COURS D'ÉCONOMÉTRIE

Professeur Philippe Deschamps

Edition 2006-2007

Université de Fribourg
Séminaire d'Econométrie
Boulevard de Pérolles 90
CH-1700 Fribourg, Suisse

TABLE DES MATIERES

Seconde partie: **Modèles économétriques à une équation**

AVANT-PROPOS

Ce cours d'économétrie de second cycle est enseigné depuis 1981 aux étudiants de troisième et de quatrième année de licence en Sciences Economiques à l'Université de Fribourg (Suisse), et, depuis 1996, aux étudiants du diplôme de Mathématiques appliquées à la Finance de l'Université de Neuchâtel (dans le cadre des accords BENEFRI).

Les notes de ce cours peuvent être imprimées et peuvent être utilisées, en tout ou en partie, comme support d'un cours de niveau équivalent, à condition:

(1) d'en avertir l'auteur à l'adresse suivante:

philippe.deschamps@unifr.ch;

(2) d'en mentionner clairement l'origine.

Elles ne peuvent pas être publiées sur un site différent de leur site d'origine:

http://mypage.bluewin.ch/Philippe_Deschamps.

Ces notes ont été composées à l'aide des logiciels $A_{\mathcal{M}}S-T_{\mathrm{E}}X$, $P_{\mathrm{I}}CT_{\mathrm{E}}X$, et $T_{\mathrm{A}}B_{\mathrm{L}}E$. L'auteur remercie Madame Edith Beck-Walser, qui a mené à bien, avec beaucoup de dévouement, la saisie informatique d'une version préliminaire du texte. Il remercie également Monsieur Roberto Cerratti pour ses commentaires constructifs, Mademoiselle Réanne Meyer pour la composition des formules des chapitres XV et XVI de la seconde partie, et Mademoiselle Brigitte Sermier pour son assistance efficace lors de la correction des épreuves.

Fribourg, été 2002.

CONNAISSANCES PRÉREQUISES

- Cours de mathématiques de première année (l'équivalent de l'ouvrage de P. Deschamps, *Cours de Mathématiques pour Economistes*, Paris, Dunod 1988).
- Probabilité, probabilité jointe, probabilité conditionnelle
- Indépendance de deux événements
- Théorème de la probabilité totale
- Variables aléatoires discrètes et continues
- Distribution et densité (cas univarié)
- Espérance mathématique et propriétés
- Variance et propriétés
- Variable aléatoire binomiale
- Variable aléatoire uniforme
- Variable aléatoire normale: propriétés et emploi des tables

PREMIÈRE PARTIE

QUELQUES NOTIONS DE BASE DU CALCUL DES PROBABILITÉS ET DE L'ANALYSE STATISTIQUE

CHAPITRE I

VECTEURS ALÉATOIRES

Définition

On peut associer à tout résultat possible ω d'une expérience aléatoire un vecteur $X(\omega) \in \mathbb{R}^k$. Si pour tout $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$, l'ensemble:

$$\{\omega \mid X_i(\omega) \leq x_i, \quad i = 1, \ldots, k\}$$

est un événement dont on peut calculer la probabilité, la fonction $X(\omega)$ est dite mesurable et X porte le nom de vecteur aléatoire. Il est discret si $X(\omega)$ prend ses valeurs dans un ensemble dénombrable, continu sinon.

1.1 Distribution jointe

Dans le cas discret et continu, elle peut s'énoncer comme:

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = P[(X_1 \leq x_1) \cap (X_2 \leq x_2) \cap \ldots \cap (X_k \leq x_k)] \quad .$$

1.2 Densité jointe

- Cas discret:

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = P[(X_1 = x_1) \cap (X_2 = x_2) \cap \ldots \cap (X_k = x_k)] \quad .$$

- Cas continu: la fonction de densité est la fonction dont l'intégrale donne la fonction de distribution. Formellement, $f_X = f_{X_1,\ldots,X_k}$ est la densité jointe du vecteur $X = (X_1,\ldots,X_k)$ si:

$$F_X(x_1,\ldots,x_k) = \int_{-\infty}^{x_k} \ldots \int_{-\infty}^{x_1} f_X(u_1,\ldots,u_k)\, du_1 \ldots du_k \quad .$$

Note

Dans tout ce qui suit, nous supposerons pour alléger la notation que $k = 2$. La généralisation à $k > 2$ est facile et les définitions pertinentes se trouvent dans la littérature. On étudiera donc un vecteur (X, Y).

Exemples

- Cas discret: Le tableau suivant donne les valeurs de deux variables X et Y et les probabilités que le couple (X, Y) prenne la valeur (x, y):

		X			
		0	1	2	
Y	0	0,20	0,20	0,10	0,5
	1	0,40	0,05	0,05	0,5
		0,60	0,25	0,15	

On obtient:
$f_{X,Y}(0,0) = 0,2$; $f_{X,Y}(0,1) = 0,4$; etc.
$F_{X,Y}(1,0) = 0,4$; $F_{X,Y}(1,1) = 0,85$; etc.

- Cas continu:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{x^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}\right)$$

(densité jointe de deux variables normales centrées indépendantes) .

En intégrant cette densité sur $[a, b] \times [c, d]$, on obtient $P[(a \leq X \leq b) \cap (c \leq Y \leq d)]$.

1.3 Densité marginale

- Cas discret:

$$\begin{aligned} f_X(x_i) &= \sum_j f_{X,Y}(x_i, y_j) \\ f_Y(y_j) &= \sum_i f_{X,Y}(x_i, y_j) \end{aligned}$$

- Cas continu:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dy$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dx$$

Exemple

Pour les densités jointes données précédemment à la section 1.2:

(a) $f_X(0) = 0,6 \quad ; \quad f_X(1) = 0,25 \quad ; \quad f_X(2) = 0,15$
$f_Y(0) = 0,5 \quad ; \quad f_Y(1) = 0,5$

(b)

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{x^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}\right) dy \\ &= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) \underbrace{\int_{-\infty}^{+\infty} \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(\frac{-y^2}{2\sigma_2^2}\right) dy}_{=1} \\ &= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) \end{aligned}$$

$$f_Y(y) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{y^2}{2\sigma_2^2}\right) \quad .$$

1.4 Densité conditionnelle

- Cas discret: les densités conditionnelles s'obtiennent à partir de la définition d'une probabilité conditionnelle $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  Donc:

$$f_{X|Y}(x_i \mid y_j) = \frac{f_{X,Y}(x_i, y_j)}{f_Y(y_j)}$$

(définie si $f_Y(y_j) \neq 0$).

- Cas continu:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad \text{si} \quad f_Y(y) \neq 0 \quad .$$

Note: cette fonction dépend d'une réalisation particulière de Y. Cette fonction est donc aléatoire car Y est aléatoire (on peut dire aussi qu'elle dépend d'un paramètre aléatoire).

Exemple pour les densités jointes données précédemment (section 1.2):

(a) Cas discret:
$f_{X|Y}(0 \mid 0) = 0,4$
$f_{X|Y}(1 \mid 0) = 0,4$
$f_{X|Y}(2 \mid 0) = 0,2$
Les valeurs de $f_{X|Y}(x \mid 1)$ sont celles d'une *autre* densité.

(b) Dans le cas continu, on avait $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. Donc $f_{X|Y}(x \mid y) = f_X(x)$

1.5 Indépendance

- Cas discret: X et Y sont indépendantes si pour tout i et pour tout j, on a:

$$f_{X,Y}(x_i, y_j) = f_X(x_i) f_Y(y_j) \quad .$$

Dans l'exemple précédent (section 1.2, cas discret), X et Y ne sont pas indépendantes, car:

$$f_{X,Y}(0,0) = 0,2 \neq f_X(0) f_Y(0) = 0,6 \cdot 0,5 \quad .$$

- Cas continu: X et Y sont indépendantes si pour tout x et pour tout y, on a:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \quad .$$

Dans l'exemple précédent (section 1.2, cas continu), on a l'indépendance.

Propriété très importante

Si X et Y sont indépendantes, alors: $E(XY) = E(X)E(Y)$. La réciproque n'est pas vraie en général!

Exercice. Démontrez la propriété précédente dans le cas continu.

1.6 Covariance

Définition

$$\mathrm{Cov}(X,Y) = E\left[\{X - E(X)\}\{Y - E(Y)\}\right] \quad .$$

Exercice

Montrez que $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$.

Propriété importante (conséquence de l'exercice)

Si X et Y sont indépendantes, alors $\mathrm{Cov}(X,Y) = 0$. La réciproque n'est pas vraie en général!

Contre exemple montrant que la réciproque n'est pas vraie.

		X			
		-1	0	$+1$	
Y	-1	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{5}{16}$
	0	$\frac{3}{16}$	0	$\frac{3}{16}$	$\frac{6}{16}$
	$+1$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{5}{16}$
		$\frac{5}{16}$	$\frac{6}{16}$	$\frac{5}{16}$	

On n'a pas l'indépendance, car

$$f_{X,Y}(0,0) = 0 \neq f_X(0) f_Y(0) = \frac{6}{16} \cdot \frac{6}{16} \quad .$$

Mais la covariance est nulle:

$$\begin{aligned} E(XY) &= 1\cdot\frac{1}{16}+0\cdot\frac{3}{16}-1\cdot\frac{1}{16}+0\cdot\frac{3}{16}+0\cdot 0 \\ &\quad +0\cdot\frac{3}{16}-1\cdot\frac{1}{16}+0\cdot\frac{3}{16}+1\cdot\frac{1}{16}=0 \\ E(X) &= -\frac{5}{16}+0+\frac{5}{16}=0 \\ E(Y) &= -\frac{5}{16}+0+\frac{5}{16}=0 \\ \Longrightarrow \quad \mathrm{Cov}(X,Y) &= E(XY)-E(X)E(Y)=0 \quad . \end{aligned}$$

1.7 Espérances conditionnelles et partielles

L'espérance conditionnelle s'évalue à partir de la densité conditionnelle.

- Cas discret: $E\left(X \mid Y=y_j\right)=\sum_i x_i f_{X|Y}\left(x_i \mid y_j\right)$
- Cas continu: $E\left(X \mid Y=y\right)=\int_{-\infty}^{+\infty} x f_{X|Y}\left(x \mid y\right) dx$

Dans l'exemple de la section 1.2 (cas discret):

$$\begin{aligned} E\left(X \mid Y=0\right) &= 0,4\cdot 0+0,4\cdot 1+0,2\cdot 2=0,8 \\ E\left(X \mid Y=1\right) &= 0,8\cdot 0+0,1\cdot 1+0,1\cdot 2=0,3 \quad . \end{aligned}$$

Propriété très importante

$$E(X)=E_Y[E\left(X \mid Y\right)] \quad .$$

Cette propriété porte le nom de "loi des espérances itérées" (Law of Iterated Expectations). Elle est analogue au théorème de la probabilité totale: une espérance inconditionnelle, tout comme une probabilité inconditionnelle, peut être évaluée à l'aide d'un arbre.

- Loi des espérances itérées dans le cas discret:

$$E(X) = \sum_j E(X \mid Y = y_j) P(Y = y_j)$$

- Loi des espérances itérées dans le cas continu:

$$E(X) = \int_{-\infty}^{+\infty} f_Y(y) \underbrace{\int_{-\infty}^{+\infty} x f_{X|Y}(x \mid y)\, dx}_{E(X|Y)} \quad dy$$

Exemple pour le cas discret (données de la section 1.2):

- On a vu que $E(X \mid Y = 0) = 0,8$ et $E(X \mid Y = 1) = 0,3$.

- Par ailleurs $P(Y = 0) = 0,5$ et $P(Y = 1) = 0,5$. $E_Y[E(X \mid Y)]$ est la moyenne des espérances conditionnelles:

$$\begin{aligned} E_Y[E(X \mid Y)] &= E(X \mid Y = 0) P(Y = 0) + E(X \mid Y = 1) P(Y = 1) \\ &= 0,8 \cdot 0,5 + 0,3 \cdot 0,5 = 0,55 \quad . \end{aligned}$$

- Il est facile de vérifier à l'aide de la densité marginale que $0,55$ est bien égale à $E(X)$:

$$\begin{aligned} E(X) &= \sum_i x_i P[X = x_i] \\ &= 0 \cdot 0,6 + 1 \cdot 0,25 + 2 \cdot 0,15 = 0,55 \quad . \end{aligned}$$

Cas particulier de l'espérance conditionnelle: l'espérance partielle

Définition

$$\begin{aligned} E(Y \mid Y \leq a) &= \sum_j y_j P(Y = y_j \mid Y \leq a) \quad \text{(cas discret)} \\ &= \int_{-\infty}^{+\infty} y f(y \mid Y \leq a)\, dy \qquad \text{(cas continu)} \\ \text{où} \quad f(y \mid Y \leq a) &= \frac{d}{dy} P(Y \leq y \mid Y \leq a) \quad . \end{aligned}$$

Propriété

- Dans le cas discret:

$$E\left(Y \mid Y \leq a\right) = \sum_{\{j:y_j \leq a\}} y_j \frac{P(Y = y_j)}{P(Y \leq a)}$$

- Dans le cas continu:

$$E\left(Y \mid Y \leq a\right) = \int_{-\infty}^{a} y \frac{f_Y(y)}{F_Y(a)} dy$$

Démonstration pour le cas continu:

$$\begin{aligned} P\left(Y \leq y \mid Y \leq a\right) &= \frac{P\left(Y \leq y \cap Y \leq a\right)}{P\left(Y \leq a\right)} \\ &= \begin{cases} \dfrac{F_Y(y)}{F_Y(a)} & \text{si} \quad y \leq a \\ 1 & \text{si} \quad y > a \end{cases} \end{aligned}$$

Donc:

$$\begin{aligned} f\left(y \mid Y \leq a\right) &= \frac{d}{dy} P\left(Y \leq y \mid Y \leq a\right) \\ &= \begin{cases} \dfrac{f_Y(y)}{F_Y(a)} & \text{si} \quad y \leq a \\ 0 & \text{si} \quad y > a \end{cases} \end{aligned}$$

et $\int_{-\infty}^{+\infty} y f\left(y \mid Y \leq a\right) dy = \int_{-\infty}^{a} y \frac{f_Y(y)}{F_Y(a)} dy.$

Exercice. Démontrez la propriété précédente dans le cas discret.

1.8 Application économique des espérances partielles (gestion de stock)

Cet exercice a pour but d'illustrer l'intérêt de la loi des espérances itérées, appliquée aux espérances partielles.

Énoncé

Un commerçant a une demande journalière aléatoire Y pour une denrée vendue par kilos. Y, mesurée en centaines de kilos, a la densité suivante:

$$\begin{aligned} f_Y(y) &= 3y^2 \quad \text{si} \quad 0 \leq y \leq 1 \\ &= 0 \quad \text{sinon} \quad . \end{aligned}$$

(Le commerçant ne peut stocker plus de 100 kilos).

Il veut commander $k \cdot 100$ kilos de cette denrée. Il l'achète 6 francs par kilo et la vend 10 francs par kilo. Quelle est la valeur de k qui maximisera l'espérance mathématique de son profit journalier?

Solution

Le profit peut s'écrire comme:

$$\begin{aligned} \Pi(k, Y) &= 1000Y - 600k \quad \text{si} \quad Y \leq k \\ &= 400k \qquad\quad \text{si} \quad Y > k \quad . \end{aligned}$$

Le profit est aléatoire. Mais son espérance ne dépend que de la variable de décision k. Il s'agit donc de calculer cette espérance et de la maximiser par rapport à k.

La loi des espérances itérées donne:

$$E(\Pi) = E(\Pi \mid Y \leq k) P(Y \leq k) + E(\Pi \mid Y > k) P(Y > k) \quad .$$

On va évaluer tour à tour chacun de ces termes. $E(\Pi \mid Y \leq k)$ dépend de:

$$\begin{aligned} E(Y \mid Y \leq k) &= \int_{-\infty}^{k} y \frac{f_Y(y)}{F_Y(k)} dy \\ &= \int_0^k \frac{y\left(3y^2\right)}{k^3} dy \\ &= \left[\frac{3}{4}\frac{y^4}{k^3}\right]_0^k = \frac{3}{4}k \quad . \end{aligned}$$

Alors:

$$\begin{aligned}
E(\Pi \mid Y \leq k) &= 1000E(Y \mid Y \leq k) - 600k \\
&= 1000\left(\frac{3}{4}k\right) - 600k = 150k \\
P(Y \leq k) &= \int_0^k 3y^2 dy = \left[\frac{3y^3}{3}\right]_0^k = k^3 \\
P(Y > k) &= 1 - k^3 \\
E(\Pi \mid Y > k) &= 1000k - 600k = 400k \quad .
\end{aligned}$$

En combinant:

$$\begin{aligned}
E(\Pi) &= (150k)\,k^3 + (400k)\left(1 - k^3\right) \\
&= -250k^4 + 400k \quad .
\end{aligned}$$

En maximisant:

$$\begin{aligned}
\frac{dE(\Pi)}{dk} &= -1000k^3 + 400 = 0 \\
&\Longrightarrow \quad k^3 = 0,4 \Longrightarrow k = (0,4)^{1/3} \approx 0,7368 \quad . \\
\frac{d^2E(\Pi)}{dk^2} &= -3000k^2 < 0 \quad .
\end{aligned}$$

CHAPITRE II

FONCTIONS DE VARIABLES ALÉATOIRES

2.1 Changement de variables (cas univarié)

Énoncé du problème

On connaît une densité $f_Y(y)$. Quelle est la densité d'une fonction strictement monotone (i.e. strictement croissante ou strictement décroissante) de Y? Si $U = h(Y)$, alors, si h est croissante:

$$\begin{aligned} P[U \leq u] &= P[h(Y) \leq u] \\ &= P[Y \leq h^{-1}(u)] \end{aligned}$$

et, si h est décroissante:

$$P[U \leq u] = P[Y \geq h^{-1}(u)].$$

Mais quelle est la densité qui donne bien cette probabilité lorsqu'on l'intègre? La réponse est donnée par le théorème du changement de variables, dont on va voir la version univariée et multivariée.

Théorème.

Supposons que la variable aléatoire continue Y ait pour densité $f_Y(y)$ et soit:

$$\mathcal{Y} = \{y \mid f_Y(y) > 0\} \quad (\mathcal{Y} \textit{ s'appelle le support de } f_Y)$$

Si $h(\cdot)$ est une fonction dérivable et strictement monotone de domaine $\mathcal{Y}$ et d'image $\mathcal{U}$, alors $U = h(Y)$ a pour densité:

$$\begin{aligned} f_U(u) &= f_Y\left(h^{-1}(u)\right) \left| \frac{dy}{du} \right| \quad \textit{pour} \quad u \in \mathcal{U} \\ &= 0 \quad \textit{sinon} \quad . \end{aligned}$$

Exemple

- Soit

$$\begin{aligned} f_Y(y) &= 2y \quad \text{si} \quad 0 \leq y \leq 1 \\ &= 0 \quad \text{sinon} \quad . \end{aligned}$$

- On cherche la densité de $U = h(Y) = -4Y + 3$. Cette fonction est dérivable et bijective.
- En résolvant $u = -4y + 3$, on obtient:

$$y = \frac{3-u}{4} \quad , \quad \text{donc} \quad \left| \frac{dy}{du} \right| = \frac{1}{4} \quad \text{et} \quad h^{-1}(u) = \frac{3-u}{4} \quad .$$

- Le théorème donne:

$$\begin{aligned} f_U(u) &= f_Y\left(\frac{3-u}{4}\right)\frac{1}{4} \\ &= 2\left(\frac{3-u}{4}\right)\frac{1}{4} \quad \text{si} \quad -1 \leq u \leq 3 \\ f_U(u) &= 0 \quad \text{sinon} \quad . \end{aligned}$$

Exercice: Soit Y la valeur d'un portefeuille en euros et $U = 1.5Y$ la valeur du même portefeuille en francs suisses. On suppose que la densité de Y est exponentielle:

$$\begin{aligned} f_Y(y) &= \beta e^{-\beta y} \quad \text{pour } y > 0 \\ &= 0 \quad \text{sinon.} \end{aligned}$$

On demande de trouver la densité de la variable U.

2.2 Changement de variables (cas multivarié)

Théorème.

Soit Y_1 et Y_2 deux variables aléatoires de densité jointe $f_{Y_1,Y_2}(y_1, y_2)$. Soit:

$$\mathcal{Y} = \{(y_1, y_2) \mid f_{Y_1,Y_2}(y_1, y_2) > 0\} \quad .$$

Soit $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = h\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ une fonction bijective de domaine $\mathcal{Y}$ et d'image $\mathcal{U}$.

Si:

(1) *les dérivées partielles de h sont continues sur $\mathcal{Y}$,*
(2) *le jacobien:*

$$J = \det \begin{pmatrix} \partial y_1/\partial u_1 & \partial y_1/\partial u_2 \\ \partial y_2/\partial u_1 & \partial y_2/\partial u_2 \end{pmatrix}$$

est non nul pour $(u_1, u_2) \in \mathcal{U}$,

alors:

$$\begin{aligned} f_{U_1,U_2}(u_1,u_2) &= \mid J \mid f_{Y_1,Y_2}[h^{-1}(u_1,u_2)] \quad \textit{pour} \quad u \in \mathcal{U} \\ &= 0 \quad \textit{sinon} \quad . \end{aligned}$$

Exemple

Densité de la somme et de la différence de variables uniformes.

$$\begin{aligned} \text{Soit} \quad f_{Y_1,Y_2}(y_1,y_2) &= 1 \quad \text{si} \quad 0 \le y_1 \le 1 \quad \text{et} \quad 0 \le y_2 \le 1 \\ &= 0 \quad \text{sinon} \quad . \end{aligned}$$

On demande la densité jointe de:

$$\begin{pmatrix} U_1 \\ \\ U_2 \end{pmatrix} = \begin{pmatrix} Y_1 + Y_2 \\ \\ Y_2 - Y_1 \end{pmatrix} \quad .$$

On peut écrire:

$$\begin{pmatrix} u_1 \\ \\ u_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \\ -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \\ y_2 \end{pmatrix}$$

$$\Longrightarrow \qquad \begin{pmatrix} y_1 \\ \\ y_2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ \\ u_2 \end{pmatrix}$$

$$\Longrightarrow \qquad J = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} = \mid J \mid \quad .$$

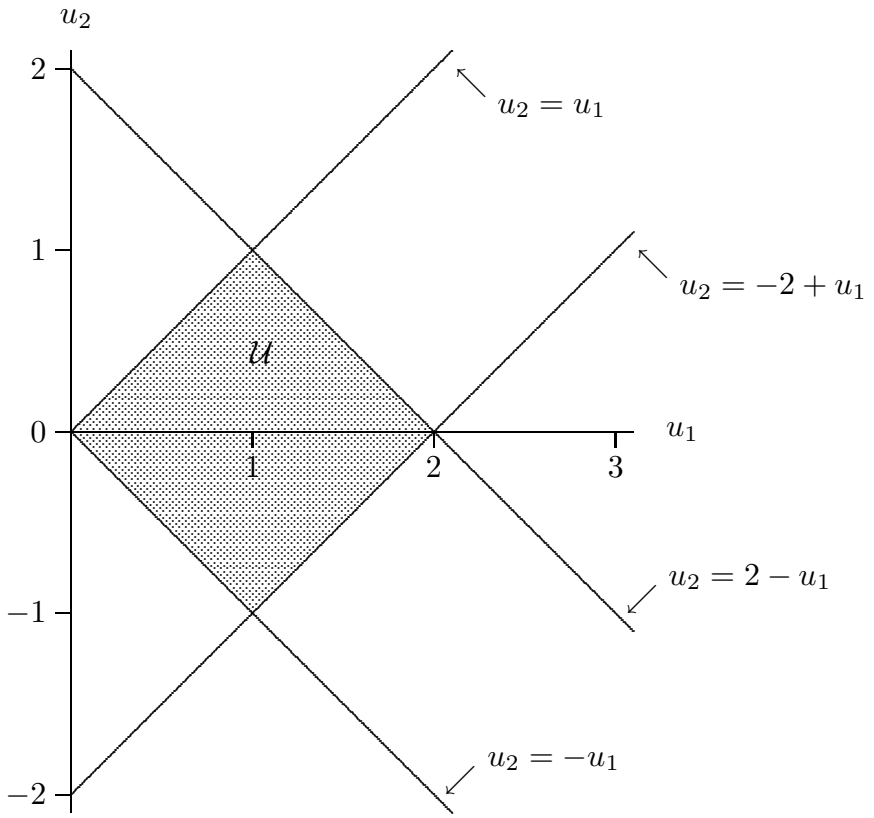
$$\begin{aligned} \text{Donc} \quad f_{U_1,U_2}(u_1,u_2) &= \frac{1}{2} \quad \text{pour} \quad u \in \mathcal{U} \\ &= 0 \quad \text{sinon} \quad . \end{aligned}$$

Mais quelle est la forme de $\mathcal{U}$? Pour déterminer la forme de $\mathcal{U}$, il faut traduire les conditions sur y_1, y_2 en un système de conditions sur u_1, u_2.

On a $y_1 = \frac{1}{2}(u_1 - u_2)$ et $y_2 = \frac{1}{2}(u_1 + u_2)$. Donc:

$$\begin{aligned} y_1 \geq 0 &\implies u_2 \leq u_1 \\ y_1 \leq 1 &\implies u_2 \geq -2 + u_1 \\ y_2 \geq 0 &\implies u_2 \geq -u_1 \\ y_2 \leq 1 &\implies u_2 \leq 2 - u_1 \end{aligned}$$

et l'ensemble $\mathcal{U}$ prend la forme indiquée sur la figure suivante:



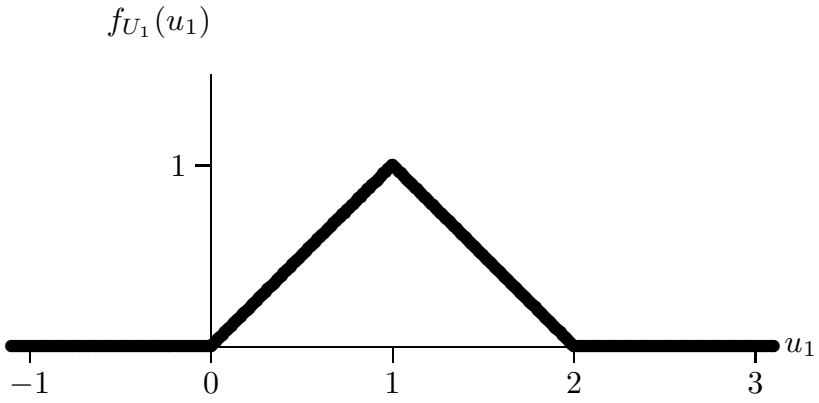
Quelle est alors la densité d'une somme de variables uniformes?

Pour calculer la densité de $Y_1 + Y_2$, il suffit de calculer la densité marginale de U_1; on voit sur la figure que si $0 \leq u_1 \leq 1$, la densité $f_{U_1,U_2}(u_1, u_2)$ est non nulle pour $-u_1 \leq u_2 \leq u_1$. Si $1 \leq u_1 \leq 2$, la densité est non nulle pour $-2 + u_1 \leq u_2 \leq 2 - u_1$.

Donc:

$$\begin{aligned} f_{U_1}(u_1) &= \int_{-u_1}^{u_1} \frac{1}{2} du_2 = \left[\frac{1}{2}u_2\right]_{-u_1}^{u_1} = u_1 \quad \text{pour} \quad 0 \leq u_1 \leq 1 \\ f_{U_1}(u_1) &= \int_{-2+u_1}^{2-u_1} \frac{1}{2} du_2 = \left[\frac{1}{2}u_2\right]_{-2+u_1}^{2-u_1} \\ &= \frac{2 - u_1}{2} - \frac{-2 + u_1}{2} = 2 - u_1 \quad \text{pour} \quad 1 \leq u_1 \leq 2 \quad . \end{aligned}$$

La densité marginale de $U_1 = Y_1 + Y_2$ a donc la forme triangulaire suivante:



2.3 La fonction génératrice des moments

Définition

Soit X une variable aléatoire. Si $E\left(e^{tX}\right)$ existe pour t dans un voisinage ouvert de zéro, la fonction génératrice des moments de X est définie comme:

$$m_X(t) = E\left(e^{tX}\right)$$

Utilité

$m_X(t)$ permet de calculer facilement les moments de X; la fonction génératrice des moments permet en outre, dans certains cas, de calculer facilement la distribution d'une somme de variables aléatoires indépendantes.

Propriétés

(1) $\dfrac{d^r}{dt^r} m_X(0) = E(X^r)$

- En effet:

$$\frac{d}{dt}E\left[e^{tX}\right] = E\left[\frac{d}{dt}e^{tX}\right] = E\left[Xe^{tX}\right] = E(X) \quad \text{si} \quad t = 0 \quad .$$

- De même:

$$\frac{d^2}{dt^2}E\left[e^{tX}\right] = E\left[\frac{d^2}{dt^2}e^{tX}\right] = E\left[X^2e^{tX}\right] = E\left(X^2\right) \quad \text{si} \quad t = 0 \quad , \quad \text{etc.}$$

(2) Si $m_X(t) = m_Y(t)$ pour tout t dans un voisinage ouvert de $t = 0$, alors $F_X(x) = F_Y(y)$ pour $x = y$

(3) Si X et Y sont indépendantes, alors $m_{X+Y}(t) = m_X(t)m_Y(t)$. En effet:

$$E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E\left(e^{tX}\right)E\left(e^{tY}\right) \quad .$$

Exemple: calcul de la fonction génératrice des moments d'une variable normale.

Soit $X \sim N\left(\mu, \sigma^2\right)$,

$$m_X(t) = E\left(e^{tX}\right) = e^{t\mu} E\left(e^{t(X-\mu)}\right) = e^{t\mu} \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{t(x-\mu)} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$
$$= e^{t\mu} \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left[(x-\mu)^2 - 2\sigma^2 t(x-\mu)\right]\right\} dx \quad .$$

Noter que

$$(x-\mu)^2 - 2\sigma^2 t(x-\mu) = (x-\mu)^2 - 2\sigma^2 t(x-\mu) + \sigma^4 t^2 - \sigma^4 t^2$$
$$= \left(x - \mu - \sigma^2 t\right)^2 - \sigma^4 t^2 \quad .$$

Donc:

$$m_X(t) = e^{t\mu} e^{\sigma^2 t^2/2} \underbrace{\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu-\sigma^2 t)^2} dx}_{=1 \text{ car intégrale d'une densité } N(\mu+\sigma^2 t, \sigma^2)}$$
$$m_X(t) = e^{t\mu + \sigma^2 t^2/2} \quad .$$

Exemple d'application: calcul des deux premiers moments $E(X)$ et $V(X)$ d'une variable normale.

Si $X \sim N(\mu, \sigma^2)$, on a vu que $m_X(t) = e^{t\mu + \frac{\sigma^2 t^2}{2}}$. Alors:

$$\frac{d}{dt} m_X(t) = \left(\mu + \sigma^2 t\right) e^{t\mu + \frac{\sigma^2 t^2}{2}} \quad \Longrightarrow \quad m'_X(0) = \mu = E(X)$$

$$\frac{d^2}{dt^2} m_X(t) = \sigma^2 e^{t\mu + \frac{\sigma^2 t^2}{2}} + \left(\mu + \sigma^2 t\right)^2 e^{t\mu + \frac{\sigma^2 t^2}{2}}$$

$$\Longrightarrow \quad m''_X(0) = \sigma^2 + \mu^2 = E(X^2)$$

$$\Longrightarrow \quad V(X) = E(X^2) - E^2(X)$$

$$= \sigma^2 + \mu^2 - \mu^2 = \sigma^2 \quad .$$

On peut, de manière analogue, calculer tous les moments de X.

Note: il existe des tables des fonctions génératrices des moments des variables les plus courantes; voir l'appendice B de Mood, Graybill, Boes, *Introduction to the Theory of Statistics*, 1974.

Exercice: Soit X une variable aléatoire ayant la distribution normale réduite $N(0,1)$. Montrez que $E(X^3) = 0$ et que $E(X^4) = 3$.

Autre exemple d'application: calcul de la distribution d'une somme de variables normales indépendantes.

Soit $X \sim N\left(\mu_x, \sigma_x^2\right)$ et $Y \sim N\left(\mu_y, \sigma_y^2\right)$ et supposons X et Y indépendantes.

$$\begin{aligned} m_{X+Y}(t) &= m_X(t)\, m_Y(t) \quad \text{(Propriété 3)} \\ &= e^{t\mu_x + \sigma_x^2 t^2/2}\, e^{t\mu_y + \sigma_y^2 t^2/2} \\ &= e^{t(\mu_x+\mu_y)+(\sigma_x^2+\sigma_y^2)t^2/2} \end{aligned}$$

$m_{X+Y}(t)$ est donc la fonction génératrice des moments d'une variable distribuée selon $N\left(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2\right)$. En vertu de la propriété 2, la distribution de $Z = X + Y$ est donc une normale de paramètres $\mu_x + \mu_y$ et $\sigma_x^2 + \sigma_y^2$.

Il est beaucoup plus facile de prouver le résultat de cette manière que par l'utilisation du théorème de changement de variables.

2.4 Fonctions de variables normales

(1) Toute combinaison linéaire de variables normales indépendantes est normale:

$$X_j \sim N\left(\mu_j, \sigma_j^2\right) \quad \text{indépendantes} \quad (j = 1, \ldots, n)$$

$$a_j \quad \text{constantes en probabilité} \quad (j = 1, \ldots, n)$$

$$\Longrightarrow \quad \sum_{j=1}^{n} a_j X_j \sim N\left(\sum_{j=1}^{n} a_j \mu_j \quad , \quad \sum_{j=1}^{n} a_j^2 \sigma_j^2\right)$$

(2) Variable Chi-Carré:

$$X_j \sim N(0,1) \quad \text{indépendantes} \quad (j = 1, \ldots, k)$$

$$\Longrightarrow \quad Y = \sum_{j=1}^{k} X_j^2 \sim \chi_k^2$$

(3) Variable t de Student:

$$X \sim N(0,1) \quad ; \quad Y \sim \chi^2_k \quad ; \quad X \quad \text{et} \quad Y \quad \text{indépendantes}$$

$$\Longrightarrow \quad Z = \frac{X}{\sqrt{Y/k}} \sim t_k$$

(4) Variable F de Fisher-Snedecor

$$X \sim \chi^2_k \quad ; \quad Y \sim \chi^2_r \quad ; \quad X \quad \text{et} \quad Y \quad \text{indépendantes}$$

$$\Longrightarrow \quad Z = \frac{X/k}{Y/r} \sim F_{k,r} \quad .$$

Notes sur ce qui précède

(1) La densité de Student est symétrique autour de 0. Elle tend vers la densité $N(0,1)$ lorsque $k \longrightarrow \infty$. Ses deux premiers moments n'existent que si $k > 2$.

(2) La densité de Fisher-Snedecor tend vers la densité d'une variable χ^2_k/k lorsque r, le nombre de degrés de liberté au dénominateur, tend vers l'infini.

(3) Les expressions des densités χ^2, Student, et Fisher peuvent être trouvées dans la littérature, notamment l'ouvrage de Mood, Graybill, Boes (en tête des tables). Elles sont compliquées et nous n'en ferons pas usage dans la première partie du cours. Elles sont obtenues à l'aide du théorème de changement de variables vu précédemment.

(4) Nos définitions précédentes permettent d'engendrer des réalisations simulées des variables en question.

Exercice. Supposons que vous disposiez d'un logiciel permettant d'engendrer des réalisations simulées de variables aléatoires normales réduites indépendantes. Comment pourriez-vous engendrer des réalisations simulées d'une variable ayant une distribution de Student avec k degrés de liberté?

CHAPITRE III

ESTIMATION PONCTUELLE

3.1 Échantillon aléatoire, estimateur, estimation

Échantillon aléatoire

Suite de variables aléatoires indépendantes ayant la même distribution (i.i.d.)

Exemple

Tailles de 100 étudiants de première année, distribuées $N(\mu, \sigma^2)$ et indépendantes: $(X_i, i = 1, \ldots, 100)$.

Estimateur

Fonction de variables aléatoires observables, ne dépendant pas de paramètres inconnus.

Exemple

$$\begin{aligned} \hat{\mu} &= \frac{\sum_{i=1}^{100} X_i}{100} \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^{100} (X_i - \hat{\mu})^2}{100} \end{aligned}$$

Estimation

Valeur prise par une telle fonction pour des réalisations particulières des variables aléatoires, soit $x_1, x_2, \ldots$

Exemple

$$\hat{\mu} = 175, \quad \hat{\sigma}^2 = 25$$

3.2 Fonction de vraisemblance

Soit $(x_1, \ldots, x_n)$ des réalisations des variables aléatoires $X_1, \ldots, X_n$.

Soit $f_X(x_1, \ldots, x_n \mid \theta_1, \ldots, \theta_k)$ la densité jointe de ces variables au point $(x_1, \ldots, x_n)$; cette densité dépend des paramètres inconnus $\theta_1, \ldots, \theta_k$. Si l'on considère cette densité jointe comme une fonction des paramètres inconnus, on l'appelle fonction de vraisemblance et l'écrit:

$$L(\theta_1, \ldots, \theta_k ; x_1, \ldots, x_n) \quad \text{ou plus simplement} \quad L(\theta_1, \ldots, \theta_k) \quad .$$

Note

Les observations x_i sont ici des *paramètres* de la vraisemblance; en d'autres termes, la vraisemblance n'est définie *qu'après* l'observation des réalisations des variables! La vraisemblance est donc une notion statistique, tandis que la densité jointe est une notion probabiliste.

3.3 Maximum de vraisemblance

Principe

On choisit comme estimations des θ_i les valeurs de ces paramètres qui maximisent $L(\theta_1, \ldots, \theta_k)$.

Interprétation dans le cas discret

On choisit comme estimations les valeurs des θ_i qui donnent la plus grande probabilité d'avoir obtenu le résultat expérimental $(x_1, \ldots, x_n)$.

Exemple 1

Une boîte contient 3 boules, qui peuvent être soit rouges, soit blanches. Le nombre de boules rouges est inconnu. On tire deux boules sans remise. On obtient 2 boules rouges. On demande d'estimer le nombre n de boules rouges que contient la boîte à l'aide du principe du maximum de vraisemblance.

Solution

La vraisemblance est donnée dans ce cas par la probabilité d'obtenir le résultat expérimental observé (tirage de 2 boules rouges), considérée comme fonction des quatre valeurs possibles du paramètre inconnu ($n = 0, 1, 2, 3$).

$$\begin{aligned} L(0) &= P(R_1 \cap R_2 \mid n=0) = 0 \\ L(1) &= P(R_1 \cap R_2 \mid n=1) = 0 \\ L(2) &= P(R_1 \cap R_2 \mid n=2) \\ &= P(R_2 \mid R_1, n=2)\, P(R_1 \mid n=2) \\ &= \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3} \\ L(3) &= P(R_1 \cap R_2 \mid n=3) = 1 \quad . \end{aligned}$$

Donc l'estimation est $\hat{n} = 3$.

Exemple 2

On demande d'estimer par maximum de vraisemblance le paramètre p d'une loi binomiale $Bi(n,p)$.

Rappel

$$\begin{aligned} n &= \text{nombre d'essais indépendants} \\ p &= \text{probabilité de succès lors de chaque essai} \\ Y &= \text{nombre de succès est} \quad Bi(n,p) \\ P(Y=r) &= C_n^r p^r (1-p)^{n-r} \end{aligned}$$

Solution

On peut écrire:

$$Y = \sum_{i=1}^{n} X_i \quad \text{où} \quad \begin{aligned} X_i &= 1 \quad \text{si l'essai } i \text{ donne un succès} \\ X_i &= 0 \quad \text{sinon} \quad . \end{aligned}$$

- On observe les réalisations $(x_1, \ldots, x_n)$. Le nombre de succès observé est $r = \sum_{i=1}^{n} x_i$
- On a:

$$f(x_1, \ldots, x_n \mid p) = p^r (1-p)^{n-r} \quad \text{(car l'ordre des réalisations est donné)}$$

- En considérant cette densité comme une fonction du paramètre inconnu p, on a:

$$L(p) = p^r (1-p)^{n-r}$$

- Pour maximiser cette fonction, il est commode de maximiser son logarithme:

$$\begin{aligned}
\log L(p) &= r\log p + (n-r)\log(1-p) \\
\frac{d\log L}{dp} &= \frac{r}{p} - \frac{n-r}{1-p} = 0 \\
&\Longrightarrow \quad \frac{r}{p} = \frac{n-r}{1-p} \Longrightarrow \quad \frac{1-p}{p} = \frac{n-r}{r} \\
&\Longrightarrow \quad \frac{1}{p} - 1 = \frac{n}{r} - 1 \quad \Longrightarrow \quad \hat{p} = \frac{r}{n} \quad .
\end{aligned}$$

- On estime donc p par le pourcentage des succès observés. On a bien un maximum car:

$$\frac{d^2\log L}{dp^2} = -\frac{r}{p^2} - \frac{n-r}{(1-p)^2} < 0 \quad .$$

Exemple 3

- On demande d'estimer par maximum de vraisemblance les paramètres μ et σ^2 d'une loi normale à partir d'un échantillon aléatoire $(X_i, i = 1, \ldots, n)$.
- On a, par définition de la densité normale:

$$f_{X_i}(x_i) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \quad .$$

- En vertu de l'indépendance:

$$f_X\left(x_1, \ldots, x_n \mid \mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \quad .$$

- En considérant cette fonction comme fonction des paramètres inconnus:

$$L\left(\mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

qui est à maximiser par rapport à μ et σ^2.

Les conditions de premier ordre s'écrivent:

$$(1)\quad \frac{\partial \log L}{\partial \mu} = \frac{2}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$(2)\quad \frac{\partial \log L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$(1) \quad \Longrightarrow \quad \sum_{i=1}^{n} x_i = n\mu, \quad \text{donc} \quad \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$(2) \quad \Longrightarrow \quad -n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\Longrightarrow \quad \sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$$

$$\Longrightarrow \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \quad \text{en remplaçant} \quad \mu \quad \text{par} \quad \hat{\mu} \quad .$$

Exercice: vérifier que l'on a bien un maximum.

Note: Par la suite, nous utiliserons toujours $\hat{\sigma}^2$ pour désigner l'estimateur de σ^2 par maximum de vraisemblance. Un autre estimateur, que nous désignerons par s^2, sera vu au début du chapitre suivant.

CHAPITRE IV

PROPRIÉTÉS DES ESTIMATEURS

4.1 Estimateur sans biais

Définition:

Un estimateur $\hat{\theta}$ de θ est dit *sans biais* si $E(\hat{\theta}) = \theta$.

Exemple:

Soit un échantillon aléatoire $(X_i, i = 1, \ldots, n)$ avec $E(X_i) = \mu$ pour tout i et $V(X_i) = \sigma^2$ pour tout i. On va montrer que:

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

et

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

sont sans biais.

En ce qui concerne la moyenne:

$$E\left(\bar{X}\right) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n} E\left(\sum X_i\right) = \frac{1}{n} \sum E\left(X_i\right) = \frac{1}{n} n\mu = \mu$$

En ce qui concerne la variance, notons que:

$$E\left[\sum \left(X_i - \bar{X}\right)^2\right] = E\left[\sum X_i^2 - n\bar{X}^2\right] = E\left(\sum X_i^2\right) - \frac{E(\sum X_i)^2}{n}$$

et que:

$$E\left(\sum X_i^2\right) = \sum E\left(X_i^2\right) = \sum \left(\sigma^2 + \mu^2\right) = n\left(\sigma^2 + \mu^2\right)$$

car $\sigma^2 = E\left(X_i^2\right) - \mu^2$, et donc $E\left(X_i^2\right) = \sigma^2 + \mu^2$.

