méthodes statistiques utilisées dans le cadre de ce modèle. Voir page 242.

Bien que cette question aille au-delà du contenu du cours, on peut se demander ce qu'apporte l'économétrie par rapport à une analyse économique théorique.

Les modèles théoriques proposent une description du fonctionnement de l'économie (ou de certains de ses marchés) au moyen d'un ensemble de relations entre variables économiques. Une fois cette description proposée, plusieurs types questions peuvent se poser. Par exemple :

1. Les relations établies par le modèle théorique existent-elles vraiment ?
2. En supposant que ce soit le cas, quelles sont les propriétés de ces relations ? Si deux variables X et Y sont mises en relation, peut-on supposer que cette dernière est linéaire ? non linéaire ? Les variables X et Y varient-elles ensemble dans le même sens ou en sens opposé ?
3. En supposant que le modèle théorique propose une relation entre deux variables X et Y exprimée au moyen d'une fonction appartenant à une classe donnée (p. ex. fonctions linéaires, log-linéaires, polynômes, *etc*), la classe proposée est-elle la bonne ?
4. En supposant que ce soit le cas, autrement dit s'il existe un élément dans la classe de fonctions qui permet d'exprimer la relation existant réellement entre X et Y, quel est cet élément ? Si par exemple la relation est linéaire (la courbe représentant la fonction reliant une variable à l'autre est une droite) quelle est la valeur de chacun des coefficients exprimant cette relation ?

Les questions ci-dessus sont de deux natures :

— Certaines (la première et la troisième) posent celle de la validité du modèle économique théorique, c'est à dire sa capacité à rendre compte correctement du fonctionnement réel de l'économie.

— Les autres questions traitent de la possibilité d'utiliser un modèle théorique pour émettre sur la nature des relations entre variables économiques des énoncés de type qualitatif (par exemple : l'augmentation d'un taux d'intérêt entraîne la baisse du taux d'inflation) ou quantitatif (par exemple : une augmentation d'1 point du taux de croissance du PIB, permet, sans changer le niveau de la dette de l'État, de diminuer de 10% le niveau des impôts directs perçus par l'État au cours des 2 prochaines années).

Les réponses à ces questions sont déterminantes. On comprend aisément qu'il est intéressant de savoir si un modèle économique théorique parvient à rendre compte correctement de la réalité d'une relation économique. Si ce n'est pas le cas, on peut le considérer comme faux, et son utilisation ne contribue pas à une meilleure compréhension des mécanismes économiques. En supposant qu'un modèle soit considéré comme adéquat, la possibilité de l'utiliser pour parvenir à des énoncés quantitatifs non-triviaux est d'un intérêt majeur pour les économistes (possibilité d'effectuer des prévisions, conduite de politiques économiques, *etc*). Or, parmi les modèles théoriques économiques formulés, peu (aucun ?) offrent une telle possibilité. Par ailleurs, ces modèles eux-mêmes ne proposent aucune méthode permettant de savoir s'ils sont justes ou faux.

L'utilisation des diverses méthodes d'inférence de l'économétrie complète la formulation d'un modèle théorique et vise à apporter des réponses à des questions du type de celles mentionnées ci-dessus, en fournissant des *estimations* des paramètres des diverses relations apparaissant dans les modèles économiques, en permettant de *tester* l'adéquation d'une formulation proposée par un modèle théorique avec la réalité. De plus, parce que ces estimations et tests sont effectués en utilisant les méthodes de l'inférence statistique, ils sont accompagnés d'une évaluation des risques

qui leur sont associés. [2]

Bibliographie

— *Cours de statistique mathématique*, Alain Monfort, Economica (coll. Économie et statistiques avancées), 3e édition, 1997
— *Statistique et économétrie. Du modèle linéaire ... aux modèles non-linéaires*, Xavier Guyon, Ellipses (coll. Universités, mathématiques appliquées), 2001
— *Advanced Econometrics*, Takeshi Amemiya, Harvard University Press, 1985

Le cours fait évidemment appel à des notions et résultats de la théorie des probabilités. Il sera utile de se référer aux ouvrages suivants :

— *Calcul des probabilités*, Dominique Foata et Aimé Fuchs, Dunod, 2003 (2e édition)
— *Cours de probabilités*, Alain Montfort, Economica, 1996 (3e édition)

Dans ce document, il sera beaucoup question de *régression linéaire*. NE PAS lire la page Wikipedia consacrée à ce sujet. [3]

À propos de la lecture de ce document

1. Ce document est un *support* pour le cours offert dans le cursus MASE de Lille 3, et est donc conçu et rédigé pour un public assistant aux cours (même si cela n'empêche quiconque voulant l'utiliser dans un autre cadre de le faire). Ce support est donc destiné à fournir un accompagnement (compléments, présentations alternatives de résultats, exemples, etc) au cours en présentiel, et à ce titre, toute personne y assistant et ayant l'intention de faire du mieux qu'elle peut (notes, compréhension, appropriation des résultats, etc) ne peut éviter sa lecture intégrale. Le rôle de ce support sera d'autant plus efficace que la lecture d'une section interviendra une première fois avant qu'elle soit abordée en présentiel, puis une seconde fois ensuite.

   En résumé :
   — lire
   — par morceaux/sections
   — dans l'ordre
   — avant le cours...
   — ... et après le cours

page:animation

2. Ce document comporte un certain nombre de graphiques animés (constitués de plusieurs images) identifiables par la barre de contrôle située sous le graphique, semblable à ceci :

   Les carrés de cette barre sont des boutons permettant de contrôler l'animation en cliquant dessus. Les symboles représentés sur ces boutons sont ceux couramment utilisés dans tous les dispositifs multimedia. Dans l'ordre de la barre, on retrouve les contrôles suivants : retour à la première image, retour à l'image précédente, lecture inversée, lecture normale, aller à l'image suivante, aller à la dernière image, diminuer la vitesse de lecture, revenir à la vitesse

2. De manière informelle, le risque d'un outil statistique dont le but est d'obtenir de l'information sur les caractéristiques du processus ayant généré les observations désigne le risque d'obtenir une information incorrecte ou trop éloignée des véritables caractéristiques de ce processus.

3. Malgré les bonnes intentions des auteurs de cette page, elle est très confuse et ne constitue ni une bonne introduction grand public au sujet, ni un article pédagogique pouvant appuyer un enseignement.

de lecture normale, augmenter la vitesse de lecture.

Pour visualiser les animations, il est indispensable d'utiliser le lecteur de fichiers PDF Adobe Reader, téléchargeable gratuitement à partir du site d'Adobe.[4] Pour des raisons de sécurité notamment, il est vivement conseillé d'utiliser la version la plus récente de cet outil. Les autres lecteurs de fichiers PDF ne vous permettront pas d'animer les graphiques. Si vous n'avez pas la possibilité de vous procurer ou d'installer Adobe Reader, un lien (http) vers un site affichant une animation pourra vous être proposé.

4. http://get.adobe.com/fr/reader

Table des matières

Chapitre 1

ch:mrls_interpr

Le modèle de régression linéaire simple : présentation

Dans ce chapitre, on étudie un des modèles les plus simples destinés à modéliser et étudier la dépendance entre deux phénomènes dont la mesure s'effectue au moyen de variables notées X et Y.

sec:mr_contexte

1.1 Le contexte et les objectifs

On suppose que deux variables d'intérêt X et Y sont éventuellement liées l'une à l'autre (c'est-à-dire non indépendantes l'une de l'autre). De plus on suppose que la relation éventuelle entre ces variables est orientée : la variable X « explique » la variable Y. Dans le contexte d'un modèle économique, cette hypothèse est courante. En effet, la plupart des modèles économiques distinguent les variables endogènes des variables exogènes : le modèle décrit comment le niveau des premières est déterminé en fonction du niveau des secondes. Notons donc, ainsi que l'exprime leur qualificatif, que le modèle économique ne dit rien sur la façon dont se détermine les niveaux des variables exogènes. On verra comment prendre en compte cette distinction faite au sein des variables dans le contexte d'un modèle économétrique.

Une façon simple de représenter la dépendance de Y envers X consiste à poser une relation linéaire entre les variables : $Y = aX + b$. Dans une représentation de ce type, la caractérisation de la dépendance de la variable Y envers la variable X, c'est à dire la façon dont les variations de X provoquent des variations de Y, est entièrement capturée par la valeur du coefficient a. Il s'agit de proposer un modèle statistique qui permette le même type de modélisation de cette dépendance et qui permette de l'étudier au moyen de techniques d'inférence statistique appropriées. Le modèle le plus simple est le modèle de régression linéaire. Dans un tel modèle, la relation entre les variables X et Y est représentée et caractérisée de manière simple, au moyen d'un petit nombre d'éléments qui constituent les *paramètres* du modèle (semblables à a et b dans l'égalité précédente). Les méthodes d'inférence développées dans le contexte de ce modèle ont pour but d'approximer ces paramètres à partir d'observations des variables X et Y.

sec:mr_construction

1.2 Heuristique de la construction du modèle

Soient X et Y deux variables décrivant chacune un phénomène dans une population. On sélectionne, par un procédé supposé aléatoire [1], n individus de cette population, et pour chacun on introduit le couple de variables mesurant les deux phénomènes étudiés : pour le i^{e} individu, on notera ce couple (X_i, Y_i). En utilisant la convention de notation qui distingue les variables de leurs réalisations, on notera (x_i, y_i) le couple des valeurs observées de X_i et de Y_i.

On souhaite reprendre l'orientation donnée à la relation entre les variables (voir la section précédente). Pour chaque individu i, la variable X_i est supposée déterminer le niveau de la variable Y_i. On appelle alors $X_1, \ldots, X_n$ variables *explicatives* et $Y_1, \ldots, Y_n$ variables *expliquées* ou variables *dépendantes*. Cette distinction sur la nature des variables est en général introduite dans la construction du modèle statistique. Dans dans la version la plus simple du modèle de régression linéaire, on suppose que les variables $X_1, \ldots, X_n$ sont non-aléatoires. Du point de vue statistique, cela revient à dire qu'au sein du modèle économétrique, les variables $X_1, \ldots, X_n$ sont fixes dans le sens où les valeurs prises par ces variables ne sont pas distribuées selon une « véritable » loi de probabilité. [2] Elles ne peuvent par conséquent qu'être simplement égales à leurs observations $x_1, \ldots, x_n$. Avec une telle hypothèse, les variables $X_1, \ldots, X_n$ sont déterminées par leurs observations et aucun autre comportement possible pour ces variables n'est admis. En dehors de ce qui est directement issu de l'observation, le modèle ne permet de déterminer aucune propriété particulière pour les variables $X_1, \ldots, X_n$. On retrouve en cela la notion de variable exogène qui existe dans un modèle économique désignant une variable dont la valeur, ou les propriétés, sont déterminées en dehors du modèle. [3] Dans la suite, on traduira cette hypothèse en utilisant indifféremment dans la notation les observations $x_1, \ldots, x_n$ de ces variables ou bien les variables $X_1, \ldots, X_n$ elles-mêmes.

Avec cette hypothèse, si on veut un modèle statistique qui reprenne l'idée de base de la dépendance linéaire de Y envers X, le modèle pourrait par exemple stipuler qu'il existe des nombres β_0 et β_1 tels que la relation

$$Y_i = \beta_0 + \beta_1 x_i \tag{1.1}$$

eq:rel_incomp

est vraie pour tout individu i. Les nombres β_0 et β_1 sont donc les paramètres du modèle qui permettent de caractériser la dépendance qui existe pour chaque individu i entre x_i et Y_i. L'hypothèse exprimée par la formulation (1.1) conduit immédiatement à un certain nombre de commentaires.

Puisque le terme de droite de l'égalité $Y_i = \beta_0 + \beta_1 x_i$ est fixe, il est clair que celui de gauche doit l'être aussi. En suivant le même raisonnement que celui que nous avons tenu pour les variables

1. Contrairement à ce qui a été présenté dans les rappels du chapitre 10, on n'a pas besoin ici de supposer que la sélection se fait par échantillonnage aléatoire simple.

2. On peut toujours considérer un nombre réel z comme une variable aléatoire Z en lui attribuant comme loi de probabilité $\mathrm{P}(Z = z) = 1$. Dans ce cas, la loi de Z n'est pas une « véritable » loi de probabilité. On dit plutôt que Z a une loi de probabilité dégénérée. Une variable aléatoire a une loi de probabilité dégénérée s'il existe un nombre réel r tel que la probabilité pour que la variable soit égale à r vaut 1. Du point de vue de leurs propriétés statistiques, de telles « variables » peuvent être considérées comme constantes. En ce sens, ce ne sont pas de « véritables » variables aléatoires.

3. Il est possible, en faisant appel à la notion probabiliste de conditionnement, d'écrire un modèle statistique dans lequel les variables $X_1, \ldots, X_n$ sont des variables aléatoires, mais dans lequel l'utilisation des méthodes d'inférence conduira à des résultats ayant la même interprétation et le même usage que ceux que nous dériverons dans le contexte plus simple utilisé ici.

$X_1, \ldots, X_n$, les variables $Y_1, \ldots, Y_n$ doivent dans ce cas avoir une distribution dégénérée et ne peuvent donc être égales à autre chose que leurs observations. Le modèle devrait alors stipuler qu'il existe des nombres β_0 et β_1 tels que $y_i = \beta_0 + \beta_1 x_i$, $\forall i = 1, \ldots, n$. L'objectif consistant à trouver des approximations des paramètres β_0 et β_1 peut être atteint d'une manière très simple, puisqu'il suffit en effet d'utiliser (par exemple) les 2 premières observations (x_1, y_1) et (x_2, y_2) pour déduire la valeur de β_0 et de β_1.

Cependant, dans quasiment toutes les situations rencontrées en pratique, on constaterait que les approximations, qu'on peut par exemple noter β_0^* et β_1^*, ainsi obtenues pour les deux paramètres ne permettent pas d'avoir l'égalité $y_i = \beta_0^* + \beta_1^* x_i$ pour tout $i = 1, \ldots, n$. De manière plus générale, il n'existe quasiment jamais de nombres β_0 et β_1 tels que $y_i = \beta_0 + \beta_1 x_i$, $\forall i = 1, \ldots, n$.

Un exemple simple permet d'en illustrer la raison. Considérons le cas d'une étude statistique dans laquelle les n individus sont des employés d'une chaîne de supermarchés occupant des postes similaires. Pour un individu i, X_i désigne l'ancienneté dans l'emploi (exprimée en mois) et Y_i le salaire mensuel de cet individu. Si on adopte l'hypothèse que pour tout individu i on a $y_i = \beta_0 + \beta_1 x_i$, alors tous les individus ayant le même nombre de mois d'ancienneté doivent nécessairement avoir le *exactement le même* salaire mensuel. Or dans la réalité, cela n'est jamais le cas. Si à ancienneté égale, des individus peuvent avoir des salaires qui diffèrent, cela revient à dire que d'autres facteurs que l'ancienneté peuvent avoir un effet dans la détermination du niveau du salaire. Dans ce sens, les n relations exprimées par (1.1) sont incomplètes (et ne peuvent représenter le phénomène observé entre l'ancienneté et le salaire pour tous les individus).

D'une manière générale, même si on souhaite modéliser une relation de forme linéaire entre une variable explicative et une variable expliquée en retenant une formulation semblable à (1.1), il faut incorporer dans la modélisation retenue le fait que le niveau de la variable expliquée n'est pas exclusivement déterminé par celui de la variable explicative. Une façon simple de compléter chacune de ces relations consiste à introduire des termes notés $\varepsilon_1, \ldots, \varepsilon_n$ de manière que

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \forall i = 1, \ldots, n \tag{1.2}$$

eq:rel_base

L'introduction de ces termes donne lieu à l'interprétation suivante. Pour tout individu i, le niveau de la variable expliquée Y_i se décompose additivement en deux termes :

page:interpr

it:interpr

1. Le terme $\beta_0 + \beta_1 x_i$ qui traduit l'idée de départ d'une relation linéaire dans laquelle la variable exogène explique la variable endogène ; dans la décomposition de Y_i, ce terme est donc la part du niveau de Y_i qui est déterminée par la valeur x_i de la variable explicative X_i.

page:eps

2. Le terme ε_i qui traduit le fait que la valeur de X_i ne détermine pas à elle seule le niveau de la variable dépendante Y_i ; dans la décomposition de Y_i, ce terme est donc la part du niveau de Y_i qui est déterminée par d'autres facteurs que la variable explicative X_i.

Avec la formulation (1.2) et les interprétations données ci-dessus, on a traduit l'idée d'une relation linéaire dans laquelle une variable en détermine une autre, tout en laissant la possibilité à des facteurs autres que la variable explicative d'avoir un effet sur le niveau de la variable expliquée. Il reste cependant à trouver un moyen de formuler l'idée que la variable explicative joue un rôle prépondérant dans la détermination de la variable expliquée, et que les autres facteurs dont on reconnaît l'existence n'ont qu'un impact négligeable sur cette dernière, et dont l'intérêt reste accessoire.

Le fait que l'on ne s'intéresse pas à l'impact qu'ont ces facteurs dans la détermination du niveau de la variable expliquée est traduit par le fait que la façon dont cet impact s'exerce n'est pas modélisé, contrairement à ce qui est fait pour décrire le rôle de la variable explicative. Plus précisément, dans une relation telle que (1.2), on ne cherche ni à identifier ce que sont ces autres facteurs, ni à mesurer chacun d'entre eux au moyen de variables. De plus, la manière dont Y_i dépendrait de ces autres facteurs n'est pas explicitement modélisée. Cela est à contraster avec le statut de la variable explicative, dont (1) on donne la définition et la signification en tant que variable, et (2) dont on stipule la façon dont elle peut affecter le niveau de la variable expliquée (l'effet de X_i sur Y_i est traduit par le terme $\beta_0 + \beta_1 X_i$).

Le fait que l'impact de ces facteurs sur la variable dépendante puisse être négligé est traduit par une nouvelle hypothèse. On supposera par la suite que pour tout individu i, en observant que $X_i = x_i$ on peut s'attendre à ce que la valeur de la variable expliquée Y_i soit $\beta_0 + \beta_1 x_i$. Cette hypothèse signifie que les facteurs autres que la variable explicative ne contribuent en rien à la valeur à laquelle on s'attend pour la variable expliquée.

Si on reprend l'exemple de la relation entre l'ancienneté dans l'emploi et le salaire, ce type d'hypothèse revient à supposer que si deux individus ont une ancienneté identique, notée x, alors on peut s'attendre à ce leurs salaires soient égaux, bien que ceux qui seront *observés* ne le soient pas nécessairement. La valeur commune attendue pour ces deux salaires est $\beta_0 + \beta_1 x$.

Il reste donc à formuler mathématiquement au sein d'un modèle statistique formellement défini, et qui servira de cadre à l'inférence menée sur les paramètres β_0 et β_1, l'ensemble des hypothèses et interprétations formulées ci-dessus.

1.3 Définition et interprétations du modèle de régression linéaire simple

sec:mreg_def

1.3.1 Définition

def:mrls1v

Définition 1.1 *Soient $(X_1, Y_1), \ldots, (X_n, Y_n)$ n couples de variables aléatoires dont les observations sont notées $(x_1, y_1), \ldots, (x_n, y_n)$. Le modèle de régression linéaire simple de Y sur X est un modèle statistique dans lequel les conditions suivantes sont satisfaites*

C1. *Les variables $X_1, \ldots, X_n$ ont une loi dégénérée :* $\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = 1$

C2. *Pour tout $i = 1, \ldots, n$ on peut écrire l'espérance de Y_i comme une fonction affine de x_i :*

$$\exists \beta_0 \in \mathbb{R},\ \exists \beta_1 \in \mathbb{R},\ \mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i, \qquad \forall i = 1, \ldots, n$$

C3. *Pour toute paire (i, j) d'éléments de $\{1, \ldots, n\}$, il existe un réel strictement positif σ tel que*

$$\mathrm{cov}(Y_i, Y_j) = \begin{cases} 0 \text{ si } i \neq j \\ \sigma^2 \text{ si } i = j \end{cases}$$

rem:vraie_loi

Remarque 1.1 Le modèle de régression linéaire simple consiste en l'ensemble des lois de probabilité possibles pour $((X_1, Y_1), \ldots, (X_n, Y_n))$ telles que les conditions exprimées par les conditions C1,

C2 et C3 sont vérifiées. Pour développer des méthodes d'inférence dans le contexte de ce modèle, on supposera que celui-ci est *bien spécifié*, c'est-à-dire que la loi de probabilité dont est issu le $2n$-uplet $((X_1, Y_1), \ldots, (X_n, Y_n))$ de variables aléatoires est bien l'une de lois du modèle. Cette loi de probabilité est désignée par le terme *vraie loi*, dans le sens où parmi toutes les lois constituant le modèle, c'est celle qui décrit la distribution de probabilité des variables aléatoires dont on observera les réalisations.

Par ailleurs, pour n'importe quelle loi de probabilité du modèle, la condition C2 implique que connaissant x_i, on peut écrire l'espérance de Y_i comme une fonction affine de x_i. Sous l'hypothèse que le modèle est bien spécifié, ceci est aussi vrai en particulier pour la vraie loi. Dans ce cas, les nombres qui permettent d'écrire $\mathrm{E}(Y_i)$ comme une fonction affine de x_i sont notés $\overline{\beta}_0$ et $\overline{\beta}_1$. On appelle ces nombres *vraies valeurs* des paramètres β_0 et β_1. Ces vraies valeurs sont inconnues et le modèle défini ci-dessus constitue le cadre dans lequel seront développées des méthodes d'inférence statistique permettant d'estimer ces vraies valeurs. R

La définition ci-dessus admet une définition équivalente, qui formalise la relation (1.2) ainsi que les remarques qu'elle a suscitées.

Propriété 1.1 *Soient $(X_1, Y_1), ..., (X_n, Y_n)$ n couples de variables aléatoires dont les observations sont $(x_1, y_1), ..., (x_n, y_n)$. On définit les n variables aléatoires $\varepsilon_1, \ldots, \varepsilon_n$ par $\varepsilon_i \equiv Y_i - \mathrm{E}(Y_i)$, $i = 1, \ldots, n$. Les conditions* C1 *à* C3 *sont satisfaites si et seulement si les conditions suivantes le sont aussi*

C′1. *La condition* C1 *est satisfaite*

C′2. $\exists \beta_0 \in \mathbb{R}, \exists \beta_1 \in \mathbb{R},\ Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$

C′3. $\exists \sigma \in]0, +\infty[,$

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 \ si\ i \neq j \\ \sigma^2 \ si\ i = j \end{cases} \qquad \forall i, j = 1, \ldots, n$$

La preuve de cette proposition est obtenue à partir de la définition des variables $\varepsilon_1, \ldots, \varepsilon_n$ et de l'égalité suivante, obtenue en supposant C1 ou C′1 vraie :

page:cov

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \mathrm{cov}(Y_i, Y_j) \tag{1.3}$$

eq:cov

Il est donc possible de définir indifféremment le modèle de régression linéaire simple par les conditions C1 à C3 ou par les conditions C′1 à C′3. Ces dernières sont plus fréquemment utilisées.

sec:mrag_interpr

1.3.2 Interprétations

La condition C1 formalise le fait que les variables $X_1, \ldots, X_n$ sont les variables explicatives dans la relation entre X et Y, et sont considérées comme fixes (voir les commentaires faits à ce propos dans la section précédente)

Dans la condition C2 l'espérance de Y_i peut s'interpréter comme la valeur attendue de Y_i. Par conséquent, l'égalité exprimée dans C2 indique que pour chaque individu i, en observant le niveau x_i de sa variable explicative, on peut déduire la valeur attendue de sa variable dépendante, donnée

par $\beta_0 + \beta_1 x_i$. Cette valeur attendue de Y_i est donc une fonction linéaire de la variable explicative. Il est important de noter que les deux nombres qui définissent cette relation sont les mêmes pour tous les individus.

C3 est une condition qui n'est pas fondamentale dans la modélisation : elle ne capture aucun des éléments qui ont motivé la construction du modèle, décrits dans la section précédente. Cette condition permet, tout en préservant les caractéristiques essentielles de ce modèle, d'en proposer une version très simple sur le plan statistique. De ce point de vue, la condition $\text{cov}(Y_i, Y_j) = 0$ si $i \neq j$ indique que les variables expliquées relatives à deux individus distincts sont des variables aléatoires non-corrélées. L'absence de corrélation entre deux variables équivaut à l'absence de toute dépendance de forme linéaire entre ces variables.

Par ailleurs, la condition $\text{cov}(Y_i, Y_i) = \sigma^2 \ \forall i = 1, \ldots, n$, qui équivaut évidemment à $\text{V}(Y_i) = \sigma^2$ $\forall i = 1, \ldots, n$, impose aux variances des n variables aléatoires $Y_1, \ldots, Y_n$ d'être *identiques.*[4] Cette propriété est appelée *homoscédasticité.*

Les termes $\varepsilon_1, \ldots, \varepsilon_n$ ont la même interprétation que celle qui en a été donnée dans la section précédente (voir le point 2 à la page 13). En utilisant la définition de ces termes et la condition C1, on voit que $\text{E}(\varepsilon_i) = 0, \forall i = 1, \ldots, n$, ce qui traduit les remarques qui ont été faites précédemment. Dans la condition C′2 on reconnaît que des facteurs distincts de la variable explicative X_i peuvent affecter le niveau de la variable dépendante Y_i. Ces facteurs sont mesurés par la variable ε_i. Cependant, on s'attend à ce que, compte tenu du niveau de la variable explicative, ces facteurs ne jouent aucun rôle dans la détermination de Y_i : la valeur attendue de ε_i est nulle, c'est-à-dire $\text{E}(\varepsilon_i) = 0$.

On appelle la variable aléatoire ε_i *terme d'erreur* associé à (x_i, Y_i) ; on notera e_i la réalisation de cette variable. Cette terminologie traduit le fait que dans le modèle de régression linéaire simple, si connaissant x_i on essaie de prévoir la valeur de Y_i, la prévision serait $\text{E}(Y_i)$, c'est-à-dire $\beta_0 + \beta_1 x_i$.[5] Par conséquent, l'erreur de prévision qui est faite est $Y_i - \text{E}(Y_i)$, c'est-à-dire ε_i. Ce terme apparaît donc ici comme un terme d'erreur. On note alors que la propriété $\text{E}(\varepsilon_i) = 0$ équivaut à ce qu'on s'attende à ne pas faire d'erreur de prévision.

Il est à noter que contrairement aux variables explicatives et expliquées, on ne dispose pas des observations de $\varepsilon_1, \ldots, \varepsilon_n$. Comme on l'a déjà mentionné dans la section précédente, les termes d'erreur sont destinés à capturer l'effet de tous les facteurs qui en dehors de la variable explicative, peuvent avoir un impact sur le niveau de la variable dépendante. Cependant, la modélisation retenue n'identifie pas explicitement ces facteurs et on n'introduit pas de variables bien définies, et bien identifiées dans la pratique, permettant de les mesurer. La variable ε_i n'est pas définie par autre chose que $\varepsilon_i = Y_i - \text{E}(Y_i)$. Compte tenu de cela et de la condition C2 qui impose $\text{E}(Y_i) = \beta_0 + \beta_1 x_i$, on voit que pour connaître la valeur e_i prise par ε_i, il n'y a d'autre moyen que d'utiliser la formule $e_i = y_i - \overline{\beta}_0 - \overline{\beta}_1 x_i$. Or le membre de droite ne peut être connu puisque les vraies valeurs $\overline{\beta}_0$ et $\overline{\beta}_1$ des paramètres sont inconnues.

Supposons momentanément que nous observons les réalisations de $\varepsilon_1, \ldots, \varepsilon_n$. Puisque par définition, on a $e_i = y_i - \overline{\beta}_0 - \overline{\beta}_1 x_i$ pour tout individu i et qu'on observe évidemment x_i et y_i pour tout i, on pourrait par simple résolution d'un système linéaire de n équations (une pour chaque individu)

4. De plus, cette même condition impose à ces variances d'exister. Même si ce problème ne sera pas abordé par la suite, l'hypothèse d'existence des variances a une importance dans le traitement statistique du modèle défini ci-dessus.

5. On assimile ici la prévision à la valeur attendue. Il est possible de justifier cela sur le plan théorique.

à 2 inconnues ($\overline{\beta}_0$ et $\overline{\beta}_1$) déduire la valeur des paramètres du modèle. Dans ce cas, la construction du modèle de régression linéaire et les méthodes statistiques qui lui sont associées n'ont plus de raison d'être. Ce qui suit n'a donc d'intérêt qu'en supposant que $e_1, \ldots, e_n$ sont inconnues.

On rappelle qu'on peut interpréter la relation $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ comme une décomposition de Y_i en « partie expliquée par x_i » + « partie non expliquée par x_i », la première étant $\beta_0 + \beta_1 x_i$ et la seconde ε_i. Intuitivement, la capacité de la variable explicative à expliquer la variable dépendante sera d'autant meilleure que l'écart entre Y_i et $\beta_0 + \beta_1 x_i$ a tendance à être petit. Si on mesure cet écart par $\left[Y_i - (\beta_0 + \beta_1 x_i)\right]^2$, la valeur attendue est $\mathrm{E}\left[(Y_i - \beta_0 - \beta_1 x_i)^2\right] = \mathrm{E}\left[\left(Y_i - \mathrm{E}(Y_i)\right)^2\right] = \mathrm{V}(Y_i) = \sigma^2$. Cela permet donc d'interpréter le paramètre σ comme une mesure de la capacité de la variable explicative à plus ou moins bien expliquer à elle seule le niveau de la variable expliquée.

Le paramètre β_0 s'interprète comme la valeur attendue de Y_i lorsque $x_i = 0$. On appelle ce paramètre *intercept*, ou ordonnée à l'origine, pour une raison exposée ci-dessous. Le paramètre β_1 a plusieurs interprétations possibles et équivalentes.

— Considérons deux individus statistiques i et j et supposons que l'on observe x_i et x_j de sorte que $x_j = x_i + 1$. On aura alors $\mathrm{E}(Y_j) - \mathrm{E}(Y_i) = \beta_0 + \beta_1(x_i + 1) - \beta_0 - \beta_1 x_i = \beta_1$. On interprète donc β_1 comme la différence entre la valeur attendue de la variable expliquée pour un individu quelconque i et la valeur attendue de cette même variable pour un individu j ayant un niveau de la variable explicative d'une unité supérieur à celui de cette variable pour l'individu i.

— Si on considère la fonction affine qui exprime la valeur de $\mathrm{E}(Y_i)$ en fonction de x_i (voir la condition C2), on a

$$\frac{\mathrm{dE}(Y_i)}{\mathrm{d}x_i} = \beta_1.$$

Par conséquent, si la variable explicative x_i augmente de Δ unités, la variation attendue de la variable dépendante sera de $\beta_1 \Delta$ unités. β_1 est appelé la *pente* du modèle.

Cette dernière interprétation fait clairement apparaître β_1 comme le paramètre d'intérêt dans ce modèle. Étant donnée la forme affine exprimant la relation entre la variable explicative et la variable expliquée, le paramètre β_1 capture à lui seul toute la dépendance de Y_i envers x_i. Les techniques d'inférence développées dans le cadre du modèle de régression linéaire simple auront pour objet β_1.

Pour terminer ce chapitre, on peut à l'aide d'un graphique représenter la manière dont le modèle de régression linéaire simple modélise la relation entre les variables et comment cette modélisation se positionne par rapport à ce qu'on observe. Pour cela, on commence à placer les observations en faisant figurer dans le plan les points de coordonnées (x_i, y_i) pour $i = 1, \ldots, n$. Cette représentation des observations des variables est appelée le *nuage de points*. On introduit ensuite la modélisation du modèle de régression linéaire simple. Celui-ci est construit en posant comme condition qu'il existe deux réels, dont les valeurs inconnues sont $\overline{\beta}_0$ et $\overline{\beta}_1$, qui permettent le lier la variable explicative (exogène) à la variable dépendante (endogène), par la relation $Y_i = \overline{\beta}_0 + \overline{\beta}_1 X_i + \varepsilon_i$, $i = 1, \ldots, n$. Si cela est le cas, les observations $(x_1, y_1), \ldots, (x_n, y_n)$ des variables du modèle, ainsi que les réalisations (non-observées) $e_1, \ldots, e_n$ des termes d'erreur doivent satisfaire $y_i = \overline{\beta}_0 + \overline{\beta}_1 x_i + e_i$, $i = 1, \ldots, n$. Cette relation entre variable exogène et variable endogène est représentée graphiquement par une droite d'équation $y = \overline{\beta}_0 + \overline{\beta}_1 x$. Pour chaque individu $i = 1, \ldots,$ "l'erreur" entre la droite issue du modèle et la réalité observée est $e_i = y_i - (\overline{\beta}_0 + \overline{\beta}_1 x_i)$, et se lit graphiquement comme la différence verticale entre y_i et la droite. Cette construction donne lieu à la figure 1.1.

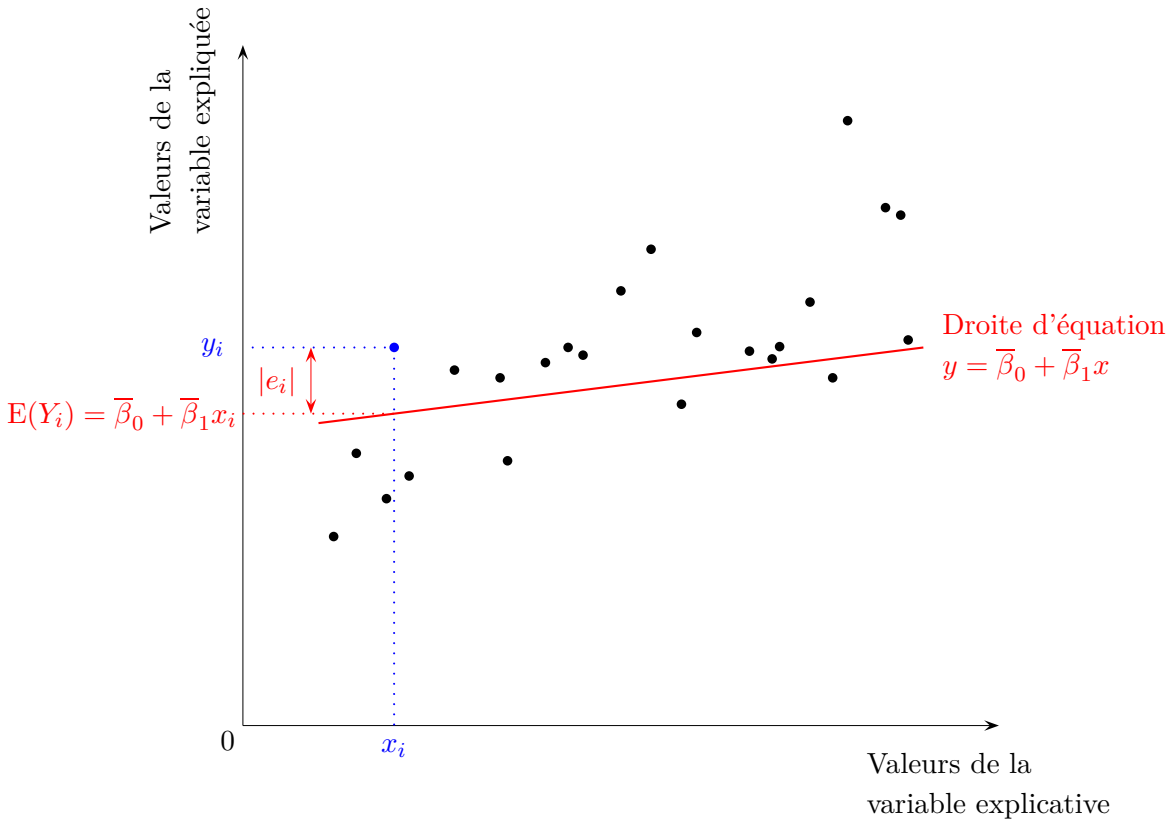


fig:g1

FIGURE 1.1: Modélisation de la relation entre variables dans le modèle de régression linéaire simple

Finalement, le tableau 1.1 de la page 20 récapitule les différents éléments du modèle introduits jusqu'à présent, en séparant ce qui est du domaine des variables du modèle de ce qui est du domaine des paramètre. Pour chaque élément, on rappelle la notation utilisée, l'interprétation qui en est faite

ainsi que les principales hypothèses qui sont formulées à son propos.

TABLE 1.1: Récapitulatif des éléments constitutifs du modèle de régression linéaire simple

VARIABLES

Notation	*Interprétation*	*Dénomination*	*Observations*	*Hypothèses*
Y_i	Variable aléatoire mesurant le phénomène à expliquer pour l'individu i	Variable expliquée, dépendante, endogène	y_i	Son espérance est une fonction affine de x_i. Toutes ces variables sont non-corrélées et ont la même variance.
X_i	Variable aléatoire mesurant la phénomène expliquant Y_i	Variable explicative, exogène	x_i	Considérée comme dégénérée : $\mathrm{P}(X_i = x_i) = 1$
ε_i	Variable aléatoire mesurant la partie Y_i qui ne peut être expliquée par X_i	Terme d'erreur associé à (x_i, Y_i)	N'est pas observée. La réalisation (non observée) de ε_i est notée e_i.	$\varepsilon_i = Y_i - \mathrm{E}(Y_i)$ Son espérance est nulle. Toutes ces variables sont non-corrélées et ont la même variance.

PARAMÈTRES

Notation	*Interprétation*	*Dénomination*	*Commentaires*
β_0	Valeur attendue de Y_i lorsqu'on observe $X_i = 0$	Ordonnée à l'origine, *intercept*	Sa vraie valeur est inconnue ; on la note $\overline{\beta}_0$.
β_1	Variation attendue de Y_i lorsque x_i augmente d'une unité	Pente	C'est le paramètre d'intérêt, qui capture entièrement la dépendance de la variable endogène envers la variable exogène. Sa vraie valeur est inconnue ; on la note $\overline{\beta}_1$.
σ	Écart-type commun des variables dépendantes	—	C'est également l'écart-type commun des termes d'erreur. Sa vraie valeur est inconnue ; on la note $\overline{\sigma}$.

RELATIONS DÉCOULANT DE LA DÉFINITION DU MODÈLE

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	$y_i = \overline{\beta}_0 + \overline{\beta}_1 x_i + e_i$	$\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$

Chapitre 2

ch:mrls_univ

Le modèle de régression linéaire simple : estimation des paramètres

Le modèle statistique défini dans la section précédente est notamment construit dans le but de fournir un cadre à des méthodes d'inférence permettant d'estimer les paramètres β_0 et β_1. En suivant le principe décrit dans le chapitre 10, on cherchera dans cette section à dégager une façon adéquate d'utiliser les variables $X_1, \ldots, X_n, Y_1, \ldots Y_n$ en vue de former un estimateur ponctuel des paramètres. L'utilisation de cet estimateur et des observations fournira l'estimation de la vraie valeur de ces paramètres.

sec:mcoint

2.1 Approche intuitive

Dans une première approche du problème, on cherche des valeurs des paramètres pour lesquelles la partie de $Y_1, \ldots, Y_n$ qui n'est pas expliquée par $X_1, \ldots, X_n$ est la plus petite possible en moyenne. Pour cela, on choisit des valeurs de β_0 et de β_1 pour lesquelles l'écart moyen entre les Y_i et les $\beta_0 + \beta_1 X_i$ est minimale.

Formellement, à tout choix d'un couple de réels (β_0, β_1) on associe les écarts entre les Y_i et les $\beta_0 + \beta_1 X_i$, qu'on note $S_i(\beta_0, \beta_1)$ et qu'on mesure par $S_i(\beta_0, \beta_1) = \left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2$, $i = 1, \ldots, n$. Choisir les valeurs β_0 et β_1 pour lesquelles la moyenne des écarts est la plus petite revient à minimiser la fonction S définie par

$$\begin{aligned} S: \quad & \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}_+ \\ & (\beta_0, \beta_1) \longmapsto S(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \end{aligned}$$

Tout estimateur de β_0, β_1 obtenu de cette manière est appelé estimateur des moindres carrés ordinaires (MCO). L'appellation « moindres carrés » provient du fait que les estimateurs sont obtenus en minimisant une moyenne de carrés $(Y_i - \beta_0 - \beta_1 X_i)^2$, $i = 1, \ldots, n$. Ces moindres carrés sont « ordinaires » car la moyenne des carrés est une moyenne ordinaires, *i.e.*, dans laquelle tous les carrés ont le même poids.

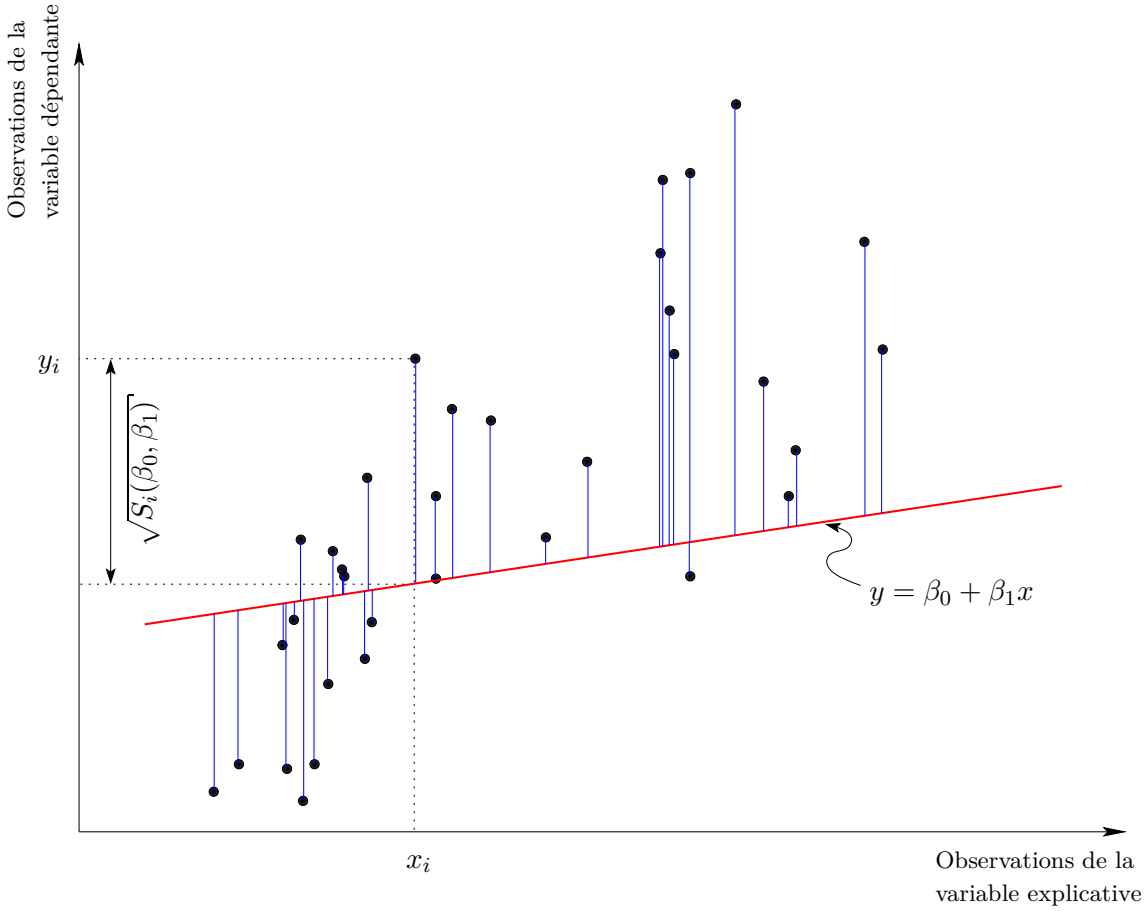


fig:minS_interpr

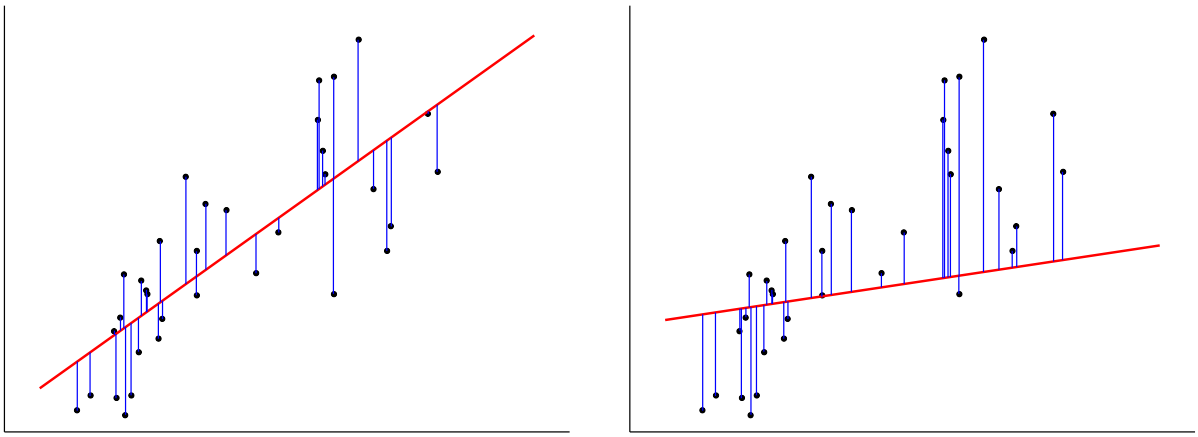
FIGURE 2.1: Interprétation graphique de la fonction S

Avant de considérer la résolution de ce problème de minimisation, on peut illustrer graphiquement l'approche suivie ici. On utilise le nuage de points qui représente dans le plan les points de coordonnées (x_i, y_i), $i = 1, \ldots, n$ (voir la figure 1.1 de la section 1.3.2). À tout couple de réels (β_0, β_1), on associe une droite d'équation $y = \beta_0 + \beta_1 x$. On peut représenter cette droite dans le même plan que celui utilisé pour le nuage de points. Au couple (β_0, β_1) choisi, et donc à la droite correspondante, on peut associer les écarts $S_1(\beta_0, \beta_1), \ldots, S_n(\beta_0, \beta_1)$ définies ci-dessus. On peut représenter sur le même graphique ces quantités. $S_i(\beta_0, \beta_1)$ est le carré de la distance $|Y_i - (\beta_0 + \beta_1 x_i)|$ entre Y_i et $\beta_0 + \beta_1 x_i$. Sur le graphique, l'écart $S_i(\beta_0, \beta_1)$ est donc le carré de la distance verticale entre Y_i et la droite d'équation $y = \beta_0 + \beta_1 x$. La figure 2.1 représente le nuage de points et, pour un couple (β_0, β_1) donné, la droite associée (en rouge) ainsi que les distances verticales entre les points du nuage et la droite (symbolisées par les barres bleues verticales).

page:minSmco

Choisir le couple (β_0, β_1) de façon à minimiser la fonction S revient d'une certaine manière à choisir la droite pour laquelle les distances verticales entre cette droite et les points du nuage sont les plus petites en moyenne. Dans ce sens, minimiser S consiste à cherche la droite qui passe au plus près des points du nuage. La figure 2.2 montre pour deux choix possibles du couple (β_0, β_1) les droites et les distances qui en résultent. On constate ainsi que sur le graphique 2.2 (a) pour lequel on a choisi $(\beta_0, \beta_1) = (16.43, 0.47)$, les distances associées à la droite sont en moyenne plus petites que sur le graphique d'à côté, construit en choisissant $(\beta_0, \beta_1) = (25.02, 0.10)$. La valeur de

la fonction S associée au graphique 2.2 (a) sera donc plus petite que celle associée au graphique 2.2 (b).



(a) $(\beta_0, \beta_1) = (16.43, 0.47)$ gauche

droite (b) $(\beta_0, \beta_1) = (25.02, 0.10)$

fig:minS_compar FIGURE 2.2: Droites et écarts $S_1(\beta_0, \beta_1), \ldots, S_n(\beta_0, \beta_1)$ associés à différents choix de (β_0, β_1)

On aborde à présent la résolution du problème de minimisation de la fonction S. On doit donc résoudre

$$\min_{(\beta_0,\beta_1)\in\mathbb{R}^2} S(\beta_0, \beta_1) \tag{2.1}$$ eq:mco

S est une fonction de (β_0, β_1) deux fois continûment dérivable. De plus c'est une fonction convexe. Par conséquent, tout extremum est un minimum, atteint en tout point $(\hat{\beta}_0, \hat{\beta}_1)$ de $\mathbb{R}^2$ satisfaisant

$$\frac{\partial S}{\partial \beta_k}(\hat{\beta}_0, \hat{\beta}_1) = 0, \quad k = 0, 1. \tag{2.2}$$ eq:cpomco

Dans la minimisation de S, il est important de distinguer deux situations.

1. S'il existe deux individus i et j pour lesquels $X_i \neq X_j$, alors S est strictement convexe, puisque c'est la somme de n fonctions strictement convexes de (β_0, β_1) : $S(\beta_0, \beta_1) = \sum_{i=1}^n S_i(\beta_0, \beta_1)$ où $S_i(\beta_0, \beta_1) = (Y_i - \beta_0 - \beta_1 X_i)^2$. L'allure de la fonction S est représentée par la figure 2.3 (sur laquelle les couleurs sur la surface varient en fonction de la valeur atteinte par S, de manière similaire à une carte géographique d'un relief).

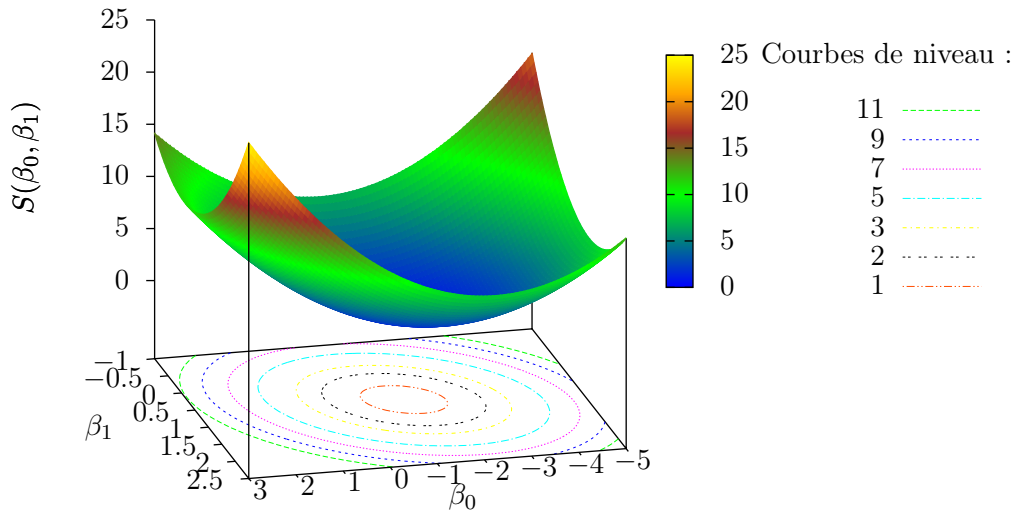


FIGURE 2.3: Allure de la fonction S (cas 1).

Cette fonction admet un unique minimum au point $(\hat{\beta}_0, \hat{\beta}_1)$, entièrement caractérisé par le système de deux équations (2.2). Comme la fonction S est un polynôme du second degré en chacun de ses arguments β_0 et β_1, chaque équation de (2.2) est un polynôme du premier degré en chacun des arguments. Autrement dit, pour trouver le minimand $(\hat{\beta}_0, \hat{\beta}_1)$ de la fonction S il suffit donc de résoudre un système de deux équations linéaires à deux inconnues.

Notons que

$$\frac{\partial S}{\partial \beta_k}(\beta_0, \beta_1) = \frac{\partial}{\partial \beta_k} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k} (Y_i - \beta_0 - \beta_1 X_i)^2, \quad k = 0, 1.$$

Par conséquent

$$\frac{\partial S}{\partial \beta_0}(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} -2(Y_i - \beta_0 - \beta_1 X_i) \tag{2.3}$$

$$\frac{\partial S}{\partial \beta_1}(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} -2X_i(Y_i - \beta_0 - \beta_1 X_i). \tag{2.4}$$

Le système (2.2) s'écrit donc

$$\begin{cases} \frac{1}{n} \sum_{i=1}^{n} (-2)(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \frac{1}{n} \sum_{i=1}^{n} (-2) X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \end{cases} \iff \begin{cases} \frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} X_i = 0 \\ \frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \frac{1}{n} \sum_{i=1}^{n} X_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} X_i^2 = 0 \end{cases}$$

La première équation est équivalente à

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}, \tag{2.5}$$

où $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ et $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. En substistuant cette expression de $\hat{\beta}_0$ dans le seconde équation, on obtient

$$\frac{1}{n}\sum_{i=1}^n X_i Y_i - \overline{X}\,\overline{Y} + \hat{\beta}_1\overline{X}^2 - \hat{\beta}_1\frac{1}{n}\sum_{i=1}^n X_i^2 = 0, \tag{2.6}$$

eq:cpob2

de sorte que si $\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 \neq 0$, on a

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2}$$

Notons que $\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$, de sorte que la condition $\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 \neq 0$ permettant de définir $\hat{\beta}_1$ est $\sum_{i=1}^n (X_i - \overline{X})^2 \neq 0$. Le membre de gauche de cette relation étant une somme à termes positifs, cette somme est nulle si et seulement si chacun de ses termes est nul :

$$\sum_{i=1}^n (X_i - \overline{X})^2 = 0 \iff (X_i - \overline{X})^2 = 0\ \forall i \iff X_i = \overline{X}\ \forall i \iff X_1 = X_2 = \cdots = X_n$$

page:ident

Or ceci est une possibilité qui a été exclue au début de ce premier point. Par conséquent, on a le résultat suivant.

th:ident

Théorème 2.1 *Dans le modèle de régression linéaire simple, s'il existe deux individus i et j tels que $X_i \neq X_j$, alors l'estimateur des moindres carrés ordinaires de (β_0, β_1) est donné par*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2}, \tag{2.7}$$

eq:mcob2

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{2.8}$$

eq:mcob1

Ce théorème est illustré par la figure 2.4.

2. Supposons maintenant que pour tous les individus i on a $X_i = x$. La fonction S est donc définie par $S(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x)^2$. Supposons que S admette un minimum au point $(\hat{\beta}_0, \hat{\beta}_1)$. On aura donc

$$\hat{S} = S(\hat{\beta}_0, \hat{\beta}_1) \leq S(\beta_0, \beta_1), \quad (\beta_0, \beta_1) \in \mathbb{R}^2.$$

Il est facile de voir que pour tout (β_0, β_1) choisi de sorte que $\beta_0 + \beta_1 x = \hat{\beta}_0 + \hat{\beta}_1 x$, la fonction S sera aussi égale à $\hat{S}$. Pour cela, on fixe β_0 arbitrairement et on choisit $\beta_1 = \frac{1}{x}(\hat{\beta}_0 + \hat{\beta}_1 x - \beta_0)$. Il existe dont une infinité de choix possibles et la fonction S admet une infinité (continuum) de minimands.

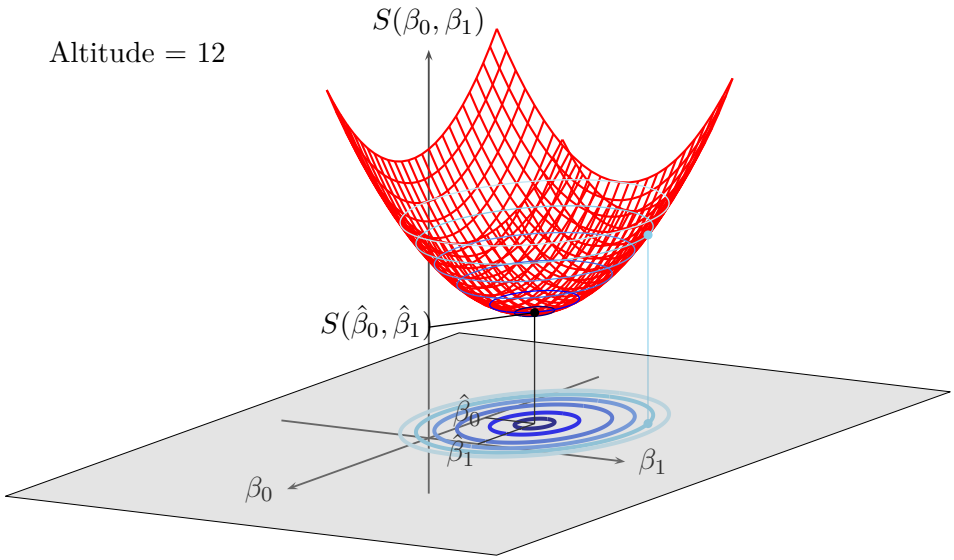
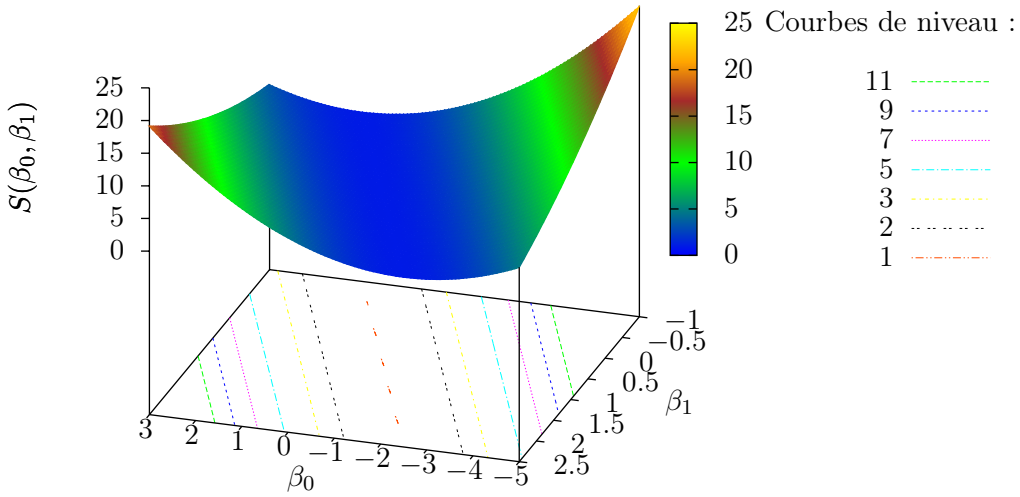


FIGURE 2.4: Les estimateurs des moindres carrés ordinaires sont obtenus en minimisant S (En baissant l'altitude du point de vue, l'animation permet de voir le dessous de la surface et le minimum de la fonction S. Celui-ci vaut ici 1 et est atteint en $(\hat{\beta}_0, \hat{\beta}_1) = (-1, 0.5)$). *Cliquez pour lancer l'animation.*

fig:minscr



pégg:sm

FIGURE 2.5: Allure de la fonction S (cas 2).

On peut caractériser l'ensemble des minimands à l'aide des observations des variables du modèle. Si on examine dans ce cas les expressions (2.3) et (2.4), on a

$$\frac{\partial S}{\partial \beta_1}(\beta_0,\beta_1) = x\frac{\partial S}{\partial \beta_0}(\beta_0,\beta_1)\,, \quad \forall\beta_0, \forall\beta_1.$$

Les deux dérivées partielles de S sont proportionnelles et les deux conditions du premier ordre (2.2), qui demeurent des conditions nécessaires et suffisantes pour que $(\hat\beta_0,\hat\beta_1)$ minimise la fonction S, sont par conséquent redondantes. Elles donnent toutes deux

$$\hat\beta_0 = \overline{Y} - \hat\beta_1 x\,.$$

La solution au problème (2.1) est donc l'ensemble de $\mathbb{R}^2$ défini par $\{(\beta_0,\beta_1)\in\mathbb{R}^2\,|\,\beta_0 = \overline{Y} - x\beta_1\}$. La figure 2.5 illustre cette situation. Les courbes de niveau de la surface de S sont projetées sur le plan ($\mathbb{R}^2$). On voit que la courbe la plus basse (celle de niveau 1), qui indique les lieux (β_0,β_1) où S atteint son minimum, est une droite de $\mathbb{R}^2$: elle coïncide avec l'ensemble des solutions $\{(\beta_0,\beta_1)\in\mathbb{R}^2\,|\,\beta_0 = \overline{Y} - x\beta_1\}$.
On résume ces résultats par le théorème suivant.

th:nident

Théorème 2.2 *Dans le modèle de régression linéaire simple, si pour tous les individus i on a $X_i = x$, alors la solution au problème (2.1) n'est pas unique. Tout élément de $\mathbb{R}^2$ de la forme $(\overline{Y} - x\hat\beta_1, \hat\beta_1)$, où $\hat\beta_1\in\mathbb{R}$, peut être considéré comme un estimateur des moindres carrés ordinaires de (β_0,β_1). On dira dans ce cas que l'estimateur des moindres carrés ordinaires n'existe pas.*

Dans ce dernier cas, il n'est pas possible de distinguer les vraies valeurs des paramètres d'autres valeurs *a priori* possibles pour ces paramètres. En effet, si $X_i = X_j$ pour tout $i, j = 1, \ldots, n$ l'espérance de la variable dépendante sera la même pour tous les individus : $X_i = X_j = x, \forall i, j = 1, \ldots, n \implies \mathrm{E}(Y_i) = \mathrm{E}(Y_j) = \overline{\beta}_0 + \overline{\beta}_1 x$, $\forall i, j = 1, \ldots, n$. Notons m la valeur commune de cette espérance. Si on choisit alors $\beta_1 = b$ et $\beta_0 = m - bx$, on aura également $\beta_0 + \beta_1 x = m = \overline{\beta}_0 + \overline{\beta}_1 x = \mathrm{E}(Y_i)$, $\forall i = 1, \ldots, n$. Autrement dit des valeurs des paramètres différentes des vraies valeurs donnent la même valeur pour l'espérance de la variable endogène et les valeurs des paramètres qui permettent d'écrire la condition C2 ne sont pas uniques. Il est par conséquent impossible de distinguer les vraies valeurs des paramètres d'autres valeurs. On dit dans ce cas que les paramètres β_0 et β_1 du modèle sont *non-identifiés*. Si les vraies valeurs des paramètres sont non-identifiées, il est normal qu'en cherchant à estimer ces paramètres on n'obtienne pas de solution unique.

page:nident

2.2 Approche théorique

sec:mcotheo

Une approche plus théorique consiste à ne considérer que les estimateurs *linéaires* de β_0 et β_1, puis à chercher dans l'ensemble de tels estimateurs ceux qui sont préférables aux autres.

Définition 2.1 *Une statistique $\tilde{\beta}_k$ est un estimateur linéaire de β_k si on peut trouver n nombres $\tilde{w}_{k1}, \ldots, \tilde{w}_{kn}$, pouvant éventuellement dépendre de $X_1, \ldots, X_n$, tels que $\tilde{\beta}_k = \sum_{i=1}^n \tilde{w}_{ki} Y_i$, $k = 0, 1$.*

rem:nuplet

Remarque 2.1 On constate qu'à tout n-uplet de nombres $(w_1, \ldots, w_n)$, on peut associer un estimateur linéaire $\sum_{i=1}^n w_i Y_i$. D'autre part, tout estimateur linéaire $\sum_{i=1}^n w_i Y_i$ est complétement caractérisé par le n-uplet $(w_1, \ldots, w_n)$. Par conséquent, choisir un estimateur linéaire de β_k revient à choisir un n-uplet de réels.

pro:mcob12lin

Propriété 2.1 *Les estimateurs des moindres carrés ordinaires $\hat{\beta}_0$ et $\hat{\beta}_1$ de β_0 et β_1 sont des estimateurs linéaires. Plus précisément, on a $\hat{\beta}_k = \sum_{i=1}^n \hat{w}_{ki} Y_i$, $k = 0, 1$, avec*

$$\hat{w}_{1i} = \frac{X_i - \overline{X}}{\sum_{j=1}^n X_j^2 - n\overline{X}^2} \tag{2.9}$$

eq:linbh1

$$\hat{w}_{0i} = \frac{1}{n} - \overline{X}\hat{w}_{1i} \tag{2.10}$$

eq:linbh0

$i = 1, \ldots, n$.

Preuve : On établit facilement que

$$\sum_{i=1}^n X_i Y_i - n\overline{X}\,\overline{Y} = \sum_{i=1}^n X_i Y_i - \overline{X}\sum_{i=1}^n Y_i = \sum_{i=1}^n (X_i - \overline{X}) Y_i$$

Par conséquent, en utilisant l'expression (2.7) de $\hat{\beta}_1$, on peut écrire

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X}) Y_i}{\sum_{i=1}^n X_i^2 - n\overline{X}^2} = \sum_{i=1}^n \hat{w}_{1i} Y_i$$

En ce qui concerne $\hat{\beta}_0$, on a à partir de (2.8)

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{w}_{1i}\overline{X}Y_i = \sum_{i=1}^{n}\Big(\frac{1}{n} - \overline{X}\hat{w}_{1i}\Big)Y_i = \sum_{i=1}^{n} \hat{w}_{0i}Y_i\,,$$

Un critère de comparaison d'estimateurs est l'erreur quadratique moyenne (EQM).

def:eqm **Définition 2.2** *Pour un estimateur $\tilde{\beta}_k$ de β_k, l'erreur quadratique moyenne (EQM) de $\tilde{\beta}_k$ est la fonction qui au couple (β_0, β_1) associe le nombre* $\mathrm{E}\big[(\tilde{\beta}_k - \beta_k)^2\big]$, $k = 0, 1$.

rem:eqm **Remarque 2.2**

— La définition 2.2 indique explicitement que l'EQM de $\tilde{\beta}_1$ est une fonction de β_1 *et* de β_0. Cependant, il peut sembler au premier abord que $\mathrm{E}\big[(\tilde{\beta}_1 - \beta_1)^2\big]$ ne dépend que de β_1. Cela n'est évidemment pas vrai. En effet, comme $\tilde{\beta}_1$ est un estimateur linéaire de β_1, on a

$$\tilde{\beta}_1 = \sum_{i=1}^{n} \tilde{w}_{1i}Y_i = \sum_{i=1}^{n} \tilde{w}_{1i}(\beta_0 + \beta_1 X_i + \varepsilon_i),$$

pour un certain n-uplet $(\tilde{w}_{11}, \ldots, \tilde{w}_{1n})$. On voit alors clairement que la variable aléatoire $\tilde{\beta}_1$ peut s'écrire non seulement en fonction de β_1, mais aussi en fonction de β_0. Par conséquent, il n'est pas surprenant que la loi de $\tilde{\beta}_1$ dépende à la fois β_0 et de β_1. Cela sera en particulier vrai pour l'EQM $\mathrm{E}\big[(\tilde{\beta}_1 - \beta_1)^2\big]$. La même remarque s'applique évidemment aux estimateurs de β_0.

— L'EQM d'un estimateur $\tilde{\beta}_k$ est une mesure de la précision de cet estimateur, puisque l'EQM s'interprète comme la distance attendue entre un estimateur $(\tilde{\beta}_k)$ et ce qu'il estime (β_k).

— Pour deux estimateurs $\check{\beta}_k$ et $\tilde{\beta}_k$ de β_k, on dit que $\check{\beta}_k$ est préférable à $\tilde{\beta}_k$ au sens de l'erreur quadratique moyenne si l'EQM de $\check{\beta}_k$ est inférieure ou égale à l'EQM de $\tilde{\beta}_k$, ceci *pour toutes les valeurs possibles des paramètres* β_0 et β_1.

— En général, il n'est pas possible de trouver des estimateurs préférables à tout autre au sens de l'EQM.
Cependant, comme on va le montrer, si on introduit un autre type de contrainte sur les estimateurs qu'on considère, alors on pourra trouver, dans le contexte du modèle de régression linéaire simple, un estimateur préférable à tout autre au sens de l'EQM.

La contrainte supplémentaire imposée aux estimateurs est d'être sans biais.

Définition 2.3

1. *Le biais d'un estimateur $\tilde{\beta}_k$ de β_k est la fonction qui à (β_0, β_1) associe le nombre* $\mathrm{E}(\tilde{\beta}_k - \beta_k)$.
2. *On dit qu'un estimateur $\tilde{\beta}_k$ de β_k est sans biais si son biais est constant et égal à 0 :* $\mathrm{E}(\tilde{\beta}_k - \beta_k) = 0$, $\forall \beta_0, \beta_1$.

Remarque 2.3

— Le premier point de la remarque 2.2 faite à propos de l'EQM peut aussi s'appliquer au biais d'un estimateur. Le biais de $\tilde{\beta}_0$ dépend de β_0 *et* de β_1.

— Un estimateur $\tilde{\beta}_k$ de β_k est sans biais si et seulement si $\mathrm{E}(\tilde{\beta}_k) = \beta_k$, $\forall \beta_0, \beta_1$.
— Si $\tilde{\beta}_k$ est un estimateur sans biais de β_k, alors son EQM coïncide avec sa variance. En effet on a dans ce cas
$$\mathrm{E}[(\tilde{\beta}_k - \beta_k)^2] = \mathrm{E}[(\tilde{\beta}_k - \mathrm{E}(\tilde{\beta}_k))^2] = \mathrm{V}(\tilde{\beta}_k).$$
— La variance d'un estimateur sans biais est donc une mesure de sa précision. Plus la variance d'un estimateur sans biais est petite, plus cet estimateur est précis.
— Par conséquent, pour comparer des estimateurs sans biais d'un même paramètre, il suffit de comparer leurs variances. Plus précisément, si $\check{\beta}_k$ et $\tilde{\beta}_k$ sont deux estimateurs sans biais de β_k, on préfèrera $\check{\beta}_k$ à $\tilde{\beta}_k$ au sens de l'EQM si $\mathrm{V}(\check{\beta}_k) \leq \mathrm{V}(\tilde{\beta}_k)$, $\forall \beta_0, \beta_1$.

Dans le modèle de régression linéaire standard, si on ne considère que des estimateurs linéaires et sans biais de β_0 et de β_1, on préfèrera ceux qui sont de variance minimale.

pro:cnsesb **Propriété 2.2** *Dans le modèle de régression linéaire simple, le biais d'un estimateur linéaire $\tilde{\beta}_k$ de β_k, défini par le n-uplet $(\tilde{w}_{k1}, \dots \tilde{w}_{kn})$, est la fonction qui au couple (β_0, β_1) associe le nombre*
$$\beta_0 \sum_{i=1}^{n} \tilde{w}_{ki} + \beta_1 \sum_{i=1}^{n} \tilde{w}_{ki} X_i - \beta_k. \tag{2.11}$$ eq:esb

Preuve : Par définition, le biais de l'estimateur $\tilde{\beta}_k = \sum_{i=1}^{n} \tilde{w}_{ki} Y_i$ de β_k est la fonction qui au couple (β_0, β_1) associe le nombre $\mathrm{E}(\sum_{i=1}^{n} \tilde{w}_{ki} Y_i) - \beta_k$. Comme les $\tilde{w}_{ki}$ ne dépendent que de $X_1, \dots, X_n$, et que ces variables ont une distribution dégénérée (condition C1), il en est de même pour les $\tilde{w}_{ki}$, et on a $\mathrm{E}(\sum_{i=1}^{n} \tilde{w}_{ki} Y_i) = \sum_{i=1}^{n} \tilde{w}_{ki} \mathrm{E}(Y_i)$. D'après la condition C2 du MLRS, on a $\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_i$ et en substituant cette expression dans l'expression du biais, ce dernier s'écrit $\sum_{i=1}^{n} \tilde{w}_{ki}(\beta_0 + \beta_1 X_i) - \beta_k$. En factorisant β_0 et β_1, on obtient (2.11).

Remarque 2.4 On note qu'en utilisant (2.11), la condition pour qu'un estimateur linéaire $\tilde{\beta}_0 = \sum_{i=1}^{n} \tilde{w}_{0i} Y_i$ de β_0 soit sans biais est
$$\beta_0 \sum_{i=1}^{n} \tilde{w}_{0i} + \beta_1 \sum_{i=1}^{n} \tilde{w}_{0i} X_i = \beta_0, \quad \forall \beta_0, \forall \beta_1. \tag{2.12}$$ eq:esbb1

Cette condition est une condition sur le n-uplet de réels $(\tilde{w}_{01}, \dots, \tilde{w}_{0n})$ qui définit $\tilde{\beta}_0$. Cette condition (2.12) s'écrit aussi
$$\beta_0 (\sum_{i=1}^{n} \tilde{w}_{0i} - 1) + \beta_1 \sum_{i=1}^{n} \tilde{w}_{0i} X_i = 0, \quad \forall \beta_0, \forall \beta_1. \tag{2.13}$$ eq:esbb1bis

Les $\tilde{w}_{0i}, \dots, \tilde{w}_{0n}$ satisfont cette condition si et seulement si ils satisfont
$$\begin{cases} \sum_{i=1}^{n} \tilde{w}_{0i} - 1 = 0, \\ \sum_{i=1}^{n} \tilde{w}_{0i} X_i = 0. \end{cases} \tag{2.14}$$ eq:cesbb1

En effet, il est clair que si (2.14) est vérifiée, alors (2.13) l'est aussi. Réciproquement, supposons (2.13) vraie. Alors pour le cas particulier $\beta_0 = 1$ et $\beta_1 = 0$, on doit avoir $\sum_{i=1}^{n} \tilde{w}_{0i} - 1 = 0$. De

même pour $\beta_0 = 0$ et $\beta_1 = 1$, on doit avoir $\sum_{i=1}^n \tilde{w}_{0i} X_i = 0$. Autrement dit, (2.14) est également vraie.

Pour les mêmes raisons, tout estimateur linéaire $\tilde{\beta}_1 = \sum_{i=1}^n \tilde{w}_{1i} Y_i$ de β_1 sera sans biais si et seulement si $\tilde{w}_{11}, \ldots, \tilde{w}_{1n}$ satisfont la condition

$$\begin{cases} \sum_{i=1}^n \tilde{w}_{1i} = 0, \\ \sum_{i=1}^n \tilde{w}_{1i} X_i - 1 = 0. \end{cases} \tag{2.15}$$

eq:cesbb2

pro:mcosansbiais **Propriété 2.3** *Dans le modèle de régression linéaire simple, les estimateurs des moindres carrés ordinaires de β_0 et de β_1 sont sans biais.*

Preuve : On utilise les expressions de $\hat{\beta}_0$ et $\hat{\beta}_1$ données par la propriété 2.1 et on montre que les conditions (2.14) et (2.15) sont satisfaites. Ainsi pour $\hat{\beta}_1$, on a $\hat{w}_{1i} = \frac{X_i - \overline{X}}{\sum_{j=1}^n X_j^2 - n\overline{X}^2}$. Donc

$$\sum_{i=1}^n \hat{w}_{1i} = \frac{\sum_{i=1}^n (X_i - \overline{X})}{\sum_{i=1}^n X_i^2 - n\overline{X}^2} = 0,$$

car le numérateur est nul, et d'autre part

$$\sum_{i=1}^n \hat{w}_{1i} X_i - 1 = \frac{\sum_{i=1}^n (X_i - \overline{X}) X_i}{\sum_{i=1}^n X_i^2 - n\overline{X}^2} - 1 = 0,$$

car le numérateur et le dénominateur du rapport sont identiques. Autrement dit $\hat{\beta}_1$ est un estimateur linéaire qui vérifie les conditions (2.15).
Quant à $\hat{\beta}_0$, on a $\hat{w}_{0i} = \frac{1}{n} - \overline{X}\hat{w}_{1i}$. Par conséquent,

$$\sum_{i=1}^n \hat{w}_{0i} - 1 = \sum_{i=1}^n \frac{1}{n} - \overline{X} \sum_{i=1}^n \hat{w}_{1i} - 1 = 1 - 0 - 1 = 0,$$

puisque on a montré que $\sum_{i=1}^n \hat{w}_{1i} = 0$. D'autre part,

$$\sum_{i=1}^n \hat{w}_{0i} X_i = \sum_{i=1}^n (\frac{1}{n} - \overline{X}\hat{w}_{1i}) X_i = \overline{X} - \overline{X} \sum_{i=1}^n \hat{w}_{1i} X_i = \overline{X} - \overline{X} = 0,$$

car on a montré ci-dessus que $\sum_{i=1}^n \hat{w}_{1i} X_i = 1$. Donc $\hat{\beta}_0$ satisfait les conditions (2.14).

Avant d'énoncer et prouver le résultat central de ce chapitre, nous avons besoin d'établir l'expression de la variance d'estimateurs linéaires de β_0 et β_1.

pro:varlin **Propriété 2.4** *Dans le modèle de régression linéaire simple, si $\tilde{\beta}_k$ est un estimateur linéaire de β_k défini par le n-uplet $(\tilde{w}_{k1}, \ldots, \tilde{w}_{kn})$, alors la variance de $\tilde{\beta}_k$ est donnée par*

$$\mathrm{V}(\tilde{\beta}_k) = \sigma^2 \sum_{i=1}^n \tilde{w}_{ki}^2. \tag{2.16}$$

eq:varlin

Preuve : Puisque $\tilde{\beta}_k$ est linéaire,

$$\mathrm{V}(\tilde{\beta}_k) = \mathrm{V}\Big(\sum_{i=1}^{n} \tilde{w}_{ki} Y_i\Big) = \sum_{i=1}^{n} \mathrm{V}(\tilde{w}_{ki} Y_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{cov}(\tilde{w}_{ki} Y_i, \tilde{w}_{kj} Y_j).$$

En utilisant le fait que les $\tilde{w}_{ki}$ ne dépendent que de $X_1, \ldots, X_n$ et que ces variables ont une distribution de probabilité dégénérée (condition C1), en appliquant les propriétés de la covariance, on a

$$\mathrm{V}(\tilde{\beta}_k) = \sum_{i=1}^{n} \tilde{w}_{ki}^2 \mathrm{V}(Y_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \tilde{w}_{ki} \tilde{w}_{kj} \mathrm{cov}(Y_i, Y_j).$$

La condition C3 définissant le modèle de régression linéaire simple implique que tous les termes de la première somme sont égaux à $\tilde{w}_{ki}^2 \sigma^2$, et que toutes les covariances de la deuxième (double) somme sont nulles. On obtient donc l'expression voulue de $\mathrm{V}(\tilde{\beta}_k)$.

cor:varlin **Corollaire 2.1** *Dans le modèle de régression linéaire simple, les variances des estimateurs des moindres carrés ordinaires $\hat{\beta}_0$ et $\hat{\beta}_1$ de β_0 et de β_1 sont*

$$\mathrm{V}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \right] \quad \textit{et} \quad \mathrm{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

Preuve : Pour $\mathrm{V}(\hat{\beta}_1)$ il suffit d'utiliser la propriété précédente, tandis que pour $\mathrm{V}(\hat{\beta}_0)$, il faut en plus utiliser la première condition de (2.15).

rem:varlin **Remarque 2.5**

— Notons que les variances des estimateurs des moindres carrés ordinaires sont inconnues, puisqu'elles dépendent de σ^2. On peut cependant les estimer (voir la section 2.5).

— La variance de $\hat{\beta}_k$ est une mesure de la distance attendue entre $\hat{\beta}_k$ et β_k. C'est donc un indicateur de la précision de l'estimateur $\hat{\beta}_k$. On constate que cette précision est affectée par deux facteurs.

1. Le premier est la variance commune σ^2 des Y_i. Plus elle est élevée, plus $\mathrm{V}(\hat{\beta}_k)$ est grande. Autrement dit, pour $X_1 = x_1, \ldots, X_n = x_n$ donnés, plus les variables dépendantes ont tendance à être dispersées autour de leur espérance, plus les estimateurs des moindres carrés ordinaires auront tendance à l'être et moins ils seront précis.
2. L'impact du second facteur est capturé par $\sum_{i=1}^{n} X_i^2 - n\overline{X}^2$. On établit facilement que ce terme s'écrit aussi $\sum_{i=1}^{n} (X_i - \overline{X})^2$ et que sous cette forme il capture la variabilité observée des variables $X_1, \ldots, X_n$ autour de leur moyenne. Plus ce terme est élevé, plus $\mathrm{V}(\hat{\beta}_k)$ est petite. Autrement dit, plus on observe de dispersion des valeurs de la variable explicative autour de sa valeur moyenne, plus les estimateurs des moindres carrés ordinaires seront précis.

   On rappelle que β_1 mesure la dépendance entre la variable expliquée et la variable explicative. Pour estimer précisément cette dépendance, on a besoin d'un échantillon dans lequel ces variations sont suffisamment élevées. En effet, si on observe peu de variabilité pour la variable explicative, on a peu d'observations sur le phénomène qu'on cherche à

représenter et à estimer, à savoir la réponse de la variable expliquée aux variations de la variable explicative.

Dans le cas limite où $X_1 = \cdots = X_n$, on n'observe aucune variation de la variable explicative, et on n'a donc aucune information sur la manière dont la variable expliquée pourrait répondre aux variations de la variable explicative. Dans ce cas, la dispersion des valeurs de la variable explicative est nulle : $\sum_{i=1}^n (X_i - \overline{X})^2 = 0$, et la variance $\mathrm{V}(\hat{\beta}_k)$ est infinie.

Cette situation correspond en fait au cas où il est impossible d'estimer β_1 (voir page 28, après le théorème 2.2). Dans ce cas limite, il est impossible, du point de vue des conditions qui définissent le modèle de régression linéaire simple, de distinguer les vraies valeurs des paramètres parmi un continuum de valeurs possibles pour ces paramètres. Par conséquent, on ne peut pas attendre d'estimateurs raisonnables, comme ceux des moindres carrés ordinaires, de fournir une estimation précise de β_0 et de β_1.

— Les deux déterminants des variances $\mathrm{V}(\hat{\beta}_0)$ et $\mathrm{V}(\hat{\beta}_1)$ décrits dans les deux points précédents peuvent s'illustrer graphiquement. Le paramètre σ^2 est la variance commune de $Y_1, \ldots, Y_n$. Cette variance indique le caractère plus ou moins dispersé de la distribution de Y_i autour de son espérance $\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_i$. Plus cette variance est grande, plus on s'attend à ce que l'écart $\big(Y_i - \mathrm{E}(Y_i)\big)^2$ soit élevé. Sur un graphique semblable à celui de la figure 1.1, cet écart est la distance verticale entre Y_i et la droite rouge. Donc dans cas où σ^2 est élevé, plus on s'attend à ce que les points soient dispersés autour de cette droite, dans la direction verticale.
L'autre déterminant des variances de $\hat{\beta}_0$ et $\hat{\beta}_1$ est $\sum_{i=1}^n (X_i - \overline{X})^2$, une mesure de la variabilité observée de $X_1, \ldots, X_n$ autour de leur moyenne $\overline{X}$. Ce terme sera d'autant plus grand que les abscisses des points (X_i, Y_i) du graphique 1.1 sont dispersées. Autrement dit $\sum_{i=1}^n (X_i - \overline{X})^2$ mesure la distance horizontale entre ces points et la droite : plus ces points sont dispersés dans la direction horizontale, plus $\sum_{i=1}^n (X_i - \overline{X})^2$ est grande.
En résumé, l'estimation de β_0 et de β_1 par moindres carrés ordinaires sera d'autant meilleure (plus précise) que la dispersion verticale entre les points (X_i, Y_i) est faible, et/ou que la dispersion horizontale de ces mêmes points est grande. Les graphiques de la figure 2.6 illustrent cette remarque.

Le théorème suivant est la propriété la plus importante de la méthode d'estimation par moindres carrés ordinaires.

th:gm **Théorème 2.3 (Gauss-Markov)** *Dans le modèle de régression linéaire simple, si les paramètres β_0 et β_1 sont identifiés, les estimateurs des moindres carrés ordinaires sont les estimateurs ayant la plus petite variance parmi tous les estimateurs linéaires sans biais de β_0 et de β_1.*

Preuve : Considérons le problème d'estimer β_1 par un estimateur linéaire et sans biais. Cela revient à considérer tous les n-uplets de réels pour lesquels les conditions (2.15) sont satisfaites. À chacun de ces n-uplets est associé un estimateur $\tilde{\beta}_1 = \sum_{i=1}^n \tilde{w}_{1i} Y_i$ de β_1, dont la variance est donnée par (2.16), c'est à dire $\mathrm{V}(\tilde{\beta}_1) = \sigma^2 \sum_{i=1}^n \tilde{w}_{1i}^2$.
Pour montrer que $\hat{\beta}_1 = \sum_{i=1}^n \hat{w}_{1i} Y_i$ est l'estimateur linéaire et sans biais de β_1 ayant une variance plus petite que celle de tout autre estimateur linéaire sans biais $\tilde{\beta}_1 = \sum_{i=1}^n \tilde{w}_{1i} Y_i$

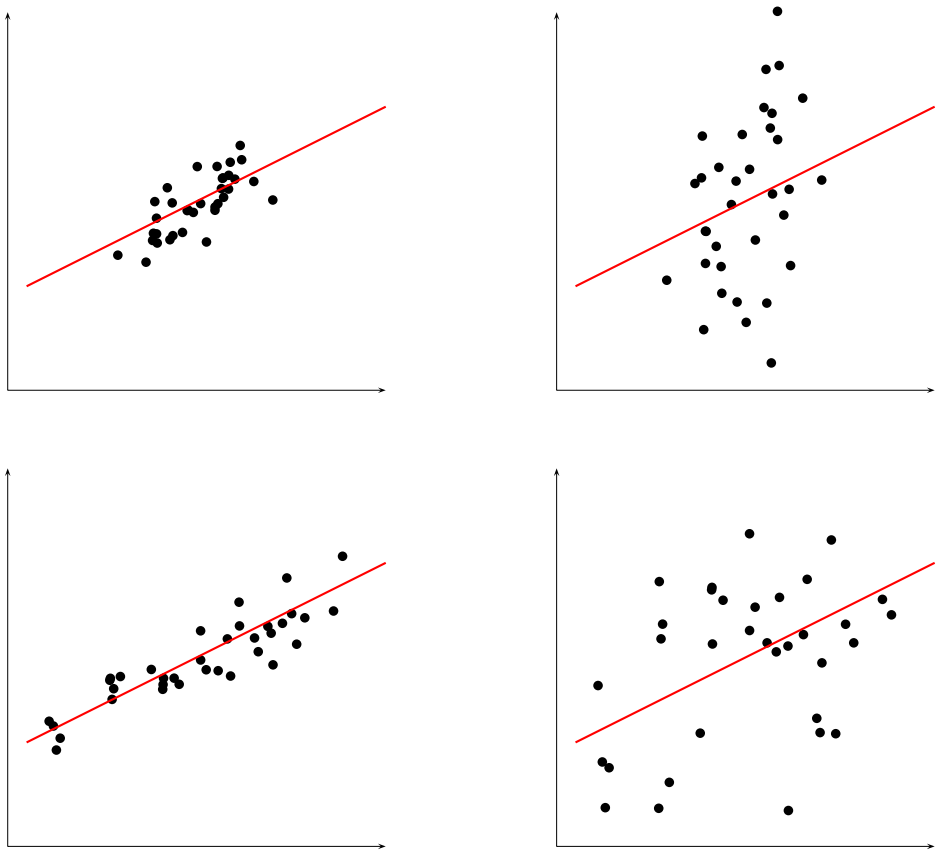


fig:nv

FIGURE 2.6: Tableau de graphiques illustrant l'impact de σ^2 et $\sum_{i=1}^{n}(X_i - \overline{X})^2$ sur l'allure du nuage de points $((X_i, Y_i), i = 1, \ldots, n)$: σ^2 est plus petit sur la colonne de gauche que sur la colonne de droite ; $\sum_{i=1}^{n}(X_i - \overline{X})^2$ est plus petit sur la ligne du haut que sur la ligne du bas. La droite rouge est celle correspondant aux vraies valeurs des paramètres β_0 et β_1. Elle est la même dans tous les cas et sert à mesurer la dispersion relative verticale des points.

de β_1, il est équivalent de montrer que le n-uplet $(\hat{w}_{11}, \ldots, \hat{w}_{1n})$ satisfait (2.15), et que pour tout autre n-uplet $(\tilde{w}_{11}, \ldots, \tilde{w}_{1n})$ satisfaisant les mêmes conditions, on a

$$\sigma^2 \sum_{i=1}^{n} \hat{w}_{1i}^2 \leq \sigma^2 \sum_{i=1}^{n} \tilde{w}_{1i}^2.$$

Autrement dit, il faut montrer que le problème

$$\min_{(w_{11},\ldots,w_{1n})\in\mathbb{R}^n} \sigma^2 \sum_{i=1}^{n} w_{1i}^2 \quad \text{sous contrainte que :} \quad \begin{array}{l} \sum_{i=1}^{n} w_{1i} = 0, \\ \sum_{i=1}^{n} w_{1i} X_i - 1 = 0, \end{array}$$

admet pour solution $\hat{w}_{11}, \ldots, \hat{w}_{1n}$.
Ce problème est évidemment équivalent à

$$\min_{(w_{11},\ldots,w_{1n})\in\mathbb{R}^n} \sum_{i=1}^{n} w_{1i}^2 \quad \text{sous contrainte que :} \quad \begin{array}{l} \sum_{i=1}^{n} w_{1i} = 0, \\ \sum_{i=1}^{n} w_{1i} X_i - 1 = 0, \end{array} \tag{2.17}$$

eq:minvar

Puisque la fonction à minimiser est convexe et dérivable en $w_{11}, \ldots, w_{1n}$, et que les contraintes sont linéaires en $w_{11}, \ldots, w_{1n}$, on peut utiliser la méthode du lagrangien pour résoudre ce problème. Le lagrangien s'écrit

$$L(w_{11}, \ldots, w_{1n}, \lambda, \gamma) = \sum_{i=1}^{n} w_{1i}^2 + \lambda \sum_{i=1}^{n} w_{1i} + \gamma(\sum_{i=1}^{n} w_{1i} X_i - 1)$$

Un n-uplet $(w_{11}^*, \ldots, w_{1n}^*)$ est une solution du problème (2.17) si et seulement si il existe deux réels λ^* et γ^* tels que

$$\begin{cases} \dfrac{\partial L}{\partial w_{1i}}(w_{11}^*, \ldots, w_{1n}^*, \lambda^*, \gamma^*) = 0, \quad i = 1, \ldots, n, \\ \dfrac{\partial L}{\partial \lambda}(w_{11}^*, \ldots, w_{1n}^*, \lambda^*, \gamma^*) = 0, \\ \dfrac{\partial L}{\partial \gamma}(w_{11}^*, \ldots, w_{1n}^*, \lambda^*, \gamma^*) = 0. \end{cases} \tag{2.18}$$

eq:lagcpoa

ou encore, en utilisant la définition de L

$$\begin{cases} 2w_{1i}^* + \lambda^* + \gamma^* X_i = 0, \quad i = 1, \ldots, n, \\ \sum_{i=1}^{n} w_{1i}^* = 0, \\ \sum_{i=1}^{n} w_{1i}^* X_i = 1 \end{cases} \tag{2.19}$$

eq:lagcpob

On somme les n premières équations du système, et on obtient

$$2 \sum_{i=1}^{n} w_{1i}^* + n\lambda^* + \gamma^* n \overline{X} = 0. \tag{2.20}$$

eq:lagcpo1

On multilplie la i^{e} des premières équations par X_i, $i = 1, \ldots, n$, et on fait la somme des n équations ainsi obtenues, ce qui donne

$$2\sum_{i=1}^{n} w_{1i}^* X_i + \lambda^* n\overline{X} + \gamma^* \sum_{i=1}^{n} X_i^2 = 0. \tag{2.21}$$

eq:lagcpo2

En utilisant la $(n+1)^{\text{e}}$ équation du système (2.19) dans (2.20) et la $(n+2)^{\text{e}}$ dans (2.21), on a les conditions suivantes

$$\begin{cases} n\lambda^* + \gamma^* n\overline{X} = 0 \\ \\ 2 + \lambda^* n\overline{X} + \gamma^* \sum_{i=1}^{n} X_i^2 = 0 \end{cases} \tag{2.22}$$

eq:lagcpo3

De la première équation de (2.22) on tire $\lambda^* = -\gamma^* \overline{X}$, qu'on substitue dans la seconde pour obtenir

$$2 + \gamma^* (\sum_{i=1}^{n} X_i^2 - n\overline{X}^2) = 0.$$

Puisqu'on a supposé β_0 et β_1 identifiés, $\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 \neq 0$ (voir le commentaire qui précède le théorème 2.1 à la page 25), et on en déduit

$$\gamma^* = \frac{-2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}.$$

Si on substitue cette expression dans la première équation de (2.22), on obtient

$$\lambda^* = \frac{2\overline{X}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}.$$

On peut finalement obtenir les expressions des w_{1i}^* en remplaçant les expressions de λ^* et γ^* qu'on vient d'obtenir dans chacune des n premières équations de (2.19) :

$$2w_{1i}^* + \frac{2\overline{X}}{\sum_{j=1}^{n} X_j^2 - n\overline{X}^2} - \frac{2X_i}{\sum_{j=1}^{n} X_j^2 - n\overline{X}^2} = 0, \quad i = 1, \ldots, n,$$

ou encore

$$w_{1i}^* = \frac{X_i - \overline{X}}{\sum_{j=1}^{n} X_j^2 - n\overline{X}^2}, \quad i = 1, \ldots, n.$$

En utilisant (2.9), on constate que $w_{1i}^* = \hat{w}_{1i}$, $i = 1, \ldots, n$, (voir la propriété 2.1). Autrement dit, l'estimateur linéaire et sans biais de β_1 ayant la plus petite variance est l'estimateur des moindres carrés ordinaires $\hat{\beta}_1$.

On obtient le résultat concernant $\hat{\beta}_0$ par le même procédé. La preuve est laissée en Exercice.

Le résultat du théorème 2.3 peut être étendu pour montrer que les estimateurs des moindres carrés ordinaires permettent d'obtenir les estimateurs les plus précis parmi tous les estimateurs linéaire et sans biais de n'importe quelle combinaison linéaire de β_0 et de β_1. Plus précisément, si l'objectif est d'estimer $c_0\beta_0 + c_1\beta_1$, où c_0 et c_1 sont des réels connus, alors le meilleur estimateur linéaire et sans biais de cette combinaison linéaire est $c_0\hat{\beta}_0 + c_1\hat{\beta}_1$. Autrement dit, le meilleur estimateur (au sens donné ci-dessus) de la combinaison linéaire est la combinaison linéaire des estimateurs. La preuve de ce résultat s'obtient par la même démarche que celle utilisée pour démontrer le théorème 2.3. Cette preuve est donc laissée en Exercice. Ce résultat permet alors d'obtenir celui du théorème 2.3 comme corollaire, en choisissant d'abord $c_0 = 0$ et $c_1 = 1$, puis $c_0 = 1$ et $c_1 = 0$.

2.3 Propriétés des estimateurs des moindres carrés ordinaires

La plupart des propriétés importantes des estimateurs des moindres carrés ordinaires ont été prouvées ci-dessus. Par conséquent, le résultat suivant consiste simplement en un résumé de ces propriétés.

th:promco **Théorème 2.4** *Dans le modèle de régression linéaire simple, si les paramètres β_0 et β_1 sont identifiés, alors*

1. *Les estimateurs des moindres carrés ordinaires sont donnés par*

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \qquad et \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2}\,.$$

2. *Ces estimateurs sont des variables aléatoires dont les variances sont données par*

$$\mathrm{V}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n X_i^2 - n\overline{X}^2}\right] \qquad et \qquad \mathrm{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n X_i^2 - n\overline{X}^2}\,.$$

3. *Les estimateurs des moindres carrés ordinaires sont linéaires et sans biais. Parmi de tels estimateurs, ce sont les estimateurs les plus précis (de variance minimale).*

Si β_0 et β_1 ne sont pas identifiés, alors la méthode des moindres carrés ordinaires ne permet pas d'estimer β_0 et β_1 séparément.

Rappelons que pour tout $(u_1, \ldots, u_n)$ et $(v_1, \ldots, v_n)$ n-uplets de réels, on a

$$\sum_{i=1}^n (u_i - \overline{u})(v_i - \overline{v}) = \sum_{i=1}^n (u_i - \overline{u}) v_i = \sum_{i=1}^n u_i v_i - n\overline{u}\,\overline{v} \tag{2.23}$$ eq:uv

En effet en développant le membre de gauche de (2.23), on a

$$\sum_{i=1}^n (u_i - \overline{u})(v_i - \overline{v}) = \sum_{i=1}^n (u_i - \overline{u}) v_i - \overline{v} \sum_{i=1}^n (u_i - \overline{u})$$

Comme $\sum_{i=1}^n (u_i - \overline{u}) = 0$, on obtient la première égalité de (2.23). Si on développe maintenant le membre du milieu de (2.23), on obtient

$$\sum_{i=1}^n (u_i - \overline{u}) v_i = \sum_{i=1}^n u_i v_i - \overline{u} \sum_{i=1}^n v_i = \sum_{i=1}^n u_i v_i - \overline{u} n \overline{v}$$

ce qui est la seconde égalité de (2.23).

On peut alors donner une autre expression de $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \tag{2.24}$$ eq:bh1alt

2.4 Mesure de la qualité de l'estimation par moindres carrés ordinaires

2.4.1 Valeurs ajustées et résidus

Définition 2.4 *Dans le modèle de régression linéaire simple, les valeurs ajustées issues de l'estimation par moindres carrés ordinaires de β_0 et de β_1 sont les n variables aléatoires notées $\hat{Y}_1, \ldots, \hat{Y}_n$, définies par $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, $i = 1, \ldots, n$.*

Remarque 2.6

1. Puisque $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs de β_0 et de β_1, on peut interpréter $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ comme un estimateur de $\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_i$. Autrement dit $\hat{Y}_i$ est l'estimation de la valeur attendue de Y_i lorsqu'on connaît X_i. En reprenant l'interprétation donnée au point 1 page 13, on peut également dire que $\hat{Y}_i$ est l'estimation de la partie de Y_i qui peut être expliquée par la valeur de X_i.
2. La valeur ajustée $\hat{Y}_i$ ne coïncide pas avec $\mathrm{E}(Y_i)$, mais on s'attend à ce qu'elle le fasse, puisque la différence attendue est nulle. En effet

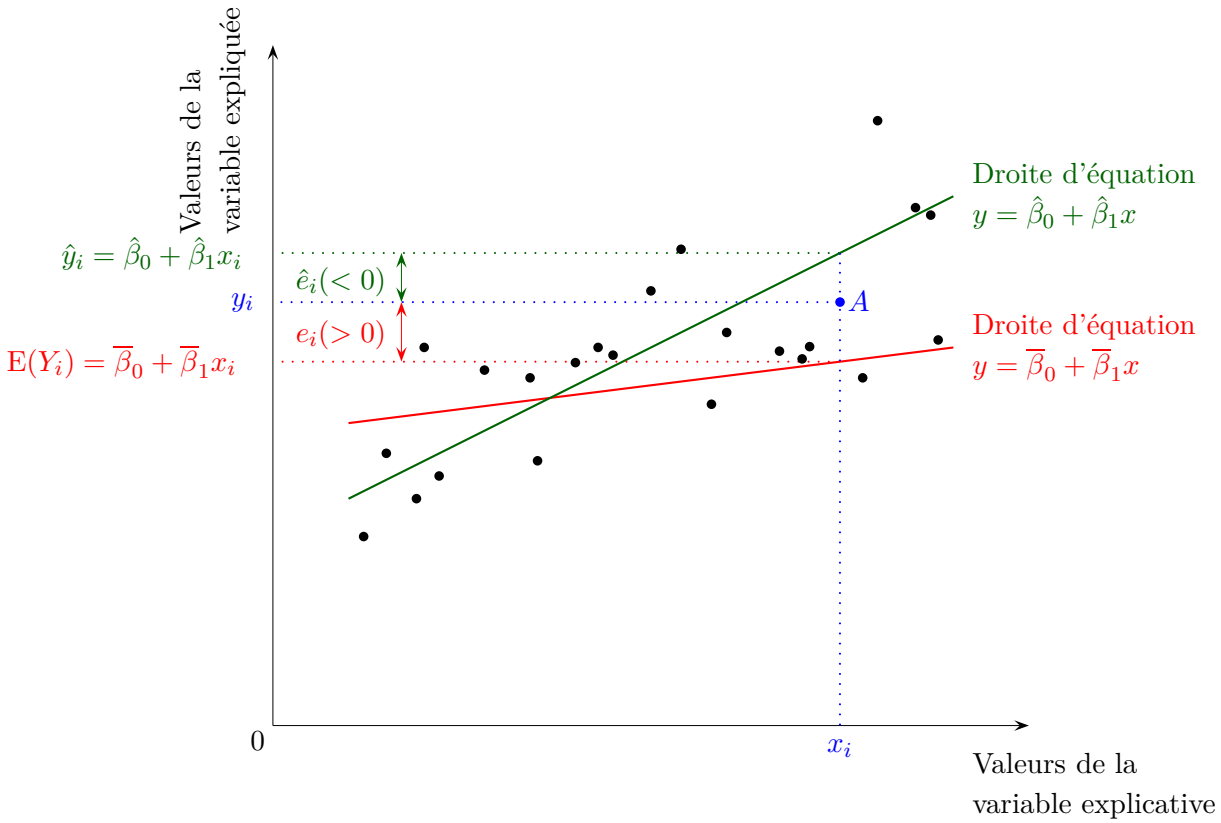
$$\begin{aligned}\mathrm{E}(\hat{Y}_i - \mathrm{E}(Y_i)) &= \mathrm{E}(\hat{\beta}_0 + \hat{\beta}_1 X_i - \beta_0 - \beta_1 X_i) \\ &= \mathrm{E}(\hat{\beta}_0) + \mathrm{E}(\hat{\beta}_1 X_i) - \mathrm{E}(\beta_0 + \beta_1 X_i) \\ &= \beta_0 + \beta_1 X_i - (\beta_0 + \beta_1 X_i) \\ &= 0\end{aligned}$$

   où l'avant dernière égalité résulte du faut que $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 et β_1, respectivement.
3. La valeur ajustée $\hat{Y}_i$ ne coïncide pas non plus avec Y_i. On donne la définition suivante de leur différence.

Définition 2.5 *Dans le modèle de régression linéaire simple, on appelle résidus de l'estimation par moindres carrés ordinaires les variables aléatoires, notés $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$, et définies par $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$.*

Remarque 2.7 Le i^{e} résidu $\hat{\varepsilon}_i$ s'interprète comme l'estimation de la partie de Y_i qu'on ne peut pas expliquer par X_i. Dans la mesure où la valeur ajustée $\hat{Y}_i$ peut être considérée comme un estimateur de $\mathrm{E}(Y_i)$, on peut considérer que $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ est un estimateur de $Y_i - \mathrm{E}(Y_i) = \varepsilon_i$.

La figure 2.7 de la page 39 illustre graphiquement les résultats de l'estimation par moindres carrés ordinaires. Sur cette figure, les couples des observations de (X_i, Y_i), $i = 1, \ldots, n$ sont représentés par des points. La droite d'équation $y = \overline{\beta}_0 + \overline{\beta}_1 x$ est celle le long de laquelle sont alignés les points de coordonnées $(x_i, \mathrm{E}(Y_i))$. Cette droite s'interprète comme la vraie droite, puisque c'est celle qui représente la vraie relation entre Y_i et X_i. La droite d'équation $y = \hat{\beta}_0 + \hat{\beta}_1 x$ contient les points de coordonnées $(x_i, \hat{y}_i)$, où $\hat{y}_i$ est la réalisation de la variable aléatoire $\hat{Y}_i$. Cette droite est entièrement caractérisée par $\hat{\beta}_0$ et $\hat{\beta}_1$. Elle représente l'estimation par moindres carrés ordinaires de la relation entre Y_i et X_i.



pagg:mco FIGURE 2.7: Représentation graphique de l'estimation par moindres carrés ordinaires.

Pour un point A représentant le couple d'observations (x_i, y_i), l'image de x_i par la fonction $y = \overline{\beta}_0 + \overline{\beta}_1 x$ est évidemment $\mathrm{E}(Y_i)$. La différence entre y_i et $\mathrm{E}(Y_i)$ est égale à la réalisation de la variable aléatoire ε_i, qu'on a notée e_i sur le graphique de la figure 2.7. Pour ces valeurs des variables X_i et Y_i, le nombre $\hat{y}_i$ correspond à l'image de x_i par la fonction $y = \hat{\beta}_0 + \hat{\beta}_1 x$. La réalisation, notée $\hat{e}_i$, de la variable aléatoire résidu $\hat{\varepsilon}_i$ est la différence entre y_i et $\hat{y}_i$.

sec:R2

2.4.2 Propriétés

Les résidus possèdent une propriété importante qu'on ré-interprétera dans le chapitre suivant.

pro:orth **Propriété 2.5** *Dans le modèle de régression linéaire simple, on a*

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \qquad et \qquad \sum_{i=1}^{n} \hat{\varepsilon}_i X_i = 0.$$

Preuve : En utilisant la définition de $\hat{\varepsilon}_i$, on constate que ces deux égalités sont une ré-écriture des conditions nécessaires (2.2) définissant les estimateurs des moindres carrés ordinaires $\hat{\beta}_0$ et $\hat{\beta}_1$ comme solutions du problème de minimisation de la fonction $S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$.

Cette propriété permet d'obtenir le résultat suivant.

th:r2 **Théorème 2.5 (Décomposition de la régression)** *Dans le modèle de régression linéaire simple, on a*

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2. \tag{2.25}$$ eq:r2

Preuve : On a

$$(Y_i-\overline{Y})^2=(Y_i-\hat{Y}_i+\hat{Y}_i-\overline{Y})^2=(Y_i-\hat{Y}_i)^2+(\hat{Y}_i-\overline{Y})^2+2(Y_i-\hat{Y}_i)(\hat{Y}_i-\overline{Y}),\quad i=1,\ldots,n.$$

Par conséquent, pour démontrer le théorème, il est suffisant de montrer que

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\overline{Y})=0.$$

La définition de $\hat{\varepsilon}_i$ permet d'écrire le membre de gauche de cette égalité comme

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\overline{Y})=\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{Y}_i-\overline{Y})=\sum_{i=1}^{n}\hat{\varepsilon}_i\hat{Y}_i-\overline{Y}\sum_{i=1}^{n}\hat{\varepsilon}_i=\sum_{i=1}^{n}\hat{\varepsilon}_i\hat{Y}_i=\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{\beta}_0+\hat{\beta}_1X_i).$$

où l'avant-dernière égalité provient de la propriété 2.5 et la dernière de la définition de $\hat{Y}_i$. En décomposant la dernière expression, on a

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\overline{Y})=\hat{\beta}_0\sum_{i=1}^{n}\hat{\varepsilon}_i+\hat{\beta}_1\sum_{i=1}^{n}\hat{\varepsilon}_iX_i.$$

Les deux égalités de la propriété 2.5 permettent de conclure $\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\overline{Y})=0$.

rem:moy_va **Remarque 2.8** Ce résultat a l'interprétation suivante. Le membre de gauche de l'égalité (2.25) est une mesure des variations des Y_i autour de leur moyenne au sein de l'échantillon des individus $i=1,\ldots,n$, ces variations étant mesurées par les (carrés des) distances entre les Y_i et leur moyenne. Pour interpréter le membre de droite, il faut remarquer que

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{\varepsilon}_i)=\overline{Y}-\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i=\overline{Y},$$

d'après la première égalité de la propriété 2.5. Par conséquent le premier terme du membre de droite de l'égalité (2.25) est une mesure des variations des $\hat{Y}_i$ autour de leur moyenne au sein de l'échantillon des individus $i=1,\ldots,n$. Quant au second terme de ce membre de droite, il est égal à $\sum_{i=1}^{n}\hat{\varepsilon}_i^2$, qui est une mesure des variations des $\hat{\varepsilon}_i$ autour de leur moyenne, celle-ci valant 0 d'après la propriété 2.5.

L'égalité (2.25) du théorème 2.5 est une décomposition des variations des Y_i en la somme des variations de $\hat{Y}_i$ et des variations des $\hat{\varepsilon}_i$.

Si on revient à l'interprétation du modèle, on rappelle que Y_i est déterminée par deux facteurs non-corrélés l'un avec l'autre : un facteur prenant la forme d'une fonction affine de la variable explicative du modèle X_i, et un facteur représenté par toutes les autres variables non-corrélées avec X_i. Par conséquent, les sources des variations des Y_i sont aussi de deux natures : il y a d'un côté la partie des variations de Y_i dues aux variations de la variable explicative, et de l'autre la partie des variations de Y_i attribuable aux variations de variables non-corrélées avec la variable explicative.

L'égalité (2.25) traduit cette distinction dans les sources des variations observées des Y_i. Le membre de gauche mesure les variations observées des Y_i. Il s'agit de la variation totale, sans que l'on chercher à distinguer la partie de ces variations attribuables à une source ou à l'autre. On appelle le terme $\sum_{i=1}^n (Y_i - \overline{Y})^2$ *variation totale*, ou *somme des carrés totaux* (SCT).

Dans le membre de droite, le premier terme $\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$ est une estimation de la part des variations des Y_i qui sont attribuables aux variations de la variable explicative. En effet $\hat{Y}_i$ est une estimation $\mathrm{E}(Y_i)$, c'est-à-dire de la partie de Y_i qui peut s'écrire entièrement comme une fonction affine de la variable explicative, *uniquement*. Par conséquent, la seule source de variabilité de $\mathrm{E}(Y_i)$ est la variabilité de X_i. L'estimation de la variabilité de $\mathrm{E}(Y_i)$ est $\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$. On appelle le terme $\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$ *variation expliquée*, ou *somme des carrés expliqués* (SCE).

Quant au second terme du membre de droite $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2$, c'est une estimation de partie des variations des Y_i qui ne peuvent être causées par des variations de X_i. C'est la partie des variations des Y_i qui reste, ou résiduelle, une fois qu'on a retranché aux variations des Y_i la part attribuable aux variations de la variable explicative. On appelle le terme $\sum_{i=1}^n \hat{\varepsilon}_i^2$ *variations résiduelles*, ou *somme des carrés des résidus* (SCR).

On peut donc ré-énoncer le théorème 2.5 de la façon suivante : dans le modèle de régression linéaire simple, on a SCT = SCE + SCR. À partir de cette égalité, on peut construire un estimateur de la capacité de la variable explicative à déterminer le niveau de la variable dépendante.

def:r2 **Définition 2.6** *Dans le modèle de régression linéaire simple, on appelle coefficient de détermination de la régression, et on note R^2 le nombre défini par*

$$R^2 = \frac{\mathrm{SCE}}{\mathrm{SCT}} = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2}$$

rem:r2 **Remarque 2.9**

1. Puisque SCT = SCE + SCR et que les trois sommes de cette égalité sont positives, on a nécessairement SCT $\geq$ SCE ≥ 0 et donc $0 \leq R^2 \leq 1$. Le rapport définissant R^2 s'interprète alors comme une proportion. Le coefficient de détermination est la part des variations observées des Y_i qu'on peut estimer être attribuables aux variations de la variable explicative. On dira alors qu'on peut estimer que $(100 \times R^2)\%$ des variations des variables $Y_1, \ldots, Y_n$ sont dues aux variations des variables explicatives $X_1, \ldots, X_n$..
2. On voit que l'égalité SCT = SCE + SCR permet de définir R^2 par $1 - \frac{\mathrm{SCR}}{\mathrm{SCT}} = 1 - \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\mathrm{SCT}}$
3. Le rapport R^2 est une mesure de la capacité des variables explicatives à faire varier, par leurs propres variations, les variables endogènes. Autrement dit, R^2 est une mesure de l'effet que les X_i peuvent avoir sur les Y_i, c'est à dire une mesure du pouvoir explicatif des X_i sur les Y_i. Plus précisément, plus R^2 est proche de 1, plus la part des variations des Y_i qu'on peut attribuer aux variations des X_i est grande. De façon équivalente, plus R^2 est proche de 1, plus la part des variations des Y_i attribées aux variables autres que $X_1, \ldots, X_n$ (et non corrélées aux X_i) est faible. Autrement dit, le principal déterminant du niveau des Y_i est le niveau des X_i. Dans ce cas, le pouvoir explicatif des variables explicatives est élevé.

   Si R^2 est proche de 0, la plus grande partie des variations des variables Y_i est attribuable aux variations résiduelles, c'est à dire aux variations des variables autres que les variables explica-

tives, et non-corrélées à celles-ci. Dans ce cas, le pouvoir explicatif des variables explicatives est faible.

Les cas extrêmes $R^2 = 0$ et $R^2 = 1$ peuvent s'énoncer (de manière équivalente) sous une forme qui permet d'obtenir directement les interprétations données ci-dessus. C'est ce qu'exprime la propriété 2.6 (voir plus bas).

4. On a justifié (dans la section 2.1) la démarche d'estimation des paramètres par minimisation de la fonction S, en notant que pour un choix donné $(\beta_0, \beta_1) \in \mathbb{R}^2$, le nombre $S(\beta_0, \beta_1)$ mesurait l'importance des facteurs autres que la variable explicative dans la détermination du niveau de la variable expliquée, via la relation $Y = \beta_0 + \beta_1 X + \varepsilon$ (voir page 22). Autrement dit, choisir d'estimer de cette manière les paramètres, revient à choisir les valeurs de ces paramètres qui maximisent la capacité de la variable explicative X à déterminer le niveau de la variable expliquée Y. Maintenant qu'on dispose, à travers le coefficient de détermination R^2, d'une estimation de cette capacité, on devrait être capable de formaliser la justification donnée à la minimisation de S. Pour cela, pour chaque couple (β_0, β_1) de valeurs possibles des paramètres, on peut mesurer la capacité de X à déterminer le niveau de Y lorsque la valeur des paramètres est (β_0, β_1) par la quantité $R^2(\beta_0, \beta_1)$ définie par

$$R^2(\beta_0, \beta_1) = 1 - \frac{\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2}{\text{SCT}}$$

On voit alors que chercher les valeurs des paramètres qui donnent la capacité maximum de X pour expliquer Y sont celles qui maximisent $R^2(\beta_0, \beta_1)$ ou encore, celles qui minimisent $S(\beta_0, \beta_1)$. Par conséquent, le couple de valeurs (β_0, β_1) pour lequel $R^2(\beta_0, \beta_1)$ est maximal est évidemment $(\hat{\beta}_0, \hat{\beta}_1)$. R

pro:R2 **Propriété 2.6** *Dans le modèle de régression linéaire simple, si $\exists i, j$ tels que $X_i \neq X_j$, alors on a*

1. $R^2 = 1 \iff \exists (\beta_0^*, \beta_1^*) \in \mathbb{R}^2, Y_i = \beta_0^* + \beta_1^* X_i, \forall i = 1, \ldots, n.$
2. $R^2 = 0 \iff \hat{\beta}_1 = 0.$

Preuve :

1. En utilisant les définitions de R^2 et de $\hat{\varepsilon}_i$ on a

$$R^2 = 1 \iff \text{SCR=0} \iff \sum_{i=1}^n \hat{\varepsilon}_i^2 = 0 \iff \hat{\varepsilon}_i = 0,\ i = 1, \ldots, n$$
$$\iff Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, n$$

où la dernière équivalence provient de l'égalité $Y_i = \hat{Y}_i + \hat{\varepsilon}_i$ et de la définition de $\hat{Y}_i$. On voit donc que l'équivalence du premier point de la propriété est obtenu en choisissant $\beta_0^* = \hat{\beta}_0$ et $\beta_1^* = \hat{\beta}_1$.

2. Toujours avec les mêmes définitions, on a

$$R^2 = 0 \iff \text{SCE=0} \iff \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2 = 0 \iff \hat{Y}_i = \overline{Y},\ i = 1, \ldots, n$$
$$\iff \hat{Y}_i = \hat{Y}_j,\ i, j = 1, \ldots, n$$

En utilisant la définition des $\hat{Y}_i$, les dernières égalités sont équivalentes à

$$\hat{\beta}_1(X_i - X_j) = 0,\ i, j = 1, \ldots, n$$

Ces n^2 égalités sont toutes vraies si et seulement si

$$X_i = X_j,\ i, j = 1, \ldots, n \text{ ou } \hat{\beta}_1 = 0$$

La première condition étant exclue, on obtient donc $R^2 = 0 \iff \hat{\beta}_1 = 0$.

rem:R2 **Remarque 2.10**

1. Le premier point de la proposition 2.6 montre clairement que lorsque $R^2 = 1$ on estime que Y_i est uniquement déterminé par X_i, $i = 1, \ldots, n$. Dans ce cas, pour tout individu i, les facteurs autres que X_i pouvant affecter le niveau de Y_i sont inexistants. Dans la formulation du modèle de régression linéaire simple, cela revient à écrire que $\varepsilon_i = 0$ pour $i = 1, \ldots, n$, et qu'on peut écrire Y_i comme une fonction affine de X_i. La condition C′2 est dans ce cas :

$$\exists \beta_0 \in \mathbb{R}, \exists \beta_1 \in \mathbb{R} \text{ t.q. } Y_i = \beta_0 + \beta_1 x_i\,, \quad i = 1, \ldots, n.$$

   La preuve de la proposition 2.6 montre que les réels β_0 et β_1 qui satisfont les n égalités de la condition C′2 sont donnés par $\hat{\beta}_0$ et $\hat{\beta}_1$, respectivement. Tous les points de coordonnées (X_i, Y_i), $i = 1, \ldots, n$, appartiennent à la droite d'équation $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

2. Le second point montre que lorsque $R^2 = 0$, on estime que Y_i n'est déterminé que par des variables autres que X_i. Autrement dit, on estime donc que lorsque X_i varie, cela n'engendre aucune variation de Y_i. Dans le contexte d'un modèle de régression linéaire simple, dans lequel on suppose que $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, cela revient à estimer que $\beta_1 = 0$. C'est précisément ce que dit l'égalité $\hat{\beta}_1 = 0$.

3. Le premier point de cette propriété laisse suggérer qu'il existe une relation entre le coefficient de détermination et le coefficient de corrélation linéaire empirique. On rappelle que le coefficient de corrélation linéaire empirique entre les variables X et Y, noté $r(X, Y)$, est défini par

$$r(X, Y) = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}$$

   Il permet d'évaluer l'intensité d'une liaison linéaire entre X et Y. On peut donc s'attendre à ce que ce coefficient soit lié au coefficient de détermination, puisque ce dernier mesure le pouvoir explicatif de X sur Y au travers d'une liaison linéaire du type $Y_i = \beta_0 + \beta_1 X_i$, perturbée par un terme ε_i. L'intensité de cette liaison linéaire sera d'autant plus forte (et donc $|r(X, Y)|$ proche de 1) que l'influence des ε_i sera faible. C'est précisément ce qu'indique le coefficient de détermination R^2. La propriété suivante formalise cette remarque.

pro:R2_rxy2 **Propriété 2.7** *Dans le modèle de régression linéaire simple, on a* $R^2 = r(X, Y)^2$.

Preuve : On rappelle que $\overline{Y}$ coïncide avec la moyenne des valeurs ajustées (voir la remarque 2.8). Par conséquent

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^n \hat{Y}_i = \frac{1}{n}\sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}$$

Donc

$$\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2=\sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1X_i-\hat{\beta}_0+\hat{\beta}_1\overline{X})^2=\hat{\beta}_1^2\sum_{i=1}^{n}(X_i-\overline{X})^2$$
$$=\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]^2\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{\left[\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})\right]^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

où pour obtenir la dernière égalité, on a utilisé l'expression (2.24) de $\hat{\beta}_1$. Le résultat découle directement de l'utilisation de cette expression dans la définition de R^2.

Remarque 2.11 (Coefficient de détermination dans un modèle où la relation n'a pas de terme constant) Finalement, remarquons que la formule du R^2 et son interprétation reposent sur la décomposition (2.25) du théorème 2.5. Celle-ci a été obtenue en utilisant les égalités de la propriété 2.5. Or la première de ces égalités n'est en général pas vérifiée dans un modèle de régression linéaire dans lequel on impose $\beta_0=0$ (Exercice : vérifier cette affirmation). Par conséquent, dans un tel modèle, les propriétés du coefficient R^2 et son interprétation ne sont plus valables. Cependant, la propriété 2.6 et les observations faites aux remarques 2.9 et 2.10 permettent de donner une autre interprétation au R^2, et à partir de là, d'en proposer une extension adaptée au cas où on suppose $\beta_0=0$. Pour cela, on envisage successivement deux contextes possibles.

Si on place dans le contexte où on suppose que la variable exogène n'a aucun pouvoir explicatif sur la variable endogène, alors $\beta_1=0$ et si on estime le modèle sous cette condition, les valeurs ajustées obtenues, notées $\hat{Y}_i^o$, sont alors $\hat{Y}_i^o=\overline{Y}$ pour tout $i=1,\ldots,n$ (Exercice : vérifier ceci). Si maintenant on se place dans le contexte où rien n'est dit *a priori* sur le pouvoir explicatif de la variable exogène, on a évidemment $\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1X_i$, $i=1,\ldots,n$. Par conséquent, on peut interpréter $(\hat{Y}_i-\hat{Y}_i^o)^2=(\hat{Y}_i-\overline{Y})^2$ comme la distance entre les i^{e} valeurs ajustées obtenues dans chacun des deux contextes, et $\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$ comme la distance totale entre toutes ces valeurs ajustées. Si le pouvoir explicatif de la variable exogène est effectivement faible (ou nul), alors l'estimation effectuée dans chacun de ces deux contextes et les valeurs ajustées correspondantes devraient être peut différentes. Autrement dit on devrait avoir dans ce cas $\hat{Y}_i$ peu différent de $\overline{Y}$ pour tout i et donc $\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$ proche de 0. Or ce terme est le numérateur du coefficient R^2 et celui-ci devrait donc également être proche de 0. Ceci est précisément l'interprétation de la valeur de R^2 donnée dans le second point de la remarque 2.9.

On peut alors maintenant reprendre ce même raisonnement lorsqu'on considère un modèle dans lequel la relation ne contient aucun terme constant, *i.e.*, $Y_i=\beta_1X_i+\varepsilon_i$, $i=1,\ldots,n$. Dans ce cas, si on suppose que la variable exogène n'a pas de pouvoir explicatif (*i.e.*, $\beta_1=0$) et qu'on utilise cette hypothèse pour l'estimation du modèle, les valeurs ajustées obtenues sont $\hat{Y}_i^o=0$ pour tout $i=1,\ldots,n$ (Exercice : montrer cela). Si maintenant aucune hypothèse n'est faite *a priori* sur ce pouvoir explicatif, alors $\hat{Y}_i=\hat{\beta}_1X_i$, $i=1,\ldots,n$, où $\hat{\beta}_1$ est obtenu en minimisant $\sum_{i=1}^{n}(Y_i-\beta_1X_i)^2$. La distance totale entre les valeurs ajustées obtenues dans chacun de ces deux contextes est donc à présent $\sum_{i=1}^{n}(\hat{Y}_i-\hat{Y}_i^o)^2=\sum_{i=1}^{n}\hat{Y}_i^2$.

Si comme précédemment, on évalue le pouvoir explicatif de la variable exogène en examinant la distance totale entre les valeurs ajustées obtenues dans ces deux contextes, on

On termine cette section en rappelant les propriétés élémentaires du coefficient de corrélation linéaire empirique.

Propriété 2.8

1. $r(Y,X) = r(X,Y) \in [-1;1]$.
2. $r(X,Y) = 1 \iff \exists a \in]0,+\infty[, \exists b \in \mathbb{R},\ Y_i = aX_i + b\ \forall i = 1,\dots,n$. *De plus,* $r(X,Y) = -1 \iff \exists a \in]-\infty,0[, \exists b \in \mathbb{R},\ Y_i = aX_i + b\ \forall i = 1,\dots,n$.

Pour démontrer ces propriétés, il est commode d'introduire la notation suivante : $\Sigma_{X,Y} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$. Ainsi, on peut réécrire $r(X,Y) = \Sigma_{X,Y}/\sqrt{\Sigma_{X,X}\Sigma_{Y,Y}}$.

Preuve :

1. La propriété de symétrie résulte directement du fait que $\Sigma_{X,Y} = \Sigma_{Y,X}$. Pour tout réel λ, on peut former

$$\Sigma_{\lambda X+Y,\lambda X+Y} = \sum_{i=1}^{n}\Big[(\lambda X_i + Y_i) - (\lambda\overline{X} + \overline{Y})\Big]^2$$

En développant, on peut écrire

$$\Sigma_{\lambda X+Y,\lambda X+Y} = \sum_{i=1}^{n}\Big[\lambda(X_i - \overline{X}) + (Y_i - \overline{Y})\Big]^2 = \lambda^2\Sigma_{X,X} + \Sigma_{Y,Y} + 2\lambda\Sigma_{X,Y} \qquad (2.26)$$

eq:cs

Cette expression permet de considérer $\Sigma_{\lambda X+Y,\lambda X+Y}$ comme un polynôme en $\lambda \in \mathbb{R}$. On note que ce polynôme est toujours positif ou nul (il peut s'exprimer comme une somme de carrés). Par conséquent, son discriminant doit nécessairement être négatif ou nul. Autrement dit, on doit avoir $4\Sigma_{X,Y}^2 - 4\Sigma_{X,X}\Sigma_{Y,Y} \leq 0$, ou encore $\frac{\Sigma_{X,Y}^2}{\Sigma_{X,X}\Sigma_{Y,Y}} \leq 1$, c'est à dire $r(X,Y)^2 \leq 1$. D'où le résultat.

2. Supposons qu'il existe des réels a et b, avec $a \neq 0$, tels que $Y_i = aX_i + b$, pour tout $i = 1,\dots,n$. On a $Y_i - \overline{Y} = a(X_i - \overline{X})$ pour tout i et on vérifie alors facilement que $\Sigma_{Y,Y} = a^2\Sigma_{X,X}$ et que $\Sigma_{X,Y} = a\Sigma_{X,X}$. Par conséquent, $r(X,Y) = 1$ si $a > 0$ et $r(X,Y) = -1$ si $a < 0$.
Supposons maintenant que $|r(X,Y)| = 1$, ou de manière équivalente, que $r(X,Y)^2 = 1$. Cela équivaut aussi à $\Sigma_{X,Y}^2 = \Sigma_{X,X}\Sigma_{Y,Y}$. Le discriminant du polynôme introduit en (2.26) est alors nul et il admet une racine unique, notée λ^*. D'après (2.26), on peut écrire

$$\sum_{i=1}^{n}\Big[\lambda^*(X_i - \overline{X}) + (Y_i - \overline{Y})\Big]^2 = 0$$

Cette somme de carrés est nulle si et seulement si tous les carrés sont nuls. On doit donc avoir $\lambda^*(X_i - \overline{X}) + (Y_i - \overline{Y}) = 0$, $\forall i = 1,\dots,n$, ou encore

$$Y_i = aX_i + b, \quad i = 1,\dots,n$$

avec $a = -\lambda^*$ et $b = \lambda^*\overline{X} + \overline{Y}$. Finalement, on étudie le signe de a. Notons que la racine λ^* est égale à $-\Sigma_{X,Y}/\Sigma_{X,X}$. Donc, sous l'hypothèse initiale que $|r(X,Y)| = 1$, on a $a > 0 \iff \Sigma_{X,Y} > 0 \iff r(X,Y) = 1$ et donc $a < 0 \iff \Sigma_{X,Y} < 0 \iff r(X,Y) = -1$.

sec:sigma2

2.5 Estimation des variances

2.5.1 Estimation de la variance des termes d'erreur

Comme on le verra dans la section suivante, on ne peut se contenter d'une simple estimation de β_0 et de β_1. On souhaite par exemple disposer d'une mesure de la précision de l'estimation obtenue. Puisque les estimateurs des moindres carrés ordinaires sont sans biais, on peut mesurer leur précision par la variance de ces estimateurs. Nous avons vu dans la propriété 2.4 à l'équation (2.16), et dans le corollaire qui suit, que les variances des estimateurs des moindres carrés ordinaires dépendent de la variance σ^2 des termes d'erreur ε_i. Or la valeur de celle-ci est inconnue. Dans cette section, on présente une façon d'estimer cette variance basée sur le résultat suivant.

Propriété 2.9 *Dans le modèle de régression linéaire simple, si les paramètres β_0 et β_1 sont identifiés, on a*

$$\mathrm{E}(\sum_{i=1}^{n}\hat{\varepsilon}_i^2) = (n-2)\sigma^2.$$

Preuve : On a

$$\begin{aligned}
\hat{\varepsilon}_i = Y_i - \hat{Y}_i &= \beta_0 + \beta_1 X_i + \varepsilon_i - \hat{\beta}_0 - \hat{\beta}_1 X_i && \text{[par définition des } \hat{Y}_i\text{]}\\
&= \beta_0 + \beta_1 X_i + \varepsilon_i - \overline{Y} + \hat{\beta}_1\overline{X} - \hat{\beta}_1 X_i && \text{[par définition de } \hat{\beta}_0\text{]}\\
&= \beta_0 + \beta_1 X_i + \varepsilon_i - \beta_0 - \beta_1\overline{X} - \overline{\varepsilon} + \hat{\beta}_1\overline{X} - \hat{\beta}_1 X_i && \text{[par calcul de } \overline{Y}\text{]}\\
&= \varepsilon_i - \overline{\varepsilon} - (X_i - \overline{X})(\hat{\beta}_1 - \beta_1)
\end{aligned}$$

Donc

$$\hat{\varepsilon}_i^2 = \varepsilon_i^2 + \overline{\varepsilon}^2 + (X_i - \overline{X})^2(\hat{\beta}_1 - \beta_1)^2 - 2\varepsilon_i(X_i - \overline{X})(\hat{\beta}_1 - \beta_1) - 2\varepsilon_i\overline{\varepsilon} + 2\overline{\varepsilon}(X_i - \overline{X})(\hat{\beta}_1 - \beta_1)$$

et

$$\begin{aligned}
\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}\varepsilon_i^2 + n\overline{\varepsilon}^2 + (\hat{\beta}_1 - \beta_1)^2\sum_{i=1}^{n}(X_i - \overline{X})^2 - 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(X_i - \overline{X})\varepsilon_i - 2n\overline{\varepsilon}^2\\
+ 2\overline{\varepsilon}(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(X_i - \overline{X}) \quad (2.27)
\end{aligned}$$

eq:sumrescarre

Le dernier terme du membre de droite est nul. D'autre part, d'après l'expression (2.24), le numérateur de $\hat{\beta}_1$ peut s'écrire

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \overline{X})Y_i &= \sum_{i=1}^{n}(X_i - \overline{X})(\beta_0 + \beta_1 X_i + \varepsilon_i)\\
&= \beta_0\sum_{i=1}^{n}(X_i - \overline{X}) + \beta_1\sum_{i=1}^{n}(X_i - \overline{X})X_i + \sum_{i=1}^{n}(X_i - \overline{X})\varepsilon_i\\
&= \beta_1\sum_{i=1}^{n}(X_i - \overline{X})X_i + \sum_{i=1}^{n}(X_i - \overline{X})\varepsilon_i
\end{aligned}$$

car $\sum_{i=1}^n (X_i - \overline{X}) = 0$. Par conséquent

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})X_i} = \beta_1 \frac{\sum_{i=1}^n (X_i - \overline{X})X_i}{\sum_{i=1}^n (X_i - \overline{X})X_i} + \frac{\sum_{i=1}^n (X_i - \overline{X})\varepsilon_i}{\sum_{i=1}^n (X_i - \overline{X})X_i}$$

et donc

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})\varepsilon_i}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

Par conséquent $\sum_{i=1}^n (X_i - \overline{X})\varepsilon_i = (\hat{\beta}_1 - \beta_1)\sum_{i=1}^n (X_i - \overline{X})^2$ et en utilisant cette expression dans (2.27), on peut écrire

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n \varepsilon_i^2 - (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^n (X_i - \overline{X})^2 - n\overline{\varepsilon}^2$$

Si à présent on calcule l'espérance, on obtient

$$\begin{aligned}
\mathrm{E}\Big(\sum_{i=1}^n \hat{\varepsilon}_i^2\Big) &= \sum_{i=1}^n \mathrm{E}(\varepsilon_i^2) + \mathrm{E}[(\hat{\beta}_1 - \beta_1)^2] \sum_{i=1}^n (X_i - \overline{X})^2 - n\mathrm{E}(\overline{\varepsilon}^2) \quad \text{[linéarité de l'espérance]} \\
&= n\sigma^2 - \mathrm{V}(\hat{\beta}_1) \sum_{i=1}^n (X_i - \overline{X})^2 - \frac{n}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathrm{E}(\varepsilon_i \varepsilon_j) \quad \text{[condition C}'4\text{; } \mathrm{E}(\hat{\beta}_1) = \beta_1\text{]} \\
&= n\sigma^2 - \sigma^2 - \frac{1}{n} \sum_{i=1}^n \mathrm{E}(\varepsilon_i^2) \quad \text{[condition C}'4\text{; expression de } \mathrm{V}(\hat{\beta}_1)\text{]} \\
&= n\sigma^2 - \sigma^2 - \sigma^2
\end{aligned}$$

cor:sigma **Corollaire 2.2** *Dans le modèle de régression linéaire simple, la variable aléatoire $\hat{\sigma}^2$ définie par*

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

est un estimateur sans biais de σ. On a $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$.

2.5.2 Estimation de la variance des estimateurs des moindres carrés ordinaires

Comme mentionnné à la remarque 2.5 les variances des estimateurs des moindres carrés ordinaires ne sont inconnues que parce que σ^2 l'est. Cependant, le résultat précédent nous permet de former des estimateurs des variances.

pro:estimvarmco **Propriété 2.10** *Dans le modèle de régression linéaire simple, si les paramètres β_0 et β_1 sont identifiés, les variables aléatoires $\hat{\mathrm{V}}(\hat{\beta}_0)$ et $\hat{\mathrm{V}}(\hat{\beta}_1)$ définies par*

$$\hat{\mathrm{V}}(\hat{\beta}_0) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] \quad \text{et} \quad \hat{\mathrm{V}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

sont des estimateurs sans biais de $\mathrm{V}(\hat{\beta}_0)$ et $\mathrm{V}(\hat{\beta}_1)$, respectivement.

Preuve : Il découle directement de l'expression des variances et du corollaire 2.2 que $\mathrm{E}[\hat{\mathrm{V}}(\hat{\beta}_k)] = \mathrm{V}(\hat{\beta}_k)$, $k = 0, 1$.

Chapitre 3

sec:univ-tests

Le modèle de régression linéaire simple : tests et régions de confiance

Dans cette section, on s'intéresse d'une autre manière aux paramètres d'intérêt du modèle de régression linéaire simple. Le problème d'inférence abordé est celui des tests d'hypothèses sur ces paramètres. La démarche sera évidemment celle rappelée à la section section 10.3.2. L'assimilation de cette section est donc un préalable à la lecture de ce chapitre.

sec:MRLSG

3.1 Contexte : le modèle gaussien

Jusqu'à présent, on a étudié le problème de l'estimation des paramètres du modèle de régression linéaire simple. Les propriétés qu'il est possible d'établir pour les estimateurs dépendent de la manière dont a été spécifié le modèle. Ainsi en s'appuyant sur les conditions C1 à C2, il a été possible de montrer que $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 et β_1. En rajoutant la condition C3, on a pu établir que parmi tous les estimateurs linéaires et sans biais de β_1, $\hat{\beta}_1$ était celui qui avait la plus petite variance.

Si on s'intéresse maintenant à un problème de test, la démarche est similaire. On peut commencer par proposer un test, puis en utilisant les conditions qui spécifient le modèle, on étudie les propriétés de ce test. Puis parmi tous les tests envisagés, on cherche éventuellement le meilleur.

On sait, comme cela est rappelé dans la section 10.3.2, que pour établir les propriétés d'un test et choisir un test optimal il est nécessaire de pouvoir effectuer des calculs de risques. Ces derniers étant définis comme des probabilités de commettre des erreurs, il faut pour cela disposer de lois permettant de faire les calculs de probabilité. Notons que dans le modèle de régression linéaire simple tel que défini par les conditions C1, C2 et C3, rien ne nous permet de faire de tels calculs, dès que ceux-ci portent sur des statistiques qui sont des fonctions de $Y_1, \ldots, Y_n$. Il faut donc compléter d'une certaine manière la définition du modèle et lui ajouter des conditions qui permettront d'effectuer le calcul des risques.

Plusieurs approches sont possibles. Celle qu'on adopte (la plus simple) consiste à introduire dans la définition même du modèle les lois permettant le calcul des probabilités d'erreurs. On modifie alors la définition du modèle de régression linéaire simple.

sec:modeleG

3.1.1 Définition du modèle gaussien

On modifie la définition du MRLS de manière à introduire explicitement dans le modèle une loi de probabilité qui permet de calculer les risques des tests qu'on utilisera.

Pour définir le nouveau modèle, on ajoute aux conditions C1 à C3 (ou C′1 à C′3) qui définissent le MRLS, la condition C″N suivante :

C″N. $(\varepsilon_1, \dots, \varepsilon_n)$ est un n-uplet gaussien

La section 9.1 regroupe tous les résultats et définitions relatifs à la loi normale et aux n-uplets gaussiens qui seront utilisés dans ce document. On rappelle en particulier (voir la définition 9.1) que n variables aléatoires forment un n-uplet gaussien si toute combinaison linéaire de ces variables définit une variable aléatoire gaussienne (*i.e.,* dont la loi de probabilité est une loi normale). On rappelle également qu'un tel n-uplet peut être vu comme un vecteur aléatoire de $\mathbb{R}^n$ (c'est à dire un vecteur de $\mathbb{R}^n$ dont les coordonnées sont aléatoires). Finalement, on rappelle que comme pour une variable aléatoire gaussienne (ou normale), la loi d'un vecteur aléatoire gaussien est entièrement caractérisée par le vecteur des espérances et par la matrice des variances-covariances (voir la remarque 9.2).

Avec la condition C″N, on considère le n-uplet $(\varepsilon_1, \dots, \varepsilon_n)$ comme les coordonnées du vecteur aléatoire ε. On a d'après la condition C′3 et la définition des termes d'erreur :

$$\mathrm{E}(\varepsilon) = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = 0_n \qquad \mathrm{V}(\varepsilon) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 I_n$$

où 0_n désigne le vecteur nul de $\mathbb{R}^n$ et I_n la matrice identité de $\mathbb{R}^n$ vers $\mathbb{R}^n$. Avec la condition supplémentaire C″N, on aura donc $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$. (Exercice : détailler ce raisonnement)

On s'aperçoit donc que le modèle de régression linéaire simple défini par les conditions C′1 à C′3 et C″N peut aussi se définir de manière équivalente par les conditions C′1, C′2 et C′N, où cette dernière est

C′N. $\exists \sigma \in]0, +\infty[,\ \varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$.

On appelle *modèle de régression linéaire simple gaussien* (MRLSG par la suite) le modèle défini par les conditions C′1, C′2 et C′N. Ce modèle servira de contexte dans lequel seront construits des tests permettant de tester des hypothèses formulées sur les paramètres d'intérêt.

Nous avons montré que le modèle de régression linéaire simple admet deux définitions équivalentes, l'une exprimée au moyen de conditions portant sur les propriétés de $Y_1, \dots, Y_n$ et l'autre au moyen de conditions portant sur les propriétés $\varepsilon_1, \dots, \varepsilon_n$. Il en est de même pour le MRLSG, ainsi que le montre la propriété suivante.

pro:CN

Propriété 3.1 *Définissons* $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ *et* $\iota_n = (1, \dots, 1)^\top \in \mathbb{R}^n$. *Les conditions* C′1, C′2 *et* C′N *sont vérifiées si et seulement si les conditions* C1 *et* CN *le sont aussi, la condition* CN *étant définie par*

CN. $\exists \beta_0 \in \mathbb{R}, \exists \beta_1 \in \mathbb{R}, \exists \sigma \in]0, +\infty[,\ \boldsymbol{Y} \sim \mathcal{N}(\beta_0 \iota_n + \beta_1 \boldsymbol{X}, \sigma^2 I_n)$.

Preuve : Exercice : montrer que si on suppose C′1, C′2 et C′N vraies, alors C1, C2 et CN le sont aussi, et réciproquement.

On peut formaliser la propriété précédente en introduisant une définition du MRLSG.

def:msrlg-univ **Définition 3.1** *Soient $(X_1, Y_1), \ldots, (X_n, Y_n)$ n couples de variables aléatoires dont les observations sont notées $(x_1, y_1), \ldots, (x_n, y_n)$. Le modèle de régression linéaire simple gaussien (MRLSG) de Y sur X est un modèle statistique dans lequel les conditions* C1 *et* CN *sont satisfaites. De manière équivalente, ce modèle est également défini par les conditions* C′1, C′2 *et* C′N.

L'animation de la figure 3.1 illustre la modélisation de la relation entre les variables explicatives et expliquées retenue dans le MRLSG. [1] Par rapport au modèle de régression linéaire simple, le MRLSG ajoute la condition que le n-uplet $(Y_1, \ldots, Y_n)$ est gaussien. Pour représenter cet ajout de condition graphiquement, on reprend le graphique de la figure 1.1 en y ajoutant une 3$^{\text{e}}$ dimension (verticale) qui permet de représenter le caractère gaussien des variables expliquées $Y_1, \ldots, Y_n$. La droite représentant dans le plan la relation entre $\mathrm{E}(Y_i)$ et x_i, pour $i = 1, \ldots, n$ est d'abord tracée. Pour chaque individu i, on représente par un $\bullet$ le couple d'observations (x_i, y_i) ainsi que, en utilisant la dimension verticale, la densité (gaussienne) de Y_i (courbe « en cloche »). Cette variable aléatoire gaussienne est d'espérance $\mathrm{E}(Y_i)$ et d'écart-type σ. On rappelle que ces deux paramètres déterminent entièrement la forme de la densité d'une variable aléatoire gaussienne. Plus précisément, l'espérance détermine l'emplacement de la courbe de la densité (plus exactement de son axe de symétrie) et l'écart-type détermine la forme de cette densité (son caractère plus ou moins aplati). Par conséquent, dans le cas du MRLSG, les densités de Y_i et Y_j sont respectivement situées autour de $\mathrm{E}(Y_i)$ et de $\mathrm{E}(Y_j)$ et ont le même forme, puisque dans ce modèle les écarts-type de Y_i et Y_j sont les mêmes.

sec:probetaG ### 3.1.2 Propriétés des estimateurs dans le modèle gaussien

L'ajout de la condition CN dans la définition du modèle de régression linéaire simple permet d'obtenir des résultats supplémentaires pour les estimateurs des moindres carrés ordinaires de β_0 et β_1, ainsi que pour l'estimateur de la variance σ^2. Il est commode d'introduire les notations suivantes :

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \quad \text{et} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \tag{3.1}$$

eq:betah_vec

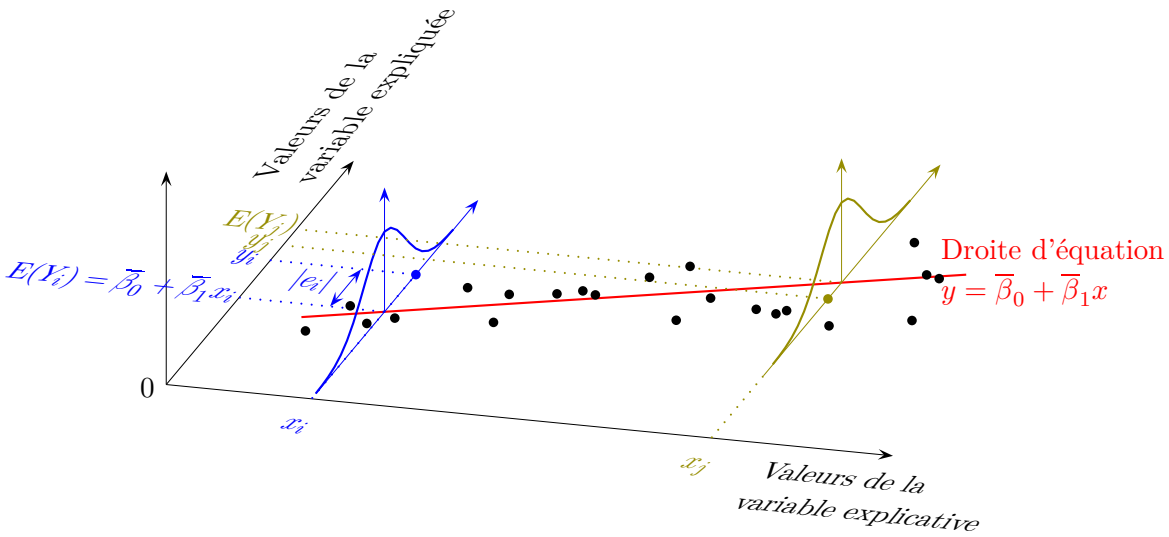
$\hat{\beta}$ est donc un vecteur aléatoire de $\mathbb{R}^2$ et β est un élément de $\mathbb{R}^2$.

pro:mcogauss **Propriété 3.2** *Dans le MRLSG, le couple $(\hat{\beta}_0, \hat{\beta}_1)$ est gaussien. On a $\hat{\beta} \sim \mathcal{N}(\beta, \mathrm{V})$ où $\mathrm{V} = \sigma^2 v$ avec*

$$v = \begin{pmatrix} \dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} & \dfrac{-\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\ \dfrac{-\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} & \dfrac{1}{\sum_{i=1}^n (X_i - \overline{X})^2} \end{pmatrix}$$

Preuve : Exercice

1. On rappelle que l'animation n'est visible qu'avec le lecteur Adobe Reader (voir page 5). Si vous ne disposez pas de ce lecteur, l'animation est visible à http://gremars.univ-lille3.fr/~torres/enseigne/ectrie/mrlsg.



Par rapport à l'individu i, la densité de l'individu j s'est déplacée, puisque d'après la condition CN elle doit être symétrique autour de l'espérance $\mathrm{E}(Y_j)$. La condition CN implique par ailleurs la condition C3, qui impose que les variances sont égales : $\mathrm{V}(Y_i) = \mathrm{V}(Y_j) = \sigma^2$. Par conséquent, la forme des densités pour les individus i et j est la même (voir la section 9.1.1).

fig:figmrlsg

FIGURE 3.1: Modélisation de la relation entre variables dans le modèle de régression linéaire gaussien

coro:mcogauss

Corollaire 3.1 *Dans le MRLSG, les estimateurs des moindres carrés ordinaires de β_0 et de β_1 sont des variables aléatoires gaussiennes. On a*

$$\hat{\beta}_0 \sim \mathcal{N}\Big(\beta_0, \sigma^2\big(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\big)\Big) \qquad \hat{\beta}_1 \sim \mathcal{N}\Big(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\Big)$$

Pour terminer cette section, on complète le résultat obtenu sur la loi de $\hat{\beta}$. On commence par rappeler des définitions introduites dans la section 9.1.

def:rap_loi

Définition 3.2

1. *La loi du χ^2 à m degrés de liberté est la loi suivie par la somme des carrés de m variables aléatoires gaussiennes $\mathcal{N}(0,1)$ indépendantes. On note cette loi $\chi^2(m)$. Autrement dit, si*

$(Z_1, \ldots, Z_m)$ est un m-uplet gaussien $\mathcal{N}(0_m, I_m)$, alors $\sum_{j=1}^m Z_j^2 \sim \chi^2(m)$.

2. *La loi de Student à m degrés de liberté est la loi de la variable aléatoire T définie par*

$$T = \frac{Z}{\sqrt{\frac{C}{m}}}$$

où Z et C sont des variables aléatoires indépendantes, avec $Z \sim \mathcal{N}(0,1)$ et $C \sim \chi^2(m)$. On note $T \sim Student(m)$.

3. *La loi de Fisher à m_1 et m_2 degrés de liberté est la loi de la variable aléatoire F définie par*

$$F = \frac{C_1/m_1}{C_2/m_2}$$

où les variables aléatoires C_1 et C_2 sont indépendantes, avec $C_k \sim \chi^2(m_k)$, $k = 1, 2$.

On admettra temporairement le résultat suivant, qui est une conséquence de la propriété 9.18.

pro:chi2 **Propriété 3.3** *Dans le MRLSG, la variable aléatoire $\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sigma^2}$ suit une loi du χ^2 à $(n-2)$ degrés de libertés. De plus, cette variable aléatoire est indépendante de $\hat{\beta}$.*

r:stu_comb_lin_beta **Corollaire 3.2** *Dans le MRLSG, quels que soient les réels a_0 et a_1 avec $a_0 \neq 0$ ou $a_1 \neq 0$, on a*

$$\frac{a_0(\hat{\beta}_0 - \beta_0) + a_1(\hat{\beta}_1 - \beta_1)}{\sqrt{a_0^2 \hat{\mathrm{V}}(\hat{\beta}_0) + a_1^2 \hat{\mathrm{V}}(\hat{\beta}_1) + 2a_0 a_1 \widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1)}} \sim Student(n-2)$$

où $\hat{\mathrm{V}}(\hat{\beta}_0)$ et $\hat{\mathrm{V}}(\hat{\beta}_1)$ sont les variances estimées de $\hat{\beta}_0$ et $\hat{\beta}_1$ dont les expressions sont données à la propriété 2.10, et $\widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1)$ la covariance estimée entre $\hat{\beta}_0$ et $\hat{\beta}_1$, définie par

$$\widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1) = \hat{\sigma}^2 \frac{-\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

Preuve : Exercice

- Montrer que $a_0\hat{\beta}_0 + a_1\hat{\beta}_1$ est une variable aléatoire gaussienne (utiliser la propriété 3.2).
- Calculer son espérance et sa variance
- Centrer et réduire cette variable afin de former une variable aléatoire qu'on note Z.
- Montrer que $(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$ et qu'elle est indépendante de Z (utiliser la propriété 3.3).
- Former un rapport ayant une loi de Student (utiliser la définition 3.2).
- Utiliser les expressions de $\mathrm{V}(\hat{\beta}_0)$, $\mathrm{V}(\hat{\beta}_1)$, $\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1)$ et de leurs estimateurs, puis simplifier afin d'obtenir le résultat.

Un cas particulier important du résultat précédent est obtenu en choisissant $(a_0, a_1) = (0, 1)$ ou bien $(a_0, a_1) = (1, 0)$.

cor:mcostudent **Corollaire 3.3** *Dans le MRLSG, on a pour $k = 0, 1$*

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_k)}} \sim Student(n-2)$$

sec:testb1

3.2 Test d'une hypothèse sur β_1

sec:test1

3.2.1 Test de significativité

Le paramètre β_1 est le paramètre essentiel dans le MRLS(G). En effet, ce modèle a été construit afin de fournir un cadre d'étude d'une relation supposée exister entre X et Y. Plus précisément, ce modèle stipule que Y est une fonction affine de X et par conséquent la manière dont Y dépend de X peut se mesurer par $\beta_1 = \frac{\mathrm{d}Y}{\mathrm{d}X}$. Une question essentielle (et qui a déjà été abordée à la section 2.4.2) est celle de l'existence d'une telle dépendance.

Pour étudier cette question, on peut poser et résoudre le problème de test suivant :

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

Si la procédure de test utilisée conduit à accepter H_0, on dira que le paramètre β_1 n'est pas *significativement différent de 0*, ou encore que β_1 n'est pas *significatif*. Par extension, lorsqu'on résout le problème de test considéré ici, on dit que l'on fait un *test de significativité de* β_1.

Résoudre ce problème consiste à se fixer un niveau $\alpha \in]0,1[$ puis à choisir une statistique $T_n = T\big((X_1,Y_1),\dots,(X_n,Y_n)\big)$ et une région $\mathscr{T}$ de $\mathbb{R}$ tels que

1. H_0 sera rejetée si et seulement si l'évènement $T_n \in \mathscr{T}$ se réalise ;
2. lorsque H_0 est vraie, la probabilité que cet évènement se réalise ne dépasse pas α.

La première condition définit le test au moyen duquel on choisit entre H_0 et H_1. En reprenant la notation de la section 10.3.2, ce test sera défini par

$$\varphi\big((X_1,Y_1),\dots,(X_n,Y_n)\big) = \begin{cases} 1 \text{ si } T_n \in \mathscr{T} \\ 0 \text{ sinon} \end{cases}$$

La seconde condition impose au test choisi d'avoir le niveau α : le risque de type 1 de ce test ne dépasse pas α.

Pour un niveau α fixé, le choix de la statistique T_n (ou, de manière équivalente, de la fonction T) et de la région $\mathscr{T}$ tel que les deux conditions ci-dessus sont satisfaites n'est pas unique. Autrement dit, pour le problème de test posé, il existe plusieurs tests de niveau α. Pour choisir parmi deux de ces tests, il faudra évaluer leurs risques de type 2 et retenir le test dont le risque de type 2 est le plus petit. De manière plus générale, en suivant l'approche due à Neyman et Pearson (voir page 249), il faut chercher parmi tous les tests de niveau α celui, s'il existe, dont le risque de type 2 est le plus faible.

Comme pour le problème de l'estimation, on abordera dans un premier temps la résolution de ce problème par une approche intuitive ; on présentera ensuite une approche théorique, guidée par l'approche usuelle des tests statistiques.

On mentionne finalement que toutes les définitions et résultats obtenus pour β_1 et le problème de test $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ se transposent directement au paramètre d'ordonnée à l'origine β_0 en changeant simplement l'indice $_1$ et indice $_0$ (sauf pour la désignation des hypothèses H_0 et H_1, bien entendu).

sec:test1int

3.2.2 Approche intuitive

Cette approche repose sur l'enchaînement suivant. β_1 n'est pas connu, mais nous pouvons en avoir une bonne estimation, fournie par $\hat{\beta}_1$, le plus précis des estimateurs linéaires sans biais de β_1. Pour décider si β_1 est nul (H_0 est la bonne hypothèse) ou pas (H_1 est la bonne hypothèse), on peut se baser sur l'observation de la valeur de $\hat{\beta}_1$. En effet, puisque ce dernier est un bon estimateur de β_1, il est probable d'observer que $\hat{\beta}_1$ est proche de 0 lorsque $\beta_1 = 0$, *i.e.*, lorsque H_0 est vraie. Autrement dit, si on est amené à observer que $\hat{\beta}_1$ est éloigné de 0, on observe un évènement dont la probabilité d'occurrence est faible lorsque H_0 est vraie. On juge alors que H_0 n'est pas vraisemblable au vu de ce qu'on observe et on rejette H_0.

Dans une telle approche, il faut se fixer un seuil s, avec $s \in]0, +\infty[$, permettant d'exprimer « $\hat{\beta}_1$ est trop éloigné de 0 » au moyen d'une inégalité telle que $|\hat{\beta}_1| > s$. En reprenant la démarche générale de construction des tests exposée dans la section 10.3.2, la statistique T_n est ici égale à $|\hat{\beta}_1|$ et la région critique $\mathscr{T}$, constituant l'ensemble des valeurs de la statistique qui sont peu vraisemblables lorsque H_0 est vraie, est $\mathscr{T} =]s, +\infty[$. L'évènement $T_n \in \mathscr{T}$ conduisant au rejet de H_0 est donc bien $|\hat{\beta}_1| > s$.

Le nombre s désigne le seuil au delà duquel $\hat{\beta}_1$ est jugé trop éloigné de 0 pour que H_0 soit une hypothèse plausible. La question du choix de s reste posée. Pour guider ce choix, on fait appel à la condition de niveau qui impose que $\mathrm{P}_{H_0}(T_n \in \mathscr{T}) \leq \alpha$ (voir l'inégalité (10.1) et les commentaires qui l'accompagnent, page 249). Cette inégalité s'écrit encore

$$\mathrm{P}_{H_0}(|\hat{\beta}_1| > s) \leq \alpha \tag{3.2}$$

eq:stud1

Choisir s de manière que le test ait un niveau α revient à résoudre en s l'inégalité (3.2). La probabilité qui en constitue le membre de gauche est déterminée par la loi de la variable aléatoire $\hat{\beta}_1$.

Les résultats obtenus dans la section 3.1 nous permettent pour n'importe quel réel $s > 0$ de calculer le membre de gauche de l'inégalité (3.2). En effet, d'après le corollaire 3.3, la loi de la variable aléatoire $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}$ est connue et on peut écrire

$$\begin{aligned}
\mathrm{P}_{H_0}(|\hat{\beta}_1| > s) &= 1 - \mathrm{P}_{H_0}(-s \leq \hat{\beta}_1 \leq s) \\
&= 1 - \mathrm{P}_{H_0}\left(\frac{-s-\beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} \leq \frac{s-\beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\right) \\
&= 1 - \mathrm{P}_{H_0}\left(\frac{-s-\beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} \leq \tau_{n-2} \leq \frac{s-\beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\right) \\
&= 1 - \left[F_{\tau(n-2)}\left(\tfrac{s-\beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\right) - F_{\tau(n-2)}\left(\tfrac{-s-\beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\right)\right]
\end{aligned}$$

où τ_{n-2} est une variable aléatoire suivant une loi de Student à $(n-2)$ degrés de liberté et où $F_{\tau(n-2)}$ désigne la fonction de répartition de cette loi.

Comme la notation P_{H_0} l'indique, cette probabilité doit être calculée en supposant H_0 vraie.

Dans ce cas, $\beta_1 = 0$ et

$$\mathrm{P}_{H_0}(|\hat{\beta}_1| > s) = 1 - \Big[F_{\tau(n-2)}\Big(\frac{s}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\Big) - F_{\tau(n-2)}\Big(\frac{-s}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\Big)\Big] = 2\Big[1 - F_{\tau(n-2)}\Big(\frac{s}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\Big)\Big]$$

où la dernière égalité provient de la symétrie autour de 0 de la densité de la loi de Student$(n-2)$. Par conséquent, la contrainte portant sur le niveau du test, exprimée par l'inégalité (3.2), s'écrit

$$\mathrm{P}_{H_0}(|\hat{\beta}_1| > s) \leq \alpha \iff 2\Big[1 - F_{\tau(n-2)}\Big(\frac{s}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\Big)\Big] \leq \alpha \iff F_{\tau(n-2)}\Big(\frac{s}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\Big) \geq 1 - \tfrac{\alpha}{2}$$

Comme $F_{\tau(n-2)}$ est continue et strictement croissante, la dernière inégalité s'écrit

$$\frac{s}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} \geq F^{-1}_{\tau(n-2)}(1 - \tfrac{\alpha}{2})$$

Le membre de droite de cette inégalité est par définition le quantile d'ordre $1 - \frac{\alpha}{2}$ de $F_{\tau(n-2)}$, ou encore le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student à $n-2$ degrés de liberté. On notera $\tau_{n-2;1-\frac{\alpha}{2}}$ ce quantile. Finalement, le test de la forme « On rejette H_0 et on accepte H_1 si on observe que $|\hat{\beta}_1| > s$ » aura le niveau α si le seuil s est choisi de sorte que

$$s \geq \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}$$

On note qu'à ce point, l'imposition de la contrainte (3.2) ne permet pas de dégager une valeur unique de s. Pour cela, on s'intéresse au risque de type 2. On rappelle que la démarche consiste à choisir parmi un ensemble de tests ayant tous un niveau α, celui (ou ceux) pour le(s)quel(s) le risque de type 2 sera toujours le plus faible.

On considère ici les tests de la forme « On rejette H_0 et on accepte H_1 si on observe que $|\hat{\beta}_1| > s$ », avec $s \geq \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}$. Pour tout test de cette forme, le risque de type 2 s'exprime comme [2]

$$\mathrm{P}_{H_1}(|\hat{\beta}_1| \leq s)$$

où la notation P_{H_1} indique que la probabilité est calculée en supposant H_1 vraie, c'est à dire en supposant $\beta_1 \neq 0$. La valeur de cette probabilité dépend de la valeur de β_1 $(\neq 0)$ choisie pour effectuer le calcul. Cependant, quelle que soit cette valeur, on voit que cette probabilité est une fonction croissante de s. [3] Par conséquent, si on cherche le test de la forme donnée ci-dessus ayant le plus petit risque de type 2, il faut choisir le seuil s le plus petit possible. Sachant que pour que le test soit de niveau α il faut que s ne soit pas plus petit que $\tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}$, on est conduit à choisir

$$s = \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}$$

Le test ainsi obtenu consiste donc à rejeter $H_0 : \beta_1 = 0$ et à accepter $H_1 : \beta_1 \neq 0$ au niveau α si on observe $|\hat{\beta}_1| > \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}$, ou, de manière équivalente, si on observe

$$\left|\frac{\hat{\beta}_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\right| > \tau_{n-2;1-\frac{\alpha}{2}}$$

2. On rappelle que le risque de type 2 est la probabilité que l'on a de rejeter H_1 lorsque cette dernière est supposée vraie (voir la section 10.3.2.3).

3. Pour toute variable aléatoire réelle U et pour toute paire de réels (s_1, s_2) tels que $s_1 < s_2$, l'évènement $U \leq s_1$ implique l'évènement $U \leq s_2$ et il est donc au moins aussi probable d'observer le second que le premier.

Ce test est appelé test de Student, qu'on définit formellement.

student_signi_bilat

Définition 3.3 *Dans le MRLSG, on appelle test de Student de niveau α de $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ le test défini par :*

On rejette H_0 et on accepte H_1 si on observe $|T| > \tau_{n-2;1-\frac{\alpha}{2}}$; on rejette H_1 et on accepte H_0 sinon

où $\tau_{n-2;1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi Student$(n-2)$, et T est la statistique définie par

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}$$

et appelée statistique de Student (ou « T » de Student) associée à H_0.

Sous la formulation précédente du test, la statistique de test est $|T|$ et la région critique est $]\tau_{n-2;1-\frac{\alpha}{2}}, +\infty[$. Pour reprendre la notation de la section 10.3.2, le test de Student de niveau α est défini par

$$\varphi((X_1, Y_1), \ldots, (X_n, Y_n)) = \begin{cases} 1 \text{ si } |T| > \tau_{n-2;1-\frac{\alpha}{2}} \\ 0 \text{ sinon} \end{cases}$$

sec:test1theo

3.2.3 Approche théorique

Dans la section précédente, le test a été introduit en partant d'un principe raisonnable : $\hat{\beta}_1$ est une bonne approximation de β_1 et si H_0 est vraie, il est peu probable d'observer un évènement tel que $|\hat{\beta}_1| > s$, pour une valeur bien choisie de s. Dans une telle approche, la forme du test (décider H_1 si $|\hat{\beta}_1| > s$) est donnée *a priori* et il reste à chercher le meilleur des tests de niveau α parmi les tests ayant cette forme.

Dans une approche théorique de construction d'un test pour résoudre le problème $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$, on ne se limite pas à chercher le meilleur des tests ayant une forme donnée, mais on cherche plutôt le meilleur test.

Comme rappelé dans la section 10.3.2.4, les tests sont évalués sur la base de leurs risques (types 1 et 2) et le meilleur test pour un problème de test donné est un test UPP au niveau α : c'est un test de niveau α dont le risque de type 2 est inférieur (ou égal) à celui de tout autre test de niveau α.

Dans le cas du MRLSG dans lequel on veut tester $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ on ne peut pas montrer qu'un test UPP au niveau α existe. Autrement dit, on ne peut exhiber un test meilleur que tous les autres. Pour lever cette « indétermination » dans le choix du test à utiliser, on utilise une approche similaire à celle suivie dans la section 2.2 pour résoudre un problème d'estimation : pour estimer un paramètre, on a cherché le meilleur estimateur dans un ensemble d'estimateurs ayant des propriétés (souhaitées) données. Ici, lorsqu'on veut résoudre un problème de test, on cherchera le meilleur test parmi tous les tests ayant des propriétés souhaitées. Parmi les bonnes propriétés qu'on peut attendre d'un test, on retrouve la notion d'absence de biais (voir la définition 10.1) On rappelle qu'un test sans biais est un test pour lequel, quelle que soit la décision considérée, il est toujours plus probable de prendre cette décision lorsqu'elle correspond à une bonne décision que

lorsqu'elle correspond à un mauvaise décision (voir les commentaires qui suivent la définition 10.1). Plus formellement, si le test est basé sur une statistique $T_n = T(X_1, Y_1, \ldots, X_n)$, et est de la forme « on décide H_1 si on observe l'évènement $T_n \in \mathscr{T}$ »[4] alors ce test est de niveau α et sans biais dès que

$$\mathrm{P}_{H_0}(T_n \in \mathscr{T}) \leq \alpha \leq \mathrm{P}_{H_1}(T_n \in \mathscr{T})$$

La démarche consistant à rechercher le/les meilleur/s test/s parmi les tests sans biais au niveau α est plus difficile à suivre que celle utilisée pour déterminer le meilleur estimateur (linéaire et sans biais) des paramètres du modèle. Aussi on ne présentera pas la preuve du résultat principal de cette section.

th:optstudent

Théorème 3.1 *Pour tester $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ dans le MRLSG, le meilleur test parmi tous les tests sans biais au niveau α est le test de Student de la définition 3.3.*

Ce résultat dit que (1) le test de Student est un test sans biais au niveau α et que (2) il n'existe pas d'autre test sans biais au niveau α — donc semblables au test de Student en ce qui concerne le risque de type 1 — dont le risque de type 2 soit plus petit que celui du test de Student. Le but étant de chercher des tests ayant les plus petits risques possibles, ce résultat est un résultat d'optimalité du test de Student, dans le contexte du MRLSG.

test-generalis-egal

3.2.4 Test d'une valeur quelconque de β_1

On s'est intéressé, pour les raisons qu'on a évoquées au début de la section 3.2.1, à un problème de test qui revenait à décider si β_1 valait 0 ou non. Même si la valeur 0 surgit naturellement dans beaucoup de problèmes de tests, on peut être intéressé par des problèmes dans lesquels la valeur testée est quelconque.

Soit b un réel connu. On veut tester $H_0 : \beta_1 = b$ contre $H_1 : \beta_1 \neq b$. En suivant l'approche développée dans la section 3.2.2, on note que si H_0 est vraie, la distance entre β_1 et b est nulle. On basera donc le test sur la distance entre $\hat{\beta}_1$ et b et on rejettera H_0 si on observe que cette distance est trop grande. Le test sera donc basé sur l'inégalité $|\hat{\beta}_1 - b| > s$. La démarche est ensuite la même que dans la section 3.2.2. On choisit d'abord s de manière que le test soit de niveau α, α étant fixé *a priori*. On doit donc résoudre en s l'inégalité $\mathrm{P}_{H_0}(|\hat{\beta}_1 - b| > s) \leq \alpha$. En effectuant les mêmes développements qu'en 3.2.2, on obtient

$$\mathrm{P}_{H_0}(|\hat{\beta}_1 - b| > s) = 1 - \left[F_{\tau(n-2)}\left(\frac{s+b-\beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\right) - F_{\tau(n-2)}\left(\frac{-s+b-\beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\right)\right]$$

et comme on suppose $H_0 : b = \beta_1$ vraie, cette probabilité est simplement $2\left[1 - F_{\tau(n-2)}\left(\frac{s}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\right)\right]$. Pour que le test soit de niveau α, il faut donc que s soit supérieur à $\tau_{n-2;1-\frac{\alpha}{2}}\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}$. Par ailleurs, la minimisation du risque de type 2 conduit à choisir $s = \tau_{n-2;1-\frac{\alpha}{2}}\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}$.

Le test consiste donc à rejeter $H_0 : \beta_1 = b$ et à accepter $H_1 : \beta_1 \neq b$ au niveau α si on observe

$$\left|\frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\right| > \tau_{n-2;1-\frac{\alpha}{2}}$$

4. Pour reprendre la notation de la section 10.3.2, le test, noté φ, est défini par $\varphi_n = 1 \iff T_n \in \mathscr{T}$.

On a définition semblable à celle introduite précédemment dans le cas où on avait choisi $b = 0$.

def:studn0 **Définition 3.4** *Dans le MRLSG, on appelle test de Student de niveau α de $H_0 : \beta_1 = b$ contre $H_1 : \beta_1 \neq b$ le test défini par :*

On rejette H_0 et on accepte H_1 si on observe $|T(b)| > \tau_{n-2;1-\frac{\alpha}{2}}$; on rejette H_1 et on accepte H_0 sinon

où $T(b)$ est la statistique définie par

$$T(b) = \frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}$$

et appelée statistique de Student (ou « T » de Student) associée à H_0.

Il reste à montrer que ce test possède de bonnes propriétés. Pour cela, définissons les variables $Z_i = Y_i - bX_i$, $i = 1, \ldots, n$ et considérons les implications de C1 et CN sur les couples $(X_1, Z_1), \ldots, (X_n, Z_n)$. La condition C1 ne portant que sur $X_1, \ldots, X_n$ est évidemment satisfaite. Par ailleurs, si CN est satisfaite, alors on en déduit (en utilisant une démarche identique à celle utilisée dans la preuve de la propriété 3.1) que le vecteur $\boldsymbol{Z}$ défini par $\boldsymbol{Z} = \boldsymbol{Y} - b\boldsymbol{X}$ est gaussien. Avec la condition C1, on calcule alors aisément

$$\begin{aligned}
\mathrm{E}(Z_i) &= \mathrm{E}(Y_i - bX_i) = \beta_0 + (\beta_1 - b)X_i \\
\mathrm{cov}(Z_i, Z_j) &= \mathrm{cov}(Y_i - bX_i, Y_j - bX_j) = \mathrm{cov}(Y_i, Y_j)
\end{aligned}$$

Par conséquent, $\exists \delta_0 \in \mathbb{R}, \exists \delta_1 \in \mathbb{R}, \exists \sigma_Z \in]0, \infty[$ t.q. $\boldsymbol{Z} \sim \mathcal{N}(\delta_0 \iota_n + \delta_1 \boldsymbol{X}, \sigma_Z^2 I_n)$. Autrement dit si on a un MRLSG pour les couples de variables $(X_1, Y_1), \ldots, (X_n, Y_n)$, alors on a aussi un MRLSG pour les couples $(X_1, Z_1), \ldots, (X_n, Z_n)$. [5] Les paramètres des deux modèles sont reliés par

$$\delta_0 = \beta_0 \qquad \delta_1 = \beta_1 - b \quad \text{et} \quad \sigma_Z = \sigma$$

Si on se place dans le MRLSG pour $(X_1, Z_1), \ldots, (X_n, Z_n)$, les résultats des sections précédentes permettent de dire que pour tester $H_0 : \delta_1 = 0$ contre $H_1 : \delta_1 \neq 0$ au niveau α, le meilleur des tests parmi les tests sans biais au niveau α est le test de Student. Il consiste à rejeter H_0 au niveau α si on observe que

$$\left| \frac{\hat{\delta}_1}{\sqrt{\hat{\mathrm{V}}(\hat{\delta}_1)}} \right| > \tau_{n-2;1-\frac{\alpha}{2}}$$

où $\hat{\delta}_1$ et $\hat{\mathrm{V}}(\hat{\delta}_1)$ sont respectivement l'estimateur des moindres carrés ordinaires de δ_1 et l'estimateur de la variance de ce dernier, obtenus par les méthodes du chapitre précédent. On a en particulier

$$\hat{\delta}_1 = \frac{\sum_{i=1}^n (Z_i - \overline{Z})X_i}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

Par définition des Z_i on a $\overline{Z} = \overline{Y} - b\overline{X}$, et le numérateur ci-dessus s'écrit

$$\sum_{i=1}^n (Z_i - \overline{Z})X_i = \sum_{i=1}^n \left[Y_i - \overline{Y} - b(X_i - \overline{X}) \right] X_i = \sum_{i=1}^n (Y_i - \overline{Y})X_i - b\sum_{i=1}^n (X_i - \overline{X})X_i$$

5. Ces deux modèles sont en fait identiques puisque la réciproque est vraie : si C1 et CN sont vérifiées pour les couples $(X_1, Z_1), \ldots, (X_n, Z_n)$, alors elles le sont aussi pour les couples $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Donc

$$\hat{\delta}_1 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})X_i - b\sum_{i=1}^n (X_i - \overline{X})X_i}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})X_i}{\sum_{i=1}^n (X_i - \overline{X})^2} - b = \hat{\beta}_1 - b$$

D'autre part

$$\hat{\delta}_0 = \overline{Z} - \hat{\delta}_1 \overline{X} = (\overline{Y} - b\overline{X}) - (\hat{\beta}_1 - b)\overline{X} = \overline{Y} - \hat{\beta}_1 \overline{X} = \hat{\beta}_0$$

Finalement, pour calculer $\hat{\mathrm{V}}(\hat{\delta}_1)$, on utilise la formule donnée à la section 2.5. On a

$$\hat{\mathrm{V}}(\hat{\delta}_1) = \frac{\hat{\sigma}_Z^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

où $\hat{\sigma}_Z^2$ est l'estimateur de la variance σ_Z^2 des Z_i présenté à la section 2.5, basé sur les résidus de l'estimation de δ_0 et de δ_1 par moindres carrés. Le i^{e} résidu est

$$Z_i - \hat{Z}_i = Y_i - bX_i - (\hat{\delta}_0 + \hat{\delta}_1 X_i)$$

En utilisant les expressions obtenues pour $\hat{\delta}_0$ et $\hat{\delta}_1$ on obtient

$$Z_i - \hat{Z}_i = Y_i - bX_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i - bX_i) = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = \hat{\varepsilon}_i$$

Par conséquent

$$\hat{\sigma}_Z^2 = \frac{1}{n-2}\sum_{i=1}^n (Z_i - \hat{Z}_i)^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\sigma}^2$$

et donc $\hat{\mathrm{V}}(\hat{\delta}_1) = \hat{\mathrm{V}}(\hat{\beta}_1)$.

On vient de montrer que la statistique sur laquelle est basé le test de Student pour tester $\delta_1 = 0$ contre $\delta_1 \neq 0$ s'écrit

$$\frac{\hat{\delta}_1}{\sqrt{\hat{\mathrm{V}}(\hat{\delta}_1)}} = \frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}$$

Le test de Student de la définition 3.4 servant à tester $\beta_1 = b$ contre $\beta_1 \neq b$ et le test de Student servant à tester $\delta_1 = 0$ contre $\delta_1 \neq 0$ sont donc définis par le même évènement (la même inégalité). De plus, puisque $\beta_1 = b \iff \delta_1 = 0$, les hypothèses testées sont les mêmes. Par conséquent les deux tests ont les mêmes risques de type 1 et de type 2 et ils conduisent toujours tous les deux à la même décision. Par conséquent, ces deux tests sont les mêmes. L'optimalité (directement déduite du théorème 3.1) obtenue dans le MRLSG pour $(X_1, Z_1), \ldots, (X_n, Z_n)$ est donc équivalente à l'optimalité du test de la définition 3.4. On a donc démontré la propriété suivante.

pro:test_stu_b

Propriété 3.4 *Dans le MRLSG, pour tester $H_0 : \beta_1 = b$ contre $H_1 : \beta_1 \neq b$, le test de Student défini par*

On rejette H_0 et on accepte H_1 au niveau α si on observe $\left|\frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}\right| > \tau_{n-2;1-\frac{\alpha}{2}}$

est le meilleur parmi les tests sans biais au niveau α.

sec:stu_test_ineg

3.2.5 Test d'une inégalité sur β_1

Dans les problèmes de test étudiés jusqu'à présent, l'hypothèse nulle spécifie que β_1 est égal à une valeur donnée b. Il existe des situations dans lesquelles ce n'est pas la valeur de β_1 qui est intéressante en soi, mais simplement son signe. On sait en effet que si β_1 est positif, alors Y varie dans le même sens que X, et en sens opposé si β_1 est négatif. Il est dans ce cas intéressant de pouvoir disposer d'un test de $H_0 : \beta_1 \leq 0$ contre $H_1 : \beta_1 > 0$. De manière plus générale, on peut être amené à tester $H_0 : \beta_1 \leq b$ contre $H_1 : \beta_1 > b$, où b est une valeur donnée et connue.

On a le résultat suivant.

pro:test_ineq

Propriété 3.5 *Dans le MRLSG, pour tester $H_0 : \beta_1 \leq b$ contre $H_1 : \beta_1 > b$, le test de Student défini par*

On rejette H_0 et on accepte H_1 au niveau α si on observe $\frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} > \tau_{n-2;1-\alpha}$

est le meilleur parmi les tests sans biais au niveau α.

On prendra soin de noter que la contrainte de niveau de ce test (le risque de type 1 ne dépasse pas α) impose d'utiliser le quantile d'ordre $1-\alpha$ de la loi de Student à $n-2$ degrés de liberté (et non le quantile d'ordre $1-\frac{\alpha}{2}$ comme auparavant).

La forme du test peut se comprendre aisément. Si H_0 est vraie, alors il est probable d'observer de petites valeurs de $\frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}$. Si ce n'est pas le cas, c'est à dire si cette statistique dépasse un certain seuil s, on décidera que H_1 est vraie. Le choix du seuil est guidé comme précédemment par la contrainte sur le risque de type 1. La difficulté supplémentaire par rapport aux problèmes de test étudiés précédemment, est qu'ici, même si H_0 est supposée vraie, la loi de $\frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}$ n'est pas connue, et on ne peut donc pas calculer le RT1. Ceci n'est pas un obstacle puisqu'il faut noter que dans l'approche classique des tests d'hypothèses, on n'a pas besoin de *calculer* de RT1, mais simplement de s'assurer qu'il est *borné supérieurement* par le niveau α qu'on a choisi, ce qu'il est facile d'obtenir ici. Pour un seuil s, le RT1 est la fonction qui à tout $\beta_1 \in]-\infty, b]$ (*i.e.*, H_0 est supposée vraie) associe la probabilité de décider H_1

$$\mathrm{P}_{H_0}\Big(\frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} > s\Big)$$

Or pour tout $\beta_1 \leq b$,

$$\frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}}$$

et donc

$$\mathrm{P}_{H_0}\Big(\frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} > s\Big) \leq \mathrm{P}_{H_0}\Big(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} > s\Big) \tag{3.3}$$

eq:ineqstu

D'après le corollaire 3.1, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} \sim$ Student$(n-2)$, et la probabilité du membre de droite de (3.3) est inférieure ou égale à α si et seulement si $s \geq \tau_{n-2;1-\alpha}$. Donc pour tout choix de s dans $[\tau_{n-2;1-\alpha}; +\infty[$, le RT1 (membre de gauche de l'inégalité 3.3) sera inférieur ou égal à α.

Pour choisir un seuil dans cet intervalle, on procède comme précédemment, en s'intéressant au risque de type 2. Celui-ci est défini par des probabilités de la forme

$$\mathrm{P}_{H_1}\Big(\frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} \leq s\Big)$$

On voit alors que si veut choisir le seuil $s \in [\tau_{n-2;1-\alpha}; +\infty[$ de manière à minimiser ces probabilités, il faut prendre la plus petite valeur possible, *i.e.*, $s = \tau_{n-2;1-\alpha}$.

On vient d'obtenir le test de la propriété 3.5. Il reste à montrer que ce test est optimal parmi les tests sans biais au niveau α. Comme précédemment, ce résultat sera admis.

3.3 Tests d'hypothèses portant sur β_0 et β_1

Jusqu'à présent, les hypothèses formulées ne portent que sur un seul des deux paramètres. Le cas de β_1 a été traité en détail et celui de β_0 se traite par une démarche identique, adaptée au paramètre d'ordonnée à l'origine.

Dans cette nouvelle section, on s'intéresse à des tests d'hypothèses qui impliquent simultanément les deux paramètres. On distingue deux cas. Dans un premier temps, les hypothèses considérées portent sur une combinaison linéaire donnée de β_0 et de β_1. Cela revient à introduire un nouveau paramètre défini par cette combinaison linéaire et le méthodes de test seront semblables à celles déjà développées.

Dans un second temps, on étudiera des problèmes de test dans lesquels H_0 et H_1 sont « bi-dimensionnelles » : elles portent simultanément sur les deux paramètres β_0 et β_1, mais chacun intervenant séparément de l'autre. Pour résoudre ce type de problème, on ne peut adapter les méthodes présentées dans les sections précédentes. En revanche, on verra qu'on peut essayer de les combiner pour aboutir à une procédure de test.

sec:test_Ha

3.3.1 Test sur une combinaison linéaire de β_0 et de β_1

3.3.1.1 Cas général : test sur la valeur de $a_0\beta_0 + a_1\beta_1$

Le corollaire 3.2, duquel on tire le résultat (corollaire 3.3) utilisé pour former les tests décrits ci-dessus, permet d'obtenir aisément un test de niveau α pour une hypothèse portant sur la valeur du paramètre γ, défini comme $\gamma = a_0\beta_0 + a_1\beta_1$, où a_0 et a_1 sont des réels connus et fixés, tous les deux non nuls (si les deux sont nuls, le problème n'est d'aucun intérêt, et si l'un des deux est nul, on est ramené à un test sur la valeur d'un seul des paramètres).

Considérons le problème de test $H_0 : \gamma = r$ contre $H_1 : \gamma \neq r$, où r est un réel connu. La démarche pour obtenir un test dans ce cas est exactement la même que celle qui a été utilisée jusqu'à présent. On part du constat que si H_0 est vraie, alors le meilleur estimateur linéaire sans biais de γ devrait être proche de r. Si on observe que ce n'est pas le cas, on décide H_1. Ce meilleur estimateur étant $\hat{\gamma} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1$, le test sera de la forme « on décide H_1 si on observe $|\hat{\gamma} - r| > s$ ». Il faut alors déterminer quels sont les seuils s pour lesquels le risque de type 1 ne dépasse pas le niveau α qu'on s'est fixé. Puis parmi tous les seuils satisfaisant cette condition, on choisira celui

pour lequel le risque de type 2 est le plus faible. Étant donné la forme de ce test, et en utilisant un argument identique à celui utilisé pour les tests décrits dans les sections précédentes, la deuxième étape conduit à choisir le plus petit des seuils satisfaisant la condition sur le risque de type 1.

La première étape repose sur le résultat 3.2 et permet d'obtenir que si on suppose H_0 vraie, alors

$$\frac{\hat{\gamma} - r}{\sqrt{\hat{\mathrm{V}}(\hat{\gamma})}} \sim \text{Student}(n-2)$$

où $\hat{\mathrm{V}}(\hat{\gamma})$ est l'estimateur sans biais de $\mathrm{V}(\hat{\gamma}) = a_0^2\mathrm{V}(\hat{\beta}_0) + a_1^2\mathrm{V}(\hat{\beta}_1) + 2a_0a_1\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1)$ obtenu en remplaçant dans cette expression les variances et la covariance par leur estimateurs sans biais respectifs. On note au passage que puisque qu'on a choisit $a_0 \neq 0$ et $a_1 \neq 0$, on a $\hat{\mathrm{V}}(\hat{\gamma}) \neq 0$, dès lors qu'il existe i, j tels que $X_i \neq X_j$ (Exercice). La condition sur le risque de type 1 d'un test de la forme « on décide H_1 si on observe $|\hat{\gamma} - r| > s$ » s'écrit

$$\mathrm{P}_{H_0}\Big(\frac{|\hat{\gamma} - r|}{\sqrt{\hat{\mathrm{V}}(\hat{\gamma})}} > \frac{s}{\sqrt{\hat{\mathrm{V}}(\hat{\gamma})}}\Big) \leq \alpha$$

Compte-tenu de ce qui précède, la probabilité dans membre de gauche est égale à $\mathrm{P}(|\tau_{n-2}| > \frac{s}{\sqrt{\hat{\mathrm{V}}(\hat{\gamma})}})$, où τ_{n-2} est une variable aléatoire suivant une loi de Student à $(n-2)$ degrés de liberté. En suivant la même approche que précédemment, l'inégalité ci-dessus équivaut à

$$s \geq \tau_{n-2;1-\frac{\alpha}{2}}\sqrt{\hat{\mathrm{V}}(\hat{\gamma})} \tag{3.4}$$ eq:rt1_stu_comb

Dans la seconde étape, on doit choisir parmi tous les seuils s satisfaisant (3.4) celui pour lequel le test de la forme donnée ci-dessus aura le plus petit risque de type 2. Comme ce risque est défini par $\mathrm{P}_{H_1}\big(|\hat{\gamma} - r| \leq s\big)$ lorsque le seuil choisi est s, il faut choisir ce dernier le plus petit possible, tout en imposant la condition (3.4) obtenue en première étape. On choisira donc le seuil $s^* = \tau_{n-2;1-\frac{\alpha}{2}}\sqrt{\hat{\mathrm{V}}(\hat{\gamma})}$. On résume la démarche par le résultat suivant.

pro:test_comb-lin **Propriété 3.6** *Dans le MRLSG, pour tester $H_0 : a_0\beta_0 + a_1\beta_1 = r$ contre $H_1 : a_0\beta_0 + a_1\beta_1 \neq r$, on utilise le test défini par*

On décide H_1 au niveau α si on observe

$$\frac{|a_0\hat{\beta}_0 + a_1\hat{\beta}_1 - r|}{\sqrt{a_0^2\hat{\mathrm{V}}(\hat{\beta}_0) + a_1^2\hat{\mathrm{V}}(\hat{\beta}_1) + 2a_0a_1\widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1)}} > \tau_{n-2;1-\frac{\alpha}{2}}$$

et on décide H_0 sinon

Ce test est le meilleur parmi les tests de niveau α ayant la forme « on décide H_1 si on observe $|a_0\hat{\beta}_0 + a_1\hat{\beta}_1 - r| > s$ ».

On peut démontrer un résultat plus fort, établissant que ce test est optimal parmi les tests sans biais. Cette optimalité peut s'obtenir facilement comme une conséquence de l'optimalité des tests de Student dans le cas d'un test sur la valeur du paramètre β_0 (voir le résultat de la propriété 3.4 transposé au cas de β_0). Pour cela, avec la notation $\gamma = a_0\beta_0 + a_1\beta_1$ introduite précédemment, on

écrit $\beta_0 = \frac{\gamma}{a_0} - \frac{a_1}{a_0}\beta_1$ (on rappelle qu'on a supposé que a_0 et a_1 sont non-nuls). Par conséquent, on a

$$\mathrm{E}(Y_i) = \frac{\gamma}{a_0} - \frac{a_1}{a_0}\beta_1 + \beta_1 X_i = \delta_0 + \delta_1 Z_i, \qquad i = 1, \ldots, n$$

avec

$$\delta_0 = \frac{\gamma}{a_0} = \beta_0 + \frac{a_1}{a_0}\beta_1, \quad \delta_1 = \beta_1, \quad Z_i = X_i - \frac{a_1}{a_0},\ i = 1, \ldots, n \tag{3.5}$$ eq:equiv-XZ

On vérifie facilement qu'avec les égalités (3.5), on a un MRLSG de Y sur X si et seulement si on a un MRLSG de Y sur Z (voir la définition 3.1). Plus précisément, la loi de $(X_1, Y_1), \ldots, (X_n, Y_n)$ satisfait les conditions C1 et CN si et seulement si la loi de $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ satisfait ces mêmes conditions (Exercice). En particulier, puisqu'on s'est initialement placé dans le MRLSG de Y sur X, la condition CN et les propriétés des n-uplets gaussiens permettent d'écrire que

$$\boldsymbol{Y} \sim \mathcal{N}(\delta_0 \iota_n + \delta_1 \boldsymbol{Z}, \eta^2 I_n) \tag{3.6}$$ eq:mrlsg-yz

$\boldsymbol{Z} = (Z_1, \ldots, Z_n)$, et où on a $\eta = \sigma$, et δ_0 et δ_1 donnés par (3.5). Comme $a_0\beta_0 + a_1\beta_1 = r \iff \delta_0 = \frac{r}{a_0}$, tester $H_0 : a_0\beta_0 + a_1\beta_1 = r$ contre $H_1 : a_0\beta_0 + a_1\beta_1 \neq r$ dans le premier modèle revient à tester $H_0 : \delta_0 = \frac{r}{a_0}$ contre $H_1 : \delta_0 \neq \frac{r}{a_0}$ dans le second. Si on se place dans le modèle initial, on peut utiliser le test de la propriété 3.4. Si on se place dans le MRLSG de Y sur Z, on constate que le problème de test est du même type que celui présenté à la section 3.2.4. Comme on est dans le contexte d'un MRLSG, la propriété 3.4 établit que pour résoudre ce problème, le test sans biais optimal est le test de Student. Si on peut montrer qu'il coïncide avec le test de la propriété 3.6, on aura montré l'optimalité annoncée dans cette propriété.

On examine donc le test de Student de $H_0 : \delta_0 = \frac{r}{a_0}$ contre $H_1 : \delta_0 \neq \frac{r}{a_0}$ dans le MRLSG de Y sur Z, afin de le ré-exprimer à l'aide des variables Y et X. D'après la propriété 3.4, ce test est défini par :

« on décide H_1 au niveau α si on observe $\frac{|\hat{\delta}_0 - \frac{r}{a_0}|}{\sqrt{\hat{\mathrm{V}}(\hat{\delta}_0)}} > \tau_{n-2;1-\frac{\alpha}{2}}$ »

où $\hat{\delta}_0$ et $\hat{\mathrm{V}}(\hat{\delta}_0)$ sont l'estimateur des moindres carrés ordinaires de δ_0 et l'estimateur sans biais de la variance de $\hat{\delta}_0$, respectivement. Le premier est donné par le théorème 2.1 et le second par le corollaire 2.1, appliqués dans le contexte du MRLSG de Y sur Z. Pour obtenir le résultat recherché (l'équivalence des tests de Student dans les deux modèles), on exprime les estimateurs en fonction des observations de X et de Y. On commence par noter que d'après (3.5), on a $\overline{Z} = \overline{X} - \frac{a_1}{a_0}$ et donc

$$Z_i - \overline{Z} = X_i - \frac{a_1}{a_0} - \overline{X} - \frac{a_1}{a_0} = X_i - \overline{X}$$

Par conséquent

$$\hat{\delta}_1 = \frac{\sum_{i=1}^n (Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^n (Z_i - \overline{Z})^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \hat{\beta}_1 \tag{3.7}$$ eq:mco-YZ

et donc

$$\hat{\delta}_0 = \overline{Y} - \overline{Z}\hat{\delta}_1 = \overline{Y} - (\overline{X} - \frac{a_1}{a_0})\hat{\beta}_1 = \hat{\beta}_0 + \frac{a_1}{a_0}\hat{\beta}_1 \tag{3.8}$$ eq:mco-YZ0

ce qui permet d'exprimer les estimateurs des paramètres du MRLSG de Y sur Z en fonction des variables Y et X.[6] Par ailleurs, le corollaire 2.1 permet d'écrire

$$\mathrm{V}(\hat{\delta}_0) = \eta^2\Big(\frac{1}{n} + \frac{\overline{Z}^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}\Big) = \eta^2\Big(\frac{1}{n} + \frac{(\overline{X} - \frac{a_1}{a_0})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\Big)$$

L'estimateur sans biais de $\mathrm{V}(\hat{\delta}_0)$ est obtenu à partir de cette expression, en remplaçant la variance inconnue η^2 par son estimateur sans biais formé à partir de la somme des carrés des résidus de l'estimation par moindres carrés ordinaires de δ_0 et δ_1 (voir le corollaire 2.2). Le i^{e} résidu est par définition $Y_i - \hat{\delta}_0 - \hat{\delta}_1 Z_i$. Compte-tenu de l'expression de Z_i (voir (3.5)) et des expressions de $\hat{\delta}_0$ et $\hat{\delta}_1$ (voir (3.7) et (3.8)), on peut écrire ce résidu comme :

$$Y_i - \hat{\delta}_0 - \hat{\delta}_1 Z_i = Y_i - (\hat{\beta}_0 + \frac{a_1}{a_0}\hat{\beta}_1) - \hat{\beta}_1(X_i - \frac{a_1}{a_0}) = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = \hat{\varepsilon}_i$$

Autrement dit, les résidus de l'estimation par moindres carrés ordinaires dans le MRLSG de Y sur Z coïncident avec ceux de l'estimation du MRLSG de Y sur X. La somme des carrés des résidus est donc aussi la même dans les deux modèles et l'estimation de η^2 est la même que celle de σ^2, *i.e.*, $\hat{\eta}^2 = \hat{\sigma}^2$. On peut alors obtenir l'expression de $\hat{\mathrm{V}}(\hat{\delta}_0)$ à partir de celle de $\hat{\mathrm{V}}(\delta_0)$ (voir ci-dessus) et on a

$$\mathrm{V}(\hat{\delta}_0) = \hat{\sigma}^2\Big(\frac{1}{n} + \frac{(\overline{X} - \frac{a_1}{a_0})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\Big)$$

La statistique de Student pour tester $H_0 : \delta_0 = \frac{r}{a_0}$ contre $H_0 : \delta_0 \neq \frac{r}{a_0}$ est $\frac{|\hat{\delta}_0 - \frac{r}{a_0}|}{\sqrt{\hat{\mathrm{V}}(\hat{\delta}_0)}}$. En utilisant les expressions obtenues ci-dessus, on a

$$\frac{|\hat{\delta}_0 - \frac{r}{a_0}|}{\sqrt{\hat{\mathrm{V}}(\hat{\delta}_0)}} = \frac{|\hat{\beta}_0 + \frac{a_1}{a_0}\hat{\beta}_1 - \frac{r}{a_0}|}{\sqrt{\hat{\sigma}^2\Big(\frac{1}{n} + \frac{(\overline{X} - \frac{a_1}{a_0})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\Big)}} \tag{3.9}$$

eq:stat-stu-yz

Or

$$\begin{aligned}\frac{1}{n} + \frac{(\overline{X} - \frac{a_1}{a_0})^2}{\sum_{i=1}^n (X_i - \overline{X})^2} &= \frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} + \frac{\big(\frac{a_1}{a_0}\big)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} - 2\frac{a_1}{a_0}\frac{\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\ &= \frac{1}{a_0^2}\Big[a_0^2\Big(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\Big) + \frac{a_1^2}{\sum_{i=1}^n (X_i - \overline{X})^2} + 2a_0 a_1 \frac{-\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}\Big]\end{aligned}$$

et donc, en utilisant les expressions de $\hat{\mathrm{V}}(\hat{\beta}_0)$ et $\hat{\mathrm{V}}(\hat{\beta}_1)$ données dans la propriété 2.10, ainsi que celle de $\widehat{\mathrm{cov}}(\hat{\beta}_0, \beta_1)$ donnée dans le corollaire 3.2, on peut écrire

$$\hat{\sigma}^2\Big(\frac{1}{n} + \frac{(\overline{X} - \frac{a_1}{a_0})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\Big) = \frac{1}{a_0^2}\Big(a_0^2\hat{\mathrm{V}}(\hat{\beta}_0) + a_1^2\hat{\mathrm{V}}(\hat{\beta}_1) + 2a_0 a_1 \widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1)\Big)$$

6. On remarquera que les estimateurs des moindres carrés obtenus dans chacun des deux MRLSG satisfont les mêmes relations que les paramètres eux-mêmes de ces deux modèles. Plus formellement, la relation entre les paramètres des modèles est donnée par (3.5), et on vient d'obtenir que $\hat{\delta}_0 = \hat{\beta}_0 + \frac{a_1}{a_0}\hat{\beta}_1$ et $\hat{\delta}_1 = \hat{\beta}_1$. On peut voir ceci comme l'illustration d'une bonne propriété de la méthode d'estimation.

On peut donc réécrire l'égalité (3.9)

$$\frac{|\hat{\delta}_0 - \frac{r}{a_0}|}{\sqrt{\hat{\mathrm{V}}(\hat{\delta}_0)}} = \frac{\frac{1}{|a_0|}|a_0\hat{\beta}_0 + a_1\hat{\beta}_1 - r|}{\sqrt{\frac{1}{a_0^2}\left(a_0^2\hat{\mathrm{V}}(\hat{\beta}_0) + a_1^2\hat{\mathrm{V}}(\hat{\beta}_1) + 2a_0a_1\widehat{\mathrm{cov}}(\hat{\beta}_0,\hat{\beta}_1)\right)}}$$
$$= \frac{|a_0\hat{\beta}_0 + a_1\hat{\beta}_1 - r|}{\sqrt{a_0^2\hat{\mathrm{V}}(\hat{\beta}_0) + a_1^2\hat{\mathrm{V}}(\hat{\beta}_1) + 2a_0a_1\widehat{\mathrm{cov}}(\hat{\beta}_0,\hat{\beta}_1)}}$$

Cette dernière égalité montre donc que la statistique de Student associée au test de $H_0 : \delta_0 = \frac{r}{a_0}$ contre $H_1 : \delta_0 \neq \frac{r}{a_0}$ dans le MRLSG de Y sur Z coïncide avec la statistique associée au test de $H_1 : a_0\beta_0 + a_1\beta_1 = r$ contre $H_1 : a_0\beta_0 + a_1\beta_1 \neq r$ dans le MRLSG de Y sur X, décrit dans la propriété 3.6. Par conséquent,

$$\frac{|\hat{\delta}_0 - \frac{r}{a_0}|}{\sqrt{\hat{\mathrm{V}}(\hat{\delta}_0)}} > \tau_{n-2;1-\frac{\alpha}{2}} \iff \frac{|a_0\hat{\beta}_0 + a_1\hat{\beta}_1 - r|}{\sqrt{a_0^2\hat{\mathrm{V}}(\hat{\beta}_0) + a_1^2\hat{\mathrm{V}}(\hat{\beta}_1) + 2a_0a_1\widehat{\mathrm{cov}}(\hat{\beta}_0,\hat{\beta}_1)}} > \tau_{n-2;1-\frac{\alpha}{2}}$$

et comme les hypothèses nulles des deux modèles sont les mêmes, le test de la propriété 3.6 et le test de Student dans le MRLSG de Y sur Z conduisent toujours chacun à la même décision que l'autre. Ces deux modèles étant équivalents, les deux tests ont les même propriétés : ils sont équivalents. Comme l'un est optimal parmi les tests sans biais, l'autre l'est également. On résume ce résultat par la propriété suivante.

ro:test_comb-lin_opt

Propriété 3.7 *Dans le MRLSG, pour tester $H_0 : a_0\beta_0 + a_1\beta_1 = r$ contre $H_1 : a_0\beta_0 + a_1\beta_1 \neq r$, le test défini par*

On décide H_1 au niveau α si on observe

$$\frac{|a_0\hat{\beta}_0 + a_1\hat{\beta}_1 - r|}{\sqrt{a_0^2\hat{\mathrm{V}}(\hat{\beta}_0) + a_1^2\hat{\mathrm{V}}(\hat{\beta}_1) + 2a_0a_1\widehat{\mathrm{cov}}(\hat{\beta}_0,\hat{\beta}_1)}} > \tau_{n-2;1-\frac{\alpha}{2}}$$

et on décide H_0 sinon

est le meilleur parmi les tests sans biais de niveau α. On l'appelle test de Student de $H_0 : a_0\beta_0 + a_1\beta_1 = r$ contre $H_1 : a_0\beta_0 + a_1\beta_1 \neq r$.

On termine cette section en mentionnant qu'on peut adapter ce test à des cas où les hypothèses alternatives sont unilatérales (on utilise alors le quantile d'ordre $1-\alpha$ de la loi Student$(n-2)$). On peut également l'adapter pour tester $H_0 : a_0\beta_0 + a_1\beta_1 \leq r$ contre $H_1 : a_0\beta_0 + a_1\beta_1 > r$. Pour cela, on procède en utilisant la même démarche que dans les sections 3.2.1 et 3.2.5. Dans tous les cas, le test obtenu est optimal parmi les tests sans biais.

sec:testEY

3.3.1.2 Un cas particulier important : test sur $\mathbf{E}(Y_i)$

On rappelle que la condition C1 implique que $\mathrm{P}(X_i = x_i) = 1$ pour tout $i = 1, \ldots, n$, et qu'avec la condition CN, on peut trouver des réels β_0 et β_1 tels que l'espérance de Y_i s'écrit $\beta_0 + \beta_1 X_i$, $i = 1, \ldots, n$. On s'intéresse alors à un seul individu (quelconque) i et on veut tester $H_0 : \mathrm{E}(Y_i) = m$

contre $H_1 : \mathrm{E}(Y_i) \neq m$, où m est un réel connu. Avec ce qui a été rappelé, cela revient à tester $H_0 : \beta_0 + \beta_1 X_i = m$ contre $H_1 : \beta_0 + \beta_1 X_i \neq m$. On constate qu'écrit sous cette forme, le problème de test sur $\mathrm{E}(Y_i)$ est un cas particulier des problèmes de test sur une combinaison linéaire de β_0 et β_1, qui ont été étudiés à la section précédente. Il suffit alors d'en appliquer directement les résultats, dans le cas particulier où $a_0 = 1$, $a_1 = X_i$ et $r = m$.

Plus explicitement, le test optimal de niveau α parmi les tests sans biais (voir la propriété 3.7) consiste à décider $H_1 : \mathrm{E}(Y_i) \neq m$ si on observe

$$\frac{|\hat{\beta}_0 + \hat{\beta}_1 X_i - m|}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_0) + X_i^2 \hat{\mathrm{V}}(\hat{\beta}_1) + 2X_i \widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1)}} > \tau_{n-2;1-\frac{\alpha}{2}}$$

En utilisant les expressions de $\hat{\mathrm{V}}(\hat{\beta}_0)$, $\hat{\mathrm{V}}(\hat{\beta}_1)$ et $\widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1)$ (voir la propriété 2.10 et le corollaire 3.2), on montre facilement (Exercice) que

$$\hat{\mathrm{V}}(\hat{\beta}_0) + X_i^2 \hat{\mathrm{V}}(\hat{\beta}_1) + 2X_i \widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1) = \hat{\sigma}^2 \Big[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^n X_j^2 - n\overline{X}^2} \Big]$$

En notant que par définition $\hat{\beta}_0 + \hat{\beta}_1 X_i = \hat{Y}_i$, et le test consiste alors à décider H_1 au niveau α si on observe que

$$|\hat{Y}_i - m| > \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 \Big[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^n X_j^2 - n\overline{X}^2} \Big]}$$

Le terme sous la racine carré est évidemment l'expression de l'estimateur sans biais de $\mathrm{V}(\hat{Y}_i)$ (Exercice).

On note qu'on ne s'intéresse ici qu'à un seul individu, et que ce test est propre à cet individu.[7] En particulier, il est tout à fait possible de décider H_0 au niveau α pour un individu i et décider H_1 au même niveau pour un autre individu $j \neq i$.

sec:test_joint

3.3.2 Test d'une hypothèse jointe sur β_0 et β_1

3.3.2.1 Présentation du problème et de l'approche

On considère à présent le problème de test de $H_0 : \{\beta_0 = 0 \text{ et } \beta_1 = 0\}$ contre $H_1 : \{\beta_0 \neq 0 \text{ ou } \beta_1 \neq 0\}$. Si on essaie de procéder comme auparavant pour dégager une forme du test qui sera utilisé, on doit considérer β_0 et β_1 simultanément, sous la forme d'un couple (β_0, β_1) de $\mathbb{R}^2$, puis décider H_1 si l'estimation $(\hat{\beta}_0, \hat{\beta}_1)$ de ce couple est trop éloigné de 0. La difficulté ici est que l'objet manipulé est de dimension 2 et donc que les lois de probabilité sous-jacentes sont bivariées.

Cependant, on peut se ramener à des objets unidimensionnels de 2 manières. On peut d'abord considérer que $(\hat{\beta}_0, \hat{\beta}_1)$ est un vecteur de $\mathbb{R}^2$ (dont les coordonnées sont aléatoires), et que pour juger si $(\hat{\beta}_0, \hat{\beta}_1)$ est trop éloigné de 0, on peut utiliser la norme euclidienne de ce vecteur $\sqrt{\hat{\beta}_0^2 + \hat{\beta}_1^2}$. Le test consistera alors à décider H_1 si cette norme dépasse un certain seuil, qu'on devra choisir de manière que le risque de type 1 du test ainsi construit ne dépasse pas le niveau donné au départ,

7. En toute rigueur, on devrait indiquer ceci en indexant par i les hypothèses H_0 et H_1 considérées ici.

puis à choisir parmi tous les seuils admissibles celui pour lequel le risque de type 2 est le plus petit possible.

Une deuxième manière de se ramener à des variables aléatoires de dimension 1 consiste à noter que même si H_0 porte sur un couple de paramètres, on peut voir cette hypothèse comme formée à l'aide de plusieurs hypothèses semblables à celles étudiées dans la section 3.3.1. En effet, si on se donne un couple de réels $a = (a_0, a_1) \neq (0,0)$, on peut former l'hypothèse nulle $H_0(a) : a_0\beta_0 + a_1\beta_1 = 0$. On voit alors que H_0 est vraie si et seulement si $H_0(a)$ est vraie pour tout choix possible de $a \neq (0,0)$. Or pour un choix donné de a, la section précédente montre qu'on sait tester de manière optimale (sans biais) $H_0(a) : a_0\beta_0 + a_1\beta_1 = 0$ contre $H_1(a) : a_0\beta_0 + a_1\beta_1 \neq 0$. Par conséquent, on peut essayer de construire un test consistant à décider H_1 s'il existe un a pour lequel on a décidé $H_1(a)$ au moyen du test optimal sans biais présenté à la section 3.3.1.

Ces deux manières de procéder conduisent au même test de H_0 contre H_1. On présentera ce test en suivant la seconde approche. En effet, elle s'appuie sur des tests déjà étudiés à la section 3.3.1, et elle paraît dès à présent avoir de bonnes propriétés puisque ces tests sur lesquels elle est fondée possèdent une propriété d'optimalité (propriété 3.7).

sec:test_fisher_univ

3.3.2.2 Test de Fisher

3.3.2.2.1 La forme du test On se donne $a = (a_0, a_1) \in \mathbb{R}^2$, avec $a \neq (0,0)$, et on introduit le problème de test de $H_0(a) : a_0\beta_0 + a_1\beta_1 = 0$ contre $H_1(a) : a_0\beta_0 + a_1\beta_1 \neq 0$. D'après la section 3.3.1, le test optimal sans biais consiste à décider $H_1(a)$ si la variable aléatoire $T(a)$ définie par

$$T(a) = \frac{|a_0\hat{\beta}_0 + a_1\hat{\beta}_1|}{\sqrt{a_0^2\hat{\mathrm{V}}(\hat{\beta}_0) + a_1^2\hat{\mathrm{V}}(\hat{\beta}_1) + 2a_0a_1\hat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1)}}$$

dépasse un certain seuil, ce seuil ne dépendant pas de a.[8] Si on considère alors la démarche présentée ci-dessus, la forme du test recherché est la suivante :

« on décide H_1 s'il existe $a \in \mathbb{R}^2$, $a \neq (0,0)$, pour lequel on observe $T(a) > s$ »

où s désigne le seuil choisi. Si on note $\mathbb{R}^2_* = \mathbb{R}^2 \setminus \{(0,0)\}$, ce test revient à décider H_1 si on observe $\max_{a\in\mathbb{R}^2_*} T(a) > s$. Pour choisir le seuil s, on procède comme d'habitude : l'imposition de la contrainte sur le niveau du risque de type 1 détermine les seuils s admissibles, puis la minimisation du risque de type 2 permet de choisir une valeur s^* de ce ce seuil.

Si le niveau qu'on se fixe pour le risque de type 1 est α, alors la contrainte sur ce risque est

$$\mathrm{P}_{H_0}(\max_{a\in\mathbb{R}^2_*} T(a) > s) \leq \alpha$$

La résolution en s de cette inégalité nécessite de connaître la loi de la variable aléatoire $\max_{a\in\mathbb{R}^2_*} T(a)$. Pour l'établir, on commence par montrer qu'il existe une forme explicite pour $\max_{a\in\mathbb{R}^2_*} T(a)$.

sec:maxim_Ta

3.3.2.2.2 La maximisation de $T(a)$ Pour obtenir la forme explicite de $\max_{a\in\mathbb{R}^2_*} T(a)$, on ré-exprime d'abord $T(a)$ en utilisant les expressions des estimateurs des variances et covariance

8. Le dénominateur de $T(a)$ n'est jamais nul, dès qu'il existe i, j tels que $X_i \neq X_j$. Exercice.

apparaissant au dénominateur :

$$\hat{\mathrm{V}}(\hat{\beta}_0) = \hat{\sigma}^2 v_{00} \qquad \hat{\mathrm{V}}(\hat{\beta}_1) = \hat{\sigma}^2 v_{11} \qquad \hat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1) = \hat{\sigma}^2 v_{01}$$

où v_{00} et v_{11} sont les éléments diagonaux la matrice v introduite dans la propriété 3.2, et v_{01} est son élément extra-diagonal, ce qui permet d'écrire

$$T(a) = \frac{|a_0\hat{\beta}_0 + a_1\hat{\beta}_1|}{\hat{\sigma}\sqrt{a_0^2 v_{00} + a_1^2 v_{11} + 2a_0 a_1 v_{01}}}$$

On remarque facilement que pour tout $a \in \mathbb{R}_*^2$ et tout réel $t \neq 0$ on a $T(a) = T(ta)$, où ta est défini par (ta_0, ta_1). Supposons qu'il existe $a^* \in \mathbb{R}_*^2$ tel que $T(a^*) = \max_{a\in\mathbb{R}_*^2} T(a)$, c'est à dire $T(a^*) \geq T(a)\ \forall a \in \mathbb{R}_*^2$. Définissons alors $a^{**} = t^* a^*$ où

$$t^* = \frac{1}{\sqrt{a_0^{*2} v_{00} + a_1^{*2} v_{11} + 2a_0^* a_1^* v_{01}}}$$

En utilisant la définition de $a^{**} = (t^* a_0^*, t^* a_0^*)$ et la définition de t^*, on constate que

$$a_0^{**2} v_{00} + a_1^{**2} v_{11} + 2a_0^{**} a_1^{**} v_{01} = t^{*2}(a_0^{*2} v_{00} + a_1^{*2} v_{11} + 2a_0^* a_1^* v_{01}) = 1 \tag{3.10}$$ eq:lastar

Mais on a également

$$T(a^{**}) = T(t^* a^*) = T(a^*) \geq T(a), \quad \forall a \in \mathbb{R}_*^2$$

Cette inégalité et (3.10) montrent que si on cherche le maximum atteint par $T(a)$ lorsque a parcourt $\mathbb{R}_*^2$, on peut se limiter à chercher le maximum atteint par $T(a)$ pour des a dans $A = \{a \in \mathbb{R}_*^2 \mid a_0^2 v_{00} + a_1^2 v_{11} + 2a_0 a_1 v_{01} = 1\}$. Plus formellement, on a $\max_{a\in\mathbb{R}_*^2} T(a) = \max_{a\in A} T(a)$ et pour trouver la forme explicite de $\max_{a\in\mathbb{R}_*^2} T(a)$, on résout le problème de maximisation $\max_{a\in A} T(a)$.

On a évidemment $T(a^*) \geq T(a) \geq 0\ \forall a \in A \iff T(a^*)^2 \geq T(a)^2\ \forall a \in A$. Par conséquent, $T(a^*)$ sera la racine positive de $\max_{a\in A} T(a)^2$. Par ailleurs, pour tout $a \in A$, le dénominateur de $T(a)^2$ est $\frac{1}{\hat{\sigma}^2}$, et puisqu'on maximise $T(a)^2$ sur A, on peut se contenter de maximiser $\frac{1}{\hat{\sigma}^2}(a_0\hat{\beta}_0 + a_1\hat{\beta}_1)^2$ sur A. Il s'agit d'un problème de maximisation de $\frac{1}{\hat{\sigma}^2}(a_0\hat{\beta}_0 + a_1\hat{\beta}_1)^2$ avec la contrainte $a_0^2 v_{00} + a_1^2 v_{11} + 2a_0 a_1 v_{01} = 1$. Puisque la fonction à maximiser et la contrainte sont continûment différentiables sur $\mathbb{R}_*^2$, on peut appliquer la méthode du lagrangien. Le lagrangien est

$$\mathcal{L}(a_0, a_1, \lambda) = \frac{1}{\hat{\sigma}^2}(a_0\hat{\beta}_0 + a_1\hat{\beta}_1)^2 + \lambda(1 - a_0^2 v_{00} + a_1^2 v_{11} + 2a_0 a_1 v_{01})$$

et $a^* = (a_0^*, a_1^*) \in \mathbb{R}_*^2$ est solution du problème s'il existe un réel λ^* tel que

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial a_0}(a_0^*, a_1^*, \lambda^*) = 0 \\ \frac{\partial \mathcal{L}}{\partial a_1}(a_0^*, a_1^*, \lambda^*) = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda}(a_0^*, a_1^*, \lambda^*) = 0 \end{cases}$$

En calculant les dérivées à partir de l'expression de $\mathcal{L}$, ce système s'écrit

$$\begin{cases} \frac{1}{\hat{\sigma}^2}(a_0^*\hat{\beta}_0 + a_1^*\hat{\beta}_1)\hat{\beta}_0 - \lambda^*(a_0^* v_{00} + a_1^* v_{01}) = 0 & \text{(3.11)} \text{ eq:dl\_da0} \\ \frac{1}{\hat{\sigma}^2}(a_0^*\hat{\beta}_0 + a_1^*\hat{\beta}_1)\hat{\beta}_1 - \lambda^*(a_0^* v_{01} + a_1^* v_{11}) = 0 & \text{(3.12)} \text{ eq:dl\_da1} \\ a_0^{*2} v_{00} + a_1^{*2} v_{11} + 2a_0^* a_1^* v_{01} = 1 & \text{(3.13)} \text{ eq:dl\_dmu} \end{cases}$$

Pour résoudre ce système, il est plus commode de le ré-exprimer de manière matricielle. On recourt à la notation introduite au début de ce chapitre à la section 3.1. En particulier, on utilisera le vecteurs $\hat{\beta}$ et a^* de $\mathbb{R}^2$ et la matrice v définis par

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \qquad a^* = \begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix} \qquad v = \begin{pmatrix} v_{00} & v_{01} \\ v_{01} & v_{11} \end{pmatrix}$$

(voir (3.1) et la propriété 3.2). Avec de telles notation, on peut écrire le membre de droite de (3.13) sous la forme $a^{*\top} v a^*$ ($^\top$ désignant l'opérateur de transposition d'une matrice ou d'un vecteur). Par ailleurs, si on empile les deux égalités (3.11) et (3.12) on peut les réécrire simultanément sous la forme

$$\frac{1}{\hat{\sigma}^2}(a_0^*\hat{\beta}_0 + a_1^*\hat{\beta}_1)\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} - \lambda^* \begin{pmatrix} a_0^* v_{00} + a_1^* v_{01} \\ a_0^* v_{01} + a_1^* v_{11} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

ou encore

$$\frac{1}{\hat{\sigma}^2}(a^{*\top}\hat{\beta})\hat{\beta} - \lambda^* v a^* = 0_2$$

Donc le système des équations (3.11) à (3.13) s'écrit

$$\begin{cases} \frac{1}{\hat{\sigma}^2}(a^{*\top}\hat{\beta})\hat{\beta} - \lambda^* v a^* = 0_2 \text{ eq:dl\_da} & (3.14) \\ a^{*\top} v a^* = 1 \text{ eq:dl\_dmuv} & (3.15) \end{cases}$$

Si on pré-multiplie l'égalité (3.14) par $a^{*\top}$, on obtient $\frac{1}{\hat{\sigma}^2}(a^{*\top}\hat{\beta})^2 - \lambda^* a^{*\top} v a^* = 0$, ou encore

$$\lambda^* = \frac{(a^{*\top}\hat{\beta})^2}{\hat{\sigma}^2 a^{*\top} v a^*} = T(a^*)^2$$

la dernière égalité provenant de la réécriture matricielle du numérateur et du dénominateur de $T(a^*)^2$. Notons que comme a^* satisfait (3.15), on a aussi

$$T(a^*)^2 = \lambda^* = \frac{(a^{*\top}\hat{\beta})^2}{\hat{\sigma}^2}$$

Autrement dit, le réel λ^* coïncide avec la valeur maximale atteinte par la fonction $T(a)^2$ à maximiser lorsque a varie dans A. Ceci est le première étape importante dans la résolution du problème de maximisation. Pour enchaîner sur la seconde étape, on note que $a^{*\top}\hat{\beta} = a_0^*\hat{\beta}_0 + a_1^*\hat{\beta}_1 = \hat{\beta}^\top a^*$. Par conséquent, en mettant en facteurs v à gauche et a^* à droite, le membre de droite de (3.14) peut s'écrire

$$\frac{1}{\hat{\sigma}^2}\hat{\beta}\hat{\beta}^\top a^* - \lambda^* v a^* = v\left[(\hat{\sigma}^2 v)^{-1}\hat{\beta}\hat{\beta}^\top - \lambda^* I_2\right] a^*$$

où I_2 est la matrice identité de dimensions $(2,2)$. On doit avoir montré auparavant que v est inversible, ce qui est bien le cas dès que $\exists i, j$ tels que $X_i \neq X_j$ (Exercice). On peut alors réécrire (3.14) :

$$v\left[(\hat{\sigma}^2 v)^{-1}\hat{\beta}\hat{\beta}^\top - \lambda^* I_2\right] a^* = 0_2 \iff \left[(\hat{\sigma}^2 v)^{-1}\hat{\beta}\hat{\beta}^\top - \lambda^* I_2\right] a^* = 0_2 \iff (B - \lambda^* I_2)a^* = 0_2$$

où B désigne la matrice $(\hat{\sigma}^2 v)^{-1}\hat{\beta}\hat{\beta}^\top$. On note alors que la dernière égalité établit que λ^* est une valeur propre de B associée au vecteur propre a^*. Autrement dit, les couples (a^*, λ^*) solutions du

système d'équations (3.11) à (3.13) sont nécessairement des couples (vecteur propre, valeur propre) de B. Ceci achève la seconde étape de la résolution.

Pour terminer, on rapproche les résultats obtenus aux deux étapes : on sait que tout a^* et tout λ^* solution du système (3.11)-(3.13) sont un vecteur et une valeur propres de B, et que $\lambda^* = T(a^*)^2 = \max_{a\in A} T(a)^2$. Par conséquent, λ^* doit être la plus grande des valeurs propres de B. Comme $B = (\hat\sigma^2 v)^{-1}\hat\beta\hat\beta^\top$, son rang est égal à celui de la matrice $\hat\beta\hat\beta^\top$. Étant donnée la forme de cette dernière, elle est nécessairement de rang 1, et donc B aussi. Par conséquent, parmi les deux valeurs propres de B, l'une d'elles est nulle. Et comme la trace d'une matrice est égale à la somme de ses valeurs propres, la valeur propre non nulle de B est égale à la trace de B. On a donc

$$\text{trace}(B) = \text{trace}((\hat\sigma^2 v)^{-1}\hat\beta\hat\beta^\top) = \text{trace}(\hat\beta^\top(\hat\sigma^2 v)^{-1}\hat\beta) = \hat\beta^\top(\hat\sigma^2 v)^{-1}\hat\beta$$

où la troisième égalité provient du fait que $\text{trace}(A_1A_2) = \text{trace}(A_2A_1)$ dès que les produits matriciels A_1A_2 et A_2A_1 sont possibles, et la dernière égalité provient du fait que $\hat\beta^\top(\hat\sigma^2 v)^{-1}\hat\beta$ est une « matrice » de dimensions (1,1) et est donc égale à sa propre trace. Les deux valeurs propres de B sont donc 0 et $\hat\beta^\top(\hat\sigma^2 v)^{-1}\hat\beta$, et λ^* est égale à la plus grande des deux. Comme v est définie positive (Exercice : démontrer que c'est le cas), on a $\hat\beta^\top(\hat\sigma^2 v)^{-1}\hat\beta > 0$, i.e. la plus grande des valeurs propres de B est celle qui est non nulle. On a donc $\lambda^* = \hat\beta^\top(\hat\sigma^2 v)^{-1}\hat\beta$. Comme on a déjà établit que $\lambda^* = T(a^*)^2$, on a

$$T(a^*)^2 = \hat\beta^\top(\hat\sigma^2 v)^{-1}\hat\beta \tag{3.16}$$ eq:maxTa

Il reste à vérifier que $T(a)^2$ atteint bien son maximum en a^*. Pour cela, on peut vérifier la concavité de la fonction $\mathcal{L}(a_0, a_1, \lambda^*)$. On admettra ici que c'est le cas.

sec:Ftest_univ

3.3.2.2.3 Le test La forme explicite obtenue pour $T(a^*)^2$ permet de construire le test décrit au début de cette section. On rappelle qu'on cherche celui-ci de la forme « On décide H_1 si on observe $\max_{a\in\mathbb{R}^2_*} T(a) > s$ ». Étant donné le résultat obtenu sur la maximisation de $T(a)$, ce test consiste à décider H_1 si $T(a^*) > s$. Si α est le niveau fixé pour le test, alors il faut choisir s de sorte que $\text{P}_{H_0}(T(a^*) > s) \leq \alpha$. Comme s est positif[9], $T(a^*) > s \iff T(a^*)^2 > s^2$ et d'après (3.16), on voit qu'il faut choisir s de sorte que

$$\text{P}_{H_0}(\hat\beta^\top(\hat\sigma^2 v)^{-1}\hat\beta > s^2) \leq \alpha \tag{3.17}$$ eq:rt1_fisher

Pour résoudre en s cette inégalité, il faut connaître la loi de la variable aléatoire $\hat\beta^\top(\hat\sigma^2 v)^{-1}\hat\beta$ lorsque H_0 est supposée vraie, c'est à dire lorsqu'on suppose que $\beta = 0$. Ceci est possible grâce au résultat suivant.

pro:test_ficher_univ

Propriété 3.8 *Dans le contexte du MRLSG,* $\frac{1}{2}(\hat\beta - \beta)^\top\hat{\text{V}}(\hat\beta)^{-1}(\hat\beta - \beta)$ *suit une loi de Fisher à* $(2, n-2)$ *degrés de liberté.*

Preuve : Exercice.

C'est une conséquence des propriétés 3.2, 9.17 et 3.3. Il faut montrer qu'on peut écrire

9. Comme $T(a^*) \geq 0$, si s était négatif, la probabilité serait égale à 1 et ne pourrait jamais être plus petite que tout niveau α dans $]0; 1[$.

$(\hat\beta-\beta)^\top\hat{\mathrm{V}}(\hat\beta)^{-1}(\hat\beta-\beta)$ sous la forme d'un rapport de la forme donnée dans la définition de la loi de Fisher (voir la définition 9.5 ou la définition 3.2).

Solution : D'après la propriété 3.2 et la propriété 3.3, si on définit la variable aléatoire $C_1=(\hat\beta-\beta)^\top\mathrm{V}(\hat\beta)^{-1}(\hat\beta-\beta)$, on a

$$C_1=(\hat\beta-\beta)^\top(\sigma^2v)^{-1}(\hat\beta-\beta)\sim\chi^2(2)$$

D'après la propriété 3.3, la variable aléatoire $C_2=(n-2)\frac{\hat\sigma^2}{\sigma^2}$ suit une loi $\chi^2(n-2)$ et est indépendante de C_1. Par définition de la loi de Fisher, on a $\frac{C_1/2}{C_1/(n-2)}\sim F(2,n-2)$. En utilisant l'expression de C_1 et de C_2, on établit facilement que

$$\frac{C_1/2}{C_1/(n-2)}=\frac12(\hat\beta-\beta)^\top(\hat\sigma^2v)^{-1}(\hat\beta-\beta)=\frac12(\hat\beta-\beta)^\top\hat{\mathrm{V}}(\hat\beta)^{-1}(\hat\beta-\beta)$$

■

Ce résultat permet de résoudre (3.17). En effet

$$\mathrm{P}_{H_0}(\hat\beta^\top(\hat\sigma^2v)^{-1}\hat\beta>s^2)=\mathrm{P}_{H_0}(\tfrac12\hat\beta^\top(\hat\sigma^2v)^{-1}\hat\beta>\tfrac12s^2)=\mathrm{P}(F_{2,n-2}>\tfrac12s^2)$$

où la dernière égalité utilise le fait que lorsqu'on suppose H_0 vraie, on a $\beta=0$ et dans ce cas il découle de la propriété 3.8 que $\frac12\hat\beta^\top(\hat\sigma^2v)^{-1}\hat\beta\sim F(2,n-2)$. Par conséquent, (3.17) équivaut à

$$\mathrm{P}(F_{2,n-2}>\tfrac12s^2)\leq\alpha$$

L'inégalité est une égalité lorsque $\frac12s^2$ est égal au quantile d'ordre $1-\alpha$ de la loi $F(2,n-2)$ et sera plus petite que α pour tout nombre supérieur à ce quantile. Par conséquent, s'il faut que s satisfasse (3.17), il faut choisir s de sorte que $\frac12s^2\geq F_{(2,n-2);1-\alpha}$, ou encore $s\in[\sqrt{2F_{(2,n-2);1-\alpha}}\,;+\infty[$. Finalement, parmi tous les seuils s dans cet intervalle, il faut choisir celui pour lequel le RT2 est le plus petit possible. Pour un choix de s donné, le RT2 est

$$\mathrm{P}_{H_1}(\max_{a\in\mathbb{R}^2_*}T(a)\leq s)$$

On constate que cette probabilité est une fonction croissante de s. Par conséquent, si on veut la minimiser, il faut choisir le seuil s le plus petit possible. Comme s doit être dans $[\sqrt{2F_{(2,n-2);1-\alpha}}\,;+\infty[$, le seuil choisi sera $s^*=\sqrt{2F_{(2,n-2);1-\alpha}}$. On a donc le résultat suivant.

pro:Ftest0_univ

Propriété 3.9 *Dans le contexte du MRLSG, parmi tous les tests de niveau α de $H_0:\{\beta_0=\beta_1=0\}$ contre $H_1:\{\beta_0\neq0$ ou $\beta_1\neq0\}$ ayant la forme*

« *on décide H_1 si on observe* $\max_{a\in\mathbb{R}^2_*}T(a)>s$ », *le test le plus puissant est celui pour lequel on choisit* $s=\sqrt{2F_{(2,n-2);1-\alpha}}$. *Ce test est équivalent à*

« *on décide H_1 si on observe* $\frac12\hat\beta^\top\hat{\mathrm{V}}(\hat\beta)^{-1}\hat\beta>F_{(2,n-2);1-\alpha}$ ».

Ce test est appelé test de Fisher de $H_0:\{\beta_0=\beta_1=0\}$ contre $H_1:\{\beta_0\neq0$ ou $\beta_1\neq0\}$.

Preuve : La première partie de la propriété est un résumé du résultat qui la précède. La seconde partie provient du fait que

$$\max_{a\in\mathbb{R}^2_*} T(a) > \sqrt{2F_{(2,n-2);1-\alpha}} \iff \frac{1}{2}\left[\max_{a\in\mathbb{R}^2_*} T(a)\right]^2 > F_{(2,n-2);1-\alpha}$$

et de $\left[\max_{a\in\mathbb{R}^2_*} T(a)\right]^2 = T(a^*)^2 = \hat{\beta}^\top \hat{\mathbb{V}}(\hat{\beta})^{-1}\hat{\beta}$.