D'autre part:

$$\begin{aligned} E\left(\sum_{i=1}^{n} X_i\right)^2 &= E\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_i X_j\right] \\ &= \sum_{i=1}^{n} E\left(X_i^2\right) + 2\underbrace{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E\left(X_i X_j\right)}_{n(n-1)/2 \quad \text{termes}} \end{aligned}$$

Mais $E\left(X_i^2\right) = \sigma^2 + \mu^2$, et, par l'indépendance:

$$E\left(X_i X_j\right) = E\left(X_i\right) E\left(X_j\right) = \mu^2 \quad .$$

Donc:

$$\begin{aligned} E\left(\sum_{i=1}^{n} X_i\right)^2 &= n\left(\sigma^2 + \mu^2\right) + \frac{2n(n-1)}{2} \quad \mu^2 \\ &= n\sigma^2 + n\mu^2 + n^2\mu^2 - n\mu^2 = n\left(\sigma^2 + n\mu^2\right) \quad . \end{aligned}$$

Donc $\dfrac{E(\sum X_i)^2}{n} = \sigma^2 + n\mu^2$, et:

$$\begin{aligned} E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] &= E\left(\sum_{i=1}^{n} X_i^2\right) - \frac{E(\sum X_i)^2}{n} \\ &= n\left(\sigma^2 + \mu^2\right) - \sigma^2 - n\mu^2 = (n-1)\,\sigma^2 \quad . \end{aligned}$$

Donc:

$$\begin{aligned} E\left(s^2\right) &= E\left[\frac{\sum\left(X_i - \bar{X}\right)^2}{n-1}\right] = \frac{1}{n-1}(n-1)\,\sigma^2 \\ &= \sigma^2 \quad , \end{aligned}$$

ce qui montre que s^2 est sans biais.

4.2 Estimateur convergent

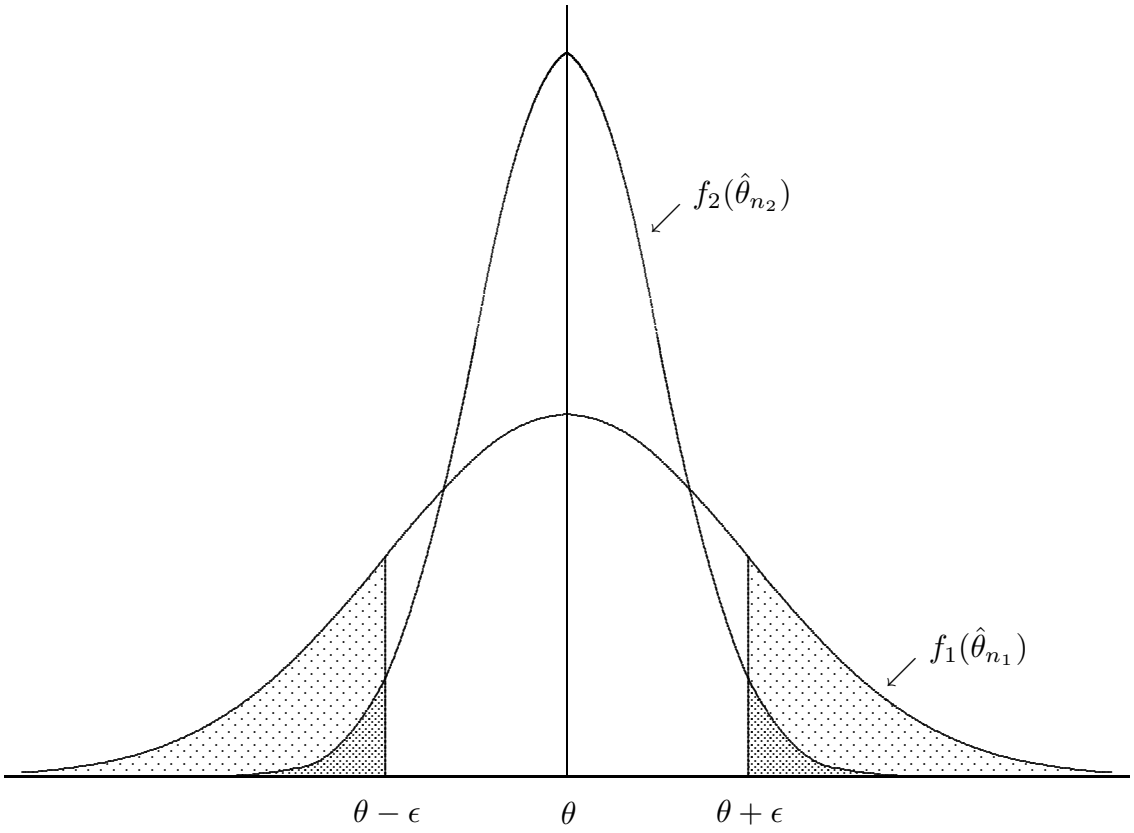
Définition

Un estimateur $\hat{\theta}_n$ de θ est dit convergent si et seulement si:

$$\lim_{n \longrightarrow \infty} P\left[\mid \hat{\theta}_n - \theta \mid > \epsilon\right] = 0 \quad \text{pour tout} \quad \epsilon > 0; \quad \text{on écrit} \quad \operatorname{plim} \hat{\theta}_n = \theta \quad .$$

Interprétation

Si $\hat{\theta}_n$ possède une densité $f(\hat{\theta}_n)$, la probabilité $P[|\ \hat{\theta}_n - \theta\ | > \epsilon]$ est la zone hachurée de la figure suivante:



Cette probabilité doit tendre vers 0 lorsque n tend vers l'infini; ceci sera le cas si les densités deviennent de plus en plus concentrées autour de θ.

Conditions suffisantes

Si $\lim_{n\longrightarrow\infty} E(\hat{\theta}_n) = \theta$ et si $\lim_{n\longrightarrow\infty} V(\hat{\theta}_n) = 0$, alors $\operatorname{plim} \hat{\theta}_n = \theta$. Ceci sera démontré au chapitre X de la deuxième partie.

Exemple

Si $(X_i, i = 1, \ldots, n)$ est un échantillon aléatoire avec $E(X_i) = \mu$, $V(X_i) = \sigma^2$, alors $\operatorname{plim} \bar{X} = \mu$, car:

$$\begin{aligned} E\left(\bar{X}\right) &= \mu \\ V\left(\bar{X}\right) &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\sigma^2\right) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n} \longrightarrow 0 \quad . \end{aligned}$$

Note

Contrairement à l'absence de biais qui est une propriété de petit échantillon (valable pour tout n), la convergence est une propriété asymptotique (valable si $n \longrightarrow \infty$).

4.3 Estimateur efficace

Un estimateur efficace est un estimateur sans biais, et de variance minimale parmi tous les estimateurs sans biais.

Définition

$$\hat{\theta} \quad \text{est efficace:} \quad \begin{cases} E(\hat{\theta}) = \theta \\ \\ V(\hat{\theta}) \leq V(\tilde{\theta}) \quad \text{si} \quad E(\tilde{\theta}) = \theta \end{cases} .$$

Interprétation

La variance d'un estimateur est une mesure de l'imprécision de notre estimation de la vraie valeur du paramètre. Un estimateur sans biais, mais de variance énorme, est inutile: on ne se trompe pas en moyenne, mais on peut se tromper énormément dans des cas individuels, c.a.d. pour certains échantillons. Il est donc important que la variance soit la plus petite possible.

Exemple

Nous prouverons au chapitre X de la seconde partie que si les X_i sont normales i.i.d., alors $\bar{X}$ est efficace.

4.4 Minimisation de l'erreur quadratique moyenne

- Que faire si l'on doit choisir entre un estimateur sans biais mais de grande variance, ou un estimateur un peu biaisé mais de petite variance?
- Réponse: on peut minimiser l'erreur quadratique moyenne:

$$EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$
$$\Big(\text{Si} \quad \hat{\theta} \quad \text{est sans biais,} \quad EQM(\hat{\theta}) = V(\hat{\theta})\Big) \quad .$$

- Justification: On va montrer que:

$$EQM(\hat{\theta}) = V(\hat{\theta}) + \text{Biais}^2(\hat{\theta}) \quad .$$

- En effet:

$$\begin{aligned} EQM(\hat{\theta}) &= E(\hat{\theta}-\theta)^2 \\ &= E\left[\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta\right]^2 \\ &= E\left[\hat{\theta}-E(\hat{\theta})\right]^2 + E\left[E(\hat{\theta})-\theta\right]^2 + 2E\left[\left(\hat{\theta}-E(\hat{\theta})\right)\left(E(\hat{\theta})-\theta\right)\right]. \end{aligned}$$

$$\begin{aligned} \text{Mais} \quad E\left[\left(\hat{\theta}-E(\hat{\theta})\right)\left(E(\hat{\theta})-\theta\right)\right] &= \left[E(\hat{\theta})-\theta\right]E\left[\hat{\theta}-E(\hat{\theta})\right] \\ &= \left[E(\hat{\theta})-\theta\right]\left[E(\hat{\theta})-E\left(E(\hat{\theta})\right)\right] \\ &= \left[E(\hat{\theta})-\theta\right]\left[E(\hat{\theta})-E(\hat{\theta})\right]=0 \quad . \end{aligned}$$

D'autre part:

$$\begin{aligned} E\left[\hat{\theta}-E(\hat{\theta})\right]^2 &= V(\hat{\theta}) \\ E\left[E(\hat{\theta})-\theta\right]^2 &= \left[E(\hat{\theta})-\theta\right]^2 = \text{Biais}^2(\hat{\theta}). \end{aligned}$$

4.5 Interprétation des propriétés

- Il est utile d'illustrer ces propriétés à l'aide d'échantillons fictifs, qui peuvent être obtenus par simulation.
- Supposons donc que l'on ait m échantillons de taille n, permettant de calculer m estimations $\hat{\theta}_i(n)$:

	échantillons		
x_{11}	x_{12}		x_{1m}
$\vdots$	$\vdots$	$\ldots$	$\vdots$
x_{n1}	x_{n2}		x_{nm}
$\hat{\theta}_1(n)$	$\hat{\theta}_2(n)$		$\hat{\theta}_m(n)$

- Si $\hat{\theta}$ est sans biais, on aura en général

$$\lim_{m\longrightarrow\infty}\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i(n)=\theta \quad \text{pour tout } n \quad .$$

- Si $\hat{\theta}$ est efficace, on aura en général

$$\lim_{m \longrightarrow \infty} \frac{1}{m} \sum_{i=1}^{m} \left(\hat{\theta}_i(n) - \bar{\hat{\theta}}(n) \right)^2 \quad \text{minimale pour tout } n \quad .$$

- Si $\hat{\theta}$ minimise l'EQM, on aura en général

$$\lim_{m \longrightarrow \infty} \frac{1}{m} \sum_{i=1}^{m} \left(\hat{\theta}_i(n) - \theta \right)^2 \quad \text{minimale pour tout } n \quad .$$

- Si $\hat{\theta}$ est convergent, on aura pour tout i:

$$\lim_{n \longrightarrow \infty} P \left[\mid \hat{\theta}_i(n) - \theta \mid > \epsilon \right] = 0 \quad .$$

On fait donc ici tendre n (et non m) vers l'infini.

Remarque: Dans ce contexte, les estimations $\hat{\theta}_i(n)$ sont des nombres *pseudo-aléatoires*, car il s'agit d'une expérience de simulation. La notation "lim" est par conséquent plus appropriée que la notation "plim".

CHAPITRE V

TESTS D'HYPOTHÈSES

5.1 Méthode des intervalles de confiance

Cette méthode est facile à appliquer lorsque l'on possède un estimateur sans biais d'un paramètre inconnu θ (soit $\hat{\theta}$ cet estimateur), et que la densité de $\hat{\theta}$ est symétrique autour de θ (par exemple normale). On cherche alors un intervalle entre les bornes duquel la vraie valeur du paramètre inconnu θ a une certaine probabilité $1-\alpha$ de se situer.

Exemple: construction d'un intervalle de confiance sur l'espérance μ d'une population normale.

- Si la variance σ^2 est connue, on a:

- échantillon $(X_1, \dots, X_n)$; $X_i \sim N(\mu, \sigma^2)$
- Valeurs observées $x_1, \dots, x_n$
- $\bar{x} = \dfrac{\sum_{i=1}^n x_i}{n}$ est une réalisation d'une variable distribuée $N(\mu, \frac{\sigma^2}{n})$
- $\dfrac{\mu - \bar{x}}{\sigma/\sqrt{n}} = \sqrt{n}\left(\dfrac{\mu - \bar{x}}{\sigma}\right)$ est donc une réalisation d'une variable distribuée $N(0,1)$.

Si $Z_{\alpha/2}$ est la valeur de la $N(0,1)$ ayant une probabilité $\alpha/2$ d'être dépassée:

$$P\left[-Z_{\alpha/2} \leq \frac{\mu - \bar{x}}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}\right] = 1-\alpha \quad , \quad \text{donc:}$$

$$P\left[\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha \quad .$$

On a une probabilité de $1-\alpha$ de ne pas se tromper lorsque l'on affirme que μ se situe entre ces 2 bornes.

- Si la variance σ^2 est inconnue, on peut l'estimer par $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$.

On peut écrire:

$$\sqrt{n}\left(\frac{\mu - \bar{x}}{s}\right) = \frac{\sqrt{n}\left(\dfrac{\mu - \bar{x}}{\sigma}\right)}{\sqrt{\dfrac{\sum (x_i - \bar{x})^2}{(n-1)\sigma^2}}}$$

On démontrera plus loin (4.3 de la seconde partie) que $\frac{\sum(x_i - \bar{x})^2}{\sigma^2}$ est distribuée χ^2_{n-1} et est indépendante de $\sqrt{n}\left(\frac{\mu - \bar{x}}{\sigma}\right)$

Alors $\sqrt{n}\left(\frac{\mu - \bar{x}}{s}\right) \sim t_{n-1}$, et l'intervalle de confiance s'écrit:

$$P\left[\bar{x} - t_{n-1;\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1;\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right] = 1 - \alpha$$

On ne rejette pas une hypothèse impliquant que μ soit intérieure aux deux bornes, on rejette une hypothèse impliquant que μ soit extérieure aux deux bornes.

5.2 Méthode générale de construction des tests

On a ici un vecteur de paramètres inconnus $\theta = (\theta_1, \ldots, \theta_k)$. On veut tester: $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ (θ_0 est un vecteur de *nombres*)

Note: rien n'empêche θ d'être une fonction d'un autre vecteur ϕ de paramètres plus fondamentaux; exemple: $k = 1$ et $\theta_1 = \phi_1 - \phi_2$, $H_0 : \theta_1 = 0$ contre $H_1 : \theta_1 \neq 0$.

Procédure de test

Elle doit conduire, soit au rejet de H_0 en faveur de H_1, soit à l'absence de rejet, en tenant compte des deux types d'erreurs possibles:

	Rejeter H_0	Ne pas rejeter H_0
H_0 vraie	Erreur de type I (prob. α)	Décision correcte (prob. $1 - \alpha$)
H_0 fausse	Décision correcte (prob. $1 - \beta$)	Erreur de type II (prob. β)

Les probabilités sont conditionnelles aux événements définissant les lignes!

- On a donc:

$$\begin{aligned} \alpha &= P(\text{rejeter} \quad H_0 \mid H_0 \quad \text{vraie}) = \text{taille du test, ou niveau} \\ \beta &= P(\text{ne pas rejeter} \quad H_0 \mid H_0 \quad \text{fausse}) \end{aligned}$$

- $1 - \beta$ s'appelle la *puissance* du test. C'est la probabilité de déceler la violation de H_0, si H_0 est fausse (probabilité conditionnelle!)
- Malheureusement, on peut montrer qu'il est impossible, en général, de minimiser α et β simultanément. La procédure générale de construction d'un test que l'on va décrire

tient compte de cet état des choses: on va, dès le départ, choisir une valeur faible de α (typiquement 0.01 ou 0.05), et, *pour cette valeur de* α, choisir un test puissant parmi les tests de taille α.

Procédure de construction

Étape 1: on se donne une probabilité α de commettre une erreur de type I (rejeter H_0 si H_0 est vraie).

Étape 2: on choisit une statistique $s(\hat{\theta}, \theta_0)$, à l'aide d'un critère tel que ceux que nous exposerons aux sections 5.3, 5.4, et 5.5. Ces critères conduisent à des tests puissants.

Étape 3: on détermine la distribution conditionnelle de $s(\hat{\theta}, \theta_0)$ sous l'hypothèse H_0, c'est-à-dire si $\theta = \theta_0$.

Étape 4: la probabilité α permet de déterminer une région d'acceptation $R_A(\alpha)$ et une région critique $R_C(\alpha)$:

$$\begin{aligned} R_A(\alpha) &= \{s \mid P(s \in R_A(\alpha) \mid H_0) = 1 - \alpha\} \\ R_C(\alpha) &= \bar{R}_A(\alpha) \quad . \end{aligned}$$

Ces régions peuvent être calculées à l'aide des résultats de l'étape 3, qui nous donne la distribution de $s = s(\hat{\theta}, \theta_0)$ sous H_0!

Étape 5: on décide de rejeter H_0 si $s(\hat{\theta}, \theta_0) \in R_C(\alpha)$.

Notes

(1) Par construction, α est alors bien la probabilité de commettre une erreur de type I (rejeter H_0 si H_0 est vraie) car on a supposé que H_0 était vraie en calculant la distribution conditionnelle de $s(\hat{\theta}, \theta_0)$ à l'étape 3.

(2) La puissance $1 - \beta$ dépend de la vraie valeur (inconnue) de θ, puisqu'elle se calcule conditionnellement à H_1, c'est-à-dire lorsque la valeur de θ n'est pas donnée à priori.

(3) Le fait de ne pas rejeter H_0 ne signifie pas démontrer H_0: cela veut seulement dire que les données ne fournissent pas suffisamment d'informations pour infirmer H_0! Il est donc plus correct de dire "on ne rejette pas H_0" que "on accepte H_0".

(4) Pour l'étape 2, il existe un assez grand nombre de critères. Les trois critères que nous allons exposer sont très employés, sont d'une applicabilité générale, et ont des propriétés d'optimalité sur le plan de la puissance. Dans certains cas les trois critères conduisent à la même statistique. Dans la plupart des cas les trois critères sont asymptotiquement équivalents.

5.3 Le critère du rapport des vraisemblances (LR)

Définition

Le rapport des vraisemblances λ est défini comme:

$$\lambda = \frac{\max_{H_0} L(\theta)}{\max_{\Omega} L(\theta)}$$

où θ est le vecteur de paramètres inconnus de vraisemblance $L(\theta)$. H_0 désigne ici l'ensemble des valeurs de θ compatibles avec l'hypothèse nulle, et Ω désigne l'ensemble de toutes les valeurs admissibles de θ.

Exemple

$$\begin{aligned} \theta &= \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \quad ; \quad H_0 \quad : \quad \left\{ \begin{pmatrix} \mu_0 \\ x \end{pmatrix} \mid x > 0 \right\} \subseteq \mathbb{R} \\ \Omega &= \left\{ \begin{pmatrix} y \\ x \end{pmatrix} \mid x > 0 \right\} \subseteq \mathbb{R}^2 \quad . \end{aligned}$$

Interprétation

- Comme la vraisemblance est une fonction positive, $\lambda \geq 0$,
- Comme un maximum contraint est inférieur à un maximum libre, $\lambda \leq 1$
- Donc $0 \leq \lambda \leq 1$; et:

  si $\lambda \approx 0$, mauvais accord entre l'observation et l'hypothèse H_0
  si $\lambda \approx 1$, bon accord entre l'observation et l'hypothèse H_0 .

- En d'autres termes, si λ est proche de 0 l'hypothèse H_0 ne paraît pas vraisemblable à la lumière des informations fournies par l'échantillon. Donc, on rejettera H_0 si λ est proche de 0.
- Problème: en-dessous de quelle valeur décidera-t-on que λ est suffisamment proche de 0 pour que l'on puisse rejeter H_0? La réponse est fournie par la procédure de test décrite plus haut. On devra choisir λ_α de telle sorte que si l'on rejette H_0 lorsque $\lambda < \lambda_\alpha$, alors la probabilité d'une erreur de type I est précisément égale à α. Le calcul de λ_α nécessite la connaissance de la distribution de λ (ou d'une fonction monotone de λ) conditionnelle à l'hypothèse H_0.

Premier exemple d'application: test sur l'espérance d'une population normale dont la variance est connue

- On a $X_i \sim N(\mu, \sigma^2)$ indépendantes ($i = 1, \cdots, n$), σ^2 connue.
- On veut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.
- On a ici $\theta = \mu$ (un seul paramètre inconnu)
- $L(\mu) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right)$

$$\begin{aligned}
\max_{H_0} L(\mu) &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu_0)^2\right) \\
\max_{\Omega} L(\mu) &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\bar{x})^2\right) \\
\lambda &= \frac{\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu_0)^2\right)}{\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\bar{x})^2\right)} \\
&= \exp\left[-\frac{1}{2\sigma^2}\left(\sum(x_i-\mu_0)^2 - \sum(x_i-\bar{x})^2\right)\right] \quad .
\end{aligned}$$

- Notons que $\sum(x_i-\mu_0)^2 = \sum(x_i-\bar{x})^2 + n(\bar{x}-\mu_0)^2$.
  En effet:

$$\sum(x_i-\mu_0)^2 = \sum(x_i-\bar{x}+\bar{x}-\mu_0)^2 = \sum(x_i-\bar{x})^2 + n(\bar{x}-\mu_0)^2 + \underbrace{2\sum(x_i-\bar{x})(\bar{x}-\mu_0)}_{=0}$$

- Donc:

$$\begin{aligned}
\lambda &= \exp\left[-\frac{1}{2\sigma^2}\left(\sum(x_i-\bar{x})^2 + n(\bar{x}-\mu_0)^2 - \sum(x_i-\bar{x})^2\right)\right] \\
&= \exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu_0)^2\right].
\end{aligned}$$

- Une fonction monotone de λ est donnée par:

$$-2\log\lambda = \frac{(\bar{x}-\mu_0)^2}{\sigma^2/n} \underset{\text{def}}{=} LR$$

($LR = -2\log\lambda$ s'appelle la *statistique* du rapport des vraisemblances)

- Si H_0 est vraie ($\mu = \mu_0$), LR est le carré d'une normale réduite! On a donc trouvé la distribution d'une fonction monotone de λ sous H_0.

Conclusion

- On a: $-2\log\lambda = \dfrac{(\bar{x}-\mu_0)^2}{\sigma^2/n}$
- On définit: $Z_{\text{obs}} = \dfrac{(\bar{x}-\mu_0)}{\sigma/\sqrt{n}}$
- Si on décide de rejeter $H_0 : \mu = \mu_0$ lorsque $Z_{\text{obs}} > Z_{\alpha/2}$ ou $Z_{\text{obs}} < -Z_{\alpha/2}$, α sera bien la probabilité d'une erreur de type I puisque $Z_{\text{obs}} \sim N(0,1)$ sous H_0.
- De façon équivalente, on rejette H_0 si $\lambda < \lambda_\alpha$ où λ_α est défini implicitement par $-2\log\lambda_\alpha = Z^2_{\alpha/2}$ (soit $\lambda_\alpha = \exp\left(-\frac{1}{2}Z^2_{\alpha/2}\right)$).

Exercice. Calculez, en fonction de μ, la puissance du test précédent lorsque α, μ_0, σ^2, et n sont donnés. Comment cette fonction de puissance se comporte-t-elle lorsque la taille n de l'échantillon tend vers l'infini?

Second exemple d'application: test sur l'espérance d'une population normale, variance inconnue

- On a toujours $X_i \sim N(\mu,\sigma^2)$ indépendantes pour $i = 1,\ldots,n$; mais σ^2 est inconnue.
- Le test est toujours $H_0 : \mu = \mu_0$ contre $H_0 : \mu \neq \mu_0$
- Ici, $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$
- Sous H_0 : la maximisation de L implique $\hat{\mu}_0 = \mu_0$ et $\hat{\sigma}_0^2 = \sum_{i=1}^n (x_i-\mu_0)^2/n$.
- Sous Ω: la maximisation de L implique $\hat{\mu} = \bar{x}$ et $\hat{\sigma}^2 = \sum_{i=1}^n (x_i-\bar{x})^2/n$ comme on l'a vu.
- Le rapport des vraisemblances s'énonce comme:

$$\begin{aligned}\lambda &= \frac{\left(2\pi\hat{\sigma}_0^2\right)^{-n/2}\exp\left[-\frac{1}{2\hat{\sigma}_0^2}\sum(x_i-\mu_0)^2\right]}{\left(2\pi\hat{\sigma}^2\right)^{-n/2}\exp\left[-\frac{1}{2\hat{\sigma}^2}\sum(x_i-\bar{x})^2\right]} \\ &= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{-n/2}, \quad \text{puisque:} \\ \sum(x_i-\mu_0)^2 &= n\hat{\sigma}_0^2 \quad ; \quad \sum(x_i-\bar{x})^2 = n\hat{\sigma}^2.\end{aligned}$$

- On a vu que:

$$\sum(x_i-\mu_0)^2 = \sum(x_i-\bar{x})^2 + n(\bar{x}-\mu_0)^2 \quad .$$

- En substituant plus haut:

$$\lambda = \left(1 + \frac{n\left(\bar{x} - \mu_0\right)^2}{\sum\left(x_i - \bar{x}\right)^2}\right)^{-n/2} \quad , \quad \text{donc :}$$

$$(n-1)\left(\lambda^{-2/n} - 1\right) = \frac{\left(\bar{x} - \mu_0\right)^2}{s^2/n} \quad \text{avec} \quad s^2 = \frac{\sum\left(x_i - \bar{x}\right)^2}{n-1} \quad .$$

- On reconnaît le carré d'une variable de student avec $n-1$ degrés de liberté sous H_0. On a donc de nouveau trouvé la distribution d'une fonction monotone de λ sous H_0.

Conclusion

- On définit $t_{\text{obs}} = \dfrac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$
- On a $(n-1)\left(\lambda^{-2/n} - 1\right) = t^2_{\text{obs}}$, soit aussi: $\lambda = \left[1 + \frac{t^2_{\text{obs}}}{n-1}\right]^{-n/2}$
- Si on décide de rejeter H_0 lorsque $t_{\text{obs}} > t_{n-1,\frac{\alpha}{2}}$, ou $t_{\text{obs}} < -t_{n-1,\frac{\alpha}{2}}$, α sera bien la probabilité de commettre une erreur de type I puisque $t_{\text{obs}} \sim t_{n-1}$ sous H_0.
- De façon équivalente, on rejette H_0 si $\lambda < \lambda_\alpha$, où:

$$\lambda_\alpha = \left[1 + \frac{t^2_{n-1,\frac{\alpha}{2}}}{n-1}\right]^{-n/2}$$

5.4. Le critère de Wald

Nous n'énoncerons ici ce critère que pour le test d'une seule hypothèse, car la généralisation aux tests joints sera vue plus tard.

Définition

Soit $L(\theta) = L(\theta_1, \cdots, \theta_k)$ la vraisemblance et soit $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ l'estimation de θ qui maximise $L(\theta)$. On s'intéresse au test:

$$H_0 : \theta_i = \theta_0 \text{ contre } H_1 : \theta_i \neq \theta_0$$
(θ_i est un élément de θ, θ_0 *est un nombre*)

La statistique de Wald est définie comme:

$$\mathcal{W} = \frac{(\hat{\theta}_i - \theta_0)^2}{\hat{V}(\hat{\theta}_i)} \quad ,$$

où $\hat{V}(\hat{\theta}_i)$ est l'estimation de la variance de $\hat{\theta}_i$ obtenue par maximisation de la vraisemblance.

Note: la vraisemblance est maximisée *sans* contraintes!

Interprétation

Il s'agit du carré d'une distance entre l'estimation de θ_i sous H_0 (à savoir θ_0) et l'estimation de θ_i sous H_1 (à savoir $\hat{\theta}_i$). On divise par la variance estimée pour tenir compte de la précision de l'estimation.

Exemple

- Soit $L(\mu, \sigma^2)$ la vraisemblance précédente (population normale, variance inconnue). Pour tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, on forme:

$$\mathcal{W} = \frac{(\hat{\mu} - \mu_0)^2}{\hat{V}(\hat{\mu})} \quad = \quad \frac{(\bar{x} - \mu_0)^2}{\hat{\sigma}^2/n}$$

où $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ est l'estimation de σ^2 par maximum de vraisemblance.

- Comme précédemment, on peut transformer la statistique $\mathcal{W}$ en une autre statistique possédant une distribution connue sous H_0, à l'aide d'une transformation monotone. En effet, comme $\hat{\sigma}^2 = \frac{n-1}{n}s^2$, on a:

$$\mathcal{W} = \frac{(\bar{x} - \mu_0)^2}{\frac{n-1}{n}\frac{s^2}{n}} = \frac{n}{n-1} \quad t^2_{\text{obs}}$$

et le critère de Wald conduit donc, dans ce cas-ci, au même test que le critère du rapport des vraisemblances (le test t).

5.5. Le critère des multiplicateurs de Lagrange

De nouveau, nous énoncerons ce critère pour le test d'une seule hypothèse; la généralisation aux tests joints sera vue plus tard.

Soit $\mathcal{L}(\theta) = \mathcal{L}(\theta_1, \ldots, \theta_k)$ la vraisemblance logarithmique $\mathcal{L} = \log_e L$. On s'intéresse au test:

$$H_0 : \theta_i = \theta_0 \text{ contre } H_1 : \theta_i \neq \theta_0 \quad .$$

Soit $\hat{\theta}_0$ l'estimation de θ par maximisation de la vraisemblance *sous la contrainte* H_0. $\hat{\theta}_0$ est obtenu en annulant les dérivées du Lagrangien:

$$\Lambda(\theta, \lambda) = \mathcal{L}(\theta) - \lambda(\theta_i - \theta_0).$$

Dans un modèle linéaire et pour des observations distribuées normalement, on peut montrer que la statistique du multiplicateur de Lagrange est égale à:

$$LM = \frac{\hat{\lambda}_0^2}{\hat{V}_0(\lambda)}$$

où $\hat{\lambda}_0$ est la valeur de λ évaluée au point $\theta = \hat{\theta}_0$ et où $\hat{V}_0(\lambda)$ est l'estimation de $V(\lambda)$ obtenue par maximisation de $\mathcal{L}$ sous H_0.

Interprétation

L'annulation de la dérivée de Λ par rapport à θ_i implique:

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \lambda$$

ce qui montre que $\hat{\lambda}_0$ est le taux de variation de la vraisemblance maximisée $\mathcal{L}(\hat{\theta}_0)$ lorsque l'on s'éloigne de la situation contrainte. Si ce taux de variation est nul, le fait de relâcher H_0 ne modifie pas la vraisemblance contrainte: cette contrainte n'apparaît donc pas comme significative.

Exemple

Soit $\mathcal{L}\left(\mu, \sigma^2\right)$ la vraisemblance logarithmique précédente:

$$\mathcal{L}\left(\mu, \sigma^2\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \quad .$$

On a vu que:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right) = \frac{n\left(\bar{x} - \mu\right)}{\sigma^2} \\ &= \lambda \quad \text{(par l'annulation de la dérivée de} \quad \Lambda) \end{aligned}$$

Donc:

$$\hat{\lambda}_0 = \left.\frac{\partial \mathcal{L}}{\partial \mu}\right|_{\mu=\mu_0, \sigma^2=\hat{\sigma}_0^2} = \frac{n\left(\bar{x} - \mu_0\right)}{\hat{\sigma}_0^2}$$

$$\text{où} \quad \hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_0\right)^2 \quad .$$

Par ailleurs:

$$V(\lambda) = \frac{1}{\sigma^4} V\left(\sum_{i=1}^{n} x_i\right) = \frac{n\sigma^2}{\sigma^4} = \frac{n}{\sigma^2} \quad , \quad \text{donc} \quad \hat{V}_0(\lambda) = \frac{n}{\hat{\sigma}_0^2} \quad .$$

$$\text{Donc} \quad LM = \frac{\dfrac{n^2 (\bar{x} - \mu_0)^2}{\hat{\sigma}_0^4}}{\dfrac{n}{\hat{\sigma}_0^2}} = \frac{n (\bar{x} - \mu_0)^2}{\hat{\sigma}_0^2} \quad .$$

Comme précédemment, on peut appliquer une transformation monotone à LM pour obtenir t_{obs}^2. En effet:

$$\begin{aligned} \hat{\sigma}_0^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2 \\ &= \frac{1}{n} \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 + n (\bar{x} - \mu_0)^2 \right] \\ &= \hat{\sigma}^2 + (\bar{x} - \mu_0)^2 \quad . \end{aligned}$$

Donc:

$$\begin{aligned} \frac{1}{LM} &= \frac{\hat{\sigma}_0^2}{n (\bar{x} - \mu_0)^2} = \frac{\hat{\sigma}^2 + (\bar{x} - \mu_0)^2}{n (\bar{x} - \mu_0)^2} \\ &= \frac{1}{n} + \frac{\hat{\sigma}^2}{n (\bar{x} - \mu_0)^2} = \frac{1}{n} + \frac{\frac{n-1}{n} s^2}{n (\bar{x} - \mu_0)^2} \\ &= \frac{1}{n} + \frac{n-1}{n} \frac{1}{t_{\text{obs}}^2} = \frac{t_{\text{obs}}^2 + n - 1}{n t_{\text{obs}}^2} \quad . \end{aligned}$$

Soit aussi:

$$LM = \frac{n t_{\text{obs}}^2}{t_{\text{obs}}^2 + n - 1} \quad .$$

5.6 Comparaison des trois critères

Rappelons que $LR = -2\log\lambda$.

- Pour le test vu précédemment:

$$H_0 : \mu = \mu_0 \text{ contre } H_1 : \mu \neq \mu_0$$
$$\text{observations } x_i \sim N(\mu, \sigma^2) \text{ indépendantes, } \sigma^2 \text{ inconnue,}$$

on a établi que:

$$\mathcal{W} = \frac{n}{n-1} t^2_{\text{obs}}$$
$$\frac{1}{LM} = \frac{1}{n} + \frac{n-1}{n} \frac{1}{t^2_{\text{obs}}}$$
$$LR = n \log\left[1 + \frac{n\left(\bar{x} - \mu_0\right)^2}{\sum \left(x_i - \bar{x}\right)^2}\right] = \log\left[1 + \frac{t^2_{\text{obs}}}{n-1}\right]^n.$$

- On a donc une relation bijective entre t^2_{obs} et chacune des trois statistiques, ce qui veut dire que chacun des trois critères conduit au même test (le test t).

- Il n'en est pas toujours ainsi: dans des situations plus compliquées, les trois statistiques $\mathcal{W}$, LM, et LR ne seront pas des fonctions bijectives les unes des autres, et leurs régions critiques seront différentes *en petit échantillon*.

- En revanche, si $n \longrightarrow \infty$, les distributions des trois statistiques sous H_0 tendront en général vers la même distribution χ^2. Ceci peut se vérifier facilement pour le test que nous venons de voir, puisque $\frac{1}{n} \longrightarrow 0$, $\frac{n-1}{n} \longrightarrow 1$, et $\left(1 + \frac{t^2_{\text{obs}}}{n-1}\right)^n \longrightarrow \exp\left(t^2_{\text{obs}}\right)$. Mais la validité de cette proposition est beaucoup plus générale!

- Quel est alors l'intérêt de l'étude de ces trois statistiques? Il réside dans leur commodité d'emploi. Celle-ci dépend du contexte:

  (a) $\mathcal{W}$ sera plus facile à employer chaque fois que le modèle est plus facile à estimer *sans* contraintes;

  (b) LM sera plus facile à employer chaque fois que le modèle est plus facile à estimer *sous* H_0;

  (c) LR nécessite l'estimation du modèle avec *et* sans contraintes; en revanche, son calcul ne nécessite que la connaissance des valeurs de la vraisemblance maximisée. Aucun calcul analytique de dérivées ni de variance n'est nécessaire.

SECONDE PARTIE

MODÈLES ÉCONOMÉTRIQUES À UNE ÉQUATION

CHAPITRE I.

LA RÉGRESSION SIMPLE: ESTIMATION PONCTUELLE

1.1 Description du problème et exemples économiques

(1) Nous partons d'une relation linéaire, spécifiée par un modèle économique. Par exemple :

La fonction de consommation :

$$C = a + bY$$

La loi de demande :

$$X = a - bP_X$$

La fonction de coût :

$$CT = a + bQ\,.$$

(2) Nous désirons estimer les paramètres a, b de ces modèles à des fins *d'analyse* ou de *prévision*. Une telle estimation est plus élaborée qu'une simple étude de corrélation. Elle peut en effet servir à répondre à des questions de politique économique telles que :

(a) comment faudrait-il modifier les dépenses gouvernementales pour augmenter le niveau de l'emploi de $x\%$? Pour réduire le taux d'inflation de $y\%$?

(b) combien une firme doit-elle produire pour maximiser son profit?

(c) Une politique de soutien du prix d'un produit agricole doit-elle prendre la forme d'un prix garanti aux producteurs (et de l'achat de toute production

invendue) ou d'un subside à ces producteurs? Les coûts respectifs de ces deux politiques alternatives dépendront de l'élasticité de la demande, qui peut être estimée par l'économètre, à partir de données sur les variables X et P_X.

Les égalités précédentes ne seront jamais vérifiées exactement par des données sur les variables C, Y, X, P_X, etc. En effet :

- l'on ne peut espérer qu'une relation linéaire exacte fournisse une description complète du comportement des agents économiques. Il est trop complexe pour cela. Il est parfois erratique.
- des erreurs aléatoires de mesure, d'agrégation, etc., sont d'ordinaire présentes dans tout échantillon. Ces erreurs ne peuvent être expliquées par un modèle déterministe.

On ajoutera donc aux fonctions précédentes un *terme d'erreur* aléatoire u, et l'on écrira:

$$C = a + bY + u$$
$$X = a - bP_X + u$$
$$CT = a + bQ + u.$$

1.2 Le modèle et ses hypothèses

1.2.1 L'équation de régression.

Nous avons donc une équation linéaire de la forme :

$$y_t = a + bx_t + u_t \quad , \quad t = 1, \ldots, n \quad .$$

L'indice t correspond à une observation particulière, par exemple l'année 1960 dans un échantillon de 20 observations annuelles.

La variable y_t s'appelle indifféremment *variable endogène*, ou *variable dépendante*, ou *variable expliquée*. La variable x_t s'appelle indifféremment *variable exogène*, ou *variable indépendante*, ou *variable explicative*. On parle aussi de *régresseur*. Le terme u_t est un terme d'erreur aléatoire inobservable.

a et b sont des paramètres à estimer. Leurs estimateurs seront notés $\hat{a}$ et $\hat{b}$.

1.2.2 Les hypothèses.

Les estimateurs $\hat{a}$ et $\hat{b}$ vont dépendre des y_t, donc des u_t : ce seront des variables aléatoires, et nous aurons besoin des moments de leur distribution. Il nous faut donc faire des hypothèses sur la distribution des u_t .

H_1 . $E(u_t) = 0$ pour tout t .

Si cette hypothèse n'était pas satisfaite, le terme d'erreur aléatoire u_t aurait une composante systématique, qui aurait dû être incluse dans la partie non aléatoire de l'équation de régression. Le modèle serait alors mal spécifié.

H_2 . $V(u_t) = E(u_t^2) = \sigma^2$ pour tout t .

Cette hypothèse implique que chaque erreur u_t ait la même variance; si les u_t ont une distribution normale, chaque u_t aura la même distribution.

Comme exemple de modèle où cette hypothèse n'est pas vérifiée, on peut citer un modèle de régression dont les observations sont des moyennes calculées à partir de nombres d'observations différents: si le modèle vrai est:

$$y_{is} = a + bx_{is} + u_{is} \quad \text{pour } i = 1, \ldots, n_s \text{ et } s = 1, \ldots, T$$

où les u_{is} sont de variance σ^2 et sont indépendantes, et si le modèle estimé est:

$$\bar{y}_s = a + b\bar{x}_s + \bar{u}_s \quad \text{pour } s = 1, \ldots, T$$

avec:

$$\bar{y}_s = \frac{\sum_{i=1}^{n_s} y_{is}}{n_s}, \quad \bar{x}_s = \frac{\sum_{i=1}^{n_s} x_{is}}{n_s}, \quad \bar{u}_s = \frac{\sum_{i=1}^{n_s} u_{is}}{n_s}$$

on vérifie aisément que la variance des $\bar{u}_s$ dépend de s.

H_3 . $\mathrm{Cov}(u_t, u_h) = 0$ $t \neq h$.

Cette hypothèse sera satisfaite si le fait que u_t prenne une certaine valeur est indépendant de la valeur prise par u_h . Elle pourrait être violée, par exemple, si y_t était la production d'un bien agricole dans une région géographique donnée t . Une autre observation, faite dans une région voisine, pourrait être influencée par des conditions météorologiques communes.

Un autre exemple de viol de cette hypothèse est le cas où les u_t sont engendrées par l'équation de récurrence $u_t = \rho u_{t-1} + \epsilon_t$, où les ϵ_t sont d'espérance nulle, de variance constante, et ne sont pas corrélées entre elles. On vérifie aisément que la covariance entre u_t et u_{t-1} dépend de ρ.

H_4 . Les x_t sont non aléatoires (on dit aussi non stochastiques).

Cette hypothèse est provisoire, destinée à simplifier les arguments présentés. Nous verrons plus loin qu'on pourrait la remplacer par l'hypothèse plus faible que $E(x_t u_t) = 0$, sans changer certains résultats. Par la loi des espérances itérées, on peut aussi supposer que $E(u_t \mid x_t) = 0$.

L'hypothèse que la covariance entre le régresseur et le terme d'erreur contemporain est nulle est violée dans le modèle suivant:

$$C_t = a + bY_t + u_t$$

$$Y_t = C_t + I_t$$

où C_t est la consommation au temps t, Y_t est le revenu national au temps t, I_t est l'investissement au temps t, et u_t est le terme d'erreur. En substituant la première équation dans la seconde et en résolvant, on s'aperçoit aisément que $E(Y_t u_t) \neq 0$.

H_5 . x_t prend au moins deux valeurs différentes. Si cette hypothèse n'était pas satisfaite, nous n'aurions pas un problème de régression : en effet, $a + bx_t$ serait constante, et $y_t = a + bx_t + u_t$ serait constante à un terme aléatoire près. Nous aurions alors le modèle $y_t = \mu + u_t$ avec $\mu = E(y_t)$.

Nous voulons trouver les paramètres $\hat{a}$, $\hat{b}$ de la droite $\hat{a} + \hat{b}x_t$ qui *approche le mieux* la dépendance des y sur les x, c'est-à-dire qui *s'écarte le moins* du nuage de points (x_t, y_t). Quels critères allons-nous employer?

Il faut, qu'en moyenne, la distance entre y_t et $\hat{a} + \hat{b}x_t$ soit minimale. Il faut donc que la valeur absolue de $\hat{u}_t = y_t - \hat{a} - \hat{b}x_t$ soit petite, pour tout t. Nous pourrions retenir comme critères :

$$(1) \qquad \min_{\hat{a},\hat{b}} \quad \max_t \quad |\hat{u}_t|$$

$$(2) \qquad \min_{\hat{a},\hat{b}} \quad \sum_t |\hat{u}_t|$$

$$(3) \qquad \min_{\hat{a},\hat{b}} \quad \sum_t \hat{u}_t^2$$

Pour des raisons de commodité, nous allons employer le troisième critère : c'est la *méthode des moindres carrés.*

La différence:

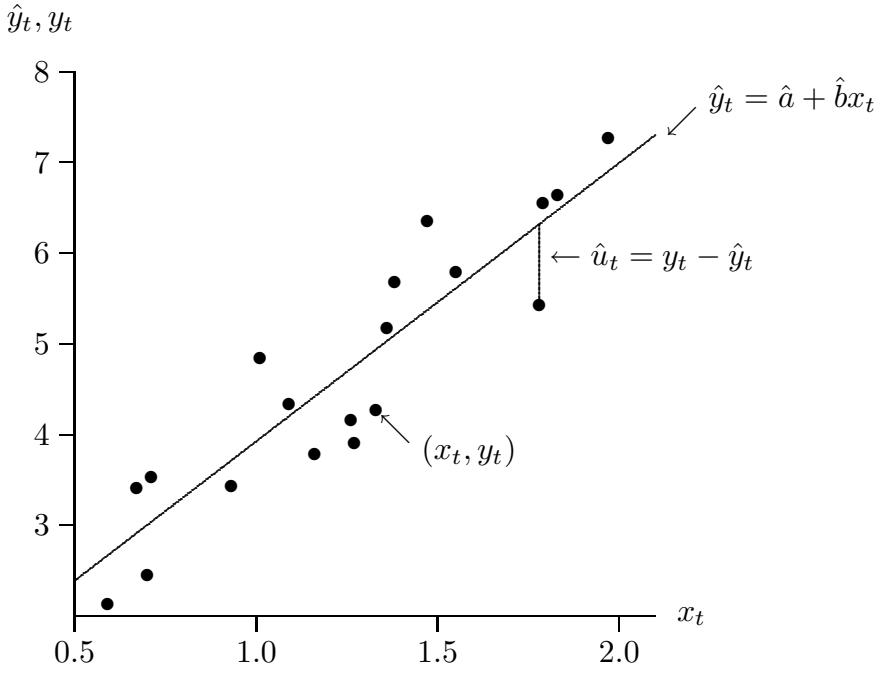
$$\hat{u}_t = y_t - \hat{a} - \hat{b}x_t$$

s'appelle un *résidu*, et est une estimation de l'erreur u_t. On peut écrire indifféremment:

$$y_t = a + bx_t + u_t$$

$$y_t = \hat{a} + \hat{b}x_t + \hat{u}_t$$

mais la première de ces relations est une hypothèse, tandis que l'autre est une identité! L'estimation par moindres carrés du modèle de régression simple sur la base d'observations (x_t, y_t) est illustrée par la figure suivante.