Pour justifier la démarche exposée dans cette section, on est parti de l'observation que $H_0 : \{\beta_0 = \beta_1 = 0\}$ est vraie si et seulement si toutes les hypothèses $H_0(a) : a_0\beta_0 + a_1\beta_1 = 0$, $a \in \mathbb{R}^2_*$ sont vraies. Autrement dit H_1 est vraie si et seulement si l'une des $H_1(a)$ est vraie. En examinant la construction du test de Fisher, on peut alors le voir comme un test dans lequel on décide H_1 si on a pu trouver un $a \in \mathbb{R}^2_*$ pour lequel on a décidé $H_1(a)$. Or la décision concernant $H_1(a)$ est prise en utilisant le test sans biais optimal présenté à la section 3.3.1 (propriété 3.7), basé sur l'inégalité $T(a) > s^*$, le seuil s^* étant défini ci-dessus. Ceci fournit une justification à l'utilisation du test de Fisher, puisqu'on peut considérer qu'il est construit en utilisant des tests sans biais optimaux.

Cependant, on ne peut pas en déduire que le test de Fisher est lui même optimal parmi les tests sans biais. En revanche, il existe des résultats établissant l'optimalité de ce test dans une certaine classe de tests. Cette propriété n'est pas abordée ici. [10]

sec:Ftest_gen_univ

3.3.2.3 Généralisations

La démarche précédente se prolonge aisément afin de dériver un test de $H_0 : \{\beta_0 = b_0 \text{ et } \beta_1 = b_1\}$ contre $H_0 : \{\beta_0 \neq b_0 \text{ ou } \beta_1 \neq b_1\}$, où b_0 et b_1 sont des nombres connus quelconques. En effet, en calquant ce qui a été fait auparavant, on note que H_0 est vraie si et seulement si l'hypothèse $H_0(a) : a_0(\beta_0 - b_0) + a_1(\beta_1 - b_1) = 0$ est vraie pour tout $(a_0, a_1) \in \mathbb{R}^2_*$. On peut aussi écrire $H_0(a)$ sous la forme $H_0(a) : a_0\beta_0 + a_1\beta_1 = c(a)$, où $c(a) = a_0b_0 + a_1b_1$. La propriété 3.7 fournit le test sans biais optimal pour tester $H_0(a)$. Ce test consiste à décider $H_1(a)$ si on observe que $T(a) > s$, où, en utilisant les notations matricielles de la section précédente, on peut écrire

$$T(a) = \frac{|a^\top\hat{\beta} - c(a)|}{\hat{\sigma}\sqrt{a^\top v a}}$$

En suivant la démarche utilisée pour dériver le test de Fisher, on construit ici un test de la forme suivante : « on décide H_1 s'il existe a tel qu'on décide $H_1(a)$ en utilisant le test basé sur $T(a)$ ». Étant donnée la forme de ce test, cela revient à décider H_1 s'il existe $a \in \mathbb{R}^2_*$ tel que $T(a) > s$, ou de manière équivalente, si $\max_{a\in\mathbb{R}^2_*} T(a) > s$. En introduisant le vecteur

$$b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

on constate que $c(a) = a^\top b$ et donc que

$$T(a) = \frac{|a^\top(\hat{\beta} - b)|}{\hat{\sigma}\sqrt{a^\top v a}}$$

10. On peut néanmoins se référer à la section 6.1.2 pour une présentation succincte de cette propriété.

En utilisant cette expression, on peut chercher à expliciter $T(a^*) = \max_{a\in\mathbb{R}^2_*} T(a)$ en suivant une démarche identique à celle de la section 3.3.2.2.2. En effectuant un changement de notation et en posant $\hat{\gamma} = \hat{\beta} - b$, la statistique $T(a)$ prend la même forme que dans la section 3.3.2.2.2, et par conséquent on en déduit que (voir (3.16))

$$T(a^*)^2 = \hat{\gamma}^\top(\hat{\sigma}^2 v)^{-1}\hat{\gamma} = (\hat{\beta} - b)^\top(\hat{\sigma}^2 v)^{-1}(\hat{\beta} - b)$$

Le reste de la démarche est identique à celui de la section 3.3.2.2.3, et on aboutit au résultat suivant, semblable à celui de la propriété 3.9.

pro:Ftest_univ **Propriété 3.10** *Dans le contexte du MRLSG, parmi tous les tests de niveau α de $H_0 : \{\beta_0 = b_0$ et $\beta_1 = b_1\}$ contre $H_1 : \{\beta_0 \neq b_0$ ou $\beta_1 \neq b_1\}$ ayant la forme*

« on décide H_1 si on observe $\max_{a\in\mathbb{R}^2_*} T(a) > s$ »,

où $T(a) = \frac{|a^\top(\hat{\beta}-b)|}{\hat{\sigma}\sqrt{a^\top v a}}$, le test le plus puissant est celui pour lequel on choisit $s = \sqrt{2F_{(2,n-2);1-\alpha}}$. Ce test est équivalent à

« on décide H_1 si on observe $\frac{1}{2}(\hat{\beta} - b)^\top\hat{\mathrm{V}}(\hat{\beta})^{-1}(\hat{\beta} - b) > F_{(2,n-2);1-\alpha}$ ».

Ce test est appelé test de Fisher de $H_0 : \{\beta_0 = b_0$ et $\beta_1 = b_1\}$ contre $H_1 : \{\beta_0 \neq b_0$ ou $\beta_1 \neq b_1\}$.

3.4 Les p-values

3.4.1 Définition

Les tests présentés dans les sections précédentes peuvent tous s'exprimer sous la forme $\mathcal{F}$ suivante :

($\mathcal{F}$) *On rejette H_0 au niveau α si on observe que $S_n > q_{n,1-\alpha}$*

où S_n est la statistique de test et $q_{n,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi suivie par S_n lorsqu'on suppose H_0 vraie.

Par exemple, si on considère le test de Student présenté dans la définition 3.3, on peut le réexprimer sous la forme $\mathcal{F}$ en posant $S_n = |T|$ et $q_{n,1-\alpha} = \tau_{n-2;1-\frac{\alpha}{2}}$. On vérifie que $\tau_{n-2;1-\frac{\alpha}{2}}$ est effectivement le quantile d'ordre $1-\alpha$ de la loi de $|T|$ lorsque $H_0 : \beta_1 = 0$ est supposée vraie. En effet, le seuil $\tau_{n-2;1-\frac{\alpha}{2}}$ a été choisi de manière que $\mathrm{P}_{H_0}(|T| > \tau_{n-2;1-\frac{\alpha}{2}}) = \alpha$. Cela équivaut évidemment à $\mathrm{P}_{H_0}(|T| \leq \tau_{n-2;1-\frac{\alpha}{2}}) = 1-\alpha$, ce qui exprime que $\tau_{n-2;1-\frac{\alpha}{2}}$ est bien le quantile d'ordre $1-\alpha$ de la loi de $|T|$ lorsque H_0 est supposée vraie. On peut montrer de manière semblable que la formulation $\mathcal{F}$ peut également s'obtenir pour les tests de la définition 3.4 et de la propriété 3.5.

Lorsqu'on exprime un test sous la forme $\mathcal{F}$, on voit que pour décider si on rejette ou pas H_0 il suffit de comparer la valeur d'une statistique de test avec le quantile d'ordre $1-\alpha$ de la loi suivie par cette statistique lorsque H_0 est supposée vraie. Nous allons montrer que sur cette base et en utilisant la relation qui lie quantiles et fonction de répartition, la règle de décision peut s'exprimer d'une manière alternative. Notons F_{H_0} la fonction de répartition de S_n lorsque H_0 est supposée vraie. Pour les tests présentés dans les sections précédentes, S_n est une variable aléatoire continue

(elle est égale soit à $|T|$, soit à T selon le cas considéré). Sa fonction de répartition est bijective de $\mathbb{R}$ dans $[0,1]$, strictement croissante. Par conséquent on a

$$S_n > q_{n,1-\alpha} \iff F_{H_0}(S_n) > F_{H_0}(q_{n,1-\alpha})$$

Par définition du quantile $q_{n,1-\alpha}$, on a $F_{H_0}(q_{n,1-\alpha}) = 1-\alpha$. En définissant la variable aléatoire $P_n = 1 - F_{H_0}(S_n)$, on peut écrire

$$S_n > q_{n,1-\alpha} \iff P_n < \alpha \tag{3.18}$$ eq:pval1

On peut donc ré-exprimer les tests des sections précédentes sous la forme $\mathcal{F}'$ suivante :

($\mathcal{F}'$) *On rejette H_0 au niveau α si on observe $P_n < \alpha$*

où $P_n = 1 - F_{H_0}(S_n)$. La variable aléatoire P_n est appelée p-value associée à la statistique S_n.

On peut illustrer graphiquement l'équivalence des évènements $S_n > q_{n,1-\alpha}$ et $P_n < \alpha$. On utilise pour cela la séquence de graphiques de la figure 3.2, sur lesquels on représente les couples de valeurs $(\alpha, q_{n,1-\alpha})$ et (s_n, p_n) où s_n et p_n sont les réalisations de la statistique de test S_n et de la p-value P_n, respectivement. Le graphique du haut représente la courbe de la fonction $1-F_{H_0}$ pour laquelle on a $p_n = 1 - F_{H_0}(s_n)$ et $\alpha = 1 - F_{H_0}(q_{n,1-\alpha})$. Ce graphique illustre que l'inégalité $s_n > q_{n,1-\alpha}$ correspond au cas où la valeur observée de la statistique S_n est dans la zone rose sur l'axe horizontal. Dans ce cas, on a bien p_n dans la zone verte sur l'axe vertical, ce qui correspond à l'inégalité $p_n < \alpha$. La réciproque est également vraie, ce qui illustre l'équivalence entre ces inégalités. Le graphique du bas permet d'illustrer le même résultat à l'aide la fonction de densité de la loi de S_n lorsque H_0 est supposée vraie. Pour lancer l'animation, cliquez sur la figure 3.2.

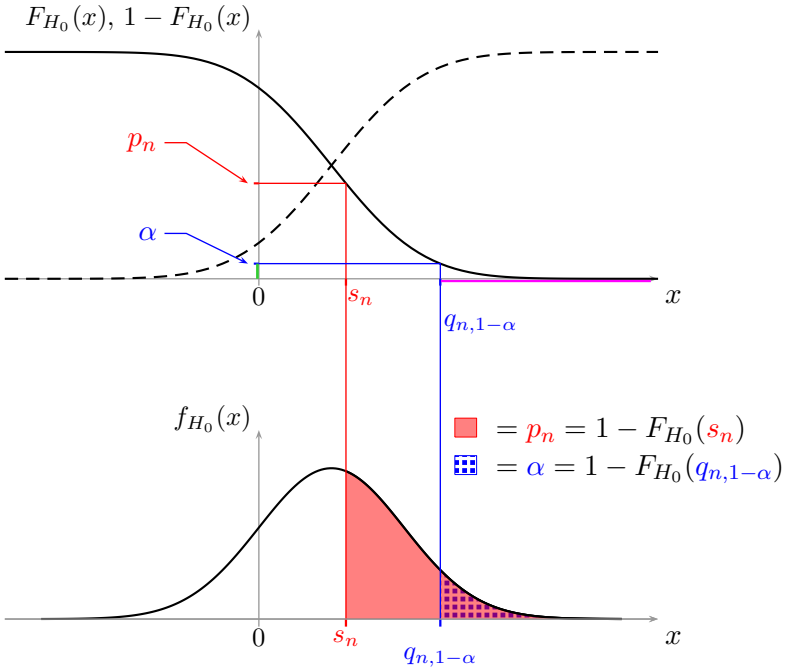
3.4.2 Interprétation

De manière générale, comme la statistique de test S_n et la p-value P_n sont en relation bijective, tout ce qui peut s'exprimer à l'aide de S_n peut s'exprimer de manière équivalente à l'aide de P_n et vice-versa. L'équivalence entre les formulations $\mathcal{F}$ et $\mathcal{F}'$ en est un exemple.

Cependant, certaines propriétés s'expriment plus aisément à l'aide d'une p-value, comprise entre 0 et 1, qu'au moyen de la statistique de test associée, dont les valeurs possibles ne sont pas nécessairement bornées.

Que H_0 soit supposée vraie ou pas et quelles que soient les observations dont on dispose, pour un test donné ayant la forme $\mathcal{F}$, H_0 sera toujours d'autant plus difficile à rejeter que le niveau choisi α est petit. En effet, indépendamment de supposer H_0 vraie ou pas, exprimé sous la forme $\mathcal{F}$ le test rejette H_0 lorsque la statistique de test dépasse le quantile d'ordre $1-\alpha$ d'une loi de probabilité. Dans le cas des tests des sections 3.2.2 et **??**, cette loi est continue, et sa fonction répartition F_{H_0} est continue, strictement croissante. Par conséquent le quantile d'ordre $1-\alpha$ de cette loi est $q_{n,1-\alpha} = F_{H_0}^{-1}(1-\alpha)$, où $F_{H_0}^{-1}$ est l'application réciproque de F_{H_0} et est donc continue strictement croissante. On en déduit que $q_{n,1-\alpha}$ est une fonction continue strictement décroissante de α. On a donc bien que plus le niveau auquel on choisit de faire le test est élevé, plus le quantile correspondant est petit, et plus il est probable d'observer $S_n > q_{n,1-\alpha}$ et donc de décider de rejeter H_0. Inversement, plus α est petit, plus $q_{n,1-\alpha}$ est grand, et moins il est probable de rejeter H_0. [11]

11. Cela est évidemment en parfaite cohérence avec la signification du niveau d'un test : c'est le risque maximal de



Cette séquence de graphiques illustre l'équivalence $s_n > q_{n,1-\alpha} \iff p_n < \alpha$. *Cliquez pour lancer l'animation.*

fig:FH0

FIGURE 3.2: Illustration de l'équivalence $S_n > q_{n,1-\alpha} \iff P_n < \alpha$

En utilisant cette propriété, on peut dégager une mesure du degré avec lequel les observations sont en faveur ou pas de H_0. On raisonne ici à observations $x_1, y_1, \ldots, x_n, y_n$ données, pour lesquelles la valeur observée de la statistique de test S_n est s_n. De la discussion du paragraphe précédent, on déduit que si on choisit pour le test un niveau suffisamment élevé, les observations dont on dispose conduiront au rejet de H_0 ; de même, si le niveau est suffisamment petit, nos observations nous conduiront à accepter H_0. Comme la fonction quantile de la loi de S_n lorsque H_0 est vraie est une fonction continue strictement croissante, il existe un unique α^* tel que $s_n = q_{n,1-\alpha^*}$, pour lequel on aura donc

$$s_n > q_{n,1-\alpha} \iff \alpha > \alpha* \tag{3.19}$$ eq:pval2

Cette équivalence exprime que les observations conduisent à rejeter H_0 au niveau α si et seulement si le niveau choisi est supérieur à α^*. Pour déterminer α^*, on note que par définition on a $q_{n,1-\alpha^*} = F_{H_0}^{-1}(1-\alpha^*)$. Donc

$$s_n = q_{n,1-\alpha^*} \iff s_n = F_{H_0}^{-1}(1-\alpha^*) \iff \alpha^* = 1 - F_{H_0}(s_n)$$

ce qui montre que α^* est la réalisation de la p-value P_n. [12]

On peut illustrer graphiquement l'inégalité (3.19). La courbe de la figure 3.3 est celle de la fonction $1 - F_{H_0}$ reliant quantile et p-value. En particulier, à s_n donnée correspond la valeur p_n (ou α^*). L'intervalle $]\alpha^*; 1]$ en rouge sur l'axe vertical représente l'ensemble des niveaux α supérieurs à p_n pour lesquels, sur la base des observations avec lesquelles on calcule les valeurs s_n et p_n, l'hypothèse H_0 est rejetée. Pour tout niveau α dans cet intervalle, le quantile $q_{n,1-\alpha}$ correspondant est dans l'intervalle $[0; s_n[$ (en rouge sur l'axe horizontal), et donc inférieur à s_n.

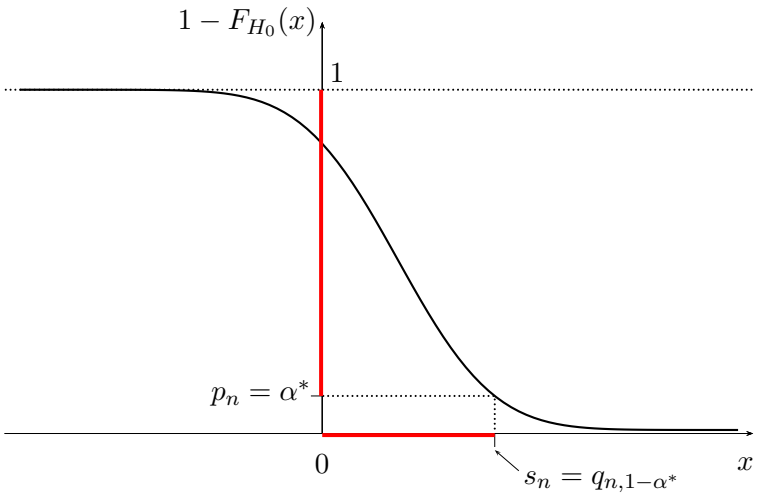


fig:pval

FIGURE 3.3: Illustration graphique de l'inégalité (3.19)

rejeter à tort H_0 qu'on est prêt à supporter. Plus ce niveau est petit, plus on souhaite se prémunir d'un rejet erroné de H_0. Pour cela, il faut rendre le rejet de H_0 plus difficile à obtenir, c'est à dire moins probable.

12. Ce dont on aurait pu se rendre compte en notant que les inégalités (3.18) et (3.19) sont les mêmes.

À observations données, la valeur de la p-value est donc le niveau du test jusqu'où on peut monter sans rejeter H_0, et au delà duquel cette hypothèse est rejetée. Si les observations sont telles que p_n est élevée (proche de 1), on peut choisir un niveau α élevé, tout en étant inférieur à p_n. Dans un tel cas, les observations ne conduisent pas à un rejet de H_0, bien que le choix d'un niveau élevé rende H_0 facile à rejeter (voir la discussion ci-dessus). C'est évidemment l'inverse qui se produit lorsque p_n est faible. Dans ce cas, α peut être petit (et donc H_0 difficile à rejeter *a priori*) mais supérieur à p_n, entraînant le rejet de H_0. Cette discussion fait apparaître que la valeur p_n de la p-value déduite des observations est une mesure du support de celles-ci pour l'hypothèse nulle. Plus p_n est grande, plus les observations « supportent » H_0.

Comme mentionné au début de cette section, tout ce qui s'énonce à partir de la p-value peut se formuler de manière équivalente en utilisant la statistique de test. Par exemple ce qui a été dit lorsque la p-value est grande peut se dire de manière équivalente sur la base de la valeur prise par S_n. Lorsque s_n est petite, on peut choisir des niveaux de tests élevés (et donc des quantiles relativement petits) sans pour autant que les observations nous amènent à rejeter H_0 à ces niveaux. Cependant, la p-value appartenant à l'intervalle $[0;1]$, il est plus facile *a priori* de savoir ce qu'est une grande p-value observée que de savoir si la valeur observée de la statistique de test est petite, cette dernière pouvant être un élément d'un ensemble non borné.

sec:rc

3.5 Régions de confiance

Après les problèmes d'estimation des paramètres et de tests d'hypothèses, on s'intéresse maintenant à la construction de régions de confiance. L'objectif est le suivant : on cherche, à partir des observations, à déterminer une région de l'espace des paramètres ayant de bonnes chances de contenir la valeur inconnue de ces paramètres. Les régions ainsi obtenues sont appelées *régions de confiance* (voir la section 10.3.3).

Dans le cas simple d'un paramètre unidimensionnel (dont la valeur est un élément de $\mathbb{R}$), cette région est donc un sous-ensemble de $\mathbb{R}$ à laquelle on donne très souvent la forme d'un intervalle. On parle dans ce cas d'*intervalle de confiance*. Celui-ci s'interprète comme une fourchette de nombres dans laquelle il est probable que se situe la valeur inconnue du paramètre. Dans le cas où on cherche à construire une région de confiance pour plusieurs paramètres à la fois, la région recherchée peut prendre diverses formes. On présentera en détail la démarche qui permet d'obtenir un intervalle de confiance pour le paramètre β_1. Pour construire un intervalle de confiance pour β_0, on réplique la même démarche, en l'adaptant au cas du paramètre β_0. On décrira ensuite une manière de former une région de confiance pour une combinaison linéaire de β_0 et de β_1. On terminera la section en proposant une région de confiance pour le couple (β_0, β_1).

L'aspect fondamental dans la construction d'une région de confiance est la « forte chance » qu'elle a de contenir la valeur inconnu du paramètre d'intérêt. Il faut donc être en mesure de calculer ces « chances ». Autrement dit, il faut pouvoir utiliser une loi de probabilité permettant de calculer la probabilité qu'une région contienne une valeur donnée du paramètre. Tout comme pour les tests, le calcul d'une telle probabilité sera rendu possible en introduisant la condition que $(Y_1, \ldots, Y_n)$ est gaussien : on se place donc dans le contexte du MRLSG.

La démarche présentée ici s'appuie entièrement sur les résultats présentés à la section 10.3.3.

Le théorème 10.2 montre qu'à toute famille de tests on peut associer une région de confiance et réciproquement. Autrement dit on peut chercher à obtenir une région de confiance en construisant une famille de tests. Par ailleurs, le corollaire 10.1 et la discussion qui suit montrent que le choix d'une région de confiance peut se faire en examinant la puissance de la famille de tests associée : pour un paramètre d'intérêt donné, on choisira la région de confiance de niveau $1-\alpha$ pour laquelle les tests de niveau α associés ont une puissance la plus élevée possible.

Grâce à ces résultats, la construction d'une région de confiance pour une paramètre donné peut se voir comme découlant directement de la construction et de l'étude des propriétés de tests d'hypothèses portant sur la valeur de ce paramètre.

sec:ic_b1

3.5.1 Intervalle de confiance pour β_1

La propriété 3.4 établit que pour toute valeur $b \in \mathbb{R}$, le test de Student est le test sans biais de niveau α le plus puissant pour tester $H_0 : \beta_1 = b$ contre $H_1 : \beta_1 \neq b$. Afin de faire le parallèle avec les résultats de la section 10.3.3, on notera φ_b ce test. On a donc

$$\varphi_{b,n} = 0 \iff \left| \frac{\hat{\beta}_1 - b}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} \right| \leq \tau_{n-2;1-\frac{\alpha}{2}} \tag{3.20}$$

eq:gerard

Le théorème 10.2 permet de conclure que la région de $\mathbb{R}$, notée $\mathscr{C}_{1,n}$, contenant toutes les valeurs b pour lesquelles l'inégalité dans (3.20) est satisfaite est une région de confiance de niveau $1-\alpha$ pour β_1. De plus, le corollaire 10.1 et le dernier paragraphe de la section 10.3.3 (page 256) impliquent que cette région de confiance est la plus exacte parmi toutes les régions de confiance sans biais de niveau $1-\alpha$ pour le paramètre β_1. Il reste à expliciter $\mathscr{C}_{1,n}$. En utilisant sa définition formelle, on a

$$\begin{aligned}
\mathscr{C}_{1,n} &= \left\{ b \in \mathbb{R} \mid \varphi_{b,n} = 0 \right\} = \left\{ b \in \mathbb{R} \mid \frac{|\hat{\beta}_1 - b|}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} \leq \tau_{n-2;1-\frac{\alpha}{2}} \right\} \\
&= \left\{ b \in \mathbb{R} \mid \hat{\beta}_1 - \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)} \leq b \leq \hat{\beta}_1 + \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)} \right\} \\
&= \left[\hat{\beta}_1 - \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)} \, ; \, \hat{\beta}_1 + \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)} \right]
\end{aligned} \tag{3.21}$$

eq:ic_beta1

On constate donc que la région de confiance de niveau $1-\alpha$ optimale (dans le sens où c'est la plus exacte parmi les régions sans biais) pour le paramètre β_1 est un intervalle de $\mathbb{R}$.

Cet intervalle est centré en $\hat{\beta}_1$, le meilleur estimateur ponctuel sans biais de β_1. Les extrémités de l'intervalle sont obtenues en ajoutant/retranchant $\tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}$ à/de $\hat{\beta}_1$. On constate donc qu'à un niveau de confiance $1-\alpha$ donné, l'estimateur ponctuel de β_1 contrôle la position de l'intervalle de confiance, et ce qui en contrôle la largeur est la précision (estimée) de cet estimateur, mesurée par $\hat{\mathrm{V}}(\hat{\beta}_1)$. On a mentionné dans la section 10.3.3 (voir page 253) que le diamètre d'une région de confiance (*i.e.*, la largeur dans le cas d'un intervalle) était une caractéristique à prendre en compte, puisqu'elle décrit le caractère informatif de cette région. On voit que dans le cas de l'intervalle $\mathscr{C}_{1,n}$ cette largeur est d'autant plus petite que l'estimation de β_1 est précise. Autrement dit, plus cette estimation est précise, plus le caractère informatif de $\mathscr{C}_{1,n}$ sera prononcé, ce qui constitue un résultat

souhaitable (et attendu) : une grande précision sur l'estimation de β_1 permet à l'intervalle $\mathscr{C}_{1,n}$ d'écarter plus de valeurs jugées non plausibles de ce paramètre (voir la discussion de la page 253). On appelle précision de l'intervalle $\mathscr{C}_{1,n}$ la largeur de cet intervalle, égale à $2 \times \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}$.

Bien que le théorème 10.2 permette de déduire que $\mathrm{P}_{\beta_1}(\beta_1 \in \mathscr{C}_{1,n}) \geq 1-\alpha$, on peut aussi obtenir cette inégalité à partir de la définition de $\mathscr{C}_{1,n}$ et du corollaire 3.3. En effet

$$\begin{aligned}
\mathrm{P}_{\beta_1}(\beta_1 \in \mathscr{C}_{1,n}) &= \mathrm{P}_{\beta_1}\Big(\hat{\beta}_1 - \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)} \leq \beta_1 \leq \hat{\beta}_1 + \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}\Big) \\
&= \mathrm{P}_{\beta_1}\Big(-\tau_{n-2;1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\mathrm{V}}(\hat{\beta}_1)}} \leq \tau_{n-2;1-\frac{\alpha}{2}}\Big) \\
&= 1-\alpha
\end{aligned}$$

où la première égalité provient de (3.21) et la dernière provient du corollaire 3.3 et de la définition de $\tau_{n-2;1-\frac{\alpha}{2}}$ comme le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-2$ degrés de liberté.

sec:ic_beta0

3.5.2 Intervalle de confiance pour β_0

Il suffit de calquer la démarche utilisée pour construire $\mathscr{C}_{1,n}$. La région de confiance sans biais la plus exacte au niveau $1-\alpha$ pour le paramètre β_0 est l'intervalle $\mathscr{C}_{0,n}$ défini par

$$\mathscr{C}_{0,n} = \Big[\hat{\beta}_0 - \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_0)}\ ;\ \hat{\beta}_0 + \tau_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\beta}_0)}\Big] \tag{3.22}$$

eq:ic_beta0

Les mêmes remarques que pour $\mathscr{C}_{1,n}$, transposées au cas du paramètre β_0, s'appliquent à $\mathscr{C}_{0,n}$.

sec:ic_beta0

3.5.3 Intervalle de confiance pour une combinaison linéaire de β_0 et de β_1

En prolongeant la démarche basée sur l'association entre région de confiance et famille de tests, il est facile de former une région de confiance de niveau $1-\alpha$ pour une combinaison linéaire $a_0\beta_0 + a_1\beta_1$ de β_0 et β_1 à partir du test de Student de la propriété 3.7. Le paramètre d'intérêt ici est $\gamma = a_0\beta_0 + a_1\beta_1$. Comme précédemment, on définit $\mathscr{C}_n$ la région de $\mathbb{R}$ formée de toutes les valeurs r pour lesquelles le test de Student de niveau α de $H_0(r) : a_0\beta_0 + a_1\beta_1 = r$ contre $H_1(r) : a_0\beta_0 + a_1\beta_1 \neq r$ conduit à décider $H_0(r)$. Formellement, en utilisant la définition de ce test de Student (voir la propriété 3.7), on a

$$\mathscr{C}_n = \Big\{ r \in \mathbb{R} \mid \frac{|a_0\hat{\beta}_0 + a_1\hat{\beta}_1 - r|}{\sqrt{a_0^2\hat{\mathrm{V}}(\hat{\beta}_0) + a_1^2\hat{\mathrm{V}}(\hat{\beta}_1) + 2a_0a_1\widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1)}} \leq \tau_{n_2;1-\frac{\alpha}{2}} \Big\}$$

Par le même raisonnement que dans la section 3.5.1, on déduit que $\mathscr{C}_n$ est la région de confiance de niveau $1-\alpha$ pour $\gamma = a_0\beta_0 + a_1\beta_1$ la plus exacte parmi toutes les régions de confiance sans biais de niveau $1-\alpha$ pour γ.

On montre que $\mathscr{C}_n$ est un intervalle de $\mathbb{R}$. En effet, si on note $\hat{\gamma} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1$, on peut écrire

$$\hat{\mathrm{V}}(\hat{\gamma}) = \hat{\mathrm{V}}(a_0\hat{\beta}_0 + a_1\hat{\beta}_1) = a_0^2\hat{\mathrm{V}}(\hat{\beta}_0) + a_1^2\hat{\mathrm{V}}(\hat{\beta}_1) + 2a_0a_1\widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1)$$

Donc

$$\begin{aligned}\mathscr{C}_n &= \Big\{r \in \mathbb{R} \mid -\sqrt{\hat{\mathrm{V}}(\hat{\gamma})}\,\tau_{n_2;1-\frac{\alpha}{2}} \leq \hat{\gamma} - r \leq \sqrt{\hat{\mathrm{V}}(\hat{\gamma})}\,\tau_{n_2;1-\frac{\alpha}{2}}\Big\} \\ &= \Big[\hat{\gamma} - \sqrt{\hat{\mathrm{V}}(\hat{\gamma})}\,\tau_{n_2;1-\frac{\alpha}{2}}\ ;\ \hat{\gamma} + \sqrt{\hat{\mathrm{V}}(\hat{\gamma})}\,\tau_{n_2;1-\frac{\alpha}{2}}\Big]\end{aligned}$$

En reprenant le cas particulier du test sur $\mathrm{E}(Y_i)$ (ce qui revient à choisir $a_0 = 1$ et $a_1 = X_i$, voir section 3.3.1.2), on peut directement obtenir un intervalle de confiance de niveau $1-\alpha$ pour la valeur attendue de Y_i. Dans un tel cas, on a $\gamma = \beta_0 + \beta_1 X_i$ et utilise la forme de $\mathscr{C}_n$ donnée ci-dessus, avec $\hat{\gamma} = \hat{\beta}_0 + \hat{\beta}_1 X_i$. Plus explicitement, on a

$$\mathscr{C}_n = \Big[\hat{\beta}_0 + \hat{\beta}_1 X_i - \sqrt{\hat{\mathrm{V}}(\hat{\beta}_0 + \hat{\beta}_1 X_i)}\,\tau_{n_2;1-\frac{\alpha}{2}}\ ;\hat{\beta}_0 + \hat{\beta}_1 X_i + \sqrt{\hat{\mathrm{V}}(\hat{\beta}_0 + \hat{\beta}_1 X_i)}\,\tau_{n_2;1-\frac{\alpha}{2}}\Big]$$

avec

$$\hat{\mathrm{V}}(\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\sigma}^2\Big(\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^n X_j^2 - n\overline{X}^2}\Big)$$

sec:rc_b0b1

3.5.4 Région de confiance pour (β_0, β_1)

Finalement, on utilise toujours la même approche pour construire une région de confiance pour le couple (β_0, β_1). Il s'agit ici de trouver une région de $\mathbb{R}^2$ qui contient simultanément les valeurs (inconnues) des deux paramètres.

Dans la section 3.3.2.3, pour $b = (b_0, b_1)$ donné, on a présenté le test de Fisher de niveau α pour tester $H_0(b) : (\beta_0, \beta_1) = (b_0, b_1)$ contre $H_1(b) : (\beta_0, \beta_1) \neq (b_0, b_1)$. Ce test peut être construit pour chaque valeur possible du couple (b_0, b_1). Le théorème 10.2 permet de conclure que la région $\mathscr{C}_n$ de $\mathbb{R}^2$ contenant tous les couples $b = (b_0, b_1)$ pour lesquels on décide $H_0(b)$ au niveau α au moyen du test de Fisher est une région de confiance de niveau $1-\alpha$ pour le couple $\beta = (\beta_0, \beta_1)$. La forme du test de Fisher est donnée dans la propriété 3.10, et on a

$$\mathscr{C}_n = \{b \in \mathbb{R}^2 \mid \frac{1}{2}(\hat{\beta} - b)^\top (\hat{\sigma}^2 v)^{-1}(\hat{\beta} - b) \leq F_{(2,n-2);1-\alpha}\}$$

La frontière de cette région est décrite par l'équation d'une ellipse de centre $\hat{\beta}$.

Chapitre 5

ch:mco

Le modèle de régression linéaire standard : définition et estimation

Dans le chapitre précédent, on a considéré un problème dans lequel on voulait construire un modèle statistique simple permettant de représenter une relation deux variables et de l'étudier. Dans ce chapitre (et les suivants), on généralise l'approche, ainsi que les résultats obtenus.

sec:mrlst_def

5.1 Définition

On se fixe ici les mêmes objectifs que dans le chapitre 1, mais en cherchant à généraliser les relations étudiées dans les chapitres précédents. Plus précisément, on cherche un modèle statistique simple permettant de représenter et d'étudier au moyen des méthodes d'inférence statistique usuelles une relation dans laquelle p variables exogènes expliquent une variable endogène. La variable endogène (ou dépendante) est notée Y comme auparavant et les variables exogènes (explicatives) sont numérotées et notées $X_1, \ldots, X_p$.

L'approche exposée à la section 1.2 reste tout à fait adaptée à ce nouveau contexte. Notamment, le modèle proposera une décomposition additive de Y en deux parties :

- une partie qui capture de manière simple (linéairement) l'influence des variables $X_1, \ldots, X_p$ sur Y ;
- une partie qui capture l'effet que des facteurs non identifiés ou non mesurés, autres que ceux mesurés par $X_1, \ldots, X_p$, peuvent éventuellement avoir sur Y.

Pour compléter la représentation recherchée dans le contexte des objectifs fixés au départ, on introduira

- des conditions qui permettent de traduire la distinction entre variables exogènes et variable endogène ;
- des conditions permettant de capturer la prédominance de l'effet des variables exogènes explicitement introduites sur celui que peuvent avoir les autres facteurs dans la détermination du niveau de la variable endogène.

En ce qui concerne les notations, on désignera par X_{ik} la variable aléatoire exprimant la mesure de la variable X_k pour l'individu i de l'échantillon, tandis que comme auparavant, Y_i est la variable

aléatoire qui exprime la mesure de la variable endogène pour ce même individu, $k = 1, \ldots, p$ et $i = 1, \ldots, n$. Les observations de ces variables seront notées x_{ik} et y_i, respectivement.

def:mrlsp **Définition 5.1** *Pour chaque individu i d'un échantillon de taille n, on dispose d'un $(p+1)$-uplet de variables aléatoires $(X_{i1}, X_{i2}, \ldots, X_{ip}, Y)$ $i = 1, \ldots, n$. Le* modèle de régression linéaire standard *(MRLS) à p variables de Y sur $(X_1, \ldots, X_p)$ est un modèle statistique dans lequel les conditions suivantes sont satisfaites*

C$_p$1. $\mathrm{P}(X_{ik} = x_{ik}, k = 1, \ldots, p, i = 1, \ldots, n) = 1$

C$_p$2. *Il existe $p+1$ réels $\beta_0, \beta_1, \ldots, \beta_p$ tels que*

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \quad \forall i = 1, \ldots, n$$

C$_p$3. *Il existe un réel strictement positif σ tel que*

$$\mathrm{cov}(Y_i, Y_j) = \begin{cases} 0 \text{ si } i \neq j \\ \sigma^2 \text{ si } i = j \end{cases}$$

pour toute paire (i, j) d'éléments de $\{1, \ldots, n\}$.

Les interprétations des trois conditions de la définition précédente sont d'une nature identique à celles qui ont été faites dans la section 1.3.2 des conditions C1, C2 et C3 (définition 1.1). Il suffit simplement de tenir compte du fait que dans le modèle introduit ci-dessus, on utilise p variables pour expliquer la variable endogène : connaissant la valeur des variables exogènes, la valeur attendue de la variable endogène s'écrit comme une fonction affine de la valeur des variables exogènes. Cette fonction est caractérisée par les $p+1$ paramètres $\beta_0, \beta_1, \ldots, \beta_p$.

La remarque faite sur la *vraie loi* et les *vraies valeurs* des paramètres (voir la remarque 1.1) s'applique également. Les paramètres admettent des vraies valeurs qu'on notera $\overline{\beta}_0, \overline{\beta}_1, \ldots, \overline{\beta}_p$. Celles-ci sont inconnues et un objectif sera d'estimer ces valeurs à partir des observations des variables du modèle.

La définition donnée ci-dessus désigne un modèle particulier dans l'ensemble des modèles de régression linéaire. Ces derniers sont des modèles caractérisés par la condition C$_p$2. Le qualificatif « standard » utilisé dans l'appellation du modèle de la définition 5.1 traduit le fait que celui-ci constitue un point de référence pour l'ensemble des modèles de régression linéaire, caractérisé par l'ajout des conditions simplificatrices C$_p$1 et C$_p$3. Il est notamment courant de comparer les propriétés des modèles régressions linéaire plus généraux avec celles du modèle standard défini ci-dessus.

Comme dans le cas du modèle simple présenté au chapitre 1, le modèle de régression linéaire standard à p variables admet une définition équivalente, faisant apparaître explicitement la décomposition recherchée de Y, rappelée avant la définition 5.1.

pro:mrlsp_eps **Propriété 5.1** *Pour chaque individu i d'un échantillon de taille n, on dispose d'un $(p+1)$-uplet de variables aléatoires $(X_{i1}, X_{i2}, \ldots, X_{ip}, Y)$ $i = 1, \ldots, n$. Les conditions* C$_p$1, C$_p$2 *et* C$_p$3 *sont satisfaites si et seulement si les conditions suivantes le sont aussi*

C$'_p$1. $\mathrm{P}(X_{ik} = x_{ik}, k = 1, \ldots, p, i = 1, \ldots, n) = 1$

C$'_p$2. *Il existe* $p+1$ *réels* $\beta_0, \beta_1, \ldots, \beta_p$ *tels que*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad \forall i = 1, \ldots, n$$

où $\varepsilon_i \equiv Y_i - \mathrm{E}(Y_i)$, $i = 1, \ldots, n$

C$'_p$3. *Il existe un réel strictement positif* σ *tel que*

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 \text{ si } i \neq j \\ \sigma^2 \text{ si } i = j \end{cases}$$

pour toute paire (i, j) *d'éléments de* $\{1, \ldots, n\}$.

On note donc que la condition C$'_p$2 permet de faire apparaître dans la définition du modèle la décomposition recherchée de la variable endogène : le niveau de cette variable s'exprime comme la somme d'une partie qui ne dépend (linéairement) que du niveau des variables exogènes explicitement introduites dans le modèle, et d'une partie qui dépend d'autres facteurs non définis, non-mesurés et/ou non-observables.

Dans la suite, qu'il s'agisse de l'interprétation ou de l'étude du modèle, on pourra atteindre les résultats et objectifs recherchés en proposant une reformulation du MRLS à p variables dans lequel les éléments sont considérés comme des vecteurs dont les coordonnées peuvent être aléatoires, et en s'appuyant sur des notions d'algèbre linéaire. Pour cela on introduit les éléments suivants :

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad X_{i\bullet} = \begin{pmatrix} X_{i0} \\ X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix} \qquad X_{\bullet k} = \begin{pmatrix} X_{1k} \\ X_{2k} \\ \vdots \\ X_{nk} \end{pmatrix}$$

pour $i = 1, \ldots, n$ et $k = 0, \ldots, p$, et avec la convention que $X_{i0} = 1$, $i = 1, \ldots, n$. Les éléments $\boldsymbol{Y}$, $X_{\bullet k}$ sont considérés comme des vecteurs de $\mathbb{R}^n$ dont les coordonnées sont aléatoires. Il en est de même pour $X_{i\bullet}$, à la différence qu'il est un vecteur de $\mathbb{R}^{p+1}$. On introduit de plus la matrice X de taille $(n, (p+1))$, dont les entrées sont aléatoires et dont les colonnes sont les vecteurs aléatoires $X_{\bullet k}$, $k = 0, \ldots, p$, de $\mathbb{R}^n$:

$$X = \begin{pmatrix} X_{10} & X_{11} & \cdots & X_{1p} \\ X_{20} & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n0} & X_{n1} & \cdots & X_{np} \end{pmatrix} = \begin{pmatrix} X_{\bullet 0} & X_{\bullet 1} & \cdots & X_{\bullet p} \end{pmatrix} = \begin{pmatrix} X_{1\bullet}^\top \\ X_{2\bullet}^\top \\ \vdots \\ X_{n\bullet}^\top \end{pmatrix}$$

On introduit également le vecteur non aléatoire β de $\mathbb{R}^{p+1}$ dont les coordonnées sont les paramètres inconnus de la relation exprimée par la condition C$_p$2 ou C$'_p$2 :

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Ces éléments permettent de reformuler les conditions C$_p$2 et C$_p$3, ou C$'_p$2 et C$'_p$3, de la manière suivante.

pro:mrslmat **Propriété 5.2**

1. *Les conditions* C_p1 *à* C_p3 *de la définition 5.1 sont équivalentes à la condition*

$$\exists \beta \in \mathbb{R}^{p+1}, \exists \sigma \in]0, \infty[,\ \mathrm{E}(\boldsymbol{Y}) = X\beta \text{ et } \mathrm{V}(\boldsymbol{Y}) = \sigma^2 I_n$$

où, comme mentionné à la section 3.1, $\mathrm{E}(\boldsymbol{Y})$ *est le vecteur de* $\mathbb{R}^n$ *dont la* i^{e} *coordonnée est* $\mathrm{E}(Y_i)$*,* $i = 1, \ldots, n$ *et* $\mathrm{V}(\boldsymbol{Y})$ *est la matrice de dimensions* (n, n) *dont la* $(i, j)^{\mathrm{e}}$ *entrée est* $\mathrm{cov}(Y_i, Y_j)$*.*

2. *Les conditions* C'_p1 *à* C'_p3 *de la propriété 5.1 sont équivalentes à la condition*

$$\exists \beta \in \mathbb{R}^{p+1}, \exists \sigma \in]0, \infty[,\ \boldsymbol{Y} = X\beta + \varepsilon \text{ et } \mathrm{V}(\varepsilon) = \sigma^2 I_n$$

où le vecteur aléatoire ε *de* $\mathbb{R}^n$ *est défini par*

$$\varepsilon = \boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y}) = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

rem:EY_LX **Remarque 5.1** En utilisant la propriété 5.2, on constate que la condition C_p2 impose au vecteur $\mathrm{E}(\boldsymbol{Y})$ de $\mathbb{R}^n$ d'être une combinaison linéaire des $p+1$ vecteurs $X_{\cdot 0}, X_{\cdot 1}, \ldots, X_{\cdot p}$ de $\mathbb{R}^n$: on doit avoir $\mathrm{E}(\boldsymbol{Y}) = \beta_0 X_{\cdot 0} + \cdots + \beta_p X_{\cdot p}$ pour des réels $\beta_0, \ldots, \beta_p$. Ces vecteurs, qui composent les colonnes de la matrice X, engendrent un sous-espace de $\mathbb{R}^n$, noté $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ et la condition C_p2 s'écrit $\mathrm{E}(\boldsymbol{Y}) \in L(X_{\cdot 0}, \ldots, X_{\cdot p})$. D'après la condition C'_p2, on peut donc décomposer le vecteur $\boldsymbol{Y}$ de $\mathbb{R}^n$ en tant que somme d'un vecteur de $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ et d'un vecteur ε de $\mathbb{R}^n$. On peut représenter graphiquement cette décomposition à l'aide de la séquence de graphiques de la figure 5.1.[1] Cette représentation fait notamment apparaître l'espace $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ comme sous-espace de $\mathbb{R}^n$, engendré par les vecteurs $X_{\cdot 0}, \ldots, X_{\cdot p}$ (et donc contenant ces vecteurs). Les graphiques montrent également que $\mathrm{E}(\boldsymbol{Y})$ appartient à ce sous-espace ainsi que l'impose la condition C_p2, mais qu'en général $\boldsymbol{Y}$ n'en fait pas partie. Finalement, la décomposition de $\boldsymbol{Y}$ en la somme de $\mathrm{E}(\boldsymbol{Y}) \in L(X_{\cdot 0}, \ldots, X_{\cdot p})$ et de $\varepsilon \in \mathbb{R}^n$ est illustrée. R

rem:decomp_EY **Remarque 5.2** Si les $p+1$ vecteurs $X_{\cdot 0}, \ldots, X_{\cdot p}$ sont linéairement indépendants, alors ils forment une base du sous-espace $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ de $\mathbb{R}^n$ de dimension $p+1$. Dans ce cas, la décomposition de $\mathrm{E}(\boldsymbol{Y})$ sur les vecteurs $X_{\cdot 0}, \ldots, X_{\cdot p}$ est unique. Les paramètres $\beta_0, \ldots, \beta_p$ de cette décomposition sont les coordonnées du vecteur $\mathrm{E}(\boldsymbol{Y})$ dans la base $X_{\cdot 0}, \ldots, X_{\cdot p}$. L'unicité résultant de l'indépendance de ces vecteurs a une conséquence importante en termes d'interprétation du modèle. En effet, ce modèle est destiné à représenter et mesurer la relation entre la variable Y et les variables $X_1, \ldots, X_p$, et pose que cette relation est linéaire, ainsi que le traduit la condition C'_p2 (ou C_p2). En particulier, la mesure de la réaction de la variable Y à une variation de la variable X_k est β_k. Si les vecteurs $X_{\cdot 0}, \ldots, X_{\cdot p}$ n'étaient pas linéairement indépendants, alors la décomposition de $\mathrm{E}(\boldsymbol{Y})$ sur ces vecteurs ne serait pas unique : on pourrait trouver des réels $\gamma_0, \ldots, \gamma_p$, avec $\gamma_l \neq \beta_l$ pour au

1. Cette séquence de graphiques est animée. Pour visualiser l'animation, reportez-vous aux indications données à la fin de l'introduction de ce document. Si vous ne disposez pas d'un lecteur de fichiers PDF permettant d'animer la séquence de graphiques, l'animation est disponible à l'url http://gremars.univ-lille3.fr/~torres/enseigne/ectrie/Cp2/.

moins un $l \in \{0, \ldots, p\}$, tels que $\mathrm{E}(\boldsymbol{Y}) = \gamma_0 X_{\cdot 0} + \cdots + \gamma_p X_{\cdot p}$. [2] On voit alors qu'il y a une ambiguïté lorsqu'on cherche à représenter, au moyen d'une condition telle que C_p2, la relation entre Y et une variable exogène X_k, puisque le lien entre ces deux variables peut être caractérisé soit par β_k soit par γ_k. Il y a dans ce cas une indétermination dans cette représentation : les paramètres d'intérêt qui relient la variable endogène aux variables exogènes ne sont pas caractérisés de manière unique par la condition C_p2. Lorsque cela se produit, on dit que ces paramètres ne sont pas identifiés. On verra quelles sont les conséquences de cela en terme d'interprétation et d'estimation de ces paramètres. ℝ

rem:rangX **Remarque 5.3** Notons que les $p+1$ vecteurs $X_{\cdot 0}, \ldots, X_{\cdot p}$ sont les colonnes de la matrice X. Par conséquent, ils sont linéairement indépendants si et seulement si la matrice X est de rang $p+1$. On voit alors qu'une condition nécessaire pour que le rang de X soit égal à $p+1$ est qu'on dispose d'un nombre d'observations n supérieur au nombre $p+1$ de paramètres qui expriment la relation entre les variables exogènes et la variable endogène. Rappelons également l'équivalence suivante qui sera par la suite : $\mathrm{rang}(X) = p+1 \iff X^\top X$ est inversible. [3] ℝ

rem:LZ **Remarque 5.4** Il est clair que si la condition C_p2 spécifie que $\mathrm{E}(\boldsymbol{Y})$ peut s'écrire $\mathrm{E}(\boldsymbol{Y}) = X\beta$, alors pour toute matrice Q de taille $(p+1, p+1)$ inversible, on peut écrire $\mathrm{E}(\boldsymbol{Y}) = XQ^{-1}Q\beta$. On peut alors définir $\delta = Q\beta$ et $Z = XQ^{-1}$. Dans ce cas, la condition C_p2 impose qu'il existe $\delta \in \mathbb{R}^{p+1}$ tel que $\mathrm{E}(\boldsymbol{Y}) = Z\delta$. Ré-écrite de cette manière, la condition C_p2 exprime que le vecteur $\mathrm{E}(\boldsymbol{Y})$ de $\mathbb{R}^n$ est un élément du sous espace $L(Z_{\cdot 0}, \ldots, Z_{\cdot p})$ de $\mathbb{R}^n$ engendré par les colonnes de la matrice Z.

On notera que $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ et $L(Z_{\cdot 0}, \ldots, Z_{\cdot p})$ sont les mêmes sous-espaces de $\mathbb{R}^n$. En effet, l'équivalence $Z = XQ^{-1} \iff X = ZQ$ montre que toute combinaison linéaire de $Z_{\cdot 0}, \ldots, Z_{\cdot p}$ est également une combinaison linéaire de $X_{\cdot 0}, \ldots, X_{\cdot p}$ et *vice versa*. Par conséquent, les conditions $\mathrm{E}(\boldsymbol{Y}) \in L(X_{\cdot 0}, \ldots, X_{\cdot p})$ et $\mathrm{E}(\boldsymbol{Y}) \in L(Z_{\cdot 0}, \ldots, Z_{\cdot p})$ sont parfaitement équivalentes. Autrement dit, en choisissant d'écrire la condition C_p2 sous la forme $\mathrm{E}(\boldsymbol{Y}) = Z\delta$ où $Z = XQ^{-1}$ pour une matrice Q connue, on ne change pas le modèle. Cela revient à choisir $Z_0, \ldots, Z_p$ comme variables explicatives au lieu de $X_0, \ldots, X_p$, les unes étant des transformations bijectives des autres, ainsi que le traduit l'équivalence $Z = XQ^{-1} \iff X = ZQ$. Ce changement de variables explicatives induit un changement de paramètres (ou une reparamétrisation).

Lorsque $X_{\cdot 0}, \ldots, X_{\cdot p}$ sont linéairement indépendants, ils forment une base de $L(X_{\cdot 0}, \ldots, X_{\cdot p})$. Puisque Q est inversible, les vecteurs $Z_{\cdot 0}, \ldots, Z_{\cdot p}$ sont également linéairement indépendants. Ils forment un autre base de $L(X_{\cdot 0}, \ldots, X_{\cdot p})$. Exprimer $\mathrm{E}(\boldsymbol{Y}) = Z\delta$ revient simplement à exprimer le vecteur $\mathrm{E}(\boldsymbol{Y})$ dans cette nouvelle base, et les éléments de δ sont les coordonnées de $\mathrm{E}(\boldsymbol{Y})$ dans la base $Z_{\cdot 0}, \ldots, Z_{\cdot p}$. ℝ

page:rang 2. Dans le cas où $X_{\cdot 0}, \ldots, X_{\cdot p}$ ne sont pas linéairement indépendants, l'un de ces vecteurs s'écrit comme une combinaison linéaire des autres. Quitte à renuméroter les variables exogènes, on peut supposer qu'il existe des nombres $c_0, \ldots, c_{p-1}$ non tous nuls tels que $X_{ip} = \sum_{k=0}^{p-1} c_k X_{ik}$ pour tout individu i. Dans ce cas, en utilisant la condition C_p2, on voit que $\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{ip-1} + \beta_p \sum_{k=0}^{p-1} c_k X_{ik} = \gamma_0 + \gamma_1 X_{i1} + \cdots + \gamma_p X_{ip}$ où $\gamma_k = \beta_k + c_k \beta_p$, $k = 0, \ldots, p-1$ et $\gamma_p = 0$.

3. Vous devriez vous assurer que ce résultat vous est connu et que vous savez le démontrer…

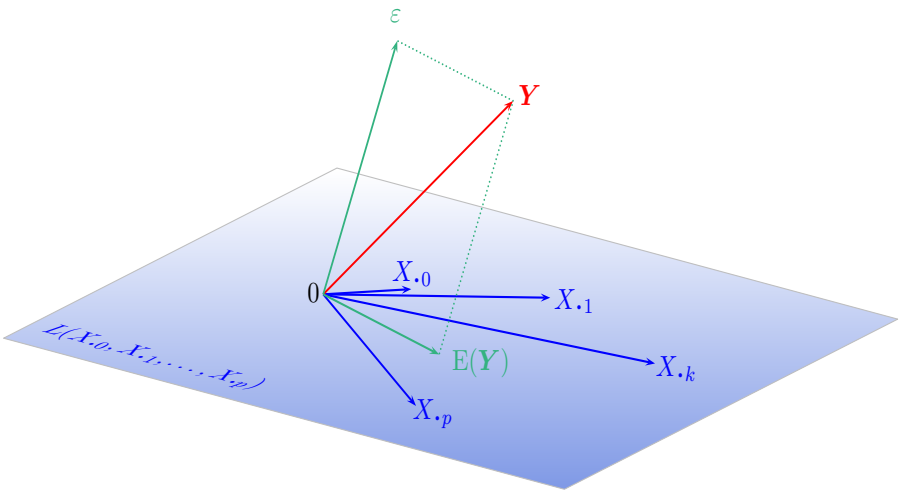


fig:decomp_EY

FIGURE 5.1: Représentation de la décomposition $\boldsymbol{Y} = \mathrm{E}(\boldsymbol{Y}) + \varepsilon$

sec:mrls_interpp

5.2 Interprétation des paramètres du modèle

Pour interpréter le rôle des paramètres dans un MRLS, on peut avoir recours au procédé suivant. On s'intéresse par exemple à β_1. Considérons deux individus identiques pour lesquels les caractéristiques mesurées par les variables exogènes sont identiques, sauf pour ce qui concerne X_1. Plus précisément, on suppose que pour les individus i et j distincts, on observe $X_{ik} = X_{jk} = x_k$ pour $k = 2, \ldots, p$ et que pour $k = 1$, on a $x_{i1} = x_{j1} + 1$. On peut déduire l'impact de cette différence sur la valeur attendue de la variable endogène de chacun de ces individus. On aura donc

$$\begin{aligned} \mathrm{E}(Y_i) &= \beta_0 + \beta_1(x_{j1} + 1) + \beta_2 x_2 + \cdots + \beta_p x_p \\ \mathrm{E}(Y_j) &= \beta_0 + \beta_1 x_{j1} + \beta_2 x_2 + \cdots + \beta_p x_p \end{aligned}$$

La différence attendue sur le niveau de la variable endogène pour les individus i et j est donc

$$\mathrm{E}(Y_i - Y_j) = \mathrm{E}(Y_i) - \mathrm{E}(Y_j) = \beta_1$$

On interprète donc le paramètre β_1 attaché à la variable explicative X_1 comme la différence attendue sur le niveau de la variable endogène entre deux individus identiques en tout point, excepté que le premier a un niveau variable X_1 d'une unité plus élevé que le second.

On obtient le même type d'interprétation en étudiant l'effet d'une variation de X_1 sur le niveau attendu de Y, toutes les autres variables étant maintenues constantes et fixées à des valeurs données $x_2, \ldots, x_p$. Pour cela, on considère $\mathrm{E}(Y_i)$ comme une fonction des variables $X_{i1}, \ldots, X_{ip}$: $\mathrm{E}(Y_i) = f(X_{i1}, \ldots, X_{ip}) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$. L'effet d'une variation de X_{i1} sur la valeur attendue de Y_i s'étudie alors au moyen de la dérivée partielle de f par rapport à X_{i1}, évaluée en $x_1, x_2, \ldots, x_p$. On obtient facilement :[4]

$$\left.\frac{\partial \mathrm{E}(Y_i)}{\partial X_{i1}}\right|_{(x_1, x_2, \ldots, x_p)} = \frac{\partial f}{\partial X_{i1}}(x_1, x_2, \ldots, x_p) = \beta_1$$

Le paramètre β_1 mesure donc l'effet des variations de X_{i1} sur la valeur attendue de Y_i. On peut obtenir un résultat plus précis sur cet effet, puisque si on fixe le niveau X_{ik} à x_k pour $k = 2, \ldots, p$, alors $\mathrm{E}(Y_i)$ dépend de linéairement X_{i1}. Par conséquent, β_1 est également la variation relative de $\mathrm{E}(Y_i)$ consécutive à un accroissement de Δ unités de X_{i1}, les niveaux des variables autres que X_{i1} étant maintenue inchangés. Formellement, en considérant comme auparavant $\mathrm{E}(Y_i) = f(X_{i1}, \ldots, X_{ip})$, on peut écrire :

$$\frac{f(x_{i1} + \Delta, x_2, \ldots, x_p) - f(x_{i1}, x_2, \ldots, x_p)}{\Delta} = \beta_1$$

Lorsqu'on choisit $\Delta = 1$, on peut interpréter β_1 comme la variation attendue de Y_i engendrée par une augmentation d'une unité du niveau de la variable X_{i1}, le niveau des autres variables restant inchangé (ou encore, *toutes choses égales par ailleurs*).

Notons que le signe de β_1 est important, puisque s'il est positif, un accroissement provoquera, toutes choses égales par ailleurs, une augmentation de la valeur attendue de Y_i, tandis que si β_1 est négatif, c'est une diminution qui sera attendue.

4. Cette dérivée ne dépend évidemment pas de l'endroit où elle est évaluée, puisque f est linéaire en chacun de ses arguments.

rem:tcpa **Remarque 5.5** Lors de l'interprétation des paramètres d'un MRLS ou lors d'un exercice théorique qui consiste à examiner l'effet d'une augmentation de l'une des variables exogènes sur le niveau attendu de la variable endogène, il est très important de raisonner « toutes choses égales par ailleurs ». Les exemples suivants montrent pourquoi.

1. Considérons un MRLS dans lequel les individus sont des maisons, la variable endogène est le prix de vente de la maison et les variables explicatives sont la surface, le nombre de pièces et l'âge de la maison. Désignons par β_1 le paramètre de la variable nombre de pièces. Si on s'intéresse au signe possible de β_1, on pourrait de manière un peu hâtive conclure qu'il est positif, puisqu'en général, une maison avec beaucoup de pièces sera vendue à un prix plus élevé qu'une maison ayant peu de pièces (donc β_1 positif). Cependant, dans le contexte du MRLS formulé dans cet exemple, ce raisonnement n'est pas valable. En effet, comme décrit ci-dessus, β_1 s'interprète comme la différence attendue entre le prix de vente de deux maisons identiques en tout point, excepté que l'une possède une pièce de plus que la précédente. Cela implique donc que dans cette comparaison, les deux maisons ont la même surface. Par conséquent, si l'une a plus de pièces que l'autre, la taille moyenne de ses pièces doit être plus petite, et peut donc avoir une valeur de vente moindre. On voit donc qu'en raisonnant toutes choses égales par ailleurs, comme il se doit, il est tout à fait plausible de penser que la hausse du nombre de pièces peut se traduire par un prix de vente attendu plus faible (donc β_1 négatif).

   Le raisonnement erroné qui conduisait à estimer que β_1 devrait être positif comportait une étape sous-jacente qui consistait à affirmer que si une maison a plus de pièces qu'une autre, elle est en général de plus grande superficie et a donc plus de valeur. On voit dans ce cas que le raisonnement envisage non seulement une augmentation du nombre de pièces, mais également une augmentation de la superficie. Ce type de raisonnement dans lequel on autorise éventuellement une variation du niveau des variables autres que celles dont on étudie l'effet sur la variable endogène n'est pas correct, dans le contexte de l'interprétation des paramètres d'un MRLS.

2. Considérons à présent un MRLS dans lequel les individus sont des villes, la variable endogène est le nombre moyen de passagers/heure dans les bus de la ville, et les variables exogènes sont le prix du ticket de bus, le prix du litre d'essence, le revenu moyen *per capita*, la superficie de la ville, le nombre d'habitants. Dans cet exemple, on désigne par β_1 le paramètre attaché à la variable prix du ticket de bus. Lors de l'interprétation de ce paramètre, on peut être tenté de dire que l'augmentation du prix du ticket de bus n'a pas le même effet sur le nombre de passagers/heure dans les petites villes que dans les grandes villes. Cependant, le MRLS décrit ici ne permet pas de mesurer le degré d'exactitude de cette affirmation. Raisonner « toutes choses égales par ailleurs » implique que l'effet d'une variation du prix du ticket de bus sur la variable endogène doit s'étudier pour des villes ayant des niveaux identiques des autres variables exogènes, et en particulier pour des villes ayant la même superficie et le même nombre d'habitants.

3. On mentionnera le caractère parfois délicat d'un raisonnement « toutes choses égales par ailleurs ». Pour illustrer cela, considérons un MRLS dans lequel les individus sont des humains, où la variable dépendante est le montant des dépenses de santé, les variables explicatives sont la zone d'habitat (rurale/urbaine) le sexe, l'âge, le carré de l'âge. Cette dernière variable est

introduite pour éventuellement capturer l'existence potentielle de liaisons non linéaires entre l'âge et les dépenses de santé. Il est clair que dans ce cas, il est difficile d'évaluer l'effet de l'augmentation de la variable âge sur la variable dépenses de santé, en voulant garder constant le niveau de la variable carré de l'âge. Dans un tel cas, plutôt que de vouloir interpréter le paramètre attaché à la variable âge, il est recommandé d'étudier, toutes choses égales par ailleurs, l'effet de l'âge (et donc des variables âge et carré de l'âge) sur les dépenses de santé. Cet effet global de l'âge se décompose alors en un effet linéaire (par l'intérmédiaire de la variable âge) et d'un effet non-linéaire (par l'intermédiaire de la variable carré de l'âge).

4. Finalement, on notera que dans le cas où X n'est pas de rang $p+1$ la mesure des effets d'une variable exogène X_k sur la variable endogène Y au moyen du paramètre β_k n'est plus possible. On a vu dans la remarque 5.2 que cela correspond au cas où les paramètres du modèle ne sont pas identifiés et qu'il existe plusieurs façons de mesurer l'effet « toutes choses égales par ailleurs » d'une augmentation de X_k sur la variable Y.

   La raison est assez proche de celle évoquée dans le point précédent. En effet, lorsque le rang de X n'est pas égal à $p+1$, au moins l'une des variables exogènes s'exprime comme une combinaison linéaires des autres (voir la remarque 5.3). En reprenant l'exemple de la note 2 du bas de la page 91 dans lequel on suppose qu'on peut écrire $X_{ip} = \sum_{k=0}^{p-1} c_k X_{ik}$ pour $i = 1, \ldots, n$, on voit que si X_{i1} augmente d'une unité, alors on a nécessairement que X_{ip} varie de c_1 unité(s). Il est donc impossible d'interpréter β_1 comme la variation attendue de Y_i lorsque X_{i1} augmente d'une unité, toutes choses égales par ailleurs. On a effectivement dans ce cas

$$\begin{aligned} \mathrm{E}(Y_i) &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \\ &= (\beta_0 + c_0\beta_p) + (\beta_1 + c_1\beta_p) X_{i1} + \cdots + (\beta_{p-1} + c_{p-1}\beta_p) X_{ip-1} \end{aligned}$$

   en exprimant la dernière variable exogène en fonction des autres. Cette écriture fait apparaître que l'effet attendu sur Y_i d'une augmentation d'une unité de la variable X_{i1}, toutes choses égales par ailleurs, est $\gamma_1 = \beta_1 + c_1\beta_p$. Les paramètres qui permettent de représenter la relation linéaire entre les variables exogènes et la variable endogène ne sont donc pas ceux apparaissant dans la condition C_p2, mais plutôt des combinaisons linéaires de ces derniers, données par $\gamma_k = \beta_k + c_k\beta_p$, $k = 0, \ldots, p-1$.

   Ceci montre que la relation C_p2, par laquelle on exprime le fait qu'on veut expliquer linéairement Y en fonction des p variables $X_1, \ldots, X_p$, est mal spécifiée, dans la mesure où lorsque les $p-1$ premières variables exogènes sont prises en compte, alors la dernière est redondante (elle est elle-même une combinaison linéaire des autres variables explicatives) et n'est donc pas nécessaire pour expliquer Y.

   Ces conclusions ne sont valables que si les $p-1$ premières variables endogènes sont elles-mêmes linéairement indépendantes. Si ce n'était pas le cas, alors on pourrait itérer le raisonnement qui vient d'être tenu à propos de X_p : on aurait une relation dans laquelle seulement $p-2$ variables exogènes expliquent Y. Si on note r le rang de la matrice X, alors la condition C_p2 est équivalente à une condition qui établit qu'il existe r réels $\gamma_1, \ldots, \gamma_r$ *uniques* pour lesquels on a

$$\mathrm{E}(Y_i) = \gamma_1 X_{ik_1} + \cdots + \gamma_r X_{ik_r}, \quad i = 1, \ldots, n \tag{5.1}$$

eq:decomp_EY_s

où $k_1, \ldots, k_r$ sont r indices distincts parmi $\{0, 1, \ldots, p\}$. Il est en effet clair que si la décomposition (5.1) est vraie, alors C_p2 est également vraie : il suffit de poser $\beta_k = \gamma_j$ s'il existe j tel que $k_j = k$ et $\beta_k = 0$ sinon. Réciproquement, si C_p2 est vraie et si $\text{rang}(X) = r$, alors parmi les $p+1$ vecteurs qui constituent les colonnes de X, il y en a r au maximum qui sont linéairement indépendants. On note $X_{\cdot k_1}, \ldots, X_{\cdot k_r}$ ces r vecteurs. Il forment une base de $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ et on a donc évidemment $L(X_{\cdot 0}, \ldots, X_{\cdot p}) = L(X_{\cdot k_1}, \ldots, X_{\cdot k_r})$. Comme la condition C_p2 établit que $\mathrm{E}(\boldsymbol{Y})$ appartient à $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ (voir la remarque 5.2), on doit avoir que $\mathrm{E}(\boldsymbol{Y})$ s'exprime de manière unique comme une combinaison linéaire des vecteurs formant une base de cet espace. Les coefficients de cette combinaison linéaire sont $\gamma_1, \ldots, \gamma_r$. ℝ

5.3 Estimation des paramètres $\beta_0, \ldots, \beta_p$

sec:estim_mco_multiv

5.3.1 La méthode des moindres carrés

sec:estim_mco_multiv

Dans cette section, on reprend le problème d'estimation des paramètres $\beta_0, \ldots, \beta_p$. La démarche exposée dans la section 2.1 peut s'appliquer ici : on cherche les valeurs des paramètres pour lesquelles les distances (mesurées par les carrés des différences) entre Y_i et la partie de Y_i expliquée par les variables exogènes ont la plus petite moyenne. Minimiser la moyenne de ces distances revient à minimiser leur somme. On est donc amené à minimiser par rapport à $\beta_0, \ldots, \beta_p$ la fonction $S(\beta_0, \ldots, \beta_p)$ définie par

$$S(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2$$

Cette fonction est continue et dérivable par rapport à chacun de ses arguments. Sa minimisation repose donc sur le calcul de ses dérivées premières et secondes. La première étape consiste à trouver un $(p+1)$-uplet $(\hat{\beta}_0, \ldots, \hat{\beta}_p)$ pour lequel chacune des dérivées premières de S s'annule :

$$\frac{\partial S}{\partial \beta_k}(\hat{\beta}_0, \ldots, \hat{\beta}_p) = 0, \quad k = 0, \ldots, p \tag{5.2}$$

eq:mcocpo1

Pour tout $k = 0, \ldots, p$, S est un polynôme de degré 2 en β_k. Donc le membre de gauche de la k^{e} équation de (5.2) est linéaire en β_l, $l = 0, \ldots, p$, $k = 0, \ldots, p$. Ces $p+1$ équations forment donc un système linéaire à $p+1$ inconnues, et une reformulation matricielle à l'aide des éléments $\boldsymbol{Y}$, β et X introduits à la section 5.1 permet d'en exprimer facilement les solutions.

Notons qu'en utilisant la notation $X_{i0} = 1$ pour tout i, pour $k = 0, \ldots, p$, on a

$$\frac{\partial S}{\partial \beta_k}(\beta_0, \ldots, \beta_p) = -2 \sum_{i=1}^{n} X_{ik}(Y_i - \beta_0 X_{i0} - \beta_1 X_{i1} - \cdots - \beta_p X_{ip}) \tag{5.3}$$

eq:derivS1

Avec les définitions de β et de $X_{i\cdot}$, $i = 1, \ldots, n$, introduites à la section 5.1, on peut écrire $\frac{\partial S}{\partial \beta_k}(\beta) = -2 \sum_{i=1}^{n} X_{ik}(Y_i - X_{i\cdot}^\top \beta)$, $k = 0, 1, \ldots, p$. On remarque que $Y_i - X_{i\cdot}^\top \beta$ est la i^{e} coordonnée du vecteur (aléatoire) $\boldsymbol{Y} - X\beta$ de $\mathbb{R}^n$. Par conséquent, on peut également écrire :

$$\frac{\partial S}{\partial \beta_k}(\beta) = -2 X_{\cdot k}^\top (\boldsymbol{Y} - X\beta) \qquad k = 0, 1, \ldots, p \tag{5.4}$$

eq:derivS2

où les vecteurs $X_{\bullet k}$, $k = 0, 1, \ldots, p$, ont été définis en 5.1. Par conséquent, pour minimiser S on cherche un $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^\top \in \mathbb{R}^{p+1}$ tel que

$$X^\top(\boldsymbol{Y} - X\hat{\beta}) = 0_{p+1} \tag{5.5}$$ eq:mcocpo2

où 0_{p+1} désigne le vecteur nul de $\mathbb{R}^{p+1}$, ou encore,

$$X^\top X\hat{\beta} = X^\top \boldsymbol{Y} \tag{5.6}$$ eq:eq_norm

Un tel $\hat{\beta}$ existe et est unique si et seulement si la matrice $X^\top X$ est inversible. Dans ce cas, on obtient

$$\hat{\beta} = (X^\top X)^{-1} X^\top \boldsymbol{Y}$$

Sous cette condition d'inversibilité, il faut vérifier que $\hat{\beta}$ réalise le minimum de S. Ce sera le cas si la matrice des dérivées secondes de S est définie positive en $\hat{\beta}$. À partir de (5.3), on calcule la $(k, l)^{\text{e}}$ entrée de cette matrice :

$$\frac{\partial^2 S}{\partial\beta_k \partial\beta_l}(\beta) = 2\sum_{i=1}^{n} X_{ik}X_{il} = 2X_{\bullet k}^\top X_{\bullet l}$$

On en déduit que la matrice des dérivées secondes de S est $2X^\top X$. On vérifie qu'elle est définie positive. Soit $a \neq 0_{p+1}$ un vecteur non nul de $\mathbb{R}^{p+1}$. On a

$$a^\top(X^\top X)a = (Xa)^\top Xa = \sum_{i=1}^{n} A_i^2 \tag{5.7}$$ eq:XtX_dp

où $A_i = \sum_{k=0}^{p} X_{ik}a_k = X_{i\bullet}^\top a$ est la i^{e} coordonnée du vecteur $A = Xa$. Donc $X^\top X$ est définie positive si et seulement si $Xa \neq 0$, pour tout $a \in \mathbb{R}^{p+1}$, $a \neq 0_{p+1}$. Or cette condition est nécessairement vérifiée puisqu'elle équivaut à la condition que $X^\top X$ est inversible (voir la remarque 5.3), ce que nous avons supposé. On a donc prouvé le résultat suivant.

pro:est_mco_mrls **Propriété 5.3** *Si la matrice X, de taille $(n, p+1)$ et d'élément constitutif X_{ik}, est de rang $p+1$, alors S admet un unique minimum, atteint en $\hat{\beta} = (X^\top X)^{-1}X^\top \boldsymbol{Y}$. On appelle $\hat{\beta}$ l'estimateur des moindres carrés ordinaires (MCO) de β. La $(k+1)^e$ coordonnée $\hat{\beta}_k$ de $\hat{\beta}$ est l'estimateur des MCO de β_k.*

Si X est de rang inférieur à $p+1$, la fonction S admet un même minimum en plusieurs points de $\mathbb{R}^{p+1}$. On dit dans ce cas que l'estimateur des MCO de β n'existe pas.

Remarque 5.6 Il est naturel que l'estimateur des MCO de β n'existe pas lorsque le rang de X n'est pas égal à $p+1$. En effet, nous avons noté à la remarque 5.2, que dans ce cas la décomposition de $\mathrm{E}(\boldsymbol{Y})$ comme combinaison linéaire des vecteurs $X_{\bullet 0}, \ldots, X_{\bullet p}$ n'était pas unique. Dans ce cas, l'équation (5.6) qui caractérise les solutions de la minimisation de S montre ces solutions sont multiples. Le résultat énoncé ci-dessus montre alors que si les paramètres $\beta_0, \ldots, \beta_p$ sont non-identifiés, alors l'estimateur de MCO de ces paramètres n'existe pas. R

Remarque 5.7 Pour l'interprétation graphique de la conséquence du rang de X sur la minimisation de S, voir les graphiques 2.3 (cas $\mathrm{rang}(X) = p+1$) et 2.5 (cas $\mathrm{rang}(X) < p+1$) du chapitre 2. R

rem:mco_et_mrls

Remarque 5.8 Il est important/intéressant de noter que pour dériver la solution du problème de minimisation de la fonction S, il n'a été fait aucun usage des conditions C_p1 à C_p3 qui définissent le MRLS. Autrement dit, la minimisation de S admet pour solution $\hat{\beta} = (X^\top X)^{-1} X^\top \boldsymbol{Y}$ que ces conditions soient vraies ou pas.[5] Le fait de se placer dans le contexte d'un MRLS n'intervient que dans le choix d'une méthode d'estimation, qui est celle qui vient d'être exposée et qui conduit à utiliser la solution du problème de minimisation de S comme estimateur du vecteur β des paramètres du modèle. On rappelle que ce choix est basé sur l'observation que la condition C_p2 impose à $\mathrm{E}(\boldsymbol{Y})$ d'être un combinaison linéaire des vecteurs constituant les colonnes de X ; on essaie alors d'approximer une telle combinaison linéaire par celle qui est la plus proche de $\boldsymbol{Y}$. Cette remarque reste également valable pour tout le contenu de la section qui suit.

En revanche, comme on le verra à la section 5.3.3, les propriétés de $\hat{\beta}$ en tant qu'estimateur de β (biais, précision, etc) dépendront de celles de $\boldsymbol{Y}$ et de X, et en particulier de la relation qui les lie l'un à l'autre. On voit donc que ces propriétés découleront bien des conditions C_p1 à C_p3 définissant le modèle. R

sec:inter_geo_mco

5.3.2 Interprétation géométrique de l'estimation par moindres carrés

L'estimateur des MCO de β obtenu en minimisant la fonction S donne lieu à des interprétations géométriques intéressantes. Pour cela, il faut reconsidérer le problème de minimisation de S, en rappelant des éléments d'algèbre élémentaires sur l'espace vectoriel $\mathbb{R}^n$ (sur $\mathbb{R}$).

On rappelle que le produit scalaire de deux vecteurs $u = (u_1, \ldots, u_n)^\top$ et $v = (v_1, \ldots, v_n)^\top$ de $\mathbb{R}^n$ est le réel noté $\langle u, v \rangle$ défini par $\langle u, v \rangle = \sum_{i=1}^n u_i v_i = u^\top v$.[6] Ce produit scalaire permet de définir la norme d'un vecteur u, notée $\|u\|$, par $\|u\| = \sqrt{\langle u, u \rangle}$.

Par conséquent, puisque $(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})$ est la i^{e} coordonnée du vecteur $\boldsymbol{Y} - X\beta$, on peut écrire

$$S(\beta_0, \ldots, \beta_p) = \|\boldsymbol{Y} - X\beta\|^2 \tag{5.8}$$

eq:Snorme

Donc le problème de minimisation de S s'écrit $\min_{\beta \in \mathbb{R}^{p+1}} \|\boldsymbol{Y} - X\beta\|^2$.

On note maintenant que par construction, tout vecteur de $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ s'écrit sous la forme $X\beta$ pour un certain $\beta \in \mathbb{R}^{p+1}$ et réciproquement, tout vecteur de $\mathbb{R}^n$ s'écrivant sous la forme $X\beta$ est dans $L(X_{\cdot 0}, \ldots, X_{\cdot p})$. Par conséquent, chercher le $\beta \in \mathbb{R}^{p+1}$ qui minimise $\|\boldsymbol{Y} - X\beta\|^2$ revient à chercher le vecteur $\hat{\boldsymbol{Y}}$ de $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ tel que

$$\|\boldsymbol{Y} - U\|^2 \geq \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 \qquad \forall U \in L(X_{\cdot 0}, \ldots, X_{\cdot p}) \tag{5.9}$$

eq:Smin_distan

Autrement dit, le problème de minimisation de S est équivalent à $\min_{U \in L(X_{\cdot 0}, \ldots, X_{\cdot p})} \|\boldsymbol{Y} - U\|^2$. Le point 5 de la propriété 9.22 (section 9.2) établit que la solution de ce problème est le vecteur $\hat{\boldsymbol{Y}}$ de $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ correspondant à la projection orthogonale de $\boldsymbol{Y}$ sur cet espace. Si les vecteurs $X_{\cdot 0}, \ldots, X_{\cdot p}$ sont linéairement indépendants, ils forment une base de cet espace. Comme ils forment les colonnes de la matrice X, celle-ci est de rang $p + 1$, et le point 4 de cette même propriété 9.22

5. La seule condition nécessaire porte sur le rang de X, ce qui n'est pas une condition permettant de définir le MRLS.

6. Comme on l'a fait jusqu'à présent, on assimile un vecteur de $\mathbb{R}^n$ au n-uplet de ses coordonnées.

permet d'écrire que la matrice associée à l'application de projection orthogonale sur $L(X_{\bullet 0}, \ldots, X_{\bullet p})$ est $P_L = X(X^\top X)^{-1}X^\top$. On a alors

$$\hat{\boldsymbol{Y}} = P_L \boldsymbol{Y} = X(X^\top X)^{-1}X^\top \boldsymbol{Y} \tag{5.10}$$ eq:yhat

Puisque par construction $\hat{\boldsymbol{Y}} \in L(X_{\bullet 0}, \ldots, X_{\bullet p})$, il doit s'écrire sous la forme $\hat{\boldsymbol{Y}} = X\hat{\beta}$ pour un certain $\hat{\beta} \in \mathbb{R}^{p+1}$. Comme la matrice X est de rang $p+1$, ou de manière équivalente, ses colonnes forment une base de $L(X_{\bullet 0}, \ldots, X_{\bullet p})$, un tel $\hat{\beta}$ est unique. L'expression de $\hat{\boldsymbol{Y}}$ donnée par (5.10) montre qu'on doit alors avoir

$$\hat{\beta} = (X^\top X)^{-1}X^\top \boldsymbol{Y}$$

ce qui est bien la solution trouvée dans la section précédente.

On vient de montrer que dans le MRLS où $\mathrm{rang}(X) = p+1$, l'estimateur des MCO de β est le vecteur $\hat{\beta}$ des coordonnées de la projection orthogonale sur le sous-espace $L(X_{\bullet 0}, \ldots, X_{\bullet p})$ de $\mathbb{R}^n$ engendré par les vecteurs $X_{\bullet 0}, \ldots, X_{\bullet p}$ contenant les observations des variables exogènes, et dont ils forment une base. Ce résultat est illustré par la séquence de graphiques de la figure 5.2.

La reformulation de S donnée en (5.8) permet une interprétation intéressante de $\hat{\boldsymbol{Y}}$. On rappelle que l'application qui à deux vecteurs u et v de $\mathbb{R}^n$ fait correspondre la norme de leur différence $\|u-v\|$ est une distance (elle est non-négative, symétrique, s'annule si et seulement si $u = v$ et satisfait l'inégalité triangulaire). Par conséquent, minimiser $S(\beta) = \|\boldsymbol{Y} - X\beta\|^2$ par rapport à $\beta \in \mathbb{R}^{p+1}$ ou encore trouver $\hat{\boldsymbol{Y}}$ satisfaisant l'inégalité (5.9) revient à chercher le vecteur de $L(X_{\bullet 0}, \ldots, X_{\bullet p})$ pour lequel la distance avec $\boldsymbol{Y}$ est la plus petite. On voit donc qu'estimer β au moyen de l'estimateur des moindres carrés revient à trouver les coefficients de la combinaison linéaire des vecteurs $X_{\bullet 0}, \ldots, X_{\bullet p}$ formant le vecteur le plus proche de $\boldsymbol{Y}$. Ces coefficients sont les composantes $\hat{\beta}_0, \ldots, \hat{\beta}_p$ de $\hat{\beta}$.

rem:proj_et_mco **Remarque 5.9** Dans le même esprit que la remarque 5.8, on note que l'interprétation géométrique qui vient d'être donnée de la méthode d'estimation par moindres carrés des paramètres du MRLS, et qui montre que la démarche d'estimation est assimilable à une projection orthogonale, peut être abstraite du contexte du MRLS. En effet, ce dernier est introduit parce qu'on veut représenter une relation entre variables, à laquelle on sait donner un sens (*i.e.*, qu'on sait interpréter). Cependant, indépendamment de tout sens qu'on pourrait donner à une telle relation, l'équivalence entre moindres carrés et projection orthogonale demeure, puisque ni la validité de la minimisation qui conduit à l'estimation par moindres carrés, ni le résultat montrant que cette minimisation revient à effectuer une projection orthogonale ne s'appuie sur le fait que les conditions C_p1 à C_p3 sont satisfaites ou pas. Par conséquent, si on se donne $q+1$ vecteurs de $\mathbb{R}^n$, notés $Z, U_1, \ldots, U_q$ de sorte que $U_1, \ldots, U_q$ soient linéairement indépendants, on peut s'intéresser à la projection orthogonale de Z sur le sous-espace de $\mathbb{R}^n$ engendré par $U_1, \ldots, U_q$. Ainsi qu'on l'a noté, cela revient à chercher la combinaison linéaire de ces vecteurs la plus proche de Z. Les coefficients qui définissent cette combinaison linéaire peuvent s'interpréter comme les estimateurs des paramètres d'une "relation" entre une "variable endogène" dont les observations seraient les coordonnées de Z et q "variables exogènes", la k^{e} d'entre elles ayant pour observations les coordonnées de U_k. En effet, si on construisait de manière artificielle (c'est à dire indépendamment de tout objectif de représenter ou d'approximer une réalité quelconque) un modèle de régression dans lequel la variable exogène est Z et les variables explicatives sont $U_1, \ldots, U_q$, l'estimation par moindres carrés des paramètres nous amènerait à résoudre un problème de minimisation dont la solution nous donnerait les coordonnées de

la projection orthogonale de Z sur l'espace engendré par $U_1, \ldots, U_q$. Il faut noter que la relation qui permettrait de définir un tel modèle est parfaitement fictive puisqu'on ne prétend pas qu'elle existe ou qu'elle approxime une relation existante. Elle sert d'auxiliaire qui permet de faire le lien entre l'estimation de ses paramètres par moindres carrés et une projection orthogonale. R

sec:propr_mco

5.3.3 Propriétés de l'estimateur des moindres carrés

Cette section présente la propriété la plus importante de $\hat{\beta}$, qui établit que l'estimateur MCO de β est optimal dans la classe des estimateurs linéaires et sans biais. Ce résultat est identique à celui obtenu dans la section 2.1. Il est cependant établi ici dans le contexte plus général d'un MRLS en utilisant une approche plus globale. Puisqu'on étudie les propriétés de $\hat{\beta}$, on supposera qu'il existe, et sans qu'on le rappelle par la suite, on se placera toujours sous la condition que $\text{rang}(X) = p+1$.

On commence par montrer que $\hat{\beta}$ appartient bien à la classe des estimateurs considérés.

def:est_lin_mrls

Définition 5.1 *Un estimateur de β est linéaire s'il peut s'écrire sous la forme $A\boldsymbol{Y}$, où A est une matrice connue, non aléatoire.*

On rappelle qu'un estimateur $\tilde{\beta}$ du vecteur des paramètres β est sans biais si quelle que soit la valeur de ce vecteur, l'espérance de $\tilde{\beta}$ est égale à cette valeur, ou formellement : $\mathrm{E}(\tilde{\beta} - \beta) = 0_{p+1}$, $\forall \beta \in \mathbb{R}^{p+1}$.

On obtient facilement la condition pour qu'un estimateur linéaire soit sans biais. Cette condition s'écrit $\mathrm{E}(A\boldsymbol{Y} - \beta) = 0_{p+1},\ \forall \beta \in \mathbb{R}^{p+1}$. Comme A et β sont non-aléatoires, cette condition s'exprime également $A\mathrm{E}(\boldsymbol{Y}) - \beta = 0_{p+1},\ \forall \beta \in \mathbb{R}^{p+1}$, ou encore, en utilisant la condition $\mathrm{C}_p2 : (AX - I_{p+1})\beta = 0_{p+1},\ \forall \beta \in \mathbb{R}^{p+1}$. Cette égalité est évidemment vraie si $AX = I_{p+1}$. Pour qu'elle soit vraie quelle que soit β dans $\mathbb{R}^{p+1}$ il est également nécessaire d'avoir $AX = I_{p+1}$. En résumé, dans le MRLS un estimateur linéaire de β est sans biais si et seulement si la matrice A qui le caractérise satisfait $AX = I_{p+1}$. On a immédiatement la propriété suivante.

pro:emco_lin_sb

Propriété 5.4 *Dans le MRLS, l'estimateur MCO défini dans la propriété 5.3 est un estimateur linéaire et sans biais de β*

Preuve : En choisissant $A = (X^\top X)^{-1} X^\top$, on voit que $\hat{\beta}$ a bien la forme donnée dans la définition 5.1. On vérifie facilement que $AX = I_{p+1}$.

Si on souhaite montrer que $\hat{\beta}$ est le meilleur dans la classe des estimateurs linéaires et sans biais, il faut établir un critère qui permette de comparer deux estimateurs dans cette classe. Ce critère doit tenir compte du fait que dans le contexte du MRLS, le paramètre est multidimensionnel (*i.e.*, dont la valeur est un $(p+1)$-uplet ou un vecteur de réels). [7] Pour aboutir à un critère de comparaison dans ce contexte, on reprend le raisonnement de la section 2.2, qui avait conduit à l'utilisation de l'erreur quadratique moyenne (EQM) pour comparer deux estimateurs d'un paramètre unidimensionnel. La justification de ce choix repose sur l'interprétation de l'EQM d'un estimateur comme un indicateur de sa précision, puisque l'EQM mesure la distance attendue entre l'estimateur et ce qu'il estime. En reprenant ce qu'on a dit dans le dernier paragraphe de la section 5.3.2, la distance permettant de définir l'EQM pour un estimateur $\tilde{\beta}$ de β se mesure par $\|\tilde{\beta} - \beta\|^2$, et l'EQM elle-même est $\mathrm{E}(\|\tilde{\beta} - \beta\|^2)$.

7. Cet aspect était également présent dans la discussion de la section 2.2, mais il n'a pas été abordé.

En utilisant ce critère, on préfèrera un estimateur $\tilde{\beta}$ à β^* si $\mathrm{E}(\|\tilde{\beta} - \beta\|^2) \leq \mathrm{E}(\|\beta^* - \beta\|^2)$ pour tout $\beta \in \mathbb{R}^{p+1}$. En utilisant les rappels faits au début de la section 5.3.2, on peut écrire

$$\mathrm{E}(\|\tilde{\beta} - \beta\|^2) = \mathrm{E}\left[\sum_{k=0}^{p} (\tilde{\beta}_k - \beta_k)^2\right] = \sum_{k=0}^{p} \mathrm{E}[(\tilde{\beta}_k - \beta_k)^2]$$

Or $\mathrm{E}[(\tilde{\beta}_k - \beta_k)^2]$ est l'EQM de l'estimateur $\tilde{\beta}_k$ de β_k. Donc l'EQM de $\tilde{\beta}$ est la somme des EQM de ses éléments. Par conséquent, le critère de comparaison introduit ici revient à préférer $\tilde{\beta}$ à β^* si

$$\frac{1}{p+1} \sum_{k=0}^{p} \mathrm{E}[(\tilde{\beta}_k - \beta_k)^2] \leq \frac{1}{p+1} \sum_{k=0}^{p} \mathrm{E}[(\beta_k^* - \beta_k)^2]$$

c'est-à-dire si, *en moyenne*, l'EQM des composantes de $\tilde{\beta}$ est plus petite que l'EQM des composantes de β^*. Avec un tel critère de comparaison, on peut être amené à préférer $\tilde{\beta}$ à β^*, bien qu'il soit possible que pour certains éléments de β l'estimateur β^* soit plus précis que $\tilde{\beta}$, dans le sens où pour il est possible d'avoir $\mathrm{E}[(\tilde{\beta}_k - \beta_k)^2] \geq \mathrm{E}[(\beta_k^* - \beta_k)^2]$ pour un certain k. Autrement dit ce critère peut amener à préferer des estimateurs précis en moyenne, mais peu précis pour quelques éléments de β.

Ce critère n'est pas satisfaisant, et on le remplace par un critère qui, au lieu d'amener à préférer un estimateur dont les EQM de chacun de ses éléments sont *en moyenne* plus petites, conduit à préférer $\tilde{\beta}$ à β^* si, *composante par composante*, le premier a une EQM plus petite que le second, ou formellement, si $\mathrm{E}[(\tilde{\beta}_k - \beta_k)^2] \leq \mathrm{E}[(\beta_k^* - \beta_k)^2]$, pour $k = 0, \ldots, p$. Cette approche conduit au critère suivant. [8]

def:best_eqm **Définition 5.2** *Soient $\tilde{\beta}$ et β^* deux estimateurs d'un même paramètre β dont la valeur est dans $\mathbb{R}^{p+1}$. On dit que $\tilde{\beta}$ est préférable (au sens de l'EQM) à β^* si pour tout $\beta \in \mathbb{R}^{p+1}$, la matrice $\mathrm{E}\big[(\beta^* - \beta)(\beta^* - \beta)^\top\big] - \mathrm{E}\big[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^\top\big]$ est semi-definie positive.*

rem:EQMcom **Remarque 5.10** Cette définition appelle plusieurs remarques.

1. Comme la $(k,l)^{\mathrm{e}}$ entrée de la matrice $(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^\top$ est $(\tilde{\beta}_k - \beta_k)(\tilde{\beta}_l - \beta_l)$, le k^{e} élément de la diagonale de la matrice $\mathrm{E}\big[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^\top\big]$ est l'EQM de $\tilde{\beta}_k$. Définissons pour tout $k = 0, \ldots, p$ le vecteur a_k de $\mathbb{R}^{p+1}$ dont la $(k+1)^{\mathrm{e}}$ coordonnée vaut 1 et toutes les autres sont nulles. Si $\tilde{\beta}$ est préférable à β^*, la définition 5.2 implique que pour tout $k = 0, \ldots, p$ et tout $\beta \in \mathbb{R}^{p+1}$ on a

$$a_k^\top \Big(\mathrm{E}\big[(\beta^* - \beta)(\beta^* - \beta)^\top\big] - \mathrm{E}\big[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^\top\big]\Big) a_k \geq 0$$

   Avec la forme donnée aux a_k, on vérifie facilement que cette inégalité revient à écrire que pour tout $k = 0, \ldots, p$ et pour tout $\beta \in \mathbb{R}^{p+1}$ on a

$$\mathrm{E}[(\beta_k^* - \beta_k)^2] \geq \mathrm{E}[(\tilde{\beta}_k - \beta_k)^2]$$

   Le critère de comparaison d'estimateurs donné à la définition 5.2 répond bien à l'objectif annoncé juste avant cette définition.

8. On rappelle que si A est une matrice dont les entrées sont des variables aléatoires, alors $\mathrm{E}(A)$ est un matrice dont la $(i,j)^{\mathrm{e}}$ entrée est l'espérance de la $(i,j)^{\mathrm{e}}$ entrée de A.

it:EQMcom

2. Cet objectif est même dépassé. En effet, on vient de montrer que si on s'intéresse à l'estimation d'une composante donnée de β, alors le critère de la définition 5.2 permet de sélectionner l'estimateur le plus précis. Si au lieu de s'intéresser à une des composantes de β on souhaite estimer une combinaison linéaire de plusieurs de ces composantes, alors le critère permettra aussi de comparer deux estimateurs et d'en sélectionner le meilleur. Pour le montrer, on se donne $p+1$ réels quelconques $c_0, \ldots, c_p$ et on considère l'estimation du nouveau paramètre γ défini comme $\gamma = c^\top \beta = c_0\beta_0 + \cdots + c_p\beta_p$, où $c = (c_0, \ldots, c_p)^\top$. Si $\tilde{\beta}$ est préférable à β^*, alors pour tout $\beta \in \mathbb{R}^{p+1}$ on doit avoir

$$c^\top \Big(\mathrm{E}\big[(\beta^* - \beta)(\beta^* - \beta)^\top\big] - \mathrm{E}\big[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^\top\big] \Big) c \geq 0$$

En développant, on a $\mathrm{E}\big[c^\top(\beta^* - \beta)(\beta^* - \beta)^\top c\big] - \mathrm{E}\big[c^\top(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^\top c\big] \geq 0$. On note que $c^\top(\beta^* - \beta) \in \mathbb{R}$, et que par conséquent l'inégalité s'écrit encore $\mathrm{E}\big[\big(c^\top(\beta^* - \beta)\big)^2\big] - \mathrm{E}\big[\big(c^\top(\tilde{\beta} - \beta)\big)^2\big] \geq 0$. Finalement, en développant les termes à l'intérieur de l'espérance et en utilisant la notation introduite ci-dessus, on a

$$\mathrm{E}\big[(c^\top\beta^* - \gamma)^2\big] - \mathrm{E}\big[(c^\top\tilde{\beta} - \gamma)^2\big] \geq 0$$

pour tout $\gamma \in \mathbb{R}$. Comme estimateur de la combinaison linéaire $\gamma = c^\top\beta$ des éléments de β, on peut considérer la même combinaison linéaire, prise sur les éléments de β^*, donnée par $\gamma^* = c^\top\beta^*$. De la même manière, on peut également former l'estimateur $\tilde{\gamma} = c^\top\tilde{\beta}$ de γ. L'inégalité ci-dessus s'écrit alors $\mathrm{E}\big[(\gamma^* - \gamma)^2\big] \geq \mathrm{E}\big[(\tilde{\gamma} - \gamma)^2\big]$. Elle montre que si pour estimer β, $\tilde{\beta}$ est préférable à β^*, alors pour estimer une combinaison linéaire de β, la combinaison linéaire formée à partir de $\tilde{\beta}$ est préférable à celle obtenue à partir de β^*.

La définition 5.2 établit alors une équivalence : $\tilde{\beta}$ est préférable à β^* si et seulement si quelle que soit la combinaison linéaire $c^\top\beta$ des éléments de β, l'estimateur $c^\top\tilde{\beta}$ de cette combinaison linéaire est préférable à l'estimateur $c^\top\beta^*$.

3. Finalement, comme on l'a déjà noté, si $\tilde{\beta}$ est un estimateur sans biais, alors sa matrice des variances-covariances $\mathrm{V}(\tilde{\beta})$ coïncide avec $\mathrm{E}\big[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^\top\big]$. Par conséquent, si $\tilde{\beta}$ et β^* sont deux estimateurs sans biais de β, $\tilde{\beta}$ est préférable à β^* si pour tout $\beta \in \mathbb{R}^{p+1}$ la $\mathrm{V}(\beta^*) - \mathrm{V}(\tilde{\beta})$ est définie positive. R

On dispose maintenant d'un critère qui permet de comparer deux estimateurs. En utilisant ce critère, on montre que dans la classe des estimateurs linéaires et sans biais considérée dans cette section, l'estimateur MCO $\hat{\beta}$ est préférable à tout autre estimateur de β dans cette classe.

th:gmm

Théorème 5.1 (Gauss-Markov) *Dans le contexte du MRLS, si $\tilde{\beta}$ est un estimateur linéaire et sans biais de β, alors la matrice $\mathrm{V}(\tilde{\beta}) - \mathrm{V}(\hat{\beta})$ est semi-definie positive pour tout $\beta \in \mathbb{R}^{p+1}$. Donc dans le MRLS, $\hat{\beta}$ est le meilleur estimateur linéaire sans biais de β.*

Preuve : Soit $\tilde{\beta}$ un estimateur linéaire sans biais de β. En utilisant le paragraphe qui précède la propriété 5.4, on peut écrire $\tilde{\beta} = \tilde{A}\boldsymbol{Y}$ pour une certaine une matrice $\tilde{A}$ non-aléatoire et telle que $\tilde{A}X = I_{p+1}$. Par ailleurs, la propriété 5.4 établit que $\hat{\beta} = \hat{A}\boldsymbol{Y}$ avec $\hat{A} = (X^\top X)^{-1}X^\top$. On calcule la matrice des variances-covariances de ces deux estimateurs. On a

$$\mathrm{V}(\tilde{\beta}) = \mathrm{V}(\tilde{A}\boldsymbol{Y}) = \tilde{A}\mathrm{V}(\boldsymbol{Y})\tilde{A}^\top = \sigma^2\tilde{A}\tilde{A}^\top$$

où la deuxième égalité provient du fait que $\tilde{A}$ est-non aléatoire (et en utilisant la propriété 9.7 de la section 9.1.2) et la troisième de la condition C_p3. De la même manière on obtient $V(\hat{\beta}) = \sigma^2 \hat{A}\hat{A}^\top$. Il faut montrer que $\sigma^2(\tilde{A}\tilde{A}^\top - \hat{A}\hat{A}^\top)$ est semi-définie positive. Comme $\sigma^2 > 0$ (voir la condition C_p3), il est équivalent de montrer que $\tilde{A}\tilde{A}^\top - \hat{A}\hat{A}^\top$ est une matrice semi-definie positive. On note que l'expression de $\hat{A}$ et la condition $\tilde{A}X = I_{p+1}$ impliquent $\tilde{A}\hat{A}^\top = \tilde{A}X(X^\top X)^{-1} = (X^\top X)^{-1}$. Par ailleurs, un calcul direct montre que $\hat{A}\hat{A}^\top = (X^\top X)^{-1}$. On en déduit donc que $\tilde{A}\hat{A}^\top = \hat{A}\hat{A}^\top$. Pour obtenir le résultat recherché, il suffit d'introduire la matrice B définie comme $B = \tilde{A} - \hat{A}$. En développant le produit $BB^\top$ et en utilisant les égalités précédentes, on a $\tilde{A}\tilde{A}^\top - \hat{A}\hat{A}^\top = BB^\top$. On voit de par sa forme que cette matrice est semi-definie positive (voir par exemple 5.7).

Le théorème 5.1 est un résultat d'optimalité pour l'estimation des paramètres du MRLS. À ce titre, c'est le résultat le plus important dans ce contexte. Il justifie le choix de la méthode permettant d'obtenir l'estimateur MCO. Ce résultat peut être légèrement généralisé en montrant que l'optimalité de l'estimateur MCO peut être étendue au cas où on s'intéresse à l'estimation de plusieurs combinaisons linéaires des paramètres $\beta_0, \ldots, \beta_p$.

Le résultat de théorème 5.1 montre directement que si on souhaite estimer la combinaison linéaire $\gamma = c^\top \beta$ des éléments de β, alors $c^\top \hat{\beta}$ est un estimateur linéaire et sans biais de γ qui est préférable à tout autre estimateur de la forme $c^\top \tilde{\beta}$, où $\tilde{\beta}$ est un estimateur linéaire et sans biais de β. En effet, pour que cela soit la cas, il faut que $V(c^\top \tilde{\beta}) \geq V(c^\top \hat{\beta})$ pour tout $\beta \in \mathbb{R}^{p+1}$. En utilisant la propriété 9.7, cette condition est équivalente à $c^\top V(\tilde{\beta})c \geq c^\top V(\hat{\beta})c$ et le théorème 5.1 montre qu'elle est satisfaite.

On généralise ce résultat en montrant non seulement que $c^\top \hat{\beta}$ est un meilleur estimateur de γ que tous les estimateurs linéaires et sans biais de γ (et pas seulement meilleur que les estimateurs de la forme $c^\top \tilde{\beta}$), mais également que ce résultat reste vrai si on veut estimer m combinaisons linéaires $c_1^\top \beta, \ldots, c_m^\top \beta$.

th:gmm2 **Théorème 5.2** *Soit C une matrice de taille $(m, p+1)$ dont les entrées sont des réels. On considère le paramètre $\Gamma = C\beta = (\gamma_1, \ldots, \gamma_m)^\top$, où $\gamma_l = c_l^\top \beta$ et $c_l^\top$ est la l^e ligne de C. L'estimateur $\hat{\Gamma}$ de Γ défini par $\hat{\Gamma} = C\hat{\beta}$ est préférable à tout autre estimateur linéaire et sans biais de Γ.*

Preuve : On a $\hat{\Gamma} = \hat{C}\boldsymbol{Y}$, où $\hat{C} = C(X^\top X)^{-1}X^\top$. Donc $\hat{\Gamma}$ est un estimateur linéaire et on vérifie aisément qu'il est sans biais : $E(\hat{\Gamma}) = C\beta = \Gamma$, quelle que soit la valeur possible de Γ. Donc, si on se donne $\tilde{\Gamma}$ un autre estimateur linéaire et sans biais de Γ, il faut montrer que $V(\tilde{\Gamma}) - V(\hat{\Gamma})$ est semi-définie positive. Soit donc $\tilde{\Gamma}$ un estimateur linéaire de Γ, de la forme $\tilde{\Gamma} = \tilde{C}\boldsymbol{Y}$. C'est un estimateur sans biais de Γ si et seulement si $\tilde{C}X\beta = C\beta$ pour tout $\beta \in \mathbb{R}^{p+1}$, ou encore $D\beta = 0_m$, $\forall \beta \in \mathbb{R}^{p+1}$, où D est la matrice $\tilde{C}X - C$. La condition d'absence de biais équivaut à ce que chaque ligne de D soit un vecteur de $\mathbb{R}^{p+1}$ orthogonal à tout vecteur de $\mathbb{R}^{p+1}$. Le seul vecteur satisfaisant cette condition est 0_{p+1}. On doit donc avoir $D = 0$, c'est à dire $\tilde{C}X = C$. On calcule maintenant les variances. En utilisant la propriété 9.7, on a $V(\hat{\Gamma}) = \sigma^2 \hat{C}\hat{C}^\top$ et $V(\tilde{\Gamma}) = \sigma^2 \tilde{C}\tilde{C}^\top$. On procède alors exactement comme dans la preuve du théorème 5.1. On note que grâce à la forme de $\hat{C}$ et à la condition sur $\tilde{C}$ garantissant l'absence de biais pour $\tilde{\Gamma}$, on a $\tilde{C}\hat{C}^\top = \hat{C}\hat{C}^\top$. Par conséquent, si on introduit

la matrice $G = \tilde{C} - \hat{C}$, on obtient

$$GG^\top = (\tilde{C} - \hat{C})(\tilde{C} - \hat{C})^\top = \tilde{C}\tilde{C}^\top - \hat{C}\hat{C}^\top$$

et on conclut comme dans la preuve du théorème 5.1.

Ce théorème permet d'obtenir tous les autres résultats d'optimalité liés à l'estimation des paramètres du MRLS. En choisissant $C = I_{p+1}$, on retrouve le théorème 5.1. En posant $C = a_k^\top$ où a_k est le vecteur de $\mathbb{R}^{p+1}$ dont la $(k+1)^{\text{e}}$ coordonnée vaut 1 et les autres 0, on obtient le résultat qui montre que $\hat{\beta}_k$ est le meilleur estimateur linéaire sans biais de β_k. En outre le théorème 5.1 permet d'obtenir des résultats nouveaux. Le résultat annoncé établissant que $c^\top\hat{\beta}$ est le meilleur parmi tous les estimateurs linéaires et sans biais de $c^\top\beta$ s'obtient en choisissant $C = c^\top$. On peut également montrer un résultat important qui généralise celui concernant l'estimation d'un élément de β. En effet, si on choisit

$$C = \begin{pmatrix} a_{k_1} \\ a_{k_2} \\ \vdots \\ a_{k_m} \end{pmatrix}$$

où $k_1, \ldots, k_m$ sont m indices parmi $\{0, \ldots, p\}$, alors

$$\Gamma = C\beta = \begin{pmatrix} \beta_{k_1} \\ \beta_{k_2} \\ \vdots \\ \beta_{k_m} \end{pmatrix}$$

est un sous-vecteur de β. Le théorème 5.2 montre que le meilleur estimateur linéaire de ce sous-vecteur est

$$\begin{pmatrix} \hat{\beta}_{k_1} \\ \hat{\beta}_{k_2} \\ \vdots \\ \hat{\beta}_{k_m} \end{pmatrix}$$

c'est-à-dire le sous-vecteur correspondant de $\hat{\beta}$.

rem:optim_mco_cbeta

Remarque 5.11 On peut noter qu'il existe une manière identique à celle de la section 2.2 permettant d'établir l'optimalité de l'estimateur des MCO dans l'ensemble des estimateurs linéaires et sans biais. On a vu dans le point 2 de la remarque 5.10 qu'un estimateur de β meilleur qu'un autre est également meilleur pour estimer n'importe quelle combinaison linéaire $c^\top\beta$ des coordonnées de β. Donnons-nous $c \in \mathbb{R}^{p+1}$ quelconque (mais non nul), et considérons l'estimation de la combinaison linéaire $\gamma = c^\top\beta$ au moyen d'un estimateur linéaire et sans biais. Cet estimateur est donc de la forme $a^\top\boldsymbol{Y}$, avec $a \in \mathbb{R}^n$ et doit satisfaire $\mathrm{E}(a^\top\boldsymbol{Y}) = \gamma$ pour tout $\gamma \in \mathbb{R}$, c'est-à-dire $a^\top X\beta = c^\top\beta$ pour tout $\beta \in \mathbb{R}^{p+1}$ (on a utilisé la condition C_p2 et la définition de γ). On doit donc avoir $a^\top X = c^\top$. L'EQM d'un tel estimateur coïncide donc avec sa variance $\mathrm{V}(a^\top\boldsymbol{Y}) = a^\top\mathrm{V}(\boldsymbol{Y})a = \sigma^2 a^\top a$ (on a utilisé la condition C_p3). Si on cherche alors le meilleur estimateur linéaire et sans biais $a^\top\boldsymbol{Y}$ de γ, il faut chercher $a \in \mathbb{R}^n$ satisfaisant $X^\top a = c$ qui minimise $a^\top a$. Formellement, on doit résoudre

$$\min_{a \in \mathbb{R}^n} a^\top a \qquad \text{s.c.q.} \qquad X^\top a = c$$

La fonction à minimiser et chacune des $p+1$ contraintes sont deux fois différentiables, et on peut caractériser les solutions de ce problème au moyen du Lagrangien $\mathcal{L}(a, \lambda) = a^\top a - \lambda^\top (X^\top a - c)$. Comme la fonction à minimiser est convexe et que les contraintes sont affines, pour toute valeur de λ la fonction $\mathcal{L}(\cdot, \lambda)$ est convexe sur $\mathbb{R}^n$. Par conséquent a^* est solution si et seulement si il existe $\lambda^* \in \mathbb{R}^{p+1}$ tel que

$$\frac{\partial \mathcal{L}}{\partial a_i}(a^*, \lambda^*) = 0, \quad i = 1, \ldots, n \qquad \text{et} \qquad \frac{\partial \mathcal{L}}{\partial \lambda_k}(a^*, \lambda^*) = 0, \quad k = 0, \ldots, p$$

On note que $\lambda^\top X^\top a = \sum_{k=0}^{p} \lambda_k \sum_{i=1}^{n} X_{ik} a_i$ et donc en utilisant l'expression de $\mathcal{L}$, ces conditions sont

$$2a_i^* - \sum_{k=0}^{p} \lambda_k^* X_{ik} = 0, \quad i = 1, \ldots, n \qquad \text{et} \qquad \sum_{i=1}^{n} X_{ik} a_i^* - c_k = 0, \quad k = 0, \ldots, p$$

Si on empile les n premières égalités, on peut les écrire sous la forme $2a^* - X\lambda^* = 0_n$ et en empilant les $p+1$ dernières, on a $X^\top a^* - c = 0_{p+1}$. Si on prémultiplie les 2 membres de $2a^* - X\lambda^* = 0_n$ par $(X^\top X)^{-1} X^\top$ on peut écrire $\lambda^* = 2(X^\top X)^{-1} X^\top a^*$. Mais si on utilise le fait que $X^\top a^* - c = 0_{p+1}$, on obtient $\lambda^* = 2(X^\top X)^{-1} c$. Avec l'expression ainsi obtenue pour λ^*, on a $2a^* - X\lambda^* = 0_n \iff a^* = X(X^\top X)^{-1} c$. Ce résultat établit donc que l'estimateur linéaire et sans biais le plus précis de $\gamma = c^\top \beta$ est $a^{*\top} \boldsymbol{Y} = c^\top (X^\top X)^{-1} X^\top \boldsymbol{Y} = c^\top \hat{\beta}$. ℝ

Bien que les théorèmes 5.1 et 5.2 constituent les résultats les plus importants à propos de l'estimation par moindres carrés dans un MRLS, on peut démontrer une propriété intéressante et complémentaire de $\hat{\beta}$. Cette propriété montre que $\hat{\beta}$ peut être obtenu en cherchant la combinaison linéaire des variables exogènes dont les observations sont les plus fortement corrélées avec celles de la variable endogène. Plus formellement, pour n'importe quels réels non-tous nuls $a_0, \ldots, a_p$, on peut introduire la variable, notée X_a, en formant une combinaison linéaire $X_a = a_0 X_0 + \cdots + a_p X_p$ des variables exogènes. Les observations de cette nouvelle variable sont $X_{1a}, \ldots, X_{na}$, avec $X_{ia} = a_0 X_{i0} + \cdots + a_p X_{ip}$, $i = 1, \ldots, n$. On peut alors, comme d'habitude, mesurer la corrélation linéaire empirique entre les variables X_a et Y au moyen du coefficient

$$r(Y, X_a) = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(X_{ia} - \overline{X}_a)}{\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2 \sum_{i=1}^{n} (X_{ia} - \overline{X}_a)^2}}$$

Ce calcul est possible pour n'importe quelle combinaison linéaire X_a. On peut alors chercher celle pour laquelle la corrélation linéaire mesurée par $r(Y, X_a)$ est de plus forte amplitude. Formellement, cela revient à chercher $a_0, \ldots, a_p$ de manière à maximiser $|r(Y, X_a)|$. Si a^* désigne le n-uplet pour lequel $|r(Y, X_{a^*})|$ est maximal, on appelle *coefficient de corrélation linéaire multiple* le nombre $r(Y, X_{a^*})$. C'est la plus forte corrélation linéaire empirique qu'il soit possible d'obtenir entre Y et une combinaison linéaire de $X_0, \ldots, X_p$.

On a le résultat suivant.

pro:mco_maxRa **Propriété 5.5** *Dans le MRLS, les réels $a_0^*, \ldots, a_p^*$ donnés par $a_k^* = \hat{\beta}_k$, $k = 0, \ldots, p$, maximisent la valeur de $r(Y, X_a)^2$.*

La preuve de ce résultat est donnée à la section 5.5.1.4.

Cette propriété apporte une justification plus formelle à la démarche d'estimation de β par laquelle on cherche à donner aux variables exogènes la plus forte capacité à déterminer le niveau de la variable endogène. Cette capacité reflète l'intensité du lien qui existe entre les deux groupes de variables. Dans le contexte du MRLS, on pose que ce lien est linéaire. Par conséquent, l'intensité du lien peut se mesurer par le coefficient de corrélation linéaire, et la propriété 5.5 montre que $\hat{\beta}$ permet de construire la combinaison linéaire de variables exogènes pour laquelle cette intensité est la plus forte.

sec:var

5.4 Valeurs ajustées. Résidus

Comme dans le modèle de régression linéaire à une seule variable exogène, les estimateurs MCO des paramètres permettent ici d'obtenir les valeurs ajustées et les résidus.

def:val_aj_res

Définition 5.3 *Dans le MRLS à p variables où* $\text{rang}(X) = p+1$*, on appelle valeurs ajustées les variables aléatoires constituant les coordonnées du vecteur noté $\hat{\boldsymbol{Y}}$, défini par $\hat{\boldsymbol{Y}} = X\hat{\beta}$. On appelle résidus les variables aléatoires constituant les coordonnées du vecteur aléatoire noté $\hat{\varepsilon}$, défini par $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.*

Les valeurs ajustées et les résidus ont la même interprétation que celle donnée dans le chapitre 2 (section 2.4.1). En particulier, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$ est la partie de Y_i qu'on estime être expliquée par les variables exogènes, alors que $\hat{\varepsilon}_i$ est sa partie complémentaire.