1.3 Les estimateurs de moindres carrés

Nous voulons donc minimiser en $\hat{a}$, $\hat{b}$ la somme de carrés :

$$S(\hat{a}, \hat{b}) = \sum \hat{u}_t^2 = \sum \left(y_t - \hat{a} - \hat{b}x_t \right)^2 .$$

Les conditions de premier ordre sont :

$$\frac{\partial S}{\partial \hat{a}} = -2 \sum \left(y_t - \hat{a} - \hat{b}x_t \right) = 0$$

$$\frac{\partial S}{\partial \hat{b}} = -2 \sum \left(y_t - \hat{a} - \hat{b}x_t \right) x_t = 0 .$$

Elles impliquent les *équations normales*:

(1) $$\sum y_t - n\hat{a} - \hat{b}\sum x_t = 0$$

(2) $$\sum x_t y_t - \hat{a}\sum x_t - \hat{b}\sum x_t^2 = 0\,.$$

En divisant (1) par n, on obtient :

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \quad \text{avec} \quad \bar{x} = \frac{\sum x_t}{n} \quad , \quad \bar{y} = \frac{\sum y_t}{n}\,.$$

En remplaçant cette valeur dans (2), il vient :

$$\sum (y_t - [\bar{y} - \hat{b}\bar{x}] - \hat{b}x_t)x_t = 0$$
$$\sum (y_t - \bar{y} - \hat{b}(x_t - \bar{x}))x_t = 0$$

$$\begin{aligned}\hat{b} &= \frac{\sum (y_t - \bar{y})x_t}{\sum (x_t - \bar{x})x_t} \\ &= \frac{\sum (y_t - \bar{y})(x_t - \bar{x})}{\sum (x_t - \bar{x})^2} \\ &= \frac{\sum x_t y_t - n\bar{x}\bar{y}}{\sum x_t^2 - n\bar{x}^2} \\ &= \frac{\sum (x_t - \bar{x})y_t}{\sum (x_t - \bar{x})^2} = \sum w_t y_t\end{aligned}$$

où :

$$w_t = \frac{(x_t - \bar{x})}{\sum (x_t - \bar{x})^2}.$$

Il est facile de vérifier, de même, que $\hat{a} = \sum z_t y_t$, avec:

$$z_t = \frac{1}{n} - \bar{x}w_t$$

Les deux estimateurs $\hat{a}$ et $\hat{b}$ sont donc des fonctions linéaires des y_t.

Les w_t et z_t possèdent des propriétés qu'il est utile de noter:

(1) $$\sum w_t = 0$$

(2) $$\sum w_t^2 = \frac{1}{\sum (x_t - \bar{x})^2}$$

$$\sum w_t x_t = 1 \tag{3}$$

$$\sum z_t = 1 \tag{4}$$

$$\sum z_t^2 = \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_t - \bar{x})^2} = \frac{\sum x_t^2}{n \sum (x_t - \bar{x})^2} \tag{5}$$

$$\sum z_t x_t = 0 \tag{6}$$

$$\sum w_t z_t = -\frac{\bar{x}}{\sum (x_t - \bar{x})^2}. \tag{7}$$

Exemple: soient les $n = 5$ observations suivantes sur les y_t et les x_t :

y_t	x_t
2	1
4	2
5	3
7	4
10	5

On a $\sum x_t = 15$, $\sum y_t = 28$, $\sum x_t^2 = 55$, $\sum x_t y_t = 103$, $\sum y_t^2 = 194$.

$$\hat{b} = \frac{103 - (15)(28)/5}{55 - (15)^2/5} = 1.9$$

$$\hat{a} = \frac{28}{5} - (1.9)\left(\frac{15}{5}\right) = -0.1 .$$

1.4 Moments des estimateurs de moindres carrés

1.4.1 Espérances mathématiques.

Nous allons vérifier que $\hat{a}$ et $\hat{b}$ sont des estimateurs sans biais de a et de b. On a

$$\begin{aligned}
\hat{a} = \sum z_t y_t &= \sum z_t(a + bx_t + u_t) \\
&= a\sum z_t + b\sum z_t x_t + \sum z_t u_t \\
&= a + 0 + \sum z_t u_t \\
\text{et} \quad E(\hat{a}) &= E(a) + \sum z_t E(u_t) = a
\end{aligned}$$

$$\begin{aligned}
\hat{b} = \sum w_t y_t &= \sum w_t(a + bx_t + u_t) \\
&= a\sum w_t + b\sum w_t x_t + \sum w_t u_t \\
&= 0 + b + \sum w_t u_t \\
\text{et} \quad E(\hat{b}) &= E(b) + \sum w_t E(u_t) = b.
\end{aligned}$$

1.4.2 Variances.

La variance de $\hat{b}$ se calcule comme :

$$\begin{aligned}
V(\hat{b}) &= E\left[\hat{b} - E(\hat{b})\right]^2 \\
&= E(\hat{b} - b)^2 .
\end{aligned}$$

Mais $\hat{b} - b = \sum w_t u_t$ comme nous l'avons montré. On a alors:

$$\begin{aligned} V(\hat{b}) &= E\left[\sum w_t u_t\right]^2 \\ &= E\left[\sum_{t=1}^{n} w_t^2 u_t^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i w_j u_i u_j\right] \\ &= \sum_{t=1}^{n} w_t^2 E\left(u_t^2\right) = \sigma^2 \sum_{t=1}^{n} w_t^2 = \frac{\sigma^2}{\sum (x_t - \bar{x})^2} \end{aligned}$$

puisque $E(u_t^2) = \sigma^2$, et puisque $E(u_i u_j) = 0$ pour $i \neq j$.

On a par ailleurs

$$\begin{aligned} V(\hat{a}) &= E(\hat{a} - a)^2 = E\left(\sum z_t u_t\right)^2 \\ &= \sigma^2 \sum z_t^2 \quad \text{par le même argument que précédemment} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_t - \bar{x})^2}\right] \\ &= \sigma^2 \frac{\sum x_t^2}{n \sum (x_t - \bar{x})^2}. \end{aligned}$$

1.4.3 Covariance.

$$\begin{aligned} \mathrm{Cov}(\hat{a}, \hat{b}) &= E(\hat{b} - b)(\hat{a} - a) \\ &= E\left[\left(\sum w_t u_t\right)\left(\sum z_t u_t\right)\right] \\ &= E\left[\sum_{t=1}^{n} w_t z_t u_t^2 + \sum_{i=1}^{n}\sum_{j \neq i} w_i z_j u_i u_j\right] \\ &= \sigma^2 \left[\sum w_t z_t\right] = \sigma^2 \left[\frac{\sum w_t}{n} - \bar{x} \sum w_t^2\right] \\ &= -\sigma^2 \frac{\bar{x}}{\sum (x_t - \bar{x})^2} \quad . \end{aligned}$$

1.5 Convergence en probabilité

On vérifie facilement à l'aide de ces moments que $\operatorname{plim}\hat{b} = b$ et $\operatorname{plim}\hat{a} = a$:

$$E\left(\hat{b}\right) = b \quad \text{et} \quad V\left(\hat{b}\right) = \frac{\sigma^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \underset{n\to\infty}{\longrightarrow} 0$$

$$E\left(\hat{a}\right) = a \quad \text{et} \quad V\left(\hat{a}\right) \underset{n\to\infty}{\longrightarrow} 0, \quad \text{car:} \quad V\left(\hat{a}\right) = \sigma^2 \frac{\sum x_t^2/n}{\sum (x_t - \bar{x})^2} \longrightarrow 0$$

sous la condition suffisante que $\lim_{n\to\infty} \frac{\sum x_t^2}{n}$ existe.

1.6 Interprétation matricielle

En réunissant toutes les observations sur l'équation de régression $y_t = a + bx_t + u_t$, il vient:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} a + \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} b + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

ou: $y = X\beta + u$.

Les équations normales peuvent s'écrire:

$$\begin{cases} n\hat{a} + \hat{b}\sum x_t = \sum y_t \\ \\ \hat{a}\sum x_t + \hat{b}\sum x_t^2 = \sum x_t y_t \end{cases}$$

ce qui implique:

$$\begin{pmatrix} n & \sum x_t \\ \sum x_t & \sum x_t^2 \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \sum y_t \\ \sum x_t y_t \end{pmatrix}$$

$$(X'X)\,\hat{\beta} = X'y \quad \Longrightarrow \hat{\beta} = (X'X)^{-1} X'y \quad .$$

La matrice $(X'X)^{-1}$ peut s'écrire:

$$\begin{pmatrix} n & \sum x_t \\ \sum x_t & \sum x_t^2 \end{pmatrix}^{-1} = \frac{1}{n\sum(x_t - \bar{x})^2} \begin{pmatrix} \sum x_t^2 & -\sum x_t \\ -\sum x_t & n \end{pmatrix}$$

$$= \frac{1}{\sum(x_t - \bar{x})^2} \begin{pmatrix} \sum x_t^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \quad .$$

On s'aperçoit qu'en multipliant cette matrice par σ^2, on obtient la matrice:

$$\begin{pmatrix} V(\hat{a}) & \mathrm{Cov}\left(\hat{a}, \hat{b}\right) \\ \mathrm{Cov}\left(\hat{a}, \hat{b}\right) & V\left(\hat{b}\right) \end{pmatrix} \quad .$$

Ceci peut être généralisé! En ajoutant des variables explicatives supplémentaires (des colonnes à la matrice X) on obtient le modèle de régression multiple.

On note l'importance de l'hypothèse H_5 : si $x_t = \alpha$ pour tout t, $\sum(x_t - \bar{x})^2 = 0$, $\det X'X = 0$ et les équations normales n'ont pas de solution unique.

1.7 Théorème de Gauss-Markov

Nous ne verrons ici qu'un cas particulier de ce théorème (une version plus générale sera vue en régression multiple).

Nous avons vu que les estimateurs de moindres carrés sont sans biais et convergents. Sont-ils de variance minimale? La réponse est: oui, dans la classe des estimateurs sans biais et linéaires. Nous allons vérifier cette propriété dans le cas de $\hat{b}$.

Un estimateur linéaire arbitraire de b peut s'écrire comme:

$$\begin{aligned}\tilde{b} &= \sum c_t y_t = \sum c_t \left(a + bx_t + u_t\right) \\ &= a\sum c_t + b\sum c_t x_t + \sum c_t u_t \quad ,\end{aligned}$$

une condition nécessaire et suffisante pour que $E\left(\tilde{b}\right) = b$ pour tout (a, b) est $\sum c_t = 0$, $\sum c_t x_t = 1$. Alors:

$$\begin{aligned}V\left(\tilde{b}\right) &= E\left(\tilde{b} - b\right)^2 = E\left(\sum c_t u_t\right)^2 \\ &= \sigma^2 \sum c_t^2 \quad .\end{aligned}$$

On va minimiser cette variance sous la contrainte $E\left(\tilde{b}\right) = b$ et montrer que la solution est $c_t = w_t$.

Comme la minimisation de $V(\tilde{b})$ est équivalente à celle de $V(\tilde{b})/\sigma^2$, le Lagrangien s'écrit:

$$\Lambda = \sum c_t^2 + \lambda_1 \sum c_t + \lambda_2 \left(\sum c_t x_t - 1\right)$$

et les conditions de premier ordre sont donc:

$$\frac{\partial \Lambda}{\partial c_t} = 2c_t + \lambda_1 + \lambda_2 x_t = 0 \quad (t = 1, \dots, n)$$

Pour éliminer λ_1 et λ_2 à l'aide des contraintes, nous pouvons utiliser:

$$\begin{aligned}\sum_{t=1}^{n} \frac{\partial \Lambda}{\partial c_t} &= 2\sum_{t=1}^{n} c_t + n\lambda_1 + \lambda_2 \sum_{t=1}^{n} x_t = 0 \\ \sum_{t=1}^{n} \frac{\partial \Lambda}{\partial c_t} x_t &= 2\sum_{t=1}^{n} c_t x_t + \lambda_1 \sum_{t=1}^{n} x_t + \lambda_2 \sum_{t=1}^{n} x_t^2 = 0 \quad .\end{aligned}$$

En utilisant les contraintes $\sum c_t = 0$, $\sum c_t x_t = 1$:

$$n\lambda_1 + \lambda_2 \sum x_t = 0$$

$$2 + \lambda_1 \sum x_t + \lambda_2 \sum x_t^2 = 0$$

$$\begin{pmatrix} n & \sum x_t \\ \sum x_t & \sum x_t^2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -2 \end{pmatrix} \quad .$$

L'inverse de la matrice des coefficients a déjà été calculée ($(X'X)^{-1}$). On peut donc calculer la solution du système comme:

$$\begin{pmatrix} \lambda_1 \\ \\ \lambda_2 \end{pmatrix} = \frac{1}{n\sum(x_t-\bar{x})^2} \begin{pmatrix} \sum x_t^2 & -\sum x_t \\ \\ -\sum x_t & n \end{pmatrix} \begin{pmatrix} 0 \\ \\ -2 \end{pmatrix}$$

$$= \begin{pmatrix} 2\bar{x}/\sum(x_t-\bar{x})^2 \\ \\ -2/\sum(x_t-\bar{x})^2 \end{pmatrix} .$$

En substituant ces valeurs dans $\frac{\partial \Lambda}{\partial c_t} = 0$:

$$2c_t = -2\frac{\bar{x}}{\sum(x_t-\bar{x})^2} + 2\frac{x_t}{\sum(x_t-\bar{x})^2}$$

$$c_t = \frac{(x_t-\bar{x})}{\sum(x_t-\bar{x})^2} = w_t .$$

Cette valeur de c_t minimise donc bien la variance sous la contrainte que l'estimateur soit sans biais.

1.8 Estimation de la variance des erreurs

Les variances et la covariance calculées dans les sections 1.4.2 et 1.4.3 dépendent du paramètre inconnu σ^2 . Une procédure naturelle serait de calculer la variance d'échantillon $\frac{1}{n}\sum(\hat{u}_t-\bar{\hat{u}})^2$, et de corriger un biais éventuel, pour arriver à un estimateur de σ^2 .

En fait, $\sum\left(\hat{u}_t-\bar{\hat{u}}\right)^2 = \sum\hat{u}_t^2$, car

$$\sum\hat{u}_t = \sum\left(y_t-\hat{a}-\hat{b}x_t\right) = \sum y_t - n\hat{a} - \hat{b}\sum x_t = 0$$

en vertu de la première équation normale (Section 1.3). Nous allons prouver que

$$E\left[\sum\hat{u}_t^2\right] = (n-2)\sigma^2$$

et que donc $s^2 = \frac{1}{n-2}\sum\hat{u}_t^2$ est un estimateur sans biais de σ^2.

Nous avons:

$$\begin{aligned}\hat{u}_t &= y_t - \hat{a} - \hat{b}x_t \\ &= a + bx_t + u_t - (\bar{y} - \hat{b}\bar{x}) - \hat{b}x_t \\ &= a + bx_t + u_t - a - b\bar{x} - \bar{u} + \hat{b}\bar{x} - \hat{b}x_t \\ &= u_t - \bar{u} + (b - \hat{b})(x_t - \bar{x})\,.\end{aligned}$$

Alors

$$\begin{aligned}\sum \hat{u}_t^2 &= \sum \Big[(u_t - \bar{u})^2 \,+\, (b - \hat{b})^2(x_t - \bar{x})^2 \,+\, 2(b - \hat{b})(x_t - \bar{x})(u_t - \bar{u})\Big] \\ &= \sum (u_t - \bar{u})^2 \,+\, (b - \hat{b})^2 \sum (x_t - \bar{x})^2 \,+\, 2(b - \hat{b}) \sum (x_t - \bar{x})(u_t - \bar{u}) \quad .\end{aligned}$$

Mais

$$\begin{aligned}\sum (x_t - \bar{x})(u_t - \bar{u}) &= \Big[\sum (x_t - \bar{x})^2\Big] \sum w_t (u_t - \bar{u}) \\ &= (\hat{b} - b) \sum (x_t - \bar{x})^2\end{aligned}$$

puisque $\sum w_t(u_t - \bar{u}) = \sum w_t u_t = \hat{b} - b$.

Donc

$$\begin{aligned}\sum \hat{u}_t^2 &= \sum (u_t - \bar{u})^2 + (b - \hat{b})^2 \sum (x_t - \bar{x})^2 \,-\, 2(b - \hat{b})^2 \sum (x_t - \bar{x})^2 \\ &= \sum (u_t - \bar{u})^2 - (b - \hat{b})^2 \sum (x_t - \bar{x})^2 \quad .\end{aligned}$$

Calculons séparément l'espérance de chacun de ces termes.

$$\begin{aligned}E\left[\sum (u_t - \bar{u})^2\right] &= E\left[\sum u_t^2 - \frac{1}{n}(\sum u_t)^2\right] = n\sigma^2 - \frac{n}{n}\sigma^2 = (n-1)\sigma^2 \\ E\left[(\hat{b} - b)^2 \sum (x_t - \bar{x})^2\right] &= \sigma^2.\end{aligned}$$

Et donc $E\left[\sum \hat{u}_t^2\right] = (n-2)\sigma^2$, Q.E.D.

On peut interpréter la division par $n-2$ de la manière suivante. Précédemment (à la section 4.1 de la première partie), nous avions vu que pour obtenir un estimateur sans biais de la variance, on devait diviser par $n-1$ la somme des carrés des déviations par rapport à la moyenne. Cette division par $n-1$ était en fait due à la présence d'une condition liant les

déviations par rapport à la moyenne: la somme de ces déviations est identiquement nulle. Dans le cas qui nous occupe, nous avons *deux* conditions liant les résidus $\hat{u}_t$, à savoir:

$$\sum_{t=1}^{n} \hat{u}_t = 0$$

$$\sum_{t=1}^{n} \hat{u}_t x_t = 0$$

Si nous connaissons $n-2$ des résidus, nous pouvons déterminer les valeurs des deux derniers à l'aide de ces conditions.

1.9 Décomposition de la variance: le coefficient de détermination

Nous allons voir que la variance totale des y, soit $\frac{\sum(y_t - \bar{y})^2}{n}$, peut être décomposée en une somme de deux variances, celle des $\hat{y}$ (partie expliquée par la régression) et celle des $\hat{u}$ (partie résiduelle). Ceci nous permettra de définir le coefficient de détermination, qui permet de mesurer la qualité de l'ajustement linéaire.

A cette fin, nous prouverons que :

$$\sum(y_t - \bar{y})^2 \quad = \quad \sum\left(\hat{y}_t - \bar{\hat{y}}\right)^2 \quad + \quad \sum \hat{u}_t^2$$

$$\text{soit} \qquad \text{SCT} \quad = \qquad \text{SCE} \qquad + \qquad \text{SCR}\,.$$

En guise d'étape préliminaire, démontrons une formule de calcul commode pour $\sum \hat{u}_t^2$.

Lemme $\quad \sum \hat{u}_t^2 = \sum(y_t - \bar{y})^2 - \hat{b}^2 \sum(x_t - \bar{x})^2$

Démonstration

$$\begin{aligned} \hat{u}_t \quad &= \quad y_t - \hat{y}_t \quad = \quad y_t - \hat{a} - \hat{b}x_t \\ &= \quad (y_t - \bar{y}) - \hat{b}(x_t - \bar{x}) \quad . \end{aligned}$$

Donc

$$\sum \hat{u}_t^2 = \sum (y_t - \bar{y})^2 \; - \; 2\hat{b} \sum (x_t - \bar{x})\,(y_t - \bar{y}) \; + \; \hat{b}^2 \sum (x_t - \bar{x})^2 \quad .$$

Mais $\quad \sum (x_t - \bar{x})\,(y_t - \bar{y}) = \hat{b} \sum (x_t - \bar{x})^2 \quad , \quad$ donc

$$\sum \hat{u}_t^2 = \sum (y_t - \bar{y})^2 - \hat{b}^2 \sum (x_t - \bar{x})^2 \quad , \quad Q.E.D.$$

Pour prouver que SCT = SCE + SCR, il suffit alors de montrer que :

$$\hat{b}^2 \sum (x_t - \bar{x})^2 = \sum (\hat{y}_t - \bar{\hat{y}})^2 \quad .$$

Mais ceci est évident car :

$$\sum (\hat{y}_t - \bar{\hat{y}})^2 = \sum (\hat{a} + \hat{b}x_t - \hat{a} - \hat{b}\bar{x})^2 \quad .$$

On définit alors le coefficient de détermination comme :

$$R^2 \quad = \quad \frac{\text{SCE}}{\text{SCT}} \quad = \quad 1 - \frac{\text{SCR}}{\text{SCT}}$$

et l'on a $0 \leq R^2 \leq 1$. Plus R^2 est proche de l'unité, plus grand est le pourcentage de la variance totale expliquée par la régression, et meilleure est donc la qualité de l'ajustement.

Mentionnons dès à présent une interprétation statistique plus fine du R^2. Nous démontrerons, en régression multiple, que si $b = 0$, $(n-2)R^2/(1-R^2)$ suit le carré d'une loi de Student avec $n-2$ degrés de liberté. Avec un seuil de signification α, le R^2 sera donc "bon" si:

$$\frac{(n-2)R^2}{1-R^2} > t^2_{n-2,\alpha/2}$$

1.10 Exemple numérique

Poursuivons l'exemple de la section 1.3. Nous avions trouvé les valeurs $\hat{a} = -0.1$ et $\hat{b} = 1.9$. On a de plus:

$$\begin{aligned}
\bar{x} &= 3 \\
\bar{y} &= 5.6 \\
\sum (x_t - \bar{x})^2 &= 10 \\
\sum (y_t - \bar{y})^2 &= 37.20 \\
\sum \hat{u}_t^2 &= 37.20 - (1.9)^2(10) = 1.10 \\
s^2 &= \frac{1.10}{3} = 0.37 \\
s_{\hat{b}}^2 &= \frac{0.37}{10} = 0.037 \\
s_{\hat{a}}^2 &= 0.37\left[\frac{1}{5} + \frac{9}{10}\right] = 0.403 \\
s_{\hat{a}\hat{b}} &= -\frac{(0.37)3}{10} = -0.11 \\
R^2 &= 1 - \frac{1.10}{37.20} = 0.97 \quad .
\end{aligned}$$

Nous pouvons présenter ces résultats comme:

$$\begin{array}{lll}
\hat{y}_t = -0.1 & + \quad 1.9 \quad x_t & (R^2 = 0.97) \\
\quad (0.635) & \quad (0.192) &
\end{array}$$

où les nombres entre parenthèses sont les estimations des écarts-types des coefficients estimés. On peut aussi les présenter comme:

$$\begin{array}{lll}
\hat{y}_t = -0.1 & + \quad 1.9 \quad x_t & (R^2 = 0.97) \\
\quad (-0.157) & \quad (9.88) &
\end{array}$$

où les nombres entre parenthèses sont les rapports entre les coefficients estimés et les estimations de leurs écarts-types. On appelle ces rapports les rapports t (*t-ratios*); ils nous serviront dans le cadre des tests d'hypothèses.

CHAPITRE II.

LA RÉGRESSION SIMPLE: INTERVALLES DE CONFIANCE ET TESTS D'HYPOTHÈSES

2.1 Tests sur les coefficients individuels

$\hat{a}$ et $\hat{b}$ ne sont que des estimateurs ponctuels de a et de b. Dans ce chapitre, nous nous efforcerons d'énoncer des jugements de probabilité du type :

$P\left[\underline{b} \leq b \leq \bar{b}\right] = 1 - \alpha$, où α est une constante appelée niveau de signification.

Un tel jugement de probabilité doit se lire :

"J'ai une probabilité de $1 - \alpha$ de ne pas me tromper lorsque j'affirme que b est compris entre $\underline{b}$ et $\bar{b}$".

Les bornes $\underline{b}$ et $\bar{b}$ vont dépendre de $\hat{b}$ et de sa variance. Elles sont donc aléatoires, au même titre que $\hat{b}$.

Elles dépendront aussi de la distribution de $\hat{b}$. Si cette distribution est symétrique autour de b, l'intervalle $\left[\underline{b}, \bar{b}\right]$ aura $\hat{b}$ comme point médian. Ce sera le plus petit intervalle ayant une probabilité $1 - \alpha$ de contenir b.

Il nous faut donc maintenant spécifier la distribution de $\hat{a}$ et $\hat{b}$, ce qui nécessite une hypothèse sur la distribution des erreurs u_t. Si nous faisons l'hypothèse de normalité :

$$H_6 \quad : \quad u_t \sim N(0, \sigma^2)$$

$\hat{a} = a + \sum z_t u_t$ et $\hat{b} = b + \sum w_t u_t$ seront normales, puisque ce sont alors des combinaisons linéaires de variables normales indépendantes.

Quelles seront alors les formes de $\underline{a}$, $\bar{a}$, $\underline{b}$ et $\bar{b}$?

Si σ^2 était connue, nous aurions :

$$\frac{b - \hat{b}}{\sigma_{\hat{b}}} \sim N(0,1) \quad \text{et} \quad \frac{a - \hat{a}}{\sigma_{\hat{a}}} \sim N(0,1)$$

$$\text{avec} \quad \sigma^2_{\hat{b}} = \frac{\sigma^2}{\sum (x_t - \bar{x})^2} \quad , \qquad \sigma^2_{\hat{a}} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_t - \bar{x})^2}\right] .$$

Nous pourrions alors écrire, par exemple,

$$P\left[-z_{\alpha/2} \leq \frac{b-\hat{b}}{\sigma_{\hat{b}}} \leq z_{\alpha/2}\right] \quad = \quad 1-\alpha$$

où $z_{\alpha/2}$ est la valeur de la variable normale réduite ayant une probabilité $\frac{\alpha}{2}$ d'être dépassée.

Nous aurions alors :

$$P\left[\hat{b} - z_{\alpha/2}\,\sigma_{\hat{b}} \leq b \leq \hat{b} + z_{\alpha/2}\sigma_{\hat{b}}\right] \quad = \quad 1-\alpha \quad .$$

Les bornes cherchées sont donc :

$$\begin{aligned} \underline{b} &= \hat{b} - z_{\alpha/2}\sigma_b \\ \text{et} \quad \bar{b} &= \hat{b} + z_{\alpha/2}\sigma_b \quad . \end{aligned}$$

En pratique, σ^2 est inconnue. Que se passe-t-il lorsqu'on la remplace par son estimation sans biais

$$s^2 \quad = \quad \frac{\sum \hat{u}_t^2}{n-2} \quad ?$$

Pour reprendre l'exemple de $\hat{b}$:

$$\frac{b-\hat{b}}{s_{\hat{b}}} \quad = \quad \frac{b-\hat{b}}{\sqrt{\dfrac{\sum \hat{u}_t^2}{n-2}\dfrac{1}{\sum (x_t-\bar{x})^2}}}$$

$$= \quad \frac{\dfrac{b-\hat{b}}{\sqrt{\sigma^2\left(\dfrac{1}{\sum (x_t-\bar{x})^2}\right)}}}{\sqrt{\dfrac{\sum \hat{u}_t^2}{\sigma^2(n-2)}}} \quad \underset{\text{def}}{=} \quad \frac{N}{D} \quad .$$

N est une variable normale réduite. Nous prouverons rigoureusement plus loin que

$$\frac{\sum \hat{u}_t^2}{\sigma^2}$$

est une variable χ^2 avec $n-2$ degrés de liberté, indépendante de la variable N. Par définition, le rapport $\frac{N}{D}$ est alors une variable Student avec $n-2$ degrés de liberté.

Donc :

$$\frac{b-\hat{b}}{s_{\hat{b}}} \sim t_{n-2} \quad \text{et, de manière analogue} \quad \frac{a-\hat{a}}{s_{\hat{a}}} \sim t_{n-2}$$

et les intervalles de confiance sont donnés par :

$$P\left[\hat{b} - t_{n-2;\frac{\alpha}{2}}\ s_{\hat{b}} \leq b \leq \hat{b} + t_{n-2;\frac{\alpha}{2}}\ s_{\hat{b}}\right] \quad = \quad 1-\alpha\,,$$

$$P\left[\hat{a} - t_{n-2;\frac{\alpha}{2}}\ s_{\hat{a}} \leq a \leq \hat{a} + t_{n-2;\frac{\alpha}{2}}\ s_{\hat{a}}\right] \quad = \quad 1-\alpha \quad .$$

Pour tester :

$$H_0 : b = b_0 \qquad \text{contre} \qquad H_1 : b \neq b_0$$

on ne rejettera pas H_0 si $b_0 \in [\underline{b}, \bar{b}]$.

Pour tester :

$$H_0 : b = b_0 \qquad \text{contre} \qquad H_1 : b > b_0$$

on rejette H_0 si $b_0 < \hat{b} - t_{n-2;\alpha}\, s_{\hat{b}}$.

Pour tester :

$$H_0 : b = b_0 \qquad \text{contre} \qquad H_1 : b < b_0$$

on rejette H_0 si $b_0 > \hat{b} + t_{n-2;\alpha} s_{\hat{b}}$.

Des procédures analogues sont évidemment valables pour le paramètre a.

2.2 Test sur les deux paramètres a et b

Il s'agit ici du test :

$$H_0 : a = a_0 \qquad \text{et} \qquad b = b_0$$

contre

$$H_1 : a \neq a_0 \qquad \text{ou} \qquad b \neq b_0 \quad , \quad \text{ou les deux.}$$

Ce test n'est *pas* équivalent à une juxtaposition des deux tests t sur chaque coefficient de régression. Une méthode bivariée s'impose, et nos intervalles de confiance deviennent des ellipses. En pratique, on passe par la variable F de Fisher-Snedecor.

La statistique à employer est:

$$F_{\text{obs}} \quad = \quad \frac{Q/2}{s^2}$$

$$\text{avec} \quad Q = \left[n(\hat{a} - a_0)^2 + 2n\bar{x}(\hat{a} - a_0)(\hat{b} - b_0) + \left(\sum x_t^2 \right) (\hat{b} - b_0)^2 \right] \quad .$$

Q est toujours positive ou nulle; elle sera d'autant plus grande que $\hat{a}$ et $\hat{b}$ diffèrent de a_0 et b_0. Or, ce sont bien les valeurs élevées d'une statistique F qui conduisent à rejeter l'hypothèse nulle. Par ailleurs, une valeur élevée de s^2 reflète une mauvaise qualité de l'ajustement statistique; il est donc logique qu'elle nous fasse hésiter à rejeter l'hypothèse H_0.

En régression multiple, nous démontrerons que si H_0 est vraie, F_{obs} a la distribution $F_{2,n-2}$. On rejettera donc H_0 si

$$F_{\text{obs}} \quad > \quad F_{2;n-2;\alpha} \, .$$

Nous montrerons aussi que F_{obs} est égale à $(n-2)/2n$ fois la statistique de Wald pour tester l'hypothèse $H_0 : (a, b) = (a_0, b_0)$ contre $H_1 : (a, b) \neq (a_0, b_0)$. Ceci fournit une première justification rigoureuse de l'emploi de cette statistique.

2.3 Test sur une combinaison linéaire des coefficients

Un estimateur sans biais d'une combinaison linéaire $\gamma = \alpha a + \beta b$ des coefficients a et b est bien sûr:

$$\hat{\gamma} = \alpha\hat{a} + \beta\hat{b}.$$

Afin de construire un intervalle de confiance pour γ, nous devons estimer la variance de $\hat{\gamma}$:

$$\begin{aligned} V(\alpha\hat{a} + \beta\hat{b}) &= \alpha^2 V(\hat{a}) + \beta^2 V(\hat{b}) + 2\alpha\beta \operatorname{Cov}(\hat{a}, \hat{b}) \\ &= \sigma^2 \left[\alpha^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_t - \bar{x})^2} \right) + \frac{\beta^2}{\sum (x_t - \bar{x})^2} - 2\frac{\alpha\beta\bar{x}}{\sum (x_t - \bar{x})^2} \right] \\ &= \sigma^2 \left[\frac{\alpha^2}{n} + \frac{(\beta - \alpha\bar{x})^2}{\sum (x_t - \bar{x})^2} \right] \quad . \end{aligned}$$

En utilisant le même raisonnement que précédemment (section 2.1.), on peut montrer que :

$$\frac{\gamma - \alpha\hat{a} - \beta\hat{b}}{s\sqrt{\dfrac{\alpha^2}{n} + \dfrac{(\beta - \alpha\bar{x})^2}{\sum(x_t - \bar{x})^2}}} \quad \sim \quad t_{n-2}$$

et un intervalle de confiance est donc donné par les deux bornes

$$\alpha\hat{a} + \beta\hat{b} \pm t_{n-2;\frac{\alpha}{2}}\, s\sqrt{\frac{\alpha^2}{n} + \frac{(\beta - \alpha\bar{x})^2}{\sum(x_t - \bar{x})^2}} \quad .$$

2.4 Prévision

Que se passerait-il si nous voulions trouver un intervalle de confiance sur une valeur *future* y_θ de y? On parlerait alors d'intervalle de prévision. Supposons par exemple que $y = a+bx+u$ soit une fonction de consommation, que nous possédions des données annuelles entre 1960 et 1981 sur la consommation et le revenu national, et que nous voulions prédire la consommation pour l'année 1982, conditionnellement à une projection x_θ du revenu national pour 1982.

Sous l'hypothèse que le modèle reste inchangé, nous aurons:

$$\begin{aligned} y_\theta &= a + bx_\theta + u_\theta \quad \text{et} \\ \hat{y}_\theta &= \hat{a} + \hat{b}x_\theta \quad \text{sera sans biais} \quad . \end{aligned}$$

La variable $y_\theta - \hat{y}_\theta = u_\theta - (\hat{a} - a) - (\hat{b} - b)x_\theta$ est normale, de paramètres:

$$\begin{aligned} E(y_\theta - \hat{y}_\theta) &= 0 \\ V(y_\theta - \hat{y}_\theta) &= E(y_\theta - \hat{y}_\theta)^2 \\ &= E(u_\theta^2) + E((\hat{a} - a) + (\hat{b} - b)x_\theta)^2 \end{aligned}$$

puisque $\hat{a}$ et $\hat{b}$ ne dépendent que de $u_1, u_2, \ldots u_n$, et que $E(u_i u_\theta) = 0, \quad i = 1, 2, \ldots, n$: On a donc bien $E(\hat{a}u_\theta) = E(\hat{b}u_\theta) = 0$.

Le premier terme de la somme est égal à σ^2. Le second terme peut être calculé à l'aide des résultats de la section 2.3, en posant $\alpha = 1$ et $\beta = x_\theta$. Nous avons donc:

$$E(y_\theta - \hat{y}_\theta)^2 \quad = \quad \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_\theta - \bar{x})^2}{\sum(x_t - \bar{x})^2}\right]$$

et les bornes de l'intervalle de prévision sont données par

$$\hat{y}_\theta \pm t_{n-2;\frac{\alpha}{2}} \quad s\sqrt{1+\frac{1}{n}+\frac{(x_\theta-\bar{x})^2}{\sum(x_t-\bar{x})^2}} \quad .$$

2.5 Exemple numérique

Reprenons l'exemple numérique du chapitre 1. Nous avons $t_{3;0.025} = 3.182$. Un intervalle de confiance sur b correspondant à $\alpha = 0.05$ sera donc donné par:

$$\left[1.9-(3.182)\sqrt{0.037},\ 1.9+(3.182)\sqrt{0.037}\right] = [1.29 \quad , \quad 2.51] \quad .$$

On rejettera donc, par exemple, l'hypothèse:

$$H_0 : b = 1.2$$

mais on ne rejettera pas l'hypothèse:

$$H_0 : b = 1.5.$$

Pour tester:

$$\begin{aligned} & H_0 : a = -0.15 \quad \text{et} \quad b = 2.5 \\ \text{contre} \quad & H_1 : a \neq -0.15 \quad \text{ou} \quad b \neq 2.5 \end{aligned}$$

on construit la statistique

$$\begin{aligned} F_{\text{obs}} &= \frac{1}{2(0.37)}\Big[5(-0.10+0.15)^2 + 2\cdot 5\cdot 3(-0.10+0.15)(1.9-2.5) \\ &\qquad + 55(1.9-2.5)^2\Big] \\ &= \frac{18.9125/2}{0.37} = 25.79 \quad . \end{aligned}$$

On a $F_{2;3;0.05} = 9.55$ et $F_{2;3;0.01} = 30.82$.
On ne rejette donc pas H_0 pour $\alpha = 0.01$, mais on la rejette pour $\alpha = 0.05$.

Un intervalle de confiance sur $y_0 = E[y \mid x = 3.5]$ a pour bornes :

$$-0.1 + (1.9)(3.5) \quad \pm \quad (3.182)(0.61)\sqrt{\frac{1}{5} + \frac{(3.5-3)^2}{10}} \quad \text{si} \quad \alpha = 0.05.$$

Ce qui donne [5.636 , 7.464].

Un intervalle de prévision sur $y_6 = a + b(6) + u_6$ au niveau de signification $\alpha = 0.01$ aura pour bornes:

$$-0.1 + (1.9)(6) \pm (5.841)(0.61)\sqrt{1 + \frac{1}{5} + \frac{(6-3)^2}{10}}$$

ce qui donne [6.175 , 16.426].

CHAPITRE III

COMPLÉMENT D'ALGÈBRE MATRICIELLE

3.1. Formes quadratiques

Soit x un vecteur $n \times 1$. Une forme quadratique est une expression du type $x'Ax$, où A est une matrice symétrique $n \times n$. Elle est dite *définie non négative* si $x'Ax \geq 0$ pour tout x; *définie positive* si $x'Ax > 0$ pour tout $x \neq 0$; *semi-définie positive* si $x'Ax \geq 0$ pour tout x et si rang $(A) \neq n$. La même terminologie s'applique à la matrice A. Rappelons sans autres commentaires quelques propriétés importantes des matrices symétriques et des matrices définies.

3.1.1 Propriétés des matrices symétriques.

Si $A = A'$:

(1) Ses valeurs propres sont toutes réelles.
(2) A deux valeurs propres différentes correspondent des vecteurs propres orthogonaux.
(3) On peut associer k vecteurs propres orthogonaux à une valeur propre de multiplicité k.
(4) Il existe une matrice C orthogonale, dont les colonnes sont les vecteurs propres de A, telle que:
$C'AC = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ où les λ_i sont les valeurs propres de A.
(5) Le rang de A est égal au nombre de valeurs propres de A non nulles.

3.1.2 Propriétés des matrices définies non négatives.

(1) Une matrice A d'ordre n est définie non négative si et seulement si (a) toutes ses valeurs propres sont non négatives, ou (b) il existe une matrice B de dimensions $m \times n$ et de rang m telle que $B'B = A$.
(2) Si A est définie non négative, alors (a) $a_{ii} \geq 0$ pour tout i, et (b) $B'AB$ est définie non négative pour toute matrice B de dimensions $n \times m$.

3.1.3 Propriétés des matrices définies positives.

(1) Si A est définie positive, alors:
- A est régulière.
- $a_{ii} > 0$ pour tout i.
- Si B est $n \times m$ et de rang m, $B'AB$ est définie positive (corollaire: $B'B$ est définie positive).

(2) A est définie positive si et seulement si:
- Il existe une matrice B régulière telle que $A = B'B$, ou:
- Toutes ses valeurs propres sont strictement positives, ou:
- Tous ses mineurs principaux sont strictement positifs, ou:
- Tous les mineurs principaux de $-A$ alternent en signe, en commençant par moins, ou:
- Il existe une matrice D régulière telle que $DAD' = I$.

3.2 Matrices symétriques idempotentes

Soit A une matrice $n \times n$ avec $A = A'$ et $AA = A$. Nous avons les résultats suivants:

3.2.1 A est régulière si et seulement si A = I.

Démonstration

Si A est régulière, prémultiplions les deux membres de $AA = A$ par A^{-1}. Cela donne:

$$A^{-1}AA = A^{-1}A,$$

soit aussi $IA = I$. La réciproque est immédiate.

3.2.2 Les valeurs propres de A sont 0 ou 1.

Démonstration

Si λ est une valeur propre de A, $Ax = \lambda x$ pour un vecteur $x \neq 0$. En prémultipliant les deux membres par A:

$$AAx = \lambda Ax,$$

donc aussi $Ax = \lambda^2 x$, en utilisant $AA = A$ et $Ax = \lambda x$; nous avons alors $\lambda x = \lambda^2 x$, ce qui démontre la propriété.

3.2.3 Le déterminant de A est 0 ou 1.

Démonstration

Evidente, car le déterminant d'une matrice est égal au produit de ses valeurs propres.

3.2.4 Le rang de A est égal à sa trace.

Démonstration

Comme A est symétrique, il existe une matrice orthogonale C telle que $C'AC = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

On a alors:

$$\begin{aligned} tr\, A = tr\, CC'A &= tr\, C'AC \\ &= tr\, \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \\ &= \text{rang de } A \end{aligned}$$

puisque $CC' = I$, et puisque les λ_i sont égaux à 0 ou 1, le nombre de uns étant le rang de A.

3.3 L'inversion en forme partagée

Soit A une matrice $n \times n$, régulière, partagée comme suit:

$$A = \begin{pmatrix} E & F \\ G & H \end{pmatrix}$$

et supposons que E et $D = H - GE^{-1}F$ soient régulières. Alors:

$$A^{-1} = \begin{pmatrix} E^{-1}(I + FD^{-1}GE^{-1}) & -E^{-1}FD^{-1} \\ -D^{-1}GE^{-1} & D^{-1} \end{pmatrix}$$

On vérifie en effet par multiplication que $AA^{-1} = I$.

3.4 Notions de dérivation matricielle

Si $\lambda = \lambda(x)$ est un scalaire et x est $1 \times n$:

$$\frac{\partial \lambda}{\partial x} = \left(\frac{\partial \lambda}{\partial x_1} \cdots\cdots \frac{\partial \lambda}{\partial x_n} \right) \quad .$$

De même, si x est $n \times 1$:

$$\frac{\partial \lambda}{\partial x} = \begin{pmatrix} \partial\lambda/\partial x_1 \\ \vdots \\ \partial\lambda/\partial x_n \end{pmatrix} \quad .$$

Si $v = v(x)$ et x sont des vecteurs (lignes ou colonnes) ayant respectivement n et m éléments:

$$\frac{\partial v}{\partial x} = \begin{pmatrix} \frac{\partial v_1}{\partial x_1} & \cdots & \frac{\partial v_n}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial v_1}{\partial x_m} & \cdots & \frac{\partial v_n}{\partial x_m} \end{pmatrix}$$

est la matrice Jacobienne de $v(x)$.