Remarque 5.12 Notons que le vecteur des valeurs ajustées $\hat{\boldsymbol{Y}}$ coïncide avec la projection orthogonale de $\boldsymbol{Y}$ sur l'espace $L(X_{\bullet 0}, \ldots, X_{\bullet p})$ (voir la section 5.3.2). On rappelle que tout vecteur u de $\mathbb{R}^n$ se décompose de manière unique en la somme d'un vecteur u_L de $L(X_{\bullet 0}, \ldots, X_{\bullet p})$ et d'un vecteur $u_{L^\perp}$ de $L(X_{\bullet 0}, \ldots, X_{\bullet p})^\perp$, et que dans cette décomposition, u_L est la projection orthogonale de u sur $L(X_{\bullet 0}, \ldots, X_{\bullet p})$ (voir la section 9.2). Donc, d'après la définition 5.3, dans le cas du vecteur $\boldsymbol{Y} \in \mathbb{R}^n$, cette décomposition est $\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \hat{\varepsilon}$. R

rem:proj_moyenne

Remarque 5.13 En appliquant le point 1 de la propriété 9.25, on obtient facilement l'égalité $\overline{\hat{Y}} = \overline{Y}$ déjà démontrée dans le chapitre 2. Si on désigne par $L(X_{\bullet 0})$ le sev de $\mathbb{R}^n$ engendré par $X_{\bullet 0}$, alors on a évidemment $L(X_{\bullet 0}) \subseteq L(X_{\bullet 0}, \ldots, X_{\bullet p})$. D'après la remarque 9.9, la projection orthogonale de $\hat{\boldsymbol{Y}}$ sur $L(X_{\bullet 0})$ est $\overline{\hat{Y}} X_{\bullet 0}$, et celle de $\boldsymbol{Y}$ sur ce même espace est $\overline{Y} X_{\bullet 0}$. Mais d'après le point 1 de la propriété 9.25, ces deux projections coïncident, et on doit donc avoir $\overline{\hat{Y}} = \overline{Y}$. R

rem:orth_epsilon

Remarque 5.14 Puisque $\hat{\boldsymbol{Y}}$ est la projection orthogonale de $\boldsymbol{Y}$ sur $L(X_{\bullet 0}, \ldots, X_{\bullet p})$ et qu'on a $\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \hat{\varepsilon}$, le vecteur des résidus est donc dans l'espace $L(X_{\bullet 0}, \ldots, X_{\bullet p})^\perp$. Comme tout vecteur de cet espace est orthogonal à n'importe quel élément de $L(X_{\bullet 0}, \ldots, X_{\bullet p})$, on a la propriété suivante. R

Propriété 5.6 *Le vecteur des résidus $\hat{\varepsilon}$ satisfait l'égalité $X^\top \hat{\varepsilon} = 0_{p+1}$.*

Preuve : Par définition de la matrice X (voir page 89), le $(k+1)^{\text{e}}$ élément du vecteur $X^\top\hat{\varepsilon}$ est $X_{\bullet k}^\top\hat{\varepsilon}$. D'après la remarque 5.14, $\hat{\varepsilon} \in L(X_{\bullet 0},\dots,X_{\bullet p})^\perp$ et donc $\hat{\varepsilon}$ est en particulier orthogonal à $X_{\bullet k}^\top$, c'est à dire $X_{\bullet k}^\top\hat{\varepsilon} = 0$. Ceci est vrai pour tout $k = 0,\dots,p$.

On peut également vérifier ce résultat par calcul. On a

$$X^\top\hat{\varepsilon} = X^\top(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = X^\top\boldsymbol{Y} - X^\top X\hat{\beta} = X^\top\boldsymbol{Y} - X^\top X(X^\top X)^{-1}X^\top\boldsymbol{Y} = 0_{p+1}$$

Finalement, on peut obtenir cette égalité en notant qu'elle correspond à la condition nécessaire (5.5) dans la minimisation de S.

Cette propriété est l'équivalent de la propriété 2.5 du chapitre 2. En particulier, comme la première ligne de $X^\top$ est $X_{\bullet 0}^\top = (1,\dots,1)$, on doit avoir $\sum_{i=1}^n \hat{\varepsilon}_i = 0$.

Le théorème 2.5 démontré dans le chapitre 2 reste valable à l'identique. Cependant, la technique de la preuve est un peu différente. Il peut être pratique de bien noter que, dans la preuve qui va être donnée (et dans d'autres à suivre), un vecteur de $\mathbb{R}^n$ dont toutes les coordonnées sont égales peut toujours s'écrire $cX_{\bullet 0}$ où $c \in \mathbb{R}$ est la valeur commune des coordonnées.

th:R2multiv **Théorème 5.3** *Dans le MRLS on a*

$$\sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i{}^2 \tag{5.11}$$ eq:decomp_reg

où $\overline{Y}$ désigne la moyenne de $Y_1,\dots,Y_n$.

Preuve : On forme le vecteur $\boldsymbol{Y} - \overline{Y}X_{\bullet 0}$ de $\mathbb{R}^n$ dont la i^{e} coordonnée est $Y_i - \overline{Y}$. On constate que le membre de gauche de l'égalité du théorème est le carré de la norme de ce vecteur. On a par ailleurs

$$\boldsymbol{Y} - \overline{Y}X_{\bullet 0} = (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) + (\hat{\boldsymbol{Y}} - \overline{Y}X_{\bullet 0}) = \hat{\varepsilon} + (\hat{\boldsymbol{Y}} - \overline{Y}X_{\bullet 0})$$

Notons que puisque $\hat{\boldsymbol{Y}}$ et $X_{\bullet 0}$ sont deux vecteurs de $L(X_{\bullet 0},\dots,X_{\bullet p})$, le vecteur $\hat{\boldsymbol{Y}} - \overline{Y}X_{\bullet 0}$ est également dans $L(X_{\bullet 0},\dots,X_{\bullet p})$. Il est donc orthogonal à $\hat{\varepsilon} \in L(X_{\bullet 0},\dots,X_{\bullet p})^\perp$. Le théorème de Pythagore donne alors

$$\|\boldsymbol{Y} - \overline{Y}X_{\bullet 0}\|^2 = \|\hat{\varepsilon}\|^2 + \|\hat{\boldsymbol{Y}} - \overline{Y}X_{\bullet 0}\|^2 \tag{5.12}$$ eq:decomp_var

ce qui est l'égalité (5.11).

L'animation de la figure 5.3 illustre la propriété d'orthogonalité entre $\hat{\varepsilon}$ et $\hat{\boldsymbol{Y}} - \overline{Y}X_{\bullet 0}$ utilisée dans la preuve ci-dessus.

Toute l'interprétation de la relation (5.11) qui a été faite dans le chapitre 2 reste entièrement valable ici. Notamment, l'égalité (5.11) permet de décomposer la mesure de la variabilité observée de la variable endogène (le membre de gauche de cette égalité) en la somme d'une partie qui est l'estimation de la variabilité due à celle des variables exogènes du modèle, et d'une partie qui est l'estimation de la variabilité due à des facteurs autres que les variables exogènes. Cette interprétation permet de définir le coefficient de détermination de la régression, noté R^2, de manière identique à la définition 2.6. On a donc

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = \frac{\|\hat{\boldsymbol{Y}} - \overline{Y}X_{\bullet 0}\|^2}{\|\boldsymbol{Y} - \overline{Y}X_{\bullet 0}\|^2} \tag{5.13}$$ eq:def_R2_mult

L'interprétation de ce coefficient est la même que celle donnée dans la remarque 2.9, et la propriété 2.6 devient la suivante.

pro:R2_mrls **Propriété 5.7** *Dans le MRLS avec* $\text{rang}(X) = p+1$*, on a*

1. $R^2 = 1 \iff \boldsymbol{Y} \in L(X_{\cdot 0}, \ldots, X_{\cdot p})$
2. $R^2 = 0 \iff \hat{\beta}_1 = \cdots = \hat{\beta}_p = 0$

Preuve :

1. $\boldsymbol{Y} \in L(X_{\cdot 0}, \ldots, X_{\cdot p}) \iff \hat{\boldsymbol{Y}} = \boldsymbol{Y} \iff R^2 = 1$, où la première équivalence vient du fait que $\hat{\boldsymbol{Y}}$ est la projection orthogonale de $\boldsymbol{Y}$ sur $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ et du point 2 de la propriété 9.22
2. Notons que $\hat{\beta}_1 = \cdots = \hat{\beta}_p = 0 \iff \hat{\boldsymbol{Y}} = X_{\cdot 0}\hat{\beta}_0$. En effet, l'implication dans un sens est évidente. Réciproquement, supposons $\hat{\boldsymbol{Y}} = X_{\cdot 0}\hat{\beta}_0$. On a donc $\text{proj}_{L(X_{\cdot 0},\ldots,X_{\cdot p})}(\boldsymbol{Y}) \in L(X_{\cdot 0})$. Comme $L(X_{\cdot 0}) \subset L(X_{\cdot 0}, \ldots, X_{\cdot p})$, le point 2 de la propriété 9.25 permet de conclure que $\text{proj}_{L(X_{\cdot 0},\ldots,X_{\cdot p})}(\boldsymbol{Y}) = \text{proj}_{L(X_{\cdot 0})}(\boldsymbol{Y})$, c'est à dire $\hat{\boldsymbol{Y}} = X_{\cdot 0}\overline{Y}$ (voir la remarque 5.13) et donc le numérateur de R^2 est nul. ■

rem:pro_R2_multiv **Remarque 5.15**

— On note que la première équivalence s'écrit $R^2 = 0 \iff \exists \beta^* \in \mathbb{R}^{p+1}$ t.q. $\boldsymbol{Y} = X\beta^*$. L'interprétation de cette équivalence est la même que celle donnée à la remarque 2.10 (point 1). En particulier, le β^* qui permet d'écrire Y_i comme une combinaison linéaire de $X_{i1}, \ldots, X_{ip}$ pour tout i est $\hat{\beta}$.

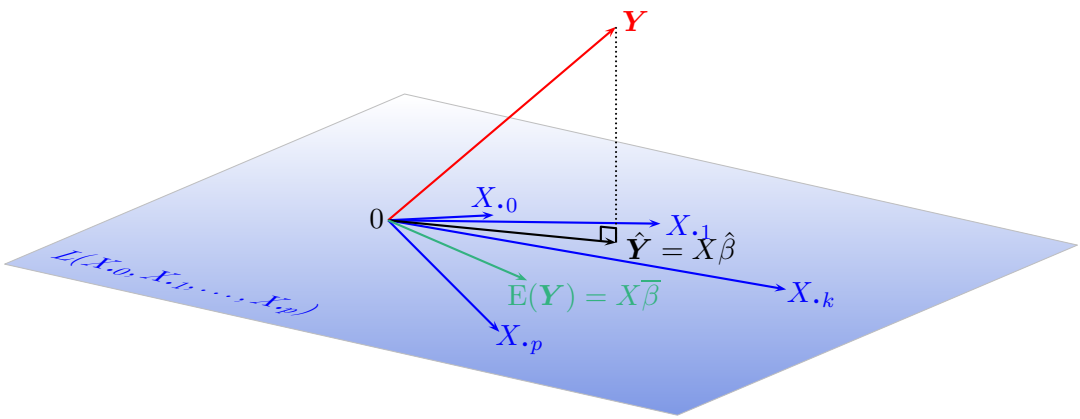
— Remarquons que si $\hat{\beta}_1 = \cdots = \hat{\beta}_p = 0$, on a $\hat{\boldsymbol{Y}} = X_{\cdot 0}\overline{Y}$. Comme on a toujours $\hat{\boldsymbol{Y}} = X_{\cdot 0}\hat{\beta}_0 + X_{\cdot 1}\hat{\beta}_p + \cdots + X_{\cdot p}\hat{\beta}_p$, et puisque la projection orthogonale sur $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ est unique[9], on doit avoir $\hat{\beta} = (\overline{Y}, 0, \ldots, 0)^\top$. En résumé, on a donc les équivalences suivantes

$$R^2 = 0 \iff \hat{\boldsymbol{Y}} \in L(X_{\cdot 0}) \iff \hat{\boldsymbol{Y}} = \overline{Y}X_{\cdot 0} \iff \hat{\beta} = (\overline{Y}, 0, \ldots, 0)^\top$$

Ici également, l'interprétation de l'équivalence dans le cas $R^2 = 0$ faite à la remarque 2.10 (point 2) reste valable. R

Les deux points de la propriété **??** et ceux de la remarque 5.15 sont illustrés par les figures 5.3 à 5.5.

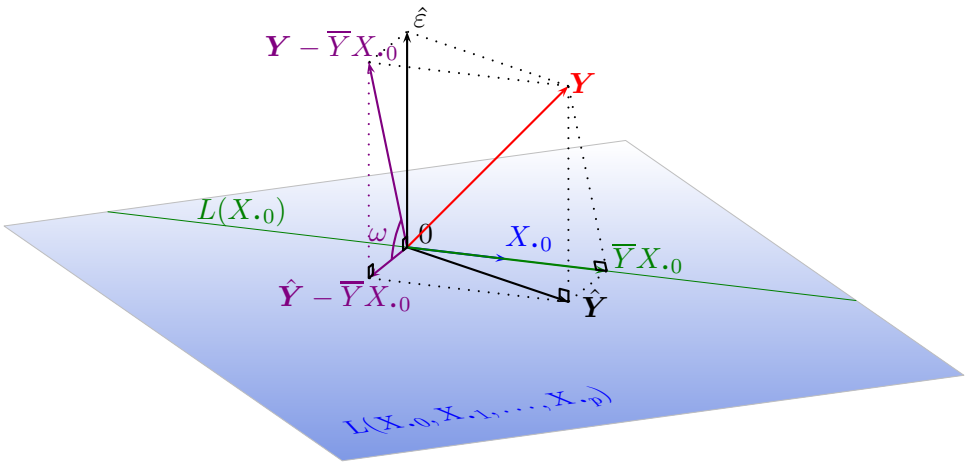
9. Ou parce que $X_{\cdot 0}, X_{\cdot 1}, \ldots, X_{\cdot p}$ sont linéairement indépendants.



Cette animation illustre les propriétés géométriques de l'estimateur MCO de β dans le MRLS. *Cliquez pour lancer l'animation.*

FIGURE 5.2: Interprétation géométrique de l'estimateur MCO de β

fig:mrls_mco



Cette animation illustre les propriétés du coefficient de détermination R^2. Cliquez pour lancer l'animation.

fig:mrls_R2

FIGURE 5.3: Illustration géométrique de la construction du coefficient de détermination R^2

Les figures 5.4 et 5.5 présentent des animations qui illustrent la propriété 5.7. Dans ces deux animations, les objets représentés sont identiques à ceux de la figure **??** (seule la perspective est différente). Les vecteurs des variables exogènes $X_{\bullet 0}, \ldots, X_{\bullet p}$ restent inchangés (et donc le plan $L(X_{\bullet 0}, \ldots, X_{\bullet p})$ aussi), et l'animation est générée par le seul déplacement de $\boldsymbol{Y}$. Ce mouvement génère évidemment un mouvement des vecteurs $\boldsymbol{Y} - \overline{Y}X_{\bullet 0}$ et $\hat{\boldsymbol{Y}} - \overline{Y}X_{\bullet 0}$, et donc une variation de l'angle qui les relie.

Sur la figure 5.4, on illustre le point 2 de la propriété 5.7. Le mouvement de $\boldsymbol{Y}$ est choisi de sorte que $\hat{\boldsymbol{Y}}$ se rapproche de $L(X_{\bullet 0})$ (le vecteur $\boldsymbol{Y}$ pivote autour de son origine, son extrémité décrivant un cercle parallèle au plan $L(X_{\bullet 0}, \ldots, X_{\bullet p})$). L'angle ω se rapproche de l'angle droit, et lorsque $\hat{\boldsymbol{Y}} \in L(X_{\bullet 0})$ (dernière image), on a $\omega = 90°$. Dans ce cas, $R^2 = 0$.

L'animation de la figure 5.5 illustre le premier point de la propriété 5.7. Le mouvement s'obtient en faisant se rapprocher $\boldsymbol{Y}$ de $L(X_{\bullet 0}, \ldots, X_{\bullet p})$. On voit alors que le vecteur $\boldsymbol{Y} - \overline{Y}X_{\bullet 0}$ se rapproche du plan, et que par conséquent l'angle qu'il forme avec ce plan, et donc *a fortiori* l'angle ω formé avec le vecteur $\hat{\boldsymbol{Y}} - \overline{Y}X_{\bullet 0}$, se rapproche de 0. Lorsque $\boldsymbol{Y} - \overline{Y}X_{\bullet 0} \in L(X_{\bullet 0})$ (dernière image), on a $\omega = 0°$. Par conséquent, dans ce cas $R^2 = 1$.

5.5 Compléments sur l'estimation de β

sec:compl_mrls

On présente dans cette section deux résultats complémentaires importants sur l'estimation de β par moindres carrés.

5.5.1 Le théorème de Frisch-Waugh

sec:FW

5.5.1.1 Motivation du résultat : MCO avec variables exogènes orthogonales

Le résultat de cette section intervient notamment lorsqu'on distingue deux groupes de variables parmi les variables exogènes du modèle, et que l'estimation des paramètres attachés aux variables d'un seul des deux groupes est privilégiée. Quitte à renuméroter les variables exogènes, on peut supposer que les deux groupes sont constitués des q premières et $p + 1 - q$ dernières variables explicatives, respectivement, et qu'on s'intéresse à l'estimation des paramètres attachés au groupe des q premières variables.

De manière à faire apparaître cette séparation, on note X_1 la matrice constituée des q premières colonnes de X et X_2 la matrice constituée des $p + 1 - q$ dernières colonnes de X, de sorte que

$$X = \left(\begin{array}{c:c} X_1 & X_2 \end{array} \right)$$

Si on effectue le partitionnement correspondant pour β on a

$$\beta = \begin{pmatrix} \beta^1 \\ \hdashline \beta^2 \end{pmatrix}$$

et on peut alors écrire $X\beta = X_1\beta^1 + X_2\beta^2$, où β^1 est le vecteur de $\mathbb{R}^q$ dont les coordonnées sont les paramètres associés aux variables dans X_1 est β^2 est le vecteur de $\mathbb{R}^{p+1-q}$ regroupant les autres

sec:FW

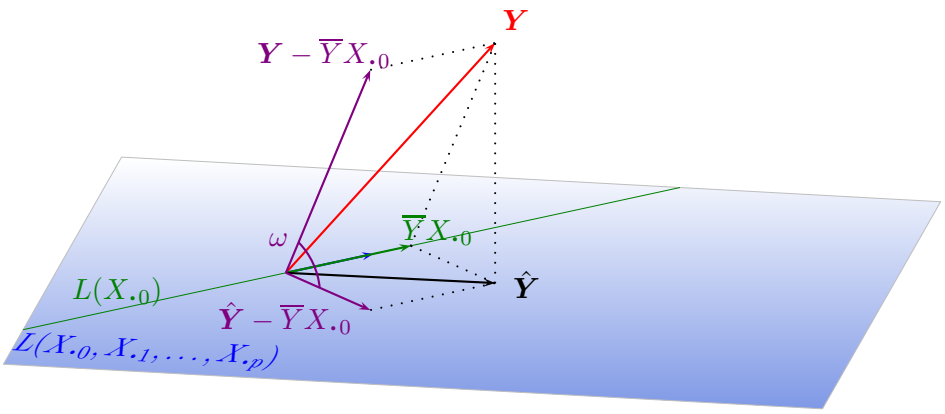


fig:R2lat

FIGURE 5.4: Interprétation graphique de la valeur de $R^2 = \cos(\omega)^2$ (R^2 tend vers 0)

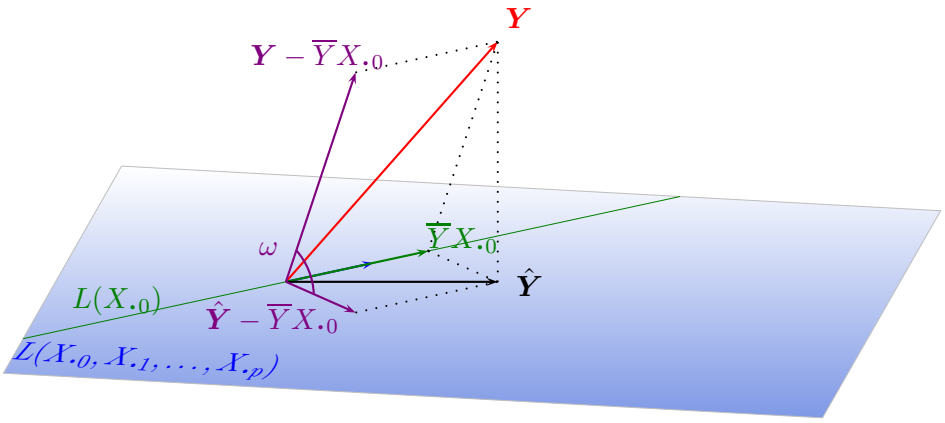


fig:R2long

FIGURE 5.5: Interprétation graphique de la valeur de $R^2 = \cos(\omega)^2$ (R^2 tend vers 1)

paramètres. La relation du modèle s'écrit donc

$$\boldsymbol{Y} = X_1\beta^1 + X_2\beta^2 + \varepsilon$$

Pour motiver le résultat à suivre, on considère d'abord le cas où chacune des variables du second groupe est orthogonale à chaque variable du premier groupe, dans le sens où

$$X_1^\top X_2 = 0$$

le 0 du membre de droite étant une matrice de dimensions $q \times (p+1-q)$ dont toutes les entrées sont nulles. Avec une telle décomposition, l'estimateur des MCO de β est

$$\begin{aligned}\hat{\beta} = \left(\begin{array}{c}\hat{\beta}^1 \\ \hdashline \hat{\beta}^2\end{array}\right) = (X^\top X)^{-1}X^\top \boldsymbol{Y} &= \left(\left(\begin{array}{c}X_1^\top \\ \hdashline X_2^\top\end{array}\right)\left(\begin{array}{c:c}X_1 & X_2\end{array}\right)\right)^{-1}\left(\begin{array}{c}X_1^\top \\ \hdashline X_2^\top\end{array}\right)\boldsymbol{Y} \\ &= \left(\begin{array}{c:c}X_1^\top X_1 & X_1^\top X_2 \\ \hdashline X_2^\top X_1 & X_2^\top X_2\end{array}\right)^{-1}\left(\begin{array}{c}X_1^\top \boldsymbol{Y} \\ \hdashline X_2^\top \boldsymbol{Y}\end{array}\right)\end{aligned}$$

Par orthogonalité des deux groupes de variables, les blocs anti-diagonaux de $X^\top X$ sont nuls et

$$\left(\begin{array}{c:c}X_1^\top X_1 & X_1^\top X_2 \\ \hdashline X_2^\top X_1 & X_2^\top X_2\end{array}\right)^{-1} = \left(\begin{array}{c:c}X_1^\top X_1 & 0 \\ \hdashline 0 & X_2^\top X_2\end{array}\right)^{-1} = \left(\begin{array}{c:c}(X_1^\top X_1)^{-1} & 0 \\ \hdashline 0 & (X_2^\top X_2)^{-1}\end{array}\right)$$

Par conséquent,

$$\left(\begin{array}{c}\hat{\beta}^1 \\ \hdashline \hat{\beta}^2\end{array}\right) = \left(\begin{array}{c}(X_1^\top X_1)^{-1}X_1^\top \boldsymbol{Y} \\ \hdashline (X_2^\top X_2)^{-1}X_2^\top \boldsymbol{Y}\end{array}\right)$$

Ceci montre alors que si on s'intéresse seulement à l'estimation de β^1, on peut faire comme si le modèle ne contenait pas les variables du second groupe et appliquer les MCO à un modèle dans lequel on aurait $\boldsymbol{Y} = X_1\beta^1 + u$, même si toutes les variables sont présentes dans le modèle de départ. On obtient ainsi $\hat{\beta}^1 = (X_1^\top X_1)^{-1}X_1^\top \boldsymbol{Y}$.

Le résultat qui sera présenté à la section suivante montre que même si on n'a pas orthogonalité entre les variables des deux groupes, on peut s'y ramener. Plus précisément, on peut effectuer une transformation sur les variables du second groupe, de sorte que :

1. chaque nouvelle variable du second groupe est orthogonale à chaque variable du premier groupe ;
2. ensemble, ces nouvelles variables et celles du premier groupe engendrent le même espace que toutes variables initiales, c'est à dire $L(X_{\bullet 0}, \ldots, X_{\bullet p})$.

Ce résultat permet donc d'estimer les paramètres liés aux variables auxquelles on s'intéresse, sans avoir à estimer les paramètres des autres variables (grâce à l'orthogonalité entre groupes de variables), et sans pour autant perdre d'information par rapport à ce qu'on obtiendrait si on estimait entièrement le modèle initial (puisque la transformation permet de « rester » dans $L(X_{\bullet 0}, \ldots, X_{\bullet p})$).

sec:thFW

5.5.1.2 Le résultat

On suppose qu'on s'intéresse à q variables exogènes, et quitte à renuméroter ces variables, on peut toujours supposer que ce sont les q premières. On a donc les partitionnements de X et de β donnés dans la section précédente.

Le résultat qui sera obtenu dans cette section repose sur la propriété 9.26 (point 3). On utilisera une décomposition de $L(X_{\bullet 0}, \ldots, X_{\bullet p})$ et afin d'alléger les expressions, on introduit la notation suivante

$$L = L(X_{\bullet 0}, \ldots, X_{\bullet p}) \qquad L_1 = L(X_1) \qquad L_2 = L(X_2)$$

où $L(X_i)$ est un raccourci pour désigner l'ev engendré par les colonnes de X_i ; par exemple $L_1 = L(X_1) = L(X_{\bullet 0}, \ldots, X_{\bullet q})$. On notera de manière naturelle P_L, P_{L_1} et P_{L_2} les matrices de projection orthogonale sur L, L_1 et L_2, respectivement. Puisqu'on travaille toujours sous l'hypothèse que X est de rang $p+1$, X_1 et X_2 sont des matrices de rang q et $p+1-q$, respectivement. Par conséquent, les vecteurs colonnes de X_i forment une base de L_i et on aura par exemple $P_{L_2} = X_2(X_2^\top X_2)^{-1}X_2^\top$ (application de la propriété 9.22, point 4).

Avec de telles notations, on constate évidemment que $L = L_1 + L_2$. Si on définit $\tilde{L}_1 = \{U \in L | U = (I - P_{L_2})V, V \in L_1\}$, l'ensemble obtenu en formant les restes de la projection orthogonale des éléments de L_1 sur L_2, alors le raisonnement de la remarque 9.11 permet ici de déduire que :

1. $\tilde{L}_1$ est un sev de $\mathbb{R}^n$ engendré par les q vecteurs linéairement indépendants constituant les colonnes de $\tilde{X}_1 = (I - P_{L_2})X_1$;
2. $L = L_2 + \tilde{L}_1$ avec L_2 et $\tilde{L}_1$ orthogonaux.

Par conséquent, puisque $P_L = P_{L_1+L_2}$,[10] le point 3 de la propriété 9.26 établit que[11]

$$P_L = P_{L_2} + P_{\tilde{L}_1}$$

Ceci permet d'écrire $P_L\boldsymbol{Y}$ de deux manières

$$P_L\boldsymbol{Y} = P_{L_1+L_2}\boldsymbol{Y} = P_{L_2}\boldsymbol{Y} + P_{\tilde{L}_1}\boldsymbol{Y}$$

Or d'une part

$$P_L\boldsymbol{Y} = P_{L_1+L_2}\boldsymbol{Y} = X\hat{\beta} = X_1\hat{\beta}^1 + X_2\hat{\beta}^2 \tag{5.14}$$

eq:fw1

et d'autre part

$$P_L\boldsymbol{Y} = P_{\tilde{L}_1}\boldsymbol{Y} + P_{L_2}\boldsymbol{Y} = \tilde{X}_1\tilde{\beta}^1 + X_2\tilde{\beta}^2 \tag{5.15}$$

eq:fw2

où $\hat{\beta}^1 \in \mathbb{R}^q$ et $\hat{\beta}^2 \in \mathbb{R}^{p+1-q}$ sont les deux sous-vecteurs de $\hat{\beta}$ contenant les estimateurs MCO des deux groupes de variables X_1 et X_2, i.e.

$$\hat{\beta} = (X^\top X)^{-1}X^\top\boldsymbol{Y} = \begin{pmatrix} \hat{\beta}^1 \\ \hdashline \hat{\beta}^2 \end{pmatrix}$$

et $\tilde{\beta}^1$ et $\tilde{\beta}^2$ sont les coordonnées de la projection de $\boldsymbol{Y}$ sur $\tilde{L}_1$ et L_2, respectivement. On a donc

$$\tilde{\beta}^1 = (\tilde{X}_1^\top\tilde{X}_1)^{-1}\tilde{X}_1^\top\boldsymbol{Y} \qquad \text{et} \qquad \tilde{\beta}^2 = (X_2^\top X_2)^{-1}X_2^\top\boldsymbol{Y}$$

On est en mesure de démontrer le résultat suivant.

10. Attention, on n'a en général pas $P_L = P_{L_1} + P_{L_2}$ ce qui précisément motive le résultat qui va suivre.
11. Il suffit d'appliquer la propriété avec $F = L_2$ et $H = L_1$.

th:FW **Théorème 5.4 (Frisch-Waugh)**

1. $\tilde{\beta}^1 = \hat{\beta}^1$
2. $M_{L_2}\boldsymbol{Y} - \tilde{X}_1\tilde{\beta}^1 = \hat{\varepsilon}$, *où* $M_{L_2} = I - P_{L_2}$

Preuve :

1. D'après (5.14) et (5.15), on a $\tilde{X}_1\tilde{\beta}^1 + X_2\tilde{\beta}^2 = X_1\hat{\beta}^1 + X_2\hat{\beta}^2$. L'égalité reste vraie en prémultipliant chacun de ses membres par $M_{L_2} = (I - P_{L_2})$. Or $M_{L_2}X_2 = 0$ et on obtient donc
$$M_{L_2}\tilde{X}_1\tilde{\beta}^1 = M_{L_2}X_1\hat{\beta}^1$$
Mais par construction $M_{L_2}X_1 = \tilde{X}_1$, et par idempotence de M_{L_2} on a $M_{L_2}\tilde{X}_1 = \tilde{X}_1$. L'égalité devient donc
$$\tilde{X}_1\tilde{\beta}^1 = \tilde{X}_1\hat{\beta}^1$$
Comme les colonnes de $\tilde{X}_1$ sont linéairement indépendantes (voir ci-dessus), on a le résultat voulu.
2. Comme $\boldsymbol{Y} = P_L\boldsymbol{Y} + \hat{\varepsilon}$, on peut écrire en utilisant l'expression de $P_L\boldsymbol{Y}$ donnée par (5.15) :
$$\boldsymbol{Y} = \tilde{X}_1\tilde{\beta}^1 + X_2\tilde{\beta}^2 + \hat{\varepsilon}$$
En prémultipliant les membres de l'égalité par M_{L_2}, on a
$$M_{L_2}\boldsymbol{Y} = \tilde{X}_1\tilde{\beta}^1 + M_{L_2}\hat{\varepsilon}$$
(tout comme dans le point précédent, on a utilisé $M_{L_2}\tilde{X}_1 = \tilde{X}_1$ et $M_{L_2}X_2 = 0$). Or $M_{L_2}\hat{\varepsilon} = \hat{\varepsilon} - P_{L_2}\hat{\varepsilon} = \hat{\varepsilon}$, puisque $L_2 \subset L$ et que $\hat{\varepsilon} \in L^\perp$. Donc $M_{L_2}\boldsymbol{Y} = \tilde{X}_1\tilde{\beta}^1 + \hat{\varepsilon}$, ce qui est le résultat recherché.

rem:fwequivmco **Remarque 5.16** Le premier point de ce résultat montre qu'on peut obtenir $\hat{\beta}^1$ en calculant $\tilde{\beta}^1 = (\tilde{X}_1^\top\tilde{X}_1)^{-1}\tilde{X}_1^\top\boldsymbol{Y}$, avec $\tilde{X}_1 = M_{L_2}X_1$. En remarquant que M_{L_2} est symétrique et idempotente, on peut également écrire $\tilde{\beta}^1 = (\tilde{X}_1^\top\tilde{X}_1)^{-1}\tilde{X}_1^\top\tilde{\boldsymbol{Y}}$ où $\tilde{\boldsymbol{Y}} = M_{L_2}\boldsymbol{Y}$. Cette dernière expression de $\tilde{\beta}^1$ permet de le voir comme l'estimateur MCO dans un modèle où la relation est $\tilde{\boldsymbol{Y}} = \tilde{X}_1\beta^1 + u$. De plus le second point du résultat peut s'écrire $\tilde{Y} - \tilde{X}_1\tilde{\beta}^1 = \hat{\varepsilon}$. Autrement dit, si on estime β^1 par MCO dans ce modèle, les résidus de cette estimation coïncident avec les résidus qu'on aurait obtenus en estimant β tout entier.

On a donc atteint l'objectif fixé initialement : si on applique les MCO à la relation $\tilde{\boldsymbol{Y}} = \tilde{X}_1\beta^1 + u$, on obtient $\hat{\beta}^1$ sans avoir à estimer $\hat{\beta}^2$. De plus, cette estimation se fait sans perdre l'information sur $\hat{\boldsymbol{Y}} = X_1\hat{\beta}^1 + X_2\hat{\beta}^2$, puisqu'elle donne les mêmes résidus que si on avait estimé β^1 et β^2. Ce dernier point est important puisqu'il permet notamment, sans avoir à estimer le modèle complet, d'obtenir l'estimateur de σ^2 (voir la section 5.6). R

Remarque 5.17 Il faut bien noter cependant que la relation $\tilde{\boldsymbol{Y}} = \tilde{X}_1\beta^1 + u$ n'est pas celle d'un modèle de régression linéaire standard. Pour le voir, il suffit de noter que
$$u = \tilde{\boldsymbol{Y}} - \tilde{X}_1\beta^1 = M_{L_2}(\boldsymbol{Y} - X_1\beta^1) = M_{L_2}(X_2\beta^2 + \varepsilon) = M_{L_2}\varepsilon$$
où la dernière égalité provient de $M_{L_2}X_2 = 0$. On peut donc calculer
$$\mathrm{V}(u) = \mathrm{V}(M_{L_2}\varepsilon) = M_{L_2}\mathrm{V}(\varepsilon)M_{L_2}^\top = \sigma^2 M_{L_2}$$

où la dernière égalité s'obtient par la symétrie et l'idempotence de M_{L_2} et par le fait que ε satisfait la condition C'_p3. On constate donc que u ne satisfait une telle condition. ℝ

rem:mcoFW **Remarque 5.18** En pratique, il s'agit de former $\tilde{X}_1$ et $\tilde{\boldsymbol{Y}}$ dans une première étape, puis dans une seconde, utiliser $\tilde{\boldsymbol{Y}}$ comme vecteur des observations de la « variable endogène » et $\tilde{X}_1$ comme matrice des observations des « variables exogènes » pour calculer un estimateur MCO de β^1. Par définition $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - P_{L_2}\boldsymbol{Y}$. On a souligné à la remarque 5.9 que toute projection orthogonale peut être vue comme une estimation par moindres carrés et réciproquement. Par conséquent, $P_{L_2}\boldsymbol{Y}$ peut s'obtenir par l'estimation MCO de relation $\boldsymbol{Y} = X_2\delta + \nu$ et $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - P_{L_2}\boldsymbol{Y}$ apparaît donc comme le vecteur des « résidus » de cette estimation. Il est en de même pour chaque colonne de $\tilde{X}_1$. Pour la dernière, par exemple, on aura $\tilde{X}_{\bullet q} = X_{\bullet q} - P_{L_2}X_{\bullet q}$, ce qui permet de l'obtenir comme le résidu de l'estimation MCO de la relation $X_{\bullet q} = X_2\delta_q + \nu_q$. ℝ

Remarque 5.19 On peut donner une interprétation intéressante du premier point du théorème 5.4. Bien que la remarque s'applique à des contextes plus généraux, on considère le cas dans lequel X_1 ne contient qu'une seule variable (différente de $X_{\bullet 0}$) et X_2 contient les p variables restantes. Le vecteur β^1 dans ce cas n'a qu'une seule coordonnée.

On rappelle que β^1 permet de mesurer les effets sur la variable Y de variations de la variable exogène X_1, *toutes choses égales par ailleurs*. Ce raisonnement toutes choses égales par ailleurs consiste à exclure dans la réponse de Y à des variations de X_1 des effets indirects qui seraient liés à des co-variations de variables explicatives. Plus précisément, si X_1 varie, alors l'effet de cette variation sur Y est constitué d'un effet direct et d'un effet indirect. L'effet direct est capturé par le coefficient attaché à la variable exogène X_1. L'effet indirect est produit par le fait qu'en faisant varier X_1, on provoque éventuellement des variations d'autres variables exogènes, qui elles-mêmes provoquent une variation de Y. En raisonnant toutes choses égales par ailleurs, on ne prend en compte que les effets directs. Ceux-ci sont mesurés par β^1 et l'estimation de ces effets est $\hat{\beta}^1$.

Le théorème 5.4 établit que $\hat{\beta}^1 = \tilde{\beta}^1$, avec $\tilde{\beta}^1 = (\tilde{X}_1^\top\tilde{X}_1)^{-1}\tilde{X}_1^\top\boldsymbol{Y}$. On constate que $\tilde{\beta}^1$ est l'estimation par moindres carrés du paramètre de la relation $\boldsymbol{Y} = \tilde{X}_1\beta^1 + u$. C'est donc l'estimation de l'effet de la variable $\tilde{X}_1$ sur la variable $\boldsymbol{Y}$. Or $\tilde{X}_1 = (I - P_{L_2})X_1$ et est par construction orthogonal au vecteur des observations de n'importe quelle variable incluse dans X_2. Autrement dit, $\tilde{X}_1$ est orthogonal à l'espace L_2. Cette propriété d'orthogonalité correspond à l'absence de lien (linéaire) entre $\tilde{X}_1$ et les variables de X_2. Pour interpréter l'orthogonalité de cette manière, il faut simplement noter que pour tout vecteur v de $\mathbb{R}^n$, on peut toujours écrire $v = v_{L_2} + v_{L_2^\perp}$, où v_{L_2} est la projection orthogonale de v sur L_2. Cette projection est une combinaison linéaire des vecteurs qui engendrent L_2, c'est à dire les colonnes de X_2. On peut donc interpréter v_{L_2} comme la partie de v qui peut s'écrire comme une combinaison linéaire des variables de X_2. Si v est orthogonal à chacune de ces variables, alors $v \in L_2^\perp$ et $v_{L_2} = 0_n$. Dans ce cas, on voit que v ne contient aucune partie qui peut s'exprimer linéairement en fonction des variables de X_2.

Si on revient au modèle de régression en appliquant cette interprétation de l'orthogonalité entre $\tilde{X}_1$ et X_2, on voit que $\tilde{X}_1$ ne peut co-varier (linéairement) avec X_2. Par conséquent, si on fait varier $\tilde{X}_1$, cette variation ne peut avoir d'effet indirect sur la variable Y, via les variations des variables dans X_2 que pourrait provoquer la variation de $\tilde{X}_1$. Donc, dans la relation $\boldsymbol{Y} = \tilde{X}_1\beta^1 + u$, on peut véritablement interpréter β^1 comme la variation de Y provoquée par la variation de $\tilde{X}_1$, sans avoir à

préciser que toutes choses sont égales par ailleurs. Par conséquent, la variation de Y provoquée par une variation de X_1 toutes choses égales par ailleurs, est identique à la variation de Y provoquée par une variation de $\tilde{X}_1$. Les estimations de ces variations sont $\hat{\beta}^1$ d'une part et $\tilde{\beta}^1$ d'autre part. Le théorème 5.4 nous dit qu'elles coïncident.

$\mathbb{R}$

Remarque 5.20 La preuve du point 1 du théorème 5.4 s'obtient également par calcul. Ainsi, en utilisant la partition de X et de β, on peut écrire (5.6) sous la forme

$$\left(\begin{array}{c} X_1^\top \\ \hdashline X_2^\top \end{array}\right)\left(\begin{array}{c:c} X_1 & X_2 \end{array}\right)\left(\begin{array}{c} \hat{\beta}^1 \\ \hdashline \hat{\beta}^2 \end{array}\right) = \left(\begin{array}{c} X_1^\top \\ \hdashline X_2^\top \end{array}\right)\boldsymbol{Y}$$

En effectuant les produits de matrices par blocs, on a

$$\left\{\begin{array}{l} X_1^\top X_1\hat{\beta}^1 + X_1^\top X_2\hat{\beta}^2 = X_1^\top \boldsymbol{Y} \\ X_2^\top X_1\hat{\beta}^1 + X_2^\top X_2\hat{\beta}^2 = X_2^\top \boldsymbol{Y} \end{array}\right.$$

Si à partir de la deuxième égalité on exprime $\hat{\beta}^2$ en fonction de $\hat{\beta}^1$ on obtient

$$\hat{\beta}^2 = (X_2^\top X_2)^{-1}X_2^\top(\boldsymbol{Y} - X_1\hat{\beta}^1)$$

En utilisant cette expression dans la première égalité, et en réarrangeant les termes, on a

$$X_1^\top(I - P_{L_2})X_1\hat{\beta}^1 = X_1^\top(I - P_{L_2})\boldsymbol{Y}$$

Comme $M_{L_2} = I - P_{L_2}$ est une matrice symétrique et idempotente, l'égalité qui vient d'être obtenue s'écrit aussi

$$(M_{L_2}X_1)^\top M_{L_2}X_1\hat{\beta}^1 = (M_{L_2}X_1)^\top M_{L_2}\boldsymbol{Y}$$

En rappelant que $\tilde{X}_1 = M_{L_2}X_1$ et $\tilde{\boldsymbol{Y}} = M_{L_2}\boldsymbol{Y}$, on peut également écrire

$$\tilde{X}_1^\top \tilde{X}_1\hat{\beta}^1 = \tilde{X}_1^\top \tilde{\boldsymbol{Y}} \tag{5.16}$$ eq:beta1_FW

En prémultipliant les deux membres de cette égalité par $(\tilde{X}_1^\top \tilde{X}_1)^{-1}$, on a bien l'égalité du point premier point du théorème. $\mathbb{R}$

On termine cette section en proposant une représentation graphique du théorème. Celle-ci est réalisée sous forme d'animation (figure (5.6)) [12]

sec:FW_appli

5.5.1.3 Une application

Une application intéressante du théorème de Wrisch-Waugh, aussi bien d'un point de vue théorique que pratique, permet d'estimer facilement les paramètres $\beta_1, \ldots, \beta_p$ qui traduisent les effets

12. L'idée de cette représentation provient d'un graphique original construit par R. Aeberhardt (http://www.crest.fr/ckfinder/userfiles/files/Pageperso/raeberhardt/FW_beamer.pdf).

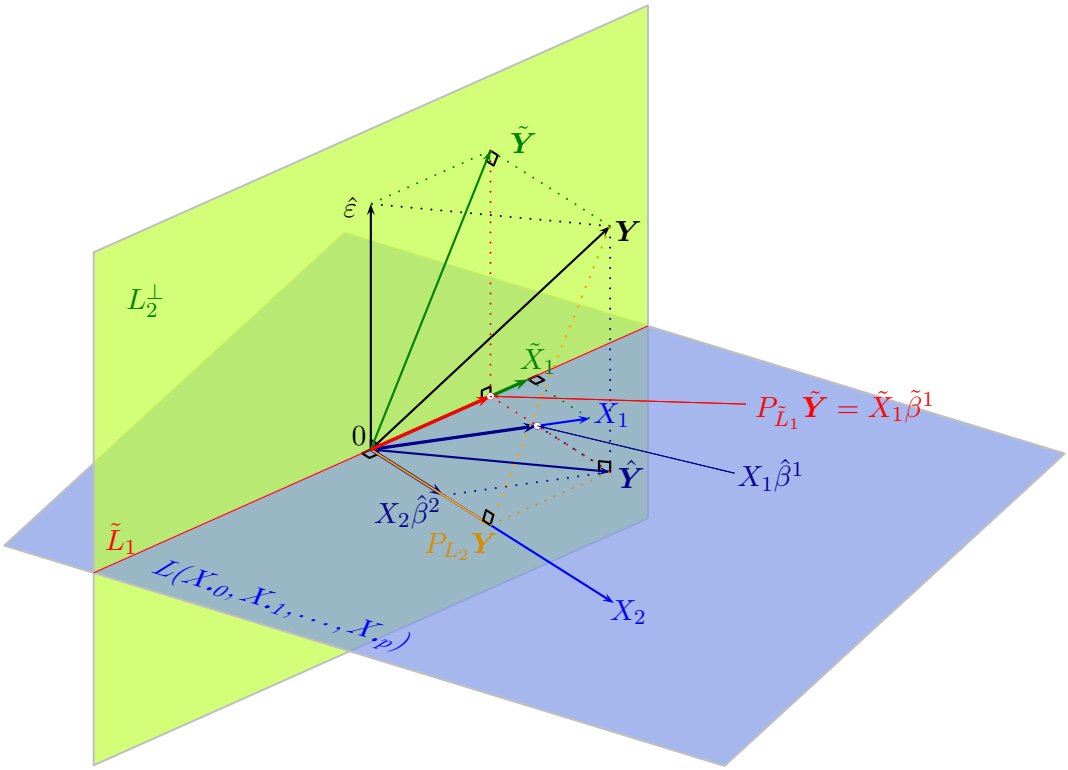


fig:FW

FIGURE 5.6: Illustration du théorème de Frisch-Waugh. *Cliquez pour lancer l'animation* (visible uniquement avec Acrobat Reader, version suffisamment récente)

des variables exogènes sur la variable endogène, sans estimer le terme constant β_0 de la relation entre les variables. Dans cette application, on posera

$$X_1 = \begin{pmatrix} X_{\bullet 1} & \cdots & X_{\bullet p} \end{pmatrix} \qquad X_2 = \begin{pmatrix} X_{\bullet 0} \end{pmatrix} = \begin{pmatrix} 1, 1, \ldots, 1 \end{pmatrix}^\top$$

et donc

$$\beta^1 = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \beta^2 = \begin{pmatrix} \beta_0 \end{pmatrix} \tag{5.17}$$ eq:beta_FW

Comme X_2 est ici le vecteur diagonal de $\mathbb{R}^n$, le sev $L_2 = L(X_2)$ est l'ensemble des vecteurs de $\mathbb{R}^n$ dont toutes les coordonnées sont égales. La projection orthogonale d'un vecteur quelconque $x \in \mathbb{R}^n$ sur un tel sous-espace, notée $P_{L_2}x$, est le vecteur dont toutes les coordonnées sont égales à $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ (voir la remarque 9.9). Autrement dit $P_{L_2}x = \overline{x}X_{\bullet 0}$. Par conséquent, en notant $M_{L_2} = I_n - P_{L_2}$, le vecteur $M_{L_2}x$ est le vecteur x dont on a remplacé chaque coordonnée par sa différence par rapport à la moyenne des coordonnées :

$$M_{L_2}x = x - P_{L_2}x = x - \overline{x}X_{\bullet 0} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} \overline{x} \\ \vdots \\ \overline{x} \end{pmatrix} = \begin{pmatrix} x_1 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{pmatrix}$$

Ce résultat est en particulier vrai pour le vecteur $\boldsymbol{Y}$ et pour chacun des vecteurs constituant les colonnes de X_1. On a donc

$$\tilde{\boldsymbol{Y}} = M_{L_2}\boldsymbol{Y} = \boldsymbol{Y} - \overline{Y}X_{\bullet 0} = \begin{pmatrix} Y_1 - \overline{Y} \\ \vdots \\ Y_n - \overline{Y} \end{pmatrix}$$

et

$$\begin{aligned} \tilde{X}_1 = M_{L_2}X_1 &= \begin{pmatrix} X_{\bullet 1} & \cdots & X_{\bullet p} \end{pmatrix} - \begin{pmatrix} \overline{X}_{\bullet 1}X_{\bullet 0} & \cdots & \overline{X}_{\bullet p}X_{\bullet 0} \end{pmatrix} \\ &= \begin{pmatrix} X_{\bullet 1} - \overline{X}_{\bullet 1}X_{\bullet 0} & \cdots & X_{\bullet p} - \overline{X}_{\bullet p}X_{\bullet 0} \end{pmatrix} \\ &= \begin{pmatrix} X_{11} - \overline{X}_{\bullet 1} & X_{12} - \overline{X}_{\bullet 2} & \cdots & X_{1p} - \overline{X}_{\bullet p} \\ X_{21} - \overline{X}_{\bullet 1} & X_{22} - \overline{X}_{\bullet 2} & \cdots & X_{2p} - \overline{X}_{\bullet p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n1} - \overline{X}_{\bullet 1} & X_{n2} - \overline{X}_{\bullet 2} & \cdots & X_{np} - \overline{X}_{\bullet p} \end{pmatrix} \end{aligned} \tag{5.18}$$ eq:X_centre

où $\overline{X}_{\bullet k} = \frac{1}{n}\sum_{i=1}^n X_{ik}$ est la moyenne des observations de la k^{e} variable explicative. Selon ce qui a été dit ci-dessus dans la remarque 5.16, pour estimer β^1, il suffit d'appliquer la méthode des moindres carrés au modèle dans lequel la variable dépendante est $\tilde{\boldsymbol{Y}}$ et les variables explicatives sont les colonnes de $\tilde{X}_1$. Dans le cas illustré ici, d'après ce qui vient d'être décrit ci-dessus, cela revient à estimer par moindres carrés les paramètres de la relation initiale où les variables (exogènes et endogène) ont préalablement été transformées en "différences à la moyenne". De manière plus

explicite, on calcule la moyenne de chacune des variables du modèle initial (endogène et exogènes) ; puis on génère de nouvelles variables en soustrayant à chaque observation de chaque variable la propre moyenne des observations de cette variable. On estime alors par moindres carrés la relation entre la nouvelle variable endogène ainsi obtenue et les nouvelles variables exogènes. Cette relation est

$$Y_i - \overline{Y} = \beta_1(X_{i1} - \overline{X}_{\cdot 1}) + \cdots + \beta_p(X_{ip} - \overline{X}_{\cdot p}) + U_i \qquad i = 1, \ldots, n \tag{5.19}$$ eq:rel_centree

où $U_i = (\varepsilon_i - \overline{\varepsilon})$. D'après le théorème 5.4, l'estimation par moindres carrés ordinaires des paramètres de (5.19) coïncide avec $(\hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$. Formellement, ces p dernières coordonnées de $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ s'écrivent $(\tilde{X}_1^\top \tilde{X}_1)^{-1} \tilde{X}_1^\top \tilde{\mathbf{Y}}$.

Dans ce même contexte, l'utilisation du point 2 du théorème permet d'obtenir un résultat supplémentaire concernant le coefficient de détermination de la régression. Pour cela, et afin d'alléger la notation, on pose $\mathbf{y} = \tilde{Y}$ et $\mathbf{x} = \tilde{X}_1$. Avec ces notations, les n relations (5.19) s'écrivent

$$\mathbf{y} = \mathbf{x}\beta^1 + U \tag{5.20}$$ eq:mrls_centre

où U est le vecteur aléatoire dont la i^{e} coordonnée est $U_i = \varepsilon_i - \overline{\varepsilon}$. Dans ce modèle, l'estimation de β^1 par moindres carrés conduit à

$$\hat{\beta}^1 = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y} \tag{5.21}$$ eq:betahat_FW

On peut définir les valeurs ajustées $\hat{\mathbf{y}} = \mathbf{x}\hat{\beta}^1$ et les résidus $\hat{U} = \mathbf{y} - \hat{\mathbf{y}}$ de cette estimation, conformément à la définition 5.3 et l'interprétation de la section 5.4. On va chercher à établir une égalité semblable à (5.12) qui permet de décomposer la variabilité observée de la variable endogène. Cette variabilité est mesurée par la somme des carrés des écarts entre les observations de cette variable et leur moyenne (c'est le membre de gauche de (5.12)). Dans le contexte d'un modèle où la relation est (5.19), ces observations sont $Y_i - \overline{Y}$, $i = 1, \ldots, n$ et leur moyenne est nulle. Donc la variabilité observée de la variable endogène est $\sum_{i=1}^n (Y_i - \overline{Y})^2$, ou, de manière identique, $\|\mathbf{y}\|^2$. Pour obtenir une décomposition de cette quantité semblable à celle du théorème 5.3, on procède comme dans la preuve de ce résultat. On a $\mathbf{y} = \mathbf{y} - \hat{\mathbf{y}} + \hat{\mathbf{y}}$. On note que $\mathbf{y} - \hat{\mathbf{y}} = \hat{U}$, et que, pour les mêmes raisons que dans la preuve du théorème 5.3, ce vecteur est orthogonal à $\hat{\mathbf{y}}$. Donc $\|\mathbf{y}\|^2 = \|\hat{U}\|^2 + \|\hat{\mathbf{y}}\|^2$ ou encore

$$\|\hat{\mathbf{y}}\|^2 = \|\mathbf{y}\|^2 - \|\hat{U}\|^2 \tag{5.22}$$ eq:yhat_centre

Comme le théorème 5.4 implique que $\hat{U} = \hat{\varepsilon}$, où, comme jusqu'à présent, $\hat{\varepsilon}$ désigne les résidus de l'estimation par moindres carrés du MRLS de départ, l'égalité ci-dessus s'écrit :

$$\|\hat{\mathbf{y}}\|^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2 - \sum_{i=1}^n \hat{\varepsilon}_i^2$$

Or on déduit de l'égalité (5.11) que le membre de gauche est $\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$, où $\hat{Y}_1, \ldots, \hat{Y}_n$ sont les valeurs ajustées dans le MRLS initial. En résumé, si on écrit les sommes de carrés sous la forme de carrés de normes, on vient d'établir que

$$\begin{aligned} \|\mathbf{y}\|^2 &= \|\boldsymbol{Y} - X_{\cdot 0}\overline{Y}\|^2 \\ \|\hat{\mathbf{y}}\|^2 &= \|\hat{\boldsymbol{Y}} - X_{\cdot 0}\overline{Y}\|^2 \\ \|\hat{U}\|^2 &= \|\hat{\varepsilon}\|^2 \end{aligned} \tag{5.23}$$ eq:ehat_centre

En utilisant la définition du coefficient de détermination de la régression dans le MRLS initial, ces égalités impliquent que ce coefficient s'écrit

$$R^2 = \frac{\|\hat{\mathbf{y}}\|^2}{\|\mathbf{y}\|^2} \tag{5.24}$$

eq:R2_centre

et coïncide par conséquent avec le coefficient de détermination calculé dans le modèle issu de la transformation des variables, dans lequel la relation est (5.19).

On termine en notant que l'équivalence $R^2 = 0 \iff \hat{\beta}_1 = \cdots = \hat{\beta}_p = 0$ de la propriété **??** s'obtient quasiment immédiatement en utilisant cette reformulation du coefficient R^2. En effet, on a

$$R^2 = 0 \iff \|\hat{\mathbf{y}}\|^2 = 0 \iff \hat{\mathbf{y}} = 0_n \iff \mathbf{x}\hat{\beta}^1 = 0_n \iff \hat{\beta}^1 = 0_p$$

où la troisième équivalence découle directement de la définition de $\hat{\mathbf{y}}$ et la dernière du fait que $\mathbf{x} = \tilde{X}_1$ est de dimensions (n, p) et de rang p.

sec:mco_maxRa

5.5.1.4 L'estimateur des moindres carrés maximise la corrélation empirique entre variables

Cette section fournit la preuve de la propriété 5.5, en s'appuyant sur le théorème de Frisch-Waugh et sur le résultat de l'application de la section précédente. On rappelle que la propriété 5.5 permet d'établir que la méthode d'estimation de β par moindres carrés conduit à rechercher la combinaison linéaire des variables exogènes pour laquelle la corrélation empirique avec la variable endogène est la plus forte.

Pour un jeu de coefficients $a_0, \ldots, a_p$ donné, on forme la variable notée X_a, définie par la combinaison linéaire des variables exogènes $X_0, \ldots, X_p$:

$$X_a = a_0 X_0 + \cdots + a_p X_p$$

Les observations de cette variable sont donc $X_{1a}, \ldots, X_{na}$, avec $X_{ia} = a_0 X_{i0} + \cdots + a_p X_{ip}$. Le coefficient de corrélation linéaire empirique entre la variable endogène Y et la variable X_a est donc

$$r(Y, X_a) = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_{ia} - \overline{X}_a)}{\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2 \sum_{i=1}^n (X_{ia} - \overline{X}_a)^2}} \tag{5.25}$$

eq:def_ryxa

où $\overline{X}_a$ désigne la moyenne (empirique) des n observations de la variable X_a. Cette construction étant possible pour tout jeu possible de coefficients $a_0, \ldots, a_p$, on peut chercher celui pour lequel la valeur absolue de $r(Y, X_a)$ est maximum. Un tel jeu identifiera la combinaison linéaire des variables exogènes dont les observations sont les plus fortement corrélées (sans tenir compte du sens de la corrélation) avec celles de la variable endogène. Une telle combinaison est celle par laquelle les variables exogènes ont le plus fort lien linéaire avec la variable endogène. C'est précisément le type de propriété que l'on cherche à obtenir dans le MRLS lorsqu'on estime $\beta_0, \ldots, \beta_p$ de manière à donner aux variables exogènes la plus forte capacité à déterminer le niveau de la variable endogène. Le résultat qu'on cherche à montrer ne devrait donc pas être une surprise, dans le sens où les estimateurs $\hat{\beta}_0, \ldots, \hat{\beta}_p$ ainsi obtenus sont ceux pour lesquels la combinaison linéaire $\hat{\beta}_0 X_0 + \cdots + \hat{\beta}_p X_p$ a le plus fort lien linéaire avec Y.

Les observations des variables étant données, la valeur de $r(Y, X_a)$ ne dépend que des coefficients $a_0, \ldots, a_p$. Par conséquent, l'objectif visé sera atteint en considérant le problème de maximisation de $|r(Y, X_a)|$ par rapport à $a_0, \ldots, a_p$, ou de manière identique, la maximisation de $r(Y, X_a)^2$ par rapport à ces mêmes coefficients.

Avant de résoudre un problème de maximisation, on notera deux propriétés de $r(Y, X_a)^2$ en tant que fonction de $a_0, \ldots, a_p$, qui nous permettront de simplifier la recherche de la solution. Notons d'abord que $r(Y, X_a)^2$ ne dépend pas de a_0. En effet, un simple calcul montre que $X_{ia} - \overline{X}_a$ ne dépend ni de a_0 ni de X_{i0}. Par conséquent, en considérant l'expression de $r(Y, X_a)$ donnée par (5.25), on déduit que ce coefficient ne dépend pas non plus de a_0. Cette propriété s'obtient également en notant que X_0 est une "variable" constante (toutes ses observations sont égales à 1). Il est donc normal qu'elle ne soit pas corrélée avec Y (ou avec n'importe quelle autre variable), et qu'elle ne contribue pas à la corrélation entre X_a et Y. Par conséquent, lorsqu'on cherche la combinaison linéaire des variables exogènes la plus fortement corrélée à Y, on peut choisir a_0 de manière arbitraire. Par commodité, on imposera $a_0 = 0$, et les combinaisons linéaires auxquelles on s'intéressera lors de la maximisation seront de la forme $X_a = a_1 X_1 + \cdots + a_p X_p$.

La seconde propriété qui sera utilisée consiste en l'égalité $r(Y, X_a)^2 = r(Y, X_{\gamma a})^2$, pour tout réel γ et pour tout jeu de réels $a_1, \ldots, a_p$. Cette égalité résulte directement de l'expression de $r(Y, X_a)$ et du fait que $X_{\gamma a} = \gamma a_1 X_1 + \cdots + \gamma a_p X_p = \gamma X_a$, et donc $X_{i(\gamma a)} = \gamma X_{ia}$, $i = 1, \ldots, n$.

Pour résoudre ce problème de maximisation, il est intéressant d'utiliser une autre formulation de $r(Y, X_a)^2$. Tout d'abord, pour tout choix de nombres $a_1, \ldots, a_p$, on introduit les vecteurs $a \in \mathbb{R}^p$ et $\mathbf{x}_a \in \mathbb{R}^n$ définis par

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} \qquad \mathbf{x}_a = \begin{pmatrix} X_{1a} - \overline{X}_a \\ \vdots \\ X_{na} - \overline{X}_a \end{pmatrix}$$

où $X_{ia} = a_1 X_{i1} + \cdots + a_p X_{ip}$ et $\overline{X}_a = \frac{1}{n} \sum_{i=1}^n X_{ia}$. En utilisant la notation de la section précédente où $\mathbf{y}$ désigne le vecteur des observations de la variable endogène, en différences par rapport à leur moyenne (voir (5.20)) :

$$\mathbf{y} = \begin{pmatrix} Y_1 - \overline{Y} \\ \vdots \\ Y_n - \overline{Y} \end{pmatrix}$$

on peut écrire

$$r(Y, X_a)^2 = \frac{(\mathbf{x}_a^\top \mathbf{y})^2}{(\mathbf{y}^\top \mathbf{y})(\mathbf{x}_a^\top \mathbf{x}_a)} = \frac{\mathbf{x}_a^\top \mathbf{y}\mathbf{y}^\top \mathbf{x}_a}{(\mathbf{y}^\top \mathbf{y})(\mathbf{x}_a^\top \mathbf{x}_a)}$$

Finalement, toujours avec la notation de la section précédente où $\mathbf{x}$ désigne la matrice dont les éléments sont les observations des variables exogènes (en excluant X_0) en différences par rapport à leurs moyennes (voir (5.18) et (5.20)), il est facile de vérifier en effectuant le produit matriciel $\mathbf{x}a$ que $\mathbf{x}_a = \mathbf{x}a$. On peut alors écrire

$$r(Y, X_a)^2 = \frac{a^\top \mathbf{x}^\top \mathbf{y}\mathbf{y}^\top \mathbf{x} a}{(\mathbf{y}^\top \mathbf{y}) a^\top \mathbf{x}^\top \mathbf{x} a} \tag{5.26}$$ eq:ryxa2

À partir de cette expression, on va montrer que l'estimateur des moindres carrés de β fournit

la solution du problème

$$\max_{a\in\mathbb{R}^p} r(Y, X_a)^2$$

Avec la remarque faite précédemment que $r(Y, X_{\gamma a})^2 = r(Y, X_a)^2$ pour tout $a \in \mathbb{R}^p$ et tout $\gamma \in \mathbb{R}$, on note immédiatement que si a^* est une solution de ce problème, alors pour tout réel non nul γ, le vecteur γa^* est également une solution. Ce problème admet donc une infinité de solutions. Afin d'en déterminer une, on imposera au vecteur solution a^* d'avoir une norme égale à 1, *i.e.*, ${a^*}^\top a^* = 1$.[13] Une fois cette solution trouvée, on pourra choisir n'importe quel $\hat{a} = \gamma a^*$ pour former la combinaison linéaire $X_{\hat{a}}$ maximisant $r(Y, X_a)^2$. Le problème à résoudre devient donc

$$\max_{a\in\mathbb{R}^p} r(Y, X_a)^2 \quad \text{s.c.q.} \quad a^\top a = 1 \tag{5.27}$$

eq:maxRa

Finalement, en utilisant l'expression (5.26) de $r(Y, X_a)^2$ et en notant que $\mathbf{y}^\top\mathbf{y} > 0$, il revient au même de résoudre

$$\max_{a\in\mathbb{R}^p} \frac{a^\top\mathbf{x}^\top\mathbf{y}\mathbf{y}^\top\mathbf{x}a}{a^\top\mathbf{x}^\top\mathbf{x}a} \quad \text{s.c.q.} \quad a^\top a = 1 \tag{5.28}$$

eq:maxRa_bis

Pour résoudre ce problème, on envisage d'abord le cas simple où on suppose que les observations des variables exogènes sont telles que $\mathbf{x}^\top\mathbf{x} = I_p$. On verra ensuite que le cas général s'obtient facilement à partir de ce cas simple.

Cas $\mathbf{x}^\top\mathbf{x} = I_p$ Avec la contrainte $a^\top a = 1$, si on suppose que $\mathbf{x}^\top\mathbf{x} = I_p$, le dénominateur de la fonction à maximiser est $a^\top\mathbf{x}^\top\mathbf{x}a = a^\top a = 1$. On cherche donc la solution de

$$\max_{a\in\mathbb{R}^p} a^\top\mathbf{x}^\top\mathbf{y}\mathbf{y}^\top\mathbf{x}a \quad \text{s.c.q.} \quad a^\top a = 1 \tag{5.29}$$

eq:maxRa_simple

La fonction à maximiser étant convexe en a, on peut utiliser la méthode du lagrangien. Pour que le vecteur a^* soit solution du problème, il faut qu'il existe un nombre λ^* tel que

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial a_k}(a^*, \lambda^*) = 0, & \forall k = 1, \ldots, p \\ \frac{\partial\mathcal{L}}{\partial\lambda}(a^*, \lambda^*) = 0 \end{cases} \tag{5.30}$$

eq:cpo_max_ryx

où la fonction lagrangien $\mathcal{L}$ est définie par $\mathcal{L}(a, \lambda) = a^\top\mathbf{x}^\top\mathbf{y}\mathbf{y}^\top\mathbf{x}a + \lambda(1 - a^\top a)$. Si pour alléger la notation on introduit la matrice Γ de dimensions (p, p) définie par $\Gamma = \mathbf{x}^\top\mathbf{y}\mathbf{y}^\top\mathbf{x}$, on a $a^\top\mathbf{x}^\top\mathbf{y}\mathbf{y}^\top\mathbf{x}a = a^\top\Gamma a = \sum_{l=1}^p\sum_{m=1}^p a_l a_m\Gamma_{lm}$, où Γ_{lm} est le $(l, m)^{\text{e}}$ élément de Γ. Par conséquent

$$\frac{\partial a^\top\mathbf{x}^\top\mathbf{y}\mathbf{y}^\top\mathbf{x}a}{\partial a_k} = \sum_{m=1}^p a_m\Gamma_{km} + \sum_{l=1}^p a_l\Gamma_{lk} = 2\sum_{l=1}^p a_l\Gamma_{kl}$$

où la seconde égalité résulte de la symétrie de la matrice Γ. D'autre part, en appliquant ce raisonnement en supposant $\Gamma = I_p$, on a

$$\frac{\partial a^\top a}{\partial a_k} = 2a_k$$

13. Cette contrainte apparaît naturellement en notant que si a^* est une solution, alors d'après ce qui vient d'être dit sur l'ensemble des solutions, le vecteur $\frac{1}{{a^*}^\top a^*}a^*$ est aussi solution ; on voit facilement que ce dernier est de norme égale à 1.

En résumé, on a obtenu

$$\frac{\partial \mathcal{L}}{\partial a_k}(a, \lambda) = 2\sum_{l=1}^{p} a_l \Gamma_{kl} - 2\lambda a_k$$

Donc la condition $\frac{\partial \mathcal{L}}{\partial a_k}(a^*, \lambda^*) = 0$ s'écrit $\sum_{l=1}^{p} a_l^* \Gamma_{kl} = \lambda^* a_k^*$. On empile ces p égalités pour obtenir les p premières équations de (5.30). Avec les notations introduites, celles-ci s'écrivent $2\Gamma a^* - 2\lambda^* a^* = 0_p$, ou encore

$$\Gamma a^* = \lambda^* a^* \tag{5.31}$$ eq:cpo_max_ryx

La dernière équation de (5.30) exprime évidement que la solution a^* satisfait la contrainte : ${a^*}^\top a^* = 1$.

Conjointement, cette contrainte et l'égalité (5.31) expriment que si un couple (a^*, λ^*) satisfait (5.30), alors il est nécessairement l'un des p couples (vecteur propre, valeur propre) de Γ.

En prémultipliant les deux membres de (5.31) par ${a^*}^\top$ et en utilisant la contrainte, on obtient

$${a^*}^\top \Gamma a^* = \lambda^* \tag{5.32}$$ eq:vp

Comme a^* est le vecteur de norme 1 pour lequel $a^\top \Gamma a$ est maximum (par définition du problème 5.29)) et qu'avec sa valeur propre associée λ^* il satisfait (5.32), on voit que cette dernière est nécessairement égale à la plus grande des valeurs propres de Γ. Cette matrice est semi-définie positive (puisqu'elle peut s'écrire sous la forme d'un produit $A^\top A$). Donc toutes ses valeurs propres sont positives ou nulles. De plus, on constate que Γ s'écrit aussi sous la forme $cc^\top$ où $c = \mathbf{x}^\top \mathbf{y}$ est un vecteur de $\mathbb{R}^p$. Le rang de Γ est donc égal à 1. Ceci implique que parmi les p valeurs propres de Γ, seule l'une d'elles est positive et les $(p-1)$ autres sont nulles. Donc la plus grande valeur propre λ^* de Γ est celle qui est non nulle. En conséquence, le résultat général établissant que la trace d'une matrice est égale à la somme de ses valeurs propres prend ici la forme $\text{trace}(\Gamma) = \lambda^*$. En utilisant les propriétés de l'opérateur trace[14], on peut écrire

$$\text{trace}(\Gamma) = \text{trace}(\mathbf{x}^\top \mathbf{y}\mathbf{y}^\top \mathbf{x}) = \text{trace}(\mathbf{y}^\top \mathbf{x}\mathbf{x}^\top \mathbf{y}) = \mathbf{y}^\top \mathbf{x}\mathbf{x}^\top \mathbf{y}$$

où la dernière égalité provient du fait que $(\mathbf{y}^\top \mathbf{x}\mathbf{x}^\top \mathbf{y})$ est une "matrice" de dimensions (1,1). On obtient donc

$$\lambda^* = \mathbf{y}^\top \mathbf{x}\mathbf{x}^\top \mathbf{y}$$

Si on substitue cette expression dans (5.31), on a

$$\mathbf{x}^\top \mathbf{y}\mathbf{y}^\top \mathbf{x} a^* = \mathbf{y}^\top \mathbf{x}\mathbf{x}^\top \mathbf{y} a^*$$

On vérifie que le vecteur $a^* = \frac{\mathbf{x}^\top \mathbf{y}}{\sqrt{\mathbf{y}^\top \mathbf{x}\mathbf{x}^\top \mathbf{y}}}$ est de norme 1 et satisfait l'égalité ci-dessus. L'ensemble des solutions au problème de maximisation dans le cas où $\mathbf{x}^\top \mathbf{x} = I_p$ est donc constitué de tous les vecteurs de $\mathbb{R}^p$ proportionnels à $\mathbf{x}^\top \mathbf{y}$.

Cas $\mathbf{x}^\top \mathbf{x}$ quelconque Dans ce cas, pour maximiser $r(Y, X_a)^2$, le problème à résoudre reste donné par (5.28). En effectuant un changement de variable adéquat, on peut donner au problème la forme (5.29).

14. $\text{trace}(AB) = \text{trace}(BA)$ pour n'importe quelles matrices A et B pour lesquelles les produits AB et BA sont définis.

Pour cela, remarquons que pour toute matrice M de dimensions (p, p) inversible, le rapport à maximiser s'écrit

$$\frac{a^\top (M^\top)^{-1} M^\top \mathbf{x}^\top \mathbf{y}\mathbf{y}^\top \mathbf{x} M M^{-1} a}{a^\top (M^\top)^{-1} M^\top \mathbf{x}^\top \mathbf{x} M M^{-1} a}$$

Parmi toutes les matrices M inversibles, on peut en choisir une pour laquelle $M^\top \mathbf{x}^\top \mathbf{x} M = I_p$ (voir plus bas), ce qui qui permet d'écrire le rapport ci-dessus sous la forme :

$$\frac{b^\top \tilde{\mathbf{x}}^\top \mathbf{y}\mathbf{y}^\top \tilde{\mathbf{x}} b}{b^\top b}$$

où $\tilde{\mathbf{x}} = \mathbf{x}M$ et $b = M^{-1}a$. Comme M est inversible et connue, chercher le a^* pour lequel le premier rapport est maximum revient à chercher le b pour lequel le second est maximum. On doit donc résoudre

$$\max_{b \in \mathbb{R}^p} \frac{b^\top \tilde{\mathbf{x}}^\top \mathbf{y}\mathbf{y}^\top \tilde{\mathbf{x}} b}{b^\top b}$$

Pour les mêmes raisons que celles précédemment avancées, on peut limiter la recherche du maximum aux vecteurs de $\mathbb{R}^p$ dont la norme est égale à 1. On est alors ramené au cas simple résolu auparavant. En utilisant les résultats obtenus dans ce cas, on obtient que le vecteur b^* qui réalise le maximum est

$$b^* = \frac{\tilde{\mathbf{x}}^\top \mathbf{y}}{\sqrt{\mathbf{y}^\top \tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top \mathbf{y}}} = \frac{M^\top \mathbf{x}^\top \mathbf{y}}{\sqrt{\mathbf{y}^\top \mathbf{x} M M^\top \mathbf{x}^\top \mathbf{y}}}$$

En utilisant la relation $b = M^{-1}a$, on déduit que le a^* recherché dans ce cas est $a^* = Mb^*$.

Pour caractériser la solution, il reste à expliciter la matrice M pour laquelle $M^\top \mathbf{x}^\top \mathbf{x} M = I_p$. Celle-ci s'obtient en diagonalisant $\mathbf{x}^\top \mathbf{x}$. Si V et Λ désignent les matrices contenant respectivement les vecteurs propres et les valeurs propres de $\mathbf{x}^\top \mathbf{x}$, on a $\mathbf{x}^\top \mathbf{x} = V \Lambda V^\top$, car $\mathbf{x}^\top \mathbf{x}$ étant symétrique, on peut choisir V de sorte que $V^\top V = I_p$, ou encore $V^{-1} = V^\top$. On écrit toute matrice diagonale D d'éléments diagonaux $d_1, \ldots, d_q$ sous la forme $D = \mathrm{diag}(d_1, \ldots, d_q)$. On a ainsi $\Lambda = \mathrm{diag}(l_1, \ldots, l_p)$ où $l_1, \ldots, l_p$ désignent les p valeurs propres de $\mathbf{x}^\top \mathbf{x}$. Si on définit la matrice $\Lambda^{1/2} = \mathrm{diag}(\sqrt{l_1}, \ldots, \sqrt{l_p})$, on vérifie facilement que $\Lambda = \Lambda^{1/2}\Lambda^{1/2}$. De manière identique, l'inverse de Λ est $\Lambda^{-1} = \mathrm{diag}(1/l_1, \ldots, 1/l_p)$ et on a $\Lambda^{-1} = \Lambda^{-1/2}\Lambda^{-1/2}$ où $\Lambda^{-1/2} = \mathrm{diag}(1/\sqrt{l_1}, \ldots, 1/\sqrt{l_p}) = (\Lambda^{1/2})^{-1}$. Introduisons alors la matrice M définie par $M = V\Lambda^{-1/2}$. On vérifie que cette matrice satisfait la condition voulue puisque

$$M^\top \mathbf{x}^\top \mathbf{x} M = \Lambda^{-1/2} V^\top V \Lambda V^\top V \Lambda^{-1/2} = \Lambda^{-1/2}\Lambda\Lambda^{-1/2} = \Lambda^{-1/2}\Lambda^{1/2}\Lambda^{1/2}\Lambda^{-1/2} = I_p$$

Avec ce choix particulier de M, le vecteur a^* qui maximise $r(Y, X_a)^2$ est donc

$$a^* = Mb^* = M\frac{M^\top \mathbf{x}^\top \mathbf{y}}{\sqrt{\mathbf{y}^\top \mathbf{x} M M^\top \mathbf{x}^\top \mathbf{y}}} = \frac{(\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{y}}{\sqrt{\mathbf{y}^\top \mathbf{x} (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y}}}$$

puisque $MM^\top = V\Lambda^{-1/2}\Lambda^{-1/2}V^\top = V\Lambda^{-1}V^\top = (\mathbf{x}^\top \mathbf{x})^{-1}$. Donc dans le cas général, l'ensemble des solutions au problème de maximisation de $r(Y, X_a)^2$ est l'ensemble des vecteurs de $\mathbb{R}^p$ proportionnels à $(\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{y}$.

Maintenant que ce problème de maximisation est résolu, il reste à conclure et montrer, ainsi qu'on l'a annoncé au début de cette section, que la combinaison linéaire des variables exogènes la plus fortement linéairement corrélée avec Y est bien $\hat{\beta}_0 X_0 + \cdots + \hat{\beta}_p X_p$. Pour cela, on remarque que grâce au résultat de la section 5.5.1.3 précédente, le jeu de p coefficients $(\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{y}$ qui maximise la corrélation linéaire empirique $r(Y, X_a)^2$ coïncide avec $\hat{\beta}_1, \ldots, \hat{\beta}_p$, ce qu'on écrit $\hat{\beta}^1 = (\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{y}$ (voir les commentaires qui suivent (5.19) et (5.20)). Donc la combinaison linéaire obtenue à l'issue de la résolution du problème de maximisation est $\hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$. La valeur de $r(Y, X_a)^2$ qui en résulte reste inchangée si on ajoute un terme constant à cette combinaison linéaire. [15] Par conséquent, la combinaison linéaire $\hat{\beta}_0 X_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$ maximise également ce coefficient de corrélation linéaire, et on obtient ainsi le résultat annoncé.

Pour terminer cette section, notons que puisque le vecteur qui maximise $r(Y, X_a)^2$ est $\hat{\beta}^1 = (\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{y}$, alors en utilisant l'expression de $r(Y, X_a)^2$ donnée par (5.26), le maximum atteint est

$$\frac{\mathbf{y}^\top \mathbf{x}(\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{y}\mathbf{y}^\top \mathbf{x}(\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{y}}{(\mathbf{y}^\top \mathbf{y})\mathbf{y}^\top \mathbf{x}(\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{x}(\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{y}} = \frac{\mathbf{y}^\top \mathbf{x}(\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{y}}$$

Il est facile de voir que le numérateur s'écrit également comme $\hat{\mathbf{y}}^\top \hat{\mathbf{y}}$, où $\hat{\mathbf{y}}$ est défini comme à la section 5.5.1.3 par $\hat{\mathbf{y}} = \mathbf{x}\hat{\beta}^1$. Autrement dit, le maximum de $r(Y, X_a)^2$ atteint en $a = \hat{\beta}^1$ est égal à

$$\frac{\hat{\mathbf{y}}^\top \hat{\mathbf{y}}}{\mathbf{y}^\top \mathbf{y}}$$

En utilisant l'équation (5.24), on constate que ce maximum, défini comme le carré du coefficient de corrélation linéaire multiple entre Y et $X_0, \ldots, X_p$ (voir la définition qui précède la propriété 5.5) coïncide avec le coefficient de détermination de la régression. Ce résultat est évidemment la généralisation de la propriété 2.7.

sec:mcocont

5.5.2 Estimation de β sous contraintes linéaires

Dans cette section, on reprend le problème d'estimation de β de la section 5.3, mais dans un modèle dans lequel on suppose que β satisfait q contraintes de la forme

$$\begin{aligned} R_{10}\beta_0 + R_{11}\beta_1 + \cdots + R_{1p}\beta_p &= r_1 \\ R_{20}\beta_0 + R_{21}\beta_1 + \cdots + R_{2p}\beta_p &= r_2 \\ &\vdots \\ R_{q0}\beta_0 + R_{q1}\beta_1 + \cdots + R_{qp}\beta_p &= r_q \end{aligned}$$

où R_{kl} et r_k sont des nombres connus, $k = 1, \ldots q$, $l = 0, \ldots, p$. En formant la matrice R dont la $(k, l+1)^{\text{e}}$ entrée est R_{kl} et le vecteur $r \in \mathbb{R}^q$ dont les coordonnées sont les réels $r_1, \ldots, r_q$, on peut écrire ces q contraintes comme

$$R\beta = r \tag{5.33}$$

eq:cont_lin

De plus, si R_k désigne le vecteur de $\mathbb{R}^{p+1}$ dont les coordonnées forment les entrées de la k^{e} ligne de R, la k^{e} contrainte s'écrit $R_k^\top \beta = r_k$.

15. Voir la remarque faite au début de cette section.

Pour que le système de contraintes ait un intérêt, on supposera que ces q contraintes ne sont pas linéairement redondantes, dans le sens où aucune d'entre-elles ne peut s'obtenir comme une combinaison linéaire des autres. Cela équivaut à l'indépendance linéaire des vecteurs $R_1, \dots, R_q$ ou encore à la condition $\text{rang}(R) = q$. On notera que ceci implique $q \leq p+1$. Ceci implique également qu'on pourra toujours trouver un $\beta \in \mathbb{R}^{p+1}$ tel que l'égalité (5.33) est satisfaite.

Imposer une contrainte telle que (5.33) revient à introduire une condition supplémentaire au modèle défini par C_p1 à C_p3. Le modèle dans lequel est imposé la contrainte (5.33) est appelé MRLS contraint, et il est défini par la condition

$$\text{C}_p^{\text{cont}} : \qquad \exists \beta \in \mathbb{R}^{p+1} \text{ t.q. } R\beta = r \text{ et } \text{E}(\boldsymbol{Y}) = X\beta, \quad \exists \sigma \in]0, \infty[\text{ t.q. } \text{V}(\boldsymbol{Y}) = \sigma^2 I_n$$

rem:cont_lin

Remarque 5.21 Comme on l'a fait dans la remarque 5.1, on peut donner une interprétation géométrique de la contrainte imposée sur β. Pour cela, on note que la condition $\text{rang}(R) = q$ imposée pour éviter la redondance des contraintes (voir ci-dessus), équivaut à ce que les q lignes de R soient les transposées de vecteurs linéairement indépendants de $\mathbb{R}^{p+1}$. On peut alors compléter cette famille de vecteurs par $p+1-q$ vecteurs linéairement indépendants de $\mathbb{R}^{p+1}$ afin de former une base de $\mathbb{R}^{p+1}$.[16] Notons Q la matrice de dimensions $(p+1-q, p+1)$ dont les lignes sont les transposées de ces vecteurs. On peut alors former la matrice A de dimensions $(p+1, p+1)$ en concaténant verticalement R et Q :

$$A = \left(\begin{array}{c} R \\ \hline Q \end{array} \right) \tag{5.34}$$

eq:complementR

Définissons alors le vecteur δ de $\mathbb{R}^{p+1}$ par

$$\delta = A\beta = \left(\begin{array}{c} R \\ \hline Q \end{array} \right) \beta = \left(\begin{array}{c} R\beta \\ \hline Q\beta \end{array} \right)$$

Par construction, les $p+1$ lignes de A forment une famille de vecteurs linéairement indépendants de $\mathbb{R}^{p+1}$, et la matrice A est de rang $p+1$, donc inversible. Par conséquent, il est équivalent de connaître β ou δ et au lieu d'écrire le modèle paramétré par β, on peut l'écrire au moyen du vecteur de paramètres δ, puisque $\text{E}(\boldsymbol{Y}) = X\beta \iff \text{E}(\boldsymbol{Y}) = XA^{-1}\delta$. On voit alors que la contrainte $R\beta = r$ revient à imposer que les q premières coordonnées de δ soient égales à celles de r et donc que δ s'écrive

$$\delta = \left(\begin{array}{c} r \\ \hline \gamma \end{array} \right)$$

pour un certain $\gamma \in \mathbb{R}^{p+1-q}$. Il est donc possible de reformuler la condition C_p^{cont} qui permet de définir le MRLS contraint. On a en effet $X\beta = XA^{-1}\delta$ et donc si la contrainte $R\beta = r$ est introduite, $X\beta$ s'écrit

$$X\beta = X \left(A_1 \vdots A_2 \right) \left(\begin{array}{c} r \\ \hline \gamma \end{array} \right) = XA_1 r + XA_2\gamma \tag{5.35}$$

eq:repar_delta

16. Un choix possible (et aisé) consiste à choisir ces vecteurs comme étant une base du noyau de R.

où A_1 et A_2 sont les sous-matrices de A^{-1} composées des q premières et $p+1-q$ dernières colonnes de A^{-1}, respectivement. La condition $\mathrm{C}_p^{\mathrm{cont}}$ impose donc en particulier

$$\exists \beta \in \mathbb{R}^{p+1} \text{ t. q. } R\beta = r \text{ et } \mathrm{E}(\boldsymbol{Y}) = X\beta \iff \exists \gamma \in \mathbb{R}^{p+1-q} \text{ t. q. } \mathrm{E}(\boldsymbol{Y}) = a + \tilde{X}\gamma \tag{5.36}$$

eq:equiv_C_p_c

où $a = XA_1r$ et $\tilde{X} = XA_2$. On voit alors que la condition $\mathrm{C}_p^{\mathrm{cont}}$ impose au vecteur $\mathrm{E}(\boldsymbol{Y})$ d'appartenir à un sous-espace affine de $\mathbb{R}^n$, qu'on notera par la suite $L_a(\tilde{X}_{\bullet 0}, \ldots, \tilde{X}_{\bullet p})$. Ce sous-espace est obtenu en translatant au moyen du vecteur a tous les vecteurs du sous-espace engendré par les colonnes de $\tilde{X}$. Appelons ce dernier $L(\tilde{X}_{\bullet 0}, \ldots, \tilde{X}_{\bullet p})$. Comme $\tilde{X} = XA_2$, chaque colonne de $\tilde{X}$ est une combinaison linéaire des colonnes de X et donc $L(\tilde{X}_{\bullet 0}, \ldots, \tilde{X}_{\bullet p}) \subset L(X_{\bullet 0}, \ldots, X_{\bullet p})$. De plus, la définition de $\tilde{X}$ montre que le rang de cette matrice est égal au rang de A_2.[17] Comme A est inversible, les $p+1-q$ dernières colonnes de A^{-1} constituant A_2 sont linéairement indépendantes. Par conséquent A_2 est de rang $p+1-q$ et $\tilde{X}$ aussi. Donc $L(\tilde{X}_{\bullet 0}, \ldots, \tilde{X}_{\bullet p})$ est un sev de $\mathbb{R}^n$ de dimension $p+1-q$.[18] R

rem:repar_C_p_cont

Remarque 5.22 Notons que la reparamétrisation qui consiste à poser $\delta = A\beta$, où A est définie par (5.34), et écrire $\mathrm{E}(\boldsymbol{Y}) = Z\delta$, avec $Z = XA^{-1}$, est un moyen d'incorporer la contrainte $R\beta = r$ dans l'écriture de $\mathrm{E}(\boldsymbol{Y})$, ainsi que le montre l'équivalence (5.36). On pourra alors prendre en compte directement cette contrainte dans l'interprétation et l'utilisation du modèle. R

Lorsqu'on cherche à estimer le vecteur des paramètres β du MRLS avec la contrainte (5.33), il est normal d'imposer que l'estimateur recherché satisfasse également cette contrainte. Pour effectuer cette estimation, on adopte la même démarche que dans la section 5.3. On minimise donc la fonction S définie en (5.8) sous la contrainte (5.33). Formellement on résoud

$$\min_{\beta \in \mathbb{R}^{p+1}} \|\boldsymbol{Y} - X\beta\|^2 \quad \text{s.c.q.} \quad R\beta = r \tag{5.37}$$

eq:min_cont

Pour résoudre ce problème, on peut adopter deux méthodes. La première consiste à noter que la fonction objectif à minimiser et la contrainte sont dérivables en chacun des $\beta_0, \ldots, \beta_p$, que l'objectif est convexe et que chacune des q contraintes est linéaire en ces mêmes arguments. Par conséquent, on peut utiliser la méthode du lagrangien pour effectuer la minimisation sous contrainte. La seconde méthode consiste à intégrer la contrainte dans la fonction à minimiser en utilisant une réécriture du modèle semblable à celle utilisée dans la remarque 5.21, permettant d'écrire $\mathrm{E}(\boldsymbol{Y})$ sous la forme donnée par l'égalité $\mathrm{E}(\boldsymbol{Y}) = a + \tilde{X}\gamma$. On présente tour à tour chacune de ces deux méthodes, qui conduisent évidemment au même résultat.

Si on utilise la première approche, on réécrit le problème comme

$$\min_{(\beta_0,\ldots,\beta_p)^\top \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} (Y_i - \beta_0 X_{i0} - \cdots - \beta X_{ip})^2 \quad \text{s.c.q.} \quad \begin{array}{l} R_{10}\beta_0 + \cdots + R_{1p}\beta_p - r_1 = 0 \\ \vdots \\ R_{q0}\beta_0 + \cdots + R_{qp}\beta_p - r_q = 0 \end{array}$$

17. Puisque $X^\top X$ est inversible, $\tilde{X}x = 0 \iff A_2 x = 0$. Les matrices $\tilde{X}$ et A_2 ont donc le même noyau. Comme elles ont aussi le même nombre de colonnes, le théorème des dimensions permet de conclure qu'elles ont le même rang.

18. On rappelle le contenu de la remarque 5.1 qui établissait qu'en l'absence de la contrainte, $\mathrm{E}(\boldsymbol{Y})$ appartient à un sev de dimension $p+1$.

Le lagrangien associé est

$$\mathcal{L}(\beta_0,\cdots,\beta_p,\lambda_1,\ldots,\lambda_q)=\sum_{i=1}^{n}(Y_i-\beta_0X_{i0}-\cdots-\beta_pX_{ip})^2+\sum_{l=1}^{q}\lambda_l(R_{l0}\beta_0+\cdots+R_{lp}\beta_p-r_l)$$

Donc pour que $\beta^*=(\beta_0^*,\ldots,\beta_p^*)^\top$ soit solution du problème, il faut trouver q réels $\lambda_1^*,\ldots,\lambda_q^*$ tels que

$$\begin{cases}\dfrac{\partial\mathcal{L}}{\partial\beta_k}(\beta_0^*,\cdots,\beta_p^*,\lambda_1^*,\ldots,\lambda_q^*)=0 & k=0,\ldots,p\\[2ex] \dfrac{\partial\mathcal{L}}{\partial\lambda_l}(\beta_0^*,\cdots,\beta_p^*,\lambda_1^*,\ldots,\lambda_q^*)=0 & l=1,\ldots,q\end{cases}\tag{5.38}$$

eq:cpo_mco_cont

Ensemble, les q dernières équations s'écrivent évidemment $R\beta^*=r$, exprimant que la solution β^* satisfait les q contraintes. En utilisant l'expression du lagrangien, la $(k+1)^{\text{e}}$ équation de ce système est

$$-2\sum_{i=1}^{n}X_{ik}(Y_i-\beta_0^*X_{i0}-\cdots-\beta_p^*X_{ip})+\sum_{l=1}^{q}\lambda_l^*R_{lk}=0$$

En reprenant la démarche de la section 5.3, la première somme de cette égalité peut s'écrire $-2X_{\bullet k}^\top(\boldsymbol{Y}-X\beta^*)$ (voir le passage de (5.3) à (5.4)). La seconde somme s'écrit $R_{\bullet k}^\top\lambda^*$, où $R_{\bullet k}$ désigne le vecteur de $\mathbb{R}^q$ constituant la k^{e} colonne de la matrice R et $\lambda^*=(\lambda_1^*,\ldots,\lambda_q^*)^\top$. Donc la $(k+1)^{\text{e}}$ équation du système (5.38) s'écrit

$$-2X_{\bullet k}^\top(\boldsymbol{Y}-X\beta^*)+R_{\bullet k}^\top\lambda^*=0$$

On constate que le premier terme de cette somme est la $(k+1)^{\text{e}}$ ligne de $-2X^\top(\boldsymbol{Y}-X\beta^*)$ et le second est la $(k+1)^{\text{e}}$ ligne de $R^\top\lambda^*$. En empilant ces $p+1$ égalités, on obtient donc $-2X^\top(\boldsymbol{Y}-X\beta^*)+R^\top\lambda^*=0_{p+1}$ et le système (5.38) s'écrit

$$\begin{cases}-2X^\top(\boldsymbol{Y}-X\beta^*)+R^\top\lambda^*=0_{p+1}\\[2ex] R\beta^*-r=0_q\end{cases}$$

De la première égalité on tire

$$\beta^*=(X^\top X)^{-1}X^\top\boldsymbol{Y}-\frac{1}{2}(X^\top X)^{-1}R^\top\lambda^*=\hat\beta-\frac{1}{2}(X^\top X)^{-1}R^\top\lambda^*\tag{5.39}$$

eq:beta_mco_co

où $\hat\beta$ est l'estimateur des moindres carrés de β dans le modèle sans la contrainte (5.33). La matrice $R(X^\top X)^{-1}R^\top$ est inversible. En effet, pour tout $x\in\mathbb{R}^q$, $x\neq 0_q$, on a $x^\top R(X^\top X)^{-1}R^\top x=z^\top(X^\top X)^{-1}z$, où $z=R^\top x$. Comme R est $(q,p+1)$ et de rang q, $x\neq 0_q$ implique $z\neq 0_{p+1}$. Et comme X est de rang $p+1$, elle est définie positive et on a $z^\top(X^\top X)^{-1}z>0$, c'est à dire $x^\top R(X^\top X)^{-1}R^\top x>0$. La matrice $R(X^\top X)^{-1}R^\top$ est donc définie positive, et par conséquent inversible. En utilisant cela et en substituant l'expression de β^* donnée par (5.39) dans l'équation $R\beta^*-r=0_q$, on a

$$\lambda^*=2\big[R(X^\top X)^{-1}R^\top\big]^{-1}(R\hat\beta-r)\tag{5.40}$$

eq:mcocont_lam

et en substituant dans (5.39), on obtient

$$\beta^*=\hat\beta-(X^\top X)^{-1}R^\top\big[R(X^\top X)^{-1}R^\top\big]^{-1}(R\hat\beta-r)\tag{5.41}$$

eq:beta_mco_co

On peut adopter une autre stratégie de résolution du problème (5.37) qui consiste à utiliser le contenu de la remarque 5.21 afin d'intégrer la contrainte $R\beta = r$ dans la fonction à minimiser $\|\boldsymbol{Y} - X\beta\|^2$. En effet, en utilisant le raisonnement de la section 5.3.2, on voit que la minimisation (5.37) consiste à chercher parmi les vecteurs de $\mathbb{R}^n$ s'écrivant sous la forme $X\beta$ avec $\beta \in \mathbb{R}^{p+1}$ tel que $R\beta = r$, celui qui est le plus proche de $\boldsymbol{Y}$. On a vu dans la remarque 5.21 que ces vecteurs sont les éléments du sous-espace affine $L_a(\tilde{X}_{\bullet 0}, \ldots, \tilde{X}_{\bullet p})$ de $\mathbb{R}^n$ et qu'ils peuvent s'écrire sous la forme $XA_1r + XA_2\gamma$ pour un γ dans $\mathbb{R}^{p+1-q}$ (voir (6.10)). Donc chercher parmi ces vecteurs celui qui est le plus proche de $\boldsymbol{Y}$ revient à chercher un $\gamma^* \in \mathbb{R}^{p+1-q}$ tel que

$$\|\boldsymbol{Y} - XA_1r - XA_2\gamma^*\|^2 \leq \|\boldsymbol{Y} - XA_1r - XA_2\gamma\|^2 \qquad \forall \gamma \in \mathbb{R}^{p+1-q}$$

Autrement dit, le problème (5.37) est équivalent à

$$\min_{\gamma \in \mathbb{R}^{p+1-q}} \|\tilde{\boldsymbol{Y}} - \tilde{X}\gamma\|^2$$

où $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - XA_1r$ et $\tilde{X} = XA_2$. Ré-écrit sous cette forme, on voit que le problème est celui qui se pose lorsqu'on souhaite estimer par moindres carrés le vecteur de paramètres γ dans un modèle caractérisé par la condition $\mathrm{E}(\tilde{\boldsymbol{Y}}) = \tilde{X}\gamma$. Dans un tel modèle, les observations de la variable endogène sont les coordonnées de $\tilde{\boldsymbol{Y}}$ et celles des variables exogènes sont les éléments de la matrice $\tilde{X}$. Dans le contexte de ce modèle, la minimisation s'effectue de manière identique à celle présentée à la section 5.3 et on obtient

$$\gamma^* = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{\boldsymbol{Y}}$$

Les deux approches étant équivalentes, elles conduisent au même vecteur du sous-espace affine $L_a(\tilde{X}_{\bullet 0}, \ldots, \tilde{X}_{\bullet p})$. On doit donc avoir $XA_1r + XA_2\gamma^* = X\beta^*$ ou encore, en se rappelant que $A^{-1} = \left(A_1 \,\vdots\, A_2\right)$ (voir la définition de A_1 et A_2 après (6.10)) :

$$XA^{-1} \left(\begin{array}{c} r \\ \hdashline \gamma^* \end{array}\right) = X\beta^*$$

Comme X est de rang $p+1$, on doit avoir $A^{-1}\delta^* = \beta^*$, où δ^* est défini comme

$$\delta^* = \left(\begin{array}{c} r \\ \hdashline \gamma^* \end{array}\right)$$

Autrement dit, les estimateurs des paramètres δ et β sont liés par la même relation que les paramètres eux-mêmes. On rappelle que ces deux paramètres correspondent à deux façons différentes d'écrire la décomposition de $\mathrm{E}(\boldsymbol{Y})$ (voir la remarque 5.22 et l'équivalence (5.36)). Les deux approches d'estimation qui viennent d'être décrites correspondent à chacune de ces deux manières de paramétrer $\mathrm{E}(\boldsymbol{Y})$.

La relation entre δ^* et β^* permet évidemment aussi de connaître l'un des deux lorsqu'on connaît l'autre. Ainsi, si on décide d'utiliser la deuxième approche qui permet d'obtenir γ^* et donc δ^*, on obtient β^* comme $\beta^* = A^{-1}\delta^*$.

sec:estim_sigma

5.6 Estimation de la variance σ^2 et de $\mathbf{V}(\hat{\beta})$

La section précédente a développé les solutions au problème d'estimation du vecteur des paramètres β qui caractérise dans le cadre d'un MRLS la manière dont les niveaux des variables exogènes affectent celui de la variable endogène.

Le modèle contient également le paramètre σ^2 supplémentaire, dont la valeur inconnue mesure la variabilité théorique de la variable endogène. L'estimation de σ^2 découle du résultat suivant.

pro:estim_sig2

Propriété 5.8 *Dans le MRLS, on a* $\mathrm{E}(\|\hat{\varepsilon}\|^2) = (n-p+1)\sigma^2$.

Preuve : $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - X\hat{\beta} = \boldsymbol{Y} - X(X^\top X)^{-1}X^\top \boldsymbol{Y} = M_X \boldsymbol{Y}$ où $M_X = I_n - X(X^\top X)^{-1}X^\top$. Par ailleurs, $\|\hat{\varepsilon}\|^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \mathrm{trace}(\hat{\varepsilon}\hat{\varepsilon}^\top)$. Par conséquent,

$$\begin{aligned}\|\hat{\varepsilon}\|^2 &= \mathrm{trace}(M_X \boldsymbol{Y}\boldsymbol{Y}^\top M_X^\top) = \mathrm{trace}(M_X \boldsymbol{Y}\boldsymbol{Y}^\top M_X) \\ &= \mathrm{trace}(M_X M_X \boldsymbol{Y}\boldsymbol{Y}^\top) = \mathrm{trace}(M_X \boldsymbol{Y}\boldsymbol{Y}^\top)\end{aligned}$$

où la deuxième égalité résulte de la symétrie de M_X, la troisième du fait que $\mathrm{trace}(AB) = \mathrm{trace}(BA)$ et la dernière de l'idempotence de M_X. Donc

$$\mathrm{E}(\|\hat{\varepsilon}\|^2) = \mathrm{E}\big(\mathrm{trace}(M_X \boldsymbol{Y}\boldsymbol{Y}^\top)\big) = \mathrm{trace}\big(\mathrm{E}(M_X \boldsymbol{Y}\boldsymbol{Y}^\top)\big)$$

La dernière égalité résulte des deux propriétés suivantes : (1) l'espérance d'une matrice dont les entrées sont des variables aléatoires est la matrice dont les entrées sont les espérances des variables aléatoires et (2) l'espérance est un opérateur linéaire. La matrice M_X ne dépend que des variables exogènes, que la condition C_p1 permet de considérer comme non-aléatoires. Par conséquent, en utilisant la condition C_p3, les propriétés de l'opérateur trace et la définition de M_X, on obtient

$$\begin{aligned}\mathrm{trace}\big(\mathrm{E}(M_X \boldsymbol{Y}\boldsymbol{Y}^\top)\big) &= \mathrm{trace}\big(M_X \mathrm{E}(\boldsymbol{Y}\boldsymbol{Y}^\top)\big) = \mathrm{trace}(\sigma^2 M_X) = \sigma^2 \mathrm{trace}(M_X) \\ &= \sigma^2 \mathrm{trace}(I_n - X(X^\top X)^{-1}X^\top) \\ &= \sigma^2\big(\mathrm{trace}(I_n) - \mathrm{trace}(X(X^\top X)^{-1}X^\top)\big) \\ &= \sigma^2\big(n - \mathrm{trace}(X^\top X(X^\top X)^{-1})\big) = \sigma^2\big(n - \mathrm{trace}(I_{p+1})\big) \\ &= \sigma^2(n-p+1)\end{aligned}$$

ce qui est le résultat recherché.

cor:esb_sigma2

Corollaire 5.1 *Dans le MRLS, la variable aléatoire* $\hat{\sigma}^2$ *définie par*

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-(p+1)}$$

est un estimateur sans biais de σ^2.

Comme dans le cas du modèle de régression linéaire simple, l'estimateur ainsi obtenu pour σ^2 permet d'obtenir l'estimation de la matrice des variances-covariance de $\hat{\beta}$ (voir la propriété 2.10).

Comme $\hat{\beta} = \hat{A}\boldsymbol{Y}$ avec $A = (X^\top X)^{-1}X^\top$ (voir la propriété 5.4), on peut appliquer la propriété 9.7 et on obtient

$$\mathrm{V}(\hat{\beta}) = \mathrm{V}(\hat{A}\boldsymbol{Y}) = \hat{A}\mathrm{V}(\boldsymbol{Y})\hat{A}^\top = \sigma^2(X^\top X)^{-1}X^\top I_n X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}$$

En tant que matrice des variances-covariances d'un vecteur aléatoire, (voir page 198), la matrice $\mathrm{V}(\hat{\beta})$ a la structure suivante :

$$\mathrm{V}(\hat{\beta}) = \begin{pmatrix} \mathrm{V}(\hat{\beta}_0) & \mathrm{cov}(\hat{\beta}_0,\hat{\beta}_1) & \mathrm{cov}(\hat{\beta}_0,\hat{\beta}_2) & \cdots & \mathrm{cov}(\hat{\beta}_0,\hat{\beta}_p) \\ \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_0) & \mathrm{V}(\hat{\beta}_1) & \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_2) & \cdots & \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_p) \\ \mathrm{cov}(\hat{\beta}_2,\hat{\beta}_0) & \mathrm{cov}(\hat{\beta}_2,\hat{\beta}_1) & \mathrm{V}(\hat{\beta}_2) & \cdots & \mathrm{cov}(\hat{\beta}_2,\hat{\beta}_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\hat{\beta}_p,\hat{\beta}_0) & \mathrm{cov}(\hat{\beta}_p,\hat{\beta}_1) & \mathrm{cov}(\hat{\beta}_p,\hat{\beta}_p) & \cdots & \mathrm{V}(\hat{\beta}_p) \end{pmatrix} \tag{5.42}$$ eq:vbeta

En utilisant alors l'expression $\mathrm{V}(\hat{\beta}) = \sigma^2(X^\top X)^{-1}$, on a nécessairement pour tout $j,k = 0,1,\ldots,p$

$$\mathrm{cov}(\hat{\beta}_j,\hat{\beta}_k) = \sigma^2 s_{j+1,k+1} \tag{5.43}$$ eq:cov_bjbk

où $s_{l,m}$ est la $(l,m)^{\mathrm{e}}$ entrée de la matrice $(X^\top X)^{-1}$.

La matrice $\mathrm{V}(\hat{\beta})$ est inconnue puisqu'elle dépend de la valeur inconnue de σ^2. Puisqu'on dispose d'un estimateur sans biais de σ^2 on peut former un estimateur de $\mathrm{V}(\hat{\beta})$.

pro:est_Vbeta **Propriété 5.9** *La matrice $\hat{\mathrm{V}}(\hat{\beta})$ définie par $\hat{\mathrm{V}}(\hat{\beta}) = \hat{\sigma}^2(X^\top X)^{-1}$ est un estimateur sans biais de $\mathrm{V}(\hat{\beta})$.*

Preuve : Ce résultat découle directement du corollaire 5.1.

Par construction, la $(j+1,k+1)^{\mathrm{e}}$ entrée de la matrice $\hat{\mathrm{V}}(\hat{\beta})$ est donc $\sigma^2 s_{j+1,k+1}$, et la propriété 5.9 montre que c'est un estimateur sans biais de $\mathrm{cov}(\hat{\beta}_j,\hat{\beta}_k)$. On note cet estimateur $\hat{\mathrm{cov}}(\hat{\beta}_j,\hat{\beta}_k)$ et on a donc

$$\hat{\mathrm{V}}(\hat{\beta}) = \begin{pmatrix} \hat{\mathrm{V}}(\hat{\beta}_0) & \hat{\mathrm{cov}}(\hat{\beta}_0,\hat{\beta}_1) & \hat{\mathrm{cov}}(\hat{\beta}_0,\hat{\beta}_2) & \cdots & \hat{\mathrm{cov}}(\hat{\beta}_0,\hat{\beta}_p) \\ \hat{\mathrm{cov}}(\hat{\beta}_1,\hat{\beta}_0) & \hat{\mathrm{V}}(\hat{\beta}_1) & \hat{\mathrm{cov}}(\hat{\beta}_1,\hat{\beta}_2) & \cdots & \hat{\mathrm{cov}}(\hat{\beta}_1,\hat{\beta}_p) \\ \hat{\mathrm{cov}}(\hat{\beta}_2,\hat{\beta}_0) & \hat{\mathrm{cov}}(\hat{\beta}_2,\hat{\beta}_1) & \hat{\mathrm{V}}(\hat{\beta}_2) & \cdots & \hat{\mathrm{cov}}(\hat{\beta}_2,\hat{\beta}_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\mathrm{cov}}(\hat{\beta}_p,\hat{\beta}_0) & \hat{\mathrm{cov}}(\hat{\beta}_p,\hat{\beta}_1) & \hat{\mathrm{cov}}(\hat{\beta}_p,\hat{\beta}_p) & \cdots & \hat{\mathrm{V}}(\hat{\beta}_p) \end{pmatrix} \tag{5.44}$$ eq:hat_vbeta

Chapitre 6

ch:mrlsgp-tests

Le modèle de régression linéaire standard : tests et régions de confiance

On a vu dans le chapitre 3 comment résoudre des problèmes de tests d'hypothèses portant sur la valeur d'un paramètre du modèle. On généralise à présent ce type de problème et on en donne une solution.

La catégorie de problèmes abordés ici étant les tests d'hypothèses, leur résolution nécessite la possibilité de calculer des probabilités, afin notamment d'évaluer les risques de type 1 et 2 (voir la section 10.3.2.3). Parmi les approches possibles offrant cette possibilité, la plus simple consiste à introduire dans la définition du modèle les lois qui serviront aux calculs de probabilités. C'est ce qui a été fait à la section 3.1 et c'est cette approche qui sera utilisée ici. Pour préciser le modèle qui sera le contexte dans lequel on cherchera des solutions aux problèmes de test qui seront posés, on introduit donc une définition semblable à la définition 3.1 de la section 3.1.

def:mrlsgp

Définition 6.1 *Le modèle de régression linéaire standard gaussien de Y sur $(X_1, \ldots, X_p)$ est le modèle statistique défini par la condition* C_p1 *et* $\mathrm{C}_p\mathrm{N}$*, où cette dernière est*

$$\mathrm{C}_p\mathrm{N}. \qquad \exists \beta \in \mathbb{R}^{p+1},\ \exists \sigma \in \left]0, +\infty\right[,\ \boldsymbol{Y} \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

On note que le modèle défini par la définition 5.1 contient le modèle gaussien défini ci-dessus. En effet, on vérifie facilement que si les conditions C_p1 et $\mathrm{C}_p\mathrm{N}$ sont satisfaites, alors les conditions C_p1 à C_p3 le sont aussi. Donc les résultats qui ont été dérivés sous les conditions qui définissent le MRLS restent valables dans le cadre du modèle gaussien défini ci-dessus.

Par ailleurs, en utilisant les résultats sur les lois normales multivariées (voir la section 9.1.2, et plus particulièrement la propriété 9.8), on voit que $\boldsymbol{Y} \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ et $\boldsymbol{Y} - X\beta \sim \mathcal{N}(0_n, \sigma^2 I_n)$ sont équivalentes. Par conséquent, d'après la définition de $\varepsilon = \boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})$, on voit que si la condition C_p1 est imposée, alors la condition $\mathrm{C}_p\mathrm{N}$ de la définition 6.1 peut être remplacée par la condition suivante :

$$\mathrm{C}_p\mathrm{N}'. \qquad \exists \beta \in \mathbb{R}^{p+1},\ \exists \sigma \in \left]0, +\infty\right[,\ \boldsymbol{Y} = X\beta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$$

L'introduction de la condition supplémentaire de normalité du vecteur $\boldsymbol{Y}$ permet d'obtenir des résultats sur les lois des estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ qu'il était impossible de dériver dans le chapitre précédent sous les seules conditions C_p1 à C_p3.

pro:loi_estim

Propriété 6.1 *Dans le modèle de régression linéaire gaussien, on a les propriétés suivantes :*

1. $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^\top X)^{-1})$
2. $(n-p-1)\dfrac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$
3. $\hat{\beta}$ *et* $\hat{\sigma}^2$ *sont indépendants*

Preuve :

1. On peut écrire $\hat{\beta} = \hat{A}\boldsymbol{Y}$ avec $\hat{A} = (X^\top X)^{-1}X^\top$. En utilisant la condition C_p1, qui assure que $\hat{A}$ est une matrice dont les entrées sont non-aléatoires, et la condition C_pN, on peut appliquer la propriété 9.8 pour déduire que $\hat{\beta}$ est un vecteur aléatoire gaussien. Il reste alors à calculer son espérance et sa variance. Comme $\hat{\beta}$ est un estimateur sans biais de β, son espérance est β. En ce qui concerne sa variance, on a :
$$\mathrm{V}(\hat{\beta}) = \mathrm{V}(\hat{A}\boldsymbol{Y}) = \hat{A}\mathrm{V}(\boldsymbol{Y})\hat{A}^\top = \hat{A}\sigma^2 I_n \hat{A}^\top = \sigma^2(X^\top X)^{-1}$$
où la deuxième égalité résulte de la condition C_p1 et de la propriété 9.7, la troisième de la condition C_pN, et la dernière de l'expression de $\hat{A}$.
2. D'après la définition de $\hat{\sigma}^2$, on peut écrire $(n-p+1)\hat{\sigma}^2 = \hat{\varepsilon}^\top\hat{\varepsilon}$, et donc $(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} = Z^\top M_X Z$, où $Z = \frac{\varepsilon}{\sigma}$ et M_X est définie comme dans la preuve de la propriété 5.8. D'après la définition de ε, on a $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$ et donc $Z \sim \mathcal{N}(0_n, I_n)$ (ceci découle des propriétés 9.8 et 9.7). Comme M_X est idempotente et de rang $n-p-1$, on applique la propriété 9.18 et on obtient le résultat.
3. Ce résultat se démontre en utilisant la propriété 9.15 et la remarque qui la suit. En effet, comme $\hat{\beta} = \hat{A}\boldsymbol{Y}$ et $\hat{\varepsilon} = M_X\boldsymbol{Y}$, avec $\boldsymbol{Y} \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, l'indépendance entre $\hat{\beta}$ et $\hat{\varepsilon}$ résulte directement de $M_X\hat{A} = 0$. On reprend alors la remarque qui suit la propriété 9.15 et on conclut que $\hat{\sigma}^2 = \frac{\hat{\varepsilon}^\top\hat{\varepsilon}}{n-p-1}$ est indépendant de $\hat{\beta}$.

sec:multiv-tests

6.1 Tests d'hypothèses linéaires sur β

sec:pb_test_fisher

6.1.1 Le problème de test

Le problème sera défini par une hypothèse nulle (H_0) qui établit que q combinaisons linéaires données des coordonnées du vecteur β prennent chacune une valeur connue, et une hypothèse alternative (H_1) qui nie l'hypothèse nulle. De manière plus précise, l'hypothèse H_0 affirme que le vecteur β satisfait les q égalités introduites dans la section 5.5.2 :

$$\begin{aligned}
R_{10}\beta_0 + R_{11}\beta_1 + \cdots + R_{1p}\beta_p &= r_1 \\
R_{20}\beta_0 + R_{21}\beta_1 + \cdots + R_{2p}\beta_p &= r_2 \\
&\vdots \\
R_{q0}\beta_0 + R_{q1}\beta_1 + \cdots + R_{qp}\beta_p &= r_q
\end{aligned}$$

où R_{kl} et r_k sont des nombres connus, $k = 1, \dots q$, $l = 0, \dots, p$. La réécriture de ces égalités sous forme matricielle se fait exactement comme dans la section 5.5.2 et on peut alors définir formellement le problème de test considéré ici :

$$H_0 : R\beta = r \qquad H_1 : R\beta \neq r \tag{6.1}$$

eq:pb_test_gen

où la matrice R et le vecteur $r \in \mathbb{R}^q$ sont définis comme à la section 5.5.2, *i.e.*,

$$R = \begin{pmatrix} R_{10} & R_{11} & \dots & R_{1p} \\ R_{20} & R_{21} & \dots & R_{2p} \\ \vdots & \vdots & \dots & \vdots \\ R_{q0} & R_{q1} & \dots & R_{qp} \end{pmatrix} \qquad r = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_q \end{pmatrix}$$

On fera la même hypothèse que dans la section 5.5.2 en supposant qu'aucune des q égalités définissant H_0 n'est redondante. Ceci implique que la matrice R est de rang q.

Si l'hypothèse H_0 est vraie, alors le vecteur des paramètres β satisfait q relations linéaires, caractérisées par R et r. On peut dégager plusieurs cas particuliers d'intérêt.

— Le cas où H_0 impose que $\beta_1 + \dots + \beta_p = r$, où r est une valeur qu'on spécifie, s'obtient en posant $R = \begin{pmatrix} 0 & 1 & 1 & \cdots & 1 \end{pmatrix}$.

— Si on veut tester que $\beta_{k_1} = \dots = \beta_{k_q} = 0$ (nullité simultanée de q paramètres), on choisira $r = 0_q$ et

$$R = \begin{pmatrix} e_{k_1+1}^\top \\ e_{k_2+1}^\top \\ \vdots \\ e_{k_q+1}^\top \end{pmatrix}$$

où e_l est le vecteur de $\mathbb{R}^{p+1}$ dont la l^{e} coordonnée est 1 et les autres sont 0. Pour le vérifier, on observe que

$$R\beta = \begin{pmatrix} e_{k_1+1}^\top \beta \\ e_{k_2+1}^\top \beta \\ \vdots \\ e_{k_q+1}^\top \beta \end{pmatrix} = \begin{pmatrix} \beta_{k_1} \\ \beta_{k_2} \\ \vdots \\ \beta_{k_q} \end{pmatrix}$$

puisque la numérotation des éléments de β commençant à 0, on a $e_{k_l+1}^\top \beta = \beta_{k_l}$, pour $l = 1, \dots, q$. Deux sous-cas importants se dégagent.

— Lorsque $q = 1$, on teste la nullité d'un paramètre. Si celui-ci est β_k, alors la matrice R est $R = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$, où le 1 est sur la $(k+1)^{\text{e}}$ colonne. Si cette hypothèse est acceptée, alors la variable X_k auquel le paramètre est attaché ne joue aucun rôle dans la détermination de Y. On retrouve le test de significativité de β_k présenté à la section 3.2.1. Ce cas particulier est traité en détail à la section 6.1.3.2 ci-dessous.

— Lorsque $q = p$ et que $k_l = l$ pour $l = 1, \dots, p$, l'hypothèse nulle H_0 impose que les paramètres attachés aux variables explicatives sont tous nuls, *i.e.*, $\beta_1 = \dots = \beta_p = 0$. La

matrice R est dans ce cas

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{6.2}$$ eq:Rsignifglob

et donc

$$R\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Si H_0 est vraie, alors aucune des variables explicatives n'a d'effet sur Y. On dit alors qu'on teste la significativité globale des paramètres. Ce cas est traité en détail dans la section 6.1.3.3 ci-dessous.

sec:Ftest

6.1.2 Le test de Fisher : dérivation du test et définition

De la même manière que ce qui a été présenté au chapitre 3, la solution au problème de tester H_0 contre H_1 peut s'obtenir par deux approches différentes. L'une (plus difficile) consiste à se donner un certain nombre de critères que doivent satisfaire les solutions (*i.e.*, les tests) recherchées, et à choisir dans l'ensemble des solutions ainsi délimité, celles qui sont les meilleures. L'autre approche consiste à proposer une solution à partir d'un enchaînement d'arguments "raisonnables" et montrer qu'elle possède de bonnes propriétés. Nous présentons cette deuxième approche.

Notons pour commencer que

$$R\beta = r \iff M(R\beta - r) = 0_q$$

où M est n'importe quelle matrice $q \times q$ inversible donnée. Décider entre H_0 et H_1 revient à décider si le vecteur $M(R\beta - r)$ est nul ou pas. Par conséquent, on peut baser un test de H_0 contre H_1 sur une estimation de la longueur $\|M(R\beta - r)\|$ de ce vecteur, ou bien de son carré $\|M(R\beta - r)\|^2 = (R\beta - r)^\top M^\top M(R\beta - r)$.[1] On sait grâce au théorème 5.2 (section 5.3.3), que le meilleur estimateur linéaire et sans biais de $M(R\beta - r)$ est $M(R\hat{\beta} - r)$. Par conséquent, on peut estimer la longueur de $M(R\beta - r)$ par la longueur de son estimateur, c'est à dire par $\|M(R\hat{\beta} - r)\|$, ou par son carré $(R\hat{\beta} - r)^\top M^\top M(R\hat{\beta} - r)$. D'après les propriétés de l'estimateur $\hat{\beta}$, si H_0 est vraie, c'est à dire si la longueur de $M(R\beta - r)$ est nulle, alors il sera probable d'observer de faibles valeurs $(R\hat{\beta} - r)^\top M^\top M(R\hat{\beta} - r)$. Par conséquent le test consistera à rejeter H_0 et accepter H_1 si on observe de grandes valeurs de la variable aléatoire $(R\hat{\beta} - r)^\top M^\top M(R\hat{\beta} - r)$. Un tel raisonnement nous donne la forme des tests qu'on considère ici : ces tests conduisent à rejeter H_0 si on observe un évènement s'écrivant $(R\hat{\beta} - r)^\top M^\top M(R\hat{\beta} - r) > s$, où s est un nombre qui sert à exprimer le fait que la variable aléatoire $(R\hat{\beta} - r)^\top M^\top M(R\hat{\beta} - r)$ prend une "grande valeur". Il reste à expliciter le choix de ce nombre ainsi que celui de la matrice M qui sera utilisée. Ces choix de M et s doivent satisfaire deux conditions :

1. On rappelle (voir la section 5.3.2) que la longueur d'un vecteur a est mesurée par la norme de ce vecteur, définie comme $\sqrt{a^\top a}$.

— M et s doivent être connus de manière à pouvoir dire si l'évènement $(R\hat{\beta}-r)^\top M^\top M(R\hat{\beta}-r) > s$ s'est réalisé ou pas
— on doit pouvoir calculer la probabilité

$$\mathrm{P}_{H_0}\Big((R\hat{\beta}-r)^\top M^\top M(R\hat{\beta}-r) > s\Big)$$

afin de garantir que cette probabilité n'excède pas le niveau qu'on aura choisi pour effectuer le test.

On procède de la manière suivante :

(a) on cherche d'abord une matrice M inversible connue pour laquelle la loi de la variable aléatoire $(R\hat{\beta}-r)^\top M^\top M(R\hat{\beta}-r)$ est connue lorsque H_0 est supposée vraie ;

(b) connaissant cette loi, on choisit $s0$ de sorte que

$$\mathrm{P}_{H_0}\Big((R\hat{\beta}-r)^\top M^\top M(R\hat{\beta}-r) > s\Big) \leq \alpha$$

On commence par noter qu'en utilisant les propriétés 6.1 et 9.8 on obtient

$$R\hat{\beta}-r \sim \mathcal{N}(R\beta-r, \sigma^2 R(X^\top X)^{-1}R^\top)$$

Si H_0 est vraie, alors l'espérance de $R\hat{\beta}-r$ est nulle. De plus, comme R est de rang q, la matrice $R(X^\top X)^{-1}R^\top$ est également de rang q, donc inversible. Définissons alors la variable aléatoire C_1 de la manière suivante :

$$C_1 = (R\hat{\beta}-r)^\top[\sigma^2 R(X^\top X)^{-1}R^\top]^{-1}(R\hat{\beta}-r)$$

On peut alors appliquer la propriété 9.17, et lorsque H_0 est vraie on a $C_1 \sim \chi^2(q)$. On va montrer que C_1 et $\hat{\sigma}^2$ sont indépendantes. Pour cela, on donne une autre expression de C_1 et on applique ensuite la propriété 6.1. Comme $R(X^\top X)^{-1}R^\top$ est symétrique et définie positive, on peut toujours écrire son inverse sous la forme $[R(X^\top X)^{-1}R^\top]^{-1} = A^\top A$ où A est une matrice $q \times q$ inversible. [2] On a donc $C_1 = (R\hat{\beta}-r)^\top \frac{A^\top A}{\sigma^2}(R\hat{\beta}-r)$. En utilisant alors la même démarche que dans la preuve du point 3 de la propriété 6.1, on voit aisément que $A(R\hat{\beta}-r)$ est indépendant de $\hat{\sigma}^2$, et donc C_1 et $(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2}$ le sont également.

Or la propriété 6.1 indique que la variable C_2 définie comme $C_2 = (n-p-1)\frac{\hat{\sigma}^2}{\sigma^2}$ suit une loi $\chi^2(n-p-1)$. Par conséquent, en utilisant la définition 9.5 de la loi de Fisher, si H_0 est vraie, alors

$$\frac{C_1/q}{C_2/(n-p-1)} \sim F(q, n-p-1)$$

Appelons F le rapport ci-dessus. En utilisant les expressions de C_1 et de C_2 on peut écrire

$$F = \frac{(R\hat{\beta}-r)^\top[\hat{\sigma}^2 R(X^\top X)^{-1}R^\top]^{-1}(R\hat{\beta}-r)}{q} \tag{6.3}$$ eq:fisherstat

2. Il suffit d'écrire $R(X^\top X)^{-1}R^\top = K\Lambda^{1/2}\Lambda^{1/2}K^\top$ où K et Λ sont respectivement les matrices des vecteurs propres et valeurs propres de $R(X^\top X)^{-1}R^\top$. On a alors $[R(X^\top X)^{-1}R^\top]^{-1} = K\Lambda^{-1/2}\Lambda^{-1/2}K^\top$ où $\Lambda^{-1/2}$ est la matrice diagonale composée de l'inverse des racines carrées des élément diagonaux de Λ. Il suffit alors de choisir $A = \Lambda^{-1/2}K^\top$.

et si H_0 est vraie, $F \sim F(q, n-p-1)$. On note que F s'écrit sous la forme $F = (R\hat{\beta}-r)^\top M^\top M(R\hat{\beta}-r)$ où $M = \frac{1}{\sqrt{q\hat{\sigma}^2}}A$. On a donc achevé le point (a) de la démarche décrite ci-dessus.

Le point (b) conduit alors à chercher le quantile d'ordre $1-\alpha$ de la loi $F(q, n-p-1)$. On le note $F_{(q,n-p-1);1-\alpha}$, et pour le déterminer, on peut se référer à la section 9.1.3.2 (voir la table à la fin de la section). Autrement dit,

$$\mathrm{P}_{H_0}(F > F_{(q,n-p-1);1-\alpha}) \leq \alpha$$

On a donc obtenu le test recherché.

def:test_fisher

Définition 6.2 *On appelle test de Fisher de $H_0 : R\beta = r$ contre $H_1 : R\beta \neq r$ le test qui consiste à décider H_1 au niveau α lorsqu'on observe $F > F_{(q,n-p-1);1-\alpha}$, où F est la variable aléatoire définie en (6.3). Dans ce contexte, la variable aléatoire F est appelée statistique de Fisher associée à H_0.*

Le test ci-dessus a été construit en utilisant une approche a priori raisonnable schématisée par les points suivants :

— l'hypothèse nulle stipule que tout vecteur de la forme $M(R\beta - r)$, où M est une matrice inversible, est de longueur nulle

— si on constate que le vecteur $M(R\hat{\beta} - r)$ est de trop grande longueur, on décide H_1.

Cette justification n'est pas suffisante et il faut montrer qu'un tel test possède de bonnes propriétés statistiques, notamment une puissance supérieure à celle de n'importe quel autre test de niveau α. Comme dans la section 3.2.3, on est confronté au problème qu'il n'existe pas de test UPP au niveau α pour le problème de tester $H_0 : R\beta = r$ contre $H_1 : R\beta \neq r$. On se restreint alors à chercher le meilleur test dans un ensemble de tests possédant des propriété souhaitées. Ici, ce n'est pas l'absence de biais qui est imposée, mais un principe d'invariance.

Ce principe peut se schématiser de la manière suivante. Supposons qu'on puisse trouver une transformation des variables du modèle qui ne change ni le modèle, ni le problème de test. Dans un tel cas, il est naturel de ne considérer que les tests qui sont invariants par rapport à un telle transformation : puisque le modèle et le problème de test restent inchangés suite à la transformation des variables, il est naturel d'imposer que les tests utilisés doivent donner la même décision, que l'on utilise les variables transformées ou les variables initiales. Les tests possédant cette propriété sont appelés *tests invariants*. On cherche alors le meilleur test parmi les test invariants, *i.e.*, le test invariant de niveau α ayant un risque de type 2 inférieur (ou égal) à celui de n'importe quel autre test invariant de niveau α.

Il est alors possible de montrer que le test de Fisher (définition 6.2) est le meilleur des tests invariants de niveau α.[3] Cette propriété permet donc de justifier l'utilisation de ce test.

À propos de la statistique de Fisher, on peut remarquer qu'elle est construite sur une estimation de $R\beta - r$. En effet, le meilleur estimateur linéaire sans biais de $R\beta - r$ est $R\hat{\beta} - r$ (théorème 5.2) et sa variance est

$$\mathrm{V}(R\hat{\beta} - r) = R\mathrm{V}(\hat{\beta})R^\top = \sigma^2 R(X^\top X)^{-1}R^\top$$

3. L'invariance étant celle par rapport à un certain nombre de transformations, dont par exemple des translations des variables endogènes

où la première égalité résulte de la propriété 9.7 et la deuxième de la propriété 6.1 (point 1). Cette variance est inconnue (car elle dépend de la valeur inconnue de σ) et peut s'estimer par

$$\hat{\mathrm{V}}(R\hat{\beta}-r) = R\hat{\mathrm{V}}(\hat{\beta})R^\top = \hat{\sigma}^2 R(X^\top X)^{-1}R^\top \tag{6.4}$$ eq:hatVRB

On constate alors que la statistique de Fisher s'écrit en fonction de l'estimateur de $R\beta - r$ et de l'estimateur de sa variance :

$$F = \frac{(R\hat{\beta}-r)^\top \left[\hat{\mathrm{V}}(R\hat{\beta}-r)\right]^{-1}(R\hat{\beta}-r)}{q} \tag{6.5}$$ eq:F_stat_V

6.1.3 Le test de Fisher pour des problèmes de test d'un intérêt particulier

On a mentionné dans la section 6.1.1 quelques formes particulières d'intérêt du problème de test général $H_0 : R\beta = r$ contre $H_1 : R\beta \neq r$. Ces cas particuliers sont caractérisés par des valeurs particulières de R et r, qui permettent de déduire de nouvelles expressions pour la statistique de Fisher F. Ces expressions sont obtenues à partir de la forme générale (6.5).

6.1.3.1 Test de nullité simultanée de q paramètres

On a vu en 6.1.1 que ce test correspond au cas où $r = 0_q$ et où chacune des q lignes de R contient un 1 sur l'une de ses colonnes et 0 partout ailleurs. Autrement dit

$$R = \begin{pmatrix} e_{k_1+1}^\top \\ e_{k_2+1}^\top \\ \vdots \\ e_{k_q+1}^\top \end{pmatrix}$$

où e_l est le vecteur de $\mathbb{R}^{p+1}$ dont la l^{e} coordonnée est 1 et les autres sont 0. Dans ce cas, l'égalité $R\beta = r$ s'écrit

$$\begin{pmatrix} \beta_{k_1} \\ \beta_{k_2} \\ \vdots \\ \beta_{k_q} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{6.6}$$ eq:test_qnuls

Cette forme particulière de R et r permet d'obtenir une forme particulière de la statistique de Fisher.

Dans ce cas, on a

$$R\hat{\beta} - r = \begin{pmatrix} e_{k_1+1}^\top \hat{\beta} \\ e_{k_2+1}^\top \hat{\beta} \\ \vdots \\ e_{k_q+1}^\top \hat{\beta} \end{pmatrix} - 0_q = \begin{pmatrix} \hat{\beta}_{k_1} \\ \hat{\beta}_{k_2} \\ \vdots \\ \hat{\beta}_{k_q} \end{pmatrix}$$

Par ailleurs, en utilisant (6.4) et l'expression de R, la matrice $\hat{\mathrm{V}}(R\hat{\beta}-r) = R\hat{\mathrm{V}}(\hat{\beta})R^\top$ a pour $(j,l)^{\mathrm{e}}$ entrée la $(k_j+1, k_l+1)^{\mathrm{e}}$ entrée de $\hat{\mathrm{V}}(\hat{\beta})$. Compte tenu de la structure de cette matrice (voir (5.44)),

cette entrée est $\hat{\text{cov}}(\hat{\beta}_{k_j}, \hat{\beta}_{k_l})$, l'estimateur de la covariance entre $\hat{\beta}_{k_j}$ et $\hat{\beta}_{k_l}$ (voir la section 5.6). En utilisant l'expression (6.5), on peut écrire dans ce cas la statistique F sous la forme :

$$F = \frac{1}{q}\begin{pmatrix}\hat{\beta}_{k_1} & \hat{\beta}_{k_2} & \cdots & \hat{\beta}_{k_q}\end{pmatrix}\begin{pmatrix}\hat{\text{V}}(\hat{\beta}_{k_1}) & \hat{\text{cov}}(\hat{\beta}_{k_1}, \hat{\beta}_{k_2}) & \cdots & \hat{\text{cov}}(\hat{\beta}_{k_1}, \hat{\beta}_{k_q}) \\ \hat{\text{cov}}(\hat{\beta}_{k_2}, \hat{\beta}_{k_1}) & \hat{\text{V}}(\hat{\beta}_{k_2}) & \cdots & \hat{\text{cov}}(\hat{\beta}_{k_2}, \hat{\beta}_{k_q}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\text{cov}}(\hat{\beta}_{k_q}, \hat{\beta}_{k_1}) & \hat{\text{cov}}(\hat{\beta}_{k_q}, \hat{\beta}_{k_2}) & \cdots & \hat{\text{V}}(\hat{\beta}_{k_q})\end{pmatrix}^{-1}\begin{pmatrix}\hat{\beta}_{k_1} \\ \hat{\beta}_{k_2} \\ \vdots \\ \hat{\beta}_{k_q}\end{pmatrix} \tag{6.7}$$

eq:Fstat_qnuls

On peut écrire cette statistique sous une forme plus compacte en désignant par $\beta^R = R\beta$ le sous-vecteur de β dont on teste la nullité, comme exprimé par (6.6). Avec ces notations, on écrit en effet (6.7) comme :

$$F = \frac{1}{q}\hat{\beta}^{R\top}\hat{\text{V}}(\hat{\beta}^R)^{-1}\hat{\beta}^R \tag{6.8}$$

eq:Fstat_qnuls_

où $\hat{\beta}^R = R\hat{\beta}$ est le vecteur des estimateurs des coordonnées de β^R et $\hat{\text{V}}(\hat{\beta}^R) = R\hat{\text{V}}(\hat{\beta})R^\top$ est la matrice de l'expression (6.7) dont les entrées sont les estimateurs des variances et covariances des coordonnées de $\hat{\beta}^R$.

À partir de la forme de la statistique de Fisher donnée par (6.8) pour tester la nullité simultanée de q paramètres, on dérive aisément les expressions de F dans les deux sous-cas évoqués dans la section 6.1.1.

sec:test_signif_ind

6.1.3.2 Test de significativité d'un paramètre

Ce cas correspond au problème de test $H_0 : \beta_k = 0$ contre $H_1 : \beta_k \neq 0$, où β_k est le paramètre d'intérêt dans ce problème. C'est un test de significativité de β_k introduit dans la section 3.2.1. On rappelle que ce type de test permet de décider si la variable exogène X_k a un effet sur la variable endogène.

La matrice R dans ce cas est réduite à une ligne ($q = 1$) et on a $R = e_{k+1}^\top$. Par conséquent $\beta^R = R\beta = \beta_k$ et $\hat{\beta}^R = \hat{\beta}_k$. De plus $\hat{\text{V}}(\hat{\beta}^R) = \hat{\text{V}}(\hat{\beta}_k)$ est la variance estimée de $\hat{\beta}_k$. À partir de (6.8) on a

$$F = \hat{\beta}_k^\top\hat{\text{V}}(\hat{\beta}_k)^{-1}\hat{\beta}_k^\top = \frac{\hat{\beta}_k^2}{\hat{\text{V}}(\hat{\beta}_k)}$$

Si H_0 est vraie, alors F suit une loi de Fisher $F(1, n-p-1)$ et on décide que H_0 est fausse au niveau α si on observe que $F > F_{(1,n-p-1);1-\alpha}$.

Notons que dans le cas du problème de test envisagé ici, la statistique de Fisher est le carré de la statistique

$$T_k = \frac{\hat{\beta}_k}{\sqrt{\hat{\text{V}}(\hat{\beta}_k)}}$$

et par conséquent, le test de Fisher est équivalent à décider H_1 au niveau α si on observe $|T_k| > \sqrt{F_{(1,n-p-1);1-\alpha}}$. Autrement dit, le test peut être effectué de manière équivalente en comparant la valeur absolue de la statistique T_k avec la quantité $\sqrt{F_{(1,n-p-1);1-\alpha}}$. Or en utilisant la définition de la loi de Student (voir la section 9.1.3.3) et la propriété 6.1, on voit que le rapport définissant la statistique T_k suit une loi de Student à $n-p-1$ degrés de libertés. Par ailleurs, en utilisant la

propriété 9.20 (voir les rappels), $\sqrt{F_{(1,n-p-1);1-\alpha}}$ coïncide avec le quantile d'ordre $1-\frac{\alpha}{2}$ de cette loi, *i.e.* $\sqrt{F_{(1,n-p-1);1-\alpha}} = \tau_{n-p-1;1-\frac{\alpha}{2}}$.

Sous cette forme, on voit que le test de Fisher n'est autre que le test de Student qui a été introduit dans le contexte du modèle de régression linéaire simple (voir la section 3.2.2). Le test de Student reste donc valide dans le contexte du modèle standard et garde toutes les propriétés qui ont été établies précédemment (voir la section 3.2.3).

En pratique, pour tester la significativité individuelle d'un paramètre du modèle, on utilise le test sous la forme Student, en comparant la valeur absolue de la statistique de Student T_k avec le quantile approprié.

sec:test_signif_glob

6.1.3.3 Test de significativité globale des paramètres

On s'intéresse maintenant au test de $H_0 : \beta_1 = \cdots = \beta_p = 0$ contre $H_1 : \exists \beta_k \neq 0, k = 1, \ldots, p$. En utilisant les développements de la section 5.5.1.3 qui illustrent une application du théorème de Frish-Waugh, on peut obtenir dans ce cas un résultat intéressant. Il est pour cela important de se rappeler le sens revêtu par H_0, donné à la fin de la section 6.1.1 : si H_0 est vraie, les variables explicatives n'ont aucun pouvoir explicatif sur la variable dépendante. Le résultat dégagé ci-dessous montre établit qu'il un lien entre le test de Fisher et le coefficient de détermination R^2.

Pour le problème de test considéré ici, la matrice R a la forme donnée par (6.2) et la statistique de Fisher est donnée par (6.8), avec $q = p$. On décide que H_0 est fausse au niveau α si on observe $F > F_{(p,n-p-1);1-\alpha}$.

Le sous-vecteur des paramètres sur lesquels porte le problème de test est :

$$\beta^R = R\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Notons que ce sous-vecteur est identique au sous-vecteur β^1 introduit en (5.17) à la section 5.5.1.3. Par conséquent, $\hat{\beta}^R = \hat{\beta}^1$ et en appliquant les résultats de cette section, on sait que

$$\hat{\beta}^R = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y}$$

où $\mathbf{x}$ et $\mathbf{y}$ sont les matrices des observations des variables exogènes et endogènes, prises en différences par rapport à leur moyenne (voir (5.18) et (5.21)). Notamment, on peut écrire

$$\mathbf{y} = (I_n - P_{L_2})\boldsymbol{Y} \qquad \text{et} \qquad \mathbf{x} = (I_n - P_{L_2})X \tag{6.9}$$

eq:XY_centre

où P_{L_2} est la matrice de projection orthogonale sur $L(X_{\cdot 0})$. On peut alors obtenir une expression de $\mathrm{V}(\hat{\beta}^R)$ qui permettra ensuite d'en obtenir une pour $\hat{\mathrm{V}}(\hat{\beta}^R)$. En effectuant un calcul analogue à celui qui a permis d'obtenir $\mathrm{V}(\hat{\beta})$ à la section 5.6, on a :

$$\begin{aligned} \mathrm{V}(\hat{\beta}^R) &= \mathrm{V}[(\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y}] = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathrm{V}(\mathbf{y}) \mathbf{x} (\mathbf{x}^\top \mathbf{x})^{-1} \\ &= (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathrm{V}(M_{L_2} \boldsymbol{Y})\, \mathbf{x} (\mathbf{x}^\top \mathbf{x})^{-1} \\ &= (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top M_{L_2} \mathrm{V}(\boldsymbol{Y}) M_{L_2}^\top \mathbf{x} (\mathbf{x}^\top \mathbf{x})^{-1} \\ &= \sigma^2 (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top M_{L_2} \mathbf{x} (\mathbf{x}^\top \mathbf{x})^{-1} \end{aligned}$$

où M_{L_2} est la matrice définie comme $M_{L_2} = I_n - P_{L_2}$, et où la dernière égalité résulte de l'idempotence et de la symétrie de cette matrice et de la condition C_pN qui donne l'expression de $V(\boldsymbol{Y}) = \sigma^2 I_n$. En utilisant à nouveau l'idempotence de M_{L_2} et la définition de $\mathbf{x}$ (voir (6.9) ci-dessus), on a $M_{L_2}\mathbf{x} = \mathbf{x}$ et par conséquent :

$$V(\hat{\beta}^R) = \sigma^2(\mathbf{x}^\top\mathbf{x})^{-1}$$

On déduit donc $\hat{V}(\hat{\beta}^R) = \hat{\sigma}^2(\mathbf{x}^\top\mathbf{x})^{-1}$. On peut alors obtenir une expression de la statistique de Fisher à partir de celles de $\hat{\beta}^R$ et de $\hat{V}(\hat{\beta}^R)$. Si on utilise l'expression (6.8) de F, on a :

$$\begin{aligned} F &= \frac{1}{p}\mathbf{y}^\top\mathbf{x}(\mathbf{x}^\top\mathbf{x})^{-1}\left[\hat{\sigma}^2(\mathbf{x}^\top\mathbf{x})^{-1}\right]^{-1}(\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top\mathbf{y} \\ &= \frac{1}{p}\frac{(\mathbf{y}^\top\mathbf{x}(\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top)(\mathbf{x}(\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top\mathbf{y})}{\hat{\sigma}^2} \\ &= \frac{\|\hat{\mathbf{y}}\|^2}{p\hat{\sigma}^2} \end{aligned}$$

où $\hat{\mathbf{y}}$ est défini comme dans la section 5.5.1.3 par $\hat{\mathbf{y}} = \mathbf{x}\hat{\beta}^R = \mathbf{x}(\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top\mathbf{y}$. Or d'après la définition de $\hat{\sigma}^2$ et les relations (5.22) et (5.23), on peut écrire :

$$p\hat{\sigma}^2 = \frac{p}{n-p-1}\|\hat{\varepsilon}\|^2 = \frac{p}{n-p-1}(\|\mathbf{y}\|^2 - \|\hat{\mathbf{y}}\|^2)$$

Par conséquent :

$$\frac{1}{F} = \frac{p}{n-p-1}\frac{\|\mathbf{y}\|^2 - \|\hat{\mathbf{y}}\|^2}{\|\hat{\mathbf{y}}\|^2} = \frac{p}{n-p-1}\left(\frac{\|\mathbf{y}\|^2}{\|\hat{\mathbf{y}}\|^2} - 1\right) = \frac{p}{n-p-1}\left(\frac{1}{R^2} - 1\right)$$

où la dernière égalité est obtenue en utilisant l'expression (5.24) du coefficient de détermination de la régression. On a donc obtenu une expression alternative de la statistique de Fisher, qu'on peut résumer par une propriété.

pro:lien_FR2

Propriété 6.2 *Dans le MRLSG, pour tester la significativité globale des paramètres $H_0 : \beta^R = 0_p$ contre $H_1 : \beta^R \neq 0_p$, où $\beta^R = (\beta_1, \ldots, \beta_p)^\top$, la statistique de Fisher s'exprime comme :*

$$F = \frac{n-p-1}{p}\frac{R^2}{1-R^2}$$

où R^2 est le coefficient de détermination.

Ce résultat est intéressant puisqu'il permet de donner une justification supplémentaire au coefficient de détermination. On rappelle que ce dernier estime le pouvoir explicatif des variables exogènes sur la variable endogène. La propriété **??** établit qu'observer $R^2 = 0$ revient à estimer que $\beta^R = 0_p$. Autrement dit, avoir $R^2 = 0$ revient à estimer que H_0 est vraie. Cette propriété de R^2 fournit donc une règle de décision alternative au test basé sur la statistique de Fisher pour choisir entre H_0 et H_1. Cette nouvelle règle s'énonce « *On décide que $H_0 : \beta^R = 0_p$ est fausse si on observe $R^2 > 0$* ». Il est donc naturel que la statistique de Fisher et R^2 soit liés puisqu'ils servent à décider si H_0 est vraie ou pas. Cependant, cette règle de décision fondée sur le coefficient R^2 peut

apparaître trop tranchée dans la mesure où elle prend comme seuil une valeur extrême (la valeur 0) de R^2. De fait, on peut facilement montrer qu'avec un seuil aussi tranché, que H_0 soit vraie ou pas, la probabilité de décider que H_0 est fausse est nulle. Autrement dit le risque de type 1 de ce test est égal à 0, tandis que son risque de type 2 est égal à 1. Ceci peut s'établir à partir du point 2 de la propriété **??** : $R^2 = 0 \iff \hat{\beta}^R = 0$. Par conséquent

$$\mathrm{P}(R^2 = 0) = \mathrm{P}(\hat{\beta}^R = 0_p) \leq \mathrm{P}(\hat{\beta}_1 = 0)$$

puisque l'évènement $\{\hat{\beta}^R = 0_p\}$ implique l'évènement $\{\hat{\beta}_1 = 0\}$. Dans le contexte du MRLSG, que H_0 ou H_1 soit vraie, $\hat{\beta}_1$ est une variable aléatoire qui suit une loi normale, et l'évènement $\{\hat{\beta}_1 = 0\}$ est par conséquent de probabilité nulle. En d'autres termes, le test qui décide que H_0 est fausse lorsqu'on observe $R^2 = 0$ revient à toujours décider que H_0 est vraie.

Si on veut baser le test sur la valeur du coefficient R^2, on peut utiliser une règle de la forme « *On décide H_1 si on observe que R^2 est suffisamment grand (proche de 1)* ». Cela revient à décider que H_0 est fausse dès qu'on estime que le pouvoir explicatif des variables exogènes sur la variable endogène est suffisamment élevée (proche de 100%). L'évènement qui conduit à décider H_1 est alors de la forme $R^2 > s$, où $s \in [0; 1]$ est le seuil à partir duquel on juge que R^2 est suffisamment grand. L'approche de Neyman et Pearson (voir la section 10.3.2.4) requiert alors de choisir s de manière que le risque de type 1 d'un tel test ne dépasse pas un niveau α fixé à l'avance :

$$\mathrm{P}_{H_0}(R^2 > s) \leq \alpha$$

Parmi tous les s satisfaisant cette inégalité, on choisit celui qu'on note s^* pour lequel le risque de type 2 est le plus petit possible. Le risque de type 2 est $\mathrm{P}_{H_1}(R^2 \leq s)$. On en déduit donc que s^* est la plus petite des valeurs s pour lesquelles $\mathrm{P}_{H_0}(R^2 > s) \leq \alpha$ et par conséquent s^* est caractérisé par

$$\mathrm{P}_{H_0}(R^2 > s^*) = \alpha$$

La propriété 6.2 permet de calculer aisément s^*. En effet, en posant $c = \frac{n-p-1}{p}$, on a

$$\begin{aligned}
\alpha &= \mathrm{P}_{H_0}(F > F_{(p,n-p-1);1-\alpha}) \\
&= \mathrm{P}_{H_0}\left(c\frac{R^2}{1-R^2} > F_{(p,n-p-1);1-\alpha}\right) \\
&= \mathrm{P}_{H_0}\left(R^2 > \frac{F_{(p,n-p-1);1-\alpha}}{c + F_{(p,n-p-1);1-\alpha}}\right)
\end{aligned}$$

Donc

$$s^* = \frac{F_{(p,n-p-1);1-\alpha}}{\frac{n-p-1}{p} + F_{(p,n-p-1);1-\alpha}}$$

On peut résumer ce résultat par la propriété suivante.

Propriété 6.3 *Dans le MRLSG, le test de Fisher pour tester la significativité globale des paramètres s'exprime de manière équivalente par « On décide que H_0 est fausse au niveau α si on observe $R^2 > \frac{F_{(p,n-p-1);1-\alpha}}{\frac{n-p-1}{p} + F_{(p,n-p-1);1-\alpha}}$ »*

Ce résultat fournit non seulement une justification supplémentaire du coefficient de détermination, mais il en permet aussi une utilisation un peu plus fine. Jusqu'à présent, on se contentait de dire que lorsque R^2 était proche de 0, on estimait que le pouvoir explicatif des variables exogènes était faible, alors que quand R^2 est proche de 1, ce pouvoir est élevé. Avec la propriété qui lie le coefficient R^2 au test de Fisher, on est capable de donner une signification plus précise de « R^2 est proche de 0 ». On peut par exemple dire que si R^2 est inférieur à $\frac{F_{(p,n-p-1);1-\alpha}}{\frac{n-p-1}{p}+F_{(p,n-p-1);1-\alpha}}$, alors non seulement R^2 est proche de 0, mais suffisamment proche pour qu'on considère que le pouvoir explicatif des variables exogène est nul.

6.1.4 Illustration de la propriété d'invariance du test de Fisher

La propriété d'invariance du test de Fisher peut s'illustrer en montrant un résultat intéressant concernant ce test. On peut montrer que pour tester $R\beta = r$ avec une matrice R et un vecteur r quelconques, il suffit de savoir faire le test dans le cas particulier où $R = I_q$ et $r = 0_q$. Pour obtenir ce résultat, on fait appel successivement à la reparamétrisation présentée à la section 5.5.2 et au théorème de Frish-Waugh (voir la section 5.5.1).

sec:invar_repar

6.1.4.1 Invariance par rapport aux reparamétrisations

On va d'abord montrer qu'on peut se restreindre à des problèmes (6.1) dans lesquels la matrice R a la forme particulière suivante :

$$\left(I_q \,\vdots\, 0_{q,p+1-q} \right)$$

Pour cela, on utilise une reparamétrisation du modèle initial qui laisse le problème de test inchangé.

Considérons le modèle de régression linéaire standard gaussien (définition 6.1) dans lequel on souhaite tester $H_0 : R\beta = r$ contre $H_1 : R\beta \neq r$, où R est une matrice de dimensions $(q, p+1)$ et r un vecteur de $\mathbb{R}^q$. Dans ce problème de test, on peut considérer que le vecteur des paramètres d'intérêt est $R\beta$ et qu'il s'agit de savoir si la valeur de ce vecteur est r ou pas.

En utilisant la démarche de la section 5.5.2, peut trouver une transformation qui permet d'écrire le modèle à l'aide du vecteur $R\beta$. En effet, on a montré dans cette section qu'on peut trouver une matrice Q connue de dimensions $(p+1-q, p+1)$ telle que la matrice

$$A = \left(\begin{array}{c} R \\ \hdashline Q \end{array} \right)$$

est inversible. On peut alors écrire

$$X\beta = XA^{-1}A\beta = Z\delta$$

où $Z = XA^{-1}$ et $\delta = A\beta$. Ces relations entre X et Z, d'une part, et entre β et δ, d'autre part, sont bijectives (puisque A est inversible). Par conséquent, la condition $\mathrm{C}_p\mathrm{N}$ qui définit le modèle de la définition 6.1 est équivalente à la condition suivante :

$$\mathrm{C}_p\mathrm{N}''. \qquad \exists \delta \in \mathbb{R}^{p+1},\ \exists \sigma \in]0, +\infty[, \quad \text{t.q. } \boldsymbol{Y} \sim \mathcal{N}(Z\delta, \sigma^2 I_n)$$

Comme mentionné à la section 5.5.2, définir le modèle de cette manière revient à changer la base de $L(X_{\cdot 0}, \ldots, X_{\cdot p})$, initialement exprimée à l'aide des colonnes de X, en l'exprimant à l'aide des colonnes de Z. Évidemment, les coordonnées de $\mathrm{E}(\boldsymbol{Y})$ dans la base initiale sont $\beta_0, \ldots, \beta_p$, tandis que dans la base constituée des colonnes de Z, ces coordonnées sont $\delta_0, \ldots, \delta_p$, le lien entre les jeux de coordonnées étant donné par $\delta = A\beta$. Ce changement de base revient à transformer les variables explicatives initiales $X_0, X_1, \ldots, X_p$ en nouvelles variables $Z_0, Z_1, \ldots, Z_p$ au moyen de la relation $Z = XA^{-1}$. Cette transformation des variables conduit à transformer les paramètres : les paramètres apparaissant suite à cette transformation sont $\delta_0, \ldots, \delta_p$ et s'obtiennent à partir des paramètres initiaux à l'aide de la relation $\delta = A\beta$. La transformation des variables et des paramètres étant bijective, le modèle initial et le modèle transformé sont identiques. Cette transformation montre simplement qu'on peut, sans changer de modèle, choisir comme on veut les variables qu'on souhaite utiliser (X ou bien Z) et donc la paramétrisation à employer (β ou δ). L'opération qui permet de passer de la formulation du modèle à l'aide de la condition $\mathrm{C}_p\mathrm{N}$, où le paramètre est β, à la formulation $\mathrm{C}_p\mathrm{N}''$ dans laquelle le paramètre est $\delta = A\beta$ s'appelle une *reparamétrisation*.

Puisque les deux conditions $\mathrm{C}_p\mathrm{N}$ et $\mathrm{C}_p\mathrm{N}''$ définissent le même modèle, on doit pouvoir formuler le problème de test de manière équivalente, quelle que soit la condition retenue pour définir le modèle. C'est en effet le cas puisqu'en utilisant la forme de la matrice A, on constate que

$$\delta = A\beta = \left(\begin{array}{c} R \\ \hdashline Q \end{array}\right)\beta = \left(\begin{array}{c} R\beta \\ \hdashline Q\beta \end{array}\right) = \left(\begin{array}{c} \delta^1 \\ \hdashline \delta^2 \end{array}\right) \tag{6.10}$$

eq:repar_delta

où $\delta^1 = R\beta$ et $\delta^2 = Q\beta$ sont des vecteurs de $\mathbb{R}^q$ et $\mathbb{R}^{p+1-q}$ respectivement. En particulier on a $R\beta = r \iff \delta^1 = r \iff D\delta = r$, où D est la matrice définie par

$$D = \left(\begin{array}{c:c} I_q & 0_{q,p+1-q} \end{array}\right) \tag{6.11}$$

eq:test_repar

Le problème de tester $H_0 : R\beta = r$ contre $H_1 : R\beta \neq r$ est donc équivalent au problème de tester $H_0'' : D\delta = r$ contre $H_1'' : D\delta \neq r$.

Puisque les problèmes de test H_0 contre H_1 d'une part, et H_0'' contre H_1'' d'autre part, sont identiques, une bonne propriété pour le test utilisé est qu'il soit invariant par rapport à la reparamétrisation du modèle et donc à l'écriture du problème de test. Autrement dit, qu'on effectue le test de H_0 contre H_1 dans le modèle défini par $\mathrm{C}_p\mathrm{N}$ (et donc une paramétrisation en β) ou bien le test de H_0'' contre H_1'' dans le modèle défini par $\mathrm{C}_p\mathrm{N}''$ (paramétrisation en δ), il est souhaitable que dans chacun des cas le test utilisé conduise à la même décision. On va montrer que le test de Fisher de la définition 6.2 possède cette propriété d'invariance.

Dans le modèle initialement formulé au moyen de $\mathrm{C}_p\mathrm{N}$, on teste H_0 contre H_1 au moyen du test de Fisher de la définition 6.2. En particulier la statistique employée F est donné par l'expression (6.3).

On se place à présent dans le modèle reparamétré défini par $\mathrm{C}_p\mathrm{N}''$, dans lequel on teste $H_0'' : \delta^1 = r$ contre $H_1'' : \delta^1 \neq r$ au moyen du test de Fisher. On va montrer que la statistique de Fisher dans ce cas est la même que celle du modèle initial, et que le test de Fisher dans ce modèle reparamétré est le même que dans le modèle initial. Considérons donc le modèle défini par $\mathrm{C}_p\mathrm{N}''$. Dans ce modèle, le problème de test se formule $H_0'' : D\delta = r$ contre $H_1'' : D\delta \neq r$ et, en utilisant la

définition 6.2 et l'expression (6.3), le test de Fisher consiste à décider que H_0 est fausse au niveau α si on observe

$$F'' = \frac{(D\hat{\delta} - r)^\top [\hat{\sigma}''^2 D(Z^\top Z)^{-1} D^\top]^{-1} (D\hat{\delta} - r)}{q} > F_{(q,n-p-1);1-\alpha} \tag{6.12}$$

eq:Fstat_repar

où $\hat{\delta} = (Z^\top Z)^{-1} Z^\top \boldsymbol{Y}$ est l'estimateur des moindres carrés du vecteur des paramètres δ et $\hat{\sigma}''^2$ est l'estimateur de σ^2 issu de l'estimation du modèle reparamétré, défini par

$$\hat{\sigma}''^2 = \frac{1}{n-p-1} \|\boldsymbol{Y} - Z\hat{\delta}\|^2$$

En utilisant la relation $Z = XA^{-1}$ on peut écrire :

$$\begin{aligned}
\hat{\delta} &= (Z^\top Z)^{-1} Z^\top \boldsymbol{Y} \\
&= (A^{-1^\top} X^\top X A^{-1})^{-1} A^{-1^\top} X^\top \boldsymbol{Y} \\
&= A(X^\top X)^{-1} A^\top A^{-1^\top} X^\top \boldsymbol{Y} \\
&= A(X^\top X)^{-1} X^\top \boldsymbol{Y} \\
&= A\hat{\beta}
\end{aligned}$$

Par conséquent

$$\hat{\sigma}''^2 = \frac{1}{n-p-1} \|\boldsymbol{Y} - XA^{-1}A\hat{\beta}\|^2 = \frac{1}{n-p-1} \|\boldsymbol{Y} - X\hat{\beta}\|^2 = \hat{\sigma}^2 \tag{6.13}$$

eq:sigsec

En utilisant ces expressions ainsi que la relation $Z = XA^{-1}$, on peut écrire la statistique de Fisher F'' dans le modèle reparamétré comme

$$F'' = \frac{(DA\hat{\beta} - r)^\top [\hat{\sigma}^2 DA(X^\top X)^{-1} A^\top D^\top]^{-1} (DA\hat{\beta} - r)}{q}$$

En notant que par construction $DA = R$, on obtient $F'' = F$. Autrement dit, qu'on effectue le test de Fisher dans une paramétrisation ou une autre, on obtiendra la même décision. Ce test est donc invariant par rapport au choix de la paramétrisation.

Ce résultat montre également que quitte à effectuer une reparamétrisation, on peut toujours se ramener à un test de $H_0 : R\beta = r$ contre $H_1 : R\beta = r$ dans lequel la matrice R a la forme de la matrice donnée par le membre de droite de (6.11).

6.1.4.2 Invariance par rapport à des translations

En utilisant le résultat d'invariance par rapport à des reparamétrisation, on montre à présent qu'on peut se limiter à des problèmes de test de type (6.1) dans lesquels le vecteur r est nul.

Puisque le MRLSG peut être défini indifféremment par $\mathrm{C}_p\mathrm{N}$ ou par $\mathrm{C}_p\mathrm{N}''$ et que le test de Fisher est invariant par rapport à la reparamétrisation conduisant à $\mathrm{C}_p\mathrm{N}''$, on choisit cette dernière caractérisation du modèle, dans lequel on teste $H_0'' : D\delta = r$ contre $H_1'' : D\delta \neq r$, où D est la matrice définie par (6.11). En particulier, dans un tel contexte, la relation entre les variables s'écrit $\boldsymbol{Y} = Z\delta + \varepsilon$ et la statistique de Fisher est dans ce cas F'', donnée par (6.12).

Considérons v un vecteur quelconque de $L(Z_{\bullet 0}, Z_{\bullet 1}, \ldots, Z_{\bullet p})$, c'est un dire un élément de $\mathbb{R}^n$ pouvant s'écrire $v = Zc$ pour un c dans $\mathbb{R}^{p+1}$. Le modèle est invariant si on translate le vecteur $\boldsymbol{Y}$ au moyen de v. En effet, la condition $\mathrm{C}_p\mathrm{N}''$ et les résultats sur les vecteurs aléatoires gaussiens (voir la section 9.1.2) impliquent $\boldsymbol{Y} + v = \boldsymbol{Y} + Zc \sim \mathcal{N}(Z\zeta, \sigma^2 I_n)$ pour un certain $\zeta = \delta + c$ et un certain σ^2. Autrement dit, le vecteur translaté $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + v$ satisfait aussi la condition $\mathrm{C}_p\mathrm{N}''$. Le modèle étant invariant par rapport à ce type de translation, on peut travailler indifféremment avec $\boldsymbol{Y}$ ou avec $\tilde{\boldsymbol{Y}}$. Dans le premier cas, le paramètre est δ tandis que dans le second, le paramètre est ζ, avec $\zeta = \delta + c$, et la relation entre les variables dans le second cas est

$$\tilde{\boldsymbol{Y}} = Z\zeta + \varepsilon \tag{6.14}$$ eq:mrlsg_transl

Comme $\zeta = \delta + c$, on a $D\delta = r \iff D\zeta = r + Dc$. Si on choisit de travailler avec la forme translatée du modèle (*i.e.*, avec la variable endogène $\tilde{Y}$ et le vecteur de paramètres ζ), le problème de test est $\tilde{H}_0 : D\zeta = \tilde{r}$ contre $\tilde{H}_1 : D\zeta \neq \tilde{r}$, où $\tilde{r} = r + Dc$. C'est bien un problème de test de la forme générale (6.1) dans un MRLSG, qu'on peut résoudre au moyen du test de Fisher. Dans ce contexte, cette statistique est obtenue à partir de l'expression générale (6.3) :

$$\tilde{F} = \frac{(D\hat{\zeta} - \tilde{r})^\top \left[\hat{\sigma}^2 D(Z^\top Z)^{-1} D^\top\right]^{-1} (D\hat{\zeta} - \tilde{r})}{q} \tag{6.15}$$ eq:Fstat_transl

où les différents estimateurs sont donnés par les formules usuelles

$$\hat{\zeta} = (Z^\top Z)^{-1} Z^\top \tilde{\boldsymbol{Y}}$$
$$\hat{\sigma}^2 = \frac{\|\tilde{\boldsymbol{Y}} - \hat{\tilde{\boldsymbol{Y}}}\|^2}{n - p - 1}$$

où $\hat{\tilde{\boldsymbol{Y}}}$ est le vecteur des valeurs ajustées dans le modèle défini par la relation (6.14), *i.e.*, $\hat{\tilde{\boldsymbol{Y}}} = Z\hat{\zeta}$.

Puisque ce qui précède est vrai pour tout vecteur $c \in \mathbb{R}^{p+1}$, on peut choisir ce vecteur de manière que $\tilde{r} = 0_q$, de manière que le problème de test à résoudre soit $\tilde{H}_0 : D\zeta = 0_q$ contre $\tilde{H}_1 : D\zeta \neq 0_q$. Il faut pour cela choisir c tel que $Dc = -r$, et en utilisant la forme particulière de D, cela revient à avoir

$$c = \begin{pmatrix} -r \\ \hdashline 0_{p+1-q} \end{pmatrix} \tag{6.16}$$ eq:c_transl

Dans ce cas particulier, on obtient l'expression de la statistique de Fisher $\tilde{F}$ à partir de celles de $\hat{\zeta}$ et $\|\tilde{\boldsymbol{Y}} - \hat{\tilde{\boldsymbol{Y}}}\|^2$. On a

$$\hat{\zeta} = (Z^\top Z)^{-1} Z^\top (\boldsymbol{Y} + Zc) = \hat{\delta} + c$$

Par conséquent,

$$D\hat{\zeta} - \tilde{r} = D\hat{\zeta} = D\hat{\delta} + Dc = D\hat{\delta} - r$$

puisque c est choisi de sorte que $\tilde{r} = 0_q$ et qu'avec un tel choix on a $Dc = -r$. Par ailleurs,

$$\hat{\tilde{\boldsymbol{Y}}} = Z\hat{\zeta} = Z\hat{\delta} + Zc$$

Donc

$$\|\tilde{\boldsymbol{Y}} - \hat{\tilde{\boldsymbol{Y}}}\|^2 = \|\boldsymbol{Y} + Zc - Z\hat{\delta} - Zc\|^2 = \|\boldsymbol{Y} - Z\hat{\delta}\|^2$$

d'où on obtient

$$\frac{\|\tilde{\boldsymbol{Y}} - \hat{\tilde{\boldsymbol{Y}}}\|^2}{n-p-1} = \hat{\sigma}''^2$$

où $\hat{\sigma}''^2$ est l'estimateur sans de σ^2 obtenu dans le modèle défini par $\mathrm{C}_p\mathrm{N}''$ (voir (6.13)). À partir de ces éléments, la statistique de Fisher $\tilde{F}$ donnée par (6.15) peut s'exprimer comme :

$$\tilde{F} = \frac{(D\hat{\delta} - r)^\top \left[\hat{\sigma}''^2 D(Z^\top Z)^{-1} D^\top\right]^{-1} (D\hat{\delta} - r)}{q}$$

On constate alors que la statistique $\tilde{F}$ dans le modèle où on effectue une translation de $\boldsymbol{Y}$ est identique à la statistique de Fisher F'' obtenue dans le modèle initial (voir l'expression (6.12) pour F''). Autrement dit, quitte à effectuer une translation de la variable endogène, on peut toujours se ramener au cas d'un problème de test général de type (6.1) dans lequel r est le vecteur nul 0_q. Pour cela, la translation s'effectue au moyen du vecteur Zc où c est donné par (6.16).

Pour terminer cette section, remarquons qu'étant donnée la forme particulière de c, le vecteur Zc à l'aide duquel on effectue la translation est $-Z^1 r$, où Z^1 est la sous-matrice de Z constituée de ses q premières colonnes :

$$Z = \left(Z^1 \,\vdots\, Z^2\right) \tag{6.17}$$ eq:partZ

où Z^1 et Z^2 sont de tailles respectives (n, q) et $(n, p+1-q)$. Comme dans le modèle initial la relation entre les variables est $\boldsymbol{Y} = Z\delta + \varepsilon$, la translation par le vecteur $-Z^1 r$ qui laisse le modèle invariant peut se voir de la manière suivante :

$$\boldsymbol{Y} = Z\delta + \varepsilon \iff \boldsymbol{Y} = Z^1\delta^1 + Z^2\delta^2 + \varepsilon \iff \boldsymbol{Y} - Z^1 r = Z^1(\delta^1 - r) + Z^2\delta^2 + \varepsilon$$

Dans la dernière égalité le membre de gauche est $\tilde{\boldsymbol{Y}}$ et celui de droite peut s'écrire sous la forme $Z\zeta + \varepsilon$ où

$$\zeta = \begin{pmatrix} \delta^1 - r \\ \hdashline \delta^2 \end{pmatrix}$$

ce qui est bien la relation $\zeta = \delta + c$ (voir le partitionnement de δ donné en (6.10)). En définissant $\zeta^1 = \delta^1 - r$ et $\zeta^2 = \delta^2$, la relation du modèle peut s'écrire

$$\tilde{\boldsymbol{Y}} = Z^1\zeta^1 + Z^2\zeta^2 + \varepsilon \tag{6.18}$$ eq:mrlsg_trans

Compte tenu de la forme de la matrice D définissant le problème de test, on voit alors que $D\delta = r \iff \delta^1 = r \iff \zeta^1 = 0_q \iff D\zeta = 0_q$.

sec:invar_proj

6.1.4.3 Transformation par projection

On va finalement montrer que n'importe quel problème de test de type (6.1) peut se ramener à un problème dans lequel la matrice R est I_q et $r = 0_q$.

Le problème à résoudre est toujours celui de tester $H_0 : R\beta = r$ contre $H_1 : R\beta \neq r$ dans le modèle défini par la condition $\mathrm{C}_p\mathrm{N}$. Les sections précédentes montrent qu'on peut, de manière équivalente, choisir de se placer dans le modèle « translaté » défini par la condition $\widetilde{\mathrm{C}_p\mathrm{N}}$ suivante :

$$\tilde{\boldsymbol{Y}} \sim \mathcal{N}(Z\zeta, \sigma^2 I_n)$$

et tester dans ce modèle $\tilde{H}_0 : D\zeta = 0_q$ contre $\tilde{H}_1 : D\zeta \neq 0_q$. On rappelle que la relation entre le modèle et le problème de test initiaux, d'une part, et le modèle et le problème de test dans lesquels on se place, d'autre part, est donnée par

$$Z = XA^{-1}, \quad \zeta = \begin{pmatrix} \delta^1 - r \\ \hdashline \delta^2 \end{pmatrix}, \quad \delta = A\beta, \quad \tilde{\boldsymbol{Y}} = \boldsymbol{Y} - Z^1 r \quad \text{et} \quad D = \left(I_q \,\vdots\, 0_{q,p+1-q} \right)$$

Le test s'effectue au moyen de la statistique de Fisher $\tilde{F}$ dont l'expression est donnée par (6.15) et qui dans le contexte étudiée est égale à

$$\tilde{F} = \frac{\hat{\zeta}^{1\top} \left[\hat{\sigma}^2 D (Z^\top Z)^{-1} D^\top \right]^{-1} \hat{\zeta}^1}{q} \tag{6.19}$$

eq:Fstat_trans

où $\hat{\sigma}^2 = \frac{\|\tilde{\boldsymbol{Y}} - Z\hat{\zeta}\|^2}{n-p-1}$, et où $\hat{\zeta}^1 = D\hat{\zeta}$ est l'estimateur des moindres carrés de ζ^1, formé en sélectionnant les q premières lignes de $\hat{\zeta} = (Z^\top Z)^{-1} Z^\top \tilde{\boldsymbol{Y}}$.

Dans le modèle défini par $\widetilde{\mathrm{C}_p\mathrm{N}}$, notons que la relation entre les variables endogène $\tilde{\boldsymbol{Y}}$ et exogènes Z s'écrit comme en (6.18) et qu'étant donnée la forme de la matrice D qui définit le problème de test, ce dernier ne porte que sur un sous-vecteur ζ^1 de ζ. En examinant la forme de la statistique de Fisher $\tilde{F}$, on constate qu'elle dépend des données à travers $\hat{\zeta}^1$ et $\hat{\sigma}^2$. L'idée consiste alors à se demander si en appliquant le test de Fisher à partir d'un modèle transformé dans lequel l'estimation de ζ^1 et σ^2 coïncide avec $\hat{\zeta}^1$ et $\hat{\sigma}^2$, la décision obtenue reste la même.

On a vu dans la section 5.5.1.2 qu'en projetant une partie des variables sur un espace généré par les autres, on obtenait un modèle dans lequel l'application de la méthode des moindres carrés permettait d'obtenir les mêmes estimateurs des paramètres attachés aux variables projetées et le même estimateur de σ^2 que dans le modèle initial. C'est le résultat établi par le théorème de Frish-Waugh (théorème 5.4) et interprété dans la remarque 5.18. En utilisant les résultats de la section 5.5.1.2, on va transformer le modèle initial défini par $\widetilde{\mathrm{C}_p\mathrm{N}}$ de manière à ne conserver que la partie δ_1 du vecteur des paramètres δ concernée par les hypothèses du problème de test.

Soit L_2 le sev de $\mathbb{R}^n$ engendré par les $p+1-q$ colonnes de la matrice Z^2 et P_{L_2} la matrice de projection orthogonale sur L_2. On procède alors comme dans la section 5.5.1.2 et on écrit à partir de la relation (6.18) :

$$\begin{aligned} (I_n - P_{L_2})\tilde{\boldsymbol{Y}} &= (I_n - P_{L_2})Z^1\zeta^1 + (I_n - P_{L_2})Z^2\zeta^2 + (I_n - P_{L_2})\varepsilon \\ &= (I_n - P_{L_2})Z^1\zeta^1 + (I_n - P_{L_2})\varepsilon \end{aligned}$$

où la deuxième égalité utilise le point 2 de la propriété 9.22. En définissant $\boldsymbol{Y}_* = (I_n - P_{L_2})\tilde{\boldsymbol{Y}}$, $Z_* = (I_n - P_{L_2})Z^1$, $\varepsilon_* = (I_n - P_{L_2})\varepsilon$ et $\gamma = \zeta^1$, la relation (6.18) du modèle implique que :

$$\boldsymbol{Y}_* = Z_*\gamma + \varepsilon_* \tag{6.20}$$

eq:mrlsg_proj

Comme le suggère la remarque 5.18, on peut considérer que cette nouvelle relation permet de définir un modèle de régression linéaire dans lequel la variable endogène est Y_* et les variables exogènes sont les colonnes de Z_*. Ce modèle n'est pas un modèle de régression linéaire standard puisque chacun des termes d'erreurs (qui sont ici les coordonnées du vecteur $\varepsilon_* = (I_n - P_{L_2})\varepsilon$) sont

des combinaisons linéaires des termes d'erreur $\varepsilon_1, \ldots, \varepsilon_n$ du modèle initial. Ces nouveaux termes d'erreur sont donc corrélés entre eux et la matrice des variances-covariances du vecteur ε_* n'est pas diagonale. Autrement dit, le modèle obtenu à partir de la relation (6.20) ne satisfait pas la condition C_p3.

Cependant, le théorème 5.4 établit que $\hat{\zeta}^1 = \hat{\gamma}$ et $\hat{\sigma}^2 = \hat{\sigma}^2_*$ où $\hat{\gamma}$ et $\hat{\sigma}^2_*$ sont les estimateurs obtenus en appliquant les moindres carrés à la relation (6.20), et donc définis par

$$\begin{aligned} \hat{\gamma} &= (Z_*^\top Z_*)^{-1} Z_*^\top \boldsymbol{Y}_* \\ \hat{\sigma}^2_* &= \frac{1}{n-p+1} \|\hat{\varepsilon}_*\|^2 \end{aligned} \tag{6.21}$$

eq:sig_star

avec $\hat{\varepsilon}_* = \boldsymbol{Y}_* - Z_* \hat{\gamma}$. Autrement dit, bien que le modèle défini à partir de la relation (6.20) ne soit pas un MRLS, il permet d'obtenir à partir des moindres carrés les mêmes estimateurs de ζ^1 et de σ^2 que le modèle initial défini par la condition $\widetilde{\mathrm{C}_p\mathrm{N}}$. Par ailleurs, comme $\boldsymbol{Y}_*$ est un vecteur aléatoire gaussien, ce modèle est gaussien. Puisque son vecteur de paramètres coïncide avec celui sur lequel porte le problème de test et que ces paramètres peuvent s'estimer par moindres carrés, on peut tenter d'appliquer dans le contexte de ce modèle le test de Fisher pour tester l'hypothèse $H_0^* : \gamma = 0_q$ contre $H_1^* : \gamma \neq 0_q$. Puisque $\gamma = \zeta^1$, ce problème de test est équivalent au problème de départ $\tilde{H}_0 : \zeta^1 = 0_q$ contre $\tilde{H}_1 : \zeta^1 \neq 0_q$.

Si on veut tester $H_0^* : \gamma = 0_q$ contre $H_1^* : \gamma \neq 0_q$ à l'aide test de Fisher dans le modèle défini par (6.20), on formule les hypothèses à tester sous la forme générale de la section 6.1.1 : $H_0^* : G\gamma = g$ et $H_1^* : G\gamma \neq g$ où évidemment $G = I_q$ et $g = 0_q$ et d'après (6.3), la statistique de Fisher pour ce problème de test est

$$\begin{aligned} F_* &= \frac{(G\hat{\gamma} - g)^\top [\hat{\sigma}^2_* G (Z_*^\top Z_*)^{-1} G^\top]^{-1} (G\hat{\gamma} - g)}{q} \\ &= \frac{\hat{\gamma}^\top [\hat{\sigma}^2_* (Z_*^\top Z_*)^{-1}]^{-1} \hat{\gamma}}{q} \\ &= \frac{\hat{\zeta}^{1\top} [\hat{\sigma}^2 (Z_*^\top Z_*)^{-1}]^{-1} \hat{\zeta}^1}{q} \end{aligned}$$

où la deuxième expression utilise les formes particulières de G et g, et la troisième les résultats $\hat{\gamma} = \hat{\zeta}^1$ et $\hat{\sigma}^2_* = \hat{\sigma}^2$ obtenus précédemment.

Grâce à des résultats sur l'inversion de matrices par blocs (voir ci-dessous), on peut montrer que

$$(Z_*^\top Z_*)^{-1} = D (Z^\top Z)^{-1} D^\top \tag{6.22}$$

eq:blocinv

Avec cette égalité, la statistique de Fisher s'écrit

$$F_* = \frac{\hat{\zeta}^{1\top} [\hat{\sigma}^2 D (Z^\top Z)^{-1} D^\top]^{-1} \hat{\zeta}^1}{q}$$

On constate donc que $F_* = \tilde{F}$, où $\tilde{F}$ est la statistique de Fisher (donnée par (6.19)) pour tester $\tilde{H}_0$ contre $\tilde{H}_1$. Autrement dit, le test de Fisher effectué dans le contexte du modèle défini par la relation (6.20) est identique au test de Fisher effectué dans le modèle initial.

On termine cette section en montrant comment on obtient l'égalité (6.22). Pour cela, on commence par utiliser le partitionnement de Z donné par (6.17), pour écrire

$$Z^\top Z = \begin{pmatrix} Z^{1\top} \\ \hdashline Z^{2\top} \end{pmatrix} \left(Z^1 \vdots Z^2 \right) = \left(\begin{array}{c:c} Z^{1\top}Z^1 & Z^{1\top}Z^2 \\ \hdashline Z^{2\top}Z^1 & Z^{2\top}Z^2 \end{array} \right)$$

La matrice $(Z^\top Z)^{-1}$ est partionnée en blocs de mêmes tailles que $Z^\top Z$ et on peut l'écrire

$$(Z^\top Z)^{-1} = \left(\begin{array}{c:c} z^{11} & z^{12} \\ \hdashline z^{21} & z^{22} \end{array} \right)$$

À partir de ces partitionnements de $Z^\top Z$ et de $(Z^\top Z)^{-1}$, on peut montrer que[4]

$$(Z^\top Z)(Z^\top Z)^{-1} = I_{p+1} \implies z^{11} = \left[Z^{1\top}Z^1 - Z^{1\top}Z^2(Z^{2\top}Z^2)^{-1}Z^{2\top}Z^1 \right]^{-1}$$

On constate alors que le membre de droite s'écrit

$$\left[Z^{1\top}(I_n - Z^2(Z^{2\top}Z^2)^{-1}Z^{2\top})Z^1 \right]^{-1} = \left[Z^{1\top}(I_n - P_{L_2})Z^1 \right]^{-1} = (Z_*^\top Z_*)^{-1}$$

par définition de Z_* et par idemptotence et symétrie de $I_n - P_{L_2}$. On peut donc écrire :

$$(Z^\top Z)^{-1} = \left(\begin{array}{c:c} (Z_*^\top Z_*)^{-1} & z^{12} \\ \hdashline z^{21} & z^{22} \end{array} \right)$$

On vérifie alors à partir de cette égalité que si D est la matrice définie par (6.11), on a

$$D(Z^\top Z)^{-1}D^\top = (Z_*^\top Z_*)^{-1}$$

6.1.4.4 Utilisation combinée des propriétés de la statistique de Fisher et illustration numérique

Les propriétés d'invariance du test de Fisher démontrées dans les sections précédentes permettent, à travers des étapes de transformation d'un modèle initial, d'écrire le problème de test de départ $H_0 : R\beta = r$ contre $H_1 : R\beta \neq r$ sous la forme d'un problème plus simple $H_0^* : \gamma = 0_q$ contre $H_1^* : \gamma \neq 0_q$, où γ est le vecteur des paramètres d'un modèle de régression. Ces étapes sont résumées ci-dessous :

4. Pour cela, réécrit l'égalité $(Z^\top Z)(Z^\top Z)^{-1} = I_{p+1}$ en exprimant le produit dans son membre de gauche à l'aide des blocs de $(Z^\top Z)$ et de $(Z^\top Z)^{-1}$ donnés ci-dessus. On résout alors l'égalité en identifiant les blocs de ce produit avec les blocs correspondant de la matrice I_{p+1}.

	Modèle	Problème de test
Modèle initial :	$\boldsymbol{Y} = X\beta + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$	$H_0 : R\beta = r,\ H_1 : R\beta \neq r$

Reparamétrisation : on définit $Z = XA^{-1}$ et $\delta = A\beta$ avec $A = \begin{pmatrix} R \\ \hdashline Q \end{pmatrix}$

	Modèle	Problème de test
Étape 1 :	$\boldsymbol{Y} = Z\delta + \varepsilon = Z^1\delta^1 + Z^2\delta^2 + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$	$H_0'' : D\delta = r,\ H_1'' : D\delta \neq r$ où $D = \left(I_q \,\vdots\, 0_{q,p+1-q} \right)$

Translation : On translate $\boldsymbol{Y}$ par le vecteur $-Z^1 r$ pour former $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - Z^1 r$

	Modèle	Problème de test
Étape 2 :	$\tilde{\boldsymbol{Y}} = Z^1(\delta^1 - r) + Z^2\delta^2 + \varepsilon$ $= Z^1\zeta^1 + Z^2\zeta^2 + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$	$\tilde{H}_0 : D\zeta = 0_q,\ \tilde{H}_1 : D\zeta \neq 0_q$

Projection : On utilise les restes de la projection orthogonale des variables sur L_2 :
$\boldsymbol{Y}_* = (I - P_{L_2})\tilde{\boldsymbol{Y}}$ et $Z_* = (I - P_{L_2})Z^1$

Pour appliquer les propriétés du test de Fisher ainsi résumées, supposons qu'on dispose d'un programme informatique à qui on fournit en entrée les observations des variables ainsi que la matrice R et le vecteur r, et qui en retour affiche la valeur de la statistique F pour tester, à partir des observations fournies, H_0 contre H_1, où les hypothèses sont définies par R et r (voir (6.1)). Plus précisément, on dispose dans un langage informatique d'une fonction appelée `FISHER` qui agit de la manière suivante :

$$(\boldsymbol{Y}, X, R, r) \xrightarrow{\texttt{FISHER}} F$$

où comme d'habitude, $\boldsymbol{Y}$ est le vecteur des observations de la variable endogène et X la matrice des observations des variables exogènes, et F est la statistique de Fisher définie par (6.3). La fonction `FISHER` doit notamment calculer l'estimation des moindres carrés $(X^\top X)^{-1}X^\top \boldsymbol{Y}$ du paramètre de la relation entre les variables endogène et exogènes, ainsi que l'estimation de la variance de la variable endogène $\frac{1}{n-(p+1)}\|\boldsymbol{Y} - X(X^\top X)^{-1}X^\top \boldsymbol{Y}\|^2$. Notons que pour cette dernière, la fonction `FISHER` doit aussi extraire les dimensions n et $(p+1)$ de la matrice X. Finalement, afin de déterminer le dénominateur de la statistique F dans son expression (6.3), cette fonction doit aussi extraire la dimension q du vecteur r.

Dans le modèle initial, les données sont dans $\boldsymbol{Y}$ et X, et la matrice R et le vecteur r définissent le problème de test. La valeur de F est donc retournée par `FISHER`$(\boldsymbol{Y}, X, R, r)$.

Pour obtenir la statistique de Fisher dans le modèle reparamétré de l'étape 1, on transforme les variables exogènes au moyen de la matrice A pour former la matrice Z des observations des variables exogènes transformées au moyen de la relation $Z = XA^{-1}$, la matrice A étant construite comme dans la section 5.5.2. Autrement dit, il faut trouver Q tel que

$$A = \begin{pmatrix} R \\ \hdashline Q \end{pmatrix}$$

est inversible. Une manière de former une telle matrice Q consiste à choisir ses lignes comme des vecteurs d'une base du noyau de R. En tant que vecteurs d'une base, ces lignes de Q seront linéairement indépendantes. En tant que vecteurs appartenant au noyau de R, ces lignes seront orthogonales à celles de R. Par conséquent, la famille des vecteurs lignes de R et de Q est libre. Donc la matrice A formée en empilant les lignes de R et de Q est inversible. Dans le modèle reparamétré, le problème de test est défini par la matrice D et le vecteur r. Par conséquent, la valeur de la statistique de Fisher F'' sera retournée par $\texttt{FISHER}(\boldsymbol{Y}, XA^{-1}, D, r)$

Si on souhaite effectuer le test dans le modèle de l'étape 2, il faut translater le vecteur $\boldsymbol{Y}$ au moyen du vecteur $-Z^1 r$, où Z^1 est la sous-matrice de Z, constituée de ses q premières colonnes. Notons que $Z^1 = ZD^\top = XA^{-1}D^\top$. Donc $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - XA^{-1}D^\top r$. Par ailleurs, dans ce modèle translaté, le problème de test est défini par la matrice D et le vecteur 0_q. Donc la valeur de la statistique de Fisher $\tilde{F}$ sera retournée par $\texttt{FISHER}(\boldsymbol{Y} - XA^{-1}D^\top r, XA^{-1}, D, 0_q)$.

Finalement, dans l'étape 3, il faut utiliser la différence entre les observations de chaque variable et leurs projections orthogonales sur l'espace engendré par les colonnes de Z^2. Ces colonnes sont les $p+1-q$ colonnes de $Z = XA^{-1}$, qu'il faut donc extraire pour former $M_{L_2} = I_n - Z^2(Z^{2^\top} Z^2)^{-1} Z^{2^\top}$ puis former les nouvelles observations des variables endogène translatée et exogènes. On aura donc

$$\boldsymbol{Y}_* = M_{L_2}\tilde{\boldsymbol{Y}} = M_2(\boldsymbol{Y} - XA^{-1}D^\top r) \qquad \text{et} \qquad Z_* = M_{L_2}Z^1 = M_2 XA^{-1}D^\top$$

Dans ce modèle projeté, le problème de test est défini par le matrice I_q et le vecteur 0_q et il faudrait donc calculer la statistique de Fisher F_* en appelant

$$\texttt{FISHER}(M_2(\boldsymbol{Y} - XA^{-1}D^\top r), M_2 XA^{-1}D^\top, I_q, 0_q)$$

Cependant, comme on l'a noté ci-dessus, la fonction `FISHER` calcule l'estimation de la variance de la variable endogène en utilisant les dimensions (nombre de lignes, nombre de colonnes) de la matrice contenant les observations des variables exogènes au moyen de la formule

$$\frac{\text{carré de la norme du vecteur des résidus}}{\text{nombre de lignes} - \text{nombre de colonnes}}$$

Dans le modèle projeté, cette matrice est celle qui est fournie en deuxième argument de la fonction `FISHER`. Elle est donc égale à $Z_* = M_2 XA^{-1}D^\top$ et ses dimensions sont (n, q). Par conséquent, la fonction `FISHER` estimera la variance de la variable endogène en divisant la somme des carrés des résidus par $n - q$. Or comme le montre l'estimateur de cette variance (6.21), le dénominateur doit être $n - p - 1$. Par conséquent, si on veut calculer dans ce contexte la statistique de Fisher F_* au moyen de la fonction `FISHER`, il faut appliquer le facteur correctif $\frac{n-p-1}{n-q}$:

$$F_* = \frac{n-p+1}{n-q}\texttt{FISHER}(M_2(\boldsymbol{Y} - XA^{-1}D^\top r), M_2 XA^{-1}D^\top, I_q, 0_q)$$

6.1.5 Autres expressions de la statistique de Fisher et interprétations du test

Dans la section 5.5.2, on a montré comment obtenir un estimateur de β lorsqu'on imposait les contraintes qui définissent H_0 dans le problème de test initial défini par (6.1). Il est donc naturel de chercher à utiliser ce type d'estimation pour construire un test de H_0 contre H_1. On va voir qu'une démarche de ce genre permet d'aboutir au test de Fisher sous une forme alternative, et d'en obtenir une autre interprétation.

6.1.5.1 Expression fondée sur la distance entre les estimateurs contraint β^* et non contraint $\hat{\beta}$

Dans la section 5.5.2 on a montré comment obtenir un estimateur de β lorsque ce vecteur de paramètres était contraint par l'égalité $R\beta = r$. Une manière de construire un test de H_0 contre H_1 consiste à examiner si l'estimation obtenue en imposant cette contrainte diffère beaucoup de celle où aucune contrainte n'est imposée. Si la contrainte est vraie, alors l'estimateur non contraint $\hat{\beta}$ devrait satisfaire approximativement cette contrainte (*i.e.*, $R\hat{\beta}$ devrait être proche de r). Par conséquent, puisque de son côté l'estimateur contraint β^* satisfait également la contrainte (par construction), on devrait dans ce cas noter peu de différence entre l'estimation non contrainte donnée par $\hat{\beta}$ et l'estimation contrainte donnée par β^*. Autrement dit, si la contrainte est satisfaite (*i.e.*, si H_0 est vraie), l'imposer ou ne pas l'imposer lors de l'estimation de β ne devrait pas changer grand chose à l'estimation obtenue.

Un critère pour décider si H_0 est vraie (la contrainte $R\beta = r$ est satisfaite) ou pas peut donc être construit à partir de la différence $\hat{\beta} - \hat{\beta}^*$: compte-tenu de ce qui vient d'être noté, observer une différence $\hat{\beta} - \hat{\beta}^*$ trop importante est peu vraisemblable si H_0 est vraie, et donc en pareil cas on rejettera H_0 et on décidera H_1.

Formellement, pour mesurer la différence entre $\hat{\beta}$ et β^*, on peut utiliser n'importe quelle mesure de distance entre ces deux vecteurs.[5] Le principe du test consiste donc à rejeter H_0 si la distance entre ces vecteurs dépasse un certain seuil.

On rappelle la relation (5.41) qui établit le lien entre l'estimateur contraint β^* de β et l'estimateur non contraint $\hat{\beta}$:

$$\beta^* = \hat{\beta} - (X^\top X)^{-1} R^\top \left[R(X^\top X)^{-1} R^\top\right]^{-1} (R\hat{\beta} - r) \tag{6.23}$$ eq:rel_est

On en déduit facilement que

$$(\hat{\beta} - \beta^*)^\top (X^\top X)(\hat{\beta} - \beta^*) = (R\hat{\beta} - r)^\top \left[R(X^\top X)^{-1} R^\top\right]^{-1} (R\hat{\beta} - r)$$

On constate alors que la statistique F donnée par (6.3) s'écrit

$$F = (\hat{\beta} - \beta^*)^\top (X^\top X)(\hat{\beta} - \beta^*) \times \frac{1}{q\hat{\sigma}^2} \tag{6.24}$$ eq:Fstat_cnc

5. La norme de la différence $\|\hat{\beta} - \hat{\beta}^*\|$ est une possibilité, mais ce n'est pas la seule. On peut vérifier que pour toute matrice M de rang $(p+1)$, la fonction qui associe à 2 vecteurs u_1 et u_2 le nombre $\|M(u_1 - u_2)\|$ est également une distance : cette fonction est non-négative, nulle si et seulement si $u_1 = u_2$, elle est symétrique et vérifie l'inégalité triangulaire.

Cette expression de la statistique du test de Fisher permet donc de lui associer l'interprétation donnée au début de cette section. En effet, on voit que la statistique s'écrit $F = \|M(\hat{\beta} - \beta^*)\|^2$ avec $M = \frac{1}{\sqrt{q}\hat{\sigma}}X$, et apparaît donc comme une mesure de la distance entre $\hat{\beta}$ et β^*. Le test de Fisher, qui conduit au rejet de H_0 si on observe que F dépasse un certain quantile, correspond bien à la forme recherchée ici : on décide de rejeter H_0 si la distance entre les estimations contrainte et non contrainte dépasse un certain seuil.

6.1.5.2 Expression fondée sur la distance entre les valeurs ajustées issues de l'estimation contrainte et de l'estimation non-contrainte

Notons que la statistique F exprimé sous la forme (6.24) est proportionnelle à

$$(\hat{\beta} - \beta^*)^\top (X^\top X)(\hat{\beta} - \beta^*) = \|X\hat{\beta} - X\beta^*\|^2 \tag{6.25}$$ eq:dist_va

qui est (le carré de) la distance entre les vecteurs $X\hat{\beta}$ et $X\beta^*$. Le premier de ces vecteurs est évidemment le vecteur des valeurs ajustées issues de l'estimation non contrainte de β. Puisque β^* est l'estimateur de β lorsque la contrainte est imposée, on peut considérer le second vecteur $X\beta^*$ comme l'estimation de $X\beta$ lorsqu'on suppose que la contrainte est vraie. Le vecteur $X\beta^*$ peut donc s'interpréter comme le vecteur des valeurs ajustées issues de l'estimation contrainte. Avec cette interprétation, on peut voir le test de Fisher comme basé sur un principe de comparaison entre les valeurs ajustées obtenues avec et sans la contrainte. Si la contrainte est vraie, c'est à dire si β satisfait l'égalité $R\beta = r$, alors le fait de ne pas imposer cette contrainte, naturellement satisfaite, lors de l'estimation, ou bien l'imposer explicitement ne devrait pas avoir beaucoup d'impact sur l'estimation obtenue. Autrement dit, si la contrainte est vraie, on peut s'attendre à ce que les valeurs ajustées contraintes et non contraintes soient proches l'une de l'autre. Sur un tel principe, on devrait donc mesurer la distance entre les deux vecteurs de valeurs ajustées $X\hat{\beta}$ et $X\beta^*$, puis décider que H_1 est vraie dès que cette distance est trop importante. Avec l'égalité (6.25), on voit à partir de l'expression de F donnée par (6.24) qu'on a

$$F = \frac{1}{q\hat{\sigma}^2}\|X\hat{\beta} - X\beta^*\|^2$$

et donc que F est d'autant plus grande que la distance entre les valeurs ajustées est importante. Le principe qui vient d'être avancé doit alors conduire à décider H_1 si la valeur observée de F est trop grande, ce qui est exactement ce à quoi conduit le test de Fisher.

6.1.5.3 Expression fondée sur la distance entre les résidus des estimation contrainte et non contrainte

Au lieu, comme on vient de le faire, de raisonner sur les vecteurs des valeurs ajustées $X\hat{\beta}$ et $X\beta^*$, on peut raisonner sur les vecteurs de résidus de l'estimation contrainte et non contrainte. Ces vecteurs sont respectivement $\varepsilon^* = \boldsymbol{Y} - X\beta^*$ et évidemment $\hat{\varepsilon} = \boldsymbol{Y} - X\hat{\beta}$. Le raisonnement est exactement le même que dans le paragraphe précédent : si la contrainte $R\beta = r$ est vraie, alors imposer explicitement la contrainte ou ne pas l'imposer ne doit pas beaucoup modifier l'estimation obtenue. Dans un tel cas, les vecteurs des résidus obtenus dans chacun des deux cas doivent être

proche l'un de l'autre. Si on observe que cela se produit, alors on décide que la contrainte n'est pas vraie et on décide H_1. Le test sera donc fondé sur la distance entre ε^* et $\hat{\varepsilon}$, qu'on peut mesurer par $\|\varepsilon^* - \hat{\varepsilon}\|^2$.

Pour montrer que le test F est basé sur ce principe, notons que d'après (6.23) et la définition de ε^*, on a

$$\begin{aligned}\varepsilon^* &= \boldsymbol{Y} - X\hat{\beta} + X(X^\top X)^{-1}R^\top\left[R(X^\top X)^{-1}R^\top\right]^{-1}(R\hat{\beta} - r)\\ &= \hat{\varepsilon} + X(X^\top X)^{-1}R^\top\left[R(X^\top X)^{-1}R^\top\right]^{-1}(R\hat{\beta} - r) \qquad (6.26)\end{aligned}$$

eq:rel_rescnc

Donc $\varepsilon^* - \hat{\varepsilon} = X(X^\top X)^{-1}R^\top\left[R(X^\top X)^{-1}R^\top\right]^{-1}(R\hat{\beta} - r)$ et

$$\begin{aligned}\|\varepsilon^* - \hat{\varepsilon}\|^2 &= (\varepsilon^* - \hat{\varepsilon})^\top(\varepsilon^* - \hat{\varepsilon})\\ &= (R\hat{\beta} - r)^\top\left[R(X^\top X)^{-1}R^\top\right]^{-1}R(X^\top X)^{-1}X^\top X(X^\top X)^{-1}R^\top\left[R(X^\top X)^{-1}R^\top\right]^{-1}(R\hat{\beta} - r)\\ &= (R\hat{\beta} - r)^\top\left[R(X^\top X)^{-1}R^\top\right]^{-1}(R\hat{\beta} - r)\end{aligned}$$

On voit alors d'après (6.24) que F s'écrit

$$F = \frac{1}{q\hat{\sigma}^2}\|\varepsilon^* - \hat{\varepsilon}\|^2$$

F est donc d'autant plus grande que la distance entre les vecteurs de résidus est importante. Le test de Fisher peut donc bien s'interpréter comme étant basé sur le principe décrit ci-dessus.

On peut noter ici que comme $X^\top\hat{\varepsilon} = 0_{p+1}$ (voir la remarque 5.14 et propriété juste après), on déduit de (6.26) que $\hat{\varepsilon}^\top\varepsilon^* = \hat{\varepsilon}^\top\hat{\varepsilon}$. Par conséquent

$$\|\varepsilon^* - \hat{\varepsilon}\|^2 = (\varepsilon^* - \hat{\varepsilon})^\top(\varepsilon^* - \hat{\varepsilon}) = {\varepsilon^*}^\top\varepsilon^* + \hat{\varepsilon}^\top\hat{\varepsilon} - 2\hat{\varepsilon}^\top\varepsilon^* = {\varepsilon^*}^\top\varepsilon^* - \hat{\varepsilon}^\top\hat{\varepsilon} = \|\varepsilon^*\|^2 - \|\hat{\varepsilon}\|^2$$

Par ailleurs, par définition on a $\hat{\sigma}^2 = \frac{1}{n-p-1}\|\hat{\varepsilon}\|^2$. On peut donc écrire

$$F = \frac{n-p-1}{q}\,\frac{\|\varepsilon^*\|^2 - \|\hat{\varepsilon}\|^2}{\|\hat{\varepsilon}\|^2}$$

6.1.5.4 Expression fondée sur la longueur du vecteur des multiplicateurs de Lagrange de l'estimation contrainte

Pour donner une dernière interprétation du test de Fisher, on utilise toujours le principe que lorsque H_0 est vraie, *i.e.*, la contrainte $R\beta = r$ est satisfaite, alors l'imposition explicite de cette contrainte dans l'estimation de β ne devrait pas avoir beaucoup d'impact, relativement à l'estimation non contrainte. D'après la section 5.5.2, l'impact, ou le « poids », de la contrainte dans l'estimation contrainte peut se mesurer au moyen de λ^*, le vecteur des multiplicateurs de Lagrange $\lambda_1^*, \ldots, \lambda_q^*$ associés aux q contraintes exprimées par l'égalité $R\beta = r$. Si H_0 est vraie, alors la contrainte ne devrait pas impacter l'estimation et les multiplicateurs devraient être proches de 0, ou encore, le vecteur λ^* devrait être proche de 0_q, le vecteur nul de $\mathbb{R}^q$.

On note que pour toute matrice M de dimensions (q, q) inversible, $\lambda^* = 0_q$ si et seulement si $M\lambda^* = 0_q$. Par conséquent, dire que λ^* est proche de 0_q revient à dire que le vecteur $M\lambda^*$ est

proche de 0. Cette proximité peut se mesurer au moyen de $\|M\lambda^*\|^2$. Comme $R(X^\top X)^{-1}R^\top$ est une matrice symétrique et définie positive (puisque R est supposé de rang q et X de rang $p+1$), on peut toujours trouver une matrice M telle que

$$M^\top M = \frac{1}{4q\hat{\sigma}^2} R(X^\top X)^{-1}R^\top$$

Avec une telle matrice, en utilisant l'expression (5.40) obtenue pour λ^*, on constate que

$$\begin{aligned}\|M\lambda^*\|^2 &= {\lambda^*}^\top M^\top M\lambda^* \\ &= \frac{4}{4q\hat{\sigma}^2}(R\hat{\beta}-r)^\top\left[R(X^\top X)^{-1}R^\top\right]^{-1}R(X^\top X)^{-1}R^\top\left[R(X^\top X)^{-1}R^\top\right]^{-1}(R\hat{\beta}-r) \\ &= F\end{aligned}$$

Ceci établit que la statistique de Fisher est une mesure de la distance entre le vecteur λ^* et 0_q : F est d'autant plus grande que cette distance est grande. Par conséquent, le test de Fisher peut effectivement s'interpréter comme basé sur un principe qui consiste à décider que H_0 est fausse si on constate que le poids des contraintes dans l'estimation contrainte est trop important.

sec:mrls-rc

6.2 Régions de confiance pour β

On aborde à présent la question de la construction de régions de confiance pour le vecteur des paramètres β. Le problème et la démarches sont exposés dans un cadre général à la section 10.3.3.

Ici, le contexte est le même que pour celui des tests d'hypothèses : on se place dans le cadre du modèle de régression linéaire standard gaussien. L'introduction de l'hypothèse de normalité pour la loi du vecteur $\boldsymbol{Y}$ (ou du vecteur des termes d'erreur) est un moyen qui permet de répondre à nécessité d'évaluer la probabilité que la région recherchée contienne la vraie valeur du vecteur des paramètres β. Dans un tel contexte, on commence par chercher une région $\mathscr{C}_n$ de $\mathbb{R}^{p+1}$, construite à partir des observations des variables du modèle, et telle que pour un $\alpha \in]0,1[$ fixé

$$\mathrm{P}_\beta(\beta \in \mathscr{C}_n) \geq 1-\alpha, \qquad \forall \beta \in \mathbb{R}^{p+1} \tag{6.27}$$

eq:rc_beta

où la notation P_β indique que la probabilité est calculée pour une valeur donnée du vecteur de paramètres β. En effet, $\mathscr{C}_n$ étant une fonction de $\boldsymbol{Y}$, la probabilité de l'évènement $\beta \in \mathscr{C}_n$ s'obtient à partir de la loi de $\boldsymbol{Y}$. La condition $\mathrm{C}_p\mathrm{N}$ montre explicitement que cette loi dépend de β. Par conséquent la probabilité de l'évènement $\beta \in \mathscr{C}_n$ dépendra de la valeur de β utilisée pour former la loi de $\boldsymbol{Y}$. La région $\mathscr{C}_n$ est appelée région de confiance de niveau $1-\alpha$ pour β.

Comme mentionné à la section 10.3.3, une condition telle que (6.27) ne permet pas de déterminer une unique région $\mathscr{C}_n$. Deux régions satisfaisant cette condition sont comparées au moyen d'un critère de précision, introduit par la définition 10.2. La précision d'une région de confiance est semblable au risque de type 2 d'un test. En effet, une région est d'autant plus précise que la probabilité qu'elle contienne une mauvaise valeur du paramètre est faible. Par comparaison, le risque de type 2 est une probabilité de choisir la mauvaise hypothèse. Grâce à la dualité existant entre tests et régions de confiance établie par le théorème 10.2, cette analogie entre précision d'une région de confiance et risque de type 2 d'un test est formalisée par le corollaire 10.1. Ce résultat

établit que rechercher la région de confiance de niveau $1-\alpha$ pour β la plus précise revient à rechercher le test de niveau α le plus puissant pour tester l'hypothèse $\beta = b$ contre l'hypothèse $\beta \neq b$, où b est un vecteur quelconque de $\mathbb{R}^{p+1}$.

L'égalité $\beta = b$ s'écrit $R\beta = r$ avec $R = I_{p+1}$ et $r = b$. D'après les résultats de la section 6.1 le test de niveau α le plus puissant est le test de Fisher. [6] Par conséquent, la région de confiance $\mathscr{C}_n$ associée à ce test est la région de confiance de niveau $1-\alpha$ pour β la plus précise. Pour expliciter $\mathscr{C}_n$, on utilise le théorème 10.2, qui permet de construire une région de confiance à partir d'une famille de tests.

Ici pour expliciter la famille de tests, on commence par construire une famille de problèmes de tests, chacun étant défini par un couple d'hypothèses portant sur le vecteurs des paramètres β. Ces problèmes seront indexés par $b \in \mathbb{R}^{p+1}$: pour chaque b, on définit les hypothèses $H_0(b) : \beta = b$ et $H_1(b) : \beta \neq b$ qui définissent un problème de test. En utilisant les résultats de la section 6.1, le test de Fisher pour tester $H_0(b)$ contre $H_1(b)$ au niveau α s'effectue en utilisant la statistique de Fisher qu'on notera ici $F(b)$. En utilisant (6.3) avec $R = I_{p+1}$ et $r = b$, on obtient

$$\begin{aligned} F(b) &= \frac{1}{q}(\hat{\beta} - b)^\top \left[\hat{\sigma}^2 (X^\top X)^{-1}\right]^{-1} (\hat{\beta} - b) \\ &= \frac{1}{q\hat{\sigma}^2} (\hat{\beta} - b)^\top (X^\top X)(\hat{\beta} - b) \end{aligned}$$

On décide donc que $H_0(b)$ est vraie au niveau α si l'évènement $F(b) \leq F_{(q,n-p-1);1-\alpha}$ se réalise.

En faisant ceci pour chaque $b \in \mathbb{R}^{p+1}$, donc pour chaque problème de test de la famille considérée, on peut construire la famille de tests de Fisher de niveau α, indexée par b et notée $\{\varphi_b \mid b \in \mathbb{R}^{p+1}\}$, où $\varphi_b = 0$ si $F(b) \leq F_{(q,n-p-1);1-\alpha}$ et $\varphi_b = 1$ sinon.

En utilisant le théorème 10.2, on peut alors associer une région de confiance de niveau $1-\alpha$ à cette famille de tests. Cette région est $\mathscr{C}_n = \{b \in \mathbb{R}^{p+1} \mid \varphi_b = 0\}$. Puisque $\varphi_b = 0 \iff F(b) \leq F_{(q,n-p-1);1-\alpha}$ on a

$$\mathscr{C}_n = \{b \in \mathbb{R}^{p+1} \mid F(b) \leq F_{(q,n-p-1);1-\alpha}\}$$

Si on utilise l'expression de $F(b)$, on peut aussi écrire

$$\mathscr{C}_n = \{b \in \mathbb{R}^{p+1} \mid \frac{1}{q\hat{\sigma}^2} (\hat{\beta} - b)^\top (X^\top X)(\hat{\beta} - b) \leq F_{(q,n-p-1);1-\alpha}\}$$

Grâce aux résultats de la section 6.1 et au théorème 10.2, la même démarche peut être suivie pour obtenir une région de confiance pour n'importe quel ensemble de q combinaisons linéaires des paramètres $\beta_0, \ldots, \beta_p$. Plus précisément, au lieu de s'intéresser aux coordonnées de β, on s'intéresse à q paramètres $\delta_1, \ldots, \delta_q$, chacun étant défini comme une combinaison linéaire donnée de $\beta_0, \ldots, \beta_p$.

6. C'est le test le plus puissant parmi ceux qui restent invariants par rapport à certaines transformations des variables.

Autrement dit les paramètres d'intérêt sont définis par

$$\begin{aligned}\delta_1 &= R_{10}\beta_0 + \cdots + R_{1p}\beta_p \\ \delta_2 &= R_{20}\beta_0 + \cdots + R_{2p}\beta_p \\ &\vdots \\ \delta_q &= R_{q0}\beta_0 + \cdots + R_{qp}\beta_p\end{aligned}$$

où les R_{kl} sont des nombres connus. On souhaite à présent construire une région de confiance au niveau $1-\alpha$ pour le vecteur de paramètres δ, où

$$\delta = \begin{pmatrix}\delta_1 \\ \delta_2 \\ \vdots \\ \delta_q\end{pmatrix}$$

Chapitre 7

chap:asy

Propriétés asymptotiques des moindres carrés

sec:asy_intro

7.1 Introduction

Dans les chapitres précédents, on a obtenu les propriétés de $\hat{\beta}$ sous les seules conditions C_p1 à C_p3. Notamment, ces conditions suffisent à montrer que $\hat{\beta}$ est l'estimateur linéaire sans biais le plus précis. En ce qui concerne les tests, on a eu besoin de rajouter la condition de normalité du vecteur $\boldsymbol{Y}$ afin de pouvoir évaluer les risques liés aux tests, et choisir des tests optimaux. Dans ce cas, il faut noter que les risques des tests sont connus de manière exacte. De même, lors de la construction de région de confiance, la probabilité pour que la région obtenue contienne la valeur (inconnue) de β est parfaitement connue (égale au niveau de confiance voulu).

La condition de normalité de $\boldsymbol{Y}$ n'est pas toujours réaliste. Dans de telles situations, on peut éventuellement spécifier une autre loi pour $\boldsymbol{Y}$ et essayer d'obtenir des tests optimaux pour les problèmes de test considérés.

Une autre manière de procéder très utilisée consiste à ne pas spécifier de loi particulière pour $\boldsymbol{Y}$, mais à imposer des conditions sur cette loi, ainsi que sur le comportement des variables exogènes, telles qu'il soit possible, lorsque la taille de l'échantillon tend vers l'infini, d'établir les lois limite de certaines statistiques. Ces lois limite sont appelées *lois asymptotiques* et l'approche reposant sur l'utilisation de ces loi est dite *asymptotique*.

Ainsi, si aucun type particulier de loi pour $\boldsymbol{Y}$ (par exemple une loi normale) n'est spécifié à l'avance, les lois des statistiques servant à la construction de tests et régions de confiance sont en général inconnues. Par conséquent, il est en général impossible de calculer les probabilités intervenant dans la construction d'un test (probabilité d'erreur de type 1) ou d'une région de confiance (probabilité que la région contienne la valeur du paramètre). Autrement dit, il est dans ce cas impossible de construire un test ou une région de confiance ayant le niveau (de risque de type 1 ou de confiance) voulu.

Cependant, lorsque la taille de l'échantillon disponible est suffisamment grande et que le modèle satisfait certaines conditions, on peut montrer que des statistiques permettant la construction de tests et de régions de confiance ont des lois asymptotiques connues. Dans ce cas, on peut

considérer que les probabilités calculées à partir de ces lois limite sont de bonnes approximations des probabilités qu'on pourrait calculer si on connaissait la véritable loi des statistiques utilisées. Dès lors, les contraintes de niveau imposées lors de la construction d'un test ou d'une région de confiance sont introduites sur les lois asymptotiques. Par exemple, si on souhaite tester l'hypothèse $H_0 : R\beta = r$ au niveau α en utilisant l'approche décrite à la section 6.1, la contrainte de niveau impose qu'on doit avoir $\mathrm{P}_{H_0}(F > s) \leq \alpha$, et le choix d'un test de plus forte puissance conduit à choisir $s = F_{q,n-p-1;1-\alpha}$ lorsqu'on suppose que $\boldsymbol{Y}$ est un vecteur gaussien. Si on lève cette supposition, la probabilité $\mathrm{P}_{H_0}(F > s)$ est en général inconnue, mais sous des conditions adéquates, on peut montrer que la statistique F possède une loi limite lorsque H_0 est vraie : la probabilité $\mathrm{P}_{H_0}(F > s)$ converge vers $1 - G_\infty(s)$ lorsque H_0 est supposée vraie, où G_∞ désigne la fonction de répartition de la loi asymptotique de F. Dans ce cas, l'approche asymptotique du test consiste à écrire la contrainte de niveau en utilisant non pas la loi inconnue de F, mais sa loi asymptotique. Ainsi, au lieu de chercher s en imposant $\mathrm{P}_{H_0}(F > s) \leq \alpha$, on imposera $1 - G_\infty(s) \leq \alpha$. L'idée est évidemment que, pourvu que n soit grand, $\mathrm{P}_{H_0}(F > s)$ et $1 - G_\infty(s)$ sont proches l'un de l'autre, et que par conséquent on peut penser qu'en imposant $1 - G_\infty(s) \leq \alpha$ on aura aussi $\mathrm{P}_{H_0}(F > s) \leq \alpha$. Cette dernière inégalité est évidemment celle qu'on souhaite avoir pour que le test soit de niveau α.

Dans les sections suivantes, on commence par obtenir des distributions asymptotiques permettant d'approximer celles de $\hat{\beta}$ et de $\hat{\sigma}$. On explicite ensuite la manière d'utiliser ces approximations afin de construire des tests et des régions de confiance pour β.

Les résultats de convergence présentés ci-dessous s'obtiennent lorsque $n \to \infty$. On étudie alors le comportement limite de suites indexées par la taille de l'échantillon et on devrait faire apparaître cette indexation dans la notation. Ainsi on devrait noter $\boldsymbol{Y}_n$, X_n, $\hat{\beta}_n = (X_n^\top X_n)^{-1} X_n \boldsymbol{Y}_n$, etc, les vecteurs et matrices formés lorsque la taille de l'échantillon est n. Cependant, afin d'alléger la notation, on s'abstient d'expliciter cette indexation ; il faudra cependant la garder à l'esprit pour bien interpréter les résultats et les conditions sous lesquelles on les obtient.

7.2 Propriétés asymptotiques de $\hat{\beta}$

7.2.1 Convergence de $\hat{\beta}$

Dans cette section, on étudie les conditions sous lesquelles la suite des estimateurs de β obtenus pour des tailles d'échantillon de plus en plus grandes converge vers la valeur du paramètre.

Intuitivement, on peut faire le raisonnement suivant. L'espérance de $\hat{\beta}_k$ est β_k. De manière générale, la variance mesure la variabilité d'une v.a. autour de son espérance. Donc si on arrive à établir que la variance de $\hat{\beta}_k$ converge vers 0 lorsque $n \to \infty$, alors on sera amené à dire que lorsque la taille de l'échantillon est très grande, $\hat{\beta}_k$ ne varie presque pas autour de β_k. Et à la limite ($n = \infty$), $\hat{\beta}_k$ ne varie plus du tout autour de β_k, *i.e.* $\hat{\beta}_k = \beta_k$.

En formalisant un peu ce raisonnement, on voudrait montrer que la limite de $\mathrm{V}(\hat{\beta}_k) = \mathrm{E}((\hat{\beta}_k - \beta_k)^2)$ est 0. On peut exprimer cette variance à l'aide de $\hat{\beta} - \beta$ puisque $\hat{\beta}_k - \beta_k = a_k^\top(\hat{\beta} - \beta)$ où a_k est le vecteur de $\mathbb{R}^{p+1}$ dont la $(k+1)^{\mathrm{e}}$ coordonnée est 1 est les p autres sont 0. On a alors

$$\mathrm{E}((\hat{\beta}_k - \beta_k)^2) = \mathrm{E}(a_k^\top(\hat{\beta} - \beta)^\top(\hat{\beta} - \beta)a_k) = a_k^\top \mathrm{E}(\|\hat{\beta} - \beta\|^2)a_k$$

Donc une condition suffisante pour que $\forall k = 0 \ldots, p$, $\mathrm{V}(\hat{\beta}_k) \to 0$ lorsque $n \to \infty$, est que $\mathrm{E}(\|\hat{\beta} - \beta\|^2) \to 0$, $n \to \infty$. C'est une condition également suffisante pour que $\mathrm{V}(a^\top \hat{\beta})$ converge vers 0 pour tout vecteur a de $\mathbb{R}^{p+1}$. Autrement dit lorsque $n \to \infty$, si $\mathrm{E}(\|\hat{\beta} - \beta\|^2) \to 0$, alors la variance de toute combinaison linéaire de $\hat{\beta}_0, \ldots, \hat{\beta}_p$ converge vers 0.

Le mode de convergence utilisé ici est la convergence en moyenne quadratique. On rappelle qu'une suite $\{Z_n\}$ de variables aléatoires converge en moyenne quadratique vers $z \in \mathbb{R}$ lorsque $\mathrm{E}((Z_n - z)^2)$ converge vers 0. La discussion ci-dessus montre donc que si $\mathrm{E}(\|\hat{\beta} - \beta\|^2) \to 0$, alors $a^\top \hat{\beta}$ converge en moyenne quadratique vers $a^\top \beta$, pour tout $a \in \mathbb{R}^{p+1}$; en particulier l'estimateur $\hat{\beta}_k$ de chaque paramètre β_k converge vers β_k en moyenne quadratique. Dans un tel cas, on dit que le vecteur $\hat{\beta}$ lui-même converge en moyenne quadratique vers $\hat{\beta}$.

Le résultat ci-dessous donne une condition suffisante pour que $\mathrm{E}(\|\hat{\beta} - \beta\|^2) \to 0$.

pro:betaconsist

Propriété 7.1 *Si lorsque $n \to \infty$, $(X^\top X)^{-1} \to 0$ alors $\mathrm{E}(\|\hat{\beta} - \beta\|^2) \to 0$ et donc $\hat{\beta}$ converge en moyenne quadratique vers β.*

Preuve : Notons que $\|\hat{\beta} - \beta\|^2 = (\hat{\beta} - \beta)^\top (\hat{\beta} - \beta) = \mathrm{trace}((\hat{\beta} - \beta)^\top (\hat{\beta} - \beta)) = \mathrm{trace}((\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top)$. Donc

$$\begin{aligned}
\mathrm{E}(\|\hat{\beta} - \beta\|^2) &= \mathrm{E}(\mathrm{trace}((\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top)) \\
&= \mathrm{trace}(\mathrm{E}((\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top) \\
&= \mathrm{trace}(\mathrm{V}(\hat{\beta})) \\
&= \sigma^2 \mathrm{trace}((X^\top X)^{-1}) \\
&= \sigma^2 \sum_{k=0}^{p} \zeta_i
\end{aligned}$$

où $\zeta_0, \ldots, \zeta_p$ sont les $p+1$ valeurs propres de $(X^\top X)^{-1}$. Comme cette matrice est définie positive, ses valeurs propres sont toutes strictement positives. Remarquons alors que la condition $(X^\top X)^{-1} \to 0$ équivaut à $\|(X^\top X)^{-1}\| = \max\{\zeta_0, \ldots, \zeta_p\} \to 0$ ou encore à $\zeta_k \to 0$, $k = 0, \ldots, p$. On a alors lorsque $n \to \infty$

$$\mathrm{E}(\|\hat{\beta} - \beta\|^2) \to 0 \iff \sum_{k=0}^{p} \zeta_i \to 0 \iff \|(X^\top X)^{-1}\| \to 0 \iff (X^\top X)^{-1} \to 0$$

■

Remarque 7.1

1. La convergence en moyenne quadratique implique la convergence en probabilité. Donc la condition $(X^\top X)^{-1} \to 0$ implique aussi que $\hat{\beta}$ converge en probabilité vers β.
2. Dans de nombreux ouvrages d'économétrie, on établit cette convergence sous la condition que $\frac{1}{n} X^\top X \to Q$ où Q est une matrice définie positive. Cette condition est plus forte que la condition $(X^\top X)^{-1} \to 0$ de la propriété 7.1 et n'est utilisée que pour obtenir une convergence en probabilité, plus faible que celle de la propriété 7.1 (qui établit une convergence en moyenne quadratique).

Montrons d'abord que la condition $\frac{1}{n}X^\top X \to Q$ avec Q définie positive implique $(X^\top X)^{-1} \to 0$. Pour cela, on note que pour n suffisamment grand, $\frac{1}{n}X^\top X$ doit être inversible puisqu'elle converge vers une matrice inversible. Par ailleurs, d'après le théorème 9.1, on doit avoir $\left(\frac{1}{n}X^\top X\right)^{-1} \to Q^{-1}$, et donc $n|||(X^\top X)^{-1}||| \to |||Q^{-1}|||$. Pour que cette convergence ait lieu, il est nécessaire que $|||(X^\top X)^{-1}||| \to 0$.

On voit ensuite que la condition usuelle $\frac{1}{n}X^\top X \to Q$ peut être trop forte puisqu'il suffit d'avoir $\frac{1}{n^\delta}X^\top X \to Q$ pour un $\delta > 0$. En effet, le même raisonnement qu'au-dessus montre que si $\frac{1}{n^\delta}X^\top X \to Q$, alors $n^\delta|||(X^\top X)^{-1}||| \to |||Q^{-1}|||$, ce qui implique $|||(X^\top X)^{-1}||| \to 0$. ☒

7.2.2 Normalité asymtptotique de $\hat{\beta}$

On s'intéresse à présent à des conditions qui établissent une loi limite pour une certaine fonction de $\hat{\beta}$. On rappelle le résultat classique en probabilité/statistique appelé théorème « central limit ». Dans sa version la plus simple, ce théorème établit que la moyenne arithmétique de n variables aléatoires indépendantes, identiquement distribuées, dont la variance (commune) existe, converge en loi après centrage et réduction vers une variable aléatoire normale centrée réduite, lorsque $n \to \infty$. Plus formellement lorsque $n \to \infty$, $\mathrm{V}(\overline{X})^{-1/2}[\overline{X} - \mathrm{E}(\overline{X})] \to Z$ en loi, où $Z \sim N(0,1)$ et $\overline{X} = \sum_{i=1}^n X_i/n$. En notant μ l'espérance commune des variables aléatoires $X_1, X_2, \ldots$, on peut aussi écrire $\mathrm{V}(\overline{X})^{-1/2}(\overline{X} - \mu) \to Z$.

On souhaite ici établir un résultat semblable pour la suite des estimateurs des moindres carrés. Pour cela, il faut noter plusieurs caractéristiques du contexte dans lequel on se place, qui nous conduiront à choisir une version adéquate du thérème « central limit », dont la version la plus simple évoquée ci-dessus n'est pas adaptée à l'objectif visé.

Premièrement, puisqu'on s'intéresse à une suite de vecteurs aléatoires, il faut une définition de la convergence en loi pour de telles suites. On ne donne pas une telle définition, mais on utilise un résultat qui caractérise cette convergence.

Deuxièmement, on verra que $\hat{\beta}$ apparaît comme une somme (et donc comme une moyenne) de vecteurs aléatoires qui ne sont pas identiquement distribués. Notamment leurs variances ne sont pas égales. Il faut donc un théorème « central limit » qui permet cette hétérogénéité.

On présente dans la section suivante les résultats de base sur lesquels on s'appuiera pour obtenir une convergence en loi de $\hat{\beta}$.

sec:clt

7.2.2.1 Convergence en loi de suites aléatoires : résultats de base

Le premier résultat permet de caractériser la convergence en loi pour des suites de vecteurs aléatoires.

th:cw

Théorème 7.1 (Cramér-Wold) *Soit $\{Z_n : n \geq 1\}$ une suite de vecteurs aléatoires de $\mathbb{R}^q$. Cette suite converge en loi vers le vecteur aléatoire Z de $\mathbb{R}^q$ si et seulement si pour tout $c \in \mathbb{R}^q$, la suite de variables aléatoires $\{c^\top Z_n : n \geq 1\}$ converge en loi vers la variable aléatoire $c^\top Z$.*

On constate que grâce à ce résultat, l'étude de la convergence en loi de suites de vecteurs aléatoires se fait en se ramenant au cas univarié.

Le second résultat permet d'obtenir une convergence en loi de suites aléatoires de lois hétérogènes (variances non égales par exemple).

th:cltl **Théorème 7.2 (Lindeberg)** *Soit* $\{Z_i, i = 1, 2, \ldots\}$ *une suite de variables aléatoires indépendantes, et telles que pour tout* i, $\mathrm{E}(Z_i) = 0$ *et* $\mathrm{V}(Z_i)$ *existe avec* $\mathrm{V}(Z_i) > 0$. *Alors si la condition*

$$\forall \eta > 0 \qquad \sum_{i=1}^{n} \mathrm{E}\Big(\mathbb{1}_{|\frac{Z_i}{V_n}|>\eta} \frac{Z_i^2}{V_n^2}\Big) \to 0, \qquad n \to \infty$$

est satisfaite, $\sum_{i=1}^{n} \frac{Z_i}{V_n}$ *converge en loi vers* $N(0,1)$, *où* $V_n^2 = \sum_{i=1}^{n} \mathrm{V}(Z_i)$.

Pour une preuve de ce théorème, voir l'ouvrage *Calcul des probabilités* de D. Foata et A. Fuchs (page 241 et suivantes, dans l'édition de 1996).

Remarque 7.2 La condition de convergence de la somme des espérances dans le théorème est appelée condition de Lindeberg.
Cette condition implique notamment que $\max_{i=1,\ldots,n} \dfrac{\mathrm{V}(Z_i)}{V_n^2} \to 0$ lorsque $n \to \infty$. En effet, pour tout n et tout $\eta > 0$ on a

$$\begin{aligned} Z_i^2 = \mathbb{1}_{|\frac{Z_i}{V_n}|>\eta} Z_i^2 + \mathbb{1}_{|\frac{Z_i}{V_n}|\leq\eta} Z_i^2 &= \mathbb{1}_{|\frac{Z_i}{V_n}|>\eta} Z_i^2 + \mathbb{1}_{Z_i^2 \leq \eta^2 V_n^2} Z_i^2 \\ &\leq \mathbb{1}_{|\frac{Z_i}{V_n}|>\eta} Z_i^2 + \eta^2 V_n^2 \end{aligned}$$

pour tout $i = 1, \ldots, n$. Donc pour tout n et tout $\eta > 0$:

$$\mathrm{V}(Z_i) = \mathrm{E}(Z_i^2) \leq \mathrm{E}\Big(\mathbb{1}_{|\frac{Z_i}{V_n}|>\eta} Z_i^2\Big) + \eta^2 V_n^2 \qquad \forall i = 1, \ldots, n$$

et donc

$$\frac{\mathrm{V}(Z_i)}{V_n^2} \leq \mathrm{E}\Big(\mathbb{1}_{|\frac{Z_i}{V_n}|>\eta} \frac{Z_i^2}{V_n^2}\Big) + \eta^2 \leq \sum_{i=1}^{n} \mathrm{E}\Big(\mathbb{1}_{|\frac{Z_i}{V_n}|>\eta} \frac{Z_i^2}{V_n^2}\Big) + \eta^2 \qquad \forall i = 1, \ldots, n$$

Par conséquent,

$$\max_{i=1,\ldots,n} \frac{\mathrm{V}(Z_i)}{V_n^2} \leq \sum_{i=1}^{n} \mathrm{E}\Big(\mathbb{1}_{|\frac{Z_i}{V_n}|>\eta} \frac{Z_i^2}{V_n^2}\Big) + \eta^2$$

Comme ceci est vrai pour tout n, on a

$$\lim_{n\to\infty} \max_{i=1,\ldots,n} \frac{\mathrm{V}(Z_i)}{V_n^2} \leq \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E}\Big(\mathbb{1}_{|\frac{Z_i}{V_n}|>\eta} \frac{Z_i^2}{V_n^2}\Big) + \eta^2 = \eta^2$$

où l'égalité est obtenue en utilisant la condition de Lindeberg. Comme ceci est vrai pour tout $\eta > 0$ on a bien

$$\max_{i=1,\ldots,n} \frac{\mathrm{V}(Z_i)}{V_n^2} \to 0, \quad n \to \infty \tag{7.1}$$

eq:limmaxvar

R

Remarque 7.3 Le théorème 7.2 montre que la condition de Lindeberg est suffisante pour avoir la convergence en loi de $\sum_{i=1}^n Z_i/V_n$. Elle est également suffisante pour avoir la condition (7.1). Un résultat dû à FELLER montre que si la suite $\{Z_i, i = 1, 2, \ldots\}$ du théorème 7.2 satisfait la condition (7.1), alors la condition de Lindeberg est également nécessaire pour la convergence en loi de $\sum_{i=1}^n Z_i/V_n$.

Autrement dit, pour toute suite $\{Z_i, i = 1, 2, \ldots\}$ de variables indépendantes dont les espérances sont nulles et les variances existent et telle que $\max_{i=1,\ldots,n} \frac{\mathrm{V}(Z_i)}{V_n^2} \to 0$ lorsque $n \to \infty$, les deux propriétés suivantes sont équivalentes :

1. $\sum_{i=1}^n Z_i/V_n$ converge en loi vers $N(0, 1)$
2. la suite $\{Z_i, i = 1, 2, \ldots\}$ satisfait la condition de Lindeberg

R

Remarque 7.4 La condition (7.1) signifie que lorsque n est grand, aucun des termes $\mathrm{V}(Z_i)$ de la somme $V_n^2 = \sum_{i=1}^n \mathrm{V}(Z_i)$ ne domine cette somme. Lorsque la somme contient un très grand nombre de termes ($n \to \infty$), alors le poids $\frac{\mathrm{V}(Z_i)}{V_n^2}$ de chaque terme de cette somme doit être négligeable (tend vers 0). R

On donne une propriété qui servira de base aux résultats de convergence de l'estimateur des moindres carrés.

pro:cltlbeta **Propriété 7.2** *Soit $\{Z_i : i = 1, 2, \ldots\}$ une suite de variables aléatoires indépendamment et identiquement distribuées d'espérance nulle et de variance égale à 1. Soit $\{\gamma_i : i = 1, 2, \ldots\}$ une suite de réels tels que pour un certain entier $\overline{n}$, $\gamma_1, \ldots, \gamma_{\overline{n}}$ sont non-tous nuls et tels que $\max_{i=1,\ldots,n} a_i^2 \to 0$, $n \to \infty$ où*

$$a_i = \sqrt{\frac{\gamma_i^2}{\sum_{j=1}^n \gamma_j^2}}, \qquad i = 1, \ldots, n \text{ et } n \geq \overline{n}$$

Alors $\sum_{i=1}^n a_i Z_i \xrightarrow{\text{loi}} N(0, 1)$, $n \to \infty$.