Dans cette notation, nous avons, si A est $n \times m$:

$$\frac{\partial (Ax)}{\partial x} = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ \vdots & \vdots & & \vdots \\ a_{1m} & a_{2m} & \dots & a_{nm} \end{pmatrix} = A' \quad .$$

De même:

$$\frac{\partial (x'A)}{\partial x} = A \quad .$$

Pour une forme quadratique, si A est $n \times n$ et symétrique, on a:

$$\frac{\partial (x'Ax)}{\partial x} = 2Ax \quad .$$

Par exemple, si $A = \left[\begin{smallmatrix} 2 & 1 \\ 1 & 3 \end{smallmatrix}\right]$, on a $x'Ax = 2x_1^2 + 2x_1x_2 + 3x_2^2$, et

$$\frac{\partial (x'Ax)}{\partial x} \quad = \quad \begin{pmatrix} 4x_1 & + & 2x_2 \\ 2x_1 & + & 6x_2 \end{pmatrix} \quad = \quad 2Ax \quad .$$

CHAPITRE IV

COMPLÉMENT D'ANALYSE STATISTIQUE MULTIVARIÉE

4.1 La loi normale multivariée

La densité normale univariée, de paramètres m et σ^2:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\ (x-m)^2\right)$$

peut être généralisée à la densité normale multivariée, fonction de densité jointe des composantes d'un vecteur aléatoire:

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}.$$

Cette généralisation est la suivante:

$$f_X\,(x) = (2\pi)^{-n/2} \quad (\det\Omega)^{-\frac{1}{2}} \quad \exp\left\{-\frac{1}{2}\,(x-\mu)'\,\Omega^{-1}\,(x-\mu)\right\}, \quad \text{où:}$$

$$\mu = \begin{pmatrix} E\,(X_1) \\ \vdots \\ E\,(X_n) \end{pmatrix} = E\,(X)$$

est le vecteur des espérances mathématiques des composantes de X, et Ω est une matrice définie positive, dite matrice de covariance, avec

$$[\Omega]_{ii} \quad = \quad V(X_i) \quad = \quad E(X_i-\mu_i)^2 \quad \text{et}$$

$$[\Omega]_{ij} \quad = \quad \mathrm{Cov}(X_i, X_j) \quad = \quad E(X_i-\mu_i)(X_j-\mu_j) \quad .$$

On a donc:

$$\Omega = E\left\{(X-\mu)(X-\mu)'\right\} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & & \vdots \\ \vdots & & \ddots & \\ \sigma_{1n} & & \dots & \sigma_{nn} \end{pmatrix}$$

on écrira $X \sim N(\mu, \Omega)$.

4.2 Fonctions linéaires et quadratiques de variables normales

4.2.1 Fonctions linéaires.

Soit $X \sim N(\mu, \Omega)$, B une matrice $m \times n$ de constantes, de rang m, et A un vecteur $m \times 1$ de constantes. Alors $Y = BX + A$ est $N(B\mu + A, B\Omega B')$.

Nous ne prouverons pas la normalité de Y. Il est néanmoins facile de calculer $E(Y)$ et la matrice de covariance $V(Y)$:

$$E(Y) = E(BX + A) = BE(X) + E(A) = B\mu + A$$

$$\begin{aligned} V(Y) &= E\left[(BX + A - B\mu - A)\,(BX + A - B\mu - A)'\right] \\ &= E\left[(BX - B\mu)\,(BX - B\mu)'\right] \\ &= BE\left[(X-\mu)\,(X-\mu)'\right]B' = B\Omega B' \quad . \end{aligned}$$

Exercice: Un portefeuille contient n actifs financiers de rendements X_i, pour $i = 1, \dots, n$. Ces rendements forment un vecteur X. X est aléatoire de distribution $N(\mu, \Omega)$. Les sommes investies dans chacun des n actifs sont de v_i, pour $i = 1, \dots, n$, et le rendement global du portefeuille est donc de $\pi = \sum_{i=1}^{n} v_i X_i$. L'utilité de ce rendement est égale à $U(\pi) = a - c\exp(-b\pi)$, où a, b, et c sont des paramètres strictement positifs. Montrez que la composition du portefeuille qui maximise l'espérance d'utilité est donnée par le vecteur $v = \frac{1}{b}\Omega^{-1}\mu$. (On utilisera la fonction génératrice des moments d'une variable normale, obtenue à la section 2.3 de la première partie.)

4.2.2 Sous-vecteurs d'un vecteur normal.

Soit $X \sim N(\mu, \Omega)$, partagé comme suit:

$$X = \begin{pmatrix} X_1 \\ \\ X_2 \end{pmatrix} \begin{matrix} n_1 \\ \\ n - n_1 \end{matrix}$$

Nous pouvons alors partager μ et Ω de la façon suivante:

$$\mu = \begin{pmatrix} \mu_1 \\ \\ \mu_2 \end{pmatrix} \begin{matrix} n_1 \\ \\ n - n_1 \end{matrix} \quad , \quad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \begin{matrix} n_1 \\ \\ n - n_1 \end{matrix}$$
$$\begin{matrix} n_1 & n - n_1 \end{matrix}$$

alors $X_1 \sim N(\mu_1, \Omega_{11})$ et $X_2 \sim N(\mu_2, \Omega_{22})$.

Démonstration

Soit B une matrice $n_1 \times n$ définie comme:

$$B \quad = \quad (I_{n_1} \quad O_{n_1 \times (n-n_1)}) .$$

Nous avons $BX = X_1$, et le théorème de la section 4.2.1 nous permet de déterminer la distribution de X_1. Nous avons $X_1 \sim N(B\mu, B\Omega B')$ avec $B\mu = \mu_1$ et

$$\begin{aligned} B\Omega B' &= (I_{n_1} \quad O_{n_1 \times (n-n_1)}) \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \begin{pmatrix} I_{n_1} \\ \\ O_{(n-n_1)\times n_1} \end{pmatrix} \\ &= \Omega_{11} . \end{aligned}$$

La dérivation de la distribution de X_2 se fait de manière analogue.

4.2.3 Formes quadratiques.

Soit $X \sim N(0, I)$ (un vecteur $n \times 1$).
Soit M une matrice non stochastique, idempotente et symétrique de rang k.
Alors $X'MX \sim \chi^2_k$.

Démonstration

En vertu des propriétés de M, il existe une matrice orthogonale C telle que

$$C'MC \quad = \quad \begin{pmatrix} I_k & O_{k\times(n-k)} \\ \\ O_{(n-k)\times k} & O_{(n-k)\times(n-k)} \end{pmatrix} .$$

Soit $Y = C'X$. Nous avons $Y \sim N(0, C'IC)$, c'est-à-dire $Y \sim N(0, I)$. Par conséquent:

$$\begin{aligned} X'MX &= X'(CC')M(CC')X \\ &= X'C(C'MC)C'X \\ &= Y'\begin{pmatrix} I_k & O \\ O & O \end{pmatrix} Y = \sum_{i=1}^{k} Y_i^2 \sim \chi_k^2 \quad . \end{aligned}$$

4.2.4 Indépendance des fonctions linéaires et des formes quadratiques.

Soit X $\sim N(0, I)$ (un vecteur $n \times 1$)
B une matrice $m \times n$ de rang m, non stochastique
M une matrice $n \times n$ idempotente et symétrique, de rang k, non stochastique.

Si $BM = O$, la forme linéaire BX est indépendante de la forme quadratique $X'MX$.

Démonstration

Soit C la matrice orthogonale de la section 4.2.3 et $Y = C'X$.
Soit alors $F = BC = (\, F_1 \quad F_2\,)$ où F_1 est $m \times k$.
On a

$$(\, F_1 \quad F_2\,) \begin{pmatrix} I_k & O \\ O & O \end{pmatrix} = BCC'MC = BMC = O,$$

ce qui implique $F_1 = O$. Alors $BX = BCY = FY = (O \quad F_2)\, Y$ ne dépend que des $n - k$ derniers éléments de Y, qui sont indépendants des k premiers, puisque $Y \sim N(0, I)$. Comme $X'MX = \sum_{i=1}^{k} Y_i^2$, la proposition est démontrée.

4.2.5 Indépendance de deux formes quadratiques.

Soit X $\sim N(0, I)$ (un vecteur $n \times 1$)
M une matrice $n \times n$ idempotente et symétrique de rang k, non stochastique
M^* une matrice $n \times n$ idempotente et symétrique de rang r, non stochastique.

Si $MM^* = O$, alors les formes quadratiques $X'MX$ et $X'M^*X$ sont indépendantes.

Démonstration

Soit C la matrice orthogonale précédente et $Y = C'X$.

Considérons alors la matrice symétrique:

$$G \quad = \quad \begin{pmatrix} G_1 & G_2 \\ G_2' & G_3 \end{pmatrix} \quad = \quad C'M^*C \quad \text{où} \quad G_1 \quad \text{est} \quad k \times k\,.$$

On a:

$$\begin{pmatrix} G_1 & G_2 \\ G_2' & G_3 \end{pmatrix} \begin{pmatrix} I_k & O \\ O & O \end{pmatrix} \quad = \quad C'M^*CC'MC \quad = \quad C'M^*MC = O$$

ce qui implique $G_1 = O$, et $G_2' = O$, donc aussi $G_2 = O$. Par conséquent:

$$X'M^*X = X'CC'M^*CC'X = Y'GY = Y' \begin{pmatrix} O & O \\ O & G_3 \end{pmatrix} Y$$

ne dépend que des $n - k$ derniers éléments de Y, qui sont indépendants des k premiers; comme $X'MX = \sum_{i=1}^{k} Y_i^2$, la proposition est démontrée.

4.3 Application: calcul de la distribution sous H_0 de la statistique t

- Test: $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$
- Echantillon: $y_i \sim N(\mu, \sigma^2)$ indépendantes.
- On a vu au chapitre V de la première partie que la statistique à employer est:

$$t_{\text{obs}} = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \quad \text{avec} \quad s^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2 \quad .$$

- Quelle est la distribution de t_{obs} si H_0 est vraie? On va montrer que $t_{\text{obs}} \sim t_{n-1}$.
- Solution: on peut écrire:

$$t_{\text{obs}} = \frac{\dfrac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{\sum (y_i - \bar{y})^2}{(n-1)\sigma^2}}} = \frac{N}{D}$$

- Si H_0 est vraie, $N \sim N(0, 1)$.
- Nous montrerons au cours que:

$$N = Bx$$
$$\frac{\sum (y_i - \bar{y})^2}{\sigma^2} = x'Mx$$

où:

$$x = \frac{y - \mu_0 i}{\sigma}$$

$$B = \frac{1}{\sqrt{n}} i'$$

$$M = I - \frac{1}{n} ii'$$

i étant un vecteur $n \times 1$ dont tous les éléments sont unitaires.

- Si H_0 est vraie, $x \sim N(0, I)$.
- Nous montrerons au cours que M est symétrique, idempotente, de rang $n - 1$; Nous montrerons de plus que BM est un vecteur nul.
- Alors le théorème de la section 4.2.3 implique que D est la racine d'une χ^2_{n-1} divisée par $n - 1$ et le théorème de la section 4.2.4 implique l'indépendance de N et de D.
- Alors, par définition, $t_{\text{obs}} \sim t_{n-1}$.

CHAPITRE V

LE MODÈLE DE RÉGRESSION MULTIPLE

5.1 Le modèle et ses hypothèses

Les notions présentées dans les deux chapitres précédents vont nous permettre de généraliser les résultats des chapitres I et II à un modèle économétrique possédant un nombre arbitraire k de variables explicatives, soit:

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \ldots + \beta_k x_{tk} + u_t \quad \text{pour} \quad t = 1, 2, \ldots, n \quad .$$

Pour prendre un exemple, il est raisonnable de supposer qu'une loi de demande comprenne comme variable explicative non seulement le prix P_Y du bien demandé, mais aussi le prix P_X d'un substitut et le revenu R du consommateur. Nous aurions alors:

$$y_t = \beta_1 + \beta_2 (P_Y)_t + \beta_3 (P_X)_t + \beta_4 R_t + u_t \quad .$$

Une formulation matricielle du modèle s'impose. Il peut s'écrire sous la forme suivante:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1k} \\ 1 & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & \ldots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

soit $y = X\beta + u$, où y est un vecteur $n \times 1$ d'observations sur la variable dépendante, X est une matrice $n \times k$ d'observations sur les variables explicatives, β est un vecteur $k \times 1$ de paramètres inconnus et u est un vecteur $n \times 1$ d'erreurs aléatoires inobservables.

Nous faisons les hypothèses suivantes:

$$\begin{aligned} &H_1 : E(u) = 0 \\ &H_2 : E(uu') = \sigma^2 I \\ &H_3 : X \text{ est non aléatoire} \\ &H_4 : rang(X) = k < n \quad . \end{aligned}$$

L'hypothèse H_2 implique que les erreurs sont de même variance, et non corrélées. Si l'hypothèse H_4 n'était pas satisfaite, il existerait une relation linéaire exacte entre certaines des colonnes de X: En substituant cette relation dans l'équation de régression, on pourrait alors supprimer un régresseur. Ceci revient à dire que le vecteur β ne pourrait pas être estimé de manière unique.

Notons que nous ne faisons pas encore d'hypothèses sur la *forme fonctionnelle* de la distribution de u.

5.2 Les estimateurs de moindres carrés

L'estimateur $\hat{\beta}$ de moindres carrés sera obtenu, comme précédemment, en minimisant la somme des carrés des résidus. Le vecteur des résidus est $\hat{u} = y - X\hat{\beta}$. Cette somme de carrés peut donc s'écrire:

$$\begin{aligned} \hat{u}'\hat{u} &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \quad . \end{aligned}$$

En utilisant les règles de la Section 3.4, on obtient:

$$\frac{\partial \hat{u}'\hat{u}}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0 \quad .$$

Comme X est de rang k, $X'X$ est définie positive, donc régulière (voir 3.1.3. (1)), et nous pouvons écrire:

$$\hat{\beta} = (X'X)^{-1}X'y \quad .$$

Par ailleurs, les conditions de second ordre pour un minimum sont satisfaites, puisque

$$\frac{\partial}{\partial \hat{\beta}}\left[\frac{\partial \hat{u}'\hat{u}}{\partial \hat{\beta}}\right] = 2X'X \quad ,$$

une matrice définie positive, ce qui montre que $\hat{u}'\hat{u}$ est convexe en $\hat{\beta}$.

5.3 Moments des estimateurs de moindres carrés

5.3.1 Espérance de $\hat{\beta}$.

$\hat{\beta}$ est un estimateur sans biais de β puisque:

$$\begin{aligned} E(\hat{\beta}) &= E\left[(X'X)^{-1}X'(X\beta+u)\right] \\ &= E\left[\beta+(X'X)^{-1}X'u\right] = \beta+(X'X)^{-1}X'E(u)=\beta \quad . \end{aligned}$$

5.3.2 Matrice de covariance de $\hat{\beta}$.

La matrice de covariance de $\hat{\beta}$ est alors:

$$\begin{aligned} V(\hat{\beta}) &= E\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] \\ &= E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}\right] \\ &= (X'X)^{-1}X'E(uu')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} = \sigma^2(X'X)^{-1} \quad . \end{aligned}$$

5.4 Le théorème de Gauss-Markov

Nous allons montrer que $\hat{\beta}$ est le plus efficace des estimateurs *linéaires* de β. Plus précisément, si $\tilde{\beta}$ est un autre estimateur linéaire sans biais de β, c'est-à-dire si $E(\tilde{\beta}) = \beta$ et $\tilde{\beta} = Ay$, les variances de ses composantes ne peuvent être inférieures à celles des composantes de $\hat{\beta}$:

$$V(\tilde{\beta}_i) \geq V(\hat{\beta}_i) \quad , \quad \text{pour} \quad i=1,2,\ldots,k \quad .$$

Démonstration

Soit donc $\tilde{\beta} = Ay$ un autre estimateur linéaire de β. Nous pouvons supposer sans perte de généralité que:

$$A = (X^{'}X)^{-1}X^{'} + C.$$

Alors:

$$\begin{aligned}\tilde{\beta} &= \left[(X^{'}X)^{-1}X^{'} + C\right](X\beta + u) \\ &= \beta + (X^{'}X)^{-1}X^{'}u + CX\beta + Cu = [I + CX]\ \beta + Au\end{aligned}$$

est un estimateur sans biais de β si et seulement si $CX = O$. Nous imposons donc cette condition, qui implique que $\tilde{\beta} = \beta + Au$.

La matrice de covariance de $\tilde{\beta}$ est alors:

$$\begin{aligned}E\left[(\tilde{\beta}-\beta)(\tilde{\beta}-\beta)^{'}\right] &= E\left[Auu^{'}A^{'}\right] \\ &= \sigma^2 AA^{'} \\ &= \sigma^2\left[(X^{'}X)^{-1}X^{'} + C\right]\left[X(X^{'}X)^{-1} + C^{'}\right] \\ &= \sigma^2\left[(X^{'}X)^{-1} + (X^{'}X)^{-1}X^{'}C^{'} + CX(X^{'}X)^{-1} + CC^{'}\right] \\ &= \sigma^2\left[(X^{'}X)^{-1} + CC^{'}\right] \quad \text{puisque} \quad CX = O \\ &= V(\hat{\beta}) + \sigma^2 CC^{'} \quad .\end{aligned}$$

Mais les éléments de la diagonale de $CC^{'}$ sont des sommes de carrés, donc non négatives.

Les variances des composantes de $\tilde{\beta}$ sont donc supérieures ou égales aux variances des composantes de $\hat{\beta}$.

5.5 L'estimation de la variance des erreurs

Comme précédemment (section 1.5) notre estimateur sans biais sera basé sur $\sum(\hat{u}_t - \bar{\hat{u}})^2 = \sum \hat{u}_t^2$ puisque $\bar{\hat{u}} = 0$. (En effet, la première ligne de la matrice $(X^{'}X)$ est le vecteur $i^{'}X$ avec $i^{'} = [1, 1 \ldots 1]$; la première composante du vecteur $X^{'}y$ est $i^{'}y$. La première équation normale s'écrit alors $i^{'}X\hat{\beta} = i^{'}y$, ou $i^{'}(y - X\hat{\beta}) = i^{'}\hat{u} = \sum \hat{u}_t = 0$). Pour trouver, comme précédemment, un estimateur sans biais de σ^2, calculons $E(\hat{u}^{'}\hat{u})$.

Nous avons

$$\begin{aligned} \hat{u} = y - X\hat{\beta} &= X\beta + u - X(X^{'}X)^{-1}X^{'}(X\beta + u) \\ &= X\beta + u - X\beta - X(X^{'}X)^{-1}X^{'}u \\ &= \left[I - X(X^{'}X)^{-1}X^{'}\right]u \underset{\text{def}}{=} Mu\,. \end{aligned}$$

On vérifie aisément que M est idempotente et symétrique.
Alors $\hat{u}^{'}\hat{u} = u^{'}M'Mu = u^{'}Mu$.

$$\begin{aligned} E(\hat{u}^{'}\hat{u}) &= E(u^{'}Mu) = E(tr\, u^{'}Mu) \quad \text{puisque}\, u^{'}Mu \quad \text{est un scalaire} \\ &= E(tr\, Muu^{'}) \quad \text{puisque} \quad trAB = trBA \\ &= trE(Muu^{'}) \quad \text{puisque la trace est une somme} \\ &= trME(uu^{'}) \quad \text{puisque } M \text{ est non aléatoire} \\ &= trM(\sigma^2 I) = \sigma^2 tr(MI) = \sigma^2 trM\,. \end{aligned}$$

$$\begin{aligned} \text{Mais} \quad trM &= trI_n - trX(X^{'}X)^{-1}X^{'} \\ &= trI_n - tr(X^{'}X)(X^{'}X)^{-1} = trI_n - trI_k \\ &= n - k\,. \end{aligned}$$

Alors $E(\hat{u}^{'}\hat{u}) = (n-k)\sigma^2$ et $s^2 = \frac{\hat{u}^{'}\hat{u}}{n-k}$ est un estimateur sans biais de σ^2.

5.6 Décomposition de la variance: les coefficients de détermination $\mathbf{R}^2$ et $\mathbf{R}^2_*$

Nous commençons, comme à la section 1.9, par démontrer une formule de calcul de $\hat{u}'\hat{u}$.

Lemme

$$\hat{u}'\hat{u} = y'y - \hat{\beta}'X'y \quad .$$

Démonstration

$$\begin{aligned}
\hat{u}'\hat{u} &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\
&= y'y - 2\hat{\beta}'X'y + \hat{\beta}'(X'X)\hat{\beta} \\
&= y'y - \hat{\beta}'X'y \quad \text{puisque} \quad (X'X)\hat{\beta} = X'y\,.
\end{aligned}$$

Nous décomposons maintenant, comme précédemment en régression simple (section 1.9), la somme des carrés totaux en une somme de deux termes:

$$\begin{aligned}
\sum (y_t - \bar{y})^2 &= \sum \left(\hat{y}_t - \bar{\hat{y}}\right)^2 + \sum \hat{u}_t^2, \quad \text{soit:} \\
\text{SCT} &= \text{SCE} \quad + \quad \text{SCR}\,.
\end{aligned}$$

Pour démontrer cette identité, notons que $\quad \sum (y_t - \bar{y})^2 = y'y - \frac{(i'y)^2}{n}$

$$\begin{aligned}
\text{et} \quad \sum (\hat{y}_t - \bar{\hat{y}})^2 &= (X\hat{\beta})'(X\hat{\beta}) - \frac{(i'X\hat{\beta})^2}{n} \\
&= \hat{\beta}'(X'X)\hat{\beta} - \frac{(i'y)^2}{n} \\
(\text{puisque} \quad i'y &= i'X\hat{\beta} + i'\hat{u} \quad \text{et} \quad i'\hat{u} = 0) \\
&= \hat{\beta}'X'y - \frac{(i'y)^2}{n} \quad .
\end{aligned}$$

Par le lemme, nous avons $y^{'}y = \hat{u}^{'}\hat{u} + \hat{\beta}^{'}X^{'}y$,

$$\text{donc} \quad \left[y^{'}y - \frac{(i^{'}y)^2}{n}\right] = \left[\hat{\beta}^{'}X^{'}y - \frac{(i^{'}y)^2}{n}\right] + \hat{u}^{'}\hat{u}\,,$$

c'est-à-dire SCT = SCE + SCR , Q.E.D.

Il faut bien noter que cette identité n'est valable que dans un modèle où la somme des résidus est nulle ($i'\hat{u} = 0$). Tel sera bien le cas lorsque le modèle de régression comporte un terme constant, puisque i' est la première ligne de X' et puisque les équations normales impliquent $X'\hat{u} = 0$.

A partir de cette identité, nous pouvons définir, **dans un modèle avec terme constant**, le coefficient de détermination comme:

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}} = \frac{\hat{\beta}^{'}X^{'}y - \frac{(i^{'}y)^2}{n}}{y^{'}y - \frac{(i^{'}y)^2}{n}} \quad .$$

Comme $\frac{\text{SCR}}{n} = \frac{\hat{u}^{'}\hat{u}}{n}$ est un estimateur biaisé de σ^2, il est préférable d'employer le *coefficient de détermination ajusté*, défini comme suit:

$$\bar{R}^2 = 1 - \frac{\text{SCR}/n-k}{\text{SCT}/n-1} = \frac{n-1}{n-k}R^2 - \frac{k-1}{n-k}$$

qui est, lui, basé sur des estimateurs sans biais des variances. Si l'on ajoute un régresseur, R^2 croîtra toujours (non strictement); ceci n'est pas le cas pour $\bar{R}^2$.

Dans un modèle sans terme constant, la somme des résidus n'est pas nécessairement nulle et la décomposition précédente (SCT = SCR +SCE) n'est donc plus valable. Le R^2 précédent n'est donc pas nécessairement compris entre 0 et 1. Néanmoins, on a toujours, en vertu du lemme:

$$y'y = \hat{\beta}'X'y + \hat{u}'\hat{u} = \hat{y}'\hat{y} + \hat{u}'\hat{u}$$

avec $\hat{y} = X\hat{\beta}$.

On peut alors définir:

$$R_*^2 = \frac{\hat{y}'\hat{y}}{y'y} = 1 - \frac{\hat{u}'\hat{u}}{y'y}$$

qui est, lui, toujours compris entre 0 et 1. Ce coefficient R_*^2 peut être utilisé dans tous les cas, tant dans un modèle sans constante que dans un modèle avec constante. Mais son interprétation est différente de celle du R^2.

Comme précédemment, nous pouvons ajuster ce dernier coefficient de détermination aux nombres de degrés de liberté, comme suit:

$$\bar{R}_*^2 = 1 - \frac{\hat{u}'\hat{u}/(n-k)}{y'y/(n-1)} = \frac{n-1}{n-k}R_*^2 - \frac{k-1}{n-k}.$$

Interprétation des coefficients de détermination:

Nous verrons plus loin que R^2 est une fonction monotone de la statistique F à employer pour tester la nullité de tous les coefficients de régression sauf la constante.

Nous verrons aussi que R_*^2 est une fonction monotone de la statistique F à employer pour tester la nullité de tous les coefficients, constante comprise.

On peut montrer que R^2 est le carré du coefficient de corrélation entre les valeurs observées y_t et les valeurs $\hat{y}_t$ calculées à l'aide de l'équation de régression estimée.

5.7 Problèmes particuliers: multicolinéarité, biais de spécification, variables muettes

5.7.1 Multicolinéarité.

(1) Comme nous l'avons déjà mentionné, l'existence d'une relation linéaire exacte entre les colonnes de X nous empêche de déterminer l'estimateur $\hat{\beta}$ de manière unique. Ce cas est un cas extrême de multicolinéarité. Mais il arrive souvent que certaines des colonnes de X présentent une dépendance linéaire *approximative*. Les conséquences de ce phénomène sont les suivantes:

- un manque de précision dans les estimations des β_i, se traduisant par de fortes variances;
- les estimations des β_i présenteront souvent des distortions importantes, dues à des raisons numériques. Le nombre de chiffres significatifs des emplacements-mémoire d'un ordinateur est en effet limité, ce qui se traduit par un manque de stabilité des programmes d'inversion matricielle, pour des matrices qui sont presque singulières.

Pour illustrer le premier point, reprenons le modèle de régression simple $y_t = a + bx_t + u_t$. Nous avons vu que

$$V(\hat{b}) = \frac{\sigma^2}{\sum(x_t - \bar{x})^2} \quad .$$

La multicolinéarité se traduira dans ce cas par une série d'observations (x_t) presque constante, c'est-à-dire par $x_t \approx \bar{x}$ pour tout t. On a alors $\sum(x_t - \bar{x})^2 \approx 0$, ce qui se traduit par une forte variance de $\hat{b}$.

(2) La multicolinéarité peut être mesurée en calculant le rapport $\frac{\lambda \max}{\lambda \min}$ de la plus grande à la plus petite valeur propre de $X'X$.

(3) Pour corriger le problème de multicolinéarité, on peut:

- soit ajouter des observations à l'échantillon quand la chose est possible; il faut néanmoins que les observations supplémentaires ne présentent pas de multicolinéarité!
- Soit introduire une information a priori. Supposons par exemple que dans la fonction de production:

$$\log Q_t = A + \alpha \log K_t + \beta \log L_t + u_t$$

les variables $\log K_t$ et $\log L_t$ soient fortement colinéaires. Si l'on sait que les rendements d'échelle sont constants ($\alpha + \beta = 1$), on peut transformer le modèle comme suit:

$$\begin{aligned} &\log Q_t = A + \alpha \log K_t + (1-\alpha) \log L_t + u_t \\ \text{ou} \quad &(\log Q_t - \log L_t) = A + \alpha(\log K_t - \log L_t) + u_t, \end{aligned}$$

ce qui a donc pour effet de supprimer un régresseur. Ceci peut résoudre le problème. Essentiellement, l'information a priori $\alpha + \beta = 1$ supplée au défaut d'information présent dans l'échantillon (tentative d'estimer trop de paramètres avec trop peu de données).

Cette information a priori peut également prendre une forme stochastique, non déterministe. Nous étudierons ce point lorsque nous verrons les méthodes bayesiennes.

5.7.2 Biais de spécification.

Examinons maintenant le problème du choix d'une forme fonctionnelle, c'est-à-dire du choix de la liste des régresseurs. Comme nous allons le montrer, l'omission d'une variable explicative a pour conséquence, en général, un biais de l'estimateur $\hat{\beta}$.

Supposons que y soit engendré par le modèle:

$$y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u \quad , \quad \text{avec} \quad \beta_2 \neq 0 \quad \text{et} \quad E(u) = 0$$

et que l'on omette les colonnes de X_2 de la liste des régresseurs. On estimerait alors par moindres carrés le modèle

$$y = X_1\beta_1 + u^* \quad \text{avec} \quad u^* = X_2\beta_2 + u$$

et par conséquent $E(u^*) = X_2\beta_2 \neq 0$. L'estimateur:

$$\hat{\beta}_1 = (X'_1 X_1)^{-1} X_1^{'} y = \beta_1 + (X'_1 X_1)^{-1} X_1^{'} u^*$$

sera biaisé.

5.7.3 Variables muettes.

Une variable muette, ou binaire (en anglais: *dummy variable*) est une variable du type

$$D_t = 1 \quad \text{si} \quad t \in T_1$$

$$D_t = 0 \quad \text{si} \quad t \notin T_1$$

où $T_1 \subseteq \{1, 2, \ldots, n\}$.

Une telle variable, incluse dans la liste des régresseurs, pourrait par exemple indiquer la présence ou l'absence de guerre, ou classifier des données selon un critère saisonnier. Pour des données mensuelles, s'il n'y pas de variations saisonnières à l'intérieur d'un même trimestre, on pourrait poser:

$D_{1t} = 1$ si t est un mois du premier trimestre, 0 sinon
$D_{2t} = 1$ si t est un mois du second trimestre, 0 sinon
$D_{3t} = 1$ si t est un mois du troisième trimestre, 0 sinon
$D_{4t} = 1$ si t est un mois du quatrième trimestre, 0 sinon.

Les quatre colonnes des régresseurs D_1, D_2, D_3, D_4 pour les 12 mois d'une année auraient alors la forme suivante:

$$\begin{matrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{matrix}$$

$$\begin{matrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}$$

$$\begin{matrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{matrix}$$

$$\begin{matrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{matrix}$$

Nous ne pourrions pas inclure de constante dans ce modèle, puisque la somme de ces quatre vecteurs est un vecteur de uns. On aurait alors colinéarité parfaite. Les coefficients des variables D_i sont en fait des constantes spécifiques à chaque saison.

Une autre possibilité serait d'inclure une constante, et de supprimer l'une des variables D_i, par exemple D_1. Les coefficients de D_2, D_3 et D_4 mesureraient alors l'effet *relatif* des facteurs saisonniers: les constantes spécifiques seraient β_1, $\beta_1 + \beta_2$, $\beta_1 + \beta_3$, $\beta_1 + \beta_4$ plutôt que β_1, β_2, β_3, β_4.

Notons aussi que les variables muettes permettent la spécification de *pentes* variables. Si $D_t = 1$ pour une période de guerre, $= 0$ sinon, et que l'on a des raisons de penser que la propension marginale à consommer β dans le modèle:

$$C_t = \alpha + \beta Y_t + u_t$$

est différente en temps de paix et en temps de guerre, on pourra estimer les paramètres du modèle:

$$C_t = \alpha + bD_tY_t + c(1 - D_t)Y_t + u_t$$

et $\hat{b}$ sera l'estimateur de la propension marginale à consommer en temps de guerre, $\hat{c}$ l'estimateur de cette propension en temps de paix.

5.8 Estimateurs par maximum de vraisemblance

Nous faisons ici l'hypothèse que le vecteur u a une distribution normale:

$$H_5 \quad u \sim N(0, \sigma^2 I) \quad .$$

Ce qui implique que $y - X\beta \sim N(0, \sigma^2 I)$.

La fonction de vraisemblance s'écrit alors:

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (y - X\beta)' (y - X\beta) \right\}$$

$$\text{et} \quad \log_e L = -\frac{n}{2} \log_e 2\pi - \frac{n}{2} \log_e \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)' (y - X\beta) \quad .$$

Nous avons alors les conditions de premier ordre suivantes:

$$\frac{\partial \log_e L}{\partial \beta} = -\frac{1}{2\sigma^2} (-2X'y + 2X'X\beta) = 0 \quad \text{(voir Section 5.2).}$$

$$\frac{\partial \log_e L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\beta)' (y - X\beta) = 0 \quad .$$

La première condition implique $\hat{\beta} = (X'X)^{-1}X'y$. En remplaçant β par $\hat{\beta}$ dans la seconde condition et en la multipliant par $2\sigma^2$, on obtient $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n}$ comme estimateur de σ^2 par **maximum de vraisemblance**.

La matrice Hessienne H s'obtient en dérivant le vecteur

$$\begin{pmatrix} -\frac{1}{2\sigma^2}\left(-2X'y+2X'X\beta\right) \\ -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\left(y-X\beta\right)'\left(y-X\beta\right) \end{pmatrix}$$

par rapport au vecteur $\begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}$. Ceci donne:

$$H = \begin{pmatrix} -\frac{(X'X)}{\sigma^2} & \frac{1}{\sigma^4}\left(-X'y+X'X\beta\right) \\ \frac{1}{\sigma^4}\left(-y'X+\beta'X'X\right) & \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\left(y-X\beta\right)'\left(y-X\beta\right) \end{pmatrix} .$$

En remplaçant β par $\hat{\beta}=(X'X)^{-1}X'y$ et σ^2 par $\hat{\sigma}^2=\frac{1}{n}(y-X\hat{\beta})'(y-X\hat{\beta})$, on obtient:

$$H = \begin{pmatrix} -\frac{(X'X)}{\hat{\sigma}^2} & O_{k\times 1} \\ O_{1\times k} & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix} .$$

qui est définie négative puisque $(X'X)$ est définie positive et $\hat{\sigma}^2>0$. Nous avons donc bien un maximum.

5.9 Exemple numérique

Une association de vignerons vaudois voudrait étudier l'influence sur la production de vin par hectare (Y) des quantités de main-d'oeuvre (X_1) et d'engrais (X_2) employées par hectare.

Une enquête est menée chez dix vignerons ($i = 1, \ldots, 10$) et l'on postule la forme fonctionnelle suivante:

$$\log Y_i = \beta_1 + \beta_2 \log X_{1i} + \beta_3 \log X_{2i} + u_i$$

où u_i est un terme d'erreur aléatoire satisfaisant nos hypothèses. Les données de l'échantillon sont résumées dans la matrice suivante:

$$\begin{pmatrix} \sum(\log Y)^2 & \sum \log Y & \sum \log Y \log X_1 & \sum \log Y \log X_2 \\ \sum \log Y & n & \sum \log X_1 & \sum \log X_2 \\ \sum \log Y \log X_1 & \sum \log X_1 & \sum(\log X_1)^2 & \sum \log X_1 \log X_2 \\ \sum \log X_2 \log Y & \sum \log X_2 & \sum \log X_2 \log X_1 & \sum(\log X_2)^2 \end{pmatrix}$$

$$= \begin{pmatrix} 19.34 & 11.8 & 7.1 & 4.1 \\ 11.8 & 10 & 2 & 2 \\ 7.1 & 2 & 7 & 1 \\ 4.1 & 2 & 1 & 7 \end{pmatrix} .$$

On a:

$$(X'X) = \begin{pmatrix} 10 & 2 & 2 \\ 2 & 7 & 1 \\ 2 & 1 & 7 \end{pmatrix}$$

$$X'y = \begin{pmatrix} 11.8 \\ 7.1 \\ 4.1 \end{pmatrix} \quad \text{et} \quad y'y = 19.34$$

$$(X'X)^{-1} = \frac{1}{432} \begin{pmatrix} 48 & -12 & -12 \\ -12 & 66 & -6 \\ -12 & -6 & 66 \end{pmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{pmatrix} 1 \\ 0.7 \\ 0.2 \end{pmatrix}$$

$$\begin{aligned} \hat{\beta}' X'y &= 17.59 \\ \hat{u}'\hat{u} &= 19.34 - 17.59 = 1.75 \\ s^2 &= 0.25 \\ R^2 &= 1 - \frac{1.75}{19.34 - \frac{(11.8)^2}{10}} = 0.677 \\ \bar{R}^2 &= \frac{9}{7}(0.677) - \frac{2}{7} = 0.585 \quad . \end{aligned}$$

Les résultats peuvent être résumés de la façon suivante (les estimations des écarts-types se trouvent entre parenthèses):

$$\begin{array}{lccccccc} \log \hat{Y} = & 1 & + & 0.7 \quad \log X_1 & + & 0.2 \quad \log X_2 & & (\bar{R}^2 = 0.585) \\ & (0.167) & & (0.195) & & (0.195). & & \end{array}$$

CHAPITRE VI

MOINDRES CARRÉS SOUS CONTRAINTES LINÉAIRES

6.1 L'estimateur de β sous contraintes

Nous dériverons dans ce chapitre l'estimateur $\hat{\beta}_c$ du vecteur β sous un système de J contraintes indépendantes, qui peut s'écrire sous la forme:

$$R\hat{\beta}_c = r \quad ,$$

où R est une matrice $J \times k$ de rang J , r est un vecteur $J \times 1$, et $\hat{\beta}_c$ est le vecteur des estimateurs de β sous contraintes.

Dans notre exemple précédent, nous pourrions vouloir imposer la contrainte que les rendements d'échelle sont constants, c'est-à-dire estimer les paramètres β_1, β_2, et β_3 de:

$$\log Y = \beta_1 + \beta_2 \log X_1 + \beta_3 \log X_2 + u \quad ,$$

sous la contrainte $\hat{\beta}_{2c} + \hat{\beta}_{3c} = 1$, où $\hat{\beta}_{2c}$ et $\hat{\beta}_{3c}$ sont les estimations contraintes de β_2 et β_3. On aurait alors:

$$R \quad = \quad (\,0 \quad 1 \quad 1\,) \quad \text{et} \quad r = 1 \quad .$$

Notons que ce problème pourrait aussi être résolu par substitution; c'est ce que nous avons fait à la section 5.7.1 (3). Mais une présentation matricielle nous sera très utile lorsque nous verrons, au chapitre 7, le test de $R\beta = r$.

Nous minimisons la somme des carrés des résidus sous les contraintes du système $R\hat{\beta}_c = r$. A cette fin, nous écrivons ce système comme $2(R\hat{\beta}_c - r) = 0$, et nous formons le Lagrangien:

$$\phi \quad = \quad (y - X\hat{\beta}_c)'(y - X\hat{\beta}_c) - 2\lambda'(R\hat{\beta}_c - r)$$

où λ' est un vecteur ligne de J multiplicateurs de Lagrange. Le système de conditions de premier ordre peut s'écrire:

$$(1) \qquad \frac{\partial \phi}{\partial \hat{\beta}_c} = -2X'y + 2(X'X)\hat{\beta}_c - 2R'\lambda = 0$$

$$(2) \qquad \frac{\partial \phi}{\partial \lambda} = -2(R\hat{\beta}_c - r) = 0 \quad .$$

En vertu de (1), on a:

$$(3) \qquad \hat{\beta}_c = \hat{\beta} + (X'X)^{-1}R'\lambda$$

$$\text{où} \quad \hat{\beta} = (X'X)^{-1}X'y \quad \text{est l'estimateur sans contraintes.}$$

En prémultipliant par R:

$$R\hat{\beta}_c = R\hat{\beta} + R(X'X)^{-1}R'\lambda$$

$$= r \quad \text{(en vertu de (2))} \; .$$

Ceci implique $\lambda = \left[R(X'X)^{-1}R'\right]^{-1}[r - R\hat{\beta}]$.

En substituant dans (3), il vient:

$$(4) \qquad \hat{\beta}_c = \hat{\beta} + (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}[r - R\hat{\beta}] \quad .$$

On constate que $\hat{\beta}_c$ (l'estimateur contraint) diffère de $\hat{\beta}$ (l'estimateur non contraint) par une fonction linéaire du vecteur $r - R\hat{\beta}$. Ce dernier vecteur sera nul si le vecteur $\hat{\beta}$ vérifie les restrictions a priori .

6.2 Efficacité de l'estimateur de β sous contraintes

Nous allons maintenant montrer que *si les restrictions a priori sont vérifiées par le vecteur* β (c.à.d. par les vraies valeurs des paramètres à estimer), l'estimateur $\hat{\beta}_c$ est au moins aussi efficace que l'estimateur $\hat{\beta}$; en particulier,

$$E(\hat{\beta}_c) = \beta \quad \text{et} \quad V(\hat{\beta}_{ic}) \leq V(\hat{\beta}_i) \quad \text{pour tout} \quad i.$$

En substituant $\hat{\beta} = \beta + (X'X)^{-1}X'u$ dans (4), il vient:

$$\begin{aligned}\hat{\beta}_c &= \beta + (X'X)^{-1}X'u + (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}\left[r - R\beta - R(X'X)^{-1}X'u\right] \\ &= \beta + \left[I - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R\right](X'X)^{-1}X'u \\ &\quad \text{sous l'hypothèse} \quad R\beta = r \\ &\underset{\text{def}}{=} \beta + A(X'X)^{-1}X'u \quad .\end{aligned}$$

Comme A est non stochastique, on a $E(\hat{\beta}_c) = \beta$,

$$\begin{aligned}\text{et} \quad V(\hat{\beta}_c) &= E\left[\hat{\beta}_c - \beta\right]\left[\hat{\beta}_c - \beta\right]' \\ &= A(X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1}A' \\ &= \sigma^2 A(X'X)^{-1}A' \quad .\end{aligned}$$

On vérifie aisément que si $V = \sigma^2(X'X)^{-1} = V(\hat{\beta})$, alors:

$$\sigma^2 A(X'X)^{-1}A' = V - VR'(RVR')^{-1}RV$$

ou: $V(\hat{\beta}_c) = V(\hat{\beta}) - VR'(RVR')^{-1}RV$.

Comme la seconde matrice de la différence est définie non négative, les éléments de sa diagonale sont non négatifs et $V(\hat{\beta}_{ic}) \leq V(\hat{\beta}_i)$, *Q.E.D.*

Exemple

Reprenons le modèle et les données de la section 5.9. Nous voulons imposer la contrainte que les rendements d'échelle sont constants. On a:

$$r = 1, \quad R = [0 \quad 1 \quad 1]$$

$$r - R\hat{\beta} = 1 - 0.7 - 0.2 = 0.1$$

$$R(X'X)^{-1}R' = \frac{1}{432} \quad (66 - 6 + 66 - 6) = \frac{10}{36}$$

et donc:

$$\hat{\beta}_c = \begin{pmatrix} 1 \\ 0.7 \\ 0.2 \end{pmatrix} + \begin{pmatrix} 36 \\ 10 \end{pmatrix} (0.1) \frac{1}{432} \begin{pmatrix} 48 & -12 & -12 \\ -12 & 66 & -6 \\ -12 & -6 & 66 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 0.7 \\ 0.2 \end{pmatrix} + \begin{pmatrix} -0.02 \\ 0.05 \\ 0.05 \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.75 \\ 0.25 \end{pmatrix} .$$

6.3 Décomposition de la somme des résidus contraints

Nous allons voir dans cette section que la somme des carrés des résidus contraints est toujours supérieure ou égale à la somme des carrés des résidus non contraints. Ceci a une conséquence sur le R^2.

Soit $\hat{u}_c = y - X\hat{\beta}_c$ le vecteur des résidus contraints. On a:

$$\begin{aligned} \hat{u}_c'\hat{u}_c &= (y - X\hat{\beta}_c)'(y - X\hat{\beta}_c) \\ &= (y - X\hat{\beta} + X\hat{\beta} - X\hat{\beta}_c)'(y - X\hat{\beta} + X\hat{\beta} - X\hat{\beta}_c) \\ &= (\hat{u} + X[\hat{\beta} - \hat{\beta}_c])'(\hat{u} + X[\hat{\beta} - \hat{\beta}_c]) \\ &= \hat{u}'\hat{u} + 2(\hat{\beta} - \hat{\beta}_c)'X'\hat{u} + (\hat{\beta} - \hat{\beta}_c)'X'X(\hat{\beta} - \hat{\beta}_c) \\ &= \hat{u}'\hat{u} + (\hat{\beta} - \hat{\beta}_c)'X'X(\hat{\beta} - \hat{\beta}_c). \end{aligned}$$

Mais le second terme de cette somme est positif ou nul, car $X'X$ est définie positive. On a donc :

$$\hat{u}_c'\hat{u}_c \geq \hat{u}'\hat{u}$$

et comme:

$$R_c^2 = 1 - \frac{\hat{u}_c'\hat{u}_c}{\sum(y_t - \bar{y})^2}$$

$$R^2 = 1 - \frac{\hat{u}'\hat{u}}{\sum(y_t - \bar{y})^2}$$

ceci implique $R_c^2 \leq R^2$.

On peut aussi noter (ceci nous sera utile au chapitre suivant) que si $u \sim N(0, \sigma^2 I)$, l'estimateur $\hat{\beta}_c$ maximise la vraisemblance sous la contrainte $R\hat{\beta}_c = r$.

CHAPITRE VII.

INFÉRENCE STATISTIQUE EN RÉGRESSION CLASSIQUE

7.1 Le test de l'hypothèse linéaire générale

Nous allons tout d'abord présenter la théorie générale du test de J contraintes indépendantes de la forme discutée plus haut. Ce test inclut comme cas particulier tous les tests mentionnés au chapitre II; nous réexaminerons ces tests à la section 7.2 dans le cadre de la régression multiple. Soit donc à tester:

$$\begin{aligned} & H_0 : R\beta = r \\ \text{contre} \quad & H_1 : R\beta \neq r \quad , \end{aligned}$$

R étant, rappelons-le, une matrice $J \times k$ de constantes connues de rang J, et r étant un vecteur $J \times 1$.