Avant de donner la preuve de ce résultat, on donne la raison pour laquelle on suppose la présence d'un tel entier $\overline{n}$. On voit que si pour un n on a $\gamma_1 = \cdots = \gamma_n = 0$, alors $a_1, \ldots, a_n$ ne sont pas définis. Supposer qu'il existe $\overline{n}$ pour lequel si $n \geq \overline{n}$, on peut trouver $i \in \{1, \ldots, n\}$ tel que $\gamma_i \neq 0$ garanti que la suite $\{a_n : n \geq \overline{n}\}$ est bien définie. Donc pour tout $n \geq \overline{n}$, $U_n = \sum_{i=1}^n a_i Z_i$ est également bien défini ainsi que la suite $\{U_n : n \geq \overline{n}\}$. Comme on ne s'intéresse qu'à la limite (en loi) de U_n, il n'est pas important que cette suite ne possède pas de termes de rang $1, 2, \ldots \overline{n} - 1$ (on peut les définir de manière arbitraire).

Preuve : Commençons par noter que $\sum_{i=1}^n a_i^2 = 1$. Posons alors $U_i = a_i Z_i$. Il faut montrer la normalité asymptotique de $\sum_{i=1}^n U_i$. Pour cela, on applique le théorème de Lindeberg (théorème 7.2). On calcule aisément $\mathrm{E}(U_i) = a_i \mathrm{E}(Z_i) = 0$ et $\mathrm{V}(U_i) = a_i^2 \mathrm{V}(Z_i) = a_i^2$, $i = 1, \ldots, n$. On définit $V_n^2 = \sum_{i=1}^n \mathrm{V}(U_i) = \sum_{i=1}^n a_i^2 = 1$. La convergence recherchée aura alors lieu si on peut montrer que la suite des U_i satisfait la condition de Lindeberg. Soit $\eta > 0$ un réel fixé et

définissons $\delta_n = \max_{i=1,\ldots,n} a_i^2$. En utilisant l'expression de U_i, on a

$$\frac{1}{V_n^2}\sum_{i=1}^{n} \mathrm{E}\Big(\mathbb{1}_{|U_i|>\eta V_n} U_i^2\Big) = \sum_{i=1}^{n} a_i^2 \mathrm{E}\Big(\mathbb{1}_{a_i^2 Z_i^2 > \eta^2} Z_i^2\Big) \leq \sum_{i=1}^{n} a_i^2 \mathrm{E}\Big(\mathbb{1}_{\delta_n Z_i^2 > \eta^2} Z_i^2\Big) \qquad (7.2)$$ eq:lc1

où l'inégalité provient de l'équivalence $a \geq b \iff \mathbb{1}_{b>c} \leq \mathbb{1}_{a>c}$, vraie pour tout réels a, b, c. Comme les Z_i sont identiquement distribués, les variables $\mathbb{1}_{\delta_n Z_i^2 > \eta^2} Z_i^2$ le sont également et toutes les espérances (7.2) sont égales à $\mathrm{E}(\mathbb{1}_{\delta_n Z_1^2 > \eta^2} Z_1^2)$. On peut donc écrire le membre de droite de l'inégalité (7.2) comme

$$\sum_{i=1}^{n} a_i^2 \mathrm{E}\Big(\mathbb{1}_{\delta_n Z_i^2 > \eta^2} Z_i^2\Big) = \mathrm{E}(\mathbb{1}_{\delta_n Z_1^2 > \eta^2} Z_1^2) \sum_{i=1}^{n} a_i^2 = \mathrm{E}(\mathbb{1}_{\delta_n Z_1^2 > \eta^2} Z_1^2) = \mathrm{E}(\mathbb{1}_{Z_1^2 > \eta^2/\delta_n} Z_1^2)$$

où la dernière égalité est vraie pour n suffisamment grand ($n \geq \overline{n}$) afin d'éviter le cas $\delta_n = 0$. Sous la condition de l'énoncé, on a $\lim_{n\to\infty} 1/\delta_n = +\infty$ et donc $\lim_{n\to\infty} \mathbb{1}_{Z_1^2 > \eta^2/\delta_n} Z_1^2 \to 0$, presque sûrement. Comme par ailleurs $\mathbb{1}_{Z_1^2 > \eta^2/\delta_n} Z_1^2 \leq Z_1^2$ pour tout $n \geq \overline{n}$, et que $\mathrm{E}(Z_1^2) = 1$, on peut appliquer le théorème de convergence dominée :

$$\lim_{n\to\infty} \mathrm{E}(\mathbb{1}_{Z_1^2 > \eta^2/\delta_n} Z_1^2) = \mathrm{E}(\lim_{n\to\infty} \mathbb{1}_{Z_1^2 > \eta^2/\delta_n} Z_1^2) = 0$$

Par conséquent le membre de gauche de (7.2) converge également vers 0 lorsque $n \to \infty$. Comme ceci est vrai pour tout $\eta > 0$, la suite $\{U_i : i = 1, 2 \ldots\}$ satisfait la condition de Lindeberg. D'après le théorème 7.2, on a donc

$$\sum_{i=1}^{n} \frac{U_i}{V_n} = \sum_{i=1}^{n} a_i Z_i \overset{\text{loi}}{\to} N(0,1), \qquad n \to \infty$$

■

Finalement, on mentionne dans cette section deux propriétés qui seront utiles pour démontrer l'équivalence entre des résultats de convergence. Elles se démontrent à l'aide de notions et résultats qui ne sont pas présentés dans ce cours. On peut en trouver une démonstration dans l'ouvrage *Cours de probabilités* d'A. Montfort.

pro:convloi_cont **Propriété 7.3 (Mann-Wald)** *Si $\{Z_n : n \geq 1\}$ est une suite de vecteurs aléatoires de $\mathbb{R}^q$ qui converge en loi vers Z, alors pour toute fonction continue $f : \mathbb{R}^q \to \mathbb{R}^m$, la suite de vecteurs aléatoires $\{f(Z_n) : n \geq 1\}$ converge en loi vers $f(Z)$.*

Ce résultat est une version pour la convergence en loi du théorème de Slutsky.

pro:convloi_mat **Propriété 7.4** *Si $\{Z_n : n \geq 1\}$ est une suite de vecteurs aléatoires de $\mathbb{R}^q$ qui converge en loi vers Z et si $\{A_n : n \geq 1\}$ est une suite de matrices (non aléatoires) de dimensions (m, q) qui converge vers A, alors la suite de vecteurs $A_n Z_n$ converge en loi vers le vecteur AZ.*

On notera en particulier que les suites AZ_n et $A_n Z_n$ ont la même limite en loi. Par conséquent, si A_n converge vers A, la loi limite de $A_n Z_n$ peut s'obtenir à partir de celle de AZ_n.

sec:asynormco

7.2.2.2 Convergence en loi de $\hat{\beta}$

On peut maintenant prouver le résultat de convergence en loi de l'estimateur $\hat{\beta}$.

Notons que dans le cas d'un modèle de régression linéaire défini par les condition C$_p$1, C$_p$2 et C$_p$3, dans lequel la matrice X ne contient qu'une seule colonne, on peut écrire $\hat{\beta} - \beta = \left(\sum_{i=1}^n X_i^2\right)^{-1} \sum_{i=1}^n X_i \varepsilon_i$. Donc

$$\frac{\sqrt{\sum_{i=1}^n X_i^2}}{\sigma}(\hat{\beta} - \beta) = \sum_{i=1}^n \frac{X_i}{\sqrt{\sum_{i=1}^n X_i^2}} \tilde{\varepsilon}_i$$

où $\tilde{\varepsilon}_i = \frac{1}{\sigma}\varepsilon_i$. En posant $Z_i = \tilde{\varepsilon}_i$ et $\gamma_i = X_i$, on est dans la condition de la propriété 7.2, dès que $\varepsilon_1, \varepsilon_2 \ldots$ forment une suite de variables aléatoires indépendantes et identiquement distribuées et que $\max_{i=1,\ldots,n} \left(X_i^2 / \sum_{i=1}^n X_i^2\right)$ converge vers 0 lorsque $n \to \infty$. Dans ce cas

$$\frac{\sqrt{\sum_{i=1}^n X_i^2}}{\sigma}(\hat{\beta} - \beta) \xrightarrow{\text{loi}} N(0,1) \qquad n \to \infty$$

Le résultat de convergence en loi de l'estimateur des moindres carrés dans le cadre de modèle avec plus d'une variable s'obtient à partir de la propriété 7.2.

th:norasybetaa

Théorème 7.3 *Soit le modèle de régression linéaire défini par les conditions* C$_p$1, C$_p$2 *et* C$_p$3. *On suppose que rang(X) = p+1 et que les conditions suivantes sont également satisfaites :*

C$_p$I. $Y_i - X_{i\cdot}^\top \beta$, $i = 1, \ldots, n$, *sont des variables aléatoires indépendantes et identiquement distribuées, pour tout n*

C$_p W$. *Lorsque* $n \to \infty$, $W^\top W$ *converge vers une matrice* Γ *définie positive, où* $W = XD^{-1}$, *et* D *est la matrice* $\text{diag}(\|X_{\cdot k}\|, k = 0, \ldots, p)$.

Si de plus $\max_{i=1,\ldots,n} \frac{X_{ik}^2}{\|X_{\cdot k}\|^2} \to 0$ *lorsque* $n \to \infty$ *pour tout* $k = 0, \ldots, p$, *alors*

$$D(\hat{\beta} - \beta) \xrightarrow{\text{loi}} N(0_{p+1}, \sigma^2 \Gamma^{-1}) \qquad n \to \infty$$

Preuve : On peut toujours écrire $X^\top X = V \Lambda V^\top$, où V et Λ sont respectivement les matrices des vecteurs et valeurs propres de $X^\top X$, les vecteurs propres étant orthonormés ($V^\top V = I_{p+1}$). Puisque $X^\top X$ est de rang $p+1$, elle est définie positive, et on peut aussi écrire $\Lambda = \Lambda^{1/2}\Lambda^{1/2}$ où $\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_k}, k = 0, \ldots, p)$ et $\lambda_0, \ldots, \lambda_p$ sont les valeurs propres de $X^\top X$. De manière analogue, on définira $\Lambda^{-1/2} = \text{diag}(1/\sqrt{\lambda_k}, k = 0, \ldots, p)$.

On note qu'avec les définitions de Γ, D et W, on peut écrire

$$\Gamma = \lim_{n\to\infty} W^\top W = \lim_{n\to\infty} D^{-1} X^\top X D^{-1} = \lim_{n\to\infty} D^{-1} V \Lambda V^\top D^{-1} = \lim_{n\to\infty} A^\top A$$

où $A = \Lambda^{1/2} V^\top D^{-1}$. Donc d'après les propriétés 7.3 et 7.4, montrer la convergence de l'énoncé revient à montrer que

$$\frac{1}{\sigma} A D(\hat{\beta} - \beta) \xrightarrow{\text{loi}} N(0_{p+1}, I_{p+1}) \tag{7.3}$$

eq:econv1

En utilisant la définition de A, on obtient $AD = \Lambda^{1/2}V^\top$ et donc (d'après C$_p$2 et la définition de $\hat{\beta}$) :

$$\begin{aligned}\frac{1}{\sigma}AD(\hat{\beta}-\beta) &= \frac{1}{\sigma}AD(X^\top X)^{-1}X^\top\varepsilon = \frac{1}{\sigma}ADV\Lambda^{-1}V^\top X^\top\varepsilon = \frac{1}{\sigma}\Lambda^{1/2}V^\top V\Lambda^{-1}V^\top X^\top\varepsilon \\ &= \frac{1}{\sigma}\Lambda^{-1/2}V^\top X^\top\varepsilon = \frac{1}{\sigma}\Lambda^{-1/2}V^\top DD^{-1}X^\top\varepsilon = \frac{1}{\sigma}\Lambda^{-1/2}V^\top DW^\top\varepsilon \\ &= \frac{1}{\sigma}(WA^{-1})^\top\varepsilon\end{aligned}$$

D'après le théorème de Cramér-Wold (théorème 7.1), montrer la convergence (7.3) revient à montrer $\frac{1}{\sigma}c^\top(WA^{-1})^\top\varepsilon \xrightarrow{\text{loi}} N(0, c^\top c)$, pour tout $c \in \mathbb{R}^{p+1}$ ($c \neq 0_{p+1}$), ou encore (d'après la propriété 7.3)

$$\frac{1}{\sigma}\frac{c^\top(WA^{-1})^\top\varepsilon}{\sqrt{c^\top c}} \xrightarrow{\text{loi}} N(0,1) \tag{7.4}$$ eq:econv2

On définit le vecteur $\gamma = WA^{-1}c \in \mathbb{R}^n$. On constate alors que d'après la définition de A et de W, on a

$$\gamma^\top\gamma = c^\top\Lambda^{-1/2}V^\top X^\top XV\Lambda^{-1/2}c = c^\top\Lambda^{-1/2}V^\top V\Lambda V^\top V\Lambda^{-1/2}c = c^\top c$$

la dernière égalité résultat de l'orthonormalité des vecteurs propres de $X^\top X$ et de l'écriture de $\Lambda = \Lambda^{1/2}\Lambda^{1/2}$. Par conséquent,

$$\frac{1}{\sigma}\frac{c^\top(WA^{-1})^\top\varepsilon}{\sqrt{c^\top c}} = \frac{1}{\sigma}\frac{\gamma^\top\varepsilon}{\sqrt{\gamma^\top\gamma}} = \sum_{i=1}^n a_i\tilde{\varepsilon}_i$$

avec $a_i = \frac{\gamma_i}{\sqrt{\gamma^\top\gamma}}$ et $\tilde{\varepsilon}_i = \varepsilon_i/\sigma$, $i = 1,\dots,n$. En utilisant la propriété 7.2, pour avoir la convergence voulue (7.4) il suffit d'avoir $\max_{i=1,\dots,n} a_i^2 \to 0$, $n \to \infty$, ce qu'on montre à présent.

Pour cela, remarquons que γ_i est l'élément de la i^{e} colonne de $\gamma^\top = c^\top{A^{-1}}^\top W^\top$ et donc $\gamma_i = c^\top{A^{-1}}^\top W_{i\cdot} = c^\top u_i$, où $W_{i\cdot}^\top$ est la i^{e} ligne de W (ou encore la i^{e} colonne de $W^\top$) et donc $u_i = {A^{-1}}^\top W_{i\cdot}$. Par conséquent, on peut écrire γ_i sous la forme $\gamma_i = c^\top u_i = \sum_{k=0}^p c_k u_{ik}$ et donc en utilisant l'inégalité de Cauchy-Schwarz [1] :

$$\gamma_i^2 \leq \sum_{k=0}^p c_k^2 \sum_{k=0}^p u_{ik}^2 = (c^\top c)u_i^\top u_i = (c^\top c)W_{i\cdot}^\top(A^\top A)^{-1}W_{i\cdot} \qquad i = 1,\dots,n$$

On en déduit $a_i^2 = \frac{\gamma_i^2}{\gamma^\top\gamma} = \frac{\gamma_i^2}{c^\top c} \leq W_{i\cdot}^\top(A^\top A)^{-1}W_{i\cdot}$, $i = 1,\dots,n$. D'après les inégalités (9.10), on a nécessairement $W_{i\cdot}^\top(A^\top A)^{-1}W_{i\cdot} \leq \xi^* W_{i\cdot}^\top W_{i\cdot}$, $i = 1,\dots,n$, où ξ^* est la plus grande des valeurs propres de $(A^\top A)^{-1}$. Donc

$$a_i^2 \leq \xi^* W_{i\cdot}^\top W_{i\cdot} \qquad i = 1,\dots,n \tag{7.5}$$ eq:econv3

1. L'inégalité de Cauchy-Schwarz établit que $|\langle u,v\rangle| \leq \|u\|\,\|v\|$ où u et v sont deux vecteurs d'un même e.v. et $\|u\|^2 = \langle u,u\rangle$. Elle est équivalente aux deux inégalités $0 \leq \|\frac{u}{\|u\|} \pm \frac{v}{\|v\|}\|^2 = 2(1 \pm \frac{\langle u,v\rangle}{\|u\|\,\|v\|})$, qu'on vérifie aisément.

Comme $W = XD^{-1}$, on doit avoir $W_{i\bullet}^\top = X_{i\bullet}^\top D^{-1} = \left(\frac{X_{i0}}{\|X_{\bullet 0}\|} \frac{X_{i1}}{\|X_{\bullet 1}\|} \cdots \frac{X_{ip}}{\|X_{\bullet p}\|}\right)$, $i = 1, \ldots, n$, et donc $W_{i\bullet}^\top W_{i\bullet} = \sum_{k=0}^p \frac{X_{ik}^2}{\|X_{\bullet k}\|^2}$, $i = 1, \ldots, n$. D'après ceci et (7.5) , on peut alors écrire

$$\max_{i=1,\ldots,n} a_i^2 \leq \xi^* \max_{i=1,\ldots,n} \sum_{k=0}^p \frac{X_{ik}^2}{\|X_{\bullet k}\|^2} \leq \xi^* \sum_{k=0}^p \max_{i=1,\ldots,n} \frac{X_{ik}^2}{\|X_{\bullet k}\|^2} \tag{7.6}$$ eq:econv4

Comme $\Gamma = \lim_{n\to\infty} A^\top A$ avec Γ inversible, on doit avoir $\Gamma^{-1} = \lim_{n\to\infty} (A^\top A)^{-1}$ (l'inversion d'une matrice est une application continue, voir théorème 9.1). Donc $\lim_{n\to\infty} \xi^*$ est la plus grande valeur propre de Γ^{-1}. Comme cette matrice est définie positive, cette valeur propre est strictement positive. Par ailleurs, comme on a supposé que lorsque $n \to \infty$, $\max_{i=1,\ldots,n} \frac{X_{ik}^2}{\|X_{\bullet k}\|^2} \to 0$ pour tout $k = 0, \ldots, p$, le membre de droite de (7.6) converge vers 0, ce qui est bien la condition recherchée. ■

Remarque 7.5 La condition $\mathrm{C}_p W$ requiert que $D^{-1}X^\top X D^{-1}$ converge vers une matrice définie positive lorsque $n \to \infty$. D'après la définition de D, il est facile de voir que la $(k,l)^{\mathrm{e}}$ entrée de $D^{-1}X^\top X D^{-1}$ est $\sum_{i=1}^n \frac{X_{ik}X_{il}}{\|X_{\bullet k}\| \|X_{\bullet l}\|}$. Cette matrice contient donc des termes semblables à des corrélations linéaires empiriques entre les variables explicatives du modèle.

Quant à la condition supplémentaire $\max_{i=1,\ldots,n} \frac{X_{ik}^2}{\|X_{\bullet k}\|^2} \to 0$ pour tout $k = 0, \ldots, p$, elle établit que pour chaque variable explicative, aucun individu ne domine le vecteur des observations de cette variable lorsque n est grand, dans le sens où la contribution maximum d'un individu à la norme de ce vecteur doit être arbitrairement petite lorsque $n \to \infty$. R

rem:falt **Remarque 7.6** La convergence du théorème 7.3 s'énonce de manière équivalente par $\frac{AD}{\sigma}(\hat\beta - \beta) \xrightarrow{\text{loi}} N(0_{p+1}, I_{p+1})$ (voir (7.3)). Cette formulation présente un intérêt particulier. En effet, notons que $\frac{AD}{\sigma} = \frac{\Lambda^{1/2}V^\top}{\sigma}$ et donc la convergence peut aussi s'écrire $\frac{\Lambda^{1/2}V^\top}{\sigma}(\hat\beta - \beta) \xrightarrow{\text{loi}} N(0_{p+1}, I_{p+1})$. Remarquons que pour tout n, $\mathrm{V}(\hat\beta - \beta) = \mathrm{V}(\hat\beta) = \sigma^2(X^\top X)^{-1}$. Donc

$$\mathrm{V}(\hat\beta)^{-1} = \frac{V\Lambda^{1/2}\Lambda^{1/2}V^\top}{\sigma^2} = \left(\frac{\Lambda^{1/2}V^\top}{\sigma}\right)^\top \left(\frac{\Lambda^{1/2}V^\top}{\sigma}\right) = \mathrm{V}(\hat\beta)^{-1/2\top} \mathrm{V}(\hat\beta)^{-1/2}$$

où $\mathrm{V}(\hat\beta)^{-1/2} = \frac{\Lambda^{1/2}V^\top}{\sigma}$ est la « racine carrée » de la variance de $\hat\beta$. Par conséquent, la convergence établie par le théorème 7.3 s'écrit aussi

$$\mathrm{V}(\hat\beta)^{-1/2}(\hat\beta - \beta) \xrightarrow{\text{loi}} N(0_{p+1}, I_{p+1})$$

R

On peut aussi établir cette convergence sous une condition sur la matrice X plus faible que dans le résultat précédent. [2]

th:norasybeta **Théorème 7.4** *Soit le modèle de régression défini par les conditions* C_p1, C_p2 *et* C_p3. *Si la condition* $\mathrm{C}_p\mathrm{I}$ *et la condition* C_pX *:*

C_pX. $\max_{i=1,\ldots,n} X_{i\bullet}^\top (X^\top X)^{-1} X_{i\bullet} \to 0$ *lorsque* $n \to \infty$

2. Cette condition est donnée par A. Monfort, *Cours de statistique mathématique*.

sont satisfaites, alors

$$\mathrm{V}(\hat{\beta})^{-1/2}(\hat{\beta}-\beta) \xrightarrow{\text{loi}} N(0_{p+1}, I_{p+1}) \qquad n \to \infty$$

où la matrice $\mathrm{V}(\hat{\beta})^{-1/2}$ *est celle définie à la remarque 7.6.*

Preuve : D'après la remarque 7.6, la convergence à établir est celle exprimée par (7.3), qui est équivalente, comme on l'a montré dans la preuve du théorème 7.3, à la convergence (7.4) :

$$\frac{1}{\sigma}\frac{c^\top (WA^{-1})^\top \varepsilon}{\sqrt{c^\top c}} \xrightarrow{\text{loi}} N(0,1) \qquad n \to \infty, \qquad \forall c \in \mathbb{R}^{p+1}, c \neq 0_{p+1}$$

où les matrices A et W sont définies dans le théorème 7.3 : $A = \Lambda^{1/2}V^\top D^{-1}$ et $W = XD^{-1}$. On sait aussi (voir la preuve du théorème 7.3) que

$$\frac{1}{\sigma}\frac{c^\top (WA^{-1})^\top \varepsilon}{\sqrt{c^\top c}} = \frac{\gamma^\top}{\sqrt{\gamma^\top \gamma}}\tilde{\varepsilon} \qquad \text{avec } \gamma = WA^{-1}c, \quad \tilde{\varepsilon} = \frac{1}{\sigma}\varepsilon$$

et que pour établir la convergence voulue, il suffit de montrer que $\max_{i=1,\dots,n} \gamma_i^2/\gamma^\top\gamma \to 0$, $n \to \infty$. En utilisant les expression de A et de W, on obtient $WA^{-1} = XV\Lambda^{-1/2}$ et donc la i^{e} colonne de $\gamma^\top = c^\top \Lambda^{-1/2}V^\top X^\top$ est

$$\gamma_i = c^\top \Lambda^{-1/2}V^\top X_{i\cdot} = c^\top v_i$$

avec $v_i = \Lambda^{-1/2}V^\top X_{i\cdot}$, $i = 1, \dots, n$. Par conséquent en utilisant l'inégalité de Cauchy-Schwarz on a :

$$\frac{\gamma_i^2}{\gamma^\top\gamma} = \frac{(c^\top v_i)\,(c^\top v_i)}{c^\top c} \leq \frac{(c^\top c)\,(v_i^\top v_i)}{c^\top c} = v_i^\top v_i = X_{i\cdot}^\top V\Lambda^{-1/2}\Lambda^{-1/2}V^\top X_{i\cdot} = X_{i\cdot}^\top (X^\top X)^{-1} X_{i\cdot}$$

$i = 1, \dots, n$. La condition $\mathrm{C}_p X$ assure donc que la condition suffisante $\max_{i=1,\dots,n} \gamma_i^2/\gamma^\top\gamma \to 0$, $n \to \infty$ pour la convergence est satisfaite. ■

Remarque 7.7 Il peut être intéressant de faire le lien entre les conditions sous lesquelles la convergence de $\mathrm{V}(\hat{\beta})^{-1/2}(\hat{\beta}-\beta)$ est obtenue dans les théorèmes 7.3 et 7.4, et notamment établir que la condition assurant la convergence du théorème 7.4 est moins forte que celle requise dans le théorème 7.3. Notons que en particulier le théorème 7.4 ne requiert pas que $\max_{i=1,\dots,n} \frac{X_{ik}^2}{\|X_{\cdot k}\|^2} \to 0$, $n \to \infty$ pour $k = 0, \dots, p$. Cependant, sous la condition $\mathrm{C}_p W$, on a

$$\mathrm{C}_p X \iff \max_{i=1,\dots,n} \frac{X_{ik}^2}{\|X_{\cdot k}\|^2} \to 0, \quad n \to \infty \quad \forall k = 0, \dots, p$$

Pour cela, remarquons que $\max_{i=1,\dots,n} X_{i\cdot}^\top (X^\top X)^{-1} X_{i\cdot} = \max_{i=1,\dots,n} X_{i\cdot}^\top D^{-1}(W^\top W)^{-1}D^{-1}X_{i\cdot}$ pour tout n. Notons ξ^{**} et ξ^* la plus petite et la plus grande valeur propre de $(W^\top W)^{-1}$.[3] En utilisant la relation (9.10) (à la fin de la section 9.3), on peut alors écrire

$$\xi^{**} \max_{i=1,\dots,n} X_{i\cdot}^\top D^{-1}D^{-1}X_{i\cdot} \leq \max_{i=1,\dots,n} X_{i\cdot}^\top (X^\top X)^{-1} X_{i\cdot} \leq \xi^* \max_{i=1,\dots,n} X_{i\cdot}^\top D^{-1}D^{-1}X_{i\cdot}$$

3. On remarque que $W^\top W = A^\top A$, où A est la matrice introduite dans la preuve du théorème 7.3. Par conséquent, la valeur propre ξ^* utilisée dans la preuve de ce théorème est la même que celle introduite ici.

pour tout n. Notons aussi que d'après la définition de D, on a

$$X_{i\cdot}^\top D^{-1} D^{-1} X_{i\cdot} = \|X_{i\cdot}^\top D^{-1}\|^2 = \sum_{k=0}^{p} \frac{X_{ik}^2}{\|X_{\cdot k}\|^2}$$

Donc la double inégalité ci-dessus s'écrit

$$\xi^{**} \max_{i=1,\dots,n} \sum_{k=0}^{p} \frac{X_{ik}^2}{\|X_{\cdot k}\|^2} \leq \max_{i=1,\dots,n} X_{i\cdot}^\top (X^\top X)^{-1} X_{i\cdot} \leq \xi^* \max_{i=1,\dots,n} \sum_{k=0}^{p} \frac{X_{ik}^2}{\|X_{\cdot k}\|^2}$$

pour tout n. Ces inégalités sont donc vraies lorsque $n \to \infty$. Sous la conditions $\mathrm{C}_p W$, ξ^* et ξ^{**} convergent respectivement vers la plus grande et plus petite des valeurs propres de la matrice définie positive Γ, toutes deux strictement positives. Par conséquent,

$$\max_{i=1,\dots,n} X_{i\cdot}^\top (X^\top X)^{-1} X_{i\cdot} \to 0 \iff \max_{i=1,\dots,n} \sum_{k=0}^{p} \frac{X_{ik}^2}{\|X_{\cdot k}\|^2} \to 0$$

lorsque $n \to \infty$. Or la convergence $\max_{i=1,\dots,n} \sum_{k=0}^{p} \frac{X_{ik}^2}{\|X_{\cdot k}\|^2} \to 0$ équivaut à $\max_{i=1,\dots,n} \frac{X_{ik}^2}{\|X_{\cdot k}\|^2} \to 0$ pour $k = 0, \dots, p$.[4] ℝ

Remarque 7.8 La normalité asymptotique de $\hat{\beta}$ apparaît souvent sous la forme

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\text{loi}} N(0_{p+1}, \sigma^2 Q^{-1}) \tag{7.7}$$

eq:norasybetaa

où Q est la matrice définie postive, définie comme $Q = \lim_{n\to\infty} \frac{X^\top X}{n}$, cette convergence étant introduite comme hypothèse. On présente le lien entre la convergence des théorèmes 7.4 et 7.3 et celle de $\sqrt{n}(\hat{\beta} - \beta)$. Notons d'abord qu'en appliquant la propriété 7.3, cette convergence équivaut à

$$\frac{1}{\sigma} Q^{1/2} \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\text{loi}} N(0_{p+1}, I_{p+1}) \tag{7.8}$$

eq:convalt

où $Q^{1/2}$ est une matrice inversible telle que $Q = {Q^{1/2}}^\top Q^{1/2}$.

Ensuite, sous l'hypothèse $\frac{X^\top X}{n} \to Q$, on peut définir $Q^{1/2}$ comme la limite de $\frac{\Lambda^{1/2} V^\top}{\sqrt{n}}$, où les matrices Λ et V sont les matrices des valeurs et vecteurs propres de $X^\top X$. Dans ce cas, en utilisant la propriété 7.4 et la convergence établie en (7.8), on a aussi

$$\frac{1}{\sigma} \frac{\Lambda^{1/2} V^\top}{\sqrt{n}} \sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sigma} \Lambda^{1/2} V^\top (\hat{\beta} - \beta) \xrightarrow{\text{loi}} N(0_{p+1}, I_{p+1}) \tag{7.9}$$

eq:convalta

La définition de la matrice $\mathrm{V}(\hat{\beta})^{-1/2}$ donnée dans le théorème 7.2 montre que la convergence (7.9) est la même que celle donnée dans le théorème, ce qui montre la cohérence des deux formes du résultat. ℝ

4. Cette dernière équivalence peut se démontrer de la manière suivante. $\max_{i=1,\dots,n} \sum_{k=1}^{q} a_{ik}^2 \leq \sum_{k=1}^{q} \max_{i=1,\dots,n} a_{ik}^2$. Donc $\max_{i=1,\dots,n} a_{ik}^2 \to 0$ pour $k = 1, \dots, q$ implique $\max_{i=1,\dots,n} \sum_{k=1}^{q} a_{ik}^2 \to 0$. Réciproquement, $\max_{i=1,\dots,n} \sum_{k=1}^{q} a_{ik}^2 \geq a_{ik}^2$, $\forall i = 1, \dots, n$, $\forall k = 1, \dots, q$. D'où $\max_{i=1,\dots,n} \sum_{k=1}^{q} a_{ik}^2 \geq \max_{i=1,\dots,n} a_{ik}^2$, $\forall k = 1, \dots, q$, et donc $\max_{i=1,\dots,n} \sum_{k=1}^{q} a_{ik}^2 \to 0$ implique $\max_{i=1,\dots,n} a_{ik}^2 \to 0$, $\forall k = 1, \dots, q$.

Remarque 7.9 Une des raisons pour lesquelles la convergence en loi de $\hat{\beta}$ est donnée sous la forme $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{\text{loi}} N(0_{p+1}, \sigma^2 Q^{-1})$, avec $Q = \lim_{n\to\infty} \frac{X^\top X}{n}$ est liée à la méthode de démonstration de ce résultat. Celle-ci est fréquemment donnée de la manière suivante.

On constate d'abord que $\hat{\beta} - \beta = (X^\top X)^{-1} X^\top \varepsilon$ (ceci découle de l'expression de $\hat{\beta}$ et de la condition C_p1). Par conséquent,

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{X^\top X}{n}\right)^{-1} \frac{X^\top \varepsilon}{\sqrt{n}}$$

Sous l'hypothèse que $\frac{X^\top X}{n}$ converge vers une matrice Q définie positive, alors si elle existe, la limite en loi de $\sqrt{n}(\hat{\beta} - \beta)$ est la même que celle de $Q^{-1}\frac{X^\top \varepsilon}{\sqrt{n}}$ (voir la propriété 7.4 et la remarque précédente). Par conséquent, pour montrer la convergence (7.7), il suffit de montrer que

$$\frac{X^\top \varepsilon}{\sqrt{n}} \xrightarrow{\text{loi}} N(0, \sigma^2 Q)$$

puisqu'alors, la propriété 7.3 permettra d'aboutir à la convergence voulue. C'est précisément cette convergence qui est établie par le théorème 7.4, sous la condition $\mathrm{C}_p X$.

Il faut cependant bien noter que la condition $\lim_{n\to\infty} \frac{1}{n} X^\top X = Q$ seule ne suffit pas pour obtenir la convergence en loi de $\frac{X^\top \varepsilon}{\sqrt{n}}$. R

Qu'on énonce la convergence en loi de $\hat{\beta}$ sous la forme du théorème 7.4 ou sous la forme (7.7), ce résultat n'a pas d'incidence pratique immédiate puisque dans les deux cas, il fait apparaître le paramètre σ dont la valeur est inconnue.

7.3 Propriétés asymptotiques de $\hat{\sigma}^2$

On établit pour $\hat{\sigma}^2$ le même type de résultats que pour $\hat{\beta}$.

7.3.1 Convergence de $\hat{\sigma}^2$

sec:convsig

On montre dans cette section que dans le contexte du modèle de régression linéaire standard, $\hat{\sigma}^2$ converge en probabilité vers σ^2 lorsque $n \to \infty$, dès qu'on suppose que $\varepsilon_1, \varepsilon_2, \ldots$ forment une suite de variables aléatoires i.i.d. Autrement dit $\hat{\sigma}^2$ est un estimateur (faiblement) convergent de σ. Le résultat de base pour obtenir ce résultat est la loi faible des grands nombres (de Khintchine) rappelée ci-dessous.

pro:weaklln **Propriété 7.5** *Si $\{Z_i : i \geq 1\}$ est une suite de variables aléatoires i.i.d. dont l'espérance commune existe (i.e., $\mathrm{E}(|Z_1|) < \infty$), alors $\frac{1}{n}\sum_{i=1}^n Z_i \xrightarrow{\text{p}} \mathrm{E}(Z_1)$, lorsque $n \to \infty$.*

On peut alors prouver le résultat de convergence de l'estimateur $\hat{\sigma}^2$.

pro:lgnsigma **Propriété 7.6** *Sous les conditions C_p1, C_p2, C_p3 et $\mathrm{C}_p\mathrm{I}$, $\hat{\sigma}^2 \xrightarrow{\text{p}} \sigma^2$ lorsque $n \to \infty$.*

Preuve : On a $\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\varepsilon}^\top\hat{\varepsilon} = \frac{1}{n-p-1}(\varepsilon^\top\varepsilon - \varepsilon^\top P_X\varepsilon)$ où $P_X = X(X^\top X)^{-1}X^\top$ (voir la section 5.6). La limite en probabilité de $\hat{\sigma}^2$ est évidemment la même que celle de $\frac{1}{n}\varepsilon^\top\varepsilon - \frac{1}{n}\varepsilon^\top P_X\varepsilon$. On a $\frac{1}{n}\varepsilon^\top\varepsilon = \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2$. Si chacun des deux termes de cette différence possède une limite finie en probabilité, alors la limite en probabilité de la différence est la différence des limites en probabilité.

Sous les conditions dans de la propriété, $\varepsilon_1^2, \varepsilon_2^2, \ldots$ forment une suite de variables aléatoires i.i.d. dont l'espérance commune est $\mathrm{E}(\varepsilon_1^2) = \sigma^2$. Par conséquent, en appliquant la loi faible des grands nombres (propriété 7.5), on a $\frac{1}{n}\,\varepsilon^\top\varepsilon \xrightarrow{\mathrm{p}} \sigma^2$.

Par ailleurs, en utilisant une démarche semblable à celle de la section 5.6, on a

$$\mathrm{E}(\varepsilon^\top P_X\varepsilon) = \sigma^2\mathrm{trace}(P_X) = \sigma^2\mathrm{trace}((X^\top X)^{-1}X^\top X) = \sigma^2(p+1)$$

Finalement, comme P_X est symétrique idempotente, ses valeurs propres distinctes sont 0 et 1. Elle est donc semi-définie positive et $\varepsilon^\top P_X\varepsilon \geq 0$ presque sûrement. On peut alors appliquer l'inégalité de Markov [5] :

$$\mathrm{P}(\frac{1}{n}\,\varepsilon^\top P_X\varepsilon > \eta) \leq \frac{1}{n\eta}\mathrm{E}(\varepsilon^\top P_X\varepsilon) = \frac{\sigma^2}{n\eta}(p+1)$$

pour tout réel $\eta > 0$. Donc $\mathrm{P}(\frac{1}{n}\,\varepsilon^\top P_X\varepsilon > \eta) \to 0$ lorsque $n \to \infty$. Comme ceci est vrai pour tout η, on a $\frac{1}{n}\,\varepsilon^\top P_X\varepsilon \xrightarrow{\mathrm{p}} 0$. Donc $\hat{\sigma}^2 \xrightarrow{\mathrm{p}} \sigma^2$. ■

sec:convloi_sigma

7.3.2 Loi asymptotique de $\hat{\sigma}^2$

On établit finalement une convergence en loi pour $\hat{\sigma}^2$. La preuve de cette convergence utilise le résultat suivant [6], par ailleurs très utile en statistique.

pro:convloi_jointe

Propriété 7.7 *Si* $\{Z_{1n} : n \geq 1\}$ *et* $\{Z_{2n} : n \geq 1\}$ *sont deux suites de vecteurs aléatoires tels que* $Z_{1n} \xrightarrow{\text{loi}} Z$ *et* $Z_{2n} \xrightarrow{\mathrm{p}} z$, *lorsque* $n \to \infty$, *où* z *est* non aléatoire, *alors le vecteur aléatoire* $Z_n = (Z_{1n}^\top, Z_{2n}^\top)^\top$ *converge en loi vers le vecteur* $(Z^\top, z^\top)^\top$, *lorsque* $n \to \infty$.

Sous les conditions de cette propriété, si Z_{1n} et Z_{2n} sont de dimensions respectives q_1 et q_2, et si $f : \mathbb{R}^{q_1+q_2} \to \mathbb{R}^q$ est une fonction continue, alors la propriété 7.3 permet d'établir que $f(Z_{1n}, Z_{2n})$ converge en loi vers $f(Z, z)$

Propriété 7.8 *Sous les conditions* C_p1, C_p2, C_p3 *et* $\mathrm{C}_p\mathrm{I}$, *et si de plus* $\omega = \mathrm{E}(\varepsilon_1^4) < \infty$, *alors* $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{\text{loi}} N(0, \omega - \sigma^4)$, $n \to \infty$.

Preuve : La loi limite de $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$ est la même que celle de $\sqrt{n}(S^2 - \sigma^2)$ où $S^2 = \frac{n-p-1}{n}\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n}\hat{\varepsilon}^\top\hat{\varepsilon}$. On montre donc que $\sqrt{n}(S^2 - \sigma^2) \xrightarrow{\text{loi}} N(0, \omega - \sigma^4)$. On note $e_i = \varepsilon_i^2$. On peut donc écrire

$$\sqrt{n}(S^2 - \sigma^2) = \sqrt{n}\Big(\frac{1}{n}\hat{\varepsilon}^\top\hat{\varepsilon} - \sigma^2\Big) = \sqrt{n}\Big(\frac{1}{n}\varepsilon^\top\varepsilon - \sigma^2 - \frac{1}{n}\varepsilon^\top P_X\varepsilon\Big) = \sqrt{n}(\overline{e} - \sigma^2) - \frac{1}{\sqrt{n}}\varepsilon^\top P_X\varepsilon$$

5. L'inégalité de Markov établit que pour une variable aléatoire Z presque sûrement positive dont l'espérance existe, on a $\mathrm{P}(Z > \eta) \leq \frac{\mathrm{E}(Z)}{\eta}$ pour tout réel $\eta > 0$. Elle se prouve en notant que $\mathrm{P}(Z > \eta) = \mathrm{E}(\mathbb{1}_{Z>\eta})$ et que $0 \leq \eta\,\mathbb{1}_{Z>\eta} \leq Z$, et donc $\mathrm{E}(\eta\,\mathbb{1}_{Z>\eta}) \leq \mathrm{E}(Z)$.

6. dont on trouve la preuve dans l'ouvrage *Cours de probabilités* d'A. Montfort.

où $\overline{e} = \frac{1}{n}\sum_{i=1}^n e_i$. Pour montrer la convergence en loi de $\sqrt{n}(S^2-\sigma^2)$, on utilise la propriété 7.7. On note $Z_{1n} = \sqrt{n}(\overline{e}-\sigma^2)$ et $Z_{2n} = \frac{1}{\sqrt{n}}\varepsilon^\top P_X \varepsilon$, et donc $\sqrt{n}(S^2-\sigma^2) = Z_{1n} - Z_{2n}$. En utilisant la même démarche que dans la preuve de la propriété 7.6, on obtient que $Z_{2n} \xrightarrow{\text{p}} 0$. Par ailleurs, on note que $e_1, e_2, \ldots$ forment une suite de variables aléatoires i.i.d. avec $\mathrm{E}(e_1) = \sigma^2$ et $\mathrm{V}(e_1) = \mathrm{E}(\varepsilon_1^4) - \left[\mathrm{E}(\varepsilon_1^2)\right]^2 = \omega - \sigma^4 < \infty$, ce qui permet d'appliquer le théorème « central limit » de Lindeberg-Lévy (pour les suites de v.a. i.i.d. dont le second moment est fini). On obtient alors $Z_{2n} \xrightarrow{\text{loi}} N(0, \omega-\sigma^4)$. Donc la propriété 7.7 permet de conclure que $(Z_{1n}, Z_{2n})^\top$ converge en loi vers $(Z,0)^\top$, où $Z \sim N(0,\omega-\sigma^4)$. La remarque qui suit la propriété donne le résultat voulu pour $\sqrt{n}(S^2-\sigma^2) = Z_{1n} - Z_{2n}$. ■

7.4 Utilisation des propriétés asymptotiques

Au début de ce chapitre, on a justifié la recherche de résultats asymptotiques

Le théorème 7.4 ainsi que les propriétés 7.6 et 7.7 permettent conjointement d'obtenir le résultat suivant.

pro:utasy **Propriété 7.9** *Dans le modèle de régression défini par* C_p1, C_p2, C_p3, $\mathrm{C}_p\mathrm{I}$, *et dans lequel on suppose* C_pX,

$$\hat{\mathrm{V}}(\hat\beta)^{-1/2}(\hat\beta-\beta) \xrightarrow{\text{loi}} N(0_{p+1}, I_{p+1}) \qquad n\to\infty$$

où $\hat{\mathrm{V}}(\hat\beta)^{-1/2} = \frac{1}{\hat\sigma}\Lambda^{1/2}V^\top$.

Preuve : C'est une application immédiate des résultats mentionnés ci-dessus.

D'un point de vue pratique, cette convergence permet de dire que sous les conditions de l'énoncé, on peut approximer la loi inconnue de $\hat\beta-\beta$ par la loi $N(0_{p+1}, \hat{\mathrm{V}})$ où $\hat{\mathrm{V}} = \left(\hat{\mathrm{V}}(\hat\beta)^{-1/2}\right)^{-1}\left[\left(\hat{\mathrm{V}}(\hat\beta)^{-1/2}\right)^{-1}\right]^\top$. On calcule que $\left(\hat{\mathrm{V}}(\hat\beta)^{-1/2}\right)^{-1} = \hat\sigma V\Lambda^{-1/2}$, et évidemment $\hat{\mathrm{V}} = \hat\sigma^2 V\Lambda^{-1/2}\Lambda^{-1/2}V^\top = \hat\sigma^2(X^\top X)^{-1} = \hat{\mathrm{V}}(\hat\beta)$. Autrement dit, sous les conditions données dans la propriété 7.9, la loi approximative de $\hat\beta-\beta$ lorsque n est grand est la loi $N(0_{p+1}, \hat\sigma^2(X^\top X)^{-1})$, ce qu'on notera $\hat\beta-\beta \overset{\text{a.}}{\sim} N(0_{p+1}, \hat\sigma^2(X^\top X)^{-1})$. Par conséquent, à chaque fois que l'utilisation de la loi de $\hat\beta$ (ou de la loi de $\hat\beta-\beta$) est requise, on pourra utiliser à la place la loi approximative $N(0_{p+1}, \hat\sigma^2(X^\top X)^{-1})$.

En particulier, si on souhaite tester $H_0 : R\beta = r$ contre $H_1 : R\beta \neq r$, on peut fonder le test sur la statistique de Fisher $F = (R\hat\beta - r)^\top\left[\hat\sigma^2 R(X^\top X)^{-1}R^\top\right]^{-1}(R\hat\beta - r)/q$, et décider H_1 si on observe que la valeur de F est trop élevée, *i.e.*, lorsque F dépasse un seuil s (voir la justification de cette démarche à la section 6.1.2). La démarche usuelle des tests requiert que le risque de type 1 associé à ce test ne dépasse pas le niveau α fixé à l'avance. Il faut alors choisir le seuil s de sorte que $\mathrm{P}_{H_0}(F > s) \leq \alpha$.[7] Pour cela, il est nécessaire de connaître la loi de F lorsqu'on suppose H_0 vraie. Cette loi est ici inconnue et l'ensemble $\mathcal{S}$ pour lequel $s \in \mathcal{S} \implies \mathrm{P}_{H_0}(F > s) \leq \alpha$ ne peut être déterminé. On peut cependant recourir aux approximations de la loi de $\hat\beta$ obtenues lorsque

7. Cette inégalité doit se comprendre comme $\mathrm{P}_{H_0}(F > s) \leq \alpha$, quelle que soit la manière de calculer la probabilité P_{H_0} lorsqu'on suppose H_0 vraie. Contrairement à ce qui se produisait dans le contexte du modèle gaussien, où lorsque H_0 était vraie, il n'y avait qu'une seule manière de calculer cette probabilité (on utilisait la loi de Fisher à $(q, n-p-1)$ degrés de liberté), il y a ici une infinité de manière de calculer cette probabilité. Ceci introduit des difficultés qu'on passe sous silence ici.

$n \to \infty$ afin d'obtenir un test de la forme voulue, *i.e.* qui consiste à décider H_1 si $F > s$, et dont le risque de type 1 est approximativement égal à α lorsque la taille de l'échantillon n est grande.

En effet si H_0 est vraie, alors (en utilisant la propriété 7.3) $R\hat{\beta} - r \overset{\text{a.}}{\sim} N(0_q, \hat{\sigma}^2 R(X^\top X)^{-1} R^\top)$ et donc (utilisation des propriétés 7.3 puis 9.17)

$$(R\hat{\beta} - r)^\top \left[\hat{\sigma}^2 R(X^\top X)^{-1} R^\top\right]^{-1} (R\hat{\beta} - r) = qF \overset{\text{a.}}{\sim} \chi^2(q) \tag{7.10}$$ eq:chi2asy

Par conséquent, lorsque n est suffisamment grand, pour tout s on a

$$\mathrm{P}_{H_0}(F > s) = \mathrm{P}_{H_0}(qF > qs) \simeq \mathrm{P}(C_q > qs)$$

où C_q est une variable aléatoire suivant la loi $\chi^2(q)$. Donc si on désigne par $\chi^2_{p;q}$ le quantile d'ordre p de la loi $\chi^2(q)$, on aura $\mathrm{P}(C_q > qs) = \alpha \iff qs = \chi^2_{1-\alpha;q}$. On a alors

$$\mathrm{P}_{H_0}(F > \frac{1}{q}\chi^2_{1-\alpha;q}) \simeq \alpha$$

lorsque n est grand et le test qui consiste à décider H_1 lorsque F dépasse le seuil $\frac{1}{q}\chi^2_{1-\alpha;q}$ a un risque de type 1 approximativement égal à α.

Remarque 7.10 Il est à noter que ce raisonnement est également vrai lorsque $\boldsymbol{Y} \sim N(X\beta, \sigma^2 I_n)$. Dans ce cas, $F \sim F(q, n-p-1)$ et également $qF \overset{\text{a.}}{\sim} \chi^2(q)$, lorsque $n \to \infty$. Par conséquent, on peut en déduire que le quantile d'ordre p de la loi $qF_{p;q,n-p-1}$ doit converger vers χ^2_q. ℝ

Les régions de confiance pour β ou pour $R\beta$ s'obtiennent par un argument similaire. Ainsi, si on définit

$$\mathscr{C}_n = \{x \in \mathbb{R}^q \mid (x - R\hat{\beta})^\top \left[\hat{\sigma}^2 R(X^\top X)^{-1} R^\top\right]^{-1} (x - R\hat{\beta}) \leq \chi^2_{1-\alpha;q}\}$$

la relation (7.10) permet d'obtenir que $\mathrm{P}_\beta(\beta \in \mathscr{C}_n) \simeq 1 - \alpha$ lorsque n est grand. Autrement dit, $\mathscr{C}_n$ est un région de $\mathbb{R}^q$ dont la probabilité de contenir $R\beta$ est approximativement égale à $1-\alpha$ lorsque n est grand. On peut donc utiliser $\mathscr{C}_n$ comme une région d'un niveau de confiance approximatif $1 - \alpha$.

Chapitre 8

Modèles avec erreurs non-sphériques : hétéroscédasticité et corrélation

sec:mcg_modele

8.1 Introduction et définition

Les modèles de régression linéaires, dans le cas où les variables exogènes sont supposées non-aléatoires, sont caractérisés par la condition C_p2 : l'espérance de la variable endogène Y s'écrit comme une fonction linéaire des variables explicatives $X_0, \ldots, X_p$.[1]

Le caractère standard du modèle de régression linéaire étudié dans les chapitres précédents provient de la condition C_p3 (ou C'_p3) : la matrice des variances-covariances du vecteur des termes d'erreur $\varepsilon = \boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})$ est proportionnelle à la matrice identité, *i.e.*, $\mathrm{V}(\varepsilon) = \sigma^2 I_n$. En termes de modélisation, cela revient à dire que les variables $Y_1, \ldots, Y_n$ (et donc les termes d'erreur correspondants) ont la même variance et sont non-corrélées.

Par léger abus de langage, on dit dans ce cas que les erreurs sont *sphériques*. La terminologie provient de la notion de loi de probabilité sphérique. La loi d'un vecteur aléatoire $Z = (Z_1, \ldots, Z_n)$, d'espérance nulle, est sphérique si sa densité f_Z satisfait la condition suivante : soient $\dot{z}$ et $\mathring{z}$ deux vecteurs de $\mathbb{R}^n$, alors $f_Z(\dot{z}) = f_Z(\mathring{z})$ si et seulement si $\sum_{i=1}^n \dot{z}_i^2 = \sum_{i=1}^n \mathring{z}_i^2$. Autrement dit la densité de Z est la même pour deux vecteurs de $\mathbb{R}^n$ si et seulement si ces deux vecteurs ont la même norme. Par conséquent pour tout $c > 0$, l'ensemble $f_Z^{-1}(c) = \{z \in \mathbb{R}^n \mid f_Z(z) = c\}$, s'il est non vide, contient tous les vecteurs de $\mathbb{R}^n$ qui ont une norme donnée. Cet ensemble coïncide donc avec une sphère de $\mathbb{R}^n$ dont le rayon est égal à la norme de n'importe quel vecteur dans $f_Z^{-1}(c)$. Graphiquement, $f_Z^{-1}(c)$ est la courbe de niveau d'altitude c de f_Z. Les courbes de niveau sont dont des sphères de $\mathbb{R}^n$. En dimension 2, une sphère est un cercle. Donc lorsque $n = 2$, les points du plan ($\mathbb{R}^2$) pour lesquels l'altitude de f_Z reste égale à c sont situés sur un cercle. Il existe un résultat établissant que si Z est un vecteur aléatoire dont la distribution est sphérique, alors la matrice des variances-covariances de Z est proportionnelle à la matrice identité I_n.[2] L'abus de langage qui a été signalé consiste à

1. Si on ne veut pas supposer que les variables exogènes sont non-aléatoires, on doit introduire la notion d'espérance conditionnelle, et les modèles de régression linéaires sont dans ce cas des modèles pour lesquels l'espérance conditionnelle de Y sachant les variables exogènes est une fonction linéaire de ces variables.

2. En toute rigueur, pour obtenir ce résultat, il faut se placer sous certaines conditions, qui permettent notamment d'assurer l'existence des variances.

assimiler la condition C'_p3 avec le fait que le vecteur ε a une distribution sphérique.

Dans de nombreuses applications du modèle de régression linéaire, on peut être amené à vouloir s'affranchir de la condition C_p3. Il existe essentiellement deux raisons pour cela (et qui peuvent se combiner l'une à l'autre).

1. Il peut arriver qu'il ne soit pas réaliste de supposer que $V(Y_i) = V(Y_j)$ pour toute paire (i, j) d'individus, et dans ce cas, les termes diagonaux de la matrice des variances-covariances du vecteur $\boldsymbol{Y}$ ne sont pas égaux. Si cela se produit, on dit qu'il y a *hétéroscédasticité*.
2. Il peut également arriver qu'on ne souhaite pas supposer *a priori* l'absence de corrélation entre Y_i et Y_j. Dans ce cas, les termes extra-diagonaux de la matrice des variances-covariances de $\boldsymbol{Y}$ ne sont pas nécessairement nuls.

Il se peut évidemment qu'on ait à la fois corrélation de la variable endogène entre deux individus et hétéroscédasticité.

En termes de définition du modèle, s'il y a corrélation et/ou hétéroscédasticité, on ne peut plus imposer *a priori* la condition C_p3 (ou C'_p3). Le modèle de régression linéaire qui sera étudié dans ce chapitre est donc défini par les conditions C_p1, C_p2 et la condition C_pV suivante :

C_pV. il existe une matrice symétrique définie positive Ω telle que $V(\boldsymbol{Y}) = \Omega$.

La condition C_pV revient à supposer que les variances $V(Y_1), \ldots, V(Y_n)$ existent et qu'on ne peut pas exprimer le niveau de la variable endogène d'un individu comme une fonction linéaire du niveau de cette variable des autres individus (voir le point 3 de la propriété 9.7).

Pour des raisons qui apparaîtront plus loin, on écrira Ω sous la forme $\Omega = \sigma^2 V$, où σ est un réel strictement positif et V est une matrice symétrique définie positive.[3] En reprenant la notion de distribution sphérique présentée au début de cette section, on dit dans ce cas que les erreurs sont *non-sphériques*.

Dans un premier temps, la condition C_pV est satisfaite avec $\Omega = \sigma^2 V$ où la matrice V est connue, ce qui équivaut à dire que $V(\boldsymbol{Y})$ est connue à une constante (positive) près. On peut alors exprimer les conditions définissant le modèle de la manière suivante :

$$\begin{array}{l} X \text{ est non aléatoire} \\ \exists \beta \in \mathbb{R}^{p+1},\ \exists \sigma \in \left]0; \infty\right[\text{ t.q. } \boldsymbol{Y} = X\beta + \varepsilon \text{ et } V(\varepsilon) = \sigma^2 V \end{array} \tag{8.1}$$

eq:def_mrl_ns_v

où ε est le vecteur aléatoire de $\mathbb{R}^n$ défini par $\varepsilon = \boldsymbol{Y} - E(\boldsymbol{Y})$, et V est une matrice connue de $M_n^*(\mathbb{R})$, l'ensemble des matrices symétriques inversibles de taille (n, n) (voir la section 9.3).

Afin de gagner en généralité, on considérera ensuite le cas où la matrice V permettant d'écrire la condition $\Omega = \sigma^2 V$ est inconnue. Dans ce cas, le modèle étudié sera caractérisé par :

$$\begin{array}{l} X \text{ est non aléatoire et} \\ \exists \beta \in \mathbb{R}^{p+1},\ \exists V \in M_n^*(\mathbb{R}),\ \exists \sigma \in \left]0; \infty\right[\text{ t.q. } \boldsymbol{Y} = X\beta + \varepsilon \text{ et } V(\varepsilon) = \sigma^2 V \end{array} \tag{8.2}$$

eq:def_mrl_ns_v

où ε est défini comme auparavant. La condition V définie positive est satisfaite dès que V est définie comme une matrice des variances-covariances d'un vecteur aléatoire de $\mathbb{R}^n$, supposée inversible.

3. Il est évidemment toujours possible d'écrire sous cette forme toute matrice définie positive Ω.

L'interprétation de β et de la condition $\mathrm{E}(\boldsymbol{Y}) = X\beta$ reste la même que celle donnée dans les sections 5.1 et 5.2.

Il est évident que le modèle de régression linéaire standard est un cas particulier du modèle défini ci-dessus, puisqu'il correspond au cas où $V = I_n$. Les estimateurs et tests présentés dans les chapitres précédents ont des propriétés (optimalité) qui ont été obtenues en utilisant la condition C_p3. Par conséquent, dans le contexte du modèle plus général étudié ici, ces propriétés ne sont plus nécessairement valides. Ce chapitre a pour objet d'étudier dans ce nouveau contexte les propriétés des estimateurs des moindres carrés ordinaires et des tests fondés sur cette méthode d'estimation. On verra que certaines (bonnes) propriétés sont perdues et on présentera des procédures d'estimation et de test adaptées qui permettent de récupérer certaines de ces propriétés.

Il faut noter que si le modèle étudié est défini par la condition (8.2), il y a ambiguïté sur σ et V : ces paramètres ne sont pas identifiés. Cette notion a été abordée dans les remarques 5.1 et suivantes, à propos du paramètre β. Plus précisément, si $\mathrm{rang}(X) = p + 1$, alors si on se donne le vecteur $\mathrm{E}(\boldsymbol{Y})$, il existe un unique $\beta \in \mathbb{R}^{p+1}$ tel que $\mathrm{E}(\boldsymbol{Y}) = X\beta$. Il n'en est pas de même pour la variance de $\boldsymbol{Y}$. Si on se donne la matrice $\mathrm{V}(\boldsymbol{Y})$, il existe plusieurs paires $(\sigma, V) \in]0;+\infty[\times M_n^*(\mathbb{R})$ telles que $\mathrm{V}(\boldsymbol{Y}) = \sigma^2 V$. Du point de vue de l'inférence, le fait qu'il n'existe pas de paire unique de valeurs pour σ et V pour laquelle la condition (8.2) est vraie rend sans objet la recherche de techniques statistiques destinées à approximer ces valeurs.

Une solution qui permet à la fois de lever cette difficulté et de trouver des méthodes d'inférence ayant de bonnes propriétés consiste à restreindre la spécification de V : dans la définition du modèle, au lieu d'imposer la condition $V \in M_n^*(\mathbb{R})$, on imposera une condition plus forte du type $V \in \mathcal{V}$, où $\mathcal{V}$ n'est qu'une partie de $M_n^*(\mathbb{R})$. Cette partie sera choisie en fonction des spécificités qu'on veut attribuer au modèle. Du choix de $\mathcal{V}$ dépendront les propriétés des estimateurs, test, etc, obtenus. Ceux-ci seront donc étudiés dans des contextes propres aux spécificités introduites dans la définition du modèle. Par exemple, si on choisit d'introduire de la corrélation dans le modèle (V non diagonale), on peut représenter cette corrélation de différentes formes, chacune amenant à une forme particulière pour V. Les estimateurs et tests auront des propriétés qui dépendent de la forme particulière retenue pour représenter la corrélation, et donc de la forme de V. Un test ayant de bonnes propriétés pour une certaine forme de corrélation peut les perdre si la corrélation est d'une autre forme.

sec:mcg_prop_mco

8.2 Propriétés des estimateurs des moindres carrés ordinaires

L'estimateur des moindres carrés ordinaires de β est obtenu en suivant la même démarche que celle présentée dans la section 5.3. Cette démarche n'utilise aucune condition sur la forme particulière de la matrice $\mathrm{V}(\boldsymbol{Y})$ et l'expression de l'estimateur reste donc $\hat{\beta} = (X^\top X)^{-1} X^\top \boldsymbol{Y}$. Il demeure linéaire puisque sa linéarité n'est pas une propriété liée à la spécification du modèle, mais une propriété provenant de son expression. Cet estimateur conserve sa propriété d'absence de biais, puisque la démonstration de cette propriété faite à la section 5.3.3 n'utilise pas la condition C_p3 (voir propriété 5.4). De manière explicite, on a le résultat suivant.

Propriété 8.1 *Dans le modèle de régression linéaire défini par les conditions C_p1, C_p2 et C_pV,*

l'estimateur des moindres carrés ordinaires de β est sans biais.

Preuve : C'est une conséquence immédiate de l'expression de $\hat{\beta}$ et des conditions C_p1 et C_p2. ■

Puisque $\hat{\beta}$ est un estimateur sans biais de β, la matrice des variances-covariances de $\hat{\beta}$ mesure la précision de cet estimateur. On a

$$V(\hat{\beta}) = (X^\top X)^{-1} X^\top V(\boldsymbol{Y}) X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1} X^\top V X (X^\top X)^{-1} \tag{8.3}$$

eq:mcg_var_hat

On constate que cette variance diffère de celle qu'on obtient dans le cas particulier du MRLS où $V = I_n$. Ceci a des conséquences aussi bien dans le domaine des tests d'hypothèses (et des régions de confiance) que dans celui de l'estimation.

Dans le cas où $\boldsymbol{Y}$ et donc ε sont des vecteurs gaussiens, $\hat{\beta}$ est également un vecteur gaussien dont la loi est $\mathcal{N}(\beta, \sigma^2 (X^\top X)^{-1} X^\top V X (X^\top X)^{-1})$. Dans la construction du test de Fisher (voir section 6.1.2), la loi la statistique F s'obtient en notant que F s'écrit sous la forme $F = \frac{n-p-1}{q} \frac{C_1}{C_2}$, et que dans le contexte du MRLSG, les variables aléatoires C_1 et C_2 sont indépendantes et suivent chacune une loi du χ^2. Plus précisément, si l'hypothèse nulle est $H_0 : R\beta = r$, on a $C_1 = (R\hat{\beta} - r)^\top \left[R\sigma^2 (X^\top X)^{-1} R^\top\right]^{-1} (R\hat{\beta} - r)$ et $C_2 = \frac{1}{\sigma^2} \hat{\varepsilon}^\top \hat{\varepsilon}$. Lorsque la condition C_p3 est relâchée et remplacée par C_pV, s'il reste vrai que $(R\hat{\beta} - r) \sim \mathcal{N}(0, R V(\hat{\beta}) R^\top)$ et donc $(R\hat{\beta} - r)^\top \left[R V(\hat{\beta}) R^\top\right]^{-1} (R\hat{\beta} - r) \sim \chi^2(q)$ lorsque H_0 est supposée vraie, on a $V(\hat{\beta}) \neq \sigma^2 (X^\top X)^{-1}$ et donc la variable aléatoire C_1 ne suit pas une loi $\chi^2(q)$. Par ailleurs, dans le même contexte, C_2 ne suit pas non plus une loi $\chi^2(n-p-1)$. En effet, le vecteur des résidus $\hat{\varepsilon} = \boldsymbol{Y} - X\hat{\beta}$ peut toujours s'exprimer sous la forme $\hat{\varepsilon} = M_X \varepsilon$, où $M_X = I_n - X(X^\top X)^{-1} X^\top$ et on a donc

$$\frac{1}{\sigma^2} \hat{\varepsilon}^\top \hat{\varepsilon} = \frac{1}{\sigma^2} \varepsilon^\top M_X \varepsilon = Z^\top \Delta Z$$

où $Z = \frac{1}{\sigma} Q^\top \varepsilon$ et Q et Δ sont les matrices des vecteurs et valeurs propres de M_X, respectivement, avec $Q^{-1} = Q^\top$. M_X est la matrice de projection orthogonale sur $L(X_{\cdot 0}, \ldots, X_{\cdot p})^\perp$, et d'après la propriété 9.24, elle a deux valeurs propres distinctes 1 et 0, la valeur propre 1 ayant un degré de multiplicité égal à la dimension du sev $L(X_{\cdot 0}, \ldots, X_{\cdot p})^\perp$. Comme X est de rang $p+1$, le sev $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ est de dimension $p+1$, et $L(X_{\cdot 0}, \ldots, X_{\cdot p})^\perp$ est donc de dimension $n-p-1$. On peut toujours supposer que les valeurs propres non nulles de M_X sont les $n-p-1$ premières, et dans ce cas on a $Z^\top \Delta Z = \sum_{k=1}^{n-p-1} Z_k^2$. Le vecteur Z est un vecteur gaussien d'espérance 0_n et de matrice de variances-covariances

$$V(Z) = V(\frac{1}{\sigma} Q^\top \varepsilon) = \frac{1}{\sigma^2} Q^\top V(\varepsilon) Q = Q^\top V Q$$

Donc, sauf dans le cas très particulier où il se trouverait que $Q^\top V Q = I_n$ (ce qui n'a aucune raison particulière de se produire), les variables aléatoires $Z_1, \ldots, Z_{n-p-1}$ qui constituent les $(n-p-1)$ premières coordonnées du vecteur Z n'ont pas une variance égale à 1 et ne sont pas indépendantes. Par conséquent il n'est pas possible d'appliquer la définition 9.4 et d'établir que $\frac{1}{\sigma^2} \hat{\varepsilon}^\top \hat{\varepsilon} = \sum_{k=1}^{n-p-1} Z_k^2$ suit une loi du χ^2.

Ce qui précède montre que toutes les procédures de test et régions de confiance développées dans les sections 6.1 et 6.2 en se basant sur l'estimation de β par moindres carrés ordinaires ne sont plus valides. Plus précisément, les tests développés dans ces sections n'ont pas le niveau de risque

de type 1 requis et les régions de confiance n'ont pas le niveau de confiance voulu, lorsqu'il sont appliqués dans un modèle où $\mathrm{V}(\boldsymbol{Y}) \neq \sigma^2 I_n$. Il conviendra donc de modifier ces procédures.

L'introduction de la condition $\mathrm{C}_p\mathrm{V}$ à la place de C_p3 a également des conséquences en matière d'estimation de β. En effet, bien que l'estimateur des moindres carrés ordinaires reste sans biais, il perd sa propriété d'optimalité parmi les estimateurs linéaires sans biais (thérorème 2.3). On prouvera ce résultat en deux temps. On commence par montrer que dans un cas particulier, $\hat{\beta}$ ne coïncide pas avec le meilleur estimateur linéaire sans biais. On construira ensuite un autre estimateur de β dont on montrera qu'il est optimal.

Supposons qu'on souhaite estimer la dernière coordonnée β_p de β au moyen d'un estimateur linéaire sans biais. On montre que dans ce cas il n'est pas optimal d'utiliser l'estimateur des moindres carrés ordinaires $\hat{\beta}_p$: il existe un autre estimateur linéaire sans biais de β_p dont la variance est plus petite que celle de $\hat{\beta}_p$. Pour cela on utilise la propriété suivante qui, dans le contexte du modèle de régression linéaire avec la condition $\mathrm{C}_p\mathrm{V}$, est analogue au résultat de la propriété de la remarque 5.11 (et s'obtient par une démarche identique).

pro:mcg_best

Propriété 8.2 *Soit c un vecteur de $\mathbb{R}^{p+1}$. Dans le modèle de régression linéaire défini par les conditions C_p1, C_p2 et C_pV, le meilleur estimateur de $c^\top\beta$ est $c^\top\tilde{\beta}$ où $\tilde{\beta} = (X^\top V^{-1}X)^{-1}X^\top V^{-1}\boldsymbol{Y}$.*

Preuve : Soit $\gamma^* = a^\top\boldsymbol{Y}$ un estimateur linéaire de $\gamma = c^\top\beta$ où a est un vecteur (non aléatoire) de $\mathbb{R}^n$ (voir la définition 5.1). γ^* est un estimateur sans biais de γ ssi $\mathrm{E}(\gamma^*) = \gamma$, $\forall\gamma \in \mathbb{R}$. En utilisant les expressions de γ^* et de γ ainsi que les conditions C_p1 et C_p2, l'absence de biais équivaut à $(a^\top X - c^\top)\beta = 0$, $\forall\beta \in \mathbb{R}^{p+1}$, ou encore $a^\top X - c^\top = 0_{p+1}^\top$. La variance d'un tel estimateur est $\mathrm{V}(\gamma^*) = \mathrm{V}(a^\top\boldsymbol{Y}) = a^\top\mathrm{V}(\boldsymbol{Y})a = \sigma^2 a^\top V a$. Par conséquent chercher le meilleur revient à chercher le vecteur $\tilde{a}$ de $\mathbb{R}^n$ tel que $\tilde{a}^\top X - c^\top = 0_{p+1}^\top$ et $\tilde{a}^\top V\tilde{a} \leq a^\top V a$ pour tout vecteur $a \in \mathbb{R}^n$ tel que $a^\top X - c^\top = 0_{p+1}^\top$. Cet estimateur sera alors donné par $\tilde{\gamma} = \tilde{a}^\top\boldsymbol{Y}$. Le vecteur $\tilde{a}$ est donc la solution du problème de minimisation

$$\min_{a\in\mathbb{R}^n} a^\top V a \quad \text{s.c.q.} \quad a^\top X - c^\top = 0_{p+1}^\top \tag{8.4}$$

eq:mcg_min

Une condition nécessaire pour que $\tilde{a}$ soit solution est qu'il existe un vecteur $\tilde{\lambda} \in \mathbb{R}^{p+1}$ tel que

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial a}(\tilde{a},\tilde{\lambda}) = 0_n & \text{(8.5)} \\ \dfrac{\partial\mathcal{L}}{\partial\lambda}(\tilde{a},\tilde{\lambda}) = 0_{p+1} & \text{(8.6)} \end{cases}$$

eq:mcg_nblue_a

eq:mcg_nblue_l

où $\mathcal{L}(a,\lambda) = a^\top V a - (a^\top X - c^\top)\lambda$ est le lagrangien associé au problème de minimisation. Les n équations (8.5) s'écrivent $V\tilde{a} - X\tilde{\lambda} = 0_n$ et (8.6) exprime la contrainte d'absence de biais $X^\top\tilde{a} - c = 0_{p+1}$.[4] D'après $\mathrm{C}_p\mathrm{V}$, la matrice V est inversible. On peut donc réécrire (8.5) comme $\tilde{a} = V^{-1}X\tilde{\lambda}$. Si on utilise cette expression de $\tilde{a}$, on peut réécrire (8.6) sous la forme $X^\top V^{-1}X\tilde{\lambda} - c = 0_{p+1}$. Comme X est de rang $p+1$, la matrice $X^\top V^{-1}X$ est inversible et on obtient alors $\tilde{\lambda} = (X^\top V^{-1}X)^{-1}c$. En substituant cette expression de $\tilde{\lambda}$

4. Voir par exemple les calculs détaillés à la remarque 5.11.

dans celle de $\tilde{a}$, on a finalement $\tilde{a} = V^{-1}X(X^\top V^{-1}X)^{-1}c$. Par conséquent l'estimateur linéaire sans biais de variance minimale est $\tilde{\gamma} = \tilde{a}^\top \boldsymbol{Y} = c^\top \tilde{\beta}$ où $\tilde{\beta}$ est défini dans l'énoncé de la propriété. Finalement, comme V est définie positive et que la fonction $a^\top X - c^\top$ qui exprime la contrainte est linéaire en a, le vecteur $\tilde{a}$ obtenu en résolvant le système (8.5)-(8.6) est bien une solution du problème de minimisation. ■

Ce résultat est à contraster avec le résultat de la remarque 5.11 qui établit que dans le cadre du modèle de régression linéaire standard (défini par les conditions C_p1, C_p2 et C_p3), le meilleur estimateur linéaire sans biais de $\gamma = c^\top\beta$ s'obtient à partir de l'estimateur des moindres carrés ordinaires de β, et est donné par $\hat{\gamma} = c^\top\hat{\beta} = c^\top(X^\top X)^{-1}X^\top\boldsymbol{Y}$. Ainsi, si on s'intéresse à la dernière coordonnée β_p de β, alors $c = (0,\ldots,0,1)^\top$ et dans ce cas le meilleur estimateur sans biais de β_p est $\tilde{\gamma} = \tilde{\beta}_p$, où $\tilde{\beta}_p$ est la dernière coordonnée de $\tilde{\beta} = (X^\top V^{-1}X)^\top XV^{-1}\boldsymbol{Y}$. De manière générale, cette coordonnée diffère de $\hat{\beta}_p$, l'estimateur des moindres carrés ordinaires de β_p, défini comme la dernière coordonnée de $\hat{\beta} = (X^\top X)^{-1}X^\top\boldsymbol{Y}$. Ce résultat montre donc que dans le modèle de régression linéaire dans lequel on a $\mathrm{C}_p\mathrm{V}$, l'estimateur des moindres carrés ordinaires ne coïncide pas avec le meilleur estimateur linéaire sans biais. On construira plus loin l'estimateur dont on montrera qu'il est optimal dans l'ensemble des estimateurs linéaires sans biais.

Finalement, on peut montrer qu'en introduisant une condition supplémentaire sur la matrice V, on retrouve le résultat de la propriété 7.1 : l'estimateur $\hat{\beta}$ reste un estimateur convergent de β. La convergence est une propriété limite de l'estimateur des moindres carrés ordinaires de β lorsque $n \to \infty$. Afin de le formuler (et le prouver) on modifie légèrement les notations afin de faire apparaître la dépendance des éléments du modèle vis-à-vis de n. Pour une taille d'échantillon n donnée, on note $V_n = \mathrm{V}(\boldsymbol{Y})$ et $\hat{\beta}_n$ l'estimateur des moindres carrés ordinaires de β.

pro:mrlns.conv.mco

Propriété 8.3 *Dans le modèle de régression linéaire défini par les conditions C_p1, C_p2 et $\mathrm{C}_p\mathrm{V}$, pour chaque taille d'échantillon n on note v_n^* la plus grande valeur propre de V_n. Si*

1. *$\exists K \in \mathbb{R}$ tel que $v_n^* \leq K$, $\forall n \in \mathbb{N}$*
2. *$(X^\top X)^{-1} \to 0$ lorsque $n \to \infty$*

alors $\hat{\beta}_n$ converge vers β en moyenne quadratique, et donc en probabilité.

Preuve : En adoptant la même démarche que dans la preuve de la propriété 7.1 et en utilisant l'expression de $\mathrm{V}(\hat{\beta})$ donnée par (8.3), on a $\mathrm{E}(\|\hat{\beta}_n - \beta\|^2) = \mathrm{trace}\big((X^\top X)^{-1}X^\top VX(X^\top X)^{-1}\big)$. En utilisant la propriété $\mathrm{trace}(AB) = \mathrm{trace}(BA)$ lorsque les produits AB et BA sont définis, on a

$$\mathrm{trace}\big((X^\top X)^{-1}X^\top VX(X^\top X)^{-1}\big) = \mathrm{trace}\big(X(X^\top X)^{-2}X^\top V\big)$$

où $(X^\top X)^{-2} = \big((X^\top X)^{-1}\big)^2$. Comme V est symétrique définie positive, on peut écrire $V = \Gamma\Upsilon\Gamma^\top$ où $\Upsilon = \mathrm{diag}(v_1,\ldots,v_n)$ est la matrice diagonale des valeurs propres de V et Γ est la matrice orthonormale des vecteurs propres de V. Donc

$$\begin{aligned}\mathrm{trace}\big(X(X^\top X)^{-2}X^\top V\big) &= \mathrm{trace}\big(X(X^\top X)^{-2}X^\top\Gamma\Upsilon\Gamma^\top\big)\\ &= \mathrm{trace}\big(\Gamma^\top X(X^\top X)^{-2}X^\top\Gamma\Upsilon\big)\\ &= \mathrm{trace}(C\Upsilon)\end{aligned}$$

où $C = \Gamma^\top X(X^\top X)^{-2}X^\top\Gamma = (c_{ij})_{i,j=1,\dots,n}$. Puisque Υ est diagonale, les termes diagonaux de $C\Upsilon$ sont $v_i c_{ii}$, $i = 1, \dots, n$. Par conséquent

$$\begin{aligned}
\operatorname{trace}(C\Upsilon) = \sum_{i=1}^n v_i c_{ii} \leq \max\{v_1, \dots, v_n\} \sum_{i=1}^n c_{ii} &= v_n^* \sum_{i=1}^n c_{ii} \\
&= v_n^* \sum_{i=1}^n \operatorname{trace}(C) \\
&= v_n^* \operatorname{trace}(\Gamma^\top X(X^\top X)^{-2}X^\top\Gamma) \\
&= v_n^* \operatorname{trace}(X(X^\top X)^{-2}X^\top\Gamma\Gamma^\top) \\
&= v_n^* \operatorname{trace}(X(X^\top X)^{-2}X^\top) \\
&= v_n^* \operatorname{trace}((X^\top X)^{-2}X^\top X) \\
&= v_n^* \operatorname{trace}((X^\top X)^{-1})
\end{aligned}$$

Si on note λ_n^* la plus grande des valeurs propres de $(X^\top X)^{-1}$ on a $0 \leq \operatorname{trace}((X^\top X)^{-1}) \leq (p+1)\lambda_n^*$ et donc $\operatorname{trace}(C\Upsilon) \leq (p+1)v_n^*\lambda_n^*$. Les deux conditions de l'énoncé garantissent que le membre de gauche converge vers 0 lorsque $n \to \infty$ (voir également la section 9.3).■

On peut introduire ici la même remarque que celle faite à propos de la convergence de $\hat\beta$ dans le contexte du modèle standard. La propriété 8.3 est souvent énoncée en remplaçant la condition 2 par la condition $\frac{X^\top X}{n} \to A$, où A est une matrice symétrique définie positive. Cette condition est plus forte que la condition utilisée dans la propriété 8.3, c'est à dire pour tout $\nu > 0$:

$$\frac{X^\top X}{n^\nu} \to A \implies (X^\top X)^{-1} \to 0$$

avec A symétrique définie positive. En effet, supposons que $\frac{X^\top X}{n^\nu} \to A$. Comme A est inversible et que l'inversion de matrice est continue (voir théorème 9.1), on doit avoir $n^\nu(X^\top X)^{-1} \to A^{-1}$. Ceci implique que $n^\nu\|(X^\top X)^{-1}\| \to \|A^{-1}\|$. Comme $\|A^{-1}\|$ est fini, on doit avoir $\|(X^\top X)^{-1}\| \to 0$, ce qui est la condition 2 de la propriété.

sec:mcg

8.3 Moindres carrés généralisés (MCG)

Les résultats de la section précédente montrent que si l'estimateur des moindres carrés ordinaires reste sans biais et convergent, sa matrice de variances-covariances n'est pas la même que dans le cas du MRLS. Il faudra donc adapter les procédures de test et de régions de confiance afin de tenir compte de la nouvelle forme de cette matrice. Mais de manière plus importante, pour estimer β, les moindres carrés ordinaires ne fournissent pas le meilleur estimateur linéaire sans biais. La propriété 8.2 semble suggérer que le meilleur estimateur linéaire sans biais est $\tilde\beta = (X^\top V^{-1}X)^{-1}X^\top V^{-1}\boldsymbol{Y}$. Évidemment, pour que $\tilde\beta$ puisse être envisagé comme estimateur de β, il faut se placer dans le cas où V est connue. On supposera que c'est le cas dans cette section. Par conséquent, le modèle est défini par la condition (8.1). Dans un premier temps, on démontre formellement que $\tilde\beta$ est bien le meilleur estimateur linéaire de β. Puis on décrit les conséquences de ce résultat sur les tests et régions de confiance construits à propos de β.

sec:mcg.blue

8.3.1 Estimation de β par MCG

Pour montrer que le meilleur estimateur linéaire sans biais de β est $(X^\top V^{-1}X)^{-1}X^\top V^{-1}\boldsymbol{Y}$, on peut adopter deux démarches distinctes, qu'on présente successivement.

La première manière d'obtenir ce résultat consiste à utiliser la propriété 8.2. Puisque $\tilde{\beta} = (X^\top V^{-1}X)^{-1}X^\top V^{-1}\boldsymbol{Y} = \tilde{A}\boldsymbol{Y}$, avec $\tilde{A} = (X^\top V^{-1}X)^{-1}X^\top V^{-1}$, on voit que $\tilde{\beta}$ est un estimateur linéaire de β. Comme de plus $\tilde{A}X = I_n$, il est également sans biais. Soit $\check{\beta} = \check{A}\boldsymbol{Y}$, un autre estimateur linéaire et sans biais de β, *i.e.*, la matrice $\check{A}$ satisfait $\check{A}X = I_n$. On veut montrer $\tilde{\beta}$ est plus précis que $\check{\beta}$, c'est à dire :

$$c^\top[\mathrm{V}(\tilde{\beta}) - \mathrm{V}(\check{\beta})]c \geq 0, \quad \forall c \in \mathbb{R}^{p+1}$$

En utilisant les propriétés des matrices de variances-covariances (voir la propriété 9.7), cette inégalité s'écrit aussi

$$\mathrm{V}(c^\top\tilde{\beta}) - \mathrm{V}(c^\top\check{\beta}) \geq 0, \quad \forall c \in \mathbb{R}^{p+1} \tag{8.7}$$

eq:min_fq_mcg

Pour un $c \in \mathbb{R}^{p+1}$ quelconque, on définit $\tilde{a} = \tilde{A}^\top c$ et $\check{a} = \check{A}^\top c$. On vérifie facilement que, d'une part,

$$\mathrm{V}(c^\top\tilde{\beta}) = \tilde{a}^\top V\tilde{a} \qquad \text{et} \qquad \mathrm{V}(c^\top\check{\beta}) = \check{a}^\top V\check{a} \tag{8.8}$$

eq:vmgc

et d'autre part

$$\tilde{a}^\top X - c^\top = 0_{p+1}^\top \qquad \text{et} \qquad \check{a}^\top X - c^\top = 0_{p+1}^\top$$

(puisque $X\tilde{A} = X\check{A} = I_n$). Or la résolution du problème de minimisation (8.4) dans la preuve de la propriété 8.2 montre qu'on a nécessairement

$$\check{a}^\top V\check{a} \geq \tilde{a}^\top V\tilde{a}$$

ce qui, étant donné (8.8), est précisément l'inégalité (8.7) voulue pour le vecteur c choisi. Mais comme ceci peut être établi pour n'importe quel choix de c, on a bien (8.7).

La seconde manière d'obtenir ce même résultat consiste à formuler un modèle équivalent à celui défini par C_p1, C_p2 et $\mathrm{C}_p\mathrm{V}$, mais dans lequel on retrouve la condition C_p3 et dont β en est le paramètre. Pour cela, choisissons une matrice M de taille (n, n) non aléatoire et inversible, quelconque. On a alors

$$\begin{aligned}
&X \text{ non aléatoire} \iff MX \text{ non aléatoire}\\
&\mathrm{E}(\boldsymbol{Y}) = X\beta \iff \mathrm{E}(M\boldsymbol{Y}) = MX\beta\\
&\mathrm{V}(\boldsymbol{Y}) = \Omega \iff \mathrm{V}(M\boldsymbol{Y}) = \tilde{\Omega}
\end{aligned}$$

où $\tilde{\Omega} = M\Omega M^\top$. De plus, on vérifie facilement que Ω est symétrique définie positive si et seulement si $\tilde{\Omega}$ est symétrique définie positive. Autrement dit, en définissant $\tilde{X} = MX$ et $\tilde{\boldsymbol{Y}} = M\boldsymbol{Y}$ on voit que X et $\boldsymbol{Y}$ satisfont C_p1, C_p2 et $\mathrm{C}_p\mathrm{V}$, si et seulement si seulement si $\tilde{X}$ et $\tilde{\boldsymbol{Y}}$ satisfont ces mêmes conditions. On constate que le vecteur β de $\mathbb{R}^{p+1}$ qui permet d'écrire la condition C_p2 pour X et $\boldsymbol{Y}$ est le même que celui qui permet d'écrire cette même condition pour $\tilde{X}$ et $\tilde{\boldsymbol{Y}}$ et que les matrices

symétriques définies positives qui permettent d'exprimer la condition C_p3 sont liées par la relation $\tilde{\Omega} = M\Omega M^\top$.

Ce qui vient d'être dit permet de considérer indifféremment l'estimation de β au sein du modèle initial relatif à X et $\boldsymbol{Y}$ ou bien au sein du modèle relatif à $\tilde{X}$ et $\tilde{\boldsymbol{Y}}$, puisque ces modèles satisfont les mêmes conditions et que β paramétrise de manière identique ces deux modèles.

Par ailleurs, l'ensemble des estimateurs linéaires et sans biais de β de l'un des modèles coïncide avec celui de l'autre. En effet, soit $\beta^* = A^*\boldsymbol{Y}$, avec $A^*X = I_n$ un estimateur linéaire sans biais de β dans le modèle initial. On peut écrire $\beta^* = A^*M^{-1}M\boldsymbol{Y} = \tilde{A}^*\tilde{\boldsymbol{Y}}$ avec $\tilde{A}^* = A^*M^{-1}$, ce qui montre que β^* est un estimateur linéaire de β dans le modèle relatif à $\tilde{X}$ et $\tilde{\boldsymbol{Y}}$. De plus, on vérifie facilement que $\tilde{A}^*\tilde{X} = I_n$, ce qui est la condition d'absence de biais d'un estimateur linéaire dans le modèle relatif à $\tilde{X}$ et $\tilde{\boldsymbol{Y}}$. La réciproque s'obtient de manière identique.

Donc pour rechercher le meilleur estimateur linéaire et sans biais de β, on peut indifféremment utiliser l'un ou l'autre modèle.

Comme tout ce qui vient d'être dit est valable pour n'importe quel choix de matrice M de taille (n, n), inversible et non-aléatoire, on peut se demander s'il en existe une telle que dans le modèle relatif à $\tilde{X} = MX$ et $\tilde{\boldsymbol{Y}} = M\boldsymbol{Y}$, il est facile de déterminer le meilleur estimateur linéaire et sans biais de β. Une manière de répondre à cette question consiste à se demander si on peut trouver une matrice M inversible telle que la matrice $\tilde{\Omega}$ des variances-covariances de $\tilde{\boldsymbol{Y}}$ prenne la forme $\tilde{\Omega} = \sigma^2 I_n$ pour un $\sigma \in]0, \infty[$. Si la réponse est positive, alors le modèle relatif à $\tilde{X}$ et $\tilde{\boldsymbol{Y}}$ est un modèle de régression linéaire standard. Dans un tel cas, le meilleur estimateur linéaire sans biais de β coïncidera avec l'estimateur des MCO de β dans ce modèle, *i.e.*, $(\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\tilde{\boldsymbol{Y}}$.

Pour trouver une telle matrice M, on part de la condition recherchée $\tilde{\Omega} = \sigma^2 I_n$, ou de manière équivalente, $\mathrm{V}(M\boldsymbol{Y}) = \sigma^2 I_n$ pour un $\sigma \in]0, \infty[$. Comme $\mathrm{V}(\boldsymbol{Y}) = \Omega$ et qu'on a choisi d'écrire Ω sous la forme $\Omega = \sigma^2 V$ pour un $\sigma \in]0, \infty[$ et V symétrique définie positive, la condition à obtenir s'écrit ceci s'écrit aussi $MVM^\top = I_n$. Puisque V est symétrique définie positive, elle peut s'écrire $V = \Gamma\Upsilon\Gamma^\top$ où Γ est une matrice orthonormée et $\Upsilon = \mathrm{diag}(\upsilon_i, i = 1, \ldots, n)$ est une matrice diagonale avec $\upsilon_i > 0$, $i = 1, \ldots, n$. On peut donc écrire $\Upsilon = \Upsilon^{1/2}\Upsilon^{1/2}$, où $\Upsilon^{1/2} = \mathrm{diag}(\sqrt{\upsilon_i}, i = 1, \ldots, n)$. Par conséquent, $\Upsilon^{-1} = (\Upsilon^{1/2}\Upsilon^{1/2})^{-1} = \Upsilon^{-1/2}\Upsilon^{-1/2}$, où $\Upsilon^{-1/2}$ est l'inverse de $\Upsilon^{1/2}$, *i.e.*, $\Upsilon^{-1/2} = \mathrm{diag}(\frac{1}{\sqrt{\upsilon_i}}, i = 1, \ldots, n)$. Si on choisit

$$M = \Upsilon^{-1/2}\Gamma^\top \tag{8.9}$$ eq:Mspher

on a bien une matrice non aléatoire inversible pour laquelle

$$MVM^\top = \Upsilon^{-1/2}\Gamma^\top V\Gamma\Upsilon^{-1/2} = \Upsilon^{-1/2}\Gamma^\top(\Gamma\Upsilon^{1/2}\Upsilon^{1/2}\Gamma^\top)\Gamma\Upsilon^{-1/2} = I_n \tag{8.10}$$ eq:vspher

(la dernière égalité utilisant l'orthonormalité de Γ).

Le choix de transformation M définie par (8.9) permet de passer du modèle de régression initial relatif à X et $\boldsymbol{Y}$ à un modèle de régression équivalent relatif à des variables $\tilde{X} = MX$ et $\tilde{\boldsymbol{Y}} = M\boldsymbol{Y}$, dans lequel $\mathrm{E}(\tilde{Y}) = \tilde{X}\beta$ et $\mathrm{V}(\tilde{\boldsymbol{Y}}) = \sigma^2 I_n$. Grâce à cette dernière condition, ce second modèle est un modèle de régression linéaire standard. On remarque que la transformation M obtenue consiste à transformer les variables initiales de manière (1) à préserver la relation donnée par la condition C_p2 (ou C'_p2) et (2) à rendre sphériques les erreurs définies dans ce nouveau modèle (voir la première

section de ce chapitre). En effet, le vecteur des erreurs dans le modèle transformé est $\tilde{\varepsilon} = \tilde{\boldsymbol{Y}} - \mathrm{E}(\tilde{\boldsymbol{Y}}) = M\boldsymbol{Y} - MX\beta$ et en utilisant (8.10) on a $\mathrm{V}(\tilde{\varepsilon}) = \mathrm{V}(M\boldsymbol{Y}) = M\mathrm{V}(\boldsymbol{Y})M^\top = \sigma^2 I_n$. Pour cette raison, on appelle *sphéricisation* (des erreurs) la transformation consistant à prémultiplier le vecteur $\boldsymbol{Y}$ et la matrice X des observations des variables par la matrice M définie en (8.9).

Comme indiqué précédemment, le modèle obtenu par sphéricisation est un modèle de régression linéaire standard pour les variables $\tilde{\boldsymbol{Y}}$ et $\tilde{X}$. Il permet d'obtenir le meilleur estimateur linéaire sans biais de β en utilisant la méthode des MCO. Cet estimateur est noté $\tilde{\beta}$ et son expression est

$$\tilde{\beta} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{\boldsymbol{Y}}$$

où $\tilde{X} = MX$, $\tilde{\boldsymbol{Y}} = M\boldsymbol{Y}$ et M est donnée par (8.9). En utilisant ces expressions, on peut écrire $M^\top M = \Gamma\Upsilon^{-1/2}\Upsilon^{-1/2}\Gamma^\top = \Gamma\Upsilon^{-1}\Gamma^\top = V^{-1}$ et donc

$$\tilde{\beta} = (X^\top M^\top MX)^{-1} X^\top M^\top M\boldsymbol{Y} = (X^\top V^{-1} X)^{-1} X^\top V^{-1}\boldsymbol{Y}$$

On constate qu'on obtient le même résultat que la première approche proposée, basée sur l'utilisation de la propriété 8.2.

On résume ces résultats sous la forme d'une propriété.

pro:mcgs **Propriété 8.4** *Dans le modèle de régression défini par les conditions C_p1, C_p2 et C_pV, le meilleur estimateur linéaire sans biais de β est $(X^\top V^{-1}X)^{-1}X^\top V^{-1}\boldsymbol{Y}$, où V est la matrice pour laquelle $\mathrm{V}(\boldsymbol{Y}) = \sigma^2 V$.*

Cet estimateur coïncide avec l'estimateur des MCO dans le modèle de régression sphéricisé relatif à $\tilde{X} = MX$ et $\tilde{\boldsymbol{Y}} = MY$, où $M = \Upsilon^{-1/2}\Gamma^\top$, Γ et Υ étant respectivement les matrices des vecteurs propres et valeurs propres de V.

On constate que l'estimateur de β obtenu consiste à appliquer les moindres carrés ordinaires au modèle obtenu par transformation des variables initiales au moyen de M. Dans ce modèle, les matrices des observations des variables sont donc $\tilde{X} = MX$ et $\tilde{\boldsymbol{Y}} = \boldsymbol{Y}$ et l'estimateur des MCO est défini comme la solution de $\min_{\beta\in\mathbb{R}^{p+1}} \|\tilde{\boldsymbol{Y}} - \tilde{X}\beta\|^2$ (voir la section 5.3.2). En utilisant la définition de $\tilde{X}$ et $\tilde{\boldsymbol{Y}}$, ceci revient à résoudre $\min_{\beta\in\mathbb{R}^{p+1}} \|M(\boldsymbol{Y} - X\beta)\|^2$. Il est évident qu'il revient au même de chercher β en minimisant $\|M(\boldsymbol{Y} - X\beta)\|$. Or M étant inversible, cette norme est une distance entre les vecteurs $\boldsymbol{Y}$ et $X\beta$.[5] Par conséquent, au sein du modèle initial relatif aux variables X et $\boldsymbol{Y}$, l'estimateur de la propriété 8.4 possède la même interprétation que l'estimateur des MCO $\hat{\beta}$ solution de $\min_{\beta\in\mathbb{R}^{p+1}} \|\boldsymbol{Y} - X\beta\|^2$: il définit l'élément de $L(X_{\cdot 0}, \ldots, X_{\cdot p})$ le plus proche de $\boldsymbol{Y}$. La différence entre cet estimateur et celui des MCO réside dans la distance utilisée pour mesurer cette proximité. Dans un cas (MCO) il s'agit de la distance "usuelle", correspondant au choix $M = I_n$, tandis que dans le second cas il s'agit d'une généralisation de cette distance usuelle, correspondant au choix $M = \Upsilon^{-1/2}\Gamma^\top$. Pour cette raison, on introduit la définition suivante.

Définition 8.1 *L'estimateur de β défini dans la propriété 8.4 est appelé estimateur des* moindres carrés généralisés. *On le note $\hat{\beta}_{MCG}$. On a donc*

$$\hat{\beta}_{MCG} = (X^\top V^{-1}X)^{-1}X^\top V^{-1}\boldsymbol{Y}$$

5. Il est facile de vérifier que si M est une matrice réelle (n, n) inversible, alors l'application $(u, v) \in \mathbb{R}^n \times \mathbb{R}^n \to \|M(u - v)\|$ est une distance sur $\mathbb{R}^n$: elle est positive si $u \neq v$ et nulle sinon, symétrique et vérifie l'inégalité triangulaire.

sec:mcg.csqces

8.3.2 Utilisations de l'estimateur MCG de β

Le fait de voir $\hat{\beta}_{\text{MCG}}$ comme un estimateur MCO dans le modèle relatif aux variables transformées $\tilde{X}$ et $\tilde{\boldsymbol{Y}}$ permet d'obtenir les propriétés de $\hat{\beta}_{\text{MCG}}$ (et des statistiques associées) directement à partir des résultats du chapitre 5 propres aux modèles de régression linéaire standards.

Cependant, lorsqu'on cherche à interpréter les résultats d'une estimation par MCG il faut bien se placer dans le modèle initial relatif aux variables X et $\boldsymbol{Y}$, puisque ces dernières sont les seules qu'on soit capable d'interpréter (contrairement donc à $\tilde{X}$ et $\tilde{\boldsymbol{Y}}$. Les notions qu'il est pertinent d'introduire et d'étudier doivent donc tenir compte de cela. C'est en particulier le cas des valeurs ajustées

sec:mcg.resid

8.3.2.1 Valeurs ajustées, résidus. Estimation de σ

La notion de valeur ajustée est introduite pour tenter d'approcher $X\beta$, la partie de la variable endogène expliquée par les variables exogènes. Si on se plaçait directement dans le modèle transformé servant à obtenir l'estimateur $\hat{\beta}_{\text{MCG}}$, on définirait le vecteur des valeurs ajustées comme $\tilde{X}\hat{\beta}_{\text{MCG}}$. Une telle définition n'a pas d'intérêt au vu de ce que qu'on souhaite approcher au moyen des valeurs ajustées. La bonne définition est donc la suivante :

Définition 8.2 *La i^e valeur ajustée issue de l'estimation par MCG est la variable aléatoire notée $\hat{Y}_i$ définie par $\hat{Y}_i = X_{i\bullet}^\top \hat{\beta}_{MCG}$. Le vecteur des valeurs ajustées est le vecteur aléatoire $\hat{\boldsymbol{Y}}$ de $\mathbb{R}^n$ dont les coordonnées sont les n valeurs ajustées. On a $\hat{\boldsymbol{Y}} = X\hat{\beta}_{MCG}$.*

On définit aussi le i^{e} résidu $\hat{\varepsilon}_i$ comme l'estimation de la partie de Y_i qu'on ne peut expliquer par les variables exogènes. On a donc dans le contexte d'une estimation par MCG $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ et le vecteur des résidus est $\hat{\varepsilon} = \boldsymbol{Y} - X\hat{\beta}_{\text{MCG}}$.

L'interprétation des valeurs ajustées et des résidus est rigoureusement identique à celle donnée à la section 2.4.1 dans le contexte de l'estimation du MRLS par MCO.

sec:mcg.test

8.3.2.2 Tests d'hypothèses

`Incomplet : ETA unknown`

Chapitre 9

Compléments

sec:nor

9.1 Lois normales et lois déduites de la loi normale

sec:nor_univ

9.1.1 Lois normales univariées

def:loi_nor1

Définition 9.1 *On dit que la variable aléatoire X suit une loi normale (ou gaussienne) s'il existe un réel μ quelconque et un réel σ strictement positif tels que la fonction de répartition F_X de X s'écrit*

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} \mathrm{d}t$$

On dit dans ce cas que X est normale (ou gaussienne) et on note $X \sim \mathcal{N}(\mu, \sigma^2)$.

rem:nor

Remarque 9.1 X est gaussienne ssi il existe $\mu \in \mathbb{R}$ et $\sigma \in]0, \infty[$ tels que X est une variable aléatoire réelle continue de densité :

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{9.1}$$

eq:densnor

Cette fonction étant une fonction de densité, on a $f_X \geq 0$, ce qu'on vérifie aisément, et

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x = \sqrt{2\pi}\,\sigma \tag{9.2}$$

eq:int_n01

pour tout μ et tout $\sigma > 0$. Cette dernière égalité sera admise. Pour en trouver la preuve, faire une recherche internet sur « intégrale de Gauss ».

Comme la densité d'une variable aléatoire permet de connaître sa loi et que f_X n'est paramétrée que par les réels μ et σ, on voit que la loi d'une v.a.r. gaussienne est connue dès que ces deux réels le sont. ℝ

La fonction f_X de la définition 9.1 possède les propriétés suivantes.

Propriété 9.1

1. $f_X(x) > 0$, $\forall x \in \mathbb{R}$
2. $\lim_{x\to-\infty} f_X(x) = 0 = \lim_{x\to\infty} f_X(x)$

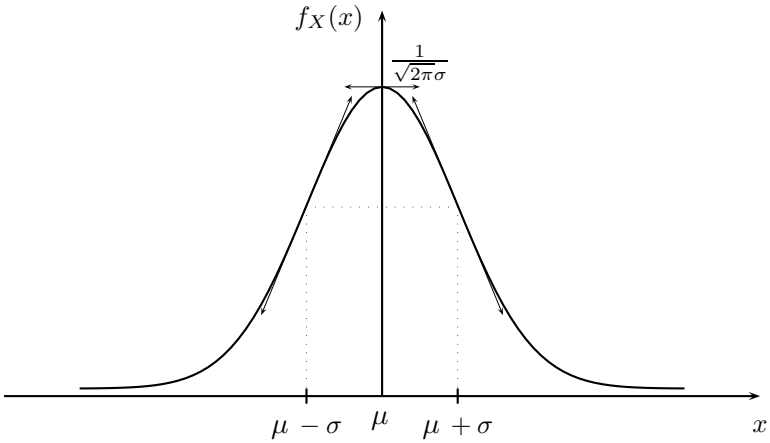


FIGURE 9.1: Courbe de la densité f_X de la v.a.r. $X \sim \mathcal{N}(\mu, \sigma^2)$

fig:dens_norm

3. *f_X admet un maximum unique en μ :* $\frac{1}{\sqrt{2\pi}\,\sigma} = f_X(\mu) > f_X(x)$, $\forall x \neq \mu$
4. *f_X possède deux points d'inflexion, en $\mu - \sigma$ et $\mu + \sigma$*
5. *f_X admet un axe de symétrie d'équation $x = \mu$:* $f_X(\mu - x) = f_X(\mu + x)$ $\forall x \in \mathbb{R}$

La preuve de ces propriétés s'obtient par une étude classique de la fonction f_X (valeurs, dérivées première et seconde, limites). Elles sont résumées par la figure 9.1.

On remarque que lorsque $\sigma \to 0$, le maximum de f_X tend vers l'infini et ses points d'inflexion tendent vers μ, tous deux de manière continue et monotone. Autrement dit, si σ tend vers 0, la forme en cloche de la courbe de f_X devient plus étroite et plus haute. Graphiquement, cette propriété est illustrée par la figure 9.2 qui représente la densité de la loi normale d'espérance 0, pour plusieurs valeurs différentes de σ. On constate que plus σ est proche de 0, plus la courbe de la densité de

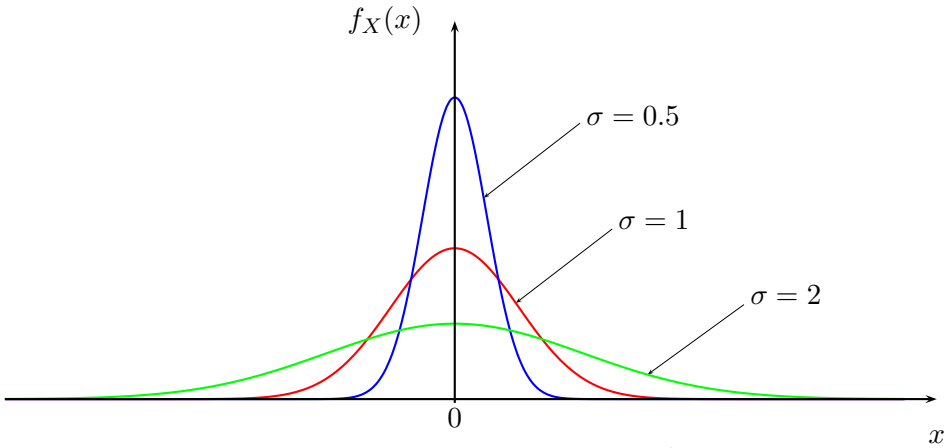


FIGURE 9.2: Forme de la densité de la loi $\mathcal{N}(0, \sigma^2)$ en fonction de σ

fig:densnor_sig

X est tassée autour de l'espérance. En conséquence, comme la surface sous la courbe de la densité vaut toujours 1 quel que soit σ, pour n'importe quel réel $a > \mu$, la probabilité pour que X dépasse a doit tendre vers 0 lorsque σ tend vers 0. C'est ce qui est illustré par la figure 9.3. La probabilité $\mathrm{P}(X > a)$ est la surface sous la courbe de densité à droite de a. Pour une valeur de σ correspondant

à la courbe rouge, cette probabilité est la somme des surfaces bleue et rouge. Pour une valeur de σ plus petite associé à la courbe bleue, cette probabilité est plus faible et n'est égale qu'à la surface bleue. Comme ceci est vrai pour tout $a > \mu$, la probabilité pour que X dépasse μ est nulle à la

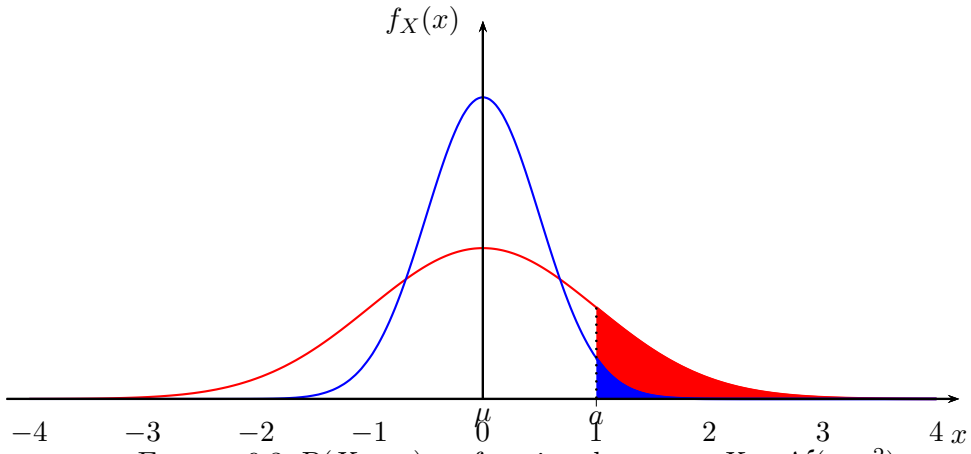


FIGURE 9.3: $\mathrm{P}(X > a)$ en fonction de σ pour $X \sim \mathcal{N}(\mu, \sigma^2)$

fig:fdrnor_sig

limite (quand $\sigma \to 0$). Par un raisonnement analogue qui exploite la symétrie de f_X autour de l'axe vertical d'équation $x = \mu$, la probabilité pour que X soit plus petite que μ est également nulle à la limite. Donc lorsque σ tend vers 0, la loi de $X \sim \mathcal{N}(\mu, \sigma^2)$ tend vers la loi d'une variable aléatoire Y telle que $\mathrm{P}(Y < \mu) = \mathrm{P}(Y > \mu) = 0$. Ce résultat déduit de l'observation des graphiques des figures 9.2 et 9.3 se démontre formellement. [1]

pro:nor_var0

Propriété 9.2 *Soit un réel μ et Y la variable aléatoire telle que $\mathrm{P}(Y = \mu) = 1$. On note F_μ la fonction de répartition de Y,* i.e., *$F_\mu(x) = 0$ si $x < \mu$ et $F_\mu(x) = 1$ sinon. Soit X une variable aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$. Lorsque σ tend vers 0, X tend en loi vers Y : $\lim_{\sigma\to 0} F_X(x) = F_\mu(x)$ pour tout $x \neq \mu$.*

Preuve : Pour montrer que $F_X(x) \to F_\mu(x)$, il faut établir la limite de l'intégrale $F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$. Pour cela, on utilise un théorème qui permet de permuter la limite et l'intégrale, et d'écrire que $\lim_{\sigma\to 0} \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = \int_{-\infty}^{x} \lim_{\sigma\to 0} f_X(t)\,\mathrm{d}t$. Le théorème utilisé est le théorème de convergence dominée. Il établit que si $\{g_n : n \geq 1\}$ est une suite de fonctions qui converge ponctuellement vers g_∞, et pour laquelle il existe une fonction g telle que $\int |g(x)|\,\mathrm{d}x < \infty$ et $|g_n(x)| \leq |g(x)|$ $\forall x \in \mathbb{R}$ et $\forall n > N$ pour un certain rang N, alors la limite de l'intégrale des g_n est égale à l'intégrale de g_∞.

Comme dans l'étude de la convergence, $F_\mu(x)$ ne prend que deux valeurs possibles selon que $x > \mu$ ou $x < \mu$, on distingue deux cas. On commence par choisir un $x > \mu$ et on va montrer que $\lim_{\sigma\to 0} F_X(x) = F_\mu(x) = 1$, ou de manière équivalente que $1 - \lim_{\sigma\to 0} F_X(x) = 1 - F_\mu(x) = 0$, ou encore que $\lim_{\sigma\to 0} \int_x^\infty f_X(t)\,\mathrm{d}t = 0$. Pour cela, on utilise le théorème de convergence dominée dans lequel $g_n(t) = \frac{1}{\sqrt{2\pi}\frac{1}{n}} e^{-\frac{1}{2}\left(\frac{t-\mu}{1/n}\right)^2}$. Autrement dit on considère la densité de X avec $\sigma = 1/n$. Il suffit de montrer que $\lim_{n\to\infty} \int_x^\infty g_n(t)\,\mathrm{d}t = 0$. On a

1. La preuve de ce résultat s'obtient facilement en utilisant les propriétés des fonctions caractéristiques. En l'absence d'une présentation de ce type de fonctions, la démonstration s'appuiera sur la définition 9.1 et sera un peu plus longue que celle habituellement proposée.

$g_n(t) \to g_\infty(t) = 0\ \forall t$. D'autre part, pour N suffisamment grand, on a $0 \leq g_n(t) \leq g_N(t)$, $\forall t \geq x$, $\forall n \geq N$. Pour le montrer, on remarque que

$$\frac{\mathrm{d}f_X}{\mathrm{d}\sigma}(t) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} \left[\left(\frac{t-\mu}{\sigma}\right)^2 - 1\right]$$

est du signe de $\left(\frac{t-\mu}{\sigma}\right)^2 - 1$. On peut toujours trouver $\sigma = \frac{1}{N}$ suffisamment petit (et donc un N suffisamment grand) tel que cette dérivée est positive en $t = x$. Elle le restera pour tout $t \geq x$ et pour tout $\sigma = \frac{1}{n}$ avec $n \geq N$. Donc pour tout $t \geq x$ on aura $n \geq N \implies g_n(t) \leq g_N(t)$. Finalement, puisque $x > \mu$, on a $\int_x^\infty g_N(t)\,\mathrm{d}t \leq \frac{1}{2}$. On peut donc appliquer le théorème de convergence dominée avec $g = g_N$ et on a $\lim_{n\to\infty} \int_x^\infty g_n(t)\,\mathrm{d}t = \int_x^\infty g_\infty(t)\,\mathrm{d}t = 0$.

Si on choisit maintenant un $x < \mu$, on peut calquer le raisonnement précédent pour obtenir $\lim_{\sigma\to 0} F_X(x) = 0$. Cependant, en utilisant la symétrie de f_X autour de μ, on peut montrer que pour tout $x < \mu$, il existe un $y > \mu$ tel que $F_X(x) = 1 - F_X(y)$. Le raisonnement précédent établit que le terme de droite converge vers 0 lorsque $\sigma \to 0$.

Cette propriété montre que pour σ suffisamment proche de 0, la loi d'une variable aléatoire $\mathcal{N}(\mu, \sigma^2)$ est arbitrairement proche de la loi de la variable aléatoire constante $Y = \mu$. On considère alors que cette dernière est une variable aléatoire gaussienne de variance nulle ($\sigma^2 = 0$). De manière générale, tout réel c peut être vu comme une variable aléatoire gaussienne $Y \sim \mathcal{N}(c, 0)$.

La propriété 9.3 ci-dessous présente les propriétés les plus utiles de la loi normale univariée. Sa preuve repose sur le résultat intermédiaire suivant.

lem:nor_univ **Lemme 9.1** *$\int_0^{+\infty} xe^{\frac{-x^2}{2}}\,\mathrm{d}x = 1$ et donc $\int_{-\infty}^{+\infty} xe^{\frac{-x^2}{2}}\,\mathrm{d}x = 0$*

Preuve : On note que la fonction $x \mapsto xe^{\frac{-x^2}{2}}$ est la dérivée de la fonction $x \mapsto -e^{\frac{-x^2}{2}}$. Par conséquent

$$\int_0^{+\infty} xe^{\frac{-x^2}{2}}\,\mathrm{d}x = -\left[e^{\frac{-x^2}{2}}\right]_0^{+\infty} = 1 - \lim_{x\to+\infty} e^{\frac{-x^2}{2}} = 1$$

Ce qui établit la première égalité. On a également

$$\int_{-\infty}^{+\infty} xe^{\frac{-x^2}{2}}\,\mathrm{d}x = \int_{-\infty}^{0} xe^{\frac{-x^2}{2}}\,\mathrm{d}x + \int_0^{+\infty} xe^{\frac{-x^2}{2}}\,\mathrm{d}x$$

Comme la fonction $x \mapsto xe^{\frac{-x^2}{2}}$ est impaire, on a $\int_{-\infty}^0 xe^{\frac{-x^2}{2}}\,\mathrm{d}x = -\int_0^{+\infty} xe^{\frac{-x^2}{2}}\,\mathrm{d}x$, d'où la seconde égalité.

pro:nor_univ **Propriété 9.3** *Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $\mathrm{E}(X) = \mu$ et $\mathrm{V}(X) = \sigma^2$*

Preuve : On a $\mathrm{E}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} xe^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\,\mathrm{d}x$. On détermine la valeur de l'intégrale, pour laquelle on effectue le changement de variable $y = \frac{x-\mu}{\sigma}$. On a

$$\int_{-\infty}^{\infty} xe^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\,\mathrm{d}x = \int_{-\infty}^{\infty} (\sigma y + \mu)e^{-\frac{y^2}{2}}\sigma\,\mathrm{d}y = \sigma^2 \int_{-\infty}^{\infty} ye^{-\frac{y^2}{2}}\,\mathrm{d}y + \sigma\mu \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,\mathrm{d}y$$

Le premier terme de cette somme est nul d'après le lemme 9.1. En utilisant l'égalité (9.2), on déduit que le second terme est égal à $\sigma\mu\sqrt{2\pi}$. On a donc $\mathrm{E}(X) = \frac{1}{\sqrt{2\pi}\sigma}\sigma\mu\sqrt{2\pi} = \mu$.