Nous allons d'abord utiliser la méthode du rapport des vraisemblances pour trouver une statistique appropriée; en utilisant les résultats de la section 4.2, nous déterminerons ensuite la distribution de cette statistique.

7.2 Dérivation de la statistique F à l'aide du critère du rapport des vraisemblances

Nous introduisons l'hypothèse:

$$H_5 : u \sim N(0, \sigma^2 I) \quad .$$

La vraisemblance s'écrit alors:

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\} \quad .$$

Rappelons la définition du rapport des vraisemblances λ:

$$\lambda = \frac{\max_{H_0} L(\beta, \sigma^2)}{\max_{\Omega} L(\beta, \sigma^2)};$$

on rejette H_0 si λ est proche de 0.

L'estimation du modèle sous H_0 et sous Ω a déjà été traitée. On avait obtenu sous H_0:

$$\hat{\beta}_c = \hat{\beta} + (X'X)^{-1}R^{'}\left[R(X'X)^{-1}R^{'}\right]^{-1}\left[r - R\hat{\beta}\right]$$

$$\hat{\sigma}_c^2 = \frac{1}{n}(y - X\hat{\beta}_c)'(y - X\hat{\beta}_c) = \frac{1}{n}\hat{u}_c'\hat{u}_c,$$

et sous Ω:

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})'(y - X\hat{\beta}) = \frac{1}{n}\hat{u}'\hat{u}.$$

Il suffit de remplacer, dans l'expression de λ, β et σ^2 par ces valeurs. En faisant les substitutions, on obtient:

$$\begin{aligned}
\lambda &= \frac{L(\hat{\beta}_c, \hat{\sigma}_c^2)}{L(\hat{\beta}, \hat{\sigma}^2)} \\
&= \frac{(2\pi)^{-n/2}(\hat{\sigma}_c^2)^{-n/2}\exp\left[-\dfrac{n\hat{\sigma}_c^2}{2\hat{\sigma}_c^2}\right]}{(2\pi)^{-n/2}(\hat{\sigma}^2)^{-n/2}\exp\left[-\dfrac{n\hat{\sigma}^2}{2\hat{\sigma}^2}\right]} \\
&= \left(\frac{\hat{\sigma}_c^2}{\hat{\sigma}^2}\right)^{-n/2} \\
&= \left(\frac{\hat{u}_c'\hat{u}_c}{\hat{u}'\hat{u}}\right)^{-n/2} \\
&= \left(\frac{\hat{u}'\hat{u} + \hat{u}_c'\hat{u}_c - \hat{u}'\hat{u}}{\hat{u}'\hat{u}}\right)^{-n/2} \\
&= \left(1 + \frac{Q}{\hat{u}'\hat{u}}\right)^{-n/2}
\end{aligned}$$

où:

$$Q = \hat{u}_c'\hat{u}_c - \hat{u}'\hat{u}.$$

Nous avons déjà démontré, à la section 6.3, que:

$$Q = (\hat{\beta} - \hat{\beta}_c)'X'X(\hat{\beta} - \hat{\beta}_c).$$

Nous montrerons au cours que, de plus:

$$Q = (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$$

et que, **si H_0 est vraie**:

$$Q = u'Lu, \quad \text{avec:}$$

$$L = X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'.$$

Nous avons donc au total quatre expressions équivalentes de la forme quadratique Q; la dernière ne vaut que sous H_0. Selon le contexte, certaines de ces expressions seront plus utiles que les autres. La dernière expression nous servira, à la section suivante, à trouver la distribution d'une fonction monotone de λ sous H_0, donnée par:

$$F = (\lambda^{-2/n} - 1)\frac{n-k}{J}.$$

7.3 Distribution sous H_0 de la statistique F

Nous invitons le lecteur à vérifier, à titre d'exercice, que la matrice L définie à la section précédente vérifie:

(1) $$L = L'$$

(2) $$LL = L$$

(3) $$\text{rang}(L) = \text{trace}(L) = J$$

(4) $$\text{si} \quad M = I - X(X'X)^{-1}X', \quad \text{alors} \quad LM = O.$$

.

Le fait que $u'Mu = \hat{u}'\hat{u}$ et les résultats de la section 4.2 impliquent alors, puisque $\frac{u}{\sigma} \sim N(0, I)$:

$$\frac{Q}{\sigma^2} = \left(\frac{u}{\sigma}\right)' L \left(\frac{u}{\sigma}\right) \sim \chi^2_J \quad \text{sous } H_0$$

$$\frac{\hat{u}'\hat{u}}{\sigma^2} = \left(\frac{u}{\sigma}\right)' M \left(\frac{u}{\sigma}\right) \sim \chi^2_{n-k}$$

et ces deux variables aléatoires sont indépendantes puisque $LM = O$.

Par conséquent:

$$F_{obs} = \frac{Q}{Js^2} = \frac{Q/J}{\hat{u}'\hat{u}/(n-k)} = \frac{Q/[\sigma^2 J]}{\hat{u}'\hat{u}/[\sigma^2(n-k)]}$$

est un rapport de deux χ^2 indépendantes divisées par leurs nombres de degrés respectifs et a la distribution $F_{J,n-k}$ sous H_0.

En utilisant:

$$\lambda = \left(1 + \frac{Q}{\hat{u}'\hat{u}}\right)^{-n/2}$$

il est facile de montrer que:

$$F_{obs} = (\lambda^{-2/n} - 1)\frac{n-k}{J}.$$

Les petites valeurs de λ correspondent donc à de grandes valeurs de F_{obs}.

En utilisant:

$$Q = \hat{u}_c'\hat{u}_c - \hat{u}'\hat{u}$$

il est facile de montrer que:

$$F_{obs} = \left(\frac{\hat{\sigma}_c^2}{\hat{\sigma}^2} - 1\right)\frac{n-k}{J}.$$

Donc pour calculer F_{obs}, il suffit d'estimer les modèles contraints et non contraints et de comparer les variances estimées.

7.4 Dérivation de la statistique F à l'aide du critère de Wald

A la section 5.4 de la première partie, nous avions énoncé la statistique de Wald pour le test d'une hypothèse portant sur un seul paramètre inconnu θ_i, et nous avions vu que cette statistique:

$$\mathcal{W} = \frac{(\hat{\theta}_i - \theta_0)^2}{\hat{V}(\hat{\theta}_i)}$$

pouvait être interprétée comme le carré d'une distance entre les estimations sous les hypothèses nulle et alternative.

Ici, nous avons un test joint de J hypothèses: celui de $H_0 : R\beta = r$ contre $H_1 : R\beta \neq r$. En posant $R\beta = \theta$, on peut considérer ce test comme celui d'une hypothèse nulle sur θ. L'expression précédente va devenir une forme quadratique, qui peut être interprétée comme le carré d'une distance dans un espace à J dimensions. L'expression précédente peut être généralisée comme suit:

$$\mathcal{W} = (R\hat{\beta} - r)'[\hat{V}(R\hat{\beta})]^{-1}(R\hat{\beta} - r)$$

où $\hat{\beta}$ est l'estimation de β par maximum de vraisemblance et où $\hat{V}(R\hat{\beta})$ est l'estimation par maximum de vraisemblance de la matrice de covariance de $R\hat{\beta}$. On a:

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\begin{aligned} V(R\hat{\beta}) &= R[\sigma^2(X'X)^{-1}]R' \\ &= \sigma^2 R(X'X)^{-1}R' \end{aligned}$$

$$\hat{V}(R\hat{\beta}) = \hat{\sigma}^2 R(X'X)^{-1}R'$$

avec $\hat{\sigma}^2 = \hat{u}'\hat{u}/n$. En substituant et en utilisant $F_{obs} = Q/(Js^2)$, on obtient:

$$\mathcal{W} = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\hat{\sigma}^2}$$

$$= \frac{J\left(\frac{Q}{J}\right)}{\frac{n-k}{n}s^2}$$

$$= \frac{Jn}{n-k}F_{obs}.$$

Donc:

$$F_{obs} = \frac{n-k}{Jn}\mathcal{W}$$

est bien une fonction monotone de la statistique de Wald.

7.5 Dérivation de F à partir du critère des multiplicateurs de Lagrange

A la section 5.5 de la première partie, nous avions formulé la statistique LM pour le test d'une hypothèse $H_0 : \theta_i = \theta_0$ comme:

$$LM = \frac{\hat{\lambda}_0^2}{\hat{V}_0(\lambda)}$$

$\hat{\lambda}_0$ étant la valeur du multiplicateur de Lagrange λ évaluée aux estimations contraintes des paramètres, et $\hat{V}_0(\lambda)$ l'estimation contrainte de la variance de λ.

Dans ce cas-ci, on a J contraintes, donc un vecteur de J multiplicateurs de Lagrange. La statistique LM va donc devenir une forme quadratique, et la variance précédente sera remplacée par une matrice de covariance.

A la section 6.1 de la seconde partie, on a vu que le vecteur des multiplicateurs de Lagrange pour la minimisation contrainte de la somme des carrés des résidus était égal à:

$$\lambda = [R(X'X)^{-1}R']^{-1}(r - R\hat{\beta}). \tag{1}$$

Comme ce vecteur ne dépend pas de paramètres inconnus, il est ici égal à $\hat{\lambda}_0$. D'autre part, comme il est proportionnel au vecteur des multiplicateurs de Lagrange pour la maximisation contrainte de la vraisemblance, on peut l'utiliser pour dériver la statistique LM (le facteur de proportionnalité se simplifie). Sa matrice de covariance est la suivante:

$$\begin{aligned} V(\lambda) &= [R(X'X)^{-1}R']^{-1}V(R\hat{\beta})[R(X'X)^{-1}R']^{-1} \\ &= \sigma^2[R(X'X)^{-1}R']^{-1}. \end{aligned}$$

Donc:

$$\hat{V}_0(\lambda) = \hat{\sigma}_0^2[R(X'X)^{-1}R']^{-1} \tag{2}$$

où $\hat{\sigma}_0^2 = \hat{u}_c'\hat{u}_c/n$.

En utilisant (1) et (2), il vient:

$$\begin{aligned} LM &= \hat{\lambda}_0'[\hat{V}_0(\lambda)]^{-1}\hat{\lambda}_0 \\ &= \frac{(R\hat{\beta}-r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-r)}{\hat{\sigma}_0^2} \\ &= \frac{Q}{\hat{\sigma}_0^2}. \end{aligned}$$

Montrons maintenant que F_{obs} est une transformation monotone de LM. On a vu à la section 7.2 que:

$$Q = (\hat{\beta}-\hat{\beta}_c)'X'X(\hat{\beta}-\hat{\beta}_c) = \hat{u}_c'\hat{u}_c - \hat{u}'\hat{u}.$$

Donc:

$$\begin{aligned} \frac{1}{LM} = \frac{\hat{\sigma}_0^2}{Q} &= \frac{\hat{\sigma}^2 + Q/n}{Q} = \frac{1}{n} + \frac{\hat{\sigma}^2}{Q} \\ &= \frac{1}{n} + \frac{\frac{n-k}{n}s^2}{J\left(\frac{Q}{J}\right)} \\ &= \frac{JF_{obs}+n-k}{nJF_{obs}} \end{aligned}$$

et donc:

$$LM = \frac{nJF_{obs}}{JF_{obs}+n-k}.$$

7.6 Cas particuliers du test de l'hypothèse linéaire générale

7.6.1 Test sur un coefficient individuel.

Si nous voulons tester:

$$\begin{aligned} & H_0 : \beta_i = \beta_i^0 \\ \text{contre} \quad & H_1 : \beta_i \neq \beta_i^0 \end{aligned}$$

la matrice R prendra la forme

$$R \quad = \quad (0 \quad 0 \quad \dots \quad 0 \quad 1 \quad 0 \quad \dots \quad 0)$$

où l'unité apparait en ième position. r est le scalaire β_i^0.

On obtient alors:

$$F_{obs} \quad = \quad \frac{(\hat{\beta}_i - \beta_i^0)^2}{s^2 \left[(X'X)^{-1}\right]_{ii}} \sim F_{1;n-k} = t_{n-k}^2$$

et la statistique

$$t_{obs} \quad = \quad \frac{(\hat{\beta}_i - \beta_i^0)}{s\sqrt{\left[(X'X)^{-1}\right]_{ii}}}$$

suit une loi de Student avec $n-k$ degrés de liberté sous H_0.

7.6.2 Test de nullité de tous les coefficients; lien avec R_*^2.

Si nous voulons tester:

$$\begin{aligned} & H_0 : \beta = 0 \\ \text{contre} \quad & H_1 : \beta \neq 0\,. \end{aligned}$$

La matrice R n'est autre que la matrice unité d'ordre k. Le vecteur r est le vecteur nul (de dimensions $k \times 1$).

On a alors:

$$F_{obs} \quad = \quad \frac{\hat{\beta}'(X'X)\hat{\beta}/k}{s^2} \sim F_{k;n-k} \quad \text{sous } H_0.$$

Il est intéressant d'établir un lien entre cette statistique et le R_*^2, car ceci nous permettra d'énoncer des valeurs critiques pour ce dernier. La statistique peut s'écrire:

$$\begin{aligned} F_{obs} &= \left(\frac{\hat{y}'\hat{y}}{\hat{u}'\hat{u}}\right)\left(\frac{n-k}{k}\right) \\ &= \left(\frac{\hat{y}'\hat{y}/y'y}{\hat{u}'\hat{u}/y'y}\right)\left(\frac{n-k}{k}\right) \\ &= \left(\frac{R_*^2}{1-R_*^2}\right)\left(\frac{n-k}{k}\right). \end{aligned}$$

Donc F_{obs} est bien une fonction monotone du R_*^2. Sa réciproque est donnée par:

$$R_*^2 = \frac{kF_{obs}}{n-k+kF_{obs}}$$

et R_*^2 est donc significatif (de manière équivalente, on rejettera H_0) si:

$$R_*^2 > \frac{kF_{k,n-k,\alpha}}{n-k+kF_{k,n-k,\alpha}}.$$

Ceci indique que le seuil critique de R_*^2 tend vers zéro lorsque le nombre d'observations n tend vers l'infini. Par exemple, un R_*^2 de $0,10$ sera significatif au seuil $\alpha = 0,05$ si $n = 122$ et $k = 2$; mais il ne le sera pas pour $k = 2$ et $n = 22$.

7.6.3 Test de nullité de tous les coefficients sauf la constante; lien avec R^2.

Le vecteur des $k-1$ derniers coefficients de régression peut s'écrire:

$$\underline{\beta} = \begin{pmatrix} \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Nous voulons tester:

$$H_0 : \underline{\beta} = 0 \quad \text{contre} \quad \underline{\beta} \neq 0.$$

L'hypothèse nulle peut s'écrire sous la forme $R\beta = r$, avec:

$$R = \left(O_{(k-1)\times 1} \quad I_{k-1} \right),$$

$$r = 0.$$

La matrice R est donc de genre $k-1 \times k$ et le vecteur r est de taille $k-1$; nous avons un cas particulier du test F avec $J = k-1$.

Nous allons montrer que la statistique peut s'écrire:

$$F_{obs} = \left(\frac{R^2}{1-R^2} \right) \left(\frac{n-k}{k-1} \right)$$

et le R^2 sera donc significatif (de manière équivalente, on rejettera H_0) si:

$$R^2 > \frac{(k-1)F_{k-1,n-k,\alpha}}{n-k+(k-1)F_{k-1,n-k,\alpha}}.$$

En effet, le vecteur des résidus dans le modèle contraint est le suivant:

$$\hat{u}_c = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$$

et la somme des carrés des résidus contraints est donc:

$$\hat{u}_c' \hat{u}_c = \sum (y_t - \bar{y})^2.$$

Par conséquent:

$$Q = \hat{u}_c'\hat{u}_c - \hat{u}'\hat{u} = \sum (y_t - \bar{y})^2 - \hat{u}'\hat{u}$$

$$\frac{Q}{\sum (y_t - \bar{y})^2} = 1 - (1 - R^2) = R^2$$

$$\frac{\hat{u}'\hat{u}}{\sum (y_t - \bar{y})^2} = 1 - R^2$$

et donc:

$$F_{obs} = \left(\frac{\hat{u}_c'\hat{u}_c - \hat{u}'\hat{u}}{\hat{u}'\hat{u}} \right) \frac{n-k}{k-1} = \frac{R^2}{1-R^2} \frac{n-k}{k-1}$$

.

7.6.4 Test sur une combinaison linéaire des coefficients.

Nous avons ici le test:

$$\begin{aligned} & H_0 : c'\beta = r \\ \text{contre} \quad & H_1 : c'\beta \neq r \end{aligned}$$

où c est un vecteur $k \times 1$ de constantes et r est un scalaire. La statistique à employer prend alors la forme suivante:

$$F_{obs} \quad = \quad \frac{(c'\hat{\beta} - r)^2}{s^2 (c'(X'X)^{-1}c)} \sim F_{1;n-k} = t_{n-k}^2$$

et la statistique:

$$t_{obs} = \frac{c'\hat{\beta} - r}{s\sqrt{c'(X'X)^{-1}c}}$$

suit donc une loi de Student avec $n - k$ degrés de liberté sous H_0.

7.6.5 Test de stabilité structurelle (Chow).

Ce test, comme on va le voir, est un cas particulier du test F. On va diviser la période de l'échantillon en deux sous-périodes de nombres d'observations $n_1 > k$ et $n_2 > k$, et étudier la stabilité des coefficients de régression d'une sous-période à l'autre. Sous l'hypothèse nulle (stabilité structurelle), les coefficients sont les mêmes; sous l'hypothèse alternative, ils sont différents.

Si l'on n'a pas de stabilité structurelle (hypothèse alternative), le modèle s'écrit:

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & O \\ O & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

où y_1 et u_1 sont $n_1 \times 1$, y_2 et u_2 sont $n_2 \times 1$, X_1 est $n_1 \times k$, X_2 est $n_2 \times k$, et β_1 et β_2 sont $k \times 1$. Sous l'hypothèse alternative, $\beta_1 \neq \beta_2$. On a ici $2k$ régresseurs. On veut tester:

$$H_0 : \beta_1 = \beta_2 \quad \text{contre} \quad H_1 : \beta_1 \neq \beta_2.$$

Sous l'hypothèse nulle, le modèle précédent peut s'écrire:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

où $\beta = \beta_1 = \beta_2$. On a ici k régresseurs.

Le nombre de contraintes imposées sous H_0 est donc de $J = k$. Le nombre de degrés de liberté dans le modèle non contraint est de $n - 2k = n_1 + n_2 - 2k$.

La statistique est donc:

$$F_{obs} = \left(\frac{\hat{u}_c' \hat{u}_c - \hat{u}' \hat{u}}{\hat{u}' \hat{u}} \right) \frac{n - 2k}{k}.$$

Le modèle contraint correspond aux hypothèses classiques avec:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

Donc, en utilisant le lemme de la section 5.6:

$$\hat{u}_c' \hat{u}_c = y'y - \hat{\beta}' X' y = y'[I - X(X'X)^{-1}X']y = y'My.$$

Dans le modèle non contraint, on a comme matrice de régresseurs:

$$X_* = \begin{pmatrix} X_1 & O \\ O & X_2 \end{pmatrix}$$

et comme vecteur de coefficients:

$$\beta_* = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Par conséquent la somme des carrés des résidus dans le modèle non contraint est de:

$$\hat{u}' \hat{u} = y'y - \hat{\beta}_*' X_*' y = y'[I - X_*(X_*'X_*)^{-1}X_*']y = y'M_*y.$$

On peut facilement voir que:

$$y'M_*y = y_1'M_1y_1 + y_2'M_2y_2$$

avec:

$$M_1 = I_{n_1} - X_1(X_1'X_1)^{-1}X_1'$$
$$M_2 = I_{n_2} - X_2(X_2'X_2)^{-1}X_2'.$$

En remplaçant, dans l'expression de F_{obs}, $\hat{u}'\hat{u}$ et $\hat{u}_c'\hat{u}_c$ par les valeurs trouvées, on obtient:

$$F_{obs} = \left(\frac{y'My - y_1'M_1y_1 - y_2'M_2y_2}{y_1'M_1y_1 + y_2'M_2y_2} \right) \left(\frac{n - 2k}{k} \right)$$

et on rejette l'hypothèse de stabilité structurelle si:

$$F_{obs} > F_{k,n-2k,\alpha}.$$

7.7 Intervalles de prévision

Supposons que nous observions k valeurs futures des k régresseurs à une période θ suivant la dernière période de l'échantillon. Ces valeurs forment un vecteur de dimension $1 \times k$, soit x'_θ.

Nous désirons, comme précédemment (section 2.4), calculer un intervalle de prévision centré sur la prévision $\hat{y}_\theta$ de la variable dépendante.

Si le modèle reste inchangé à la période θ, on a:

$$y_\theta = x'_\theta \beta + u_\theta$$

avec:

$$E(u_\theta u_1) = \cdots = E(u_\theta u_n) = 0$$

et:

$$\hat{y}_\theta = x'_\theta \hat{\beta}.$$

Sous l'hypothèse $u_\theta \sim N(0, \sigma^2)$, trouvons la distribution de l'erreur de prévision:

$$y_\theta - \hat{y}_\theta = u_\theta - x'_\theta(\hat{\beta} - \beta) \quad .$$

C'est une variable normale de paramètres:

$$\begin{aligned} E(y_\theta - \hat{y}_\theta) &= 0 \\ V(y_\theta - \hat{y}_\theta) &= E(u_\theta^2) + E(x'_\theta(\hat{\beta} - \beta))^2 - 2\,\mathrm{Cov}(u_\theta, x'_\theta(\hat{\beta} - \beta)) \quad . \end{aligned}$$

Mais la covariance est nulle, puisque $\hat{\beta}$ ne dépend que des erreurs $u_1, u_2, \ldots, u_n$ de l'échantillon qui sont indépendantes de u_θ par hypothèse. On a alors:

$$\begin{aligned} V(y_\theta - \hat{y}_\theta) &= \sigma^2 + E\left[x'_\theta(\hat{\beta} - \beta)(\hat{\beta} - \beta)' x_\theta\right] \\ &= \sigma^2 + \sigma^2 x'_\theta (X'X)^{-1} x_\theta \quad . \end{aligned}$$

Considérons alors les variables

$$V = \frac{y_\theta - \hat{y}_\theta}{\sigma\sqrt{1 + x'_\theta (X'X)^{-1} x_\theta}}$$

$$\text{et} \quad W = \sqrt{\frac{\hat{u}'\hat{u}}{\sigma^2(n-k)}} \quad .$$

V est une variable $N(0,1)$. $\frac{\hat{u}'\hat{u}}{\sigma^2}$ est une variable χ^2 avec $n-k$ degrés de liberté, puisque $\frac{u}{\sigma} \sim N(0,1)$, $\hat{u}'\hat{u} = u'Mu$ et rang $M = n-k$ (section 4.2).

Les deux sont indépendantes puisque V ne dépend que de u_θ et de:

$$(\hat{\beta} - \beta) = (X'X)^{-1}X'u$$

et que:

$$(X'X)^{-1}X'\left[I - X(X'X)^{-1}X'\right] = O.$$

Nous pouvons en déduire que

$$t_{obs} \quad = \quad \frac{V}{W} \quad = \quad \frac{y_\theta - \hat{y}_\theta}{s\sqrt{1 + x'_\theta(X'X)^{-1}x_\theta}} \sim t_{n-k}$$

et l'intervalle de prévision cherché a pour bornes

$$\hat{y}_\theta \quad \pm \quad t_{n-k;\frac{\alpha}{2}} \quad s\sqrt{1 + x'_\theta(X'X)^{-1}x_\theta} \quad .$$

7.8 Exemple numérique

Reprenons le modèle et les données de la Section 5.9.

7.8.1 Testons l'hypothèse que la quantité d'engrais X_2 ne contribue pas à la production de vin.

Nous avons:

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0 \quad .$$

$$t_{obs} \quad = \quad \frac{0.2}{0.5\sqrt{66/432}} \quad = \quad 1.023 \quad .$$

Comme $t_{7;0.025} = 2.365 > 1.023$, nous ne rejetons pas H_0 au seuil de signification $\alpha = 0.05$.

7.8.2 Testons maintenant l'hypothèse

$$H_0 : \beta_1 = 1, \quad \beta_2 = 1, \quad \beta_3 = 0$$

$$\text{contre} \quad H_1 : \beta_1 \neq 1 \quad \text{ou} \quad \beta_2 \neq 1 \quad \text{ou} \quad \beta_3 \neq 0 \quad .$$

Ceci donne:

$$F_{obs} = \frac{1}{3(0.25)} \left(\begin{pmatrix} 0 & -0.3 & 0.2 \end{pmatrix} \begin{pmatrix} 10 & 2 & 2 \\ 2 & 7 & 1 \\ 2 & 1 & 7 \end{pmatrix} \begin{pmatrix} 0 \\ -0.3 \\ 0.2 \end{pmatrix} \right)$$

$$= 1.053 < 4.35 = F_{3;7;0.05} \quad .$$

On ne rejette donc pas l'hypothèse H_0.

7.8.3 Si nous voulons tester:

$$H_0 : \beta_1 = 0.5 \quad \text{et} \quad \beta_2 = 0.5$$

$$H_1 : \beta_1 \neq 0.5 \quad \text{ou} \quad \beta_2 \neq 0.5 \quad .$$

Nous construisons la statistique:

$$F_{obs} = \frac{432}{2(0.25)} \left(\begin{pmatrix} 0.5 & 0.2 \end{pmatrix} \begin{pmatrix} 48 & -12 \\ -12 & 66 \end{pmatrix}^{-1} \begin{pmatrix} 0.5 \\ 0.2 \end{pmatrix} \right)$$

$$= 5.949 > 4.74 = F_{2;7;0.05} \quad .$$

On rejette donc H_0.

7.8.4 Si nous voulons tester l'hypothèse que la production de vin ne dépend pas des facteurs X_1 et X_2, nous avons:

$$H_0 : \beta_2 = 0 \quad \text{et} \quad \beta_3 = 0$$

$$H_1 : \beta_2 \neq 0 \quad \text{ou} \quad \beta_3 \neq 0 \quad .$$

Ceci donne:

$$F_{obs} \quad = \quad \frac{R^2/2}{(1-R^2)/7} \quad = \quad 7.332 > 4.74 \quad = \quad F_{2;7;0.05} \quad .$$

On rejette donc l'hypothèse d'indépendance.

7.8.5 Enfin, si nous voulons tester l'hypothèse que les rendements d'échelle sont constants:

$$H_0 : \beta_2 + \beta_3 = 1$$

$$H_1 : \beta_2 + \beta_3 \neq 1 \quad .$$

Nous avons $c' = (\,0 \quad 1 \quad 1\,)$ et $r = 1$.

$$\text{On a} \quad c'(X'X)^{-1}c \quad = \quad \frac{1}{432}\left((\,0 \quad 1 \quad 1\,) \begin{pmatrix} 48 & -12 & -12 \\ -12 & 66 & -6 \\ -12 & -6 & 66 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right)$$

$$= \quad \frac{120}{432} \quad .$$

Ceci donne

$$t_{obs} \quad = \quad \frac{1-0.7-0.2}{(0.5)\sqrt{120/432}} \quad = \quad \frac{0.1}{(0.5)(0.527)}$$

$$= \quad 0.379 < t_{7;0.025} = 2.365 \quad .$$

Nous ne rejetons donc pas l'hypothèse de rendements constants.

7.8.6 Supposons qu'un onzième vigneron vaudois engage 2 unités de main-d'oeuvre (X_1) et emploie 3 unités d'engrais (X_2). Entre quelles bornes sa production de vin aura-t-elle 95 chances sur 100 de se situer? On a:

$$\begin{aligned} \log_e 2 &= 0.69315 \\ \log_e 3 &= 1.09861 \\ \log_e \hat{y}_{11} &= 1 + (0.7)(0.69315) + (0.2)(1.09861) \\ &= 1.70493 \end{aligned}$$

$$x'_{11}(X'X)^{-1}x_{11} = \frac{1}{432}\left((1 \quad 0.69315 \quad 1.09861) \begin{pmatrix} 48 & -12 & -12 \\ -12 & 66 & -6 \\ -12 & -6 & 66 \end{pmatrix} \begin{pmatrix} 1 \\ 0.69315 \\ 1.09861 \end{pmatrix} \right) = 0.2482.$$

Alors les bornes de l'intervalle sont

$$1.70493 \quad \pm \quad (2.365)(0.5) \quad \sqrt{1.2482} \quad \text{soit} \quad [0.384 \quad ; \quad 3.026]$$

et la production y_{11} a 95 chances sur 100 de se situer dans l'intervalle

$$[1.468 \quad ; \quad 20.616] \quad (\text{valeur médiane} = \exp\left(\frac{0.384 + 3.026}{2}\right) = 5.5) \quad .$$

CHAPITRE VIII

MOINDRES CARRÉS GÉNÉRALISÉS: LA MÉTHODE DE AITKEN

8.1 Introduction

Dans beaucoup de modèles économétriques, l'hypothèse que les erreurs sont de variance constante et ne sont pas corrélées entre elles ne peut pas être faite. C'est ainsi que dans notre exemple numérique précédent, la production de vin par hectare de deux agriculteurs voisins pourrait fort bien être influencée par des conditions exogènes (météorologiques ou autres) communes, ce qui se traduirait par une corrélation des erreurs.

Que se passerait-il si l'on appliquait la méthode des moindres carrés ordinaires à un tel modèle? Nous verrons plus loin que les estimateurs $\hat{\beta}_i$ obtenus seraient toujours sans biais, mais qu'ils seraient inefficaces; de plus, les estimateurs de leurs variances seraient biaisés.

La méthode de Aitken permet heureusement de remédier dans une large mesure à cet état de choses.

8.2 Exemples

8.2.1 Agrégation des données.

On veut estimer les paramètres du modèle $y = X\beta + u$ avec $E(u) = 0$ et $E(uu') = \sigma^2 I$, mais l'on ne dispose que de données agrégées $\bar{y}$ et $\bar{X}$ avec $\bar{y} = Gy,\ \bar{X} = GX$. Pour prendre un exemple, supposons que les données que l'on possède soient les moyennes des deux premières observations, des trois suivantes et des quatre dernières. La matrice G a alors la forme suivante:

$$G = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix} .$$

On estimerait le vecteur β sur la base du modèle:

$$Gy = GX\beta + Gu$$

soit aussi:

$$\bar{y} = \bar{X}\beta + \bar{u}.$$

La matrice de covariance de $\bar{u}$ est donc:

$$\begin{aligned} E(\bar{u}\bar{u}') &= E(Guu'G') = \sigma^2 GG' \\ &= \sigma^2 \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/4 \end{pmatrix} \end{aligned}$$

qui n'est pas une matrice scalaire.

Ceci est le problème d'hétéroscédasticité, qui sera étudié au chapitre IX.

8.2.2 Erreurs autorégressives.

Un autre exemple de modèle de régression où la matrice de covariance des erreurs n'est pas scalaire est le modèle à erreurs autorégressives, où $E(u_t u_{t-s}) = \rho^s \sigma^2$ avec $\mid \rho \mid < 1$. Ce modèle sera traité en détail au chapitre IX.

8.2.3 Equations simultanées.

Ce modèle très employé, est dû à A. Zellner ("Seemingly unrelated regressions and tests for aggregation bias", *Journal of the American Statistical Association* 57 (1962), pp. 348–368). Nous avons les N équations de régression suivantes:

$$y_i = X_i \beta_i + u_i \quad \text{pour } i = 1, \ldots, N$$

ou, sous forme matricielle:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} X_1 & O & \ldots & O \\ O & X_2 & \ldots & O \\ \vdots & & & \vdots \\ O & & \ldots & X_N \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_N \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix} .$$

où les y_i sont des vecteurs $T \times 1$, les X_i sont des matrices $T \times k_i$, les β_i sont des vecteurs $k_i \times 1$, et les u_i sont des vecteurs $T \times 1$. On fait l'hypothèse $E(u_i u_j') = \sigma_{ij} I$. Nous avons donc l'absence de corrélation dans le temps, mais pas entre les équations (les erreurs de

deux équations différentes sont corrélées à la même période). Si l'on écrit l'équation de régression précédente comme $y = X\beta + u$, la matrice de covariance du vecteur u s'écrit:

$$E(uu') = E\begin{pmatrix} u_1u_1' & \dots & u_1u_N' \\ \vdots & \vdots & \vdots \\ u_Nu_1' & \dots & u_Nu_N' \end{pmatrix} = \begin{pmatrix} \sigma_{11}I_T & \dots & \sigma_{1N}I_T \\ \vdots & \vdots & \vdots \\ \sigma_{1N}I_T & \dots & \sigma_{NN}I_T \end{pmatrix}$$

et n'est donc ni diagonale, ni scalaire.

8.3 L'estimateur de Aitken et ses propriétés

Nous avons donc le modèle général:

$$y = X\beta + u$$

avec $E(u) = 0$ et $E(uu') = \sigma^2\Omega$, où Ω est une matrice définie positive, supposée (temporairement) connue. Pour des raisons de commodité, nous utiliserons parfois la notation $V = \sigma^2\Omega$.

Nous allons voir qu'il existe une transformation linéaire du modèle, soit une application $(y, X, u) \longrightarrow (y^*, X^*, u^*)$ telle que u^* vérifie les hypothèses du modèle de régression classique. On peut alors appliquer la méthode des moindres carrés ordinaires au modèle transformé.

Comme la matrice Ω est symétrique, il existe une matrice orthogonale C telle que $C'\Omega C = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \underset{\text{def}}{=} \Delta$, où les λ_i sont les valeurs propres de Ω. Comme Ω est définie positive, $\lambda_i > 0$ pour tout i. Définissons alors

$$\Delta^{-1/2} = \text{diag}\left(\frac{1}{\sqrt{\lambda}_1}, \dots, \frac{1}{\sqrt{\lambda}_n}\right).$$

Nous avons $\Delta^{-1/2}\Delta\Delta^{-1/2} = I$, soit aussi $\Delta^{-1/2}C'\Omega C\Delta^{-1/2} = I$, ou $T\Omega T' = I$ avec $T = \Delta^{-1/2}C'$.

Il est facile alors de montrer que T définit une transformation linéaire du modèle (et donc en particulier des erreurs) qui permet de retrouver les hypothèses faites en régression classique.

En prémultipliant $y = X\beta + u$ par T, on obtient en effet $y^* = X^*\beta + u^*$ avec $u^* = Tu$. Calculons la matrice de covariance de u^*. On a

$$E(u^*u^{*'}) = E(Tuu'T') = TE(uu')T' = \sigma^2(T\Omega T') = \sigma^2 I \quad .$$

Notons enfin que $\Omega^{-1} = T'T$. On obtient, en effet, en prémultipliant l'égalité $T\Omega T' = I$ par T^{-1} et en la postmultipliant par $(T')^{-1}$:

$$\Omega = T^{-1}(T')^{-1} \quad , \quad \text{soit} \quad \Omega^{-1} = \left[T^{-1}(T')^{-1}\right]^{-1} = T'T \quad .$$

Si l'on applique la méthode des moindres carrés ordinaires au modèle transformé $Ty = TX\beta + Tu$, on obtient:

$$\hat{\beta}_{mcg} = (X'T'TX)^{-1}X'T'Ty$$

soit aussi:

$$\begin{aligned}\hat{\beta}_{mcg} &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \\ &= (X'V^{-1}X)^{-1}X'V^{-1}y\end{aligned}$$

et l'on a:

$$V(\hat{\beta}) = E(\hat{\beta}-\beta)(\hat{\beta}-\beta)' = \sigma^2(X'T'TX)^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1} \quad .$$

Un estimateur sans biais de σ^2 est obtenu comme auparavant par:

$$\begin{aligned}s^2 &= \frac{1}{n-k}\ \hat{u}^{*'}\hat{u}^* \\ &= \frac{1}{n-k}\ (y^* - X^*\hat{\beta}_{mcg})'(y^* - X^*\hat{\beta}_{mcg}) \\ &= \frac{1}{n-k}\ (y - X\hat{\beta}_{mcg})'T'T(y - X\hat{\beta}_{mcg}) \\ &= \frac{1}{n-k}\ (y - X\hat{\beta}_{mcg})'\Omega^{-1}(y - X\hat{\beta}_{mcg}) \quad .\end{aligned}$$

Passons maintenant au problème de l'étude des propriétés de $\hat{\beta}_{mco} = (X'X)^{-1}X'y$ lorsque $E(u) = 0$ et $E(uu') = \sigma^2\Omega$. Cet estimateur sera toujours sans biais (la démonstration est exactement la même que précédemment). Mais il ne sera pas efficace. En effet, puisque le modèle $y^* = X^*\beta + u^*$ satisfait les hypothèses du modèle de régression classique, le théorème de Gauss-Markov lui est applicable; l'estimateur $\hat{\beta}_{mcg}$ est donc, pour ce modèle, le plus efficace des estimateurs linéaires sans biais. Or, $\hat{\beta}_{mcg} \neq \hat{\beta}_{mco}$ si $\Omega \neq I$.

Il y a plus grave. Lorsque $\Omega \neq I$, nous allons montrer que $V(\hat{\beta}_{mco}) \neq \sigma^2(X'X)^{-1}$. La formule classique n'est donc plus applicable. En effet, nous avons

$$\begin{aligned}V(\hat{\beta}_{mco}) &= E(\hat{\beta}_{mco}-\beta)(\hat{\beta}_{mco}-\beta)' \\ &= E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}\right] \\ &= \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1} \quad \neq \quad \sigma^2(X'X)^{-1} \quad .\end{aligned}$$

8.4 La prévision dans le modèle de Aitken

Nous avons donc le modèle $y = X\beta + u$, avec $E(u) = 0$ et $E(uu') = \sigma^2\Omega = V$. Nous voulons prédire une valeur future y_θ de la variable dépendante, conditionnellement à un vecteur futur d'observations sur les k variables explicatives. Si le modèle reste inchangé et si u_θ est l'erreur future, nous pouvons écrire:

$$y_\theta = x'_\theta\beta + u_\theta$$

$$\text{avec} \quad E(u_\theta) = 0, \quad E(u_\theta^2) = \sigma_\theta^2 \quad \text{et} \quad E(u_\theta u) = w$$

(w est un vecteur colonne de taille n).

La connaissance du vecteur w des covariances entre l'erreur future et les erreurs de l'échantillon va nous permettre de définir un préviseur de y_θ plus efficace que la valeur calculée $x'_\theta\hat{\beta}_{mcg}$. En effet, la connaissance de ces covariances et l'estimation des erreurs de l'échantillon à l'aide des résidus permet souvent de faire une inférence statistique portant sur l'erreur future u_θ. Les résultats de cette section sont dus à A. Goldberger, "Best linear unbiased prediction in the generalized linear regression model", *Journal of the American Statistical Association* 57 (1962), pp. 369–375.

Nous voulons trouver un préviseur linéaire de la forme $p = c'y$, où le vecteur c doit être choisi de façon à minimiser la variance $\sigma_p^2 = E(y_\theta - p)^2$, sous la contrainte que $E(y_\theta - p) = 0$. Comme $y_\theta - p = (x'_\theta - c'X)\beta - (c'u - u_\theta)$, cette contrainte s'écrit sous forme vectorielle comme $x'_\theta = c'X$. Nous avons donc un système de k contraintes. Quant à la variance à minimiser, elle peut s'écrire:

$$\begin{aligned}
\sigma_p^2 &= E(y_\theta - p)^2 \\
&= E(y_\theta - p)(y_\theta - p)' \quad \text{puisque} \quad p \quad \text{est un scalaire} \\
&= E(c'u - u_\theta)(c'u - u_\theta)' \quad \text{puisque} \quad x'_\theta - c'X = 0 \\
&= E(c'uu'c + u_\theta^2 - 2c'uu_\theta) \\
&= c'Vc + \sigma_\theta^2 - 2c'w.
\end{aligned}$$

Le Lagrangien peut s'écrire:

$$\mathcal{L}(c, \lambda) = c'Vc - 2c'w - 2(c'X - x'_\theta)\lambda$$

et le système de conditions de premier ordre:

$$\frac{\partial \mathcal{L}}{\partial c} = 2Vc - 2X\lambda - 2w = 0 \quad .$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -2X'c + 2x_\theta = 0$$

s'écrit sous forme matricielle comme:

$$\begin{pmatrix} V & X \\ X' & O \end{pmatrix} \begin{pmatrix} c \\ -\lambda \end{pmatrix} = \begin{pmatrix} w \\ x_\theta \end{pmatrix}$$

En utilisant la formule d'inversion en forme partagée, la solution de ce système peut s'écrire:

$$\begin{pmatrix} \hat{c} \\ -\hat{\lambda} \end{pmatrix} = \begin{pmatrix} V^{-1}\left[I - X(X'V^{-1}X)^{-1}X'V^{-1}\right] & V^{-1}X(X'V^{-1}X)^{-1} \\ (X'V^{-1}X)^{-1}X'V^{-1} & -(X'V^{-1}X)^{-1} \end{pmatrix} \begin{pmatrix} w \\ x_\theta \end{pmatrix}$$

ou, en effectuant le produit:

$$\hat{c} \quad = \quad V^{-1}\left[I - X(X'V^{-1}X)^{-1}X'V^{-1}\right] w + V^{-1}X(X'V^{-1}X)^{-1}x_\theta$$

et

$$\hat{p} \quad = \quad \hat{c}'y \quad = \quad w'V^{-1}y + x'_\theta(X'V^{-1}X)^{-1}X'V^{-1}y - w'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y$$

$$= \quad x'_\theta\hat{\beta}_{mcg} + w'V^{-1}\hat{u}_{mcg} \quad \text{avec} \quad \hat{u}_{mcg} = y - X\hat{\beta}_{mcg} \quad .$$

On s'aperçoit donc que le meilleur préviseur linéaire sans biais s'obtient en ajoutant à la valeur calculée $x'_\theta\hat{\beta}_{mcg}$ un terme correcteur $w'V^{-1}\hat{u}_{mcg}$, qui dépend notamment du vecteur w des covariances entre les erreurs passées et l'erreur future, et du vecteur de résidus $\hat{u}_{mcg}$.

Afin de trouver le gain d'efficacité entraîné par l'adjonction de ce terme correcteur, nous substituons l'expression précédemment obtenue pour $\hat{c}$ dans la formule $\sigma_p^2 = \sigma_\theta^2 - 2c'w + c'Vc$. On a:

$$\hat{c} = Mw + P'Q^{-1}x_\theta$$

avec:

$$P = X'V^{-1}$$
$$Q = X'V^{-1}X$$
$$M = (V^{-1} - P'Q^{-1}P).$$

On vérifie par ailleurs par simple multiplication que:

$$Q^{-1}PVP' = I$$

$$Q^{-1}PVM = O$$

$$M'VM = M \quad .$$

Alors:

$$\hat{c}'V\hat{c} = w'M'VMw + w'M'VP'Q^{-1}x_\theta + x_\theta'Q^{-1}PVMw + x_\theta'Q^{-1}PVP'Q^{-1}x_\theta$$

$$= w'Mw + x_\theta'Q^{-1}x_\theta \quad .$$

De même:

$$\hat{c}'w = w'Mw + x_\theta'Q^{-1}Pw$$

et donc, en substituant plus haut:

$$\sigma_{\hat{p}}^2 = \sigma_\theta^2 - w'Mw + x_\theta'Q^{-1}x_\theta - 2x_\theta'Q^{-1}Pw \quad .$$

Soit maintenant $\bar{p} = x_\theta'\hat{\beta}_{mcg}$. On vérifie aisément que $\bar{p} = \bar{c}'y$ avec $\bar{c} = P'Q^{-1}x_\theta$. En remplaçant c par $\bar{c}$ dans la formule de σ_p^2, il vient:

$$\sigma_{\bar{p}}^2 = \sigma_\theta^2 - 2\bar{c}'w + \bar{c}'V\bar{c}$$

$$= \sigma_\theta^2 - 2x_\theta'Q^{-1}Pw + x_\theta'Q^{-1}PVP'Q^{-1}x_\theta$$

$$= \sigma_\theta^2 - 2x_\theta'Q^{-1}Pw + x_\theta'Q^{-1}x_\theta$$

$$= \sigma_{\hat{p}}^2 + w'Mw \quad .$$

Nous allons montrer que la matrice M est définie non négative . Comme V^{-1} est définie positive, il existe une matrice B régulière telle que $V^{-1} = B'B$ (voir 3.1.3). Nous pouvons alors écrire:

$$\begin{aligned} M &= V^{-1} - P'Q^{-1}P \\ &= V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} \\ &= B'\left[I - BX(X'B'BX)^{-1}X'B'\right]B \\ &\underset{\text{def}}{=} B'NB \quad . \end{aligned}$$

On vérifie par simple multiplication que N est symétrique et idempotente. Elle est alors définie non négative, puisque ses valeurs propres sont 0 ou 1. Alors $M = B'NB$ est définie non négative . Par conséquent, $w'Mw \geq 0$, et $\sigma^2_{\hat{p}} \leq \sigma^2_{\bar{p}}$.