Par ailleurs, puisqu'on vient d'établir que $\mathrm{E}(X) = \mu$, on a $\mathrm{V}(X) = \mathrm{E}\big[(X-\mu)^2\big] = \int_{-\infty}^{\infty}(x-\mu)^2 f_X(x)\,\mathrm{d}x$. En utilisant cette expression de f_X et en effectue le changement de variable $y = (x-\mu)/\sigma$, on peut écrire

$$\mathrm{V}(X) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty}\sigma^2 y^2 e^{-\frac{y^2}{2}}\sigma\,\mathrm{d}y = \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2}}\,\mathrm{d}y$$

On calcule la valeur de l'intégrale. On commence par remarquer que la fonction à intégrer est paire. Par conséquent son intégrale est égale à $2\int_0^{\infty} y^2 e^{-\frac{y^2}{2}}\,\mathrm{d}y$. On effectue ensuite une intégration par parties où l'intégrale à calculer s'écrit $\int_0^{\infty} u'(y)v(y)\,\mathrm{d}y$ avec $u'(y) = ye^{-\frac{y^2}{2}}$ et $v(y) = y$. On obtient alors $u(y) = -e^{-\frac{y^2}{2}}$ et $v'(y) = 1$. Par conséquent

$$2\int_0^{\infty} y^2 e^{-\frac{y^2}{2}}\,\mathrm{d}y = 2\left[-ye^{-\frac{y^2}{2}}\right]_0^{\infty} - 2\int_0^{\infty} -e^{-\frac{y^2}{2}}\,\mathrm{d}y$$

Le premier terme de cette différence est $-2\lim_{y\to+\infty} ye^{-\frac{y^2}{2}} = 0$. Le second est égal à $\sqrt{2\pi}$ (en appliquant l'égalité (9.2)). On en déduit donc que

$$\mathrm{V}(X) = \frac{\sigma^2}{\sqrt{2\pi}}\sqrt{2\pi} = \sigma^2$$

La remarque 9.4 et la propriété 9.3 montrent que la loi d'une variable aléatoire gaussienne est entièrement caractérisée par son espérance et sa variance. Parmi toutes les possibilités pour ces deux moments, on distingue le cas correspondant à la loi normale $\mathcal{N}(0,1)$. Cette loi est appelée loi normale centrée réduite. C'est une loi pour laquelle la l'espérance et la variance sont "normalisée" à 0 et 1, respectivement. Elle joue un rôle central dans l'étude et la manipulation des lois normales, ainsi qu'on le montrera plus bas.

pro:nor_lin_EV

Propriété 9.4

1. *Si* $X \sim \mathcal{N}(\mu,\sigma^2)$ *alors* $X+b \sim \mathcal{N}(\mu+b,\sigma^2)$ *pour tout réel* b.
2. *Si* $X \sim \mathcal{N}(0,\sigma^2)$ *alors* $aX \sim \mathcal{N}(0,a^2\sigma^2)$ *pour tout réel* a.

Preuve :

1. On a $\mathrm{P}(X+b \le x) = \mathrm{P}(X \le x-b)$ et en effectuant le changement de variable $u = t+b$ on a

$$\mathrm{P}(X \le x-b) = \int_{-\infty}^{x-b}\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\,\mathrm{d}t = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2}\left(\frac{u-(\mu+b)}{\sigma}\right)^2}\,\mathrm{d}u$$

   Donc la variable $X+b$ admet une densité correspondant à celle d'une variable $\mathcal{N}(\mu+b,\sigma^2)$, d'où le résultat.

2. Considérons d'abord le cas $a > 0$. On a $\mathrm{P}(aX \le x) = \mathrm{P}(X \le x/a)$ et en effectuant le changement de variable $u = at$ on a

$$\mathrm{P}(X \le x/a) = \int_{-\infty}^{x/a}\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}\,\mathrm{d}t = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\,a\sigma}e^{-\frac{1}{2}\left(\frac{u}{a\sigma}\right)^2}\,\mathrm{d}u$$

Donc la variable aX admet une densité correspondant à celle d'une variable $\mathcal{N}(0, a^2\sigma^2)$, d'où le résultat.

On suppose maintenant $a < 0$. On a $\mathrm{P}(aX \leq x) = \mathrm{P}(X \geq x/a)$ et en utilisant la symétrie de la densité de X autour de 0 on a

$$\mathrm{P}(X \geq x/a) = \int_{x/a}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2} \mathrm{d}t = \int_{-\infty}^{-x/a} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2} \mathrm{d}t$$

On peut à présent effectuer le changement de variable $u = -at$ et on a

$$\int_{-\infty}^{-x/a} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2} \mathrm{d}t = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\,(-a\sigma)} e^{-\frac{1}{2}\left(\frac{u}{-a\sigma}\right)^2} \mathrm{d}u$$

Donc la variable aX admet une densité correspondant à celle d'une variable $\mathcal{N}(0, a^2\sigma^2)$, d'où le résultat.

Finalement, si $a = 0$ la variable aléatoire aX est égale à 0. En utilisant la propriété 9.2 et la remarque qui suit, on a $aX \sim \mathcal{N}(0, 0)$, d'où le résultat.

En appliquant les résultats de la propriété 9.4 avec $a = \frac{1}{\sigma}$ et $b = -\mu$, on obtient un résultat important qui montre que dès qu'on a une loi normale $\mathcal{N}(\mu, \sigma^2)$ quelconque, on peut toujours se ramener à la loi normale centrée réduite $\mathcal{N}(0, 1)$.

cor:nor_lin_EV **Corollaire 9.1** *Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors la variable aléatoire $Z = \frac{X-\mu}{\sigma}$ suit une loi $\mathcal{N}(0, 1)$.*

La propriété 9.4 permet d'obtenir facilement un résultat important sur les loi normales univariées.

pro:nor_lin **Propriété 9.5** *Pour n'importe quelle paire de réels (a, b), la variable aléatoire définie par $Y = aX + b$ est gaussienne avec $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.*

Preuve : On note que $Y = a(X - \mu) + (a\mu + b)$. D'après le point 1 de la propriété 9.4, la variable aléatoire $(X-\mu)$ suit une loi $\mathcal{N}(0, \sigma^2)$. D'après le point 2 de cette même propriété, $a(X-\mu)$ suit une loi $\mathcal{N}(0, a^2\sigma^2)$. En appliquant une nouvelle fois le point 1, on déduit que la variable $a(X - \mu) + (a\mu + b)$ suit une loi $\mathcal{N}(a\mu + b, a^2\sigma^2)$.

En appliquant le résultat précédent avec $a = \sigma$ et $b = \mu$, on a la réciproque du corollaire 9.1 : à partir d'une variable aléatoire $Z \sim \mathcal{N}(0, 1)$ on peut obtenir n'importe quelle loi normale $\mathcal{N}(\mu, \sigma^2)$.

cor:nor_lin **Corollaire 9.2** *Si $Z \sim \mathcal{N}(0, 1)$, alors $X = \sigma Z + \mu$ suit une loi normale $\mathcal{N}(\mu, \sigma^2)$.*

Les corollaires 9.1 et 9.2 montrent le rôle primordial joué par la loi $\mathcal{N}(0, 1)$ dans l'étude et l'utilisation des lois normales. Ce résultat a un équivalent dans le contexte des lois normales multivariées.

sec:lnor_mult ### 9.1.2 Lois normales multivariées

La notion de variable gaussienne s'étend à des n-uplets de variables aléatoires.

def:normul **Définition 9.1** *Soient $X_1, \ldots, X_n$, n variables aléatoires. Le n-uplet $X = (X_1, \ldots, X_n)$ est un n-uplet gaussien, si pour tout n-uplets $(a_1, \ldots, a_n)$ de réels, la combinaison linéaire $a_1X_1 + \cdots + a_nX_n$ est une variable aléatoire gaussienne.*

Dans le cas où on manipule des n-uplets de variables aléatoires, il est commode de voir ce n-uplet comme un vecteur aléatoire de $\mathbb{R}^n$, *i.e.*, un vecteur dont les coordonnées sont des variables aléatoires. Dans ce cas, on dit que X est un vecteur gaussien de taille n et on écrit ses coordonnées en colonne :

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \qquad \text{ou } X = (X_1, \dots, X_n)^\top$$

De la définition 9.1, on tire immédiatement le résultat suivant.

pro:normul **Propriété 9.6** *Si X est un vecteur gaussien, alors tout entier $m \in \{1, \dots, n\}$ et pour tout choix d'indices $i_1, \dots, i_m$ dans $\{1, \dots, n\}$, le sous-vecteur $(X_{i_1}, \dots, X_{i_m})^\top$ de X est gaussien. En particulier, chacune des variables aléatoires qui composent X est une variable aléatoire gaussienne.*

Preuve : Toute combinaison linéaire de $X_{i_1}, \dots, X_{i_m}$ est une combinaison linéaire de $X_1, \dots, X_n$. Le cas particulier des variables qui composent X s'obtient en choisissant $m = 1$.

La réciproque de ce résultat n'est pas vraie. En général, un vecteur formé à partir de variables aléatoires gaussiennes n'est pas nécessairement gaussien, ainsi que le montre l'exemple suivant.

Soit Z une variable aléatoire $\mathcal{N}(0,1)$ et X une variable aléatoire indépendante de Z telle que $\mathrm{P}(X = -1) = \mathrm{P}(X = 1) = \frac{1}{2}$. Considérons la variable aléatoire $Y = XZ$. On a

$$\begin{aligned} \mathrm{P}(Y \le x) &= \mathrm{P}(XZ \le x) = \mathrm{P}(XZ \le x, X = 1) + \mathrm{P}(XZ \le x, X = -1) \\ &= \mathrm{P}(Z \le x, X = 1) + \mathrm{P}(Z \ge -x, X = -1) \\ &= \mathrm{P}(Z \le x)\mathrm{P}(X = 1) + \mathrm{P}(Z \ge -x)\mathrm{P}(X = -1) = \frac{1}{2}[\mathrm{P}(Z \le x) + \mathrm{P}(Z \ge -x)] \\ &= \mathrm{P}(Z \le x) \end{aligned}$$

où la dernière égalité résulte de la symétrie de la loi $\mathcal{N}(0,1)$ autour de 0. On en déduit que Y a la même loi que Z, *i.e.*, $Y \sim \mathcal{N}(0,1)$. Or le couple (Y, Z) ne peut être gaussien. En effet, s'il l'était, la variable aléatoire $Y + Z$ devrait être gaussienne. Or

$$\mathrm{P}(Y + Z = 0) = \mathrm{P}((X + 1)Z = 0) = \mathrm{P}(X = -1) = \frac{1}{2}$$

ce qui serait impossible si $Y + Z$ était gaussienne.

em:loinormul2moments **Remarque 9.2** D'après la définition ci-dessus, la loi d'un vecteur gaussien est caractérisée par les lois de toutes ses combinaisons linéaires (il faut vérifier que ces lois sont des lois gaussiennes). Comme on l'a remarqué dans la section 9.1.1, une loi gaussienne est caractérisé par son espérance et sa variance (ou son écart-type). Par conséquent, la loi d'une combinaison linéaire $a_1X_1 + \cdots + a_nX_n$ est caractérisée par

$$\begin{aligned} \mathrm{E}(a_1X_1 + \cdots + a_nX_n) &= a_1\mu_1 + \cdots + a_n\mu_n \\ \text{et } \mathrm{V}(a_1X_1 + \cdots + a_nX_n) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \omega_{ij} \end{aligned}$$

où $\mu_i = \mathrm{E}(X_i)$ et $\omega_{ij} = \mathrm{cov}(X_i, X_j)$, $i, j = 1, \ldots, n$. Donc pour un choix donné de $a_1, \ldots, a_n$, la loi de $a_1X_1 + \cdots + a_nX_n$ est entièrement caractérisée par les nombres μ_i et ω_{ij}, $i, j = 1, \ldots, n$. Comme c'est le cas pour toute combinaison linéaire, on voit que la loi du vecteur X doit être entièrement caractérisée par ces mêmes nombres. Si on définit

page:vcov

- le vecteur $\mathrm{E}(X)$ dont la i^{e} coordonnée est $\mathrm{E}(X_i)$
- la matrice $\mathrm{V}(X)$ dont le $(i,j)^{\mathrm{e}}$ élément est $\mathrm{cov}(X_i, X_j)$

alors la loi d'un vecteur gaussien X est entièrement caractérisée par le vecteur $\mathrm{E}(X)$ et la matrice $\mathrm{V}(X)$. Par conséquent, on note $X \sim \mathcal{N}_n(\mathrm{E}(X), \mathrm{V}(X))$. Si on a $\mathrm{E}(X) = \mu$ et $\mathrm{V}(X) = V$, on note $X \sim \mathcal{N}_n(\mu, V)$.

La matrice $\mathrm{V}(X)$ est appelé matrice des variances-covariances de X. Ses éléments diagonaux sont les variances des variables $X_1, \ldots, X_n$ et ses éléments hors diagonale sont les covariances entre ces variables. Cette matrice possède les propriétés suivantes. R

pro:mat_v_cov

Propriété 9.7 *Soit X un vecteur aléatoire (pas forcément gaussien). On a*

1. $\mathrm{V}(X) = \mathrm{E}\big[(X - \mathrm{E}(X))(X - \mathrm{E}(X))^\top\big] = \mathrm{E}(XX^\top) - \mathrm{E}(X)[\mathrm{E}(X)^\top]$ *où pour toute matrice de variables aléatoires $A = (A_{ij})_{i=1,\ldots,p, j=1,\ldots,q}$, $\mathrm{E}(A)$ est la matrice des espérances dont le $(i,j)^e$ élément est $\mathrm{E}(A_{ij})$.*
2. *Si B est une matrice de nombres réels de taille $p \times n$, alors $\mathrm{V}(BX) = B\mathrm{V}(X)B^\top$.*
3. *$\mathrm{V}(X)$ est symétrique semi-définie positive. Elle est définie positive si et seulement si quels que soient les réels $a_0, \ldots a_n$, l'égalité $\mathrm{P}(a_0 + a_1X_1 + \cdots + a_nX_n = 0) = 1$ implique $a_0 = a_1 = \cdots = a_n = 0$.*

Preuve :

1. D'après la définition de $\mathrm{E}(X)$, le i^{e} élément de $X - \mathrm{E}(X)$ est $X_i - \mathrm{E}(X_i)$. Donc le $(i,j)^{\mathrm{e}}$ élément de $(X - \mathrm{E}(X))(X - \mathrm{E}(X))^\top$ est $Y_{ij} = (X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))$. Le $(i,j)^{\mathrm{e}}$ élément de $\mathrm{E}\big[(X - \mathrm{E}(X))(X - \mathrm{E}(X))^\top\big]$ est $\mathrm{E}(Y_{ij}) = \mathrm{cov}(X_i, X_j)$, ce qui montre la première égalité. Pour montrer la seconde, on se sert de l'expression $\mathrm{cov}(X_i, X_j) = \mathrm{E}(X_iX_j) - \mathrm{E}(X_i)\mathrm{E}(X_j)$ et en procédant comme avant, on montre que c'est le $(i,j)^{\mathrm{e}}$ élément de $\mathrm{E}(XX^\top) - \mathrm{E}(X)[\mathrm{E}(X)^\top]$.
2. Ce résultat s'obtient en utilisant l'expression de la variance donnée dans le point précédent de cette propriété.
3. La symétrie de $\mathrm{V}(X)$ résulte de celle de la covariance : $\mathrm{cov}(X_i, X_j) = \mathrm{cov}(X_j, X_i)$ $\forall i, j$.

   Par ailleurs, pour tout vecteur λ de $\mathbb{R}^n$, le point précédent permet d'écrire que $\lambda^\top V(X)\lambda = \mathrm{V}(\lambda^\top X)$. Or $\lambda^\top X$ est la variable aléatoire $\lambda_1X_1 + \cdots + \lambda_nX_n$ et la variance d'une variable aléatoire n'est jamais négative. Par conséquent $\lambda^\top V(X)\lambda \geq 0$, $\forall \lambda \in \mathbb{R}^n$.

   Finalement, supposons que l'implication $\mathrm{P}(a_0 + a_1X_1 + \cdots + a_nX_n = 0) = 1 \implies a_0 = a_1 = \cdots = a_n = 0$ soit vraie. Soit $\lambda \in \mathbb{R}^n$ avec $\lambda \neq 0_n$. On a $\lambda^\top \mathrm{V}(X)\lambda = \mathrm{V}(\lambda^\top X) \geq 0$. Supposons que $\mathrm{V}(\lambda^\top X) = 0$. Dans ce cas, il existe un réel c tel que $\mathrm{P}(\lambda^\top X = c) = 1$. D'après l'implication supposée vraie, on devrait avoir $c = \lambda_1 = \cdots = \lambda_n = 0$ ce qui contredit l'hypothèse posée $\lambda \neq 0_n$. Par conséquent, $\mathrm{V}(\lambda^\top X) > 0$ $\forall \lambda \in \mathbb{R}^n$, $\lambda \neq 0_n$.

Supposons maintenant que $\mathrm{V}(X)$ soit définie positive et qu'on puisse trouver des réels $a_0, \ldots, a_n$ non tous nuls, tels que $\mathrm{P}(a_0 + a_1X_1 + \cdots + a_nX_n = 0) = 1$. En posant $a = (a_1, \ldots, a_n)^\top$, on a alors $a^\top \mathrm{V}(X)a = \mathrm{V}(a^\top X) = 0$ avec $a \neq 0_n$, ce qui contredit l'hypothèse de départ que $\mathrm{V}(X)$ est définie positive.

En utilisant la remarque de la fin de la section précédente, on voit qu'un vecteur donné de $\mathbb{R}^n$ est un vecteur gaussien. Plus précisément, pour tout vecteur $c \in \mathbb{R}^n$, on peut définir le vecteur aléatoire A tel que $\mathrm{P}(A = c) = 1$. Pour tout vecteur $a \in \mathbb{R}^n$ on a $\mathrm{P}(a^\top A = a^\top c) = 1$, on peut considérer que la "variable aléatoire" $a^\top A$ est gaussienne avec $a^\top A \sim \mathcal{N}(a^\top c, 0)$ (voir la remarque qui suit la propriété 9.3). Cela montre que toute combinaison linéaire de A est une variable aléatoire gaussienne et donc que A est un vecteur gaussien.

En utilisant la définition 9.1 et le point 2 de la propriété 9.3, on obtient facilement le résultat suivant.

pro:norlin **Propriété 9.8** *Si X est gaussien de taille n, alors pour tout vecteur $a \in \mathbb{R}^m$ et toute matrice A de taille $m \times n$, le vecteur $Y = a + AX$ est un vecteur gaussien de taille m.*

Propriété 9.9 *Soit $\mu \in \mathbb{R}^n$ et V une matrice symétrique semi-définie positive, de taille $n \times n$. On définit le vecteur aléatoire $X = \mu + AZ$ où $Z \sim \mathcal{N}_n(0_n, I_n)$ et A est une matrice $n \times n$ telle que $AA^\top = V$. On a $X \sim \mathcal{N}_n(\mu, V)$.*

Preuve : L'existence de la matrice A est garantie par le fait que V est symétrique semi-definie positive. Comme Z est gaussien, la propriété 9.8 implique que $X = \mu + AZ$ est un vecteur gaussien. On calcule alors $\mathrm{E}(X) = \mu + A\mathrm{E}(Z) = \mu$ et $\mathrm{V}(X) = \mathrm{V}(AZ) = A\mathrm{V}(Z)A^\top = AA^\top = V$.

pro:nordens **Propriété 9.10** *Si X est un n-uplet gaussien $\mathcal{N}_n(\mu, V)$ et V est définie positive, alors*

$$\mathrm{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f_X(t_1, t_2 \ldots, t_n)\, \mathrm{d}t_1, \mathrm{d}t_2 \ldots \mathrm{d}t_n$$

où la fonction f_X est définie par

$$\begin{aligned} f_X(t_1, \ldots, t_n) &= \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left(-\frac{1}{2}(t-\mu)V^{-1}(t-\mu)\right) \\ &= \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}(t_i - \mu_i)(t_j - \mu_j)\right) \end{aligned}$$

et $t = (t_1, \ldots, t_n)^\top$ et v_{ij} est le $(i,j)^e$ élément de V^{-1}.

Dans la propriété précédente, la fonction f_X est la fonction de densité de X. La condition V définie positive garantit que $|V| \neq 0$ et donc que V^{-1} existe.

L'expression de la densité d'un vecteur aléatoire X gaussien formalise la remarque faite précédemment que la loi de ce vecteur est entièrement caractérisée par $\mathrm{E}(X)$ et $\mathrm{V}(X)$.

Une propriété fondamentale de la loi normale multivariée établit que pour cette loi l'indépendance et la non-corrélation sont équivalentes. On commence par rappeler la notion d'indépendance pour des vecteurs aléatoires.

def:varindep **Définition 9.2** *Soit $(X_1, \ldots, X_n)$ un n-uplet de variables aléatoires. On dit que $X_1, \ldots, X_n$ sont indépendantes (ou indépendantes dans leur ensemble) si pour tout entier $m \in \{1, \ldots, n\}$ et tout m-uplet $(i_1, \ldots, i_m)$ d'éléments distincts de $\{1, \ldots, n\}$ on a*

$$\mathrm{P}(X_{i_1} \leq x_1, \ldots, X_{i_m} \leq x_m) = \mathrm{P}(X_{i_1} \leq x_1) \times \cdots \times \mathrm{P}(X_{i_m} \leq x_m) \tag{9.3}$$ eq:indep

pour tout $(x_1, \ldots, x_m) \in \mathbb{R}^m$.

Autrement dit, $X_1, \ldots, X_n$ sont indépendantes si la loi jointe de tout m-uplet de variables distinctes prises parmi $X_1, \ldots, X_n$ est égale au produit des lois marginales des variables de ce m-uplet.

Un conséquence immédiate de cette définition est que les variables de tout m-uplet formé à partir de n variables indépendantes (distinctes) sont aussi indépendantes.

Il ne faut pas confondre l'indépendance des variables $X_1, \ldots, X_n$ avec leur indépendance deux à deux, qui établit que pour n'importe quelle paire de variables distinctes prises parmi $X_1, \ldots, X_n$, on a $\mathrm{P}(X_i \leq a, X_j \leq b) = \mathrm{P}(X_i \leq a)\mathrm{P}(X_j \leq b)$, $\forall (a, b) \in \mathbb{R}^2$. Il ne faut pas non plus assimiler l'indépendance de $X_1, \ldots, X_n$ à la condition $\mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^n \mathrm{P}(X_i \leq x_i)$, $\forall (x_1, \ldots, x_n) \in \mathbb{R}^n$.

On présente un résultat intermédiaire sur les variables indépendantes qui permet de simplifier la preuve de la propriété 9.12.

lem:norindep **Lemme 9.2** *Soit $X = (X_1, \ldots, X_n)$ un n-uplet de variables aléatoires (pas forcément gaussien).*

1. *Si pour tout $i = 1, \ldots, n$, il existe un réel c_i tel que $\mathrm{P}(X_i = c_i) = 1$ (*i.e.*, la variable X_i est constante), alors $X_1, \ldots, X_n$ sont indépendantes.*
2. *Si pour un $r \in \{1, \ldots, n\}$, il existe des réels $c_{r+1}, \ldots, c_n$ pour lesquels $\mathrm{P}(X_{r+1} = c_{r+1}) = \cdots = \mathrm{P}(X_n = c_n) = 1$ (autrement dit, les $n - r$ dernières variables de X sont constantes), alors $X_1, \ldots, X_n$ sont indépendantes si et seulement si $X_1, \ldots, X_r$ le sont.*

Preuve :

1. Pour tout $x \in \mathbb{R}$ et tout indice $i = 1, \ldots, n$, on a $\mathrm{P}(X_i \leq x) = 0$ ou 1, selon que $x < c_i$ ou $x \geq c_i$. Pour tout m, tout choix de $(x_1, \ldots, x_m) \in \mathbb{R}^m$ et tout choix d'indices $i_1, \ldots, i_m$ on a :

$$\mathrm{P}(X_{i_1} \leq x_1, \ldots, X_{i_m} \leq x_m) = \begin{cases} 0 = \displaystyle\prod_{k=1}^m \mathrm{P}(X_{i_k} \leq x_k) \text{ si } \exists k \text{ t.q. } x_k < c_k \\ \\ 1 = \displaystyle\prod_{k=1}^m \mathrm{P}(X_{i_k} \leq x_k) \text{ sinon} \end{cases}$$

   où ces égalités résultent de $\mathrm{P}(A \cap B) = 0$ si $\mathrm{P}(A) = 0$, et $\mathrm{P}(A \cap B) = \mathrm{P}(B)$ si $\mathrm{P}(A) = 1$.

2. Supposons que $X_1, \ldots, X_n$ soient indépendantes. Alors d'après la remarque qui suit la définition 9.2, les variables $X_1, \ldots, X_r$ le sont aussi. Réciproquement, supposons que $X_1, \ldots, X_r$ soient indépendantes. La condition (9.3) est satisfaite à chaque fois que tous les indices $i_1, \ldots, i_m$ sont pris dans $\{1, \ldots, r\}$. De plus, cette condition est également satisfaite si tous ces indices sont choisis dans $\{r + 1, \ldots, n\}$, d'après le

point précédent de ce lemme. Pour que $X_1, \dots, X_n$ soient indépendantes, il reste par conséquent à vérifier que la condition (9.3) est vraie lorsque parmi $i_1, \dots, i_m$ il y a des indices à la fois dans $\{1, \dots, r\}$ et dans $\{r+1, \dots, n\}$. Soient donc m un entier dans $\{1, \dots, n\}$, un m-uplet de réels $(x_1, \dots, x_m)$ et des indices $i_1, \dots, i_m$, choisis dans $\{1, \dots, n\}$ de sorte que l d'entre eux (les l premiers par exemple) soient dans $\{1, \dots, r\}$ et $m-l$ soient dans $\{r+1, \dots, n\}$. Afin d'alléger la notation, pour tout $k = 1, \dots, m$, on note A_k l'évènement $(X_{i_k} \leq x_k)$. Si pour un indice $k \in \{l+1, \dots, m\}$ on a $\mathrm{P}(A_k) = 0$, alors $\mathrm{P}(A_1 \cap \dots \cap A_m) = 0 = \prod_{k=1}^m \mathrm{P}(A_k)$. Dans le cas contraire où $\mathrm{P}(A_{l+1}) = \dots = \mathrm{P}(A_m) = 1$, on a :

$$\mathrm{P}(A_1 \cap \dots \cap A_m) = \mathrm{P}(A_1 \cap \dots \cap A_l) = \prod_{j=1}^{l} \mathrm{P}(A_j) \prod_{k=l+1}^{m} \mathrm{P}(A_k) = \prod_{i=1}^{m} \mathrm{P}(A_i)$$

où la première égalité vient de $\mathrm{P}(A_k) = 1$, $k = l+1, \dots, m$, et la deuxième du fait que par hypothèse et choix des indices, $X_{i_1}, \dots, X_{i_l}$ sont indépendantes. Dans les deux cas, la conditions (9.3) est bien vérifiée et $X_1, \dots, X_n$ sont indépendantes.

pro:norindepend **Propriété 9.11** *Soit $X \sim \mathcal{N}_n(\mu, V)$. Les coordonnées $X_1, \dots, X_n$ de X sont des variables aléatoires indépendantes si et seulement si la matrice des variances-covariances V est diagonale.*

Preuve : Si $X_1, \dots, X_n$ sont indépendantes, alors elles sont non-corrélées et les éléments hors diagonaux de V sont nuls : V est donc diagonale.
Supposons à présent que V soit diagonale et notons $\sigma_1^2, \dots, \sigma_n^2$ ses éléments diagonaux. Afin de couvrir le cas où V n'est pas définie positive, on permet que certains éléments de sa diagonale puissent être nuls. Quitte à renuméroter les coordonnées de X (ce qui ne change rien au fait que sa matrice des variances-covariances soit diagonale), on suppose que pour un certain $r \in \{0, 1, \dots, n\}$, on a $\sigma_k^2 > 0$ pour $k = 1, \dots, r$ et $\sigma_k^2 = 0$ pour $k = r+1, \dots, n$. (Si $r = n$, aucun élément diagonal de V n'est nul.) Dans ce cas, puisque $V(X_i) = \sigma_i^2$, les variables aléatoires $X_{r+1}, \dots, X_n$ sont constantes. Si $r = 0$, alors toutes les variables du n-uplet X sont constantes et on peut appliquer le point 1 du lemme 9.2 pour conclure que $X_1, \dots, X_n$ sont indépendantes. Si $r > 0$ alors d'après le point 2 de ce même lemme, $X_1, \dots, X_n$ sont indépendantes si $X_1, \dots, X_r$ le sont. On va donc montrer que c'est la cas. Soit m un entier dans $\{1, \dots, r\}$ et $i_1, \dots, i_m$ des éléments de $\{1, \dots, r\}$. D'après la propriété 9.6, $Y = (X_{i_1}, \dots, X_{i_m})^\top$ est un vecteur gaussien et chacune des coordonnées de Y est une variable aléatoire gaussienne. On a $\mathrm{E}(Y) = (\mu_{i_1}, \dots, \mu_{i_m})^\top$. De plus, $\mathrm{V}(Y)$ est une matrice diagonale dont les éléments diagonaux sont $\sigma_{i_1}^2, \dots, \sigma_{i_m}^2$. Puisque $i_k \in \{1, \dots, r\}$, on a forcément $\sigma_{i_k}^2 > 0$, $k = 1, \dots, m$. Donc la matrice $\mathrm{V}(Y)$ est définie positive. Son déterminant est égal au produit de ses éléments diagonaux : $|\mathrm{V}(Y)| = \prod_{k=1}^m \sigma_{i_k}^2$. De plus, $\mathrm{V}(Y)^{-1}$ est la matrice diagonale dont les éléments diagonaux sont l'inverse des éléments diagonaux de $\mathrm{V}(Y)$. Pour tout $a \in \mathbb{R}^m$ on a $a^\top \mathrm{V}(Y)^{-1} a = \sum_{k=1}^m (a_k/\sigma_{i_k})^2$. En utilisant la propriété 9.10, on peut écrire

$$\mathrm{P}(X_{i_1} \leq x_1, \dots, X_{i_m} \leq x_m) = \frac{1}{(2\pi)^{\frac{m}{2}} |\mathrm{V}(Y)|^{\frac{1}{2}}} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_m} e^{-\frac{1}{2} \sum_{k=1}^m \left(\frac{t_i - \mu_{i_k}}{\sigma_{i_k}} \right)^2} \, \mathrm{d}t_1 \dots \mathrm{d}t_m$$

On a

$$(2\pi)^{\frac{m}{2}}|\mathrm{V}(Y)|^{\frac{1}{2}} = \left[(2\pi)^m \prod_{k=1}^{m} \sigma_{i_k}^2\right]^{\frac{1}{2}} = \left[\prod_{k=1}^{m} 2\pi\sigma_{i_k}^2\right]^{\frac{1}{2}} = \prod_{k=1}^{m} \sqrt{2\pi\sigma_{i_k}^2}$$

et

$$e^{-\frac{1}{2}\sum_{k=1}^{m}\left(\frac{t_i-\mu_{i_k}}{\sigma_{i_k}}\right)^2} = \prod_{k=1}^{m} e^{-\frac{1}{2}\left(\frac{t_i-\mu_{i_k}}{\sigma_{i_k}}\right)^2}$$

Donc on peut écrire

$$\begin{aligned}
\mathrm{P}(X_{i_1} \leq x_1, \ldots, X_{i_m} \leq x_m) &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_m} \frac{1}{\prod_{k=1}^{m}\sqrt{2\pi\sigma_{i_k}^2}} \prod_{k=1}^{m} e^{-\frac{1}{2}\left(\frac{t_i-\mu_{i_k}}{\sigma_{i_k}}\right)^2} \,\mathrm{d}t_1 \ldots \mathrm{d}t_m \\
&= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_m} \prod_{k=1}^{m} \frac{1}{\sqrt{2\pi\sigma_{i_k}^2}} e^{-\frac{1}{2}\left(\frac{t_i-\mu_{i_k}}{\sigma_{i_k}}\right)^2} \,\mathrm{d}t_1 \ldots \mathrm{d}t_m \\
&= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_m} f_{X_{i_1}}(t_1) \times \cdots \times f_{X_{i_m}}(x_m) \,\mathrm{d}t_1 \ldots \mathrm{d}t_m \\
&= \int_{-\infty}^{x_1} f_{X_{i_1}}(t_1) \,\mathrm{d}t_1 \times \cdots \times \int_{-\infty}^{x_m} f_{X_{i_m}}(x_m) \,\mathrm{d}t_m
\end{aligned}$$

où $f_{X_{i_k}}(t_i) = \frac{1}{\sqrt{2\pi\sigma_{i_k}^2}} e^{-\frac{1}{2}\left(\frac{t_i-\mu_{i_k}}{\sigma_{i_k}}\right)^2}$ est la densité de la variable gaussienne X_{i_k}, $k = 1, \ldots, m$ (voir la définition 9.1). On en déduit que

$$\mathrm{P}(X_{i_1} \leq x_1, \ldots, X_{i_m} \leq x_m) = \mathrm{P}(X_{i_1} \leq x_1) \times \cdots \times \mathrm{P}(X_{i_m} \leq x_m)$$

Ceci étant vrai pour tout m, tout $x_1, \ldots, x_m$ et tout $i_1, \ldots, i_m$, on en conclut que $X_1, \ldots, X_r$ sont indépendantes.

Bien qu'il soit en général faux de dire qu'un vecteur formé de variables gaussiennes est gaussien, il existe un cas particulier et important dans lequel ceci est vrai.

pro:norindep **Propriété 9.12** *Si $Z_1, \ldots, Z_n$ sont n variables aléatoires gaussiennes indépendantes, alors le vecteur $Z = (Z_1, \ldots, Z_n)$ est gaussien.*

Un cas particulier de la propriété précédente s'obtient lorsque les variables indépendantes $Z_1, \ldots, Z_n$ sont toutes de loi $\mathcal{N}(0,1)$. Dans ce cas, $\mathrm{E}(Z) = 0_n$ et $\mathrm{V}(Z) = I_n$, où I_n est la matrice identité $n \times n$, et on a $Z \sim \mathcal{N}_n(0_n, I_n)$.

Les propriétés 9.8 et 9.12 permettent de montrer que pour tout vecteur $\mu \in \mathbb{R}^n$ et toute matrice symétrique définie positive V de taille $n \times n$, on peut construire un vecteur aléatoire gaussien X tel que $X \sim \mathcal{N}_n(\mu, V)$ à partir de variables aléatoires $Z_1, \ldots, Z_n$ indépendantes, de loi $\mathcal{N}(0,1)$.

On termine cette section en présentant un résultat semblable à la propriété 9.11 valable pour des sous-vecteurs d'un vecteur aléatoire gaussien.

def:vec_indep **Définition 9.3** *Soient $X = (X_1, \ldots, X_p)$ et $Y = (Y_1, \ldots, Y_q)$ deux vecteurs aléatoires. On dit que X et Y sont indépendants si la loi de (X, Y) est égale au produit de la loi de X par celle de Y :*

$$\mathrm{P}(X_1 \leq a_1, \ldots, X_p \leq a_p, Y_1 \leq b_1, \ldots, Y_q \leq b_q) = \mathrm{P}(X_1 \leq a_1, \ldots, X_p \leq a_p) \times \mathrm{P}(Y_1 \leq b_1, \ldots, Y_q \leq b_q)$$

pour n'importe quels réels $a_1, \ldots, a_p, b_1, \ldots, b_q$.

pro:vec_nor_indep **Propriété 9.13** *Soient X et Y deux sous-vecteurs d'un vecteur gaussien. On forme le vecteur gaussien $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$. X et Y sont indépendants si et seulement si la matrice des variances-covariances de Z est bloc-diagonale, de la forme*

$$\mathrm{V}(Z) = \left(\begin{array}{c:c} \mathrm{V}(X) & 0 \\ \hdashline 0 & \mathrm{V}(Y) \end{array}\right)$$

Preuve : Si X et Y sont indépendants, alors toutes les covariances entre une coordonnée de X et une coordonnée de Y sont nulles. Par définition de la matrice $\mathrm{V}(Z)$ (voir page 198 avant la propriété 9.7), ce sont précisément ces covariances qui constituent les blocs extra-diagonaux de $\mathrm{V}(Z)$. Cette matrice a donc dans ce cas la forme donnée dans l'énoncé.
Supposons maintenant que cette matrice soit bloc-diagonale. Sa matrice inverse sera également bloc-diagonale et en utilisant un procédé semblable à celui utilisé dans la preuve de la propriété 9.11, on peut séparer de manière multiplicative la densité de Z en une partie qui ne dépend que des coordonnées de X d'une part, et une partie qui ne dépend que des coordonnées de Y, d'autre part. On arrive ainsi à écrire une égalité semblable à celle de la définition 9.3.

rem:vec_nor_indep **Remarque 9.3** Dans le résultat précédent, on constate qu'il y a indépendance entre deux sous-vecteurs d'un vecteur gaussien dès que toutes les covariances qu'il est possible de former à partir des couples de leurs coordonnées sont nulles. Ceci peut s'exprimer à partir de la matrice de covariance entre le vecteur X et le vecteur Y, notée $\mathrm{cov}(X, Y)$ et dont l'élément constitutif est $\mathrm{cov}(X_i, Y_j)$, où X_i et Y_j sont respectivement la i^{e} coordonnée de X et la j^{e} coordonnée de Y. Cette matrice est donc de dimensions $p \times q$ où p est la taille du vecteur X et q celle de Y. En utilisant la même méthode que dans la preuve de la propriété 9.7, on montre que $\mathrm{cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))^\top\right] = \mathrm{E}(XY^\top) - \mathrm{E}(X)\mathrm{E}(Y)^\top$. Il est alors facile de montrer que $\mathrm{cov}(Y, X) = \mathrm{cov}(X, Y)^\top$.
La condition nécessaire et suffisante pour que les deux sous-vecteurs X et Y de la proposition précédente soient indépendants est que la matrice $\mathrm{cov}(X, Y)$ soit nulle.

Notons que la matrice $\mathrm{cov}(X, Y)$ possède les propriétés suivantes, qui s'obtiennent directement en utilisant sa définition, ainsi que les propriétés des covariances entre variables aléatoires.

pro:cov_vec **Propriété 9.14** *Soient X et Y deux vecteurs aléatoires de $\mathbb{R}^n$ et $\mathbb{R}^m$ respectivement. La matrice $\mathrm{cov}(X, Y)$ définie dans la remarque ci-dessus satisfait les propriétés suivantes :*

1. $\mathrm{cov}(Y, X) = \mathrm{cov}(X, Y)^\top$

2. $\mathrm{cov}(X, a + AY) = \mathrm{cov}(X, Y)A^\top$ *et donc (en utilisant le premier point)* $\mathrm{cov}(b + BX, Y) = B\mathrm{cov}(X, Y)$
3. $\mathrm{cov}(X, X) = \mathrm{V}(X)$

où a et b sont des vecteurs (non aléatoires) de $\mathbb{R}^{q_a}$ et $\mathbb{R}^{q_b}$ respectivement, et A et B sont des matrices réelles (non aléatoires) de dimensions $q_a \times n$ et $q_b \times n$ respectivement.

Les deux propriétés précédentes et la définition de la matrice $\mathrm{cov}(X, Y)$ permet d'obtenir le résultat suivant.

ro:indep_formlin_nor **Propriété 9.15** *Soit $Z \sim \mathcal{N}_n(\mu, V)$ où V peut s'écrire $V = \sigma^2 I_n$ pour un certain réel σ. Soient A et B deux matrices réelles de dimensions respectives $m_A \times n$ et $m_B \times n$. Si $AB^\top = 0$, alors les vecteurs aléatoires AZ et BZ sont indépendants.*

Preuve : Définissons le vecteur aléatoire Y par

$$Y = \left(\begin{array}{c} AZ \\ \hdashline BZ \end{array}\right) = CZ$$

où $C = \left(\begin{array}{c} A \\ \hdashline B \end{array}\right)$. Il est facile de vérifier que toute combinaison linéaire des coordonnées de Y s'écrit comme une combinaison linéaire des coordonnées de Z. Par conséquent, Y est un vecteur gaussien et étant donnée sa forme, on a

$$\mathrm{V}(Y) = \left(\begin{array}{c:c} \mathrm{V}(AZ) & \mathrm{cov}(AZ, BZ) \\ \hdashline \mathrm{cov}(BZ, AZ) & \mathrm{V}(BZ) \end{array}\right)$$

En utilisant la propriété 9.13, AZ et BZ seront indépendants si et seulement si $\mathrm{cov}(AZ, BZ) = 0$. D'après la propriété 9.14, on a $\mathrm{cov}(AZ, BZ) = A\mathrm{cov}(Z, Z)B^\top = A\mathrm{V}(Z)B^\top$. Comme $\mathrm{V}(Z) = \sigma^2 I_n$, on a $\mathrm{cov}(AZ, BZ) = \sigma^2 AB^\top = 0$.

Remarque 9.4 Notons que sous les conditions de la propriété 9.15, toute fonction de AZ est indépendante de toute fonction de BZ. Un cas important dans lequel ce résultat est utilisé est celui où les matrices A et B telles que $AB = 0$ sont symétriques et idempotentes. Dans ce cas, si $Z \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$ alors les formes quadratiques $Z^\top AZ$ et $Z^\top BZ$ sont des variables aléatoires indépendantes l'une de l'autre. De même la variable aléatoire $Z^\top AZ$ est indépendante du vecteur BZ. Le point 3 de la propriété 6.1 est basé sur cette remarque.

9.1.3 Lois dérivées de la loi normale

Dans de nombreuses applications, on est amené à utiliser des variables aléatoires construites comme des fonctions de plusieurs variables aléatoires gaussiennes.

9.1.3.1 La loi du χ^2

sec:chi2

def:chi2 **Définition 9.4** *La loi du χ^2 à ν degrés de liberté est la loi de la variable aléatoire $C = \sum_{i=1}^{\nu} Z_i^2$, où $Z_1, \ldots, Z_\nu$ sont des variables aléatoires $\mathcal{N}(0, 1)$ indépendantes. On note $C \sim \chi^2(\nu)$*

Notons que si on considère les variables $Z_1, \ldots, Z_\nu$ comme les coordonnées du vecteur aléatoire $Z = (Z_1, \ldots, Z_\nu)^\top \in \mathbb{R}^\nu$, alors la variable C de la définition est le carré de norme de Z. Donc la définition dit que si $Z \sim \mathcal{N}_\nu(0_\nu, I_\nu)$ alors $\|Z\|^2 \sim \chi^2(\nu)$.

Propriété 9.16

1. *Soit* $C \sim \chi^2(\nu)$. *On a*
   (a) $\mathrm{P}(C \leq x) = 0$, $\forall x \leq 0$ *et pour* $x > 0$,
   $$\mathrm{P}(C \leq x) = \int_0^x \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} t^{\frac{\nu}{2}-1} e^{-\frac{t}{2}} \,\mathrm{d}t$$
   i.e. C *admet sur* $\mathbb{R}_+$ *une densité* $f_C(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}$ *où* Γ *est la fonction gamma définie sur* $\mathbb{R}_+$ *par* $x \mapsto \Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \,\mathrm{d}t$.
   (b) $\mathrm{E}(C) = \nu$ *et* $\mathrm{V}(C) = 2\nu$
2. *Soient* $C_1 \sim \chi^2(\nu_1)$ *et* $C_2 \sim \chi^2(\nu_2)$ *indépendantes. Alors* $C_1 + C_2 \sim \chi^2(\nu_1 + \nu_2)$.

Détermination des valeurs de la distribution du χ^2 au moyen d'un tableur et du logiciel R :

	Probabilité $\mathrm{P}(C_\nu \leq x)$	Quantile q_p : $\mathrm{P}(C_\nu \leq q_p) = p$
Tableur	`1 - LOI.KHIDEUX(`x`;`ν`)`	`LOI.KHIDEUX.INVERSE(`p`;`ν`)`
R	`pchisq(`x`,`ν`)`	`qchisq(`p`,`ν`)`

Soient $X_1, \ldots, X_\nu$ des variables aléatoires indépendantes telles que $X_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$, avec $\sigma_k > 0$ pour $k = 1, \ldots, \nu$. On a $Z_k = \frac{X_k - \mu_k}{\sigma_k} \sim \mathcal{N}(0, 1)$ d'après le corollaire 9.1. De plus, $Z_1, \ldots Z_\nu$ sont indépendantes et d'après la propriété 9.12, le vecteur aléatoire $Z = (Z_1, \ldots Z_\nu)^\top$ est gaussien $\mathcal{N}_\nu(0_\nu, I_\nu)$ et donc $\|Z\|^2 \sim \chi^2(\nu)$. On note X le vecteur aléatoire de $\mathbb{R}^p$ de coordonnées $X_1, \ldots, X_\nu$, μ le vecteur de coordonnées $\mu_1, \ldots, \mu_\nu$ et V la matrice des variances-covariances de X, autrement dit

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_\nu \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_\nu \end{pmatrix} \qquad V = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_\nu^2 \end{pmatrix}$$

Il est facile de vérifier que $(X - \mu)^\top V^{-1}(X - \mu) = \sum_{k=1}^\nu Z_k^2 = \|Z\|^2$. On a donc le résultat suivant.

cor:nor_indep_chi2

Corollaire 9.3 *Si* $X_1, \ldots, X_\nu$ *sont* ν *variables aléatoires indépendantes, gaussiennes de variances non nulles, alors* $(X - \mu)^\top V^{-1}(X - \mu) \sim \chi^2(\nu)$ *où* μ *et* V *sont respectivement l'espérance et la matrice des variances-covariances du vecteur* $X = (X_1, \ldots, X_\nu)^\top$.

Le corollaire ci-dessus montre que dès que des variables gaussiennes ont une matrice des variances-covariances V diagonale, alors la forme quadratique $(X - \mu)^\top V^{-1}(X - \mu)$ est une variable aléatoire suivant une loi du χ^2. On a un résultat plus général qui exploite la propriété centrale de l'équivalence entre l'indépendance et la non corrélation pour les variables gaussiennes.

Dans le cas où un vecteur gaussien X n'a pas une matrice des variances-covariances V diagonale (et donc les variables qui constituent les coordonnées de X ne sont pas indépendantes), on peut

effectuer un changement de base pour lequel le vecteur X exprimé dans cette nouvelle base a une matrice des variances-covariances diagonale. Le changement de base étant une opération linéaire, grace aux propriétés 9.7 et 9.8, après changement de base on a toujours un vecteur gaussien, mais dont la matrice des variances-covariances est diagonale. Comme on l'a montré dans le corollaire 9.3, on peut alors lui associer une forme quadratique dont la loi est une loi du χ^2. On énonce et prouve ce résultat de manière formelle.

pro:nor2chi **Propriété 9.17** *Si $X \sim \mathcal{N}_\nu(\mu, V)$ avec V inversible, alors $(X-\mu)^\top V^{-1}(X-\mu) \sim \chi^2(\nu)$.*

Preuve : Soit P la matrice de passage telle que $\Lambda = P^{-1}VP$ est diagonale. Comme V est réelle symétrique, on peut choisir P orthonormée, *i.e.*, telle que $P^\top P = I_\nu$. Dans ce cas $P^{-1} = P^\top$ et $\Lambda = P^\top VP$. Si on définit le vecteur aléatoire de Y de $\mathbb{R}^\nu$ par $Y = P^\top(X-\mu)$, alors d'après la propriété 9.8, Y est un vecteur gaussien, et en utilisant la propriété 9.7, on a $\mathrm{E}(Y) = P^\top \mathrm{E}(X-\mu) = 0_\nu$ et $\mathrm{V}(Y) = P^\top \mathrm{V}(X-\mu)P = P^\top VP = \Lambda$. Autrement dit $Y \sim \mathcal{N}_\nu(0_\nu, \Lambda)$. Comme Λ est diagonale, en utilisant le corollaire 9.3 on a $Y^\top \Lambda^{-1} Y \sim \chi^2(\nu)$. Or en utilisant le fait que $P^{-1} = P^\top$, on a $\Lambda^{-1} = P^\top V^{-1} P$. Donc $Y^\top \Lambda^{-1} Y = (X-\mu)^\top P \Lambda^{-1} P^\top (X-\mu) = (X-\mu)^\top V^{-1}(X-\mu)$. D'où le résultat.

On constate que Y et X désignent le même vecteur de $\mathbb{R}^\nu$ exprimé dans les deux différentes bases, pour lesquelles la matrice de passage est P

Il existe un autre cas important dans lequel une forme quadratique en un vecteur gaussien a une distribution du χ^2.

pro:chi2_idempot **Propriété 9.18** *Soit A une matrice symétrique idempotente (*i.e.*, $A^2 = A$) de dimensions $\nu \times \nu$. Soit $X \sim \mathcal{N}_\nu(0_\nu, I_\nu)$. On a $X^\top AX \sim \chi^2(r)$ où r est le rang de A.*

Preuve : Comme A est symétrique, on peut trouver un matrice de passage P orthonormée et une matrice diagonale Λ telles que $\Lambda = P^\top AP$, ou encore telles que $A = P\Lambda P^\top$. Comme P est orthonormée, on a $A^2 = P\Lambda P^\top P\Lambda P^\top = P\Lambda^2 P^\top$. Comme $A^2 = A$ on doit avoir $P\Lambda^2 P^\top = P\Lambda P^\top$ ou encore $\Lambda^2 = \Lambda$ (puisque P est inversible). Comme Λ est diagonale, cette égalité équivaut à $\lambda_i^2 = \lambda_i$, $i = 1, \ldots, \nu$ où λ_i est le i^{e} élément diagonal de Λ. Par conséquent $\lambda_i \in \{0, 1\}$ pour tout i. On note q ($q \leq \nu$) le nombre d'éléments diagonaux de Λ égaux à 1. Quitte à changer l'ordre des lignes de P et de Λ, on peut toujours supposer que les q premiers éléments diagonaux de Λ sont égaux à 1 et les $\nu - q$ derniers égaux à 0. Schématiquement, la matrice Λ a la forme

$$\Lambda = \left(\begin{array}{c:c} I_q & 0 \\ \hdashline 0 & 0 \end{array}\right)$$

Son rang est évidemment égal à q. De plus, comme $\Lambda = P^\top AP$ et que P est inversible (de rang p), Λ a le même rang que A. On a donc $q = r$. On définit à présent le vecteur $Y = P^\top X$. En appliquant les propriétés 9.7 et 9.8 et le fait que P est orthonormée, on déduit que $Y \sim \mathcal{N}_\nu(0_\nu, I_\nu)$. On peut alors écrire $X^\top AX = X^\top P\Lambda P^\top X = Y^\top \Lambda Y = \sum_{i=1}^r Y_i^2$, où la dernière égalité provient de la forme de Λ. En appliquant la définition 9.4, on obtient le résultat voulu.

Remarque 9.5 La réciproque de ce résultat est également vraie : si pour une matrice A symétrique et un vecteur $X \sim \mathcal{N}_\nu(0_\nu, I_\nu)$, la variable aléatoire $X^\top AX$ suit une loi du $\chi^2(r)$, alors A est idempotente et de rang r. Ces résultats sont des cas particuliers d'un résultat plus général appelé théorème de COCHRAN.

sec:fisher

9.1.3.2 La loi de Fisher

def:fisher

Définition 9.5 *La loi de Fisher à (n, d) degrés de liberté est la loi de la variable aléatoire $F = \frac{C_n/n}{C_d/d}$, où $C_n \sim \chi^2(n)$ et $C_d \sim \chi^2(d)$ sont indépendantes. On note $F \sim F(n, d)$*

Propriété 9.19 *Si $F \sim F(n, d)$ alors :*

1. $\mathrm{P}(F \leq x) = 0$ *si* $x \leq 0$ *et*

$$\mathrm{P}(F \leq x) = \frac{\Gamma(\frac{n}{2} + \frac{d}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{d}{2})} n^{\frac{n}{2}} d^{\frac{d}{2}} \int_0^x \frac{t^{\frac{n}{2}-1}}{(d+nt)^{\frac{n+d}{2}}} \mathrm{d}t$$

   si $x > 0$.

2. $\mathrm{E}(F)$ *n'est définie que pour* $n > 2$ *et on a* $\mathrm{E}(F) = \frac{n}{n-2}$. *La variance de F n'est définie que pour* $n > 4$ *et on a* $\mathrm{V}(F) = \frac{2n^2(n+d+2)}{n(d-4)(d-2)^2}$.
3. $\frac{1}{F} \sim F(d, n)$.

Détermination des valeurs de la distribution de Fisher au moyen d'un tableur et du logiciel R :

	Probabilité $\mathrm{P}(F \leq x)$	Quantile q_p : $\mathrm{P}(F \leq q_p) = p$
Tableur	`1 - LOI.F(`x`;`n`;`d`)`	`INVERSE.LOI.F(`$1-p$`;`n`;`d`)`
R	`pf(`x`,`n`,`d`)`	`qf(`p`,`n`,`d`)`

sec:student

9.1.3.3 La loi de Student (et loi de Cauchy)

Définition 9.6 *La loi de Student à ν degrés de libertés est la loi de la variable aléatoire T_ν définie par*

$$T_q = \frac{Z}{\sqrt{\frac{C}{\nu}}}$$

où $Z \sim \mathcal{N}(0, 1)$ et $C \sim \chi^2(\nu)$ sont indépendantes. On note $T_\nu \sim \tau(\nu)$.

pro:loiT

Propriété 9.20 *Si $T_\nu \sim \tau(\nu)$ alors :*

1. $\mathrm{P}(T_\nu \leq x) = \frac{1}{\sqrt{\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \int_{-\infty}^x (1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}} \mathrm{d}t$
2. $\mathrm{E}(T_\nu)$ *n'existe que si* $\nu > 1$ *et on a* $\mathrm{E}(T_\nu) = 0$. *La variance de T_ν n'est définie que pour* $\nu > 2$ *et on a* $\mathrm{V}(T_\nu) = \frac{\nu}{\nu-2}$.
3. $(T_\nu)^2 \sim F(1, \nu)$.

Si $\nu = 1$, la loi de T_1 est la loi de Cauchy (T_1 est le ratio de deux lois $\mathcal{N}(0,1)$ indépendantes). Cette variable ne possède aucun moment.

Détermination des valeurs de la distribution de Student au moyen d'un tableur et du logiciel R :

	Probabilité $\mathrm{P}(T_\nu \leq x)$	Quantile q_p : $\mathrm{P}(T_\nu \leq q_p) = p$
Tableur	`1-LOI.STUDENT(`x`;`ν`;1)`	`LOI.STUDENT.INVERSE(`$2(1-p)$`;`ν`)`
R	`pt(`x`,`ν`)`	`qt(`p`,`ν`)`

sec:proj

9.2 Projection orthogonale

On présente dans cette section les principaux résultats liés au problème de la projection orthogonale d'un espace sur un autre. Si on se donne un ensemble E et une partie F de E, l'opération de projection consiste à associer à tout $x \in E$ un élément y de F qu'on peut interpréter comme une approximation de x. Il existe évidemment plusieurs manières de réaliser une projection. Les possibilités offertes dépendent des structures qu'on donne aux ensembles E et F. On considérera ici que E a une structure d'espace vectoriel et que F est un sous-espace de E.

Avec une telle structure, on sait que si G est un sous-espace de E supplémentaire de F, alors tout élément (vecteur) x de E se décompose de manière unique comme la somme d'un élément de F et d'un élément de G : $x = x_F + x_G$, avec $x_F \in F$ et $x_G \in G$. Un manière de projeter x sur F est d'associer à x l'élément x_F de F dans la décomposition de x sur F et G. Dans ce cas, on appelle x_F la projection de x sur F, parallèlement à G. Ce mécanisme est illustré par le graphique de la figure 9.4. E est l'espace $\mathbb{R}^3$ et F est un hyperplan de $\mathbb{R}^3$ passant par l'origine. Le supplémentaire G de F est n'importe quelle droite de $\mathbb{R}^3$ passant par l'origine, et qui n'appartient pas à F. On choisit n'importe quel vecteur x dans l'espace et on fait apparaître sa décomposition en la somme d'un élément x_G de G (en rouge) et d'un élément x_F de F (en bleu). La figure montre que la projection de x sur F parallèlement à G s'obtient en se déplaçant de x (dans l'espace) vers l'hyperplan F dans une direction qui est parallèle à la droite G. On voit que cette projection est la "coordonnée" de x dans F.

Lorsqu'on dote E d'un produit scalaire, pour un sev donné F de E il existe un choix particulier de supplémentaire qui permet de définir la projection orthogonale de x sur F. Le sous-ensemble de E formé de tous les vecteurs de E orthogonaux à F (c'est à dire tous les vecteurs de E orthogonaux à n'importe quel vecteur de F) est appelé l'orthogonal de F et noté $F^\perp$. Formellement, on définit $F^\perp = \{x \in E \mid \langle x, y\rangle = 0 \ \forall y \in F\}$. Cet ensemble $F^\perp$ possède les propriétés suivantes.

pro:proj1

Propriété 9.21 (Propriétés de $F^\perp$) *Soit F un sous espace de E avec $\dim(F) = p$ et $\dim(E) = n$, n fini.*

it:proj1

1. *Soient $f_1, \ldots, f_p$ des vecteurs de E constituant une base de F. Alors $F^\perp = \{x \in E \mid \langle x, f_i\rangle = 0, i = 1, \ldots, p\}$*
2. *$F^\perp$ est un sous-espace de E*
3. *$F \cap F^\perp = \{0_E\}$*

it:suplortho

4. *$E = F \oplus F^\perp$: pour tout vecteur x de E, on peut trouver une paire unique $(x_1, x_2) \in F \times F^\perp$ telle que $x = x_1 + x_2$.*

it:orthorth

5. *Si H est un sev de E tel que $H \subseteq F$, alors $F^\perp \subseteq H^\perp$.*

Preuve :

1. Si x est dans $F^\perp$ il est orthogonal à tout vecteur de F et donc en particulier orthogonal aux vecteurs de la base de F. Si x est orthogonal à tous les vecteurs de la base de F, alors par (bi)linéarité du produit scalaire, il est orthogonal à toute combinaison linéaire des vecteurs de cette base, c'est à dire à tout vecteur de F.
2. On vérifie que $0_E \in F^\perp$ et que si x et y sont deux vecteurs dans $F^\perp$, alors pour des scalaires quelconques α et β le vecteurs $\alpha x + \beta y$ est orthogonal à F : pour tout

vecteur z de F, on a $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle = 0$ puisque x et y sont tous les deux orthogonaux à z. Donc $\alpha x + \beta y$ est orthogonal à z.

3. Si $x \in F \cap F^\perp$ on doit avoir que x est orthogonal à lui même : $\langle x, x\rangle = 0$. D'où $x = 0_E$.
4. La preuve de cette propriété repose sur les deux résultats suivants :
   (a) tout ev de dimension finie possède une base de vecteurs orthogonaux
   (b) on peut compléter une famille de p vecteurs orthogonaux d'un espace de dimension n par $n - p$ vecteurs orthogonaux pour former une base de cet espace

   En utilisant le point (a), on peut trouver une base orthogonale $e_1, \ldots, e_p$ de F et par le point (b), on peut la compléter par $e_{p+1}, \ldots, e_n$ pour obtenir une base de E. On note G l'espace engendré par $e_{p+1}, \ldots, e_n$. Tout vecteur x de E, dont les coordonnées sont $x_1, \ldots, x_n$ dans la base $e_1, \ldots, , e_n$, s'écrit de manière unique comme la somme $x = x_F + x_G$ avec $x_F = \sum_{i=1}^{p} x_i e_i \in F$ et $x_G = \sum_{i=p+1}^{n} x_i e_i \in G$. Autrement dit $E = F \oplus G$.

   On vérifie que $G = F^\perp$. Si $x \in G$, il est clair que $x \in F^\perp$. Choisissons $x \in F^\perp$ et montrons que $x \in G$. On a $x \in F^\perp$ ssi $\langle e_i, x\rangle = 0, \forall i = 1, \ldots, p$. Or $\langle e_i, x\rangle = \sum_{j=1}^{n} x_j \langle e_i, e_j\rangle = x_i \|e_i\|^2$, par orthogonalité des $e_1, \ldots, e_n$. Donc $x \in F^\perp$ ssi $x_i\|e_i\|^2 = 0, \forall i = 1, \ldots, p$. Comme les vecteurs de la base ne peuvent être nuls, $x \in F^\perp \iff x_i = 0, \forall i = 1, \ldots, p$. Donc tout vecteur de $F^\perp$ est dans G.
5. Si $x \in F^\perp$ alors il est orthogonal à tout vecteur de F, et donc en particulier orthogonal à tout vecteur de H. Donc x est dans $H^\perp$.

On voit donc que $F^\perp$ est un supplémentaire de F et la *projection orthogonale* sur F d'un vecteur x de E est la projection de x sur F, parallèlement à $F^\perp$. Cette projection est donc est l'unique vecteur x_F de F pour lequel on a $x = x_F + x_{F^\perp}$ avec $x_{F^\perp} \in F^\perp$. Dans un tel contexte, on introduit l'application qui associe à tout x de E sa projection orthogonale sur F ; on note cette application $\mathrm{proj}_F(x)$: si $x = x_F + x_{F^\perp}$, alors $\mathrm{proj}_F(x) = x_F$.

pro:PF **Propriété 9.22 (Propriétés de proj_F)**

it:lin_PF 1. proj_F *est une application linéaire.*

it:ident_PF 2. $\mathrm{proj}_F(x) = x$ *si* $x \in F$ *et* $\mathrm{proj}_F(x) = 0_E$ *si* $x \in F^\perp$.

it:composPF 3. $(\mathrm{proj}_F \circ \cdots \circ \mathrm{proj}_F)(x) = \mathrm{proj}_F(x)$.

it:PF 4. *On se donne* $\mathcal{B}$ *une base de* E. *Soit* $f_1, \ldots, f_p$ *une base de* F *et* $A = (a_{ij})_{i=1,\ldots,n, j=1,\ldots,p}$ *la matrice dont la* j^e *colonne contient les* n *coordonnées de* f_j *dans la base* $\mathcal{B}$. *On désigne par* X *et* X_F *les* n*-uplets de scalaires désignant les coordonnées de* x *et* x_F *dans la base* $\mathcal{B}$, *respectivement. Dans cette base, la matrice représentant* proj_F *est* $A(A^\top A)^{-1}A^\top$. *Les coordonnées de* $x_F = \mathrm{proj}_F(x)$ *sont donc* $X_F = A(A^\top A)^{-1}A^\top X$.

it:PFmin 5. *Pour tout* $x \in E$, *on note* $\|x\|$ *la norme de* x *induite par le produit scalaire :* $\|x\| = \sqrt{\langle x, x\rangle}$. *Pour* x *quelconque dans* E, $\mathrm{proj}_F(x)$ *est solution du problème* $\min_{y\in F} \|x - y\|$ *:* $\forall y \in F$, $\|x - y\| \geq \|x - \mathrm{proj}_F(x)\|$.

Preuve :

1. Soient x et y deux vecteurs de E. Il existe $(x_1, x_2) \in F \times F^\perp$ et $(y_1, y_2) \in F \times F^\perp$ uniques, tels que $x = x_1 + x_2$ et $y = y_1 + y_2$. Donc pour des scalaires quelconques α et β on a

$$\alpha x + \beta y = \alpha(x_1 + x_2) + \beta(y_1 + y_2) = (\alpha x_1 + \beta y_1) + (\alpha x_2 + \beta y_2) \tag{9.4}$$ eq:proj1

Puisque F et $F^\perp$ sont des ev, $(\alpha x_1 + \beta y_1) \in F$ et $(\alpha x_2 + \beta y_2) \in F^\perp$, et l'égalité (9.4) est donc l'unique décomposition du vecteur $\alpha x + \beta y$ de E sur F et $F^\perp$ (voir le point 4 de la propriété 9.21). Donc par définition $(\alpha x_1 + \beta y_1)$ est la projection orthogonale de $(\alpha x + \beta y)$ sur F et on a

$$\text{proj}_F(\alpha x + \beta y) = \alpha \text{proj}_F(x) + \beta \text{proj}_F(y)$$

2. Si $x \in F$, l'unique décomposition de x sur F et $F^\perp$ est $x = x + 0_E$. Donc $\text{proj}_F(x) = x$. On procède de même pour montrer que $\text{proj}_F(x) = 0_E$ si $x \in F^\perp$.
3. Par définition, $\text{proj}_F(x) \in F$. D'après le point qui précède, $\text{proj}_F(\text{proj}_F(x)) = \text{proj}_F(x)$.
4. Soit $x \in E$ et x_F sa projection orthogonale sur F. Comme $x_F \in F$, il existe des scalaires $\lambda_1, \ldots, \lambda_p$ pour lesquels on a $x_F = \lambda_1 f_1 + \cdots + \lambda_p f_p$. En utilisant les coordonnées de $f_1, \ldots, f_p$ dans la base $\mathcal{B} = (b_1, \ldots, b_n)$, on peut écrire

$$x_F = \left(\sum_{i=1}^p \lambda_i a_{1i}\right) b_1 + \cdots + \left(\sum_{i=1}^p \lambda_i a_{ni}\right) b_n$$

ou encore

$$X_F = \begin{pmatrix} \sum_{i=1}^p \lambda_i a_{1i} \\ \vdots \\ \sum_{i=1}^p \lambda_i a_{ni} \end{pmatrix} = \sum_{i=1}^p \lambda_i \begin{pmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{pmatrix} = A\lambda$$

où $\lambda = (\lambda_1, \ldots, \lambda_p)^\top$. Par conséquent, la décomposition $x = x_F + x_{F^\perp}$ donne pour les coordonnées : $X = X_F + X_{F^\perp} = A\lambda + X_{F^\perp}$. On a alors $A^\top X = A^\top A\lambda + A^\top X_{F^\perp}$. Or, par construction de A,

$$A^\top X_{F^\perp} = \begin{pmatrix} \langle f_1, x_{F^\perp} \rangle \\ \vdots \\ \langle f_p, x_{F^\perp} \rangle \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \tag{9.5}$$ eq:cond_ortho

puisque $x_{F^\perp}$ est orthogonal aux vecteurs de la base de F. Donc $A^\top X = A^\top A\lambda$. Par construction, A est de plein rang colonne, donc de rang p et par conséquent $A^\top A$ est également de rang p, de dimensions $p \times p$. Elle est donc inversible et on doit avoir $\lambda = (A^\top A)^{-1} A^\top X$. On en déduit que $X_F = A\lambda = A(A^\top A)^{-1} A^\top X$.
5. Soit $x \in E$. Minimiser $\|x - y\|$ revient à minimiser $\|x - y\|^2$. Pour tout $y \in F$, on a

$$\begin{aligned} \|x - y\|^2 &= \|x - \text{proj}_F(x) + \text{proj}_F(x) - y\|^2 \\ &= \|x - \text{proj}_F(x)\|^2 + \|\text{proj}_F(x) - y\|^2 + 2\langle x - \text{proj}_F(x), \text{proj}_F(x) - y \rangle \end{aligned}$$

Comme $\text{proj}_F(x)$ et y sont dans F, le vecteur $\text{proj}_F(x) - y$ l'est également. De plus $x - \text{proj}_F(x) \in F^\perp$. Donc $\langle x - \text{proj}_F(x), \text{proj}_F(x) - y\rangle = 0$ et

$$\|x - y\|^2 = \|x - \text{proj}_F(x)\|^2 + \|\text{proj}_F(x) - y\|^2 \geq \|x - \text{proj}_F(x)\|^2$$

On conclut en notant que $\text{proj}_F(x) \in F$.

Remarque 9.6 Les p égalités (9.5) sont appelées *condition d'orthogonalité*. Elles établissent que $x - x_F$ est un vecteur orthogonal à chacun des vecteurs f_i de la base de F. En appliquant le point 1 des propriétés 9.21, cela équivaut à dire que $x - x_F$ est orthogonal à tout vecteur de F. En utilisant alors l'unicité de la décomposition de $x = x_F + x_{F^\perp}$ sur $F \oplus F^\perp$, on constate que x_F est le seul vecteur de F pour lequel $x - x_F$ est orthogonal à F. On a donc établi le résultat qui suit.

Propriété 9.23 (Caractérisation de x_F) *Sous les conditions de la propriété 9.21, la projection orthogonale de x sur F est l'unique élément x_F de E tel que :*

1. $x_F \in F$
2. $(x - x_F) \perp y,\ \forall y \in F$

cor:proj_emboit **Corollaire 9.4** *Soient F et H des sev de E tels que $H \subseteq F$. Alors pour tout $x \in E$ on a* $\|x - \text{proj}_H(x)\| \geq \|x - \text{proj}_F(x)\|$

Preuve : $H \subseteq F$ implique $\min_{y\in H} \|x - y\| \geq \min_{y\in F} \|x - y\|$ pour tout $x \in E$. Il suffit de remarquer que d'après le dernier point de la propriété précédente $\text{proj}_H(x)$ et $\text{proj}_F(x)$ sont les solutions respectives de ces deux minimisations.

Remarque 9.7 Comme toute les matrice représentant une application linéaire, la matrice représentant proj_F dépend de la base retenue. Dans bien des cas, la base de E est donnée et F est un sev de E engendré par p vecteurs linéairement indépendants de E et ceux-ci sont pris comme base de F. La matrice A contient donc les coordonnées de ces vecteurs dans la base initiale de E. On note P_F la matrice représentant proj_F : $P_F = A(A^\top A)^{-1}A^\top$.

Remarquons que quelle que soit la base de E choisie, si on note X le n-uplet des coordonnées de x dans cette base, on peut toujours écrire

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ X_{p+1} \\ \vdots \\ X_n \end{pmatrix} \tag{9.6}$$ eq:projcoord

Ceci ne correspond pas en général à la décomposition $x = x_F + x_{F^\perp}$. En effet, dans le membre de droite de (9.6), le premier terme est les coordonnées d'un vecteur qui appartient au sev de E engendré par les p premiers vecteurs de sa base, et le second est un vecteur du sev engendré par les $n - p$ derniers vecteurs de cette base. Ces deux sev ne coïncident pas forcément avec F et $F^\perp$.

Cependant, si on s'arrange pour choisir une base de E telle que ses p premiers éléments constituent une base de F et les $n-p$ derniers constituent une base de $F^\perp$, alors dans cette base, la décomposition (9.6) coïncide avec la décomposition $x = x_F + x_{F^\perp}$. Plus précisément, avec ce choix adéquat de base de E, les coordonnées de x_F sont le n-uplet X_F^* dont les p premiers éléments sont les p premières coordonnées de x dans cette base et les $n-p$ deniers sont égaux à 0 : si X^* est le n-uplet des coordonnées de x dans base, on doit avoir

$$X_F^* = \begin{pmatrix} X_1^* \\ \vdots \\ X_p^* \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Avec un tel choix de base, la matrice représentative de proj_F est donc de la forme

$$\left(\begin{array}{c|c} I_p & 0_{p,n-p} \\ \hline 0_{n-p,p} & 0_{n-p,n-p} \end{array}\right)$$

où I_p est la matrice identité d'ordre p. Si on note Λ cette matrice, on vérifie qu'on a bien $X_F^* = \Lambda X^*$.

Un moyen d'obtenir la base recherchée consiste à changer la base de F par une base $\mathcal{U}_F = (u_1, \ldots, u_p)$ de vecteurs orthonormés et de la compléter par $n-p$ vecteurs orthonormés $u_{p+1}, \ldots, u_n$ pour former une base de E (c'est toujours possible, voir la preuve du point 4 de la propriété 9.21). Ces $n-p$ derniers vecteurs sont orthogonaux aux p premiers et constituent une base de $F^\perp$.

Dans la base initiale, la matrice associée à proj_F est $P_F = A(A^\top A)^{-1}A^\top$ et dans la nouvelle base, cette matrice est Λ. Si on note Q la matrice de passage de la base initiale à la nouvelle base, on doit avoir $P_F = Q\Lambda Q^{-1}$. La matrice Q contient les coordonnées des vecteurs de la nouvelle base dans l'ancienne. Comme les vecteurs de la nouvelle base sont orthonormés, on doit avoir $\langle u_i, u_j\rangle = 1$ si $i = j$ et 0 sinon. Si on désigne par q_{ij} l'élément de la i^{e} ligne et j^{e} colonne de Q, alors l'élément à la même position dans $Q^\top Q$ est $\sum_{k=1}^n q_{ki}q_{kj}$, ce qui, par construction de Q, coïncide avec $\langle u_i, u_j\rangle$. Par conséquent on a $Q^\top Q = I_n$, ou encore $Q^{-1} = Q^\top$. Ce qui permet d'écrire la relation suivante entre les matrices représentatives de proj_F dans la base initiale et la nouvelle base orthonormée : $P_F = Q\Lambda Q^\top$ ou, de manière équivalente $\Lambda = Q^\top P_F Q$. Comme Λ est diagonale, on voit que le changement de base qui permet de représenter proj_F est celui qui permet de diagonaliser sa matrice représentative P_F. Autrement dit, les valeurs propres de P_F sont les éléments diagonaux de Λ, et sont donc égales à 1 (avec un degré de multiplicité p) et 0 (avec un degré de multiplicité $n-p$). [2] ☒

rem:proj_comp **Remarque 9.8** On peut définir l'application qui associe à tout $x \in E$ le reste $x_{F^\perp} = x - x_F$ de sa projection orthogonale sur F. Il est facile de voir que cette application est linéaire. Si on note

2. Notons que bien que dans la totalité des utilisations que nous faisons de ces résultats, le corps de scalaires sur lequel est construite la structure d'ev de E est $\mathbb{R}$, dans le cas général, ce corps est quelconque. Par conséquent 1 désigne l'élément neutre pour la multiplication de scalaires et 0 désigne l'élément neutre pour l'addition des scalaires.

P_F la matrice $A(A^\top A)^{-1}A^\top$ représentant l'application proj_F, on voit que $I - P_F$ est la matrice qui représente l'application associant x le vecteur $x_{F^\perp}$ (où I est la matrice identité). En effet, puisque $X = X_F + X_{F^\perp}$ et que $X_F = P_F X$ on a nécessairement $X_{F^\perp} = (I - P_F)X$.

lem:proj_const

Remarque 9.9 Un cas intéressant d'application en statistique est celui où $E = \mathbb{R}^n$ et $F = \mathbb{R}\iota = \{x \in \mathbb{R}^n \mid x = c\,\iota, c \in \mathbb{R}\}$, où ι est le vecteur diagonal de $\mathbb{R}^n$, *i.e.*, celui dont toutes les coordonnées valent 1. F est le sev de $\mathbb{R}^n$ qui contient tous les vecteurs proportionnels à ι, *i.e.* dont les coordonnées sont égales. Par conséquent, si x est un vecteur de $\mathbb{R}^n$, sa projection orthogonale sur F sera un vecteur x_F proportionnel à ι : on aura $x_F = c\,\iota$ pour un certain réel c. On va montrer que le facteur de proportionnalité c est égal à la moyenne des coordonnées de x.

On a évidemment que ι est une base de F et la matrice A de la propriété 9.22 est

$$A = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

D'après le point 4 de la propriété 9.22, la matrice associée à la projection orthogonale sur F est $P_F = A(A^\top A)^{-1}A^\top$. On calcule sans difficulté que $A^\top A = n$, et donc que

$$P_F = \frac{1}{n}AA^\top = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

Soit $X = (X_1, \ldots, X_n)^\top$ le n-uplet des coordonnées de x. La projection orthogonale de x sur F est le vecteur x_F dont les coordonnées sont données par

$$X_F = P_F X = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i \\ \vdots \\ \sum_{i=1}^n X_i \end{pmatrix} = \begin{pmatrix} \overline{X} \\ \overline{X} \\ \vdots \\ \overline{X} \end{pmatrix} = \overline{X}\,\iota$$

où $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ est la moyenne des coordonnées de x. ℝ

La propriété suivante résume les propriétés importantes de la matrice P_F.

pro:propr_PF

Propriété 9.24 (Propriétés de P_F) *Soit F un sev de rang p d'un ev E. Dans une base de E, la matrice $P_F = A(A^\top A)^{-1}A^\top$ représente la projection orthogonale de E sur F, notée proj_F, où A est la matrice des coordonnées des vecteurs d'une base de F. La matrice P_F possède les propriétés suivantes :*

1. *P_F est symétrique*
2. *P_F est de rang p*
3. *P_F est idempotente*

4. *P_F a deux valeurs propres distinctes : 0 et 1. Le degré de multiplicité de la valeur propre 1 est p.*

Preuve :

1. On le vérifie à partir de l'expression de P_F.
2. Par construction, A est de plein rang colonne, égal à p.
3. On le vérifie à partir de l'expression de P_F ou bien en utilisant le point 3 de la propriété 9.22.
4. Ceci a été démontré dans le paragraphe qui suit le corollaire 9.4. On peut le démontrer en utilisant les méthodes usuelles de diagonalisation d'une matrice. Comme P_F est idempotente, toute valeur propre λ doit satisfaire $P_F x = \lambda x$. D'une part, on a (comme pour toute matrice) $P_F^2 x = P_F(P_F x) = P_F \lambda x = \lambda P_F x = \lambda^2 x$ et d'autre part, comme P_F est idempotente, on a aussi $P_F x = \lambda^2 x$. On doit donc avoir $\lambda^2 = \lambda$ pour toute valeur propre de P_F, ce qui implique que les seules valeurs propres possibles sont 0 ou 1. Comme la trace d'une matrice est égale à la somme de ses valeurs propres, la trace de P_F est égal au nombre de valeurs propres égales à 1. On a $\mathrm{tr}(P_F) = \mathrm{tr}(A(A^\top A)^{-1}A^\top) = \mathrm{tr}((A^\top A)^{-1}A^\top A) = \mathrm{tr}(I_p) = p$.

pro:proj_iter **Propriété 9.25** *Soient F et H deux sev de E tels que $H \subseteq F$. Pour tout $x \in E$ on a*

it:proj_iter 1. $\mathrm{proj}_F(\mathrm{proj}_H(x)) = \mathrm{proj}_H(\mathrm{proj}_F(x)) = \mathrm{proj}_H(x)$

it:proj_contr 2. $\mathrm{proj}_F(x) \in H \iff \mathrm{proj}_F(x) = \mathrm{proj}_H(x)$

Preuve :

1. Soit $x \in E$. Notons $x_H = \mathrm{proj}_H(x)$. Comme $x_H \in H$, on a également $x_H \in F$. D'après le point 2 de la propriété 9.22, on a alors $\mathrm{proj}_F(x_H) = x_H$, ou encore $\mathrm{proj}_F(\mathrm{proj}_H(x)) = \mathrm{proj}_H(x)$.
   Notons maintenant $x_F = \mathrm{proj}_F(x)$. On a donc $x = x_F + x_{F^\perp}$, où $x_{F^\perp} \in F^\perp$. D'après le point 5 de la propriété 9.21, on a aussi $x_{F^\perp} \in H^\perp$. En utilisant le point 1 (linéarité de la projection) de la propriété 9.22, on peut écrire
   $$\mathrm{proj}_H(x) = \mathrm{proj}_H(x_F) + \mathrm{proj}_H(x_{F^\perp})$$
   D'après le point 2 de cette même propriété, on a $\mathrm{proj}_H(x_{F^\perp}) = 0$. Donc on doit avoir $\mathrm{proj}_H(x) = \mathrm{proj}_H(x_F)$, ou encore $\mathrm{proj}_H(x) = \mathrm{proj}_H(\mathrm{proj}_F(x))$.
2. Par construction, $x_H \in H$. Par conséquent, si $x_F = x_H$, alors $x_F \in H$. Réciproquement, supposons que $x_F \in H$. Dans ce cas, en appliquant le point 2 de la propriété 9.22, on a $\mathrm{proj}_H(x_F) = x_F$. Mais en utilisant ce qui vient d'être démontré, on a aussi $\mathrm{proj}_H(x_F) = x_H$. En utilisant l'unicité de la décomposition sur des espaces supplémentaires, on a nécessairement $x_F = x_H$.

Cette propriété (dite des projections emboîtées) peut être illustrée graphiquement. Le graphique de la figure 9.6 reprend celui de la figure 9.5, en y rajoutant un sev H de E tel que $H \subseteq F$. Sur la figure, H est une droite (en vert) de E appartenant au plan F. La projection orthogonale de x sur H est x_H. On voit qu'elle coïncide avec la projection orthogonale de x_F sur H (cette opération de projection étant symbolisée par les traits jaunes).

pro:projFH **Propriété 9.26** *Soient F et H deux sev de E. On note $H+F$ le sev dont les éléments s'expriment comme la somme d'un élément de F et d'un élément de H.*

1. *F et H sont orthogonaux si et seulement si* $\text{proj}_F(\text{proj}_H(x)) = 0_E \ \forall x \in E$.
2. *Si F et H sont orthogonaux* $\text{proj}_{F+H}(x) = \text{proj}_F(x) + \text{proj}_H(x) \ \forall x \in E$.

it:FWp 3. $\text{proj}_{F+H} = \text{proj}_F + \text{proj}_{\tilde{H}}$ *où $\tilde{H}$ est l'espace orthogonal à celui obtenu en projetant tous les éléments de H sur F, i.e. $\tilde{H} = \{x - \text{proj}_F(x), x \in H\}$.*

Preuve :

1. Supposons F et H orthogonaux. Comme $\text{proj}_H(x) \in H$, d'après 9.22 (point 2) on a nécessairement $\text{proj}_F(\text{proj}_H(x)) = 0_E$. Réciproquement, supposons $\text{proj}_F(\text{proj}_H(x)) = 0_E$ pour tout $x \in E$. C'est en particulier vrai pour $x \in H$. Dans ce cas, toujours d'après 9.22 (point 2), on a $\text{proj}_F(\text{proj}_H(x)) = \text{proj}_F(x) = 0_E$. Ceci implique (encore d'après 9.22, point 2), que $x \in F^\perp$. Comme ceci est vrai pour tout $x \in H$, on a le résultat voulu.

2. Pour tout $x \in E$, il existe un unique $x_0 \in F+H$ et un unique $x_1 \in (F+H)^\perp$ tel que $x = x_0 + x_1$ et on a $\text{proj}_{F+H}(x) = x_0$. Comme par définition $x_0 \in F+H$, on peut écrire $x_0 = x_{0F} + x_{0H}$ avec $x_{0F} \in F$ et $x_{0H} \in H$. Par ailleurs, on a par linéarité de proj_H (voir propriété 9.22, point 1) :

$$\text{proj}_H(x) = \text{proj}_H(x_{0F} + x_{0H} + x_1) = \text{proj}_H(x_{0F}) + \text{proj}_H(x_{0H}) + \text{proj}_H(x_1) = x_{0H}$$

car d'une part $x_{0F} \in H^\perp$ et $x_{0H} \in H$ (et on applique la propriété 9.22, point 2), et d'autre part, $x_1 \in (F+H)^\perp \subset H^\perp$ (d'après la propriété 9.21 (point 5)). On a de la même manière $\text{proj}_F(x) = x_{0F}$, et le résultat est démontré.

3. La preuve s'obtient en montrant que $F+H$ s'écrit aussi $F+\tilde{H}$. En effet, si cela est établit, on aura nécessairement $\text{proj}_{F+H} = \text{proj}_{F+\tilde{H}}$, et comme par construction $\tilde{H}$ est orthogonal à F, on pourra appliquer le point précédent pour obtenir le résultat voulu. On a $\tilde{H} = \{y = x - \text{proj}_F(x), x \in H\}$ l'ensemble des vecteurs pouvant s'écrire comme la différence d'un $x \in H$ avec sa projection orthogonale sur F. Il est facile de voir de $\tilde{H}$ est un sev de E, et que F et $\tilde{H}$ sont orthogonaux. On montre maintenant que $F+H = F+\tilde{H}$. En effet, soit $x = x_F + x_H \in F+H$. On a

$$x_F + x_H = x_F + \text{proj}_F(x_H) + [x_H - \text{proj}_F(x_H)]$$

Or $x_F + \text{proj}_F(x_H) \in F$ et $x_H - \text{proj}_F(x_H) \in \tilde{H}$. Donc $x \in F+\tilde{H}$. Soit à présent $x = x_F + x_{\tilde{H}} \in F+\tilde{H}$. Par définition de $\tilde{H}$ on a $x_F + x_{\tilde{H}} = x_F + [x_H - \text{proj}_F(x_H)] = [x_F - \text{proj}_F(x_H)] + x_H$, pour un certain $x_H \in H$. Comme $x_F - \text{proj}_F(x_H) \in F$ et $x_H \in H$, on a bien $x \in F+H$. On a donc montré que $F+H = F+\tilde{H}$. Par conséquent, $\text{proj}_{F+H} = \text{proj}_{F+\tilde{H}}$ avec F et $\tilde{H}$ orthogonaux, et d'après le point précédent de la propriété, on a aussi $\text{proj}_{F+\tilde{H}} = \text{proj}_F + \text{proj}_{\tilde{H}}$

Remarque 9.10 Si on note P_F et P_H les matrices de projection orthogonale sur F et H, respectivement (voir la propriété 9.24), on peut réécrire les deux premiers points de la propriété 9.26 de la manière équivalente suivante :

1. F et H sont orthogonaux si et seulement si leurs matrices de projection orthogonales P_F et P_H sont orthogonales, i.e. $P_F P_H = 0$
2. Si F et H sont orthogonaux, alors $P_{F+H} = P_F + P_H$

Par ailleurs, le sev $\tilde{H}$ introduit dans la preuve du troisième point de la propriété s'écrit $\tilde{H} = \{(I - P_F)x, x \in H\}$. C'est le sev obtenu en appliquant la transformation $I - P_F$ aux élément de H. On peut le noter $\tilde{H} = (I - P_F)H$ Par conséquent, la matrice de projection orthogonale sur $\tilde{H}$ peut s'écrire $P_{(I-P_F)H}$. Le résultat établit alors que $P_{F+H} = P_F + P_{(I-P_F)H}$.

rem:fwproj **Remarque 9.11** Le dernier point de la propriété précédente est particulièrement utile dans le cas où $f_1, \ldots, f_p$ forment un base de F et $h_1, \ldots, h_q$ forment une base de H. L'espace $F + H$ est donc l'ensemble de tous les vecteurs s'écrivant comme $\sum_{i=1}^p x_i f_i + \sum_{j=1}^q y_j h_j$. On peut supposer les vecteurs $f_1, \ldots, f_p, h_1, \ldots, h_q$ linéairement indépendants et dans ce cas, ils forment une base de $G = F + H$, qui est de dimension $p + q$. En utilisant les résultats sur les espaces vectoriels, on peut compléter la famille $f_1, \ldots, f_p$ par q vecteurs $\tilde{h}_1, \ldots, \tilde{h}_q$ de manière que $f_1, \ldots, f_p, \tilde{h}_1, \ldots, \tilde{h}_q$ forment une base de G ; et d'après le procédé de Gram-Schmidt, ces q vecteurs peuvent être choisis orthogonaux à f_i, $i = 1, \ldots, p$, avec plus précisément : $\tilde{h}_j = h_j - \text{proj}_F(h_j)$.
On peut vérifier que tout vecteur de $\tilde{H}$ s'écrit comme une combinaison linéaire de $\tilde{h}_1, \ldots, \tilde{h}_q$. En effet, $y \in \tilde{H} \iff y = x - \text{proj}_F x$ pour un $x \in H$. Mais x doit nécessairement s'écrire $x = \sum_{j=1}^q x_j h_j$, et par linéarité de la projection, on a

$$y = \sum_{j=1}^q x_j h_j - \text{proj}_F(\sum_{j=1}^q x_j h_j) = \sum_{j=1}^q x_j h_j - \sum_{j=1}^q x_j \text{proj}_F(h_j) = \sum_{j=1}^q x_j \left[h_j - \text{proj}_F(h_j)\right] = \sum_{j=1}^q x_j \tilde{h}_j$$

On peut également vérifier que ces vecteurs sont linéairement indépendants. En effet, si on note P_F la matrice de la projection orthogonale sur F, on a $\tilde{h}_j = M_F h_j$, où $M_F = I - P_F$. Soient $a_1, \ldots, a_q$ des scalaires tels que $\sum_{j=1}^q a_j \tilde{h}_j = 0$. On a dans ce cas

$$\sum_{j=1}^q a_j M_F h_j = M_F \sum_{j=1}^q a_j h_j = 0$$

Autrement dit, $\sum_{j=1}^q a_j h_j$ appartient au noyau de M_F. Par construction de M_F, ce noyau est évidemment F. Donc on doit avoir $\sum_{j=1}^q a_j h_j \in F$, ce qui implique que (1) soit $\sum_{j=1}^q a_j h_j = 0$, (2) soit $\sum_{j=1}^q a_j h_j$ est une combinaison linéaire des éléments de la base de F. Comme on a supposé au départ que $(f_1, \ldots, f_p)$ et $(h_1, \ldots, h_q)$ étaient des familles linéairement indépendantes, la seconde possibilité est exclue. On doit donc avoir $\sum_{j=1}^q a_j h_j = 0$, et par indépendance linéaire des $h_1, \ldots, h_q$, on a aussi $a_1 = \cdots = a_q = 0$. En résumé $\sum_{j=1}^q a_j \tilde{h}_j = 0 \implies a_1 = \cdots = a_q = 0$. Les vecteurs $\tilde{h}_1, \ldots, \tilde{h}_q$ sont donc linéairement indépendants.
En combinant les deux résultats précédents, on conclut que $\tilde{h}_1, \ldots, \tilde{h}_q$ forment une base de $\tilde{H}$. Par conséquent, on a $G = F + \tilde{H}$. Donc la projection sur $G = F + H$ peut également se voir comme la projection sur $F + \tilde{H}$, et comme par construction les vecteurs de la base de $\tilde{H}$ sont orthogonaux à ceux de F, F et $\tilde{H}$ sont orthogonaux et on peut appliquer le résultat du deuxième point de la propriété.
En résumé, en posant $\mathcal{F} = (f_1, \ldots, f_p)$ et $\mathcal{H} = (h_1, \ldots, h_q)$, la preuve consiste à changer la base $(\mathcal{F}, \mathcal{H})$ de $G = F + H$ en une base $(\mathcal{F}, \tilde{\mathcal{H}})$ de manière que $\tilde{\mathcal{H}}$ soit orthogonale à $\mathcal{F}$. Pour y parvenir,

on utilise le procédé de Gram-Schmidt qui consiste à prendre le reste de la projection orthogonale de $\mathcal{H}$ sur $\mathcal{F}$.[3]

3. Dans une telle formulation, on utilise un raccourci conceptuel permettant d'assimiler un espace vectoriel aux vecteurs de sa base.

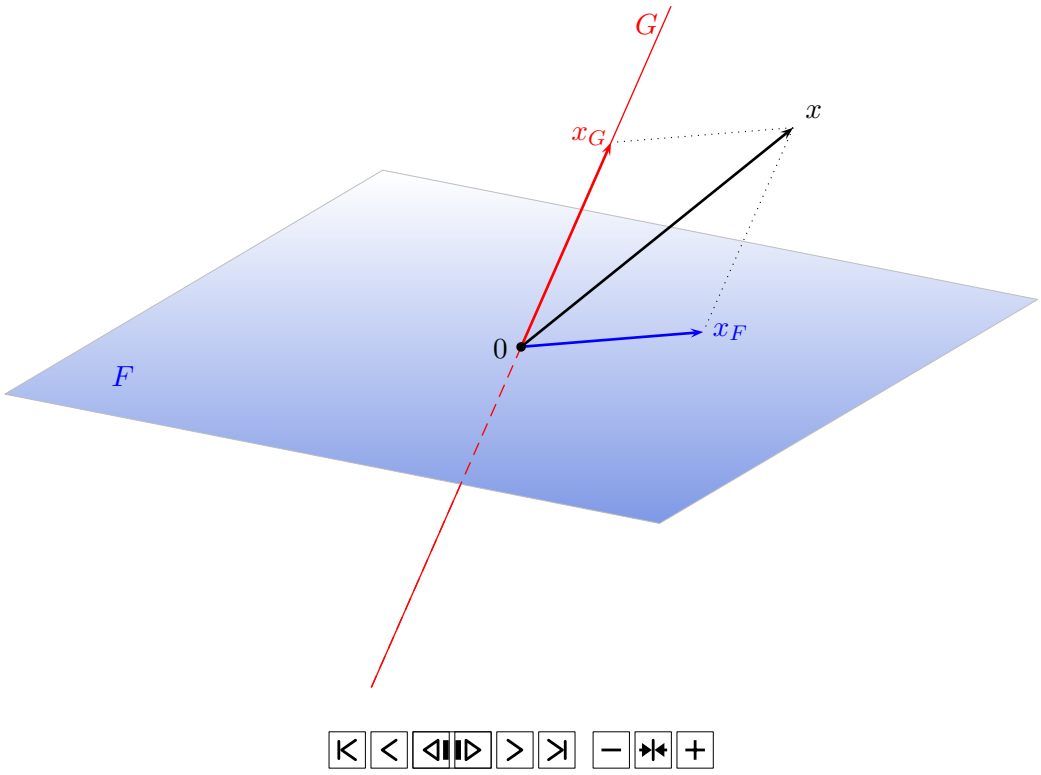


fig:proj

FIGURE 9.4: Projection d'un vecteur x de $E = \mathbb{R}^3$ sur un hyperplan F, parallèlement à G

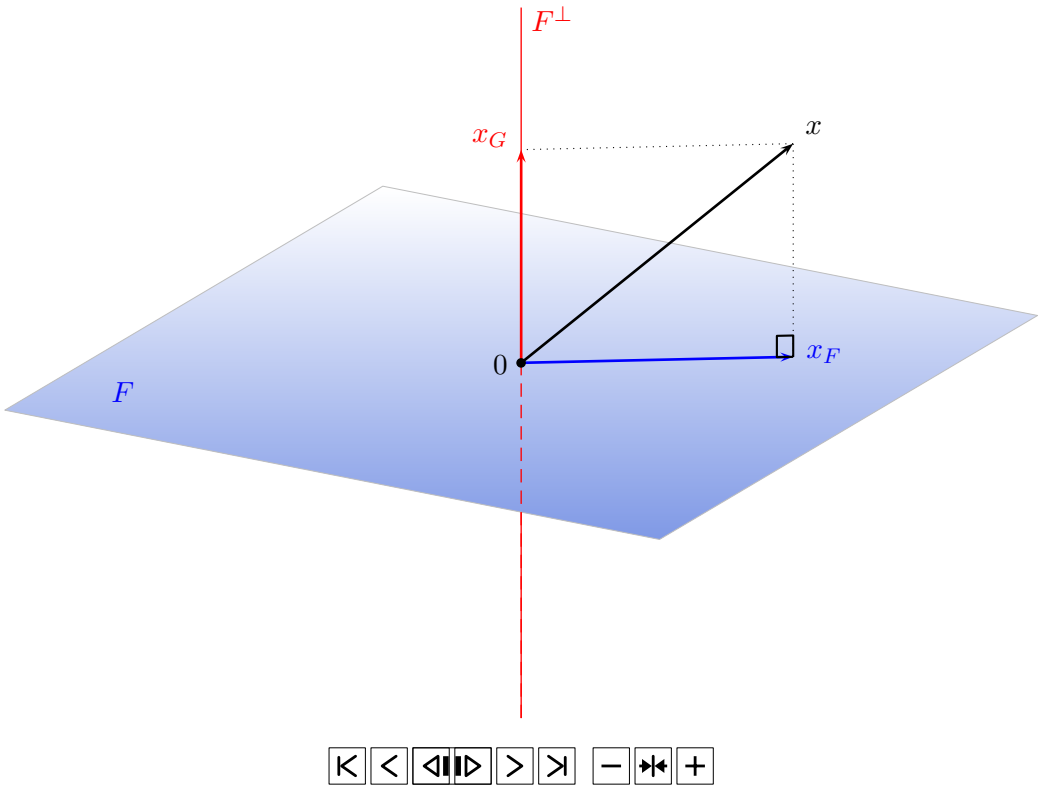


fig:projorth

FIGURE 9.5: Projection orthogonale d'un vecteur x de $E = \mathbb{R}^3$ sur un hyperplan F

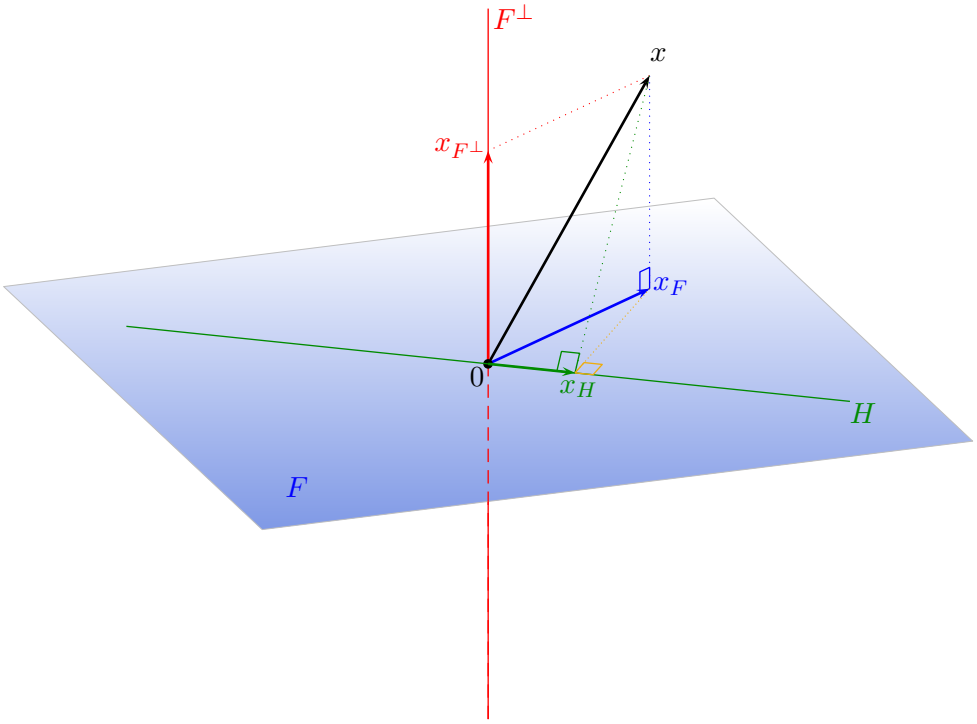


fig:projproj

FIGURE 9.6: Illustration des projections emboîtées : si $H \subseteq F$ alors $\text{proj}_H(\text{proj}_F(x)) = \text{proj}_H(x)$

sec:normat

9.3 Normes matricielles

On aura besoin d'étudier la convergence de suites de matrices. Les matrices qu'on considère seront supposées carrées. On peut dire qu'une suite de matrices $\{A_n : n \geq 1\}$ de dimensions $q \times q$ converge vers une matrice A si pour tout $i, j = 1, \ldots, q$, la suite $\{a_{n,ij} : n \geq 1\}$ converge vers a_{ij}. Même si elle convient parfaitement, cette définition présente l'inconvénient que, sans autre caractérisation, on est obligé d'étudier la convergence de q^2 suites.

Pour des vecteurs de $\mathbb{R}^m$ (c'est à dire pour des matrices $m \times 1$), on étudie la convergence au moyen de la convergence de la suite des normes : la suite de vecteurs $\{a_n : n \geq 1\}$ converge vers $a \in \mathbb{R}^m$ si et seulement si la suite $\{\|a_n - a\| : n \geq 1\}$ converge vers 0. Cette caractérisation de la convergence est justifiée par le fait que la norme de $a_n - a$ est aussi la distance entre a_n et sa limite a. On voit alors qu'on n'est pas tenu d'étudier la limite de chaque suite $\{a_{ni} : n \geq 1\}$ pour $i = 1, \ldots, m$, mais seulement celle formée à partir de la norme. Il faut pour cela disposer d'une norme. On étudie dans cette section les manières de définir une norme sur des espaces vectoriels de matrices, afin de pouvoir caractériser la convergence de suites de matrices.

sec:norm.mat.def

9.3.1 Définition et propriétés

On note $M_q(\mathbb{R})$ l'espace vectoriel des matrices carrées de taille $q \times q$ et à entrées réelles. Une norme sur $M_q(\mathbb{R})$ est appelée norme matricielle.

def:norme.mat

Définition 9.7 *Une norme matricielle est une application* $|||\cdot||| : M_q(\mathbb{R}) \to \mathbb{R}$ *satisfaisant*

1. *(positivité)* $|||A||| \geq 0$ *et* $|||A||| = 0 \iff A = 0_{(q,q)}$
2. *(homogénéité)* $|||aA||| = |a| \; |||A||| \; \forall a \in \mathbb{R}$
3. *(sous-additivité)* $|||A + B||| \leq |||A||| + |||A|||$
4. *(sous-multiplicativité)* $|||AB||| \leq |||A||| \; |||B|||$

pour n'importe quelles matrices A et B dans $M_q(\mathbb{R})$.

On constate qu'une norme matricielle est une norme satisfaisant la propriété de sous-multiplicativité.

Remarque 9.12 Parmi toutes les normes matricielles, on peut mentionner celle notée $|||\cdot|||_\infty$ définie par $|||A|||_\infty = \max\{a_{ij}, i, j = 1, \ldots, q\}$

rem:normat.puiss

Remarque 9.13 La sous-multiplicativité des normes sur $M_q(\mathbb{R})$ a une conséquence importante. En effet, on doit avoir $|||A^2||| \leq |||A|||^2$ et donc $|||A^n||| \leq |||A|||^n$ pour tout $n \geq 1$. R

rem:normat1

Remarque 9.14 Lorsqu'on dispose d'une norme matricielle, on peut caractériser la convergence de $\{A_n : n \geq 1\}$ vers A par la convergence de $|||A_n - A|||$ vers 0. Remarquons que la sous-additivité permet d'établir, comme avec toute norme, que si $A_n \to A$, alors $|||A_n||| \to |||A|||$.[4] R

4. Il suffit simplement d'utiliser la sous-additivité pour borner supérieurement $|||A_n||| = |||(A_n - A) + A|||$ et $|||A||| = |||(A_n - A) + A_n|||$, ce qui équivaut alors à $\big| \; |||A_n||| - |||A||| \; \big| \leq |||A_n - A|||$.

rem:oplimmat **Remarque 9.15** La convergence ainsi définie satisfait les opérations usuelles importantes sur les limites. En particulier, si $\{A_n : n \geq 1\}$ et $\{B_n : n \geq 1\}$ sont deux suites de $M_q(\mathbb{R})$ telles que $A_n \to A$ et $B_n \to B$, alors $A_n + B_n \to A + B$, $\lambda A_n \to \lambda A$ et $A_n B_n \to AB$. Ces propriétés se montrent en utilisant les quatre conditions qui définissent la norme utilisée pour établir les convergences. Par exemple,

$$|||A_n B_n - AB||| = |||(A_n - A)B_n + A(B_n - B)||| \leq |||A_n - A|||\ \ |||B_n|||\ + |||\ A|||\ \ |||B_n - B||| \to 0$$

où l'inégalité est obtenue par sous-additivité et sous-multiplicativité.

Remarque 9.16 La convergence de A_n vers A étant définie à l'aide d'une norme matricielle, on peut envisager que cette convergence puisse se produire pour un certain choix de la norme, mais que si on change de norme, la convergence n'ait plus lieu. Cependant, il est possible de montrer que si on se donne deux normes matricielles $|||\cdot|||_1$ et $|||\cdot|||_2$ sur $M_q(\mathbb{R})$, alors on peut trouver deux réels strictement positifs μ et ν tels que

$$\mu\ |||\ A|||_1 \leq |||A|||_2 \leq \nu\ |||\ A\ |||_1 \qquad \forall A \in M_q(\mathbb{R})$$

Autrement dit si $|||A_n - A|||_1 \to 0$, on aura également $|||A_n - A|||_2 \to 0$. Ce résultat est important puisqu'il suffit par exemple de savoir démontrer la convergence d'une suite de matrices avec une norme bien choisie pour que cette convergence ait aussi lieu avec n'importe quelle autre norme. R

Remarque 9.17 En utilisant la remarque précédente, on peut faire le lien entre la convergence d'une suite de matrice $\{A_n : n \geq 1\}$, exprimée au moyen d'une norme matricielle, et la convergence de chacune des q^2 suites $\{a_{ij,n} : n \geq 1\}$, $i, j = 1, \ldots, q$. En effet, considérons la norme matricielle $|||A||| = \max\{\sum_{j=1}^q |a_{ij}|, i = 1, \ldots, q\}$ (Exercice : vérifier que c'en est bien une). Si avec cette norme $A_n \to 0_{q,q}$, alors toutes les suites $\{a_{ij,n} : n \geq 1\}$ convergent vers 0. En effet $A_n \to 0_{q,q} \iff \max\{\sum_{j=1}^q |a_{ij,n}|, i = 1, \ldots, q\} \to 0$. Cela signifie que pour tout $\varepsilon > 0$ il existe n^* tel que $n > n^* \implies \max\{\sum_{j=1}^q |a_{ij,n}|, i = 1, \ldots, q\} < \varepsilon$. Or

$$\begin{aligned}\max\{\sum_{j=1}^q |a_{ij,n}|, i = 1, \ldots, q\} < \varepsilon \iff \sum_{j=1}^q |a_{ij,n}| < \varepsilon \quad \forall i = 1, \ldots, q \\ \implies |a_{ij,n}| < \varepsilon \quad \forall j = 1, \ldots, q \quad \forall i = 1, \ldots, q\end{aligned}$$

Donc si $n > n^*$, on a $|a_{ij,n}| < \varepsilon$ pour tout $i, j = 1, \ldots, q$, ce qui équivaut à la convergence vers 0 de toutes les suites $\{a_{ij,n} : n \geq 1\}$. D'après la remarque précédente, ce qui vient d'être établi avec la norme matricielle particulière $|||A||| = \max\{\sum_{j=1}^q |a_{ij}|, i = 1, \ldots, q\}$ peut s'établir avec n'importe quelle autre norme matricielle.

Le rayon spectral d'une matrice joue un rôle important dans l'étude des propriétés des matrices. Ce rayon spectral est relié à la notion de norme par un ensemble de résultats (voir notamment la section 9.3.2).

Définition 9.8 *Le rayon spectral d'une matrice $A \in M_q(\mathbb{R})$ est le réel $\rho(A)$ défini par* $\rho(A) = \max\{|\lambda| : Ax = \lambda x, x \in \mathbb{R}^q\}$.

Le rayon spectral est donc la plus grande valeur propre en module.

pro:ray.spec **Propriété 9.27** *Pour toute norme sur $M_q(\mathbb{R})$ et toute matrice $A \in M_q(\mathbb{R})$, on a $\rho(A) \leq |||A|||$. De plus, pour toute matrice $A \in M_q(\mathbb{R})$ et $\varepsilon > 0$, il existe une norme matricielle (dépendant de A et de ε) notée $|||\cdot|||_{A,\varepsilon}$ telle que $|||A|||_{A,\varepsilon} < \rho(A) + \varepsilon$.*

Preuve : On ne prouve que la première partie de la propriété. On se donne n'importe quelle norme $|||\cdot|||$ sur $M_q(\mathbb{R})$. Par définition de $\rho(A)$, il existe une valeur propre λ de A telle que $|\lambda| = \rho(A)$. Pour cette valeur propre et le vecteur propre x associé, on a $Axx^\top = \lambda xx^\top$ et donc

$$|||A||| \, |||xx^\top||| \geq |||Axx^\top||| = |||\lambda xx^\top||| = |\lambda| \, |||xx^\top|||$$

où l'inégalité provient de la sous-multiplicativité et la dernière égalité résulte de l'homogénéité (voir la définition 9.7). Par définition de $\rho(A)$, l'inégalité s'écrit aussi $|||A||| \, |||xx^\top||| \geq \rho(A) \, |||xx^\top|||$. Comme x est un vecteur propre de A, $x \neq 0_q$ et d'après la positivité de la norme, $|||xx^\top||| \neq 0$, ce qui donne le résultat voulu. ■

rem:ray.spec **Remarque 9.18** Ce résultat montre donc que le rayon spectral de A est l'infimum de $|||A|||$ lorsqu'on parcourt toutes les normes possibles sur $M_q(\mathbb{R})$. En particulier, $\rho(A) < 1$ si et seulement si il existe une norme $|||\cdot|||$ sur $M_q(\mathbb{R})$ telle que $|||A||| < 1$. En effet, soit A une matrice de $M_q(\mathbb{R})$. Si $|||A||| < 1$ pour une certaine norme, alors la propriété précédente implique $\rho(A) < 1$. Réciproquement, si $\rho(A) < 1$, alors on peut trouver $\varepsilon > 0$ tel que $\rho(A) + \varepsilon < 1$. La propriété 9.27 établit qu'il existe une norme matricielle $|||\cdot||| = |||\cdot|||_{A,\varepsilon}$ sur $M_q(\mathbb{R})$ pour laquelle $|||A||| < \rho(A) + \varepsilon < 1$.

Propriété 9.28 *Soit $A \in M_q(\mathbb{R})$. On a $A^n \to 0_{q,q}$ lorsque $n \to \infty$ si et seulement si $\rho(A) < 1$.*

Preuve : Supposons que $\rho(A) < 1$. On peut trouver $\varepsilon > 0$ tel que $\rho(A) + \varepsilon < 1$, et d'après la propriété 9.27, on peut trouver une norme $|||\cdot|||_{A,\varepsilon}$ sur $M_q(\mathbb{R})$ telle que $|||A|||_{A,\varepsilon} < 1$. On a alors $|||A^n|||_{A,\varepsilon} \leq |||A|||^n_{A,\varepsilon} \to 0$ lorsque $n \to \infty$. Donc A^n converge vers $0_{q,q}$. Par l'équivalence des normes sur $M_q(\mathbb{R})$, cette convergence a également lieu avec n'importe quelle autre norme. Supposons que $A^n \to 0_{q,q}$ lorsque $n \to \infty$. Soit λ la valeur propre de A pour laquelle $\rho(A) = |\lambda|$ et x le vecteur propre associé. On a $A^n x = \lambda^n x$. Si on note $M(x)$ la matrice de $M_q(\mathbb{R})$ dont les q colonnes sont toutes égales à x, on a $A^n M(x) = \lambda^n M(x)$. Comme $A^n \to 0_{q,q}$, on doit aussi avoir $B_n = A^n M(x) \to 0_{q,q}$ (voir la remarque 9.15). Comme $|||B_n||| = \rho(A)^n \, |||M(x)|||$, on ne peut avoir $|||B_n||| \to 0$ que si $\rho(A) < 1$. ■

Remarque 9.19 La remarque 9.18 permet de reformuler cette propriété de la manière suivante : $A^n \to 0_{q,q}$ si et seulement si il existe une norme $|||\cdot|||$ sur $M_q(\mathbb{R})$ pour laquelle $|||A||| < 1$.

pro:conv.An **Propriété 9.29** *Soit $A \in M_q(\mathbb{R})$. Si $\rho(A) < 1$, alors $(I_q - A)$ est inversible et $(I_q - A)^{-1} = \sum_{i=0}^{\infty} A^i$.*

Preuve : Supposons que $\rho(A) < 1$. Il existe une norme $|||\cdot|||$ sur $M_q(\mathbb{R})$ telle que $|||A||| < 1$. Soit $x \in \mathbb{R}^q$ tel que $(I_q - A)x = 0_q$, ou encore $x = Ax$. En définissant $M(x)$ la matrice de $M_q(\mathbb{R})$ dont les q colonnes sont toutes égales à x, on a $x = Ax \iff M(x) = AM(x)$. Donc $|||M(x)||| \leq |||A||| \, |||M(x)|||$, ou encore $(1 - |||A|||) \, |||M(x)||| \leq 0$. Comme $|||A||| < 1$, ceci n'est possible que si $|||M(x)||| = 0$, c'est-à-dire $x = 0_q$. Autrement dit, $(I_q - A)x =$

$0_q \implies x = 0_q$, c'est à dire $I_q - A$ est inversible. Définissons $B_n = I_q + A + A^2 \cdots + A^n$. Notons que $(I_q - A)B_{n-1} = I_q - A^n$, ou encore $B_{n-1} = (I_q - A)^{-1}(I_q - A^n)$. Par conséquent $B_n - (I - A)^{-1} = -(I - A)^{-1}A^{n+1}$ et donc

$$|||B_n - (I - A)^{-1}||| \leq |||(I - A)^{-1}||| \; |||A^{n+1}|||$$

Comme $\rho(A) < 1$, $A^{n+1} \to 0$ lorsque $n \to \infty$ (propriété 9.29). Par conséquent, le membre de droite de l'inégalité converge vers 0, et donc le membre de gauche aussi, *i.e.*, $B_n \to (I_q - A)^{-1}$, ou encore $\sum_{i=0}^{\infty} A^i = (I_q - A)^{-1}$.

rem:conv.An **Remarque 9.20** Le résultat précédent peut aussi s'énoncer : si $\rho(I_q - A) < 1$, alors A est inversible et $A^{-1} = \sum_{i=0}^{\infty}(I_q - A)^i$. En effet, il suffit d'appliquer la propriété 9.29 à la matrice $B = I_q - A$.

cor:inv.mat **Corollaire 9.5** *Si A et B sont deux matrices carrées de $M_q(\mathbb{R})$ telles que A est inversible et $|||A - B||| < \frac{1}{|||A^{-1}|||}$, alors B est inversible.*

Preuve : On a

$$|||A - B||| < \frac{1}{|||A^{-1}|||} \iff |||A - B||| \;\; |||A^{-1}||| < 1$$

La sous-multiplicativité de la norme implique que dans ce cas, on a aussi $|||(A-B)A^{-1}||| < 1$, *i.e.*, $|||I_q - BA^{-1}||| < 1$. La propriété précédente implique alors que BA^{-1} est inversible. Par conséquent, B l'est aussi. ■

Remarque 9.21 Le résultat du corollaire précédent établit que si A est une matrice inversible, alors toute matrice suffisamment proche[5] est également inversible. Le corollaire montre que "suffisamment proche" signifie à une distance inférieure à $\frac{1}{|||A^{-1}|||}$. 🄁

La propriété suivante s'obtient facilement et permet de prouver le théorème 9.1

pro:cont.inv.mat **Propriété 9.30** *Soit $A \in M_q(\mathbb{R})$ telle que $|||A||| < 1$. Alors $I_q + A$ est inversible et on a*

$$|||(I_q + A)^{-1}||| \leq \frac{1}{1 - |||A|||} \qquad et \qquad |||(I_q + A)^{-1} - I_q||| \leq \frac{|||A|||}{1 - |||A|||} \tag{9.7}$$

eq:cont.inv.mat

Preuve : On a $(I_q + A) = (I_q - B)$ avec $B = -A$ et $|||B||| = |||A||| < 1$, donc $\rho(B) < 1$. On peut alors appliquer la propriété 9.29, ce qui donne l'inversibilité de $(I_q + A)$. Cette même propriété donne

$$(I_q + A)^{-1} = (I_q - B)^{-1} = \sum_{i=0}^{\infty} B^i = \sum_{i=0}^{\infty}(-A)^i$$

Par conséquent

$$|||(I_q + A)^{-1}||| = |||\sum_{i=0}^{\infty}(-A)^i||| \leq \sum_{i=0}^{\infty}|||A|||^i = \frac{1}{1 + |||A|||}$$

5. La proximité étant définie au moyen de la distance induite par la norme : la distance entre A et B est $|||A - B|||$.

(où l'inégalité s'obtient par sous-additivité et sous-multiplicativité de $|||\cdot|||$). On a obtenu la première inégalité de l'énoncé. Par ailleurs, et pour les mêmes raisons que ci-dessus, on a

$$|||(I_q + A)^{-1} - I_q||| = ||| \sum_{i=1}^{\infty} (-A)^i ||| \leq \sum_{i=1}^{\infty} |||A|||^i = \sum_{i=0}^{\infty} |||A|||^i - 1 = \frac{1}{1 - |||A|||} - 1$$

ce qui est la seconde inégalité de l'énoncé. ■

On peut à présent énoncer et prouver le résultat suivant, qui est l'un des plus importants de cette section.

th:cont.inv.mat **Théorème 9.1** *Soit φ l'application définie sur l'ensemble $M_q^*(\mathbb{R})$ des matrices inversibles de $M_q(\mathbb{R})$ et à valeurs dans $M_q^*(\mathbb{R})$, définie par $A \longmapsto \varphi(A) = A^{-1}$. L'application φ est continue sur $M_q^*(\mathbb{R})$:*

$$\forall A \in M_q^*(\mathbb{R}), \quad H_n \to 0_{q,q} \implies \varphi(A + H_n) \to \varphi(A)$$

ou encore,

$$\forall A \in M_q^*(\mathbb{R}), \quad |||H_n||| \to 0 \implies |||(A + H_n)^{-1} - A^{-1}||| \to 0$$

Preuve : On montre facilement que φ est continue en I_q. Pour cela, considérons $H_n \to 0_{q,q}$. Cela signifie qu'à partir d'un certain rang, les matrices $I_q + H_n$ seront toutes inversibles. En effet, puisque $H_n \to 0_{q,q}$, $|||H_n||| = |||(I_q + H_n) - I_n||| \to 0$ et donc on peut trouver n_1 tel que pour tout $n > n_1$ on a $|||(I_q + H_n) - I_n||| < 1$. En utilisant la remarque 9.20, on déduit que $I_q + H_n$ est inversible pour tout $n > n_1$. Pour de tels n, on peut alors former $(I_q + H_n)^{-1}$ et puisque $|||H_n||| < 1$, la seconde inégalité (9.7) permet d'écrire

$$|||(I_q + H_n)^{-1} - I_q||| \leq \frac{|||H_n|||}{1 - |||H_n|||}$$

Donc lorsque $|||H_n||| \to 0$, on a aussi $|||(I_q + H_n)^{-1} - I_q||| \to 0$, ce qui équivaut à la continuité de φ en $A = I_q$. Pour $A \in M_q^*(\mathbb{R})$ quelconque, on obtient l'inversibilité de $A + H_n$ de la manière suivante. On a $H_n = (H_n + A) - A$. Comme A est inversible et que $H_n \to 0$, on peut trouver n_2 tel que

$$|||H_n||| = |||(H_n + A) - A||| < \frac{1}{|||A^{-1}|||}$$

pour tout $n > n_2$. Le corollaire 9.5 permet de conclure que $A + H_n$ est inversible pour tout $n > n_2$. Pour de tels n, on peut écrire

$$(A + H_n)^{-1} = \left[A(I_q + A^{-1}H_n)\right]^{-1} = (I_q + A^{-1}H_n)^{-1}A^{-1} = (I_q + \tilde{H}_n)^{-1}A^{-1}$$

où $\tilde{H}_n = A^{-1}H_n$. Donc

$$||| (A + H_n)^{-1} - A^{-1}||| = |||((I_q + \tilde{H}_n)^{-1} - I_q)A^{-1}||| \leq |||(I_q + \tilde{H}_n)^{-1} - I_q ||| \, |||A^{-1}||| \quad (9.8)$$

eq:cont.inv.mat

On a alors

$$|||H_n||| \to 0 \implies |||\tilde{H}_n||| \to 0 \implies |||(I_q + \tilde{H}_n)^{-1} - I_q||| \to 0 \implies |||(A + H_n)^{-1} - A^{-1}||| \to 0$$

où la première implication provient de la sous-multiplicativité de $|||\cdot|||$, la seconde provient de la continuité de φ en I_q et la dernière de l'inégalité (9.8). Cette succession d'implications établit la continuité de φ en A. ■

Ce résultat est très important puisqu'il permet de conclure que si pour une suite $\{A_n\}$ de matrices de $M_q(\mathbb{R})$, on a $A_n \to A$ où A est une matrice inversible, alors les A_n sont inversibles à partir d'un certain rang, et on a $A_n^{-1} \to A^{-1}$ (il suffit d'appliquer le théorème 9.1 avec $H_n = A_n - A$).

sec:norm.mat.subord

9.3.2 Norme subordonnée

On a noté que toutes les normes sur $M_q(\mathbb{R})$ étaient équivalentes. On présente ici des manières de construire de telles normes. Cette construction consiste à introduire des normes matricielles à partir d'une norme sur $\mathbb{R}^q$.

pro:normat

Propriété 9.31 *Soit $\|\cdot\|$ une norme sur l'espace $\mathbb{R}^q$. L'application $|||\cdot||| : M_q(\mathbb{R}) \to \mathbb{R}$ définie par*

$$|||A||| = \sup\{\frac{\|Ax\|}{\|x\|} \text{ t.q. } x \in \mathbb{R}^q, x \neq 0\}$$

est une norme sur $M_q(\mathbb{R})$. On l'appelle norme subordonnée à la norme $\|\cdot\|$ sur $\mathbb{R}^q$.

rem:normat

Remarque 9.22 On constate que $\sup\{\frac{\|Ax\|}{\|x\|} \text{ t.q. } x \in \mathbb{R}^q, x \neq 0\} = \sup\{\|Ax\| \text{ t.q. } x \in \mathbb{R}^q, \|x\| = 1\}$. Par conséquent,

$$|||A||| = \sup\{\|Ax\| \text{ t.q. } x \in \mathbb{R}^q, \|x\| = 1\}$$

ce qu'on notera $|||A||| = \sup_{\|x\|=1} \|Ax\|$.

On constate aussi que $|||A||| = \sup_{x \in B_1} f(x)$ où $B_1 = \{x \in \mathbb{R}^q, \|x\| = 1\}$ et $f(x) = \|Ax\|$. Comme B_1 est compact et que f est continue, le supremum de la définition ci-dessus est atteint pour un $x^* \in \mathbb{R}^q$ tel que $\|x^*\| = 1$. Autrement dit, $\|Ax^*\| \geq \|Ax\|$ $\forall x \in B_1$, et $\|Ax^*\| \in \{\|Ax\| \text{ t.q. } x \in \mathbb{R}^q, \|x\| = 1\}$. ☒

Preuve de la propriété 9.31 : On a évidemment la positivité de $|||A|||$. Supposons que $|||A||| = 0$. D'après la définition, $\|Ax\| = 0$ pour tout x t.q. $\|x\| = 1$, et donc $\|Ax\| = 0$ pour tout $x \in \mathbb{R}^q$. Comme $\|\cdot\|$ est une norme, on doit avoir $Ax = 0$ $\forall x \in \mathbb{R}^q$. Donc le noyau de A est $\mathbb{R}^q$ et A est de rang 0. D'où $A = 0$. Par ailleurs, pour n'importe quel $a \in \mathbb{R}$:

$$|||aA||| = \sup_{\|x\|=1} \|aAx\| = \sup_{\|x\|=1} |a|\|Ax\| = |a| \sup_{\|x\|=1} \|Ax\| = |a| \ |||A|||$$

ce qui montre l'homogénéité. Pour vérifier la sous-additivité, notons qu'en utilisant la sous-additivité sur la norme $\|\cdot\|$, on a pour tout $x \in \mathbb{R}^q$:

$$\|(A+B)x\| = \|Ax + Bx\| \leq \|Ax\| + \|Bx\|$$

Donc

$$|||A+B||| \leq \sup_{\|x\|=1} (\|Ax\| + \|Bx\|) \leq \sup_{\|x\|=1} \|Ax\| + \sup_{\|x\|=1} \|Bx\| = |||A||| + |||B|||$$

On montre enfin la sous-multiplicativité. Notons que par définition de $|||A|||$, on a $\|Ax\| \leq |||A||| \ \|x\|$ pour tout $0_q \neq x \in \mathbb{R}^q$. De plus, si $x = 0_q$, cette inégalité reste vraie. Soient donc

A et B dans $M_q(\mathbb{R})$ et $x \in \mathbb{R}^q$ tel que $\|x\| = 1$. On pose $y = Bx$. D'après ce qu'on vient de montrer, $\|y\| \leq |||B|||\ \|x\|$ et on peut alors écrire

$$\|ABx\| = \|Ay\| \leq |||A|||\ \|y\| = |||A|||\ \|Bx\| \leq |||A|||\ |||B|||\ \|x\| = |||A|||\ |||B|||$$

où la dernière égalité provient du fait qu'on a choisi x tel que $\|x\| = 1$. Puisque ces relations sont vraies quel que soit le choix d'un tel x, on a

$$|||AB||| = \sup_{\|x\|=1} \|ABx\| \leq |||A|||\ |||B|||$$

■

La norme définie ci-dessus dépend du choix de la norme sur $\mathbb{R}^q$. Cependant, toutes le normes sur $\mathbb{R}^q$ étant équivalentes, les normes subordonnées sur $M_q(\mathbb{R})$ sont également équivalentes. Du point de vue qui nous intéresse, cela signifie que si A_n converge vers A avec une norme sur $M_q(\mathbb{R})$, alors cette convergence a également lieu avec n'importe quelle autre norme sur $M_q(\mathbb{R})$.

Parmi toutes les normes subordonnées qu'on peut construire, celle obtenue lorsque la norme sur $\mathbb{R}^q$ est la norme euclidienne usuelle $\|x\| = \left(\sum_{i=1}^q x_i^2\right)^{1/2}$ joue un rôle particulier. Le résultat suivant donne une expression de cette norme.

pro:normatvp

Propriété 9.32 *La norme sur $M_q(\mathbb{R})$ subordonnée à la norme euclidienne usuelle sur $\mathbb{R}^q$ est donnée par*

$$|||A||| = \sqrt{\xi^*}$$

où ξ^* est la plus grande valeur propre de $A^\top A$.

Preuve : On note que par construction $A^\top A$ est définie positive et donc $\xi^* \geq 0$, ce qui justifie la racine carrée. Pour démontrer le résultat, il suffit de montrer que $|||A|||^2 = \xi^* = \max\{\xi_i, i = 1, \ldots, q\}$. D'après la définition de $|||A|||$, et la remarque 9.22, il existe un $x^* \in \mathbb{R}^q$ avec $\|x^*\| = 1$ tel que

$$|||A|||^2 = \|Ax^*\|^2 \geq \|Ax\|^2, \quad \forall x \in \mathbb{R}^q, \|x\| = 1$$

Ceci montre que x^* est solution du problème $\max_{x\in\mathbb{R}} \|Ax\|^2$ sous contrainte que $\|x\| = 1$. La contrainte s'écrit évidemment de manière équivalente $\|x\| = 1$. Donc on résoud $\max_{x\in\mathbb{R}} \|Ax\|^2$ s.c.q. $\|x\|^2 = 1$. En utilisant les résultats usuels d'optimisation, il doit alors exister un réel ξ^* tel que

$$\frac{\partial \mathcal{L}}{\partial x_i}(x^*, \xi^*) = 0, \qquad i = 1, \ldots, q \tag{9.9}$$

eq:lagnormat

où $\mathcal{L}(x, \xi) = \|Ax\|^2 - \xi(\|x\|^2 - 1)$ est le lagrangien associé au problème d'optimisation. On note que

$$\mathcal{L}(x, \xi) = x^\top A^\top Ax - \xi(x^\top x - 1) = \sum_{j=1}^q \sum_{k=1}^q x_j x_k b_{jk} - \xi(\sum_{j=1}^q x_j^2 - 1)$$

où les réels b_{jk} sont les entrées de la matrice $B = A^\top A$. À partir de cette expression, on calcule

$$\frac{\partial \mathcal{L}}{\partial x_i}(x, \xi) = \sum_{j=1}^q \frac{\partial}{\partial x_i}\left(x_j \sum_{k=1}^q x_k b_{jk}\right) - 2\xi x_i$$

Or

$$\frac{\partial}{\partial x_i}\left(x_j \sum_{k=1}^{q} x_k b_{jk}\right) = \begin{cases} x_j b_{ji} \text{ si } j \neq i \\ \sum_{k=1}^{q} x_k b_{ik} + x_i b_{ii} \text{ sinon} \end{cases}$$

Donc

$$\sum_{j=1}^{q} \frac{\partial}{\partial x_i}\left(x_j \sum_{k=1}^{q} x_k b_{jk}\right) = \sum_{\substack{j=1 \\ j\neq i}}^{q} x_j b_{ji} + \sum_{k=1}^{q} x_k b_{ik} + x_i b_{ii} = \sum_{j=1}^{q} x_j b_{ji} + \sum_{k=1}^{q} x_k b_{ik} = 2\sum_{k=1}^{q} x_k b_{ik}$$

où la dernière égalité provient de la symétrie de $B = A^\top A$. Par conséquent

$$\frac{\partial \mathcal{L}}{\partial x_i}(x, \xi) = 2\sum_{k=1}^{q} x_k b_{ik} - 2\xi x_i$$

Les q conditions (9.9) s'écrivent alors $2Bx^* - 2\xi^* x^* = 0_q$, ou encore $Bx^* = \xi^* x^*$. Cette égalité montre que toute solution x^* du problème d'optimisation est une valeur propre de la matrice B. Le multiplicateur de Lagrange ξ^* est la valeur propre de B associée à x^*. Par conséquent, x^* est à chercher parmi les q vecteurs propres de B. C'est celui qui donne à la fonction $x^\top Bx$ qu'on cherche à maximiser sa plus grande valeur. Comme $B = A^\top A$ est symétrique, les vecteurs propres sont orthonormés et on a pour tout couple (x, ξ) de vecteur et valeur propres de B :

$$x^\top Bx = \xi x^\top x = \xi$$

Autrement dit, pour chaque vecteur propre la fonction à maximiser est égale à la valeur propre associée. Le vecteur propre qui donne la plus grande valeur à la fonction qu'on cherche à maximiser est donc celui qui est associé à la plus grande valeur propre de B. Autrement dit, x^* est le vecteur propre associé à la valeur propre $\xi^* = \max\{\xi_1, \ldots, \xi_q\}$. Donc $|||A|||^2 = \xi^*$. ■

Remarque 9.23 Lorsque A est symétrique, alors $B = A^\top A = A^2$ et la plus grande valeur propre de $A^\top A$ est le carré de $\lambda^* = \max\{|\lambda_1|, \ldots, |\lambda_q|\}$, où $\lambda_1, \ldots, \lambda_q$ sont les valeurs propres de A. Ceci découle du fait que la i^{e} valeur propre de A^2 est λ_i^2 et que la plus grande valeur propre de A^2 est donc λ^{*2}. Donc dans le cas où A est symétrique, $|||A||| = \lambda^*$.

Si de plus A est (semi-) définie positive, alors $\lambda^* = \max\{\lambda_1, \ldots, \lambda_q\}$ et la norme de A coïncide avec sa plus grande valeur propre. R

Remarque 9.24 Par définition de $|||A|||$, on aura $\|Ax\| \leq |||A|||\, \|x\|$ pour tout $x \in \mathbb{R}^q$. Dans le cas où la norme est subordonnée à la norme euclidienne usuelle sur $\mathbb{R}^q$, alors $\|Ax\| \leq \sqrt{\xi^*}\, \|x\|$, où ξ^* est la plus grande des valeurs propres de $A^\top A$.

En examinant la preuve de la propriété 9.32, on constate que si on cherche x de manière à minimiser *minimiser* $\|Ax\|$ sous contrainte que $\|x\| = 1$, alors la solution x^{**} est le vecteur propre associé à ξ^{**}, la plus petite valeur propre de $A^\top A$, et on a $\|Ax^{**}\|^2 = \xi^{**}$. Par conséquent, $\|Ax\| \geq \sqrt{\xi^{**}}\|x\|$ pour tout $x \in \mathbb{R}^q$.

En résumé, si on note respectivement ξ^* et ξ^{**} la plus grande et la plus petite des valeurs propres de $A^\top A$, alors

$$\sqrt{\xi^{**}}\|x\| \leq \|Ax\| \leq \sqrt{\xi^*}\|x\| \qquad \forall x \in \mathbb{R}^q$$

ou de manière équivalente

$$\xi^{**} x^\top x \leq x^\top A^\top A x \leq \xi^* x^\top x \qquad \forall x \in \mathbb{R}^q \tag{9.10}$$ eq:spect

$\mathbb{R}$

sec:derivmat

9.4 Sur les dérivées de fonctions matricielles

9.4.1 Définition

Soit $\mathcal{A}$ un vecteur de $\mathbb{R}^m$ dont les coordonnées sont des fonctions réelles, toutes définies sur $\mathbb{R}^p$. $\mathcal{A}$ est donc un vecteur de fonctions. On représente $\mathcal{A}$ par

$$\begin{aligned} \mathcal{A} : \mathbb{R}^p &\longrightarrow \mathbb{R}^m \\ x &\longmapsto A(x) \end{aligned}$$

La j^{e} coordonnée de $\mathcal{A}$ est une fonction de x, notée a_j définie par

$$\begin{aligned} a_j : \mathbb{R}^p &\longrightarrow \mathbb{R} \\ x &\longmapsto a_j(x) \end{aligned}$$

On peut donc écrire

$$\mathcal{A}(x) = \begin{pmatrix} a_1(x) \\ a_2(x) \\ \vdots \\ a_m(x) \end{pmatrix} = \begin{pmatrix} a_1(x_1, \ldots, x_p) \\ a_2(x_1, \ldots, x_p) \\ \vdots \\ a_m(x_1, \ldots, x_p) \end{pmatrix}$$

Lorsque chacune des fonctions a_j est dérivable par rapport à chacune des coordonnées de x, la dérivée de $\mathcal{A}$ par rapport à x est $\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}x}$ et définie par la matrice de dimensions $(p \times m)$ dont le $(j,k)^{\text{e}}$ élément est $\frac{\partial a_k}{\partial x_j}$. Autrement dit

$$\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}x} = \begin{pmatrix} \dfrac{\partial a_1}{\partial x_1} & \dfrac{\partial a_2}{\partial x_1} & \cdots & \dfrac{\partial a_m}{\partial x_1} \\ \dfrac{\partial a_1}{\partial x_2} & \dfrac{\partial a_2}{\partial x_2} & \cdots & \dfrac{\partial a_m}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial a_1}{\partial x_p} & \dfrac{\partial a_2}{\partial x_p} & \cdots & \dfrac{\partial a_m}{\partial x_p} \end{pmatrix}.$$

En ce qui concerne les dimensions de $\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}x}$, on retiendra en particulier que cette matrice a autant de lignes que x.

9.4.2 Cas particuliers

1. $m = 1$: $\mathcal{A}$ est une fonction reélle de p variables réelles. Dans ce cas, $\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}x}$ est le vecteur de $\mathbb{R}^p$ défini par

$$\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}x} = \begin{pmatrix} \dfrac{\partial \mathcal{A}}{\partial x_1} \\ \dfrac{\partial \mathcal{A}}{\partial x_2} \\ \vdots \\ \dfrac{\partial \mathcal{A}}{\partial x_p} \end{pmatrix}$$

2. $\mathcal{A}(x) = Ax$, où A est une matrice réelle de dimensions $(m \times p)$ dont on note a_{jk} le $(j,k)^{\text{e}}$ élément. La j^{e} ligne de $\mathcal{A}(x)$ est $a_j(x) = \sum_{l=1}^{p} a_{jl}x_l$. Par conséquent,

$$\frac{\partial a_j}{\partial x_k} = \frac{\partial}{\partial x_k}\left(\sum_{l=1}^{p} a_{jl}x_l\right) = \sum_{l=1}^{p} \frac{\partial}{\partial x_k}(a_{jl}x_l) = a_{jk}.$$

Par conséquent le $(j,k)^{\text{e}}$ élément de $\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}x}$ est a_{kj}, et donc $\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}x} = A^\top$. Nous venons de démontrer le résultat suivant.

pro:alg_lin

Propriété 9.33 *Si A est une matrice de dimensions $(m \times n)$ et x un vecteur de $\mathbb{R}^p$, on a*

$$\frac{\mathrm{d}}{\mathrm{d}x}(Ax) = A^\top$$

3. $\mathcal{A}(x) = x^\top Ax$, où A est une matrice réelle de dimensions $(p \times p)$ dont on note a_{jk} le $(j,k)^{\text{e}}$ élément. On peut se restreindre aux cas où la matrice A est symétrique. En effet, on a

$$x^\top Ax = \frac{x^\top Ax}{2} + \frac{x^\top Ax}{2} = \frac{x^\top Ax}{2} + \frac{x^\top A^\top x}{2}$$

car $x^\top Ax$ étant de dimension (1×1), on a $x^\top Ax = (x^\top Ax)^\top = x^\top A^\top x$. Donc

$$x^\top Ax = x^\top \frac{(A + A^\top)}{2} x = x^\top Bx$$

où $B = \frac{A+A^\top}{2}$ est une matrice symétrique. Si A n'était pas symétrique, il suffirait d'utiliser $x^\top Bx$ au lieu de $x^\top Ax$.

On peut écrire $x^\top Ax = \sum_{j=1}^{p}\sum_{l=1}^{p} a_{jl}x_jx_l$ et donc

$$\frac{\mathrm{d}}{\mathrm{d}x}(x^\top Ax) = \begin{pmatrix} \frac{\partial}{\partial x_1}\left(\sum_{j=1}^{p}\sum_{l=1}^{p} a_{jl}x_jx_l\right) \\ \frac{\partial}{\partial x_2}\left(\sum_{j=1}^{p}\sum_{l=1}^{p} a_{jl}x_jx_l\right) \\ \vdots \\ \frac{\partial}{\partial x_p}\left(\sum_{j=1}^{p}\sum_{l=1}^{p} a_{jl}x_jx_l\right) \end{pmatrix}.$$

Si on étudie la k^{e} coordonnée de ce vecteur, on a

$$\begin{aligned}
\frac{\partial}{\partial x_k}\left(\sum_{j=1}^{p}\sum_{l=1}^{p} a_{jl}x_jx_l\right) &= \frac{\partial}{\partial x_k}\left(\sum_{\substack{j=1\\ j\neq k}}^{p}\sum_{l=1}^{p} a_{jl}x_jx_l + x_k\sum_{l=1}^{p} a_{kl}x_l\right)\\
&= \frac{\partial}{\partial x_k}\left(\sum_{\substack{j=1\\ j\neq k}}^{p}\left(x_j\sum_{l=1}^{p} a_{jl}x_l\right)\right) + \frac{\partial}{\partial x_k}\left(x_k\sum_{l=1}^{p} a_{kl}x_l\right)\\
&= \sum_{\substack{j=1\\ j\neq k}}^{p}\left(x_j\sum_{l=1}^{p}\frac{\partial}{\partial x_k}(a_{jl}x_l)\right) + x_k\sum_{l=1}^{p}\frac{\partial}{\partial x_k}(a_{kl}x_l) + \sum_{l=1}^{p} a_{kl}x_l\\
&= \sum_{\substack{j=1\\ j\neq k}}^{p} a_{jk}x_j + a_{kk}x_k + \sum_{j=1}^{p} a_{kj}x_j\\
&= \sum_{j=1}^{p} a_{jk}x_j + \sum_{j=1}^{p} a_{kj}x_j\,.
\end{aligned}$$

Comme A est symétrique, on a pour $a_{jk} = a_{kj}$ pour tout couple d'indices (j,k). Par conséquent, la dernière expression ci-dessus s'écrit aussi $2\sum_{j=1}^{p} x_ja_{kj}$ et donc la k^{e} coordonnée de $\dfrac{\mathrm{d}}{\mathrm{d}x}(x^\top Ax)$ est

$$\frac{\partial}{\partial x_k}\left(\sum_{j=1}^{p}\sum_{l=1}^{p} a_{jl}x_jx_l\right) = 2\sum_{j=1}^{p} a_{kj}x_j\,.$$

Par conséquent,

$$\frac{\mathrm{d}}{\mathrm{d}x}(x^\top Ax) = \begin{pmatrix} 2\sum_{j=1}^{p} a_{1j}x_j\\ 2\sum_{j=1}^{p} a_{2j}x_j\\ \vdots\\ 2\sum_{j=1}^{p} a_{kj}x_j\\ \vdots\\ 2\sum_{j=1}^{p} a_{pj}x_j \end{pmatrix} = 2Ax.$$

Si on résume les résultats, on a la propriété suivante.

pro:alg_quad

Propriété 9.34 *Si A est une matrice de dimensions $(p\times p)$ et x un vecteur de $\mathbb{R}^p$, alors*

(a) $\dfrac{\mathrm{d}}{\mathrm{d}x}(x^\top Ax) = 2Bx$ *où* $B = \frac{1}{2}(A + A^\top)$.

(b) Si A est symétrique, $A = B$, et $\dfrac{\mathrm{d}}{\mathrm{d}x}(x^\top Ax) = 2Ax$.

Chapitre 10

ch:rap_inf

Rappels sur la démarche de l'inférence statistique

Cette section permet de rappeler les principes de l'inférence statistique. On précise l'objet d'étude dans une démarche inférentielle et on rappelle les différentes notions de base (population, variable, échantillon, paramètre). On présente une justification (intuitive) des méthodes d'inférence adoptées (utilisation de statistiques) en insistant sur le lien qui existe entre une caractéristique donnée d'une distribution de probabilité et les propriétés d'un échantillon de variables aléatoires issues de cette distribution. Ces rappels sont faits dans un contexte univarié.

10.1 Objectif d'une démarche inférentielle et notions de base

it:X

1. On s'intéresse à une caractéristique donnée d'une population. Pour simplifier, on supposera que cette caractéristique peut se mesurer au moyen d'une variable notée X.[1]

   Exemple : (1) les salaires des employés en France ; (2) la tension de rupture de câbles d'ascenseur ; (3) la taille des enfants dans les classes de cours préparatoire dans le Nord ; (4) l'accès à internet à domicile pour les ménages français.

it:repart

2. Quelle que soit la caractéristique que mesure X, les valeurs prises par cette variable dans la population étudiée ont une certaine répartition.[2] Notamment, pour chaque nombre réel a cette répartition exprime la proportion d'individus de la population pour lesquels la variable X est inférieure ou égale à a. On peut alors définir la *fonction de répartition* de X qui, à chaque réel a associe cette proportion. On notera F_X cette fonction.

   Exemple : En reprenant les exemples du premier point, F_X décrit tour à tour (1) la répartition des salaires en France ; (2) comment la tension de rupture est répartie dans la population des câbles d'ascenseur testés ; (3) la répartition de la taille au sein de la po-

1. On rappelle qu'en statistique, la population désigne l'ensemble de tous les individus statistiques qu'il est possible de considérer. De plus, les variables servent à décrire une population en mesurant une caractéristique des individus de la population. Certaines caractéristiques peuvent être de dimension supérieure à 1 et on utilisera dans ce cas plusieurs variables simultanément.

2. Si une *variable* est une *manière* de mesurer une caractéristique des individus d'une population, les *valeurs* prises par cette variable sont les *mesures* faites en utilisant la variable.

pulation des enfants de CP dans le Nord ; (4) la répartition de la variable indiquant si un ménage a un accès internet à domicile, parmi la population des ménages français.

3. Dire qu'on s'intéresse à une caractéristique d'une population signifie qu'on s'intéresse à la façon dont sont réparties les mesures de cette caractéristique au sein de la population. Autrement dit, si cette caractéristique est mesurée par X, on s'intéresse à la fonction de répartition F_X de X.

4. Dans bien des cas, on ne s'intéresse pas à la fonction de répartition de X toute entière, mais seulement à certaines de ses propriétés.

   Exemple : (1) la *dispersion* des salaires en France ; (2) la tension *minimale* de rupture de câbles d'ascenseur ; (3) la taille *moyenne* des enfants dans les classes de cours préparatoire dans le Nord ; (4) la *proportion* des ménages français ayant un accès internet à leur domicile.

   Les exemples ci-dessus illustrent les propriétés les plus fréquemment étudiées d'une fonction de répartition : les valeurs extrêmes (la tension *minimale* de rupture), la tendance centrale (*la moyenne* des tailles, ou encore la *proportion* de ménages) et la dispersion (la *dispersion* des salaires).

5. En statistique, ces propriétés se mesurent au moyen de divers indicateurs. Par exemple, la tendance centrale de la fonction de répartition de X est une valeur autour de laquelle se regroupent les valeurs prises par X dans la population ; on mesure typiquement cette tendance centrale par la *médiane* ou l'*espérance* de X.[3]

   La dispersion, qui décrit le caractère plus ou moins regroupé des valeurs de X autour d'une tendance centrale, se mesure fréquemment par la variance de X.[4]

page 15: param

6. Un indicateur qui mesure une propriété donnée d'une fonction de répartition est appelé *paramètre* de cette fonction de répartition. La valeur d'un paramètre est un nombre réel qu'on peut calculer dès qu'on connaît la fonction de répartition.

   Exemple : Si on s'intéresse à la tendance centrale de F_X, le paramètre considéré peut être la médiane $\mathrm{Me}(X)$. La valeur du paramètre est la valeur de la médiane qu'on calcule à partir de F_X grâce à la formule $\mathrm{Me}(X) = \inf\{a \in \mathbb{R} \mid \ F_X(a) \geq \frac{1}{2}\}$.

page 15: parint

7. Avec les notions introduites ci-dessus, on se place dans une situation où on s'intéresse à une propriété de F_X, mesurée par un paramètre. Celui-ci est alors appelé *paramètre d'intérêt* et on le notera θ.

8. Si on peut observer la valeur de la variable X pour chaque individu de la population, alors on connaît la fonction de répartition F_X (voir comment au point 2). Par conséquent, d'après ce qui vient d'être dit, on peut calculer la valeur du paramètre d'intérêt (voir le point 6).

9. La possibilité d'observer la valeur de la variable pour chaque individu signifie effectuer un recensement de la population. Ceci ne peut dans bien des cas être envisagé. Les raisons pour cela sont multiples : le recensement peut être trop coûteux (notamment en temps) lorsque le

3. L'espérance de X, notée $\mathrm{E}(X)$, s'interprète comme la valeur attendue de X. Ce nombre s'interprète également comme la moyenne de X au sein de la population. La médiane de X, notée $\mathrm{Me}(X)$, est la plus petite valeur telle que la proportion d'individus dans la population pour laquelle X est supérieure à $\mathrm{Me}(X)$ est d'au moins 50%. Formellement, on a $\mathrm{Me}(X) = \inf\{a \in \mathbb{R} \mid F_X(a) \geq \frac{1}{2}\}$.

4. La variance de X, notée $\mathrm{V}(X)$ est une mesure de la distance moyenne entre les valeurs de X dans la population et l'espérance de X.

nombre d'individus dans la population est grand (par exemple lorsque la population est celle de tous les employés en France) ; le recensement peut conduire à la destruction des individus (par exemple lorsqu'on mesure à quelle tension le câble d'ascenseur a rompu), *etc.*

it:ech

10. Cette impossibilité implique aussi l'impossibilité de calculer (et donc connaître) la valeur de θ. On se contente alors, pour des raisons qui seront développées plus bas (à partir du point 17), de mesurer la caractéristique pour un *sous-ensemble* de la population. Les individus composant ce sous-ensemble peuvent être choisis de différentes manières. Nous nous contenterons de mentionner que la manière de choisir ces individus a des conséquences importantes sur les propriétés des méthodes statistiques qui seront employées par la suite.

11. La manière de choisir les individus dans une population que nous retiendrons est appelée *échantillonnage aléatoire simple*. Celle-ci procède de la manière suivante. On choisit *au hasard*[5] un premier individu dans la population et on effectue pour cet individu une mesure au moyen de la variable X. Cette mesure est une valeur notée x_1 de X. On « remet » l'individu dans la population et on répète l'étape précédente, ce qui fournit une seconde mesure x_2. En répétant cette opération n fois, on dispose de n mesures (ou observations) $x_1, \ldots, x_n$ de la variable X.

    Exemple : La variable X est la mesure de la taille d'un enfant de CP dans le Nord. On choisit au hasard un premier enfant inscrit en CP dans le Nord, on mesure sa taille et on la note x_1. On « remet » l'enfant dans l'ensemble des enfants de CP dans le Nord et on en choisit au hasard un deuxième ; on note sa taille x_2 ; *etc.*

12. De manière évidente, ces observations ne peuvent être connues avec certitude à l'avance, puisqu'elles dépendent de *qui sont* les n individus choisis dans la population. Par conséquent, $x_1, \ldots, x_n$ sont considérées comme les réalisations de variables aléatoires $X_1, \ldots, X_n$.[6]

13. Les propriétés de ces variables aléatoires sont assez faciles à déduire de la façon dont elles sont introduites.
    — X_i sert à mesurer la caractéristique du i^{e} individu choisi. Cet individu est choisi de manière tout à fait indépendante des autres. En effet, savoir que des individus $j, k, \ell, \ldots$ ont été choisis et savoir que pour ces individus on a effectué les mesures $x_j, x_k, x_\ell, \ldots$ de la caractéristique n'affecte pas la probabilité pour qu'on fasse quelque mesure particulière que ce soit chez l'individu i. Autrement dit, le fait de connaître les réalisations de $X_j, X_k, X_\ell, \ldots$ n'affecte pas la loi de probabilité de X_i. On dit que les variables aléatoires $X_1, \ldots, X_n$ sont *indépendantes*. Ceci signifie qu'il n'existe aucune liaison d'aucune sorte entre ces variables aléatoires. On peut donc étudier les propriétés de l'une d'elles en écartant les autres sans que cette étude soit affectée par cette mise à l'écart.
    — Considérons alors la i^{e} de ces variables aléatoires, X_i, et essayons de trouver sa loi de probabilité, caractérisée par sa fonction de répartition $F_{X_i}(a) = \mathrm{P}(X_i \leq a)$, $a \in \mathbb{R}$. On note d'abord que l'individu i étant choisi au hasard, les valeurs possibles pour X_i sont les

5. Choisir au hasard un objet dans un ensemble signifie ici que n'importe quel objet a la même probabilité d'être choisi que n'importe quel autre objet dans cet ensemble.

6. D'une manière un peu vague, on définit une variable aléatoire comme une grandeur pouvant varier en fonction du résultat d'une expérience aléatoire. Ici l'expérience est le choix d'un individu au hasard dans la population. La grandeur est la mesure de la caractéristique étudiée chez l'individu qui sera choisi. Cette mesure dépend clairement de *quel est* l'individu qui a été choisi au hasard, donc du résultat de l'expérience aléatoire.

mêmes que les valeurs possibles pour X. De plus, pour n'importe quel réel a, on sait qu'il y a une proportion égale à $F_X(a)$ d'individus dans la population pour lesquels la variable X est inférieure ou égale à a. Par conséquent, si on choisit un individu au hasard dans la population, disons le i^{e}, alors la probabilité pour que la mesure de la caractéristique pour cet individu soit inférieure ou égale à a est précisément égale à $F_X(a)$.[7] Formellement, on a $\mathrm{P}(X_i \leq a) = F_X(a)$. Ceci étant vrai quelque soit le choix de a, on a $F_{X_i} = F_X$. Ceci étant vrai pour tout i, on a $F_{X_1} = \cdots = F_{X_n} = F_X$. Autrement dit, les variables aléatoires $X_1, \ldots, X_n$ ont la même loi de probabilité. On dit que $X_1, \ldots, X_n$ sont *identiquement distribuées*.

14. On note que la variable X a une répartition dans la population qui est identique à celle des variables aléatoires $X_1, \ldots, X_n$. Or nous n'avons à aucun moment considéré la variable X comme aléatoire. Celle-ci a été simplement définie comme la façon de mesurer la caractéristique d'intérêt (voir le point 1 ci-dessus).

    Cependant, en utilisant le même raisonnement que celui introduit dans le point précédent, il est facile de déduire que si on considère la variable aléatoire $\tilde{X}$ désignant la mesure de la caractéristique étudiée pour un individu choisi au hasard dans la population, alors la fonction de répartition de $\tilde{X}$ est la même que celle de X. Il n'y a donc pas lieu de différencier $\tilde{X}$ et X : ces deux variables servent à mesurer la même chose et ont la même fonction de répartition.

    On comprend alors pourquoi on a $F_{X_i} = F_X$. En effet, X est la mesure de la caractéristique étudiée pour individu quelconque choisi au hasard dans la population, et X_i désigne *la même chose*, mais lorsqu'on convient que l'individu choisi au hasard désigne le i^{e} des n individus extraits de la population.

    Autrement dit, l'expérience aléatoire qui consiste à choisir au hasard un individu, qu'on appellera i, dans la population et qui permet de définir X_i est une réplique identique de l'expérience qui consiste à choisir au hasard un individu quelconque et qui permet de définir X. La fonction de répartition, et donc la loi de probabilité, de ce deux variables est par conséquent la même.

15. On rappelle que des variables aléatoires forment un *échantillon aléatoire simple* si elles sont indépendantes et identiquement distribuées. Cette définition fait apparaître $X_1, \ldots, X_n$ comme un échantillon aléatoire simple. De plus, comme la loi commune de ces n variables aléatoires est celle de X, on dit que $X_1, \ldots, X_n$ *forment un échantillon aléatoire simple de* X.

    On rappelle également qu'un *tirage dans une loi de probabilité* P est un nombre qui est la réalisation d'une variable aléatoire dont la loi est P. Par conséquent, les observations $x_1, \ldots, x_n$ étant les réalisations de variables aléatoires indépendantes ayant toutes la même loi, sont considérées comme des tirages dans cette loi. Celle-ci étant celle de X, on interprétera $x_1, \ldots, x_n$ comme n tirages indépendants dans la loi de X, ou encore dans F_X.[8]

7. Le raisonnement utilisé ici est identique à celui — bien connu — concernant des boules dans une urne : si une urne contient une proportion p de boules blanches, alors la probabilité pour qu'une boule choisie au hasard dans l'urne soit blanche est égale à p. Les raisonnements de ce type permettent de voir des proportions comme des probabilités et *vice versa*.

8. On ne fera pas de différence entre une probabilité et sa fonction de répartition. Par conséquent, une loi de probabilité désigne aussi bien l'une que l'autre.

16. **Résumé des points qui précèdent.** On ne peut faire de recensement et calculer θ. On constitue, par une méthode de sélection appelée échantillonage aléatoire simple, un n-uplet de variables aléatoires $X_1, \ldots, X_n$, indépendantes et identiquement distribuées. La loi de probabilité commune de ces variables est la même que celle de X. Donc $X_1, \ldots, X_n$ forment un échantillon aléatoire simple de X. En conséquence, les réalisations $x_1, \ldots, x_n$ de ces variables aléatoires constituent n tirages indépendants dans la loi F_X de X.

10.2 Présentation du principe de l'inférence statistique

it:inf

17. Il reste à expliquer pourquoi, devant l'impossibilité d'effectuer un recensement et de calculer la valeur du paramètre d'intérêt θ, on procède de la manière décrite ci-dessus. Plus précisément, quel est le but qu'on peut se fixer en matière de connaissance que l'on peut avoir de θ ? Que peut-on faire avec les observations de $X_1, \ldots, X_n$ pour atteindre ce but ?

    La valeur de θ ne pouvant se calculer, on peut tout au mieux l'approximer. Si on constitue un échantillon aléatoire simple de X, c'est dans ce but. Il faut donc montrer qu'il est possible, à partir de $X_1, \ldots, X_n$, de construire des méthodes permettant d'approximer la valeur inconnue du paramètre d'intérêt θ. Cet objectif est celui que se fixent (sous diverses formes) toutes les méthodes de l'*inférence statistique*.

18. Le principe de base de l'inférence statistique est qu'*un échantillon constitué au moyen de tirages dans une loi contient de l'information sur cette loi*. Reconnaissant ce principe, il est naturel de chercher des méthodes qui permettent d'extraire cette information. Toute méthode construite dans ce but est une *méthode d'inférence statistique*.

page:echant-moy

19. Il est évident que si $X_1, \ldots, X_n$ sont indépendantes et ont comme loi commune F_X, alors la loi de toute variable aléatoire s'exprimant comme une fonction T des variables $X_1, \ldots, X_n$ sera déduite de F_X.

    > *Exemple :* Soit une variable aléatoire X suivant la loi $\mathscr{B}(p)$. Soit $X_1, \ldots, X_n$ un échantillon aléatoire simple de X. Soit la fonction $T : \mathbb{R}^n \to \mathbb{R}$ définie par $T(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$. On forme la variable aléatoire $Y = T(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i$. La loi de Y sera déduite de celle de $X_1, \ldots, X_n$, c'est à dire F_X. En particulier, il est bien connu que l'espérance de Y est la même que celle de X, c'est à dire p. On sait également que $nY = \sum_{i=1}^n X_i$ suit une loi $\mathscr{B}(n; p)$. Par conséquent, les valeurs probables pour nY (resp. pour Y) se situeront autour de $n \times p$ (resp. p).

    Toute variable aléatoire s'exprimant comme une fonction de $X_1, \ldots, X_n$ seulement est appelée *statistique*. L'étude de la façon dont la loi d'une statistique dépend de celle de $X_1, \ldots, X_n$ s'appelle la théorie de l'échantillonnage. Dans une telle approche, on connaît F_X et on déduit la loi de $T(X_1, \ldots, X_n)$.

20. Le point qui précède rappelle qu'il existe un lien de F_X vers les propriétés de $X_1, \ldots, X_n$ et donc un lien de θ vers les propriétés de $X_1, \ldots, X_n$. On peut essayer d'obtenir un lien dans l'autre sens : des propriétés de $X_1, \ldots, X_n$, peut-on inférer quelque chose sur θ ?

    > *Exemple :* Dans l'exemple précédent, supposons que le paramètre d'intérêt soit $\theta = p$, dont on ignore la valeur. Supposons qu'on dispose de $n = 100$ observations de variables aléatoires $X_1, \ldots, X_{100}$ et qu'avec ces observations on ait obtenu la valeur 0,31 pour la

variable Y. Autrement dit, les observations $x_1, \ldots, x_{100}$ sont telles que $\frac{1}{100}\sum_{i=1}^{100} x_i = 0{,}31$. On ne connaît pas la valeur de θ, mais, en reprenant la remarque de l'exemple précédent, les observations nous disent que ce paramètre a certainement une valeur pour laquelle observer que Y vaut 0,31 est un évènement probable. Cela élimine donc comme valeurs plausibles de θ celles qui sont trop éloignées de 0,31.

Cet exemple illustre la démarche de l'inférence statistique, consistant à inférer de l'observation d'un échantillon des énoncés sur les propriétés de la loi dont il est issu, et en particulier sur la paramètre d'intérêt de cette loi. En termes plus généraux, cette démarche infère les propriétés (*inconnues*) d'une population à partir des propriétés (*observées*) d'une partie de la population (cette partie étant constituée des individus sélectionnés).

21. On peut illustrer le contenu de cette section par l'expérience suivante. Considérons une variable aléatoire X qui suit une loi normale ayant une espérance et une variance données. Au moyen de techniques de génération de nombres aléatoires, on effectue indépendamment 200 tirages aléatoires de nombres réels, de sorte que pour chaque tirage la probabilité pour que le nombre soit compris entre a et b est

$$\mathrm{P}(a \leq X \leq b) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x$$

pour toute paire de réels a et b ($a < b$), où μ et σ^2 sont les valeurs de l'espérance et de la variance de X que l'on s'est donné, respectivement. Les variables aléatoires $X_1, \ldots, X_{200}$, qui désignent les nombres qui seront tirés lors des 200 tirages, forment donc un échantillon aléatoire simple de X. On calcule ensuite la moyenne de ces nombres : $\frac{1}{200}\sum_{i=1}^{200} X_i$.

On effectue cette expérience en choisissant tour à tour diverses valeurs de l'espérance. Dans le tableau ci-dessous, on reproduit les résultats obtenus.

Expérience	Valeur de l'espérance	Moyenne observée
1	-10	-10.50
2	-5	-4.75
3	-1	-1.20
4	0	0.03
5	1	1.08
6	5	5.08
7	10	9.94

On constate que la valeur de la moyenne varie en fonction du choix de la valeur retenue pour l'espérance et, plus spécifiquement, que la première est proche de la seconde. Cela illustre la dépendance de la loi d'une statistique (ici la moyenne) vis à vis de la loi des variables de l'échantillon (en particulier l'espérance de cette loi).

Considérons maintenant l'expérience « réciproque » dans laquelle on ne se donne plus diverses valeurs de l'espérance μ, mais où on suppose plutôt que celle-ci est inconnue, mais qu'elle est égale à l'une des valeurs de l'ensemble $\{-10, -5, -1, 0, 1, 5, 10\}$. On se donne en revanche la valeur observée de la moyenne $\frac{1}{200}\sum_{i=1}^{200} X_i$ des nombres qui ont été tirés, et sur cette base, on essaie de « deviner » (ou estimer, en langage statistique) quelle est la valeur de l'espérance.

Au vu de l'expérience précédente, qui illustre le fait que la moyenne a tendance à être proche de l'espérance, si on observe que la moyenne vaut $-10{,}5$ on dira que la valeur de l'espérance

est sans doute égale à -10. Si la valeur observée de la moyenne est 1,08 on dira plutôt que la valeur de l'espérance est certainement 1, et ainsi de suite.

22. Cet exemple illustre la démarche de l'inférence statistique dans laquelle le problème consiste à « deviner » à partir d'observations la valeur inconnue d'un paramètre de la loi d'où sont issues ces observations. La raison pour laquelle cette démarche est basée sur l'utilisation d'observations est que celles-ci ont un comportement qui est entièrement déterminé par la loi dont elles proviennent, et que par conséquent ces observations peuvent nous restituer de l'information sur cette loi, ou plus particulièrement sur un paramètre d'intérêt de cette loi.

    Il est crucial de noter l'importance du modèle statistique dans une démarche d'inférence. Un modèle statistique (voir plus loin pour un définition un peu plus formelle) est une description des propriétés de la loi de probabilité dont sont issues les observations, ainsi qu'une description de la manière dont elles en sont issues. Une telle description permettra alors l'étude des propriétés de la statistique formée à partir des observations (par exemple, cette statistique a-t-elle tendance à prendre des valeurs proches du paramètre d'intérêt ?). Le modèle sert donc de cadre d'analyse des propriétés de statistiques destinées à approximer le paramètre d'intérêt. Si on ne se donne aucune information sur la manière dont les observations sont reliées à une loi de probabilité, il est impossible de savoir comment ces observations peuvent nous restituer de l'information à propos de cette loi, ou à propos de l'un des paramètres d'intérêt de cette loi.

10.3 Les problèmes d'inférence usuels

Il est commun de distinguer trois problèmes d'inférence. Cette section présente chacun d'eux. La notation est la suivante. Le paramètre d'intérêt est noté θ. Sa vraie valeur est un nombre réel inconnu appartenant à un ensemble Θ appelé ensemble des valeurs possibles du paramètre. θ est le paramètre d'une loi de probabilité qui sera celle d'une variable aléatoire X, et qu'on notera $F_X(\,\cdot\,;\theta)$. Pour toute valeur possible $\theta_0 \in \Theta$ du paramètre θ, le nombre $F_X(x;\theta_0)$ désigne la probabilité de l'évènement $(X \leq x)$, calculée lorsque la valeur du paramètre de la loi de X est θ_0.

Exemple : Si on s'intéresse par exemple à estimer la proportion p de ménages français ayant un accès internet à domicile, θ coïncide avec p : $\theta = p$. L'ensemble des valeurs possibles est $\Theta = [0;1]$. La loi de probabilité dont θ est le paramètre est la loi de Bernoulli $\mathscr{B}(\theta)$. Donc $F_X(x;\theta_0)$ désignera la fonction de répartition d'une variable aléatoire $X \rightsquigarrow \mathscr{B}(\theta_0)$:

$$F_X(x;\theta_0) = \mathrm{P}(X \leq x) = \begin{cases} 0, \text{ si } x < 0 \\ 1-\theta_0, \text{ si } 0 \leq x < 1 \\ 1, \text{ si } x \geq 1 \end{cases}$$

Exemple : On s'intéresse à la variable X qui mesure la variation du pouvoir d'achat d'une catégorie de français sur une période donnée. On suppose que la distribution de cette variable au sein de la population considérée suit une loi normale $\mathcal{N}(\mu,\sigma^2)$. Cela signifie qu'en choisissant un individu au hasard dans la population, la probabilité d'observer que la variation de son pouvoir d'achat est comprise entre a et b est

$$\mathrm{P}(a \leq X \leq b) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x$$

Ici, on peut considérer que le paramètre d'intérêt θ est le couple (μ, σ^2), qui paramétrise la loi $\mathcal{N}(\mu, \sigma^2)$ de la variable étudiée. L'ensemble Θ sera naturellement $\mathbb{R}\times\ ]0;+\infty[$. Pour un couple donné $\theta_0 = (\mu_0, \sigma_0^2)$ de Θ, la notation $F_X(x;\theta_0)$ désignera donc la fonction de répartition de loi $\mathcal{N}(\mu_0, \sigma_0^2)$:

$$F_X(x;\theta_0) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu_0)^2}{2\sigma_0^2}\right) \mathrm{d}t\,, \qquad x \in \mathbb{R}$$

On dispose de n mesures $X_1, \ldots, X_n$ de la variable X, obtenues par un procédé d'échantillonnage aléatoire. Dans cette présentation générale, on supposera que $X_1, \ldots, X_n$ constituent un échantillon aléatoire simple de X : $X_1, \ldots, X_n$ sont indépendantes et suivent toutes la même loi que X. On rappelle la distinction faite entre les variables aléatoires $X_1, \ldots, X_n$ et les réalisations (ou valeurs observées) de ces variables, notées $x_1, \ldots, x_n$.

Lorsqu'un problème d'inférence en formulé, la solution qu'on lui apporte est appelé *méthode d'inférence*.[9] Il est essentiel de mentionner que quel que soit le problème d'inférence considéré (voir les sections qui suivent), il existe à chaque fois plusieurs méthodes d'inférence. Par conséquent, se posera le choix de la « bonne » méthode. Pour que cette question ait un sens, il faut alors être en mesure de comparer plusieurs méthodes d'inférence disponibles pour résoudre un même problème d'inférence. Lorsque de tels moyens de comparaison sont établis, on peut alors être capable de retenir les meilleures méthodes d'inférence.

Les propriétés d'une méthode d'inférence particulière sont étudiées relativement au problème d'inférence posé. Autrement dit une même méthode d'inférence peut être bonne pour un problème et mauvaise pour un autre. Ceci pose donc la question du *cadre* dans lequel on pose le problème d'inférence et dans lequel on analyse les propriétés d'une méthode d'inférence destinée à résoudre le problème posé. Ce cadre est appelé *modèle statistique*. Un modèle statistique est défini comme l'ensemble des lois qu'on considère *a priori* possibles pour les variables $X_1, \ldots, X_n$.

page:modele

Dans le premier exemple ci-dessus, les variables constituant l'échantillon sont des variables aléatoires de Bernoulli. La i^{e} d'entre elles, X_i, indique si l'individu i possède un accès internet à domicile (dans ce cas on observera $X_i = 1$) ou non (et on observera alors $X_i = 0$). Le modèle est l'ensemble des lois pour $X_1, \ldots, X_n$ telles que ces variables sont indépendantes et identiquement distribuées, chacune d'entre elles ayant une loi de Bernoulli $\mathscr{B}(\theta)$, $\theta \in]0, 1[$. Dans le second exemple, X_i est la mesure de la variation du pouvoir d'achat du i^{e} individu de l'échantillon au cous de la période donnée. C'est *a priori* un réel quelconque. L'ensemble des loi possibles est l'ensemble des lois pour $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon un loi normale $\mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ et $\sigma^2 \in]0, \infty[$.

sec:est-ponct-intro

10.3.1 Estimation

L'objectif d'un problème d'estimation est d'approximer la valeur inconnue du paramètre θ. Cela peut se faire de deux manières :

1. on peut approximer θ par une valeur isolée dans Θ ;
2. on peut approximer le paramètre en cherchant une région de Θ choisie de sorte qu'elle

9. Ces méthodes portent des noms particuliers selon la catégorie de problème d'inférence qu'on souhaite résoudre (voir les sections qui suivent).

contienne la valeur inconnue de θ avec une probabilité élevée, et ceci quelle que soit cette valeur inconnue.

Le reste de cette section est consacré au premier type d'estimation. Le second sera abordé à la section 10.3.3.

On parle d'*estimation ponctuelle* de θ lorsque l'objectif correspond au premier des deux cas ci-dessus. Pour estimer θ, on utilise un *estimateur*. Un estimateur est une variable aléatoire T_n obtenue comme une fonction T de $X_1, \ldots, X_n$ à valeurs dans Θ et formée dans le but de fournir des approximations de θ. On a

$$\begin{aligned} T : \quad & \mathbb{R}^n \to \Theta \\ & (u_1, \ldots, u_n) \mapsto T(u_1, \ldots, u_n) \end{aligned}$$

et $T_n = T(X_1, \ldots, X_n)$.

Une approximation de θ obtenue en utilisant un estimateur est appelée *estimation* de θ. C'est une valeur $t_n \in \Theta$ obtenue à partir de l'estimateur T_n de la manière suivante : $t_n = T(x_1, \ldots, x_n)$. C'est donc la valeur prise par la variable aléatoire T_n lorsque les observations sont $x_1, \ldots, x_n$.

> *Exemple :* On reprend l'exemple de l'accès à internet à domicile. θ est le paramètre d'une loi de Bernoulli. Pour n ménages choisis au hasard, on introduit les variables aléatoires $X_1, \ldots, X_n$ de la manière suivante : $X_i = 1$ si le i^{e} ménage choisi a un accès internet à domicile et $X_i = 0$ sinon, $i = 1, \ldots, n$. La proportion θ de ménages qui dans la population ont un accès à internet à domicile peut être estimée par la proportion de ménages qui dans l'échantillon ont un accès à internet à domicile. L'estimateur T_n utilisé dans ce cas est cette proportion : $T_n = \frac{1}{n}\sum_{i=1}^n X_i$. La fonction T est donc définie par $T(u_1, \ldots, u_n) = \frac{1}{n}\sum_{i=1}^n u_i$. Si on observe $X_1 = x_1, \ \ldots, X_n = x_n$, alors l'estimation t_n obtenue à partir l'estimateur T_n est le nombre $t_n = \frac{1}{n}\sum_{i=1}^n x_i$.

Pour un même problème d'estimation (même paramètre à estimer à partir des mêmes mesures $X_1, \ldots, X_n$), on peut utiliser plusieurs estimateurs. On les compare usuellement au moyen de leur biais et de leur précision.

Ces notions sont définies au moyen de l'espérance de T_n (ou d'une fonction de T_n). Cette variable aléatoire étant une fonction de $X_1, \ldots, X_n$, (on rappelle que $T_n = T(X_1, \ldots, X_n)$) l'espérance $\mathrm{E}(T_n)$ se calcule à partir de la loi de $X_1, \ldots, X_n$. Ces variables formant un échantillon aléatoire simple de X, la loi de $X_1, \ldots, X_n$ est donnée par la loi de X. Par conséquent, $\mathrm{E}(T_n)$ se calcule à partir de la loi de X, c'est à dire $F_X(\,\cdot\,;\theta)$. Puisque celle-ci dépend de θ, il en sera de même pour $\mathrm{E}(T_n)$. Autrement dit, il existe autant de façons de calculer $\mathrm{E}(T_n)$ qu'il y a de lois *a priori* possibles pour X, et donc au moins autant de façons qu'il y a de valeurs *a priori* possibles pour θ. Pour cette raison, on voit $\mathrm{E}(T_n)$ comme une fonction de θ, et pour l'indiquer, on note cette espérance $\mathrm{E}_\theta(T_n)$. [10]

Soit T_n un estimateur de θ. On dit que T_n est un estimateur sans biais si pour toute valeur possible θ_0 du paramètre θ on a $\mathrm{E}_{\theta_0}(T_n) - \theta_0 = 0$. Si ce n'est pas le cas, l'estimateur T_n de θ est biaisé et son biais en θ_0 est $\mathrm{E}_{\theta_0}(T_n) - \theta_0$. Le biais d'un estimateur est donc sa tendance à s'écarter de la valeur du paramètre qu'il estime. Pour cette raison, on peut préférer un estimateur sans biais à un estimateur biaisé. Le biais fournit donc un moyen de comparer des estimateurs.

10. L'exemple du point 19 (page 239) illustre la dépendance de l'espérance d'une statistique envers les paramètres de la loi des variables à partir desquelles elle est formée. Dans cet exemple, les variables sont de loi de Bernoulli $\mathscr{B}(\theta)$ et la statistique $T_n = \frac{1}{n}\sum_{i=1}^n X_i$ formée à partir de ces variables a pour espérance θ, qui dépend de manière évidente de θ.

La précision d'un estimateur T_n de θ se mesure au moyen de son erreur quadratique moyenne. Celle-ci est définie comme la fonction qui à la valeur $\theta_0 \in \Theta$ associe le nombre $\mathrm{E}_{\theta_0}\left[(T_n - \theta_0)^2\right]$, où l'espérance est calculée en utilisant la loi $F_X(\,\cdot\,;\theta_0)$. L'erreur quadratique moyenne s'interprète donc comme une mesure de la distance attendue entre un estimateur et la valeur du paramètre qu'il estime. Un estimateur sera d'autant plus précis qu'il a tendance à ne pas s'écarter de la valeur du paramètre à estimer. Parmi deux estimateurs, on préfère donc en général un celui qui la plus petite erreur quadratique moyenne.

Lorsque le biais ou l'erreur quadratique moyenne ou tout autre propriété intéressante d'un estimateur sont trop complexes à calculer, on examine parfois les propriétés *asymptotiques* de T_n. Ces propriétés sont celles que l'on obtient lorsque la taille de l'échantillon est arbitrairement grande ($n \to \infty$). Grace à de puissants théorèmes (par exemple le théorème « central limit »), les propriétés limites (celles qu'on détermine lorsque $n \to \infty$) de T_n sont souvent plus faciles à calculer que les propriétés valables pour n fini, quelconque.

On dira par exemple que T_n est un estimateur *convergent* de θ si la limite en probabilité de T_n est égale à θ :

$$\mathrm{P}_{\theta_0}\left(|T_n - \theta_0| \ > \epsilon\right) \to 0, \quad n \to \infty, \qquad \forall \epsilon > 0, \forall \theta_0 \in \Theta.$$

où la notation P_{θ_0} indique que la probabilité est calculée lorsqu'on suppose que la valeur du paramètre θ est θ_0, et donc que la loi de X est $F_X(\,\cdot\,;\theta_0)$.

La raison pour laquelle on étudie les propriétés asymptotiques d'un estimateur T_n de θ peut être illustrée de la manière suivante. Si les conditions d'application d'un théorème « central limit » sont satisfaites, alors ce théorème dit typiquement que la différence entre la fonction de répartition de T_n et la fonction de répartition d'une loi connue. On dit que cette loi connue est la « loi limite » de T_n. Dans beaucoup de cas, cette loi limite est une loi normale. Par conséquent, pourvu que n soit suffisamment grand, la différence entre les deux fonctions de répartition est aussi petite qu'on veut. Dans ce cas, sous des conditions appropriées (qu'il faut prendre soin d'établir), les propriétés qu'on obtient en utilisant l'une des deux fonctions de répartition est aussi proche qu'on le veut des mêmes propriétés obtenues en utilisant l'autre fonction de répartition. Ainsi, supposons qu'on s'intéresse au biais de l'estimateur T_n. Selon la définition ci-dessus, si la valeur du paramètre θ est θ_0, le biais vaut $\mathrm{E}_{\theta_0}(T_n) - \theta_0$. Plaçons-nous dans le cas où on ne sait pas calculer $\mathrm{E}_{\theta_0}(T_n)$. Sous les conditions décrites dans ce paragraphe, pourvu que n soit suffisamment grand, la différence entre cette espérance et l'espérance de T_n calculée en utilisant la loi limite est aussi petite qu'on le souhaite. Or il est souvent établi que l'espérance calculée à partir de la loi limite est égale à θ_0. Par conséquent, si n est suffisamment grand, $\mathrm{E}_{\theta_0}(T_n)$ est aussi proche que l'on veut de θ_0, ou encore, le biais $\mathrm{E}_{\theta_0}(T_n) - \theta_0$ est aussi petit que l'on veut.

En conclusion, sous des conditions apropriées qui permettent d'utiliser des résultats tels que le théorème « central limit », si n est suffisamment grand, calculer les propriétés asymptotiques d'un estimateur revient quasiment à calculer les véritables propriétés de cet estimateur. La difficulté de cette approche réside dans le fait qu'il est difficile (et même souvent impossible) d'établir que la taille n de l'échantillon dont on dispose est effectivement suffisamment grande pour que l'approximation faite en utilisant une loi limite est satisfaisante.

sec:test

10.3.2 Test d'hypothèse

10.3.2.1 Problème de test

Un problème de test est un problème dans lequel il faut décider parmi deux hypothèses mutuellement exclusives, chacune concernant la valeur θ du paramètre d'intérêt, celle qu'on considère comme étant vraie. Ces hypothèses sont notées H_0 et H_1 et appelées respectivement hypothèse nulle et hypothèse alternative.

Exemple : Si θ désigne une proportion, on peut avoir avoir à choisir entre les hypothèses H_0 : $\theta \leq \frac{1}{2}$ et $H_1 : \theta > \frac{1}{2}$ celle qu'on considère comme vraie.

Quelle que soit l'hypothèse considérée comme vraie, on suppose qu'il y en a toujours une (et une seule) qui est vraie en réalité, c'est à dire qui est compatible avec la vraie valeur θ.

Résoudre un problème de test se dit *tester* H_0 *contre* H_1.

sec:rap_test

10.3.2.2 Test statistique

Le procédé par lequel on choisit entre H_0 et H_1 est appelé *test* (ou encore règle de décision, ou règle de classification). Un *test statistique* est un test dans lequel la décision est prise sur la base de l'observation d'un échantillon $X_1, \ldots, X_n$ (de manière à utiliser l'information que l'échantillon apporte à propos du paramètre et donc à propos des hypothèses formulées). Les décisions possibles sont « *on décide que* H_0 *est vraie* » d'une part, et « *on décide que* H_1 *est vraie* », d'autre part. On peut alors définir formellement un test comme une application φ définie sur $\mathbb{R}^n$ et à valeur dans $\{0, 1\}$ qui indique la décision prise en fonction de l'échantillon obtenu. Plus précisément, cette fonction est définie ainsi : $\varphi(X_1, \ldots, X_n) = k$ si et seulement si sur la base de l'échantillon $X_1, \ldots, X_n$ on décide que H_k est vraie, $k = 0, 1$. La *variable aléatoire* $\varphi_n = \varphi(X_1, \ldots, X_n)$ s'interprète comme la *règle* utilisée pour prendre une décision sur la base de l'échantillon $X_1, \ldots, X_n$. Si on a observé $X_1 = x_1, \ldots, X_n = x_n$, la *décision* prise au moyen du test φ est l'élément de $\{0, 1\}$ qu'on note $\varphi(x_1, \ldots, x_n)$.

Un test statistique permet de prendre une décision à propos de la vraie valeur de θ en utilisant certaines propriétés de l'échantillon. Pour construire un test, on cherche en général s'il existe une propriété de l'échantillon qui change — si possible fortement — selon qu'on considère H_0 ou bien H_1 comme étant vraie. Cette propriété est mesurée par une statistique T_n formée à partir des variables composant l'échantillon : $T_n = T(X_1, \ldots, X_n)$.

Exemple : Le paramètre θ est la proportion de ménages français disposant d'un accès internet à domicile. Autrement dit, si on mesure cette caractéristique d'un ménage choisi au hasard par X, on notera $X = 1$ l'évènement qui se réalise si le ménage dispose d'un accès internet à son domicile et $X = 0$ son contraire. On a évidemment $X \sim \mathscr{B}(\theta)$. Admettons que l'on veuille tester $H_0 : \theta \leq \frac{1}{2}$ contre $H_1 : \theta > \frac{1}{2}$ et que pour cela on dispose d'un échantillon aléatoire simple $X_1, \ldots, X_n$ de X.

Les hypothèses portent sur une proportion d'individus dans la population. On sait que cette proportion et la même proportion calculée dans l'échantillon ont tendance à être semblables. Autrement dit, une propriété de l'échantillon qui changera selon que H_0 est vraie ou pas est la proportion de variables $X_1, \ldots, X_n$ dans l'échantillon qui prendront la valeur 1. En effet, si H_0 est vraie, c'est à dire si $\theta \leq \frac{1}{2}$, il est probable que cette proportion sera elle même proche de plus petite que $\frac{1}{2}$; si au contraire H_0 est fausse, alors il sera probable d'observer une valeur plus

grande que $\frac{1}{2}$ pour cette même proportion. On choisit donc de mesurer cette propriété au moyen de la statistique $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, c'est à dire de la proportion d'individus de l'échantillon ayant un accès à internet à domicile.

Ainsi, on pourra partitionner l'ensemble, noté $\mathbb{T} \subseteq \mathbb{R}$, des valeurs possibles de la variable aléatoire T_n de la façon suivante : $\mathbb{T} = \mathscr{T} \cup \overline{\mathscr{T}}$, où $\overline{\mathscr{T}} = \mathbb{T} \setminus \mathscr{T}$. L'ensemble $\mathscr{T}$ est l'ensemble des valeurs les plus vraisemblables de T_n lorsque H_1 est vraie et l'ensemble $\overline{\mathscr{T}}$ est donc l'ensemble des valeurs les plus vraisemblables de T_n lorsque H_0 est vraie. Dans cette présentation, on voit qu'il est donc souhaitable de faire dépendre le choix de T_n et de $\mathscr{T}$ des hypothèses H_0 et H_1 posées.

Le principe d'un test consiste alors à comparer le comportement probable de T_n en supposant successivement que H_0 est vraie, puis que H_1 est vraie, avec le comportement observé de T_n. On décidera que H_1 est vraie si le comportement observé de T_n est plus proche de celui qui est probable lorsqu'on suppose que H_1 est vraie, que du comportement de T_n qui est probable lorsque H_0 est supposée vraie. Un test φ sera donc de la forme $\varphi(X_1, \ldots, X_n) = 1$ si et seulement si $T_n \in \mathscr{T}$, et $\varphi(X_1, \ldots, X_n) = 0$ sinon. L'ensemble $\mathscr{T}$ est donc l'ensemble des valeurs de T_n conduisant à une acceptation de H_1 et donc à un rejet de H_0 par le test φ. On appelle $\mathscr{T}$ la *région critique* du test φ.

Exemple : Dans l'exemple précédent, puisqu'il est vraisemblable d'observer une valeur de T_n grande par rapport à $\frac{1}{2}$ lorsque H_1 est vraie, l'ensemble $\mathscr{T}$ peut être choisi de la forme $\mathscr{T} =]\frac{1}{2} + d; 1]$ pour un certain réel positif d. Le test correspondant sera donc de la forme $\varphi(X_1, \ldots, X_n) = 1$ ssi $1 \geq T_n > \frac{1}{2} + d$. Si cet évènement est observé, la valeur atteinte par T_n est une valeur trop peu vraisemblable au regard de ce à quoi on s'attendrait si H_0 était vraie. On décide alors qu'elle ne l'est pas, mais que c'est H_1 la vraie hypothèse.

À ce point, la question est : quand considère-t-on qu'une valeur donnée de T_n est vraisemblable lorsque H_1 (ou H_0) est vraie ? Le critère qui permet de répondre à cette question est basé sur un calcul de risques.

sec:risques

10.3.2.3 Calcul des risques

Un test conduit forcément à prendre l'une des deux décisions suivantes : on considère que H_0 est vraie, ou bien on considère que H_1 est vraie. Évidemment, l'hypothèse considérée comme vraie à l'issue du test ne l'est pas forcément, c'est à dire n'est pas forcément vérifiée par la vraie valeur du paramètre. Autrement dit, il est possible de prendre une mauvaise décision. Deux types d'erreur amenant à une mauvaise décision sont possibles :

— l'erreur de type 1 : décider de considérer que H_1 est vraie alors que H_0 est la vraie hypothèse ;

— l'erreur de type 2 : décider de considérer que H_0 est vraie alors que H_1 est la vraie hypothèse.

Pour savoir si on a commis une erreur, il faut comparer sa décision avec la réalité. Cela exige de connaître cette dernière et par conséquent de savoir quelle est l'hypothèse vérifiée par la vraie valeur θ du paramètre. Or cette dernière est inconnue et par conséquent on ne peut pas savoir sans ambiguïté quelle est, parmi H_0 et H_1, l'hypothèse qui est vraie. Par conséquent, quel que soit le test utilisé, quel que soit le choix de $\mathscr{T}$ et quelle que soit la décision prise, on ne peut jamais savoir si cette décision prise conduit à une erreur.

On peut en revanche calculer la probabilité de commettre une erreur en envisageant tour à tour que H_0, puis H_1, est vraie.

1. Supposons que H_0 est vraie. Il y aura dans ce cas erreur (de type 1) si la décision consiste à considérer que H_1 est vraie, autrement dit si l'évènement $T_n \in \mathscr{T}$ se réalise. Par conséquent, la probabilité de commettre une erreur de type 1 est la probabilité $\mathrm{P}(T_n \in \mathscr{T})$, calculée en supposant que H_0 est vraie. On appelle cette probabilité risque de type 1 (ou RT1).
2. Supposons que H_1 est vraie. Par le même argument que dans le point précédent, la probabilité de commettre une erreur de type 2 est $\mathrm{P}(T_n \notin \mathscr{T})$, calculée en supposant H_1 vraie. On appelle cette probabilité risque de type 2 (ou RT2).

Notons que ces risques sont des *nombres*, compris entre 0 et 1 (ce sont des probabilités).

Remarquons aussi que RT1 $\neq 1 -$ RT2, bien que les évènements $(T_n \in \mathscr{T})$ (qui sert à définir le RT1) et $(T_n \notin \mathscr{T})$ (qui sert à définir le RT2) soient contraires. Cela provient du fait que la probabilité de l'un n'est pas calculée sous les mêmes conditions que la probabilité de l'autre. Dans le calcul de $\mathrm{P}(T_n \in \mathscr{T})$ on suppose que H_0 est vraie, alors que pour calculer $\mathrm{P}(T_n \in \overline{\mathscr{T}})$, on suppose H_1 vraie. On traduit souvent cela en notant P_{H_1} ou P_{H_0} pour indiquer si un calcul de probabilité se fait en supposant H_1 ou H_0 vraie, respectivement. Ainsi, on peut écrire le RT1 comme $\mathrm{P}_{H_0}(T \in \mathscr{T})$. On voit ainsi aisément que même si $\mathrm{P}_{H_0}(T \in \mathscr{T}) = 1 - \mathrm{P}_{H_0}(T \notin \mathscr{T})$, en général on a $\mathrm{P}_{H_0}(T \in \mathscr{T}) \neq 1 - \mathrm{P}_{H_1}(T \notin \mathscr{T})$, c'est-à-dire RT1 $\neq 1 -$ RT2.

sec:comp_tests

10.3.2.4 Comparaison de tests. Choix d'un test

Pour un problème de test donné, il existe évidemment plusieurs tests possibles. Par exemple, on peut construire un test basé sur une statistique T_n et une région $\mathscr{T}$ de valeurs possibles pour T_n, qui décide donc H_1 si $T_n \in \mathscr{T}$. Il existe autant de tests de cette forme qu'il existe de choix possibles pour $\mathscr{T}$. Par exemple, si $\mathscr{T}$ est un intervalle de la forme $]t^*, \infty[$, il y a autant d'intervalles possibles qu'il y a de choix possibles pour t^*. Par ailleurs, on peut également construire des tests basés sur une statistique S_n autre que T_n et utiliser un test qui décide H_1 si $S_n \in \mathscr{S}$ pour une certaine région $\mathscr{S}$.

Pour choisir entre deux tests, on se base sur leurs probabilités de commettre des erreurs. Deux tests différents n'auront en général pas les mêmes risques. Comme des risques sont des probabilités de se tromper (de prendre de mauvaises décisions), des deux tests on préfèrera celui dont les RT1 et RT2 sont les plus petits. De manière plus générale, pour un problème de test donné, on sera tenté de choisir parmi tous les tests possibles celui dont les risques de type 1 et de type 2 sont plus petits que ceux de n'importe quel autre test. Une telle approche butte sur la non-existence d'un tel test dans les situations qui présentent un intérêt. Pour se rendre compte de cette non-existence, on peut considérer l'exemple suivant.

> *Exemple :* On reprend l'exemple précédent dans lequel on s'intéresse à la valeur de la proportion θ de ménages français ayant un accès à internet à domicile, à propos de laquelle on souhaite tester $H_0 : \theta \leq \frac{1}{2}$ contre $H_1 : \theta > \frac{1}{2}$. Puisque $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, la proportion de ménages de l'échantillon ayant un accès à internet à domicile, a tendance à être proche de θ, on s'attend à ce que T_n prenne une valeur significativement plus grande que $\frac{1}{2}$ si H_1 est vraie. Par conséquent, on peut choisir un ensemble $\mathscr{T}$ (ensemble de valeurs probables pour T_n lorsque H_1 est vraie) de la forme $\mathscr{T} =]\frac{1}{2} + d, 1]$, où d est un nombre que l'on se donne. Ainsi, le test est $\varphi_{T,\mathscr{T}} = 1 \iff T_n > \frac{1}{2} + d$; il consiste à rejeter l'hypothèse que la proportion des ménages français ayant internet est plus petite que $\frac{1}{2}$ lorsque cette proportion, observée sur l'échantillon, est significativement

plus grande que $\frac{1}{2}$. Pour choisir d, il est raisonnable de retenir la valeur pour laquelle les risques du test correspondant seront les plus petits possibles. Pour un choix de d donné, le RT1 est $\mathrm{P}_{H_0}(T_n > \frac{1}{2} + d)$ et le RT2 est $\mathrm{P}_{H_1}(T_n \leq \frac{1}{2} + d)$. Il est facile de voir que pour diminuer le RT1 il faut choisir de grandes valeurs de d, alors que pour diminuer le RT2, il faut choisir de petites valeurs de d. Il est donc impossible de choisir d de manière à minimiser simultanément les deux risques. Autrement dit, soient d_1 et d_2 deux réels tels que $0 < d_1 < d_2 < \frac{1}{2}$. On considère le test dans lequel on rejette H_0 lorsque $T_n > \frac{1}{2} + d_1$ ainsi que le test dans lequel on rejette H_0 lorsque $T_n > \frac{1}{2} + d_2$. Le RT1 du premier test sera plus grand que celui du second, alors que son RT2 sera plus petit.

Cet exemple montre que dans un cas particulier, si on s'intéresse à des tests basés sur une statistique donnée (la statistique T_n mesurant la proportion dans l'échantillon), ayant une forme donnée (le test rejette H_0 lorsque la statistique prend une valeur plus grande qu'un certain seuil) alors il n'existe pas de test ayant des risques plus petits que les autres tests. Un résultat plus général, donné par le théorème 10.1 ci-dessous, établit que pour pouvoir minimiser simultanément les deux risques d'un test, alors il faut qu'il existe un évènement qui soit à la fois quasiment certain (de probabilité 1) lorsque H_0 est vraie et quasiment impossible (de probabilité nulle) lorsque H_1 est vraie. Dans un tel cas, il suffit de noter si l'évènement a été observé. S'il l'a été, il est incohérent de supposer que H_1 puisse être vraie, puisque si c'était le cas, cet évènement serait impossible à observer ; or il l'a été et on décide donc que H_0 est vraie. On a un raisonnement semblable lorsque cet évènement n'est pas observé. Pour qu'un tel évènement existe, il faut que les hypothèses définissant le problème de test attribuent chacune à la loi des variables $X_1, \ldots, X_n$ des propriétés tellement dissemblables de l'autre, pour qu'on puisse décider « à coup sûr » (*i.e.*, avec une probabilité nulle de ce tromper) celle qui est vraie.[11] Des problèmes dans lesquels on serait amené à formuler des hypothèses aussi dissemblables sont sans intérêt pratique.

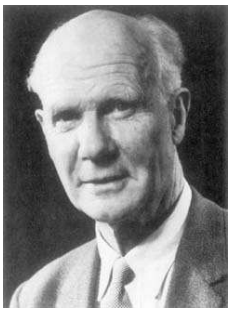
th:cs_zero_risque

Théorème 10.1 *Soit un problème de test défini par une hypothèse nulle H_0 et une hypothèse alternative H_1. Une condition nécessaire pour qu'il existe un test ayant des RT1 et RT2 inférieurs à ceux de n'importe quelle autre test est qu'il existe un évènement A dont la probabilité vaut 1 lorsque H_0 est supposée vraie, et 0 lorsque H_1 est supposée vraie.*

Preuve : Supposons qu'un tel test existe et notons-le φ^*. Son RT1 est $\mathrm{P}_{H_0}(\varphi^*(X_1, \ldots, X_n) = 1)$. Puisque son RT1 est inférieur à celui de n'importe quel autre test, il est en particulier inférieur au test noté φ_0 et qui consiste à toujours accepter H_0, *i.e.*, $\varphi_0(X_1, \ldots, X_n) = 0$, $\forall X_1, \ldots, X_n$. Or puisqu'il ne rejette jamais H_0, le RT1 de φ_0 est nul (φ_0 valant toujours 0, la probabilité que φ_0 vaille 1 est nulle). Le RT1 de φ^* vaut 0, puisque c'est un nombre positif qui doit être inférieur au RT1 de φ_0. Donc $\mathrm{P}_{H_0}(\varphi^*(X_1, \ldots, X_n) = 1) = 0$, ou encore $\mathrm{P}_{H_0}(\varphi^*(X_1, \ldots, X_n) = 0) = 1$. On introduit à présent le test φ_1 qui consiste à toujours rejeter H_0, *i.e.*, $\varphi_1(X_1, \ldots, X_n) = 1$, $\forall X_1, \ldots, X_n$. Le RT2 de φ_1 est nul et par un raisonnement semblable au précédent, on doit avoir $\mathrm{P}_{H_1}(\varphi^*(X_1, \ldots, X_n) = 0) = 0$. Définissons alors l'évènement A comme étant « Les observations sont telles qu'on accepte H_0 avec le test φ^* », ou encore $A = \{X_1, \ldots, X_n \mid \varphi^*(X_1, \ldots, X_n) = 0\}$. D'après ce qui précède on voit que la probabilité de A calculée en supposant H_0 vraie vaut 1, tandis que lorsqu'elle est calculée en supposant H_1 vraie, elle vaut 0.

11. Il apparaît dans la preuve du théorème 10.1 que les tests minimisant les deux risques sont nécessairement des tests pour lesquels ces risques sont nuls.


Jerzy NEYMAN (1894-1981)

Egon PEARSON (1895-1980)

FIGURE 10.1: J. NEYMAN et E. PEARSON

cap:Neyman-Pearson

L'approche usuelle utilisée pour lever cette indétermination qui porte sur le choix d'un test à partir des RT1 et RT2 a été proposée par J. NEYMAN et E. PEARSON (voir figure 10.1). Cette approche consiste à s'assurer que pour un problème de test donné et pour tout test envisagé pour le résoudre, le RT1 de ce test ne dépasse pas une certaine valeur, notée α et appelée *niveau* du test. La contrainte qui impose que le RT1 d'un test φ ne dépasse pas le niveau α s'écrit

page:np

$$\mathrm{P}_{H_0}(\varphi_n = 1) \leq \alpha \tag{10.1}$$

eq:np

Si le test φ est construit à partir d'une statistique T_n et prend la forme $\varphi_n = 1 \iff T_n \in \mathscr{T}$, l'inégalité (10.1) est $\mathrm{P}_{H_0}(T_n \in \mathscr{T}) \leq \alpha$. Pour un niveau α, tout test satisfaisant l'inégalité ci-dessus est appelé *test de niveau* α. Puisque pour tout problème de test d'une hypothèse H_0 contre une hypothèse H_1 on doit se donner une valeur pour α et utiliser un test de niveau α, on dit qu'*on teste* H_0 *contre* H_1 *au niveau* α.

On voit alors par exemple que si le niveau α est fixé, pour des tests de la forme $\varphi_n = 1 \iff T_n \in \mathscr{T}$, il existe des choix de $\mathscr{T}$ qui ne sont pas autorisés car ils conduiraient à une violation de la contrainte imposée par l'inégalité (10.1). On note que dans l'approche de Neyman-Pearson, on ne choisit pas directement $\mathscr{T}$, mais on fixe α *d'abord*, et *ensuite* on choisit $\mathscr{T}$ de manière que (10.1) soit satisfaite.

Le choix de α reste arbitraire. Il convient cependant de noter que d'après la contrainte (10.1), α représente la valeur maximale que le RT1 ne doit pas dépasser. Un risque étant une probabilité

de se tromper, on souhaite en général que cette probabilité ne soit pas trop élevée. Aussi dans la pratique courante des tests, on retient pour α les valeurs « standard » 0,1, 0,05 ou 0,01.

Exemple : On reprend l'exemple précédent dans lequel on s'intéresse à la valeur de la proportion de ménages français ayant un accès à internet à domicile. En suivant ce qui a été dit, on utilisera pour tester au niveau $\alpha = 0,05$ $H_0 : \theta = \frac{1}{2}$ contre $H_1 : \theta > \frac{1}{2}$ un test de la forme $\varphi(X_1, \ldots, X_n) = 1 \iff T_n \in]\frac{1}{2} + d; 1]$. La région critique $\mathscr{T}$ est de la forme $]\frac{1}{2} + d; 1]$ et choisir un test pour décider entre H_0 et H_1 revient à choisir la valeur de d. Si on veut que la contrainte (10.1) portant sur le niveau soit satisfaite, il faut que l'on ait

$$\mathrm{P}_{H_0}(1 \geq T_n > d + \tfrac{1}{2}) \leq 0,05 \qquad (10.2)$$

eq:npe

ou encore $\mathrm{P}_{H_0}(T_n > d + \frac{1}{2}) \leq 0,05$ puisque, T_n étant une proportion, on a toujours $T_n \leq 1$. On rappelle que la notation P_{H_0} indique que le calcul de probabilité doit se faire en supposant H_0 vraie, autrement dit en supposant que $\theta = \frac{1}{2}$, ou encore que $X \sim \mathscr{B}(\frac{1}{2})$. Avec cette supposition, la loi commune de $X_1, \ldots, X_n$ est évidemment $\mathscr{B}(\theta)$ et il est facile d'en déduire la loi de $T_n = \frac{1}{n}\sum_{i=1}^n X_i$, qui nous permettra de calculer $\mathrm{P}_{H_0}(T_n > d + \frac{1}{2})$ pour n'importe quelle valeur de d. En effet

$$\mathrm{P}_{H_0}(T_n > d + \tfrac{1}{2}) = \mathrm{P}_{H_0}(\tfrac{1}{n}\sum_{i=1}^n X_i > d + \tfrac{1}{2}) = \mathrm{P}_{H_0}(\sum_{i=1}^n X_i > nd + \tfrac{n}{2})$$
$$= 1 - \mathrm{P}_{H_0}(\sum_{i=1}^n X_i \leq nd + \tfrac{n}{2})$$

On sait que si H_0 est vraie, $X_1, \ldots, X_n$ sont iid $\mathscr{B}(\frac{1}{2})$ et donc $\sum_{i=1}^n X_i \sim \mathscr{B}(n; \frac{1}{2})$. Ainsi la probabilité $\mathrm{P}_{H_0}(\sum_{i=1}^n X_i \leq nd + \frac{n}{2})$ est égale à la fonction de répartition de la loi binômiale de paramètres n et $\frac{1}{2}$, évaluée en $nd + \frac{n}{2}$. Cette fonction est parfaitement connue et cette probabilité peut donc être calculée pour n'importe quelle valeur de d. Notons $b_{(n,\frac{1}{2})}$ cette fonction. Dans ce cas, pour que l'inégalité (10.2) soit satisfaite, il faut que

$$1 - b_{(n,\frac{1}{2})}(nd + \tfrac{n}{2}) \leq 0,05$$

ou encore $b_{(n,\frac{1}{2})}(nd + \frac{n}{2}) \geq 0,95$. Il faut donc choisir d de manière que $nd + \frac{n}{2}$ soit supérieur ou égal au quantile d'ordre 95% de la loi $\mathscr{B}(n, \frac{1}{2})$. On constate donc que toutes les valeurs de d ne sont pas autorisées si on veut qu'un test de la forme retenue ait un niveau 0,05.

Cependant, même si la contrainte (10.2) sur le RT1 du test exclut certaines valeurs de d, elle ne permet pas d'en déterminer une de manière unique. Pour cela, il faudra prolonger l'approche en considérant le RT2 : parmi toutes les valeurs de d pour lesquelles la contrainte (10.2) est satisfaite, on choisira celle pour laquelle le RT2 est le plus faible.

page:NP

Pour un problème de test donné, la comparaison de deux tests ne peut se faire que si ces deux tests satisfont la même contrainte (10.1) sur leurs RT1. Autrement dit, dans l'approche de Neyman-Pearson, on ne peut comparer deux tests s'ils n'ont pas le même niveau. Parmi deux tests de même niveau, on préfèrera celui pour lequel le RT2 est systématiquement le plus faible. Autrement dit, si φ^* et φ sont deux tests pour lesquels $\mathrm{P}_{H_0}(\varphi_n^* = 1) \leq \alpha$ et $\mathrm{P}_{H_0}(\varphi_n = 1) \leq \alpha$, on préfère φ^* à φ si $\mathrm{P}_{H_1}(\varphi_n^* = 0) \leq \mathrm{P}_{H_1}(\varphi_n = 0)$. Cette inégalité s'écrit aussi $\mathrm{P}_{H_1}(\varphi_n^* = 1) \geq \mathrm{P}_{H_1}(\varphi_n = 1)$ et on voit que lorsque φ^* et φ sont deux tests de même niveau, on préfère φ^* à φ si la probabilité de prendre une bonne décision lorsque H_1 est vraie est plus élevée avec le test φ^* qu'avec le test φ. Pour tout test φ, on appelle *puissance* la probabilité de décider H_1 calculée en supposant que H_1 est vraie. Par conséquent, parmi deux tests de niveau α, on préfère celui ayant la plus puissance la plus grande, ou encore, le test le plus puissant.

Dans l'approche de Neyman-Pearson, on est donc conduit à imposer sur l'ensemble des tests considérés la contrainte de niveau (10.1), et dans l'ensemble des tests satisfaisant cette contrainte, on cherche celui/ceux qui a/ont le RT2 le plus petit possible. S'il existe un test de niveau α ayant un RT2 qui n'est jamais strictement plus élevé que celui de n'importe quel autre test de niveau α, on dit qu'il est *uniformément plus puissant* (UPP) au niveau α. La recherche de test UPP pour un problème de test posé est la recherche d'un instrument de résolution du problème dont les propriétés sont optimales. L'optimalité est dans ce cas définie par le niveau minimal du RT2, compte-tenu de la borne supérieure (la valeur α) imposée au RT1.

Pour les problèmes de tests qui sont fréquement posés, il n'existe pas de test UPP. Le problème du choix du meilleur (en termes de risques) test est à nouveau posé. Pour pouvoir y répondre, on restreint la famille des tests au sein de laquelle on cherche le meilleur. Les restrictions se font en imposant aux tests qui seront éligibles de satisfaire un certain nombre de bonnes propriétés. Parmi celles-ci, la condition d'absence de biais est souvent imposée. Cette notion est définie de la manière suivante.

def:test_ss_biais **Définition 10.1** *Soit φ un test pour tester H_0 contre H_1. On dit que φ est sans biais si*

$$\mathrm{P}_{H_1}(\varphi_n = 1) \geq \mathrm{P}_{H_0}(\varphi_n = 1) \tag{10.3}$$ eq:test_ss_bia

L'inégalité définissant l'absence de biais revient à dire qu'il est plus probable de décider H_1 lorsqu'elle est vraie que lorsqu'elle est fausse. Notons également que cette même égalité s'écrit aussi

$$\mathrm{P}_{H_0}(\varphi_n = 0) \geq \mathrm{P}_{H_1}(\varphi_n = 0) \tag{10.4}$$ eq:test_ss_bia

(en utilisant le fait que $\mathrm{P}_{H_k}(\varphi_n = 0) = 1 - \mathrm{P}_{H_k}(\varphi_n = 1)$, $k = 0, 1$). L'absence de biais signifie donc également qu'il est plus probable de décider H_0 lorsqu'elle est vraie que lorsqu'elle est fausse. Autrement dit, quelle que soit la décision à laquelle on s'intéresse (décider H_0 ou décider H_1), la probabilité de prendre cette décision est toujours plus grande lorsque cette décision correspond à la bonne décision. Ceci montre que l'absence de biais pour un test est une propriété souhaitable puisqu'elle revient à imposer qu'il est plus probable de prendre une bonne décision qu'une mauvaise.

On peut alors, dans les cas où il n'existe pas de test UPP, chercher le test le plus puissant parmi tous les tests sans biais. Dans cette approche, la contrainte de niveau du test continue d'être imposée. Plus précisément, on ne considère dans un tel cas que les tests sans biais de niveau niveau α, c'est à dire les tests pour lesquels les inégalités (10.1) et (10.3) sont satisfaites. Dans l'ensemble de ces tests, on cherche celui dont le risque de type 2 est le plus faible (ou, de manière équivalente, la puissance est la plus élevée). Un moyen d'obtenir que les deux inégalités soient satisfaites consiste à imposer la contrainte

$$\mathrm{P}_{H_1}(\varphi_n = 1) \geq \alpha \geq \mathrm{P}_{H_0}(\varphi_n = 1) \tag{10.5}$$ eq:niv_et_ss_b

sec:est-ens-intro ### 10.3.3 Estimation par région de confiance

Plutôt que d'approximer θ par une seule valeur (l'estimation ponctuelle de θ) comme on l'a fait à la section 10.3.1, on peut vouloir construire, en utilisant les données de l'échantillon, une partie (ou une région) de Θ ayant une grande probabilité de contenir la valeur inconnue du paramètre θ.

Dans le cas où θ est unidimensionnel ($\Theta \subseteq \mathbb{R}$), on recherche souvent cette partie sous la forme d'un intervalle. L'objectif dans ce cas est de trouver une fourchette de valeurs ayant de grandes chances d'encadrer la valeur inconnue du paramètre θ.

On cherche donc ici un outil qui permet d'approximer la valeur inconnue d'un paramètre par un ensemble de valeurs plausibles, par opposition aux estimateurs, qui réalisent une approximation de ce même paramètre à l'aide d'une valeur isolée. Cette distinction est parfois soulignée en parlant d'*estimateur ponctuel* pour les estimateurs, et d'*estimateur ensembliste* pour les régions de confiance qui vont être présentées dans cette section.

La démarche consiste à se donner une probabilité élevée, qu'on note $1-\alpha$ ($\alpha \in]0;1[$ est donc petit), et, en utilisant un échantillon $X_1, \ldots, X_n$, on cherche à obtenir une région de Θ, notée $\mathscr{C}_n = \mathscr{C}(X_1, \ldots, X_n)$ telle que la probabilité que $\mathscr{C}_n$ contienne la valeur inconnue du paramètre θ est d'au moins $1-\alpha$, quelle que soit cette valeur inconnue :

$$\mathrm{P}_\theta(\theta \in \mathscr{C}_n) \geq 1-\alpha, \quad \forall \theta \in \Theta \tag{10.6}$$ eq:rct

On note dans l'inégalité ci-dessus que la probabilité est indexée par θ. La raison est que la réalisation ou non de l'évènement $\theta \in \mathscr{C}_n$ dépendra des réalisations des variables $X_1, \ldots, X_n$. Par conséquent, la probabilité d'un tel évènement (membre de gauche de l'inégalité (10.6)) sera calculée à partir de la loi de $X_1, \ldots, X_n$. Or cette dernière dépend précisément de θ. Par conséquent, la valeur de cette probabilité dépend également de la valeur de θ, ce qu'on indique donc en notant P_θ.

L'inégalité (10.6) s'interprète de la manière suivante. On ne connaît pas la vraie valeur du paramètre θ, mais on peut examiner ce qui se passe pour n'importe quelle valeur possible de ce paramètre. Envisageons le cas où celle-ci est θ_0, un élément quelconque de Θ. Dans ce cas particulier, on requiert de la région $\mathscr{C}_n$ qu'elle contienne cette valeur θ_0 (qui est la bonne valeur du paramètre) avec une probabilité au moins égale à $1-\alpha$. Comme indiqué dans le paragraphe précédent, cette probabilité dépend de la valeur du paramètre. On a supposé ici que cette valeur est θ_0, et par conséquent la probabilité que la région $\mathscr{C}_n$ contienne θ_0 doit se calculer avec θ_0 comme valeur du paramètre. Cette probabilité est donc $\mathrm{P}_{\theta_0}(\theta_0 \in \mathscr{C}_n)$. On impose alors à la région $\mathscr{C}_n$ de satisfaire $\mathrm{P}_{\theta_0}(\theta_0 \in \mathscr{C}_n) \geq 1-\alpha$. Ceci s'obtient avec la supposition initiale que la valeur du paramètre est θ_0. Mais comme cette dernière est en réalité inconnue, on requiert que ce raisonnement et la condition imposée à $\mathscr{C}_n$ soient valables quelle que soit la valeur possible du paramètre, c'est à dire pour n'importe quelle valeur θ_0 *a priori* possible pour ce paramètre. C'est exactement ce qu'exprime l'inégalité (10.6).

Définition 10.1 *On appelle région de confiance de niveau $1-\alpha$ pour θ toute partie (aléatoire) $\mathscr{C}_n$ de Θ pour laquelle l'inégalité (10.6) est satisfaite. On appelle $1-\alpha$ le niveau de confiance de $\mathscr{C}_n$.*

Notons que l'inégalité (10.6) s'écrit également

$$\inf_{\theta \in \Theta} \mathrm{P}_\theta(\theta \in \mathscr{C}_n) \geq 1-\alpha$$

On appelle le membre de gauche de cette inégalité *probabilité de couverture* de la région $\mathscr{C}_n$.

Construire une région de confiance consiste à délimiter dans l'ensemble Θ des valeurs possibles du paramètre une région dans laquelle se trouve la vraie valeur avec une grande probabilité ($1-$

α). Étant donnée cette démarche, on peut considérer que la délimitation opérée par une région de confiance est assimilable au montant d'information que cette région apporte sur la valeur du paramètre qu'elle cherche à recouvrir. Le nombre $1-\alpha$ peut alors s'interpréter comme le niveau de confiance qu'on souhaite attribuer à cette information. Les points de Θ qui ne sont pas dans une région de confiance donnée ne sont pas considérés comme des valeurs plausibles du paramètre. Le volume de points ainsi écartés est d'autant plus grand que celui de cette région est petit, et donc l'information sur l'endroit de Θ dans lequel il est plausible de trouver la valeur du paramètre est d'autant plus « grande » que le volume de la région est petit.

page:info-rc

Bien qu'on ne formalise pas cette idée, on peut facilement en voir la raison à travers l'exemple suivant. Supposons que le paramètre d'intérêt θ soit la probabilité d'un évènement donné (par exemple la probabilité qu'un ménage choisi au hasard dispose d'une connexion à internet à son domicile). Dans ce cas, on a par construction $\Theta = [0;1]$. Si on choisit $\mathscr{C}_n = [0;1]$, on a évidemment $\mathrm{P}_\theta(\theta \in \mathscr{C}_n) = 1$, $\forall \theta \in \Theta$ et il est donc clair que pour tout niveau de confiance possible $1-\alpha$, la région $[0;1]$ satisfait la condition (10.6). L'intervalle $[0;1]$ est donc une région de confiance pour θ au niveau $1-\alpha$. Cependant on voit bien que tout en possédant un niveau de confiance aussi haut (*i.e.*, aussi proche de 1) qu'on le souhaite, cette région coïncide avec l'ensemble des valeurs a priori possibles pour la probabilité θ et n'apporte donc aucune information sur la vraie valeur du paramètre, autre que celle dont on disposait déjà, à savoir que le paramètre θ étant une probabilité, sa vraie valeur est nécessairement dans l'intervalle $[0;1]$.

Un autre exemple permettant d'illustrer la même idée est le suivant. Supposons que pour un paramètre θ et un niveau de confiance donné $1-\alpha$ nous soyons parvenus à construire une région de confiance $\mathscr{C}_n$. Il est clair que toute partie de Θ contenant $\mathscr{C}_n$ est également une région de confiance de niveau $1-\alpha$ pour θ. En effet, pour toute partie $\mathscr{C}'_n$ de Θ telle que $\mathscr{C}_n \subseteq \mathscr{C}'_n$, l'évènement $\theta \in \mathscr{C}_n$ implique l'évènement $\theta \in \mathscr{C}'_n$ et donc $\mathrm{P}_\theta(\theta \in \mathscr{C}'_n) \geq \mathrm{P}_\theta(\theta \in \mathscr{C}_n)$. Comme la plus petite des deux probabilités est supérieure ou égale à $1-\alpha$ pour tout $\theta \in \Theta$, elles le sont toutes les deux. Par conséquent, $\mathscr{C}'_n$ satisfait la condition (10.6). Cependant, si le niveau de confiance requis est de $1-\alpha$, parmi les deux régions $\mathscr{C}_n$ et $\mathscr{C}'_n$ ayant ce niveau, on préfèrera $\mathscr{C}_n$ à $\mathscr{C}'_n$, puisqu'avec un même niveau de confiance, et tout en étant contenue dans $\mathscr{C}'_n$, la région $\mathscr{C}_n$ délimite dans Θ un ensemble de valeurs possibles pour θ moins volumineux que $\mathscr{C}'_n$. La région $\mathscr{C}_n$ est donc plus informative que $\mathscr{C}'_n$ à propos de la vraie valeur du paramètre θ.

Même si dans les exemples ci-dessus le volume d'une région est une caractéristique à prendre en compte pour évaluer le montant d'information qu'une région de confiance apporte sur la valeur du paramètre qu'elle cherche à recouvrir, il ne constitue pas un critère utilisable de manière suffisamment générale pour choisir parmi plusieurs régions de confiance. En effet, dans ces exemples les régions comparées sont emboîtées et dans ce cas la comparaison des volumes est facile. Dans le cas général, c'est un critère d'*exactitude* qui est retenu pour comparer des régions de confiance (et choisir celles qu'il est optimal d'utiliser).

def:rc_prec

Définition 10.2 *Soient $\mathscr{C}_n$ et $\mathscr{C}^*_n$ deux régions de confiance de même niveau pour θ. On dit que $\mathscr{C}^*_n$ est plus exacte que $\mathscr{C}_n$ si*

$$\mathrm{P}_{\theta_1}(\theta_0 \in \mathscr{C}^*_n) \leq \mathrm{P}_{\theta_1}(\theta_0 \in \mathscr{C}_n), \qquad \forall \theta_0 \in \Theta, \forall \theta_1 \in \Theta, \theta_0 \neq \theta_1$$

Le terme $\mathrm{P}_{\theta_1}(\theta_0 \in \mathscr{C}^*_n)$ est la probabilité que $\mathscr{C}^*_n$ contienne la valeur θ_0 du paramètre, calculée en

supposant que celle-ci est θ_1. Dans un tel calcul, si on suppose que θ_1 est la valeur du paramètre, alors lorsque l'évènement $\theta_0 \in \mathscr{C}_n^*$ se réalise, $\mathscr{C}_n^*$ contient une valeur du paramètre qui n'est pas la bonne. Donc $\mathrm{P}_{\theta_1}(\theta_0 \in \mathscr{C}_n^*)$ est la probabilité que la région $\mathscr{C}_n^*$ contienne une mauvaise valeur du paramètre. On a la même interprétation pour la région $\mathscr{C}_n$. Par conséquent, la définition ci-dessus dit que parmi deux régions de confiance de même niveau, la plus exacte est celle ayant la plus petite probabilité de contenir toute valeur erronée du paramètre. On préfèrera et choisira donc la plus exacte de ces deux régions.

Puisque le critère d'exactitude permet de comparer des régions de confiance, il est naturel de chercher à déterminer la meilleure parmi toutes les régions d'un niveau donné. L'approche est donc semblable à celle des tests, puisqu'ici, sous la contrainte que les régions considérées aient toutes le niveau choisi $1-\alpha$, on sélectionnera parmi celles-ci la plus exacte (comparer cette approche avec celle de Neyman-Pearson retenue pour les tests, décrite page 250). Plus précisément, pour une région de confiance $\mathscr{C}_n$, lorsque $\theta_0 \neq \theta_1$, on peut assimiler $\mathrm{P}_{\theta_1}(\theta_0 \in \mathscr{C}_n)$ à la probabilité d'une mauvaise décision : comme on l'a dit dans le paragraphe précédent, c'est la probabilité que la région $\mathscr{C}_n$ contienne une mauvaise valeur du paramètre. Cette probabilité est donc assimilable aux risques d'un test. La démarche de choix d'une région de confiance consistant à maximiser son exactitude est donc identique à la démarche de choix d'un test par minimisation de ses risques. Dans les deux cas (test ou région de confiance) une contrainte sur la probabilité d'une bonne décision est imposée.

L'analogie entre tests et régions de confiance est formalisée par des résultats qui établissent une correspondance entre ces deux outils d'inférence. Ils permettent notamment de calquer l'analyse des propriétés des régions de confiance sur celles des tests. Le premier de ces résultats est le théorème 10.2 ci-dessous ; il montre que construire une région de confiance de niveau $1-\alpha$ pour un paramètre est équivalent à construire une famille de tests de niveau α pour tester des hypothèses sur la valeur de ce paramètre. Le second résultat est le corollaire 10.1, qui montre que chercher la région la plus exacte revient à chercher le test le plus puissant.

th:rc-tests

Théorème 10.2

1. *À tout $\theta_0 \in \Theta$ on associe φ_{θ_0}, un test de niveau α pour tester $H_0 : \theta = \theta_0$ contre une hypothèse alternative quelconque. Comme auparavant on definit $\varphi_{\theta_0,n}$ par $\varphi_{\theta_0}(X_1, \ldots, X_n)$. La partie $\mathscr{C}_n$ de Θ définie par*

$$\mathscr{C}_n = \{\theta_0 \in \Theta \mid \varphi_{\theta_0,n} = 0\} \tag{10.7}$$

eq:rc-tests1

est une région de confiance de niveau $1-\alpha$ pour θ.

2. *Soit $\mathscr{C}_n$ une région de confiance de niveau $1-\alpha$ pour θ. Soit θ_0 un élément de Θ. La fonction φ_{θ_0} définie par*

$$\varphi_{\theta_0}(X_1, \ldots, X_n) = \begin{cases} 1 \text{ si } \theta_0 \notin \mathscr{C}_n \\ 0 \text{ sinon} \end{cases} \tag{10.8}$$

eq:rc-tests

est un test de niveau α pour tester $H_0 : \theta = \theta_0$ contre n'importe quelle alternative. [12]

12. La raison pour laquelle on ne désigne pas les hypothèses alternatives dans ce théorème (et qu'on peut donc les choisir comme on veut) est que les résultats qu'il contient ne concernent que le niveau d'un test, c'est à dire la probabilité d'un évènement, calculée en supposant vraie l'hypothèse nulle. Il n'est donc pas nécessaire de considérer les hypothèses alternatives.

Preuve : 1. Soit θ_0 un élément quelconque de Θ. Notons que par construction de $\mathscr{C}_n$, on a $\theta_0 \in \mathscr{C}_n \iff \varphi_{\theta_0,n} = 0$ et donc

$$\mathrm{P}_{\theta_0}(\theta_0 \in \mathscr{C}_n) = \mathrm{P}_{\theta_0}(\varphi_{\theta_0,n} = 0) = 1 - \mathrm{P}_{\theta_0}(\varphi_{\theta_0,n} = 1)$$

Comme le test est de niveau α pour tester H_0, cette probabilité est supérieure ou égale à $1-\alpha$ (voir l'inégalité (10.1)).

2. Soit une région de confiance $\mathscr{C}_n$ de niveau $1-\alpha$ pour le paramètre θ. Soit φ_{θ_0} le test défini par (10.8). On a

$$\mathrm{P}_{\theta_0}(\varphi_{\theta_0}(X_1,\ldots,X_n) = 1) = \mathrm{P}_{\theta_0}(\theta_0 \notin \mathscr{C}_n) \leq \alpha$$

où l'inégalité est obtenue en notant que la région de confiance $\mathscr{C}_n$ est de niveau $1-\alpha$ (voir l'inégalité (10.6)).

Ce théorème montre qu'à toute région de confiance de niveau $1-\alpha$ on peut associer une famille de tests de niveau α et réciproquement. Plus précisément, si $\mathscr{C}_n$ est une région de confiance, le théorème montre qu'on peut lui associer une famille $\{\varphi_{\theta_0} \mid \theta_0 \in \Theta\}$ de tests de niveau α pour les hypothèses nulles de la forme $H_0 : \theta = \theta_0$ au moyen de la relation (10.8). En particulier, si on dispose d'une région de confiance $\mathscr{C}_n$ de niveau $1-\alpha$ pour le paramètre θ, alors on a automatiquement un test de niveau α pour tester $H_0 : \theta = \theta_0$ qui consiste à décider que H_0 est fausse si la région $\mathscr{C}_n$ ne contient pas θ_0. Réciproquement, si pour tout $\theta_0 \in \Theta$ on dispose d'un test φ_{θ_0} de niveau α pour tester $H_0 : \theta = \theta_0$, alors on a automatiquement une région de confiance de niveau $1-\alpha$ pour θ en formant l'ensemble de toutes les valeurs θ_0 pour lesquelles on décide d'accepter l'hypothèse nulle $H_0 : \theta = \theta_0$ avec le test φ_{θ_0}.

La correspondance entre tests et régions de confiance établie par le théorème 10.2 peut s'utiliser pour montrer que l'approche par laquelle on compare des régions de confiance sur la base de leur exactitude est identiques à celle utilisée pour comparer les tests. En effet, en utilisant la définition 10.2 et le théorème 10.2, on peut montrer le corollaire suivant, qui établit que pour comparer l'exactitude de deux régions de confiance, il suffit de comparer la puissance des tests associés.

cor:rc-tests **Corollaire 10.1** *Soient $\mathscr{C}_n$ et $\mathscr{C}_n^*$ deux régions de confiance de niveau $1-\alpha$ pour un paramètre θ, et soient $\{\varphi_{\theta_0} \mid \theta_0 \in \Theta\}$ et $\{\varphi^*_{\theta_0} \mid \theta_0 \in \Theta\}$ les familles de tests qui leur sont respectivement associées. $\mathscr{C}_n^*$ est plus exacte que $\mathscr{C}_n$ si et seulement si pour tout $\theta_0 \in \Theta$, la puissance de φ_{θ_0} n'excède jamais celle de $\varphi^*_{\theta_0}$.*

Preuve : Soit $\theta_0 \in \Theta$ et le problème de test $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$. Si la puissance du test φ_{θ_0} n'excède jamais celle de $\varphi^*_{\theta_0}$, la probabilité que $\varphi^*_{\theta_0}$ décide H_1 calculée en supposant H_1 vraie doit être supérieure ou égale à la probabilité correspondante pour φ_{θ_0}. Dire que H_1 est supposée vraie revient à dire que ces probabilités doivent être calculées pour des valeurs θ_1 du paramètre différentes de θ_0. Par conséquent, on doit avoir $\mathrm{P}_{\theta_1}(\varphi^*_{\theta_0,n} = 1) \geq \mathrm{P}_{\theta_1}(\varphi_{\theta_0,n} = 1)$, $\forall \theta_1 \neq \theta_0$, ou encore

$$\mathrm{P}_{\theta_1}(\varphi^*_{\theta_0,n} = 0) \leq \mathrm{P}_{\theta_1}(\varphi_{\theta_0,n} = 0), \quad \forall \theta_1 \neq \theta_0$$

Par construction des tests, on a $\varphi_{\theta_0,n} = 0 \iff \theta_0 \in \mathscr{C}_n$ et donc l'inégalité ci-dessus s'écrit

$$\mathrm{P}_{\theta_1}(\theta_0 \in \mathscr{C}_n^*) \leq \mathrm{P}_{\theta_1}(\theta_0 \in \mathscr{C}_n), \quad \forall \theta_1 \neq \theta_0$$

Comme ceci est vrai pour tout θ_0, on en déduit que $\mathscr{C}^*$ est plus exacte que $\mathscr{C}$.

Réciproquement, si $\mathscr{C}_n^*$ est plus exacte que $\mathscr{C}_n$, alors pour tout θ_0 et θ_1 dans Θ on a l'inégalité suivante :

$$\mathrm{P}_{\theta_1}(\varphi^*_{\theta_0,n} = 1) = \mathrm{P}_{\theta_1}(\theta_0 \notin \mathscr{C}_n^*) \geq \mathrm{P}_{\theta_1}(\theta_0 \notin \mathscr{C}_n) = \mathrm{P}_{\theta_1}(\varphi^*_{\theta_0,n} = 1)$$

où les égalités proviennent de l'association entre régions de confiance et tests (voir le théorème 10.2). Pour θ_0 fixé, cette inégalité est vraie pour tout $\theta_1 \neq \theta_0$. Par conséquent, on peut écrire

$$\mathrm{P}_{H_1}(\varphi^*_{\theta_0,n} = 1) \geq \mathrm{P}_{H_1}(\varphi^*_{\theta_0,n} = 1)$$

ce qui établit que $\varphi^*_{\theta_0}$ est plus puissant que φ_{θ_0}.

Le corollaire 10.1 s'utilise surtout pour établir que si on dispose d'un test φ_0^* de niveau α pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ qui est UPP, alors la région de confiance de niveau $1-\alpha$ pour le paramètre $1-\alpha$ la plus exacte est celle associée au test φ_0^*.

On a mentionné dans la section précédente qu'il est en général impossible de construire des tests UPP, et qu'on est alors amené à chercher le test le plus puissant dans un ensemble donné de tests, formé en imposant des propriétés souhaitées aux tests qui le composent. Parmi ces propriétés, on a introduit l'absence de biais. Soit φ_{θ_0} un test sans biais de niveau α pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$; la région de confiance $\mathscr{C}_n$ associée à φ_{θ_0} satisfait la propriété suivante :

$$\mathrm{P}_{\theta_0}(\theta_0 \in \mathscr{C}_n) \geq \mathrm{P}_{\theta_1}(\theta_0 \in \mathscr{C}_n) \qquad \forall \theta_0, \theta_1 \in \Theta \tag{10.9}$$

eq:rc_ss_biais

Cette inégalité est l'inégalité (10.4) qui caractérise l'absence de biais de φ_{θ_0}, réécrite en faisant appel à l'équivalence $\theta_0 \in \mathscr{C}_n \iff \varphi_{\theta_0,n} = 0$ (voir le théorème 10.2) et en notant que si on suppose que $\theta = \theta_1$, alors H_1 est supposée vraie. L'inégalité (10.9) peut s'interpréter comme une condition d'absence de biais de la région $\mathscr{C}_n$. Dans cette inégalité, la valeur θ_0 du paramètre est celle qui est utilisée pour calculer les probabilités. Par conséquent, cette valeur est supposée être la vraie valeur du paramètre. La valeur $\theta_1 \neq \theta_0$ est donc une valeur erronée. L'inégalité établit alors qu'il est plus probable que la région $\mathscr{C}_n$ contienne la vraie valeur du paramètre qu'une valeur erronée de ce paramètre. Ce type de propriété fait donc partie des propriétés souhaitables qu'on peut vouloir imposer lorsqu'on cherche à construire une région de confiance.

page:rc_ss_biais