CHAPITRE IX

L'AUTOCORRÉLATION ET L'HÉTÉROSCÉDASTICITÉ

9.1 Erreurs autorégressives d'ordre un

Cette hypothèse a été introduite pour remédier au problème suivant. Il arrive fréquemment, dans les séries chronologiques, que les résidus présentent une allure cyclique: soit un résidu positif tend à être suivi par un résidu positif, et un résidu négatif par un résidu négatif; soit les signes des résidus successifs alternent. Le premier cas correspond à une autocorrélation positive des erreurs; le second cas, à une autocorrélation négative.

Dans un modèle de consommation par exemple, la présence d'une autocorrélation positive des erreurs pourrait traduire une certaine inertie du comportement des agents: une consommation supérieure à la normale aurait tendance à se poursuivre durant plusieurs périodes successives. La présence d'une autocorrélation négative pourrait traduire un phénomène oscillatoire, l'individu compensant par une consommation moindre à la période t un excès de consommation à la période $t-1$.

Dans un cas comme dans l'autre, l'hypothèse de non corrélation des erreurs est violée. Il faut alors appliquer la méthode de Aitken. Mais il est nécessaire pour cela de décrire formellement cette dépendance des erreurs, c'est-à-dire de postuler une forme explicite de la matrice de covariance des erreurs. On fait donc les hypothèses suivantes:

$$u_t = \rho u_{t-1} + \epsilon_t, \quad \text{avec:}$$

$$\begin{aligned} |\,\rho\,| &< 1 \\ E(\epsilon_t) &= 0 \quad \text{pour tout} \quad t, \\ E(\epsilon_t \epsilon_s) &= \sigma_\epsilon^2 \quad (t = s) \\ &= 0 \quad (t \neq s) \quad . \end{aligned}$$

L'erreur u_t possède donc une composante systématique ρu_{t-1} et une composante purement aléatoire ϵ_t.

9.2 La matrice de covariance des erreurs

On la calcule facilement en résolvant l'équation de récurrence $u_t = \rho u_{t-1} + \epsilon_t$. Comme $u_{t-1} = \rho u_{t-2} + \epsilon_{t-1}$, on obtient:

$$\begin{aligned} u_t &= \rho(\rho u_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= \rho^2 u_{t-2} + \rho\epsilon_{t-1} + \epsilon_t \\ &= \rho^2(\rho u_{t-3} + \epsilon_{t-2}) + \rho\epsilon_{t-1} + \epsilon_t \\ &= \rho^3 u_{t-3} + \rho^2\epsilon_{t-2} + \rho\epsilon_{t-1} + \epsilon_t \end{aligned}$$

soit, en remontant indéfiniment dans le temps:

$$u_t = \sum_{i=0}^{\infty} \rho^i \epsilon_{t-i}$$

ce qui implique:

$$\begin{aligned} E(u_t) &= \sum_{i=0}^{\infty} \rho^i E(\epsilon_{t-i}) = 0 \\ E(u_t^2) &= E(\epsilon_t^2) + \rho^2 E(\epsilon_{t-1}^2) + \rho^4 E(\epsilon_{t-2}^2) + \dots \\ &= \sigma_\epsilon^2(1 + \rho^2 + \rho^4 + \dots) \\ &= \frac{\sigma_\epsilon^2}{1-\rho^2} \quad . \end{aligned}$$

De même:

$$\begin{aligned} E(u_t u_{t-1}) &= E(u_{t-1}(\rho u_{t-1} + \epsilon_t)) \\ &= \rho E(u_{t-1}^2) = \frac{\rho\sigma_\epsilon^2}{1-\rho^2} = \rho\sigma_u^2 \quad . \end{aligned}$$

$$\begin{aligned} E(u_t u_{t-2}) &= E(u_{t-2}(\rho^2 u_{t-2} + \rho\epsilon_{t-1} + \epsilon_t)) \\ &= \rho^2 E(u_{t-2}^2) = \rho^2\sigma_u^2 \\ E(u_t u_{t-s}) &= \rho^s\sigma_u^2 \quad . \end{aligned}$$

Nous avons donc établi que

$$E(uu') = \sigma_u^2\,\Omega = \sigma_u^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & \rho & 1 \end{pmatrix}$$

9.3 Transformation des données (ρ connu)

Si le coefficient d'autorégression ρ est connu, la méthode de Aitken appliquée au modèle $y = X\beta + u$ fournit le meilleur estimateur linéaire sans biais de β, qui est $\hat{\beta}_{mcg} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$. Calculons l'inverse de la matrice Ω.

On vérifie par simple multiplication que:

$$\Omega^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{pmatrix} \quad .$$

Comme nous l'avons vu plus haut, il est avantageux de calculer $\hat{\beta}_{mcg}$ de la façon suivante: On trouve d'abord une matrice T telle que $\Omega^{-1} = T'T$; on applique ensuite les moindres carrés ordinaires à l'équation $Ty = TX\beta + Tu$. On vérifie également par multiplication que T est donnée par:

$$T = \frac{1}{\sqrt{1-\rho^2}} \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1 & 0 & \dots & 0 & 0 \\ 0 & -\rho & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{pmatrix} .$$

Nous pouvons laisser tomber le facteur multiplicatif qui se simplifie, apparaissant à gauche et à droite dans l'équation transformée. Nous pouvons donc retenir comme formule de transformation d'une colonne z de la matrice des données $[y \quad X]$ la règle suivante:

$$z^* = \begin{pmatrix} (\sqrt{1-\rho^2})z_1 \\ z_2 - \rho z_1 \\ z_3 - \rho z_2 \\ \vdots \\ z_n - \rho z_{n-1} \end{pmatrix}$$

et appliquer les moindres carrés ordinaires aux données transformées.

9.4 Estimation du coefficient d'autorégression

9.4.1 Méthode de Cochrane-Orcutt.

Cette méthode est la plus employée. On commence par appliquer les moindres carrés ordinaires pour obtenir un vecteur $\hat{u}$ de résidus, soit $\hat{u} = [I - X(X'X)^{-1}X']y$. On obtient ensuite $\hat{\rho}$ en regressant $\hat{u}_t$ sur $\hat{u}_{t-1}$. Ceci donne:

$$\hat{\rho} = \frac{\sum_{t=2}^{n} \hat{u}_t \hat{u}_{t-1}}{\sum_{t=2}^{n} \hat{u}_{t-1}^2} \quad .$$

On applique alors la formule des moindres carrés généralisés en remplaçant ρ par $\hat{\rho}$ dans l'expression de la matrice Ω. Soit donc:

$$\hat{\Omega} = \begin{pmatrix} 1 & \hat{\rho} & \dots & \hat{\rho}^{n-2} & \hat{\rho}^{n-1} \\ \hat{\rho} & 1 & \dots & \hat{\rho}^{n-3} & \hat{\rho}^{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{\rho}^{n-2} & \hat{\rho}^{n-3} & \dots & 1 & \hat{\rho} \\ \hat{\rho}^{n-1} & \hat{\rho}^{n-2} & \dots & \hat{\rho} & 1 \end{pmatrix} \quad .$$

On calcule $\hat{\hat{\beta}} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y$. Ceci fournit un nouveau vecteur de résidus $\hat{\hat{u}} = y - X\hat{\hat{\beta}}$. Ce nouveau vecteur peut servir à calculer une nouvelle estimation de ρ, soit $\hat{\hat{\rho}}$. Cette dernière peut servir à calculer une troisième estimation de β, et ainsi de suite. On peut poursuivre cette procédure jusqu'à la convergence des estimations de ρ.

9.4.2 Méthode de Durbin.

Réécrivons l'équation de régression sous la forme suivante:

$$y_t = \sum_{j=1}^{k} \beta_j X_{jt} + u_t \quad .$$

En retardant d'une période et en multipliant par ρ:

$$\rho y_{t-1} = \sum_{j=1}^{k} (\rho\beta_j) X_{jt-1} + \rho u_{t-1} \quad .$$

En soustrayant cette équation de la première, on obtient, puisque $u_t - \rho u_{t-1} = \epsilon_t$:

$$y_t = \rho y_{t-1} + \sum_{j=1}^{k} \beta_j X_{jt} - \sum_{j=1}^{k} (\rho\beta_j) X_{jt-1} + \epsilon_t$$

qui est une équation de régression comportant $2k+1$ régresseurs. Comme les ϵ_t vérifient les hypothèses du modèle de régression classique, on applique la méthode des moindres carrés ordinaires pour estimer ρ. (Son estimateur est celui du coefficient de y_{t-1}). Comme y_{t-1} est un régresseur stochastique (il dépend de ϵ_{t-1}), nous verrons plus loin que l'estimateur ainsi obtenu n'est pas sans biais.

On remplace alors, comme précédemment, ρ par $\hat{\rho}$ dans l'expression de Ω, et applique la formule des moindres carrés généralisés.

Notons que l'estimateur $\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ s'appelle parfois l'estimateur *Aitken-pur* ; $\hat{\hat{\beta}} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y$ s'appelle alors l'estimateur *Aitken-réalisable.*

9.5 La statistique de Durbin-Watson

Elle permet de tester l'hypothèse nulle que $\rho = 0$, contre les hypothèses alternatives $\rho \neq 0$, ou $\rho > 0$, ou $\rho < 0$. Sa distribution n'a pas pu être déterminée indépendamment de la forme de la matrice X. Il existe donc une zone de valeurs de cette statistique pour lesquelles on ne pourra rejeter ni l'hypothèse nulle, ni l'hypothèse alternative.

La statistique de Durbin-Watson est définie comme:

$$d_{obs} = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2}$$

où les $\hat{u}_t$ sont les résidus des moindres carrés ordinaires.

Nous allons étudier ses propriétés lorsque n tend vers l'infini.

Plus précisément, nous montrerons que si n est suffisamment grand d_{obs} est approximativement égale à 2 lorsque $\rho = 0$; à 0 lorsque $\rho = 1$; et à 4 lorsque $\rho = -1$. En effet,

$$\begin{aligned} d_{obs} &= \frac{\sum_{t=2}^{n}\hat{u}_t^2 + \sum_{t=2}^{n}\hat{u}_{t-1}^2 - 2\sum_{t=2}^{n}\hat{u}_t\hat{u}_{t-1}}{\sum_{t=1}^{n}\hat{u}_t^2} \\ &\approx \frac{2\sum_{t=2}^{n}\hat{u}_t^2 - 2\sum_{t=2}^{n}\hat{u}_t\hat{u}_{t-1}}{\sum_{t=2}^{n}\hat{u}_t^2}, \end{aligned}$$

puisque:

$$\begin{aligned} \sum_{t=2}^{n}\hat{u}_t^2 &\approx \sum_{t=2}^{n}\hat{u}_{t-1}^2 \\ \sum_{t=1}^{n}\hat{u}_t^2 &\approx \sum_{t=2}^{n}\hat{u}_t^2 \quad . \end{aligned}$$

Il est raisonnable de supposer que lorsque n tend vers l'infini, $\frac{1}{n-1}\sum_{t=2}^{n}\hat{u}_t^2$ tend vers σ_u^2 et $\frac{1}{n-1}\sum_{t=2}^{n}\hat{u}_t\hat{u}_{t-1}$ tend vers $\mathrm{Cov}(u_t, u_{t-1}) = \rho\sigma_u^2$. On a alors, en divisant numérateur et dénominateur par $n-1$:

$$d_{obs} \approx \frac{2\sigma_u^2 - 2\rho\sigma_u^2}{\sigma_u^2} = 2(1-\rho)$$

ce qu'il fallait montrer.

Les valeurs de d_{obs} qui sont proches de 2, nous conduisent donc à ne pas rejeter $\rho = 0$; celles qui sont proches de 0, à rejeter $\rho = 0$ en faveur de $\rho > 0$; celles qui sont proches de 4, à rejeter $\rho = 0$ en faveur de $\rho < 0$. La table des valeurs critiques fournit deux valeurs, d_U et d_L, pour chaque combinaison de nombres d'observations (n) et de nombres de variables explicatives ($k' = k - 1$). La zone $d_L < d_{obs} < d_U$ est une zone d'incertitude, de même que la zone $4 - d_U < d_{obs} < 4 - d_L$. Pour ces valeurs de d_{obs}, on ne pourra rejeter ni $\rho = 0$, ni $\rho \neq 0$.

Les règles de décision sont résumées dans le tableau suivant (l'hypothèse nulle est toujours $H_0 : \rho = 0$):

H_1	$d < d_L$	$d_L \leq d < d_U$	$d_U \leq d < 4 - d_U$	$4 - d_U \leq d < 4 - d_L$	$4 - d_L \leq d$
$\rho > 0$	Rejeter H_0	Incertain	Ne pas rejeter H_0		
$\rho < 0$	Ne pas rejeter H_0			Incertain	Rejeter H_0
$\rho \neq 0$	Rejeter H_0	Incertain	Ne pas rejeter H_0	Incertain	Rejeter H_0

Note importante: Le test de Durbin-Watson ne peut pas être employé lorsque les régresseurs incluent des variables endogènes retardées.

9.6 La prévision dans le modèle à erreurs autorégressives

Nous avons vu à la Section 8.4 que le meilleur préviseur linéaire sans biais d'une valeur future y_θ de la variable dépendante était $\hat{p} = x'_\theta \hat{\beta}_{mcg} + w' V^{-1} \hat{u}$, avec $w = E(u_\theta u)$, $V = E(uu')$ et $\hat{u} = y - X\hat{\beta}_{mcg}$. Nous allons illustrer cette règle de prévision dans le modèle à erreurs autorégressives d'ordre un, en supposant $\theta = n + 1$. Le vecteur w prend la forme:

$$w = \begin{pmatrix} E(u_1 u_{n+1}) \\ E(u_2 u_{n+1}) \\ \vdots \\ E(u_n u_{n+1}) \end{pmatrix} = \sigma_u^2 \begin{pmatrix} \rho^n \\ \vdots \\ \rho^2 \\ \rho \end{pmatrix} = \rho \sigma_u^2 \begin{pmatrix} \rho^{n-1} \\ \vdots \\ \rho \\ 1 \end{pmatrix} \quad .$$

Mais $\sigma_u^2 [\rho^{n-1} \ldots \rho \, 1]$ est la dernière ligne de V. Comme $VV^{-1} = I$, nous avons: $\sigma_u^2 [\rho^{n-1} \ldots \rho \, 1] \, V^{-1} = [0 \ldots 0 \, 1]$ et donc: $w' V^{-1} = \rho \sigma_u^2 [\rho^{n-1} \ldots \rho \, 1] V^{-1} = \rho [0 \ldots 0 \, 1]$. Par conséquent, $w' V^{-1} \hat{u} = \rho \hat{u}_n$. La formule précédente s'écrit alors:

$$\hat{p} = x'_{n+1} \hat{\beta}_{mcg} + \rho \hat{u}_n \quad .$$

L'interprétation de cette formule est immédiate. On ajoute à la valeur calculée $x'_{n+1} \hat{\beta}_{mcg}$ un terme correcteur qui aura le signe du dernier résidu de l'échantillon si le coefficient de corrélation entre deux erreurs successives est positif, le signe contraire sinon.

9.7 Le problème de l'hétéroscédasticité

Nous avons déjà rencontré ce problème à la section 8.2.1. Lorsqu'il se rencontre sous cette forme, il est très facile à traiter: la matrice $E(uu')$ est en effet connue, égale à $\sigma^2 \text{diag}(k_1, \ldots, k_n)$ où les k_i sont des constantes positives connues.

La matrice de transformation à utiliser est alors bien entendu diag $(\frac{1}{\sqrt{k_1}}, \ldots, \frac{1}{\sqrt{k_n}})$: Il suffit de multiplier les $k+1$ données correspondant à la t-ième observation par $\frac{1}{\sqrt{k_t}}$ pour retrouver une matrice de covariance scalaire.

Il existe bien sûr d'autres formes d'hétéroscédasticité. Il peut être raisonnable de supposer que la variance des erreurs augmente avec la valeur absolue de l'un des régresseurs, soit, par exemple, que $E(u_t^2) = \sigma^2 X_t^2$. Il suffit alors de multiplier les données correspondant à la t-ième observation par $\frac{1}{\sqrt{X_t^2}}$.

Plus généralement, nous allons voir qu'une hétéroscédasticité des erreurs peut être induite par des variations aléatoires des coefficients de régression, en illustrant cette situation à l'aide d'un exemple simple. Soit donc le modèle:

$$y_t = a + bx_t + u_t$$

et supposons que $b = b^* + \epsilon_t$, où b^* est constant en probabilité et où ϵ_t est une erreur aléatoire avec $E(\epsilon_t) = 0$, $V(\epsilon_t) = \sigma_\epsilon^2$, $E(\epsilon_t \epsilon_s) = 0$ pour $t \neq s$, et $E(u_t \epsilon_t) = 0$. On peut alors écrire:

$$\begin{aligned} y_t &= a + (b^* + \epsilon_t)x_t + u_t \\ &= a + b^* x_t + (u_t + \epsilon_t x_t) \\ &= a + b^* x_t + v_t \end{aligned}$$

avec $v_t = u_t + \epsilon_t x_t$. On a $E(v_t) = 0$, $E(v_t v_s) = 0$ pour $t \neq s$, mais:

$$\begin{aligned} E(v_t^2) &= E(u_t^2) + x_t^2 E(\epsilon_t^2) \\ &= \sigma_u^2 + x_t^2 \sigma_\epsilon^2 \end{aligned}$$

dépend de l'indice t.

Une solution possible, en grand échantillon, est de poser:

$$\hat{v}_t^2 = \alpha + \beta x_t^2 + \eta_t$$

où $\hat{v}_t$ est un résidu de la régression de y_t sur x_t par moindres carrés ordinaires. On estime α et β par MCO et on estime $\sigma_t^2 = E(v_t^2)$ par $\hat{\alpha} + \hat{\beta} x_t^2$. On utilise ensuite les moindres carrés pondérés pour estimer a et b^*.

9.8 Les tests de diagnostic

9.8.1 Analyse des autocorrélations.

On définit les coefficients d'autocorrélation empiriques des résidus $\hat{u}_t$ des moindres carrés comme:

$$R_s = \frac{\sum_{t=s+1}^{n} \hat{u}_t \hat{u}_{t-s}}{\sum_{t=1}^{n} \hat{u}_t^2}.$$

L'interprétation de R_s est la suivante:

- $\frac{1}{n}\sum_{t=s+1}^{n} \hat{u}_t \hat{u}_{t-s}$ est une estimation de $\mathrm{Cov}(u_t, u_{t-s})$;
- $\frac{1}{n}\sum_{t=1}^{n} \hat{u}_t^2$ est une estimation de $V(u_t)$, supposée égale à $V(u_{t-s})$;
- R_s est donc une estimation du coefficient de corrélation entre u_t et u_{t-s}, à savoir:

$$r_s = \frac{\mathrm{Cov}(u_t, u_{t-s})}{\sqrt{V(u_t)V(u_{t-s})}}.$$

L'étude du comportement des coefficients d'autocorrélation permet par exemple de distinguer un processus autorégressif (AR) d'un processus dit "à moyenne mobile" (MA).

Pour le processus autorégressif d'ordre un:

$$u_t = \rho u_{t-1} + \epsilon_t,$$

on a vu à la section 9.2 que:

$$\begin{aligned} V(u_t) = V(u_{t-s}) &= \sigma_u^2 \\ \mathrm{Cov}(u_t, u_{t-s}) &= \rho^s \sigma_u^2, \quad \text{et donc:} \\ r_s = \frac{\mathrm{Cov}(u_t, u_{t-s})}{\sqrt{V(u_t)V(u_{t-s})}} &= \rho^s. \end{aligned}$$

Le coefficient d'autocorrélation *théorique* décroît donc géométriquement avec s. Un tel comportement de la fonction d'autocorrélation empirique R_s est donc indicatif d'erreurs autorégressives.

Pour un processus à moyenne mobile d'ordre un:

$$u_t = \epsilon_t + \rho \epsilon_{t-1}$$

où les ϵ_t sont des erreurs fondamentales avec $E(\epsilon_t) = 0$ pour tout t, $E(\epsilon_t^2) = \sigma_\epsilon^2$ pour tout t, et $E(\epsilon_t \epsilon_{t-s}) = 0$ pour $s > 0$, on a:

$$\begin{aligned} E(u_t u_{t-1}) &= E(\epsilon_t + \rho\epsilon_{t-1})(\epsilon_{t-1} + \rho\epsilon_{t-2}) \\ &= E(\epsilon_t\epsilon_{t-1}) + \rho E(\epsilon_t\epsilon_{t-2}) + \rho E(\epsilon_{t-1}^2) + \rho^2 E(\epsilon_{t-1}\epsilon_{t-2}) \\ &= \rho\sigma_\epsilon^2 \end{aligned}$$

et, comme on le vérifie aisément:

$$E(u_t u_{t-s}) = 0 \quad \text{pour } s > 1.$$

Par conséquent:

$$r_s = \frac{\text{Cov}(u_t, u_{t-s})}{\sqrt{V(u_t)V(u_{t-s})}} = \frac{\rho}{1+\rho^2} \quad \text{si } s = 1;$$
$$= 0 \quad \text{si } s > 1.$$

Ces observations peuvent être généralisées à des processus d'ordre supérieur au premier. Plus généralement, un comportement du type:

$$R_s \neq 0 \quad \text{pour } 1 \leq s \leq \tau$$
$$R_s \approx 0 \quad \text{pour } s > \tau$$

sera indicatif d'erreurs à moyenne mobile; tandis que la convergence vers zéro sera graduelle pour un processus autorégressif.

9.8.2 Le test de Breusch-Godfrey (autocorrélation).

Ce test permet, lorsque les erreurs sont autorégressives d'ordre p:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \epsilon_t$$

de tester l'hypothèse:

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0$$

contre:

$$H_1 : (\rho_1, \rho_2, \ldots, \rho_p) \neq (0, 0, \ldots, 0).$$

Contrairement au test de Durbin-Watson, le test de Breusch-Godfrey peut être employé lorsque l'équation de régression contient des variables endogènes retardées $(y_{t-1}, y_{t-2}, \ldots)$ comme variables explicatives.

La statistique est obtenue en appliquant le principe des multiplicateurs de Lagrange (critère LM) dans le contexte du maximum de vraisemblance pour un modèle à erreurs autorégressives.

On a montré, à l'aide d'études de simulation, que ce test est également capable de déceler des erreurs à moyenne mobile. Il peut donc être considéré comme un test général de misspécification dynamique, ce qui le rend très utile.

Nous ne verrons la dérivation formelle de la statistique que dans un cas simple, au chapitre XIV. Cette statistique est facile à interpréter intuitivement: on peut montrer que

cette statistique est *identique* à la statistique LM utilisée pour tester la nullité jointe des ρ_i dans l'équation de régression auxiliaire:

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \rho_1 \hat{u}_{t-1} + \cdots + \rho_p \hat{u}_{t-p} + \eta_t$$

où les $\hat{u}_{t-s}$ sont les résidus de la régression de y_t sur $(1, x_{t2}, \ldots, x_{tk})$ par MCO. Cette statistique a été vue à la section 7.5.

Si H_0 est vraie, on peut montrer que la distribution *limite* (lorsque $n \to \infty$) de cette statistique est une χ^2_p. Cette distribution limite a néanmoins tendance à sous-estimer les valeurs critiques de petit échantillon (ceci a été montré à l'aide d'études de simulation).

Pour cette raison, on utilise souvent une version F de la statistique (test F de $H_0 : \rho_1 = \cdots = \rho_p = 0$ dans l'équation auxiliaire). Les études de simulation ont montré que ceci est préférable lorsque la taille de l'échantillon est faible.

9.8.3 Le test de Koenker (hétéroscédasticité).

Rappelons qu'à la section 9.7, nous avions vu que des variations aléatoires d'un coefficient de régression pouvaient se traduire par une hétéroscédasticité du type:

$$V(u_t) = \alpha + \beta x_t^2$$

où x_t est une variable explicative du modèle estimé.

Si de telles variations aléatoires portent sur plusieurs coefficients d'un modèle de régression multiple, ceci conduit naturellement à l'hypothèse:

$$V(u_t) = \alpha + \beta_1 x_{t1}^2 + \cdots + \beta_p x_{tp}^2$$

ou même, plus généralement:

$$V(u_t) = \alpha + \gamma(\beta_1 x_{t1} + \cdots + \beta_p x_{tp})^2.$$

En pratique, un test acceptable est obtenu en remplaçant $(\beta_1 x_{t1} + \cdots + \beta_p x_{tp})^2$ par $\hat{y}_t^2$, où $\hat{y}_t$ est la valeur calculée en appliquant les MCO à l'équation pour laquelle on veut tester l'hétéroscédasticité des erreurs. On peut donc utiliser un test F de $H_0 : \gamma = 0$ dans l'équation de régression auxiliaire:

$$\hat{u}_t^2 = \alpha + \gamma \hat{y}_t^2 + \eta_t.$$

Cette statistique est basée sur des critères heuristiques, et n'est pas nécessairement la meilleure.

9.8.4 Le test de Bera-Jarque (normalité).

Pour une variable normale $Y \sim N(0,1)$, il est facile de montrer à l'aide de la fonction génératrice des moments que:

$$E(Y^3) = 0 \quad \text{et} \quad E(Y^4) = 3.$$

Si $X \sim N(\mu, \sigma^2)$, $Y = (X - \mu)/\sigma \sim N(0,1)$, et donc:

$$\frac{E(X - E(X))^3}{\sigma^3} = 0,$$

$$\frac{E(X - E(X))^4}{\sigma^4} = 3.$$

La variance σ^2 peut être estimée par:

$$m_2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2.$$

De manière analogue, $E(X - E(X))^3$ peut être estimé par:

$$m_3 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^3,$$

et $E(X - E(X))^4$ peut être estimé par:

$$m_4 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^4.$$

Une déviation de la normalité sera donc indiquée par:

$$\frac{m_3}{(m_2)^{3/2}} \neq 0 \quad \text{et} \quad \frac{m_4}{(m_2)^2} \neq 3.$$

Bera et Jarque ont montré que sous l'hypothèse de normalité, la statistique:

$$n\left[\frac{1}{6}\left(\frac{m_3}{(m_2)^{3/2}}\right)^2 + \frac{1}{24}\left(\frac{m_4}{(m_2)^2} - 3\right)^2\right]$$

a une distribution limite χ^2 avec 2 degrés de liberté lorsque $n \to \infty$.

Nous verrons au chapitre XI que même si les erreurs ne sont pas normales, tous les tests vus précédemment restent approximativement valables (l'approximation est bonne si n est grand). Donc une violation de la normalité a moins d'importance qu'une violation de la sphéricité (à savoir une autocorrélation et/ou une hétéroscédasticité) qui indique, elle, une mauvaise formulation du modèle.

9.9 Exemple numérique

Nous voulons trouver les meilleures estimations linéaires sans biais de a et de b dans le modèle:

$$y_t = a + bx_t + u_t \quad \text{avec} \quad u_t = 0.6\, u_{t-1} + \epsilon_t$$

$$E(\epsilon_t) = 0, \quad V(\epsilon_t) = \sigma^2, \quad E(\epsilon_t \epsilon_s) = 0 \quad (t \neq s)$$

sur la base des données suivantes:

y_t	x_t
8	3
12	6
14	10
15	12
15	14
18	15

On demande en plus la meilleure estimation linéaire sans biais de $y_7 = a + 20b + u_7$.

La matrice X s'écrit:

$$\begin{pmatrix} 1 & 3 \\ 1 & 6 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \\ 1 & 15 \end{pmatrix} .$$

Nous transformons le vecteur y et les deux colonnes de cette matrice selon la règle énoncée à la section 9.3. Ceci donne, puisque $\rho = 0.6$:

$$X^* = \begin{pmatrix} 0.8 & 2.4 \\ 0.4 & 4.2 \\ 0.4 & 6.4 \\ 0.4 & 6.0 \\ 0.4 & 6.8 \\ 0.4 & 6.6 \end{pmatrix} \quad \text{et} \quad y^* = \begin{pmatrix} 6.4 \\ 7.2 \\ 6.8 \\ 6.6 \\ 6.0 \\ 9.0 \end{pmatrix}$$

On vérifie que:

$$(X^*)'X^* = \begin{pmatrix} 1.44 & 13.92 \\ 13.92 & 190.16 \end{pmatrix}$$

$$(X^*)'y^* = \begin{pmatrix} 19.36 \\ 228.92 \end{pmatrix}$$

$$\text{et} \quad \hat{\beta}_{mcg} = (\,(X^*)'X^*\,)^{-1}\,(X^*)'y^* = \begin{pmatrix} 6.1817 \\ 0.7513 \end{pmatrix} \quad .$$

Calculons maintenant le préviseur de y_7 si $x_7' = [1 \quad 20]$. On a:

$$x_7'\hat{\beta}_{mcg} = 6.1817 + (20)(0.7513) = 21.208 \quad .$$

Comme $\hat{u}_6 = 18 - 6.1817 - (15)(0.7513) = 0.5485$, ceci donne:

$$\hat{p} = 21.208 + (0.6)(0.5485) = 21.537.$$

9.10 Introduction aux méthodes semi-paramétriques

Nous avons vu que si $E(uu') = V \neq \sigma^2 I$, la matrice de covariance de l'estimateur de β par moindres carrés ordinaires est égale à:

$$V(\hat{\beta}_{mco}) = (X'X)^{-1}(X'VX)(X'X)^{-1}.$$

Il est possible d'utiliser cette information pour estimer les variances exactes des éléments de $\hat{\beta}_{mco}$ lorsque $V \neq \sigma^2 I$. Ceci donne:

(1) dans le cas de l'hétéroscédasticité seule: l'estimateur de White ("White heteroscetasticity consistent covariance matrix estimator")

(2) dans le cas général où l'on peut avoir hétéroscédasticité *et* autocorrélation: l'estimateur de Newey-West ("Newey-West heteroscedasticity and autocorrelation consistent covariance matrix estimator")

Dans le premier cas, on estime V par:

$$\hat{V} = \begin{pmatrix} \hat{u}_1^2 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{u}_n^2 \end{pmatrix}$$

Dans le second cas, on estime directement $X'VX$ (et non pas V) par une méthode spectrale. Pour une introduction, voir Hamilton, *Time-Series Analysis*, chapitre 10. La méthode nécessite le choix d'une fonction de pondération ("kernel function") et d'un paramètre de troncation ("window width").

En pratique ces méthodes ne donnent de bons résultats que lorsque la taille de l'échantillon est assez grande. Par ailleurs l'estimateur $\hat{\beta}_{mco}$ reste inefficace.

CHAPITRE X.

ÉLÉMENTS DE THÉORIE STATISTIQUE ASYMPTOTIQUE

10.1 Introduction

Les propriétés des estimateurs que nous avons rencontrés lors de l'étude des moindres carrés ordinaires et lors de celle des moindres carrés généralisés si $E(uu')$ est une matrice connue étaient toutes valables quelle que soit la taille n de l'échantillon. Sous l'hypothèse de normalité des erreurs, nous avons pu déterminer leur distribution de façon exacte, en fonction de n. Mais ces distributions exactes prennent vite une forme très complexe lorsque la méthode d'estimation devient plus élaborée, comme c'est le cas pour la méthode *Aitken-réalisable*. Leur étude nécessite des outils théoriques que nous ne pouvons passer en revue ici; l'application empirique de ces résultats dits de *petit échantillon* fait appel à des techniques numériques coûteuses et complexes; de plus, les moments de ces distributions de petit échantillon n'existent pas toujours!

Fort heureusement, la situation devient souvent beaucoup plus simple à la limite, lorsque la taille de l'échantillon tend vers l'infini. C'est ainsi que nous pourrons montrer que lorsque la taille de l'échantillon tend vers l'infini, la distribution de l'estimateur *Aitken-réalisable* tend vers une loi normale. Nous pourrons alors nous baser sur cette loi pour effectuer des tests approximatifs, dits *tests asymptotiques*.

La théorie que nous allons exposer dans ce chapitre sera aussi utilisée pour étudier certains estimateurs proposés lorsque les régresseurs sont stochastiques, notamment dans le cadre des modèles dynamiques et dans celui des systèmes d'équations simultanées.

Elle peut aussi être employée pour faire des tests d'hypothèses dans un modèle de régression linéaire dont les erreurs ne sont pas distribuées normalement, et pour lequel les hypothèses du chapitre VII de cette seconde partie ne sont par conséquent pas vérifiées.

10.2 Convergence en probabilité

Soit (X_n) une suite de variables aléatoires. Cette suite converge en probabilité vers un nombre a si et seulement si:

$$\lim_{n \longrightarrow \infty} P\left[\mid X_n - a \mid > \epsilon\right] = 0 \quad \text{pour tout} \quad \epsilon > 0, \quad \text{aussi petit soit-il.}$$

On écrira alors:

$$\operatorname*{plim}_{n\longrightarrow\infty} X_n = a, \quad \text{ou} \quad X_n \underset{p}{\longrightarrow} a$$

.

Lorsque cette propriété est vérifiée, les densités des X_n tendent vers une densité dont toute la masse est concentrée au point a (distribution dégénérée).

Lorsque a est un paramètre inconnu et X_n un estimateur de a, l'estimateur est dit convergent si $\operatorname{plim}_{n\longrightarrow\infty} X_n = a$.

Si X_n est non aléatoire, la limite en probabilité se réduit à une limite habituelle.

10.3 Inégalité de Chebychev

Enoncé.

Soit X une variable aléatoire continue avec $E(X) = \mu$ et $V(X) = \sigma^2 < \infty$. Pour tout nombre réel $\epsilon > 0$, X vérifie l'inégalité suivante, dite inégalité de Chebychev:

$$P[|X - \mu| > \epsilon] \leq \frac{\sigma^2}{\epsilon^2}.$$

Démonstration

Si X est une variable continue de densité $f_X(x)$, on a par définition de sa variance:

$$\begin{aligned}
\sigma^2 &= \int_{\mathbb{R}} (x-\mu)^2 f_X(x)dx \\
&= \int_{\{x:|x-\mu|>\epsilon\}} (x-\mu)^2 f_X(x)dx + \int_{\{x:|x-\mu|\leq\epsilon\}} (x-\mu)^2 f_X(x)dx \\
&\geq \int_{\{x:|x-\mu|>\epsilon\}} (x-\mu)^2 f_X(x)dx \\
&\geq \epsilon^2 \int_{\{x:|x-\mu|>\epsilon\}} f_X(x)dx = \epsilon^2 P[|X-\mu|>\epsilon]
\end{aligned}$$

10.4 Loi faible des grands nombres

Enoncé. *Soit (Y_n) une suite de variables aléatoires avec $E(Y_n) = \mu$ et $\lim_{n\to\infty} V(Y_n) = 0$. Alors $\operatorname{plim} Y_n = \mu$.*

Démonstration Par l'inégalité de Chebychev, on a, pour tout n et tout $\epsilon > 0$:

$$P[|Y_n - \mu| > \epsilon] \leq \frac{V(Y_n)}{\epsilon^2}.$$

Si $V(Y_n) \to 0$, ceci implique:

$$\lim P[|Y_n - \mu| > \epsilon] \leq \lim \frac{V(Y_n)}{\epsilon^2} = 0.$$

Comme une probabilité ne peut pas être strictement négative, la limite de la probabilité est nulle, ce qui implique le résultat.

Corollaire (généralisation). *Soit* (X_n) *une suite de variables aléatoires. Si:*

$$\lim E(X_n) = \mu \quad et \quad \lim V(X_n) = 0,$$

alors $\operatorname{plim} X_n = \mu$.

Il suffit en effet de poser $Y_n = X_n - E(X_n)$ et d'appliquer le résultat précédent.

Application: Supposons que $X_1, X_2, \ldots, X_n$ soient indépendamment et identiquement distribuées avec $E(X_i) = \mu$, $V(X_i) = \sigma^2$ et considérons la moyenne d'échantillon $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. On a $E(\bar{X}_n) = \mu$ et $\lim_{n\to\infty} V(\bar{X}_n) = 0$, donc $\operatorname{plim}_{n\longrightarrow\infty} \bar{X}_n = \mu$.

10.5 Convergence en distribution

Soit (X_n) une suite de variables aléatoires, et soit (F_{X_n}) la suite de leurs fonctions de distribution. La suite (X_n) converge en distribution vers la variable aléatoire X^*, de distribution F_{X^*}, si et seulement si:

$$\lim_{n\longrightarrow\infty} F_{X_n}(x) = F_{X^*}(x)$$

chaque fois que F_{X^*} est continue en x. On écrira alors:

$$\operatorname*{dlim}_{n\longrightarrow\infty} X_n = X^*, \quad \text{ou} \quad X_n \underset{d}{\longrightarrow} X_*.$$

Ce type de convergence est plus faible que le précédent. Sa principale application est le théorème central limite, que nous verrons plus loin.

Comme exemple, prenons la moyenne $\bar{X}_n$ de n observations X_i indépendantes, d'espérances nulles et de variances unitaires. La loi faible des grands nombres implique $\operatorname{plim} \bar{X}_n = 0$. La suite $\sqrt{n}\bar{X}_n$ ne converge pas en probabilité, mais bien en distribution; on verra par la suite que la distribution limite est normale.

Les moments de la distribution limite F_{X^*} s'appellent moments asymptotiques de X_n. On parle en particulier de l'espérance asymptotique d'un estimateur, ou de sa variance asymptotique; on peut parler de même d'un estimateur asymptotiquement sans biais, ou asymptotiquement efficace. Il est très important de noter que ces moments asymptotiques

ne sont pas définis comme les limites des moments des distributions F_{X_n}, mais bien comme les moments de la distribution limite F_{X^*}! Ceci pour deux raisons: les moments des F_{X_n} peuvent ne pas exister; et les F_{X_n} peuvent ne pas être entièrement caractérisées par leurs moments. Nous pouvons illustrer la première raison en mentionnant que la variance d'une variable Student à un degré de liberté n'existe pas; la seconde en mentionnant que la distribution lognormale (distribution de $Y = e^X$ avec $X \sim N(\mu, \sigma^2)$) n'est pas entièrement caractérisée par ses moments.

Exercice: Soit $n = 10000$ et $m = 1000$. Supposons que l'on ait engendré par simulation nm observations indépendantes x_{ij} de distribution uniforme sur l'intervalle $[-1, 1]$, pour $i = 1, \ldots, n$ et $j = 1, \ldots, m$. On calcule, pour $j = 1, \ldots, m$, les moyennes $\bar{x}^j = n^{-1} \sum_{i=1}^{n} x_{ij}$. A quoi ressemblera l'histogramme des $\bar{x}^j$? A quoi ressemblera l'histogramme des $\sqrt{n}\bar{x}^j$?

10.6 Propriétés des modes de convergence

10.6.1 Relation entre limite en probabilité et limite en distribution.

Enoncé. *Soit* (X_n, Y_n) *une suite de paires de variables aléatoires. Si* $\operatorname{plim}(X_n - Y_n) = 0$ *et* $\operatorname{dlim} Y_n = Y^*$, *alors* $\operatorname{dlim} X_n = Y^*$.

Cette propriété possède une réciproque partielle. Si $\operatorname{dlim} X_n = a$ et $\operatorname{dlim} Y_n = a$, avec a constante, alors $\operatorname{plim}(X_n - Y_n) = 0$. Cette réciproque est intuitivement évidente puisqu'une constante a une distribution dégénérée.

Mentionnons qu'une même distribution limite de X_n et de Y_n n'implique pas que $\operatorname{plim}(X_n - Y_n) = 0$, lorsque cette distribution limite n'est pas dégénérée. En effet, si les X_n et les Y_n possèdent une distribution commune normale réduite, et que X_n est indépendante de Y_n pour tout n, on a $F_{X_n - Y_n} = N(0, 2)$ pour tout n. Par conséquent, $\operatorname{dlim}(X_n - Y_n) \sim N(0, 2)$. Mais ceci n'implique nullement que $\operatorname{plim}(X_n - Y_n) = 0$, puisque pour tout $\epsilon > 0$, et pour tout n, $P[\mid X_n - Y_n \mid > \epsilon] \neq 0$.

10.6.2 Théorème de Slutsky.

Ce théorème établit la préservation des limites en probabilité par les fonctions continues:

Enoncé. *Si* $\operatorname{plim} X_n = a$ *et* $g(X_n)$ *est continue en* a, *alors* $\operatorname{plim}[g(X_n)] = g[\operatorname{plim}(X_n)] = g(a)$.

Il est important de noter que la fonction g ne peut dépendre de n. Ce théorème possède les généralisations suivantes (on définit la limite en probabilité d'une matrice comme la matrice contenant les limites en probabilité des éléments):

(1) Si (A_n) et (B_n) sont deux suites de matrices conformes pour l'addition, alors $\operatorname{plim}(A_n + B_n) = \operatorname{plim}(A_n) + \operatorname{plim}(B_n)$ si $\operatorname{plim}(A_n)$, $\operatorname{plim}(B_n)$ existent.

(2) Si (A_n), (B_n) sont deux suites de matrices conformes pour la multiplication et si $\text{plim}(A_n)$, $\text{plim}(B_n)$ existent, on a: $\text{plim}(A_n B_n) = \text{plim}(A_n)\,\text{plim}(B_n)$.

(3) Si (A_n) est une suite de matrices régulières et si $\text{plim}(A_n)$ existe et est régulière, alors: $\text{plim}(A_n^{-1}) = (\text{plim}\, A_n)^{-1}$.

10.6.3 Convergence en distribution de fonctions de variables aléatoires.

Enoncé.

(1) *Si g est continue et si* $\text{dlim}\, X_n = X$, *alors* $\text{dlim}\, g(X_n) = g(X)$

(2) *Supposons que* $\text{dlim}\, Y_n = Y^*$ *et que* $\text{plim}\, X_n = a$, *avec a constante. Alors:*

$$\text{dlim}(X_n + Y_n) = a + Y^*$$

$$\text{dlim}(X_n Y_n) = aY^*$$

$$\text{dlim}\left(\frac{Y_n}{X_n}\right) = \frac{Y^*}{a} \quad \textit{si } a \neq 0.$$

Dans le cas de convergence en distribution vers une normale, on peut énoncer une généralisation multivariée de ce résultat. Nous admettrons qu'une suite de vecteurs aléatoires $X^{(n)} = (X_1^{(n)}, \ldots, X_m^{(n)})$ converge en distribution vers un vecteur normal multivarié $X = (X_1, \ldots, X_m)$ si toute combinaison linéaire $\sum_{i=1}^m \lambda_i X_i^{(n)}$ converge en distribution vers $\sum_{i=1}^m \lambda_i X_i$. Supposons alors que l'on ait une suite de matrices $A^{(n)}$ convergeant en probabilité vers A et que la suite des vecteurs $X^{(n)}$ converge en distribution vers un vecteur $X \sim N(0, I)$. La suite $A^{(n)}X^{(n)}$ converge en distribution vers un vecteur ayant la distribution $N(0, AA')$.

10.7 Fonction caractéristique et convergence en distribution

Nous aurons, lorsque nous verrons le théorème central limite, à déterminer la distribution limite d'une somme de variables aléatoires. Calculer la distribution d'une somme $X + Y$, connaissant la distribution jointe de X et Y, est en règle générale un problème très difficile. Le passage par les fonctions caractéristiques permet souvent de simplifier les choses.

Si l'on dénote par i l'unité imaginaire ($i^2 = -1$), la fonction caractéristique d'une variable aléatoire X est définie comme:

$$\begin{aligned} \phi_X(t) &= E\left[e^{itX}\right] \\ &= E\left[\cos(tX)\right] + iE\left[\sin(tX)\right], \quad \text{en vertu des propriétés du} \\ &\qquad \text{nombre complexe} \quad e^{itX}. \end{aligned}$$

Avant de donner un exemple de fonction caractéristique, mentionnons quatre de ses propriétés:

(1) La fonction caractéristique d'une variable aléatoire existe toujours.
En effet, $\cos(tX)$ et $\sin(tX)$ sont des fonctions périodiques, donc bornées pour toute

valeur de tX; l'espérance mathématique d'une fonction bornée existe toujours. Nous ne pourrions en dire autant pour $E(e^{tX})$ par exemple.

(2) La fonction caractéristique de X caractérise entièrement la distribution de X.

(3) Si X et Y sont deux variables aléatoires indépendantes, alors: $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

En effet,
$$\begin{aligned}\phi_{X+Y}(t) &= E\left[e^{it(X+Y)}\right] \\ &= E\left[e^{itX}e^{itY}\right] \\ &= E\left[e^{itX}\right]\ E\left[e^{itY}\right]\end{aligned}$$

par l'hypothèse d'indépendance.

Cette propriété facilite le calcul de la distribution de $X+Y$. Si le produit des fonctions caractéristiques est la fonction caractéristique d'une distribution connue, cette distribution est celle de $X+Y$.

(4) Soit (X_n) une suite de variables aléatoires, et soit (ϕ_{X_n}) la suite de leurs fonctions caractéristiques. Supposons que:

(i) $\lim_{n\longrightarrow\infty}\phi_{X_n}(t) = \phi(t)$

(ii) $\phi(t)$ soit continue pour $t = 0$.

Alors $\phi(t)$ est une fonction caractéristique, celle de $\operatorname{dlim} X_n$. Plus précisement:

a) $\operatorname{dlim} X_n = X^*$, et

b) $E\left[e^{itX^*}\right] = \phi(t)$.

Cette dernière propriété nous permettra de démontrer le théorème central limite. Mais à titre d'exemple, nous allons tout d'abord calculer la fonction caractéristique d'une variable normale.

Soit donc $X \sim N(\mu, \sigma^2)$. On a $E\left[e^{itX}\right] = e^{it\mu}E\left[e^{it(X-\mu)}\right]$. Pour calculer $E\left[e^{it(X-\mu)}\right]$, faisons le changement de variable $y = x - \mu$. On a $dy = dx$, et donc:

$$\begin{aligned}E\left[e^{it(X-\mu)}\right] &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{ity}e^{-\frac{y^2}{2\sigma^2}}dy \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}(y^2-2\sigma^2ity)}dy \\ &= \frac{1}{\sigma\sqrt{2\pi}}e^{i^2t^2\sigma^2/2}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}(y^2-2\sigma^2ity+i^2t^2\sigma^4)}dy \\ &= \frac{1}{\sigma\sqrt{2\pi}}e^{-t^2\sigma^2/2}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}(y-it\sigma^2)^2}dy \quad .\end{aligned}$$

Faisons maintenant le changement de variable $v = y - it\sigma^2$. On a $dv = dy$, et donc:

$$\begin{aligned} E\left[e^{it(X-\mu)}\right] &= e^{-t^2\sigma^2/2} \quad \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{v^2}{2\sigma^2}} dv \\ &= e^{-t^2\sigma^2/2} \\ \text{Par conséquent} \quad \phi_X(t) &= e^{it\mu} e^{-t^2\sigma^2/2} \\ &= e^{it\mu - t^2\sigma^2/2} \quad . \end{aligned}$$

10.8 Versions du théorème central limite

10.8.1 Variables indépendantes, identiquement distribuées.

L'énoncé qui va suivre porte le nom de théorème de Lindeberg-Levy. Il s'applique à des variables aléatoires indépendantes et identiquement distribuées. Il permet notamment de traiter le problème de l'approximation d'une binomiale par une normale.

Théorème. *Soit (Z_i) une suite de variables indépendantes et identiquement distribuées avec $E(Z_i) = \mu$ et $V(Z_i) = \sigma^2$. Soit:*

$$X_i = \frac{Z_i - \mu}{\sigma}$$

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i = \frac{\sqrt{n}(\bar{Z} - \mu)}{\sigma}$$

On a $\operatorname{dlim} S_n \sim N(0,1)$.

Démonstration

Puisque, en général:

$$e^X = 1 + X + \frac{X^2}{2} + \frac{X^3}{3!} + \ldots,$$

on a, en appliquant cette formule à $Y_j = \frac{X_j}{\sqrt{n}}$:

$$\phi_{Y_j}(t) = E\left[e^{itY_j}\right] = 1 + itE(Y_j) + \frac{(it)^2}{2} E(Y_j^2) + \ldots .$$

Mais, puisque $E(Y_j) = 0$ et $E(Y_j^2) = \frac{1}{n}$, ceci implique:

$$\phi_{Y_j}(t) = 1 + 0 + \frac{(it)^2}{2n} + \ldots .$$

Si n est grand, on peut négliger les termes d'ordre supérieur à 2, et donc:

$$\phi_{Y_j}(t) \approx 1 - \frac{t^2}{2n} \quad .$$

Puisque les Y_j sont indépendantes, la fonction caractéristique de leur somme est le produit des fonctions caractéristiques des Y_j. Par conséquent:

$$\phi_{S_n}(t) \approx (1 - \frac{t^2}{2n})^n \quad \text{pour} \quad n \quad \text{grand} \quad .$$

Pour pouvoir appliquer la quatrième propriété des fonctions caractéristiques, nous calculons maintenant:

$$\lim_{n \longrightarrow \infty} (1 - \frac{t^2}{2n})^n.$$

Comme:

$$(1 - \frac{t^2}{2n})^n = (1 + \frac{(-t^2/2)}{n})^n$$

et comme:

$$\lim_{n \longrightarrow \infty} (1 + \frac{X}{n})^n = e^X$$

on a:

$$\lim_{n \longrightarrow \infty} \phi_{S_n}(t) = e^{-t^2/2}$$

qui est continue au point $t = 0$. Nous reconnaissons la fonction caractéristique d'une variable $N(0,1)$; par conséquent $\operatorname{dlim} S_n \sim N(0,1)$.

Terminons cette section en montrant que ce théorème permet d'approcher une binomiale par une normale. Soit donc Y une variable aléatoire prenant comme valeur le nombre de succès rencontré lors de n tirages effectués avec remise (et donc indépendants), la probabilité d'obtenir un succès lors de l'un quelconque de ces tirages étant égale à p. Nous pouvons écrire: $Y = \sum_{i=1}^{n} Z_i$, où Z_i est une variable aléatoire prenant la valeur 1 avec la probabilité p, la valeur 0 avec la probabilité $(1-p)$. On vérifie immédiatement que $E(Z_i) = p$ et $V(Z_i) = p(1-p)$. Par conséquent, $E(Y) = np$ et $V(Y) = np(1-p)$. Donc, si l'on définit:

$$X_i = \frac{Z_i - p}{\sqrt{p(1-p)}}$$

on a:

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i = \frac{Y - np}{\sqrt{np(1-p)}}.$$

Le théorème central limite est applicable, et $\operatorname{dlim} S_n \sim N(0,1)$. Si n est suffisamment grand, on peut alors approcher une binomiale de paramètres n et p par une normale d'espérance np et de variance $np(1-p)$.

10.8.2 Variables indépendantes, non identiquement distribuées.

Cette seconde version s'applique à des variables indépendantes, mais de distributions non identiques. Pour illustrer son importance, rappelons que dans le modèle de régression simple $y_t = a + bx_t + u_t$, nous avions démontré que $\hat{b} - b = \sum_{t=1}^{n} w_t u_t$ avec $w_t = \frac{x_t - \bar{x}}{\sum (x_t - \bar{x})^2}$. L'estimateur de b par moindres carrés est donc, à une constante près, une somme de variables aléatoires $w_t u_t$. Mais ces variables ne sont pas identiquement distribuées puisque $w_t \neq w_s$ pour $t \neq s$.

Le théorème suivant, dont on trouvera l'énoncé dans Judge et al., *The Theory and Practice of Econometrics*, 1985, p. 156, remplace l'hypothèse de distributions identiques par une condition sur les troisièmes moments des variables. Nous nous bornerons par la suite à faire l'hypothèse que cette condition est vérifiée, chaque fois que nous aurons besoin du théorème. Nous énoncerons ce théorème sous sa forme vectorielle, sans le démontrer.

Théorème.

Soit (Z_t) une suite de vecteurs aléatoires indépendants avec $E(Z_t) = 0$, et $V(Z_t) = E(Z_t Z_t') = \Phi_t$. Supposons que les deux conditions suivantes soient vérifiées:

(1) $\lim_{n \longrightarrow \infty} \frac{1}{n} \sum_{t=1}^{n} \Phi_t = \Phi$, *avec Φ définie positive*

(2) $E(Z_{it} Z_{jt} Z_{kt}) < \infty$ *pour tout i, j, k, t.*

Alors, si $S_n = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_t$, on a $\operatorname{dlim} S_n \sim N(0, \Phi)$.

Exercice. Pour le modèle de régression simple $y_t = a + bx_t + u_t$ sous les hypothèses du chapitre I de la seconde partie, trouvez la distribution limite de $\sqrt{n}(\hat{b} - b)$, où $\hat{b}$ est l'estimateur de b par moindres carrés ordinaires. Comment interpréter ce résultat?

10.8.3 Différences de martingales.

Lorsque nous étudierons les modèles dynamiques, nous aurons à examiner la convergence en distribution de suites de vecteurs aléatoires de la forme $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_t$, où les vecteurs Z_t sont dépendants entre eux. Nous devrons alors utiliser une généralisation des théorèmes précédents. Une telle généralisation existe dans le cas où la dépendance prend une forme particulière, celle des *différences de martingales*.

Définition:

Une suite $(Z_t)_{t=1}^{\infty}$ de variables aléatoires, ou de vecteurs aléatoires, est une différence de martingale si:

$$E(Z_t) = 0 \quad \text{pour tout } t;$$

$$E(Z_t \mid Z_{t-1}, Z_{t-2}, \ldots, Z_1) = 0 \quad \text{pour tout } t.$$

Exemple:

Dans le cadre des modèles à variables endogènes retardées, nous rencontrerons des suites (Z_t) de la forme $Z_t = u_t u_{t-1}$, où les u_t sont indépendantes, d'espérance nulle, et identiquement distribuées. Il est facile de vérifier que les Z_t forment une différence de martingale:

$$E(Z_t) = E(u_t u_{t-1}) = E(u_t)E(u_{t-1}) = 0$$

$$\begin{aligned} E(Z_t \mid Z_{t-1}, \ldots, Z_1) &= E(Z_t \mid Z_{t-1}) \\ &= E(u_t u_{t-1} \mid u_{t-1} u_{t-2}) \\ &= E_{u_{t-1}} E(u_t u_{t-1} \mid u_{t-1} u_{t-2}, u_{t-1}) \\ &= E_{u_{t-1}} E(u_t u_{t-1} \mid u_{t-1}, u_{t-2}) \\ &= E_{u_{t-1}} u_{t-1} E(u_t \mid u_{t-1}, u_{t-2}) = 0 \end{aligned}$$

La troisième égalité résulte de la loi des espérances itérées, et la quatrième vient du fait que la connaissance de $u_{t-1}u_{t-2}$ et de u_{t-1} est équivalente à celle de u_{t-1} et de u_{t-2}, sauf si $u_{t-1} = 0$; mais si $u_{t-1} = 0$, l'espérance est nulle et l'égalité est donc vérifiée.

Le théorème suivant est énoncé dans Hamilton, *Time-Series Analysis*, 1994, p. 194. Il suppose l'existence des quatre (et non plus trois) premiers moments.

Théorème.

Soit (Z_t) une différence de martingale. Si:

(1) *Les matrices de covariance $V(Z_t)$ sont définies positives;*

(2) $$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} V(Z_t) = \Phi, \quad \textit{une matrice définie positive;}$$

(3) $$E(Z_{it} Z_{jt} Z_{lt} Z_{mt}) < \infty \quad \textit{pour tout } t,i,j,l,m;$$

(4) $$\frac{1}{n} \sum_{t=1}^{n} Z_t Z_t' \underset{p}{\longrightarrow} \Phi$$

alors:

$$\operatorname{dlim} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_t \sim N(0, \Phi)$$

Exercice: On pose le modèle $y_t = by_{t-1} + u_t$ où les u_t sont indépendantes, d'espérances nulles, et identiquement distribuées. Si $b = 0$ et si $\hat{b} = \sum_t y_{t-1} y_t / \sum_t y_{t-1}^2$, montrez que la distribution limite de $\sqrt{n}\hat{b}$ est normale réduite.

10.9 L'Inégalité de Rao-Cramer

Commençons par fournir le fil directeur de cette section et de la suivante. L'inégalité de Rao-Cramer, que nous démontrerons, fournit une borne inférieure de la variance d'un estimateur sans biais. Une généralisation vectorielle de cette inégalité mène à la matrice d'information, dont l'inverse est la matrice de covariance asymptotique du vecteur des estimateurs par maximum de vraisemblance. Cette matrice permet alors d'effectuer des tests asymptotiques même lorsque l'on ne connaît pas la distribution de petit échantillon des estimateurs de maximum de vraisemblance, comme c'est le cas dans beaucoup de modèles non linéaires. La matrice d'information possède donc un intérêt double, à la fois théorique (efficacité asymptotique) et pratique (calcul de covariances asymptotiques).

Les démonstrations de cette section utiliseront l'hypothèse que les observations sont indépendantes et identiquement distribuées; mais des résultats analogues peuvent être prouvés sous des hypothèses plus générales.

Lemme.

Supposons que θ soit scalaire et soit $(X_1, X_2, \ldots, X_n)$ un échantillon aléatoire de fonction de vraisemblance:

$$L(x, \theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Supposons que L soit deux fois différentiable, et que:

$$\frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} L(x, \theta) dx = \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} L(x, \theta) dx.$$

Alors:

$$V\left(\frac{\partial \log L(x, \theta)}{\partial \theta}\right) = E\left(\frac{\partial \log L(x, \theta)}{\partial \theta}\right)^2 = -E\left(\frac{\partial^2 \log L(x, \theta)}{\partial \theta^2}\right).$$

Démonstration

Puisque $L(x, \theta)$ peut être considérée comme la densité jointe de l'échantillon, on a $\int_{\mathbb{R}^n} L(x, \theta) dx = 1$. En dérivant par rapport à θ, ceci donne:

$$\frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} L(x, \theta) dx = 0 = \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} L(x, \theta) dx.$$

Mais $\frac{\partial L}{\partial \theta} = \frac{\partial \log L}{\partial \theta} L$. On a donc aussi:

$$\int_{\mathbb{R}^n} \frac{\partial \log L(x, \theta)}{\partial \theta} L(x, \theta) dx = E\left(\frac{\partial \log L(x, \theta)}{\partial \theta}\right) = 0.$$

En dérivant une nouvelle fois par rapport à θ, il vient:

$$\int_{\mathbb{R}^n} \left[\frac{\partial^2 \log L}{\partial \theta^2} L + \frac{\partial \log L}{\partial \theta} \frac{\partial L}{\partial \theta} \right] dx = 0,$$

$$\text{ou encore:} \quad \int_{\mathbb{R}^n} \frac{\partial^2 \log L}{\partial \theta^2} L dx + \int_{\mathbb{R}^n} \left(\frac{\partial \log L}{\partial \theta} \right)^2 L dx = 0 \quad .$$

Soit aussi, puisque $E\left(\frac{\partial \log L}{\partial \theta}\right) = 0$:

$$V\left(\frac{\partial \log L}{\partial \theta}\right) = E\left(\frac{\partial \log L}{\partial \theta}\right)^2 = -E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right).$$

Inégalité de Rao-Cramer. *Soit $\hat{\theta} = \hat{\theta}(x)$ un estimateur sans biais de θ. On a l'inégalité:*

$$V(\hat{\theta}) \geq -\frac{1}{E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right)}$$

Démonstration

Comme $\hat{\theta}$ est sans biais, on a:

$$\theta = E(\hat{\theta}) = \int_{\mathbb{R}^n} \hat{\theta} L(x, \theta) dx \quad .$$

En dérivant par rapport à θ, il vient:

$$\begin{aligned} 1 &= \int_{\mathbb{R}^n} \hat{\theta} \frac{\partial L}{\partial \theta} dx = \int_{\mathbb{R}^n} \hat{\theta} \frac{\partial \log L}{\partial \theta} L dx \\ &= \operatorname{cov}\left(\hat{\theta}, \frac{\partial \log L}{\partial \theta}\right) \quad \text{puisque} \quad E\left(\frac{\partial \log L}{\partial \theta}\right) = 0 \quad . \end{aligned}$$

D'autre part, en vertu de l'inégalité générale $(\operatorname{cov}(X,Y))^2 \leq V(X)V(Y)$, nous avons:

$$1 = \operatorname{cov}^2\left(\hat{\theta}, \frac{\partial \log L}{\partial \theta}\right) \leq V(\hat{\theta}) \quad V\left(\frac{\partial \log L}{\partial \theta}\right) \quad ,$$

ou, en vertu du lemme:

$$1 \leq -V(\hat{\theta})\, E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) \quad . \quad \text{L'inégalité cherchée s'ensuit.}$$

Pour illustrer ce résultat, reprenons le problème de l'estimation par maximum de vraisemblance de l'espérance mathématique μ d'une variable normale, discuté à la section 3.3 de la première partie. Nous avions trouvé:

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2} \sum (x_i - \mu)$$

et donc

$$\frac{\partial^2 \log L}{\partial \mu^2} = -\frac{n}{\sigma^2} \quad .$$

En vertu de l'inégalité précédente, on a alors $V(\hat{\mu}) \geq \frac{\sigma^2}{n}$ si $E(\hat{\mu}) = \mu$. Mais nous savons que $E(\bar{x}) = \mu$ et $V(\bar{x}) = \frac{\sigma^2}{n}$. Nous concluons que cet estimateur est efficace.

Notons qu'un estimateur peut être efficace sans que sa variance atteigne cette borne inférieure!

10.10 La matrice d'information

Préoccupons-nous maintenant de l'estimation d'un vecteur aléatoire

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}$$

Soit $\hat{\theta}$ un estimateur sans biais de θ.

Nous admettons sans démonstration les généralisations suivantes des résultats précédents:

$$\begin{aligned} E\left(\frac{\partial \log L}{\partial \theta}\right) &= 0 \quad (\text{un vecteur } k \times 1) \\ V\left(\frac{\partial \log L}{\partial \theta}\right) &= -E\left(\frac{\partial^2 \log L}{\partial \theta \partial \theta'}\right) \quad (\text{une matrice } k \times k) \\ &\underset{\text{def}}{=} R(\theta) \quad . \end{aligned}$$

La matrice $R(\theta)$ s'appelle matrice d'information. Nous la supposerons régulière.

En lieu et place de $\operatorname{cov}\left(\hat{\theta}, \frac{\partial \log L}{\partial \theta}\right) = 1$, nous écrivons:

$$E\left(\hat{\theta}\, \frac{\partial \log L}{\partial \theta'}\right) = I \quad (\text{une matrice } k \times k)$$

et par conséquent:

$$V\begin{pmatrix} \hat{\theta} \\ \dfrac{\partial \log L}{\partial \theta} \end{pmatrix} = \begin{pmatrix} V(\hat{\theta}) & I \\ I & R(\theta) \end{pmatrix} .$$

Cette dernière matrice est définie non négative, étant une matrice de covariance. Afin d'arriver à une généralisation vectorielle de l'inégalité de Rao-Cramer, considérons un vecteur colonne arbitraire a. Comme la matrice est définie non négative, on a:

$$(\, a' \quad -a' R^{-1}(\theta)\,) \begin{pmatrix} V(\hat{\theta}) & I \\ I & R(\theta) \end{pmatrix} \begin{pmatrix} a \\ -R^{-1}(\theta)a \end{pmatrix} \geq 0$$

soit en effectuant et en simplifiant:

$$a' \left[V(\hat{\theta}) - R^{-1}(\theta) \right] a \geq 0 .$$

Donc la matrice $V(\hat{\theta}) - R^{-1}(\theta)$ est définie non négative. On a en particulier $V(\hat{\theta}_i) \geq \left[R^{-1}(\theta)\right]_{ii}$ pour tout i.

Illustrons maintenant ce résultat. Nous avons vu à la Section 5.8 que dans le modèle $y = X\beta + u$ avec $u \sim N(0, \sigma^2 I)$, la matrice $\frac{\partial^2 \log L}{\partial \theta \partial \theta'}$ prenait la forme:

$$H = \begin{pmatrix} -\dfrac{(X'X)}{\sigma^2} & -\dfrac{1}{\sigma^4} X' u \\ -\dfrac{1}{\sigma^4} u' X & \dfrac{n}{2\sigma^4} - \dfrac{1}{\sigma^6} u' u \end{pmatrix} .$$

Par conséquent:

$$R(\theta) = -E(H) = \begin{pmatrix} \dfrac{(X'X)}{\sigma^2} & O_{k\times 1} \\ O_{1\times k} & \dfrac{n}{2\sigma^4} \end{pmatrix} \quad \text{puisque} \quad E(u'u) = n\sigma^2 .$$

Donc $R^{-1}(\theta)$ est diagonale par blocs, et pour tout estimateur sans biais $\tilde{\beta}$ de β, la matrice $V(\tilde{\beta}) - \sigma^2 (X'X)^{-1}$ est définie non négative en vertu du résultat précédent, lorsque les erreurs sont distribuées normalement. Mais si $\hat{\beta} = (X'X)^{-1}X'y$, $V(\hat{\beta})$ est précisément égale à $\sigma^2 (X'X)^{-1}$.

La "borne inférieure" est atteinte par cette matrice: nous concluons que sous l'hypothèse de normalité, $\hat{\beta} = (X'X)^{-1}X'y$ n'est pas seulement le meilleur estimateur *linéaire* sans biais. C'est aussi le meilleur estimateur sans biais parmi tous les estimateurs, qu'ils soient linéaires ou non.

10.11 Propriétés asymptotiques des estimateurs par maximum de la vraisemblance

10.11.1 Cas scalaire.

Nous avons ici le cas de l'estimation d'un seul paramètre θ. La vraisemblance s'écrit $L(x,\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ comme précédemment, et l'estimateur $\hat{\theta}$ est une solution de l'équation $\frac{\partial \log L(x,\theta)}{\partial \theta} = 0$.

On démontre que sous des hypothèses assez générales, et qui n'impliquent pas la normalité, l'estimateur $\hat{\theta}$ est convergent, asymptotiquement normal, asymptotiquement sans biais, et asymptotiquement efficace. En effet, sous ces hypothèses:

$$\operatorname{plim} \hat{\theta} = \theta$$
$$\operatorname{dlim} \sqrt{n}(\hat{\theta} - \theta) \sim N\left(0, \operatorname{plim} \frac{n}{-E\left(\frac{\partial^2 \log L(x,\theta)}{\partial \theta^2}\right)}\right)$$

La borne inférieure est donc atteinte par la variance asymptotique de $\hat{\theta}$.

10.11.2 Cas vectoriel.

Dans le cas où θ est un vecteur, on démontre sous des hypothèses semblables aux précédentes les généralisations suivantes. Soit $\hat{\theta}$ le vecteur des estimateurs par maximum de vraisemblance. Alors:

$$\operatorname{plim} \hat{\theta} = \theta$$
$$\operatorname{dlim} \sqrt{n}(\hat{\theta} - \theta) \sim N\left(0, \operatorname{plim} nR^{-1}(\theta)\right)$$

où:

$$R(\theta) = -E\left[\frac{\partial^2 \log L(x,\theta)}{\partial \theta \partial \theta'}\right]$$

est la matrice d'information vue précédemment.

10.12 Distribution asymptotique du rapport des vraisemblances

10.12.1 Introduction.

Rappelons que la méthode du rapport des vraisemblances, vu à la section 5.3 de la première partie, se résume ainsi: Dans le test $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$, on calcule

$$\lambda = \frac{\max_{H_0} L(x,\theta)}{\max_{\Omega} L(x,\theta)}.$$

On a les inégalités $0 \leq \lambda \leq 1$.

On rejette H_0 si $\lambda < \lambda_0$, où λ_0 est un nombre strictement compris entre 0 et 1 et choisi en fonction d'un niveau de signification α.

Dans les cas que nous avons traités jusqu'ici, cette méthode nous a permis de trouver une règle de décision valable pour de petits échantillons, et faisant appel à une statistique possédant une distribution connue (Student, par exemple). Mais, il existe de nombreux modèles non linéaires où ceci n'est pas le cas. On doit alors se contenter de tests asymptotiques. Il est donc intéressant de connaître la distribution asymptotique d'une fonction de λ.

10.12.2 Cas scalaire.

Lorsque le vecteur θ n'a qu'une seule composante, nous allons montrer que sous H_0, $\operatorname{dlim}(-2\log_e \lambda) \sim \chi^2_{(1)}$. Notre démonstration utilise l'hypothèse que les observations sont indépendantes et identiquement distribuées, mais le résultat peut être généralisé.

Soit $\hat{\theta}$ l'estimateur de θ par maximum de vraisemblance. Nous commençons par faire un développement de $\log L(x,\theta_0)$ autour de $\hat{\theta}$ (théorème de Taylor). Ceci donne:

$$\log L(x,\theta_0) - \log L(x,\hat{\theta}) = (\theta_0 - \hat{\theta}) \left.\frac{\partial \log L(x,\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} + \frac{1}{2}(\theta_0 - \hat{\theta})^2 \left.\frac{\partial^2 \log L(x,\theta)}{\partial \theta^2}\right|_{\theta=\theta^*}$$

où θ^* est un point de l'intervalle ouvert reliant θ_0 et $\hat{\theta}$.

Comme $\left.\frac{\partial \log L}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$ par définition de $\hat{\theta}$, nous pouvons réécrire cette équation comme:

$$\log \frac{L(x,\theta_0)}{L(x,\hat{\theta})} = \frac{1}{2}(\theta_0 - \hat{\theta})^2 \left.\frac{\partial^2 \log L}{\partial \theta^2}\right|_{\theta=\theta^*},$$

soit aussi:

$$\begin{aligned} -2\log\lambda &= \left[\sqrt{n}(\hat{\theta}-\theta_0)\right]^2 \left(-\frac{1}{n}\right) \left(\left.\frac{\partial^2 \log L}{\partial \theta^2}\right|_{\theta=\theta^*}\right) \\ &= \left[\sqrt{n}(\hat{\theta}-\theta_0)\right]^2 \left(-\frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^2 \log f(x_i|\theta)}{\partial \theta^2}\right|_{\theta=\theta^*}\right) \\ &\underset{\text{def}}{=} \left[\sqrt{n}(\hat{\theta}-\theta_0)\right]^2 k^2 \quad . \end{aligned}$$

Comme $\hat{\theta}$ est convergent, on a, sous l'hypothèse H_0, $\operatorname{plim}\hat{\theta} = \theta_0$. Comme θ^* est compris entre θ_0 et $\hat{\theta}$, ceci implique:

$$\begin{aligned}\operatorname{plim} k^2 &= \operatorname{plim}\left(-\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \log f(x_i|\theta)}{\partial\theta^2}\right|_{\theta=\theta_0}\right)\\ &= -E\left(\left.\frac{\partial^2 \log f(x_i|\theta)}{\partial\theta^2}\right|_{\theta=\theta_0}\right)\end{aligned}$$

sous l'hypothèse que les termes $\left.\frac{\partial^2 \log f(x_i|\theta)}{\partial\theta^2}\right|_{\theta=\theta_0}$ sont de variance finie (ils sont en effet identiquement distribués). Ceci est une conséquence des résultats de la section 10.4. De plus, comme nous l'avons vu:

$$\operatorname{dlim}\left[\sqrt{n}(\hat{\theta}-\theta_0)\right] \sim N\left(0, \operatorname{plim}\frac{1}{-\frac{1}{n}E\left(\left.\frac{\partial^2 \log L}{\partial\theta^2}\right|_{\theta=\theta_0}\right)}\right)$$

sous l'hypothèse H_0.

Comme:

$$\operatorname{plim}\left[-\frac{1}{n}\,E\left(\left.\frac{\partial^2 \log L}{\partial\theta^2}\right)\right|_{\theta=\theta_0}\right] = -E\left(\left.\frac{\partial^2 \log f(x_i|\theta)}{\partial\theta^2}\right|_{\theta=\theta_0}\right) = \operatorname{plim} k^2,$$

ceci implique:

$$\operatorname{dlim}\left[\sqrt{n}(\hat{\theta}-\theta_0)\right] \sim N\left(0, \frac{1}{\operatorname{plim} k^2}\right) \quad .$$

Alors, en vertu des résultats de la section 10.6:

$$\begin{aligned}\operatorname{dlim}(-2\log\lambda) &= \operatorname{dlim}\left[\sqrt{n}(\hat{\theta}-\theta_0)\right]^2 \operatorname{plim} k^2\\ &= X^2 \operatorname{plim} k^2 \quad \text{où} \quad X \sim N\left(0, \frac{1}{\operatorname{plim} k^2}\right) \quad .\end{aligned}$$

Définissons maintenant $Y = (\operatorname{plim} k)X$. Comme $Y \sim N(0,1)$, $Y^2 = (\operatorname{plim} k^2)X^2 = \operatorname{dlim}(-2\log\lambda)$ est $\chi^2_{(1)}$, ce qu'il fallait démontrer.

10.12.3 Cas vectoriel.

Nous avons un vecteur θ de k paramètres à estimer et nous voulons tester l'hypothèse $H_0 : \theta_1 = \theta_1^*$ contre $H_1 : \theta_1 \neq \theta_1^*$ où θ_1 est un sous-vecteur de θ de dimension q. On montre alors que $\operatorname{dlim}(-2\log_e\lambda) \sim \chi^2_{(q)}$.

10.13 Exemple d'application dans un modèle à erreurs autorégressives

Dans le modèle de régression classique, nous avons vu, sous l'hypothèse de normalité des erreurs, que $\hat{\beta}_{mco}$ est normal quelle que soit la taille de l'échantillon. De plus, le rapport des vraisemblances λ permet de dériver un test F d'une hypothèse linéaire; ce test est, lui aussi, valable pour tout n. La distribution de Student permet de calculer des intervalles de confiance.

Dans le modèle des moindres carrés généralisés où $E(uu') = \sigma^2\Omega$, nous avons les mêmes résultats lorsque Ω est connue. Par contre, si Ω est inconnue, nous n'avons plus de résultats valables en petit échantillon. Mais si u est un vecteur normal, on peut dériver l'estimateur de β par maximum de la vraisemblance. Cet estimateur n'est pas normal car c'est une fonction non linéaire des erreurs. Néanmoins, on peut en trouver la distribution asymptotique à l'aide des résultats précédents.

Pour le modèle à erreurs autorégressives:

$$y = X\beta + u, \quad \text{avec} \quad u_t = \rho u_{t-1} + \epsilon_t$$

où les ϵ_t sont indépendantes de distribution $N(0, \sigma_\epsilon^2)$ et où X est non aléatoire, l'estimateur par maximum de vraisemblance a été étudié par Beach et MacKinnon, "A maximum likelihood procedure for regression with autocorrelated errors", *Econometrica* 46 (1978), 51–58. Nous allons brièvement discuter les résultats de ces auteurs.

Rappelons que $E(uu') = V = \sigma_u^2\Omega$, où Ω est la matrice de la section 9.2, et que $\sigma_\epsilon^2 = (1-\rho^2)\sigma_u^2$. En utilisant la définition de la densité normale multivariée, on peut écrire:

$$\log L(\beta, \sigma_\epsilon^2, \rho) = K + \frac{1}{2}\log\det V^{-1} - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta)$$

et, en utilisant les règles de dérivation matricielle de la section 3.4:

$$\frac{\partial \log L}{\partial \beta} = -X'V^{-1}X\beta + X'V^{-1}y$$

En annulant ce vecteur de dérivées, on obtient:

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

soit le même résultat qu'en moindres carrés généralisés lorsque Ω est connue.

Les dérivées par rapport à ρ et σ_ϵ^2 sont plus compliquées. Il serait superflu d'en donner les détails ici, puisque ces derniers ce trouvent dans l'article précédemment cité. Il nous suffira de mentionner que la maximisation de L par rapport à ρ implique la solution d'une équation cubique, qui possède toujours une solution comprise entre -1 et $+1$.

Le but de cette section étant d'illustrer les résultats du présent chapitre, nous allons énoncer la matrice d'information et son utilité dans le contexte de ce modèle. Appelons $\theta = (\beta, \sigma_\epsilon^2, \rho)$. Beach et MacKinnon montrent que, si X est non stochastique:

$$R(\theta) = \begin{pmatrix} (X'V^{-1}X) & O_{k\times 1} & O_{k\times 1} \\ O_{1\times k} & A & C \\ O_{1\times k} & C & B \end{pmatrix}$$

où A, B, et C sont des scalaires. Alors:

$$R^{-1}(\theta) = \begin{pmatrix} (X'V^{-1}X)^{-1} & O_{k\times 2} \\ O_{2\times k} & \begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1} \end{pmatrix}$$

et le théorème vu à la section 10.10 implique:

$$\operatorname{dlim} \sqrt{n}(\hat{\theta} - \theta) \sim N(0, \operatorname{plim} nR^{-1}(\theta)).$$

Comme tout sous-vecteur d'un vecteur normal multivarié est normal multivarié, on peut donc écrire:

$$\operatorname{dlim} \sqrt{n}(\hat{\beta} - \beta) \sim N(0, \operatorname{plim} n(X'V^{-1}X)^{-1}).$$

Nous avons vu que les estimateurs par maximum de vraisemblance sont convergents, et que les limites en probabilité sont préservées par les fonctions continues. Donc, si on remplace, dans la définition de V, ρ et σ_ϵ^2 par leurs estimateurs pour obtenir $\hat{V}$, on obtient:

$$\operatorname{plim} \hat{V} = V$$

$$\operatorname{plim} n(X'\hat{V}^{-1}X)^{-1} = \operatorname{plim} n(X'V^{-1}X)^{-1}$$

et par conséquent:

$$\operatorname{dlim} \sqrt{n}(\hat{\beta} - \beta) \sim N(0, \operatorname{plim} n(X'\hat{V}^{-1}X)^{-1}).$$

On peut donc approcher la distribution de $\hat{\beta}$ par une normale $N(\beta, (X'\hat{V}^{-1}X)^{-1})$.

Pour tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$, on peut calculer le rapport des vraisemblances λ en estimant deux fois le modèle: une fois par MCO (ceci donne l'estimation sous H_0) et une fois par la méthode de Beach et MacKinnon (ceci donne l'estimation sans contrainte). λ est le rapport des vraisemblances maximisées. Le théorème de la section 10.11 implique alors que $\operatorname{dlim} -2\log\lambda \sim \chi^2_{(1)}$ lorsque H_0 est vraie, puisqu'il n'y a qu'une seule contrainte sous H_0. Ceci fournit des valeurs critiques approximatives. Ce test n'est valable qu'en grand échantillon mais ne présente pas les zones d'incertitude de la statistique de Durbin-Watson.

Il faut bien noter que les résultats du chapitre X sont d'une applicabilité très générale; cette section n'a présenté qu'une illustration de ces résultats.

CHAPITRE XI.

PROPRIÉTÉS ASYMPTOTIQUES DES ESTIMATEURS DE MOINDRES CARRÉS ORDINAIRES

11.1 Convergence en probabilité

Nous montrerons dans cette section que $\hat{\beta} = (X^{'}X)^{-1}X^{'}y$ est un estimateur convergent de β dans le modèle classique $y = X\beta + u$, sous les hypothèses suivantes:

(H1) $$E(u) = 0$$

(H2) $$E(uu^{'}) = \sigma^2 I$$

(H3) X est non stochastique de rang $k < n$

(H4) $$\lim_{n \longrightarrow \infty} \frac{1}{n} X^{'}X = \Sigma_{XX}, \quad \text{une matrice définie positive.}$$

Comme X est non stochastique, on a:

$$\begin{aligned} E(X^{'}u) &= X^{'}E(u) = 0 \\ V\left[(X^{'}u)_i\right] &= V\left[\sum_{t=1}^{n} X_{ti}u_t\right] = \sigma^2 \sum_{t=1}^{n} X_{ti}^2 \\ \text{et} \quad V\left[\frac{1}{n}(X^{'}u)_i\right] &= \sigma^2 \frac{\sum X_{ti}^2}{n^2} \quad . \end{aligned}$$

Mais $\frac{\sum X_{ti}^2}{n}$ converge par l'hypothèse (H4) vers un nombre fini. Nous concluons que $V\left[\frac{1}{n}(X^{'}u)_i\right]$ tend vers zéro quand n tend vers l'infini. Donc les composantes de $\frac{1}{n}X^{'}u$

vérifient $E\left[\frac{1}{n}(X^{'}u)_i\right] = 0$, et $\lim_{n\longrightarrow\infty} V\left[\frac{1}{n}(X^{'}u)_i\right] = 0$. Ceci montre (section 10.4) que $\text{plim}(\frac{1}{n}X^{'}u) = 0$. On a alors, en appliquant le théorème de Slutsky:

$$\begin{aligned}
\text{plim}\,\hat{\beta} &= \text{plim}\left[\beta + (X^{'}X)^{-1}X^{'}u\right] \\
&= \beta + \text{plim}\left[(X^{'}X)^{-1}X^{'}u\right] \\
&= \beta + \text{plim}\left[\left(\frac{1}{n}X^{'}X\right)^{-1}\frac{1}{n}X^{'}u\right] \\
&= \beta + \text{plim}\left[\frac{1}{n}(X^{'}X)\right]^{-1}\text{plim}\left(\frac{1}{n}X^{'}u\right) \\
&= \beta + \Sigma_{XX}^{-1}.O_{k\times 1} = \beta \quad .
\end{aligned}$$

11.2 Normalité asymptotique

Tous les tests d'hypothèses exposés au chapitre VII l'ont été en supposant la normalité des erreurs. Qu'en est-il si l'on ne fait pas d'hypothèses spécifiques sur la distribution du vecteur u? Nous allons voir qu'un théorème central limite nous permet d'établir la normalité asymptotique de $\hat{\beta} = (X^{'}X)^{-1}X^{'}y$. Si la taille de l'échantillon est suffisamment grande, on peut alors se baser sur la distribution normale pour faire des tests asymptotiques sur le vecteur β. On raisonne en pratique comme si la variance des erreurs était connue: on utilisera donc la loi normale au lieu de la loi de Student, la loi χ^2 au lieu de la loi F.

Théorème.

Supposons que les hypothèses (H1) à (H4) soient vérifiées, et soit α_t la t-ième colonne de la matrice $X^{'}$. Définissons les vecteurs $Z_t = u_t\alpha_t$ et supposons que $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_t$ vérifie un théorème central limite. Alors, pour $\hat{\beta} = (X^{'}X)^{-1}X^{'}y$:

(a) $\text{dlim}\,\sqrt{n}(\hat{\beta} - \beta) \sim N(0, \sigma^2\Sigma_{XX}^{-1})$.
(b) *Si* $\text{plim}\,(\frac{1}{n}\,u^{'}u) = \sigma^2$, *on a* $\text{plim}(\frac{1}{n}\hat{u}^{'}\hat{u}) = \sigma^2$ *avec* $\hat{u} = y - X\hat{\beta}$.

Démonstration

(a) Notons d'abord que $E(Z_t) = 0$ et $V(Z_t) = \sigma^2\alpha_t\alpha_t^{'}$.
Par conséquent:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}V(Z_t) = \lim_{n\to\infty}\frac{\sigma^2}{n}\sum_{t=1}^{n}\alpha_t\alpha_t' = \lim_{n\to\infty}\frac{\sigma^2}{n}X^{'}X = \sigma^2\Sigma_{XX},$$

qui est finie et définie positive par l'hypothèse (H4). En vertu du théorème central limite, on a:

$$\text{dlim}\,\frac{1}{\sqrt{n}}X^{'}u = \text{dlim}\,\frac{1}{\sqrt{n}}\sum_{t=1}^{n}Z_t \;\sim\; N(0,\sigma^2\Sigma_{XX}).$$

Notons ensuite que $\sqrt{n}(\hat{\beta}-\beta) = (\frac{1}{n}X^{'}X)^{-1}\frac{1}{\sqrt{n}}X^{'}u$ et appliquons les résultats de la section 10.6. Ceci donne:

$$\begin{aligned}
\text{dlim}\,\sqrt{n}(\hat{\beta}-\beta) &= \text{plim}\left(\frac{1}{n}X^{'}X\right)^{-1}\text{dlim}\left(\frac{1}{\sqrt{n}}X^{'}u\right)\\
&\sim N\left(0,\Sigma_{XX}^{-1}\left(\sigma^2\Sigma_{XX}\right)\Sigma_{XX}^{-1}\right)\\
&\sim N\left(0,\sigma^2\Sigma_{XX}^{-1}\right) \quad .
\end{aligned}$$

(b) Pour démontrer la seconde partie du théorème, rappelons que:

$$\hat{u}^{'}\hat{u} = u^{'}\left[I - X(X^{'}X)^{-1}X^{'}\right]u.$$

Donc:

$$\frac{\hat{u}^{'}\hat{u}}{n} = \frac{1}{n}u^{'}u - \left(\frac{1}{n}X^{'}u\right)^{'}\left(\frac{1}{n}X^{'}X\right)^{-1}\left(\frac{1}{n}X^{'}u\right), \quad \text{et:}$$

$$\text{plim}\left(\frac{\hat{u}^{'}\hat{u}}{n}\right) = \text{plim}\left(\frac{1}{n}u^{'}u\right) - O_{1\times k}.\Sigma_{XX}^{-1}.O_{k\times 1} = \sigma^2$$

en vertu du théorème de Slutsky et de l'hypothèse faite dans l'énoncé.

Exercice. Calculez la distribution limite, sous l'hypothèse nulle $H_0 : R\beta = r$, de la statistique de Wald vue à la section 7.4 de la seconde partie.

CHAPITRE XII.

PROPRIÉTÉS ASYMPTOTIQUES DES ESTIMATEURS D'AITKEN

Le théorème que nous allons démontrer dans ce chapitre est un cas particulier d'application au modèle à erreurs autorégressives d'un théorème plus général, s'appliquant à tout estimateur "Aitken-réalisable". Il montre que si l'on remplace Ω par un estimateur convergent de cette matrice dans la formule de $\hat{\beta}_{mcg}$, on obtient un estimateur de β qui a la même distribution limite que $\hat{\beta}_{mcg}$.

Théorème.

Soit le modèle $y = X\beta + u$ *avec* $E(u) = 0$,

$$E(uu') = \sigma^2\Omega = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & & \\ \vdots & & \ddots & \vdots \\ \rho^{n-1} & & \cdots & 1 \end{pmatrix} \quad , \quad \textit{et } X \textit{ non stochastique.}$$

Soit $\hat{\rho}$ *un estimateur convergent de* ρ *et supposons que* $\lim_{n\longrightarrow\infty} \frac{1}{n}(X'\Omega^{-1}X) = Q$ *soit une matrice définie positive. Soit* T *la matrice de transformation de la section 9.3* ($T'T = \Omega^{-1}$), *soit* $[X'T']_t$ *la t-ième colonne de* $X'T'$, *et supposons que les vecteurs* $Z_t = (Tu)_t[X'T']_t$ *vérifient un théorème central limite.*

Considérons les deux estimateurs:

$$\begin{aligned} \hat{\beta} &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \quad , \quad et \\ \hat{\hat{\beta}} &= (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y \end{aligned}$$

$$o\grave{u} \quad \hat{\Omega} = \begin{pmatrix} 1 & \hat{\rho} & \cdots & \hat{\rho}^{n-1} \\ \hat{\rho} & 1 & & \\ \vdots & & \ddots & \vdots \\ \hat{\rho}^{n-1} & & \cdots & 1 \end{pmatrix}$$

Sous les hypothèses additionnelles que:

$$\operatorname{plim} \frac{1}{n}(X'\hat{\Omega}^{-1}X) = \lim \frac{1}{n}(X'\Omega^{-1}X) = Q$$

$$\operatorname{plim} \frac{1}{\sqrt{n}}(X'\hat{\Omega}^{-1}u - X'\Omega^{-1}u) = 0$$

$$\operatorname{plim} \frac{1}{n}u'u = \sigma^2$$

on a les résultats suivants:

(1) $\operatorname{dlim} \sqrt{n}(\hat{\beta} - \beta) = \operatorname{dlim} \sqrt{n}(\hat{\hat{\beta}} - \beta) \sim N(0, \sigma^2 Q^{-1})$
(2) $\operatorname{plim} s^2 = \sigma^2$, *avec:*

$$s^2 = \frac{1}{n-k}(y - X\hat{\hat{\beta}})'\hat{\Omega}^{-1}(y - X\hat{\hat{\beta}}).$$

Démonstration

Notons tout d'abord que $\sqrt{n}(\hat{\beta} - \beta) = (\frac{1}{n}X'\Omega^{-1}X)^{-1}\frac{1}{\sqrt{n}}X'\Omega^{-1}u$ et que:

$$X'\Omega^{-1}u = X'T'Tu = \sum_{t=1}^{n} Z_t.$$

On a $E(Z_t) = 0$; d'autre part, comme $E(Tu)_t^2 = \sigma^2$ et comme $\sum_{t=1}^{n}[X'T']_t[X'T']_t' = X'\Omega^{-1}X$,

$$\lim \frac{1}{n}\sum_{t=1}^{n} E(Z_t Z_t') = \lim \frac{\sigma^2}{n}(X'\Omega^{-1}X) = \sigma^2 Q.$$

Par conséquent, en vertu du théorème central limite, $\operatorname{dlim} \frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_t \sim N(0, \sigma^2 Q)$.

Donc:

$$\operatorname{dlim} \sqrt{n}\left(\hat{\beta} - \beta\right) = \operatorname{plim}\left(\frac{1}{n}X'\Omega^{-1}X\right)^{-1} \operatorname{dlim}\left(\frac{1}{\sqrt{n}}X'\Omega^{-1}u\right)$$
$$\sim N\left(0, Q^{-1}(\sigma^2 Q)Q^{-1}\right) = N\left(0, \sigma^2 Q^{-1}\right) \quad .$$

Pour montrer que l'estimateur "Aitken-réalisable" a la même distribution que l'estimateur "Aitken-pur", nous pouvons appliquer le résultat de la section 10.6.1. En effet:

$$\sqrt{n}(\hat{\hat{\beta}} - \beta) = (\frac{1}{n}X'\hat{\Omega}^{-1}X)^{-1}\frac{1}{\sqrt{n}}X'\hat{\Omega}^{-1}u,$$

$$\operatorname{dlim}(\frac{1}{\sqrt{n}}X'\hat{\Omega}^{-1}u) \sim N(0, \sigma^2 Q)$$

et donc:

$$\operatorname{dlim}\sqrt{n}(\hat{\hat{\beta}} - \beta) \sim N(0, \sigma^2 Q^{-1}).$$

Pour démontrer la seconde partie du théorème, notons que la limite en probabilité de:

$$\frac{1}{n-k}(y - X\hat{\beta})'\Omega^{-1}(y - X\hat{\beta})),$$

est égale à σ^2. La démonstration est exactement la même que celle de la section 11.2: il suffit de remplacer y par Ty et X par TX. Comme $\operatorname{plim}\hat{\hat{\beta}} = \operatorname{plim}\hat{\beta} = \beta$ et comme $\operatorname{plim}\hat{\Omega} = \Omega$, le théorème de Slutsky implique $\operatorname{plim} s^2 = \sigma^2$.

Ces résultats ont été obtenus sans faire l'hypothèse de normalité des erreurs, puisque nous avons utilisé un théorème central limite. Il est toutefois très intéressant de noter que nous venons d'obtenir la **même** distribution limite que celle de la section 10.13, où nous avions fait l'hypothèse de normalité des erreurs pour dériver l'estimateur par maximum de vraisemblance; il est facile en effet de vérifier que:

$$\sigma^2 Q^{-1} = \operatorname{plim} n(X'V^{-1}X)^{-1}$$

où $V = E(uu')$. Les matrices de covariance asymptotiques sont donc les mêmes; puisqu'une distribution normale est entièrement caractérisée par les deux premiers moments, ceci implique bien l'égalité des distributions limites.

Nous avons donc l'équivalence asymptotique d'une méthode simple (celle d'Aitken) et d'une méthode plus compliquée (celle du maximum de vraisemblance).

CHAPITRE XIII.

RÉGRESSEURS STOCHASTIQUES

13.1 Introduction: types de régresseurs stochastiques

Dans tous les développements précédents, X était non stochastique par hypothèse. Ceci n'étant pas réaliste, il nous faut maintenant examiner les propriétés de la méthode des moindres carrés ordinaires dans le cas où cette hypothèse n'est pas vérifiée.

Nous pourrons distinguer trois types de régresseurs stochastiques.

Dans le premier cas, la matrice X est indépendante du vecteur u. Les estimateurs MCO sont alors convergents, sans biais, et ont la distribution limite vue au chapitre XI sous l'hypothèse d'un théorème central limite. De plus, lorsque les erreurs sont normales, les statistiques t_{obs} et F_{obs} vues précédemment au chapitre VII ont les distributions t et F sous l'hypothèse nulle, même en petit échantillon.

Dans le second cas, X dépend de u, mais les régresseurs ne sont pas corrélés avec les erreurs **contemporaines**. Les estimateurs MCO ne sont pas sans biais, mais ils sont convergents. Ils ont la distribution limite vue au chapitre XI sous l'hypothèse d'un théorème central limite. Les distributions des statistiques t_{obs} et F_{obs} vues précédemment au chapitre VII ne sont t et F que si la taille de l'échantillon tend vers l'infini. Nous n'examinerons pas ce second cas dans le présent chapitre, mais nous l'étudierons plus tard dans le cadre des modèles à variables endogènes retardées.

Dans le troisième cas, certains régresseurs sont corrélés avec l'erreur contemporaine. Alors les estimateurs MCO ne sont pas convergents, et on doit utiliser la méthode des variables instrumentales, qui sera vue dans le présent chapitre.

13.2 Régresseurs stochastiques indépendants du vecteur des erreurs

Nous allons voir que si X est stochastique, mais indépendante de u, l'estimateur de β par moindres carrés ordinaires garde beaucoup de propriétés désirables. Il est toujours sans biais, et convergent. De plus, toutes les propriétés asymptotiques démontrées précédemment dans le cadre du modèle classique restent valides.

Dans la première partie de cette section, nous n'utiliserons que les hypothèses suivantes, qui sont compatibles avec l'indépendance de X et de u, mais n'impliquent pas cette indépendance:

(H_1) $E(u|X) = 0$

(H_2) $E(uu^{'}|X) = \sigma^2 I$

(H_3) $\text{plim}(\frac{1}{n}u^{'}u) = \sigma^2$

(H_4) $\text{plim}(\frac{1}{n}X^{'}X) = \lim E(\frac{1}{n}X^{'}X) = \sum_{XX}$ est définie positive .

Rappelons tout d'abord la loi des espérances itérées (section 1.7) de la première partie:

Lemme 13.1.

$$E(X) = E_Y E(X|Y) \quad .$$

Ce résultat peut aussi être appliqué aux vecteurs et matrices aléatoires. Nous démontrons maintenant une propriété fondamentale pour la suite.

Lemme 13.2. *Sous les hypothèses* (H_1), (H_2) *et* (H_4), $\text{plim}(\frac{1}{n}X^{'}u) = 0$.

Démonstration:

En vertu de la section 10.4, il suffit de montrer que:

$$E(\frac{1}{n}\sum X_{ti}u_t) = 0 \quad \text{et} \quad V(\frac{1}{n}\sum X_{ti}u_t) \longrightarrow 0.$$

Mais:

$$E(X_{ti}u_t) = E_{X_{ti}}E(X_{ti}u_t|X_{ti}) = E_{X_{ti}}X_{ti}E(u_t|X_{ti}) = 0$$

par l'hypothèse (H_1) et le lemme 13.1. Par ailleurs:

$$V(X_{ti}u_t) = E(X_{ti}^2u_t^2) = E_{X_{ti}}E(X_{ti}^2u_t^2|X_{ti}) = E_{X_{ti}}X_{ti}^2E(u_t^2|X_{ti}) = \sigma^2E(X_{ti}^2)$$

en vertu de l'hypothèse (H_2). L'hypothèse (H_4) garantit que $E(X_{ti}^2) < \infty$; donc $V(X_{ti}u_t) < \infty$, et $V(\frac{1}{n}\sum X_{ti}u_t) \longrightarrow 0$. L'estimateur $\hat{\beta} = (X^{'}X)^{-1}X^{'}y$ vérifie alors les propriétés suivantes:

Théorème 13.3. *$\hat{\beta}$ est un estimateur sans biais de β.*

Démonstration:

$$\begin{aligned} E(\hat{\beta}) &= \beta + E\left[(X'X)^{-1}X'u\right] \\ &= \beta + E_X\{E\left[(X'X)^{-1}X'u|X\right]\} \\ &= \beta + E_X\left[(X'X)^{-1}X'\right]E(u|X) = \beta \quad . \end{aligned}$$

Théorème 13.4. *$\hat{\beta}$ est un estimateur convergent de β.*

La démonstration est identique à celle donnée à la section 11.1, en vertu du lemme 13.2.

Théorème 13.5. *Soit α_t la t-ième colonne de la matrice X' (un vecteur $k \times 1$) et supposons que les vecteurs $C_t = u_t\alpha_t$ vérifient un théorème central limite. Alors:*

(1) $\operatorname{dlim}\sqrt{n}(\hat{\beta}-\beta) \sim N(0, \sigma^2 \sum_{XX}^{-1})$
(2) $\operatorname{plim}(\frac{1}{n}\hat{u}'\hat{u}) = \sigma^2$.

Démonstration :

Notons que:

$$E(C_t) = E(u_t\alpha_t) = E_X E(u_t\alpha_t|X) = E_X\alpha_t E(u_t|X) = 0.$$

De même:

$$E(C_tC_t') = E(u_t^2\alpha_t\alpha_t') = E_X E(u_t^2\alpha_t\alpha_t'|X) = E_X(\alpha_t\alpha_t')E(u_t^2|X) = \sigma^2 E(\alpha_t\alpha_t').$$

Par conséquent:

$$\lim\frac{1}{n}\sum_{t=1}^{n}E(C_tC_t') = \sigma^2\lim E(\frac{1}{n}\sum_{t=1}^{n}\alpha_t\alpha_t') = \sigma^2\lim E(\frac{1}{n}X'X) = \sigma^2\Sigma_{XX}.$$

On a alors, comme auparavant (section 11.2):

$$\operatorname{dlim}\frac{1}{\sqrt{n}}X'u = \operatorname{dlim}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}C_t \sim N(0, \sigma^2\Sigma_{XX})$$

$$\operatorname{dlim}\sqrt{n}(\hat{\beta}-\beta) = \operatorname{plim}(\frac{1}{n}X'X)^{-1}\operatorname{dlim}(\frac{1}{\sqrt{n}}X'u) \sim N(0, \sigma^2\Sigma_{XX}^{-1}).$$

La démonstration du point (2) est identique à celle donnée précédemment.

Si nous faisons maintenant l'hypothèse d'indépendance $f(X,u) = f_1(X)f_2(u)$, les distributions **conditionnelles à X** des statistiques t_{obs} et F_{obs} vues au chapitre VII ne dépendront que des nombres de degrés de liberté et seront donc les mêmes que les distributions inconditionnelles. Les valeurs critiques des lois t et F leur seront donc applicables quelle que soit la taille de l'échantillon, lorsque les erreurs sont normales.

13.3 Régresseurs stochastiques dépendants des erreurs contemporaines

Si $\text{plim}(\frac{1}{n}X^{'}u) \neq 0$, on vérifie aisément que $\text{plim}\,\hat{\beta}_{mco} \neq \beta$. Il est important de signaler que la présence d'une seule composante non nulle dans le vecteur $\text{plim}(\frac{1}{n}X^{'}u)$ peut rendre *toutes* les composantes de $\hat{\beta}_{mco}$ non convergentes. Supposons en effet que:

$$\text{plim}\left(\frac{1}{n}X^{'}u\right) = \begin{pmatrix} c \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{avec} \quad c \neq 0 \quad .$$

On a alors:

$$\text{plim}\,\hat{\beta}_{mco} = \beta + c\begin{pmatrix} s_1 \\ \vdots \\ s_k \end{pmatrix}$$

où les s_i sont les composantes de la première colonne de Σ_{XX}^{-1}. Comme, en général, aucun des s_i n'est nul, aucune composante de $\hat{\beta}_{mco}$ ne convergera vers la composante correspondante de β.

Exercice. Dans le modèle $y_t = by_{t-1} + u_t$ avec $u_t = \epsilon_t + \rho\epsilon_{t-1}$, supposons que les ϵ_t soient d'espérance nulle, de variance constante, et non corrélés entre eux. Montrez que la covariance entre y_{t-1} et u_t n'est pas nulle. Quelles sont les conséquences de cette constatation?

13.3.1 La méthode des variables instrumentales.

Cette méthode est un cas particulier de la méthode des moments généralisés (GMM); voir Hamilton, *Time Series Analysis*, 1994, chapitre 14.

Supposons que $\text{plim}(\frac{1}{n}X^{'}u) \neq 0$. Nous construisons alors une matrice Z de dimensions $n \times r$, avec $r \geq k$, possédant les propriétés suivantes:

$$\begin{array}{lrcl}
(H_1) & E(u|Z) & = & 0 \\
(H_2) & E(uu^{'}|Z) & = & \sigma^2 I \\
(H_3) & \text{plim}(\frac{1}{n}Z^{'}X) & = & \Sigma_{ZX} \quad \text{est de rang } k \\
(H_4) & \text{plim}(\frac{1}{n}Z^{'}Z) & = & \lim E(\frac{1}{n}Z^{'}Z) = \Sigma_{ZZ} \quad \text{est définie positive.}
\end{array}$$

Nous supposons en outre comme auparavant que:

$$(H_5) \quad \text{plim}(\frac{1}{n}u^{'}u) = \sigma^2$$
$$(H_6) \quad \text{plim}(\frac{1}{n}X^{'}X) \quad \text{et} \quad \text{plim}(\frac{1}{n}X^{'}u) \quad \text{existent.}$$

L'idée de base est la suivante. Définissons $P_Z = Z(Z'Z)^{-1}Z'$; cette matrice $n \times n$ est symétrique, idempotente, de rang r. Si l'on applique la transformation P_Z au modèle $y = X\beta + u$ et les moindres carrés ordinaires au modèle transformé, on obtient l'estimateur de β par variables instrumentales:

$$\hat{\beta}_{VI} = (X'P_ZX)^{-1}X'P_Zy$$

Si l'on a le même nombre d'instruments et de régresseurs, $r = k$, et la matrice $X'Z$ est carrée et en général régulière. Alors:

$$\hat{\beta}_{VI} = \left[(X^{'}Z)(Z^{'}Z)^{-1}(Z^{'}X)\right]^{-1}(X^{'}Z)(Z^{'}Z)^{-1}Z^{'}y = \left(Z^{'}X\right)^{-1}Z^{'}y \quad .$$

Pour simplifier les démonstrations, nous supposerons dans le reste de cette section que $r = k$. Mais les résultats qui vont suivre ne dépendent pas de cette hypothèse.

13.3.2 Convergence en probabilité.

Lemme 13.6. *Sous les hypothèses* (H_1), (H_2) *et* (H_4), $\text{plim}(\frac{1}{n}Z^{'}u) = 0$.

La démonstration est identique à celle du Lemme 13.2.

Théorème 13.7. *$\hat{\beta}_{VI}$ est un estimateur convergent de β.*

Démonstration:
Comme $(Z^{'}X)^{-1}Z^{'}y = (Z^{'}X)^{-1}(Z^{'}X\beta + Z^{'}u) = \beta + (Z^{'}X)^{-1}Z^{'}u$, $\text{plim}\,\hat{\beta}_{VI} = \beta + \text{plim}(\frac{1}{n}Z^{'}X)^{-1}\,\text{plim}(\frac{1}{n}Z^{'}u) = \beta + \Sigma_{ZX}^{-1} \cdot 0 = \beta \quad .$

13.3.3 Convergence en distribution.

Théorème 13.8.

Soit α_t la t-ième colonne de $Z^{'}$ et supposons que les vecteurs $C_t = u_t\alpha_t$ vérifient un théorème central limite. Alors:

(1) $\text{dlim}\,\sqrt{n}(\hat{\beta}_{VI} - \beta) \sim N(0, \sigma^2\,\text{plim}\,n(X'P_ZX)^{-1}) = N(0, \sigma^2\Sigma_{ZX}^{-1}\Sigma_{ZZ}(\Sigma_{ZX}^{-1})^{'})$

(2) $\text{plim}(\frac{1}{n}\hat{u}'\hat{u}) = \sigma^2$, *avec* $\hat{u} = y - X\hat{\beta}_{VI} \quad .$

Démonstration:

Nous avons une fois de plus $E(C_t) = 0$ et $\lim \frac{1}{n} \sum_{t=1}^{n} E(C_t C_t^{'}) = \sigma^2 \Sigma_{ZZ}$ (voir la démonstration du théorème 13.5). Donc, comme $\frac{1}{\sqrt{n}} Z^{'} u = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} C_t$, on a:

$$\operatorname{dlim} \frac{1}{\sqrt{n}} Z^{'} u \sim N(0, \sigma^2 \Sigma_{ZZ})$$

et par conséquent:

$$\operatorname{dlim} \sqrt{n}(\hat{\beta}_{VI} - \beta) = \operatorname{plim}(\frac{1}{n} Z^{'} X)^{-1} \operatorname{dlim}(\frac{1}{\sqrt{n}} Z^{'} u) \sim N(0, \sigma^2 \Sigma_{ZX}^{-1} \Sigma_{ZZ} (\Sigma_{ZX}^{-1})^{'}).$$

Pour démontrer la seconde partie du théorème, notons que:

$$\hat{u} = y - X(Z^{'} X)^{-1} Z^{'} y = \left[I - X(Z^{'} X)^{-1} Z^{'} \right] u,$$

puisque $y = X\beta + u$. Alors:

$$\hat{u}^{'} \hat{u} = u^{'} u - u^{'} Z(X^{'} Z)^{-1} X^{'} u - u^{'} X(Z^{'} X)^{-1} Z^{'} u + u^{'} Z(X^{'} Z)^{-1} (X^{'} X)(Z^{'} X)^{-1} Z^{'} u.$$

Les hypothèses H_3, H_5 et H_6 ainsi que le Lemme 13.6 impliquent alors $\operatorname{plim}(\frac{1}{n} \hat{u}^{'} \hat{u}) = \operatorname{plim}(\frac{1}{n} u^{'} u) = \sigma^2$. Ce théorème permet donc, une fois de plus, de baser des tests asymptotiques sur la distribution normale ou χ^2. La matrice de covariance asymptotique du vecteur $\hat{\beta}_{VI}$ est estimée par $\frac{\hat{u}^{'} \hat{u}}{n} (Z^{'} X)^{-1} (Z^{'} Z)(X^{'} Z)^{-1}$.

Notons que si $r > k$, l'inverse de Σ_{ZX} n'existe pas car cette matrice n'est pas carrée; mais l'autre expression de la matrice de covariance asymptotique, à savoir:

$$\sigma^2 \operatorname{plim} n(X' P_Z X)^{-1}$$

reste valable, puisque $X' P_Z X$ est d'ordre k et de rang $\min(k, r) = k$. Par ailleurs, les deux expressions sont bien équivalentes lorsque $r = k$, puisque:

$$\operatorname{plim} n(X' P_Z X)^{-1} = \operatorname{plim} \left[(\frac{1}{n} X' Z)(\frac{1}{n} Z' Z)^{-1} (\frac{1}{n} Z' X) \right]^{-1}.$$

Notons enfin que la validité de la méthode des variables instrumentales peut être établie sous des hypothèses plus générales que celles de cette section.

13.3.4 Choix des variables instrumentales.

Il est très important de noter qu'il existe en général une infinité de matrices Z vérifiant les hypothèses (H_1) à (H_4). Il y aura donc aussi une infinité d'estimateurs par variables instrumentales! Cet estimateur garantit la convergence, mais ne vérifie pas le théorème de Gauss-Markov; et le choix des variables instrumentales doit donc être basé sur des critères d'efficacité asymptotique. On peut retenir, comme critère heuristique, celui qui fait choisir une variable instrumentale (colonne de Z) fortement corrélée avec la colonne correspondante de X, tout en satisfaisant $\text{plim}(\frac{1}{n}Z'u) = 0$. Nous utiliserons ce principe lorsque nous étudierons les variables endogènes retardées.

On peut aussi souvent choisir Z de telle manière que la distribution asymptotique du théorème 13.8 soit la même que celle de l'estimateur par maximum de vraisemblance. Ceci est intéressant car l'estimateur par variables instrumentales (qui est linéaire) est souvent plus facile à calculer que l'estimateur par maximum de vraisemblance (voir par exemple la section 10.13).

CHAPITRE XIV.

INTRODUCTION AUX MODÈLES DYNAMIQUES

14.1 Retards échelonnés

On a ici un modèle de la forme suivante:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + \ldots + b_k x_{t-k} + u_t$$

La variable dépendante est donc une combinaison linéaire des valeurs présentes et passées de la variable explicative. Nous fournirons deux interprétations économiques de ce modèle:

(a) Dans le cadre d'une fonction de consommation, il correspondrait à l'hypothèse que la consommation présente dépend du revenu espéré. Ce dernier est une combinaison linéaire des revenus observés, présents et passés. Il existe donc une sorte d'inertie dans le comportement du consommateur.

(b) Dans le cadre d'un modèle d'investissement, faisons les hypothèses suivantes:

(i) La valeur désirée des stocks, y_t^*, est proportionnelle à la valeur prévue des ventes, x_t^*, à un terme d'erreur v_t près. Donc:

$$y_t^* = \alpha x_t^* + v_t \quad . \tag{1}$$

(ii) L'investissement (variation de stock entre les périodes t et $t-1$) est régi par le mécanisme suivant (ajustement partiel):

$$y_t - y_{t-1} = \beta(y_t^* - y_{t-1}) \quad \text{avec} \quad 0 < \beta < 1 \quad . \tag{2}$$

On comble donc à la période t une fraction β de la différence entre le stock effectif précédent, y_{t-1}, et le stock désiré, y_t^* .

(iii) La valeur prévue des ventes est régie par le mécanisme suivant (anticipations adaptives):

$$x_t^* = x_{t-1}^* + \gamma(x_{t-1} - x_{t-1}^*) \quad \text{avec} \quad 0 < \gamma < 1 \quad . \tag{3}$$

On comble donc à la période t un pourcentage γ de l'erreur de prévision faite à la période $t-1$.

Nous allons montrer que les équations (1), (2) et (3) mènent à un modèle à retards échelonnés.
Résolvons tout d'abord l'équation de récurrence (3). Ceci donne:

$$\begin{aligned} x_t^* &= \gamma x_{t-1} + (1-\gamma)x_{t-1}^* \\ &= \gamma x_{t-1} + (1-\gamma)[\gamma x_{t-2} + (1-\gamma)x_{t-2}^*] \\ &= \gamma x_{t-1} + \gamma(1-\gamma)x_{t-2} + (1-\gamma)^2 x_{t-2}^* \end{aligned}$$

et l'on obtient, après une infinité de substitutions, la règle de prévision suivante, dite de "lissage exponentiel":

$$x_t^* = \gamma \sum_{i=1}^{\infty} (1-\gamma)^{i-1} x_{t-i} \quad . \tag{4}$$

Si nous résolvons maintenant (2) en y_t^*:

$$y_t^* = \frac{1}{\beta}\left[y_t - (1-\beta)y_{t-1}\right] \quad . \tag{5}$$

Par ailleurs, (1) et (4) impliquent

$$y_t^* = \alpha\gamma \sum_{j=1}^{\infty} (1-\gamma)^{i-1} x_{t-i} + v_t \quad . \tag{6}$$

En égalisant les membres de droite de (5) et de (6), on obtient finalement:

$$y_t = (1-\beta)y_{t-1} + \alpha\beta\gamma \sum_{i=1}^{\infty} (1-\gamma)^{i-1} x_{t-i} + u_t \quad . \tag{7}$$

Cette dernière équation est linéaire dans les variables explicatives, et ne comporte plus que des variables observables. Elle comporte néanmoins une infinité de régresseurs! On peut évidemment supprimer les x_{t-i} pour i grand. Mais ceci ne résout que partiellement le problème, car il y a peu de degrés de liberté: le nombre de paramètres à estimer reste grand, et l'on perd une observation par variable retardée. De plus, les x_{t-i} risquent d'être fortement colinéaires.

Les méthodes de Koyck et d'Almon on été proposées pour résoudre ce problème.

14.2 La méthode de Koyck

Soit donc le modèle général:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + \ldots + b_k x_{t-k} + u_t \quad .$$

On fait l'hypothèse que les poids b_i sont géométriquement décroissants, soit $b_i = \lambda^i b_0$ avec $0 < \lambda < 1$. Par conséquent:

$$\begin{aligned} y_t &= a + b_0 x_t + \lambda b_0 x_{t-1} + \lambda^2 b_0 x_{t-2} + \ldots + \lambda^k b_0 x_{t-k} + u_t \\ \text{et} \quad y_{t-1} &= a + b_0 x_{t-1} + \lambda b_0 x_{t-2} + \lambda^2 b_0 x_{t-3} + \ldots + \lambda^k b_0 x_{t-k-1} + u_{t-1} \\ \lambda y_{t-1} &= \lambda a + \lambda b_0 x_{t-1} + \lambda^2 b_0 x_{t-2} + \ldots + \lambda^{k+1} b_0 x_{t-k-1} + \lambda u_{t-1} \end{aligned}$$

que nous soustrayons pour obtenir:

$$y_t - \lambda y_{t-1} = (a - \lambda a) + b_0 x_t - \lambda^{k+1} b_0 x_{t-k-1} + (u_t - \lambda u_{t-1}) \quad .$$

Si k est suffisamment grand, $\lambda^{k+1} \approx 0$, et nous pouvons alors retenir comme modèle:

$$y_t = a^* + \lambda y_{t-1} + b_0 x_t + u_t^* \quad .$$

Nous n'avons donc plus que deux régresseurs et une constante. Il faut noter:

(a) que cette transformation peut aussi s'appliquer à un nombre infini de retards;

(b) que l'on peut retrouver l'équation de départ à partir d'estimations de λ et de b_0 obtenues grâce au modèle transformé;

(c) que $E(y_{t-1} u_t^*) \neq 0$. Nous sommes donc dans le cas traité à la section 13.3: les estimateurs par moindres carrés ordinaires ne sont pas convergents. Ce problème sera examiné plus bas, lorsque nous traiterons des variables endogènes retardées.

Appliquons la méthode de Koyck à notre problème d'investissement. Nous avions:

$$y_t = (1 - \beta) y_{t-1} + \alpha\beta\gamma \sum_{i=1}^{\infty} (1 - \gamma)^{i-1} x_{t-i} + u_t \quad .$$

Donc:

$$y_{t-1} = (1 - \beta) y_{t-2} + \alpha\beta\gamma \sum_{i=1}^{\infty} (1 - \gamma)^{i-1} x_{t-i-1} + u_{t-1}$$

et:

$$y_t - (1-\gamma)y_{t-1} = (1-\beta)y_{t-1} + \alpha\beta\gamma x_{t-1} - (1-\beta)(1-\gamma)y_{t-2} + [u_t - (1-\gamma)u_{t-1}],$$

soit aussi:

$$y_t = (2-\beta-\gamma)y_{t-1} + \alpha\beta\gamma x_{t-1} - (1-\beta)(1-\gamma)y_{t-2} + u_t^* \quad .$$

Appelons $\hat{a}_1$, $\hat{a}_2$, $\hat{a}_3$ les estimations des coefficients de cette équation. Pour estimer les paramètres du modèle de départ, il faudrait résoudre le système:

$$\begin{aligned} \hat{a}_1 &= 2 - \hat{\beta} - \hat{\gamma} \\ \hat{a}_2 &= \hat{\alpha}\hat{\beta}\hat{\gamma} \\ \hat{a}_3 &= -(1-\hat{\beta})(1-\hat{\gamma}) = \hat{\beta} + \hat{\gamma} - 1 - \hat{\beta}\hat{\gamma} \end{aligned}$$

$\hat{\alpha}$ peut être obtenu comme $\dfrac{\hat{a}_2}{1-\hat{a}_1-\hat{a}_3}$. Il est dit *identifiable.*

Mais $\hat{\beta}$ et $\hat{\gamma}$ ne le sont pas. On ne peut déterminer que leur somme et leur produit.

14.3 La méthode d'Almon

L'hypothèse faite par Koyck que les poids $b_0 \ldots b_k$ sont géométriquement décroissants est très restrictive. L'idée d'Almon est d'utiliser une approximation polynomiale de la fonction décrivant le comportement réel des b_i. On choisit, en pratique, un polynôme de degré supérieur d'au moins une unité au nombre de points stationnaires de cette fonction. Si, par exemple, l'on pense que cette fonction a la forme d'un U ou d'un U renversé, on choisira une approximation quadratique:

$$b_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2$$

que l'on substitue dans le modèle précédent:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + \ldots + b_k x_{t-k} + u_t$$

pour obtenir:

$$\begin{aligned} y_t = a + \alpha_0 x_t + (\alpha_0 + \alpha_1 + \alpha_2) x_{t-1} + (\alpha_0 + 2\alpha_1 + 4\alpha_2) x_{t-2} \\ + \ldots + (\alpha_0 + k\alpha_1 + k^2\alpha_2) x_{t-k} + u_t \end{aligned}$$

$$\begin{aligned} &= a + \alpha_0 \left(\sum_{i=0}^{k} x_{t-i} \right) + \alpha_1 \left(\sum_{i=1}^{k} i x_{t-i} \right) + \alpha_2 \left(\sum_{i=1}^{k} i^2 x_{t-i} \right) + u_t \\ &= a + \alpha_0 Z_{1t} + \alpha_1 Z_{2t} + \alpha_2 Z_{3t} + u_t \quad . \end{aligned}$$

Les paramètres de cette équation peuvent alors être estimés par moindres carrés ordinaires, et les estimations des b_i peuvent être calculées à l'aide de l'approximation polynomiale. Notons aussi que cette technique se prête particulièrement bien à l'introduction de contraintes additionnelles sur les b_i. Supposons que l'on veuille imposer $b_1 = 1$. On a donc $1 = \alpha_0 + \alpha_1 + \alpha_2$. En substituant, il vient:

$$y_t = a + (1 - \alpha_1 - \alpha_2) Z_{1t} + \alpha_1 Z_{2t} + \alpha_2 Z_{3t} + u_t$$

ou:

$$y_t - Z_{1t} = a + \alpha_1 (Z_{2t} - Z_{1t}) + \alpha_2 (Z_{3t} - Z_{1t}) + u_t \quad .$$

Soit:

$$y_t^* = a + \alpha_1 Z_{1t}^* + \alpha_2 Z_{2t}^* + u_t \quad .$$

14.4 L'opérateur de retard

L'opérateur de retard est défini par:

$$L x_t = x_{t-1} \quad .$$

Cet opérateur peut être traité comme une variable algébrique ordinaire. En effet:

$$\begin{aligned} L^j x_t &= L \ldots L x_t = x_{t-j} \\ L^j L^k x_t &= L^{j+k} x_t = x_{t-j-k} \\ L^j \left(a_1 x_{1t} + a_2 x_{2t} \right) &= a_1 L^j x_{1t} + a_2 L^j x_{2t} \end{aligned}$$

Nous pouvons alors écrire:

$$\sum_j \mu_j x_{t-j} = \sum_j \mu_j L^j x_t = \mu(L) x_t$$

où:

$$\mu(L) \underset{\text{def}}{=} \mu_0 + \mu_1 L + \mu_2 L^2 + \mu_3 L^3 + \ldots \quad .$$

est traité comme un polynôme algébrique en L. Si les racines de $\mu(L) = 0$ sont strictement supérieures à l'unité en valeur absolue, on peut définir l'opérateur réciproque $\mu^{-1}(L)$ comme:

$$y_t = \mu^{-1}(L)x_t \quad \text{si} \quad \mu(L)y_t = x_t \quad .$$

Exercice: Soit $\mu(L) = \mu_0 + \mu_1 L$, $\gamma(L) = \gamma_0 + \gamma_1 L$, et $\rho(L) = 1 - \rho L$ pour $-1 < \rho < 1$. Trouvez la forme des séries chronologiques $[\mu(L) + \gamma(L)]x_t$, $[\mu(L)\gamma(L)]x_t$ et $[\rho^{-1}(L)]x_t$.

L'intérêt de la recherche d'un tel opérateur réciproque peut être illustré par l'exemple suivant. L'équation:

$$y_t = a + \mu y_{t-1} + bx_t + u_t$$

peut s'écrire comme:

$$\mu(L)y_t = a + bx_t + u_t$$

avec $\mu(L) = 1 - \mu L$. Elle permet d'estimer l'espérance de y_t conditionnelle à ses valeurs passées et à x_t, à savoir $E(y_t \mid y_{t-1}, x_t) = a + \mu y_{t-1} + bx_t$. Il s'agit donc d'une modélisation à court terme, car conditionnelle au passé immédiat de y_t. Mais dans le cas où x_t est un instrument de politique économique, il peut être plus intéressant d'estimer:

$$E(y_t \mid x_t, x_{t-1}, x_{t-2}, \dots)$$

qui est conditionnelle aux seules valeurs présentes et passées de l'instrument. Cette nouvelle espérance peut être calculée à l'aide de l'opérateur réciproque, car:

$$\begin{aligned} E(y_t \mid x_t, x_{t-1}, x_{t-2}, \dots) &= \mu^{-1}(L)a + b\mu^{-1}(L)x_t \\ &= \frac{a}{1-\mu} + b\mu^{-1}(L)x_t \\ &= \frac{a}{1-\mu} + b(x_t + \mu x_{t-1} + \mu^2 x_{t-2} + \dots) \end{aligned}$$

Pour illustrer un autre emploi de l'opérateur de retard, appliquons-le à la transformation de Koyck. Nous avons:

$$\begin{aligned} y_t &= a + b_0 \sum_j \lambda^j x_{t-j} + u_t = a + b_0 \sum_j \lambda^j L^j x_t + u_t \\ &= a + b_0 \left(1 + \lambda L + \lambda^2 L^2 + \lambda^3 L^3 + \dots\right) x_t + u_t \\ &= a + \frac{b_0}{1 - \lambda L} x_t + u_t \quad , \end{aligned}$$

soit aussi:

$$(1 - \lambda L)y_t = (1 - \lambda L)a + b_0 x_t + (1 - \lambda L)u_t$$

et

$$y_t = \lambda y_{t-1} + a^* + b_0 x_t + (u_t - \lambda u_{t-1}) \quad .$$

14.5 Résolution d'équations linéaires de récurrence stochastiques

Présentons maintenant une méthode générale de résolution d'une équation du type $\mu(L)y_t = \gamma(L)u_t$, où u_t est une erreur aléatoire. Il s'agit de calculer les coefficients du polynôme $\dfrac{\gamma(L)}{\mu(L)}$. Nous commencerons par un exemple.

Soit $\gamma(L) = 2 + 3L + 4L^2$ et $\mu(L) = 1 - 0.75L + 0.125L^2$. Comme les racines de $\mu(L)$ sont 2 et 4, on a:

$$\mu(L) \quad = \quad \left(1 - \frac{L}{4}\right)\left(1 - \frac{L}{2}\right)$$

$$\frac{1}{\mu(L)} = \frac{1}{\left(1 - \frac{L}{4}\right)\left(1 - \frac{L}{2}\right)} = \frac{A\left(1 - \frac{L}{4}\right) + B\left(1 - \frac{L}{2}\right)}{\left(1 - \frac{L}{4}\right)\left(1 - \frac{L}{2}\right)}$$

où A et B sont déterminés par la condition $A\left(1 - \frac{L}{4}\right) + B\left(1 - \frac{L}{2}\right) = 1$ pour tout L. Ceci implique $A = 2$ et $B = -1$, comme on le voit facilement en posant $L = 0$ et $L = 1$. Par conséquent:

$$\begin{aligned}
\frac{1}{\mu(L)} &= \frac{2}{1 - \frac{L}{2}} - \frac{1}{1 - \frac{L}{4}} \\
&= 2\left[1 + \left(\frac{1}{2}L\right) + \left(\frac{1}{2}L\right)^2 + \ldots\right] - \left[1 + \left(\frac{1}{4}L\right) + \left(\frac{1}{4}L\right)^2 + \ldots\right] \\
&= 1 + \frac{3}{4}L + \frac{7}{16}L^2 + \frac{15}{64}L^3 \ldots \quad .
\end{aligned}$$

et donc:

$$\frac{\gamma(L)}{\mu(L)} = (2 + 3L + 4L^2)(1 + .75L + .4375L^2 + \ldots) = 2 + 4.5L + 7.125L^2 + \ldots \quad .$$

Ceci peut être facilement généralisé. Si le polynôme normalisé $\mu(L) = (1-\alpha L)(1-\beta L) = 0$ a deux racines réelles distinctes $1/\alpha$ et $1/\beta$, on aura:

$$\frac{1}{\mu(L)} = \frac{1}{(1 - \alpha L)(1 - \beta L)} = \frac{A(1 - \alpha L) + B(1 - \beta L)}{(1 - \alpha L)(1 - \beta L)}$$

où A et B sont choisis tels que $A(1-\alpha L)+B(1-\beta L)=1$ pour tout L. Ceci implique:

$$A=\frac{\beta}{\beta-\alpha}$$

$$B=\frac{-\alpha}{\beta-\alpha}$$

et donc:

$$\begin{aligned}\frac{1}{\mu(L)} &= \frac{A}{1-\beta L}+\frac{B}{1-\alpha L}\\ &= A(1+\beta L+\beta^2L^2+\dots)+B(1+\alpha L+\alpha^2L^2+\dots)\\ &= (A+B)+(\beta A+\alpha B)L+(\beta^2A+\alpha^2B)L^2+\dots\\ &= \frac{1}{\beta-\alpha}\sum_{i=1}^{\infty}(\beta^i-\alpha^i)L^{i-1}.\end{aligned}$$

Dans le cas d'une racine réelle double $1/\alpha$, on obtient:

$$\begin{aligned}\frac{1}{\mu(L)} &= \frac{1}{(1-\alpha L)^2}\\ &= (1+\alpha L+\alpha^2L^2+\dots)(1+\alpha L+\alpha^2L^2+\dots)\\ &= 1+2\alpha L+3\alpha^2L^2+4\alpha^3L^3+\dots\\ &= \sum_{i=0}^{\infty}(i+1)\alpha^iL^i\end{aligned}$$

Dans le cas de deux racines complexes conjuguées, on peut employer le premier développement en utilisant les propriétés des nombre complexes.

On peut aussi utiliser un développement de Taylor autour de $L=0$; la dérivation précédente a l'avantage d'être constructive, et de mettre en évidence le lien entre $1/\mu(L)$ et les racines de $\mu(L)=0$.

14.6 Distribution rationnelle des retards

Nous sommes maintenant prêts à définir la distribution rationnelle des retards. On l'écrit sous la forme:

$$y_t=a+\mu(L)x_t+u_t$$

avec:

$$\mu(L) = \frac{\gamma(L)}{w(L)} = \frac{\gamma_0 + \gamma_1 L + \ldots + \gamma_k L^k}{w_0 + w_1 L + \ldots + w_\ell L^\ell} \quad .$$

On normalise en posant $w_0 = 1$.

Cette formulation est très générale, car toute structure des coefficients peut être approchée par ce rapport de deux polynômes. Nous pouvons en effet rendre l'approximation plus fine en augmentant k, ℓ, ou k et ℓ.

On constate facilement que la structure des retards postulée par Almon correspond à $w(L) = 1$ (donc $\ell = 0$), et $\gamma_i = a_0 + a_1 i + a_2 i^2 + \ldots + a_s i^s$. Celle de Koyck correspond à $\gamma(L) = b_0$, et $w(L) = 1 - \lambda L$ (donc $k = 0, \ell = 1$).

14.7 Variables endogènes retardées

Lors de l'application de la transformation de Koyck, nous avons fait apparaître des variables endogènes retardées dans le membre de droite de l'équation de régression. Il est important de mettre en évidence les conséquences de leur présence parmi les variables explicatives d'un modèle.

Cette section n'étant qu'une introduction au problème, nous nous contenterons ici d'étudier un modèle très simple, qui est le suivant:

$$y_t = b y_{t-1} + u_t$$

avec $-1 < b < 1$ et diverses hypothèses sur l'erreur u_t.

Un modèle beaucoup plus général sera étudié au chapitre XV. Les conclusions obtenues dans ce modèle plus général, qui comprendra plusieurs régresseurs dont certains sont des variables endogènes retardées, sont très semblables et les méthodes d'analyse sont les mêmes.

On obtient aisément, par substitutions successives, la forme suivante:

$$y_t = u_t + b u_{t-1} + b^2 u_{t-2} + \cdots = \sum_{j=0}^{\infty} b^j u_{t-j} \quad .$$

14.7.1 Erreurs sphériques.

Supposons que $E(u) = 0$ et $E(uu') = \sigma^2 I$. On a alors $E(y_{t-1}u_t) = 0$, et si $V(y_{t-1}u_t)$ existe, on a $\text{plim}(\frac{1}{n}\sum y_{t-1}u_t) = 0$. L'estimateur de b par moindres carrés ordinaires est alors convergent. Mais il n'est pas sans biais puisque $\hat{b} = b + \sum_{t=2}^{n+1} w_t u_t$ avec

$$w_t = \frac{y_{t-1}}{\sum_{j=2}^{n+1} y_{j-1}^2}$$

qui dépend de u_t via le dénominateur.

La distribution limite de $\sqrt{n}(\hat{b}_{mco} - b)$ est la distribution normale habituelle:

$$\sqrt{n}(\hat{b}_{mco} - b) \underset{d}{\longrightarrow} N(0, \sigma^2 \Sigma_{XX}^{-1}) = N(0, \frac{\sigma^2}{\text{plim } \frac{1}{n} \sum_{t=2}^{n+1} y_{t-1}^2})$$

pour autant que les hypothèses de la section 10.8.3 soient vérifiées. En particulier, la suite $(Z_t) = (y_{t-1}u_t)$ doit être une différence de martingale. Tel est bien le cas ici sous l'hypothèse d'indépendance des erreurs. En effet:

$$E(y_{t-1}u_t) = E_{y_{t-1}} y_{t-1} E(u_t \mid y_{t-1}) = 0$$

$$\begin{aligned} E(y_{t-1}u_t \mid y_{t-2}u_{t-1}, y_{t-3}u_{t-2}, \dots) = \\ E_{u_{t-1},u_{t-2},\dots} E(y_{t-1}u_t \mid y_{t-2}u_{t-1}, y_{t-3}u_{t-2}, \dots; u_{t-1}, u_{t-2}, \dots) = \\ E_{u_{t-1},u_{t-2},\dots} \left[y_{t-1} E(u_t \mid y_{t-2}u_{t-1}, y_{t-3}u_{t-2}, \dots; u_{t-1}, u_{t-2}, \dots) \right] = 0 \end{aligned}$$

Il est facile de démontrer (voir Hamilton, *Time Series Analysis*, 1994, p. 122) que l'estimateur de b par maximum de vraisemblance est le même que l'estimateur de b par moindres carrés ordinaires lorsque les erreurs sont normales.

14.7.2 Erreurs à moyenne mobile.

Il s'agit d'erreurs de la forme:

$$u_t = \epsilon_t + \rho\epsilon_{t-1} \quad \text{avec} \quad E(\epsilon) = 0, \quad E(\epsilon\epsilon') = \sigma^2 I \quad .$$

Comme nous l'avons vu, ces erreurs résultent d'une transformation de Koyck appliquée à un modèle à retards échelonnés. On vérifie immédiatement que sous les hypothèses habituelles,

$$\begin{aligned} \text{plim} \left(\frac{1}{n} \sum y_{t-1}u_t \right) &= E\left(y_{t-1}u_t\right) = E\left[\left(u_{t-1} + bu_{t-2} + \dots\right) u_t\right] \\ &= E\left(u_t u_{t-1}\right) = E\left[\left(\epsilon_t + \rho\epsilon_{t-1}\right)\left(\epsilon_{t-1} + \rho\epsilon_{t-2}\right)\right] = \rho\sigma^2 \neq 0 \quad . \end{aligned}$$

Donc l'estimateur $\hat{b} = \frac{\sum y_t y_{t-1}}{\sum y_{t-1}^2}$ n'est pas convergent. Calculons sa limite en probabilité. Notons d'abord que $y_t = by_{t-1} + \epsilon_t + \rho\epsilon_{t-1}$, et donc:

$$\sum y_t y_{t-1} \quad = \quad b \sum y_{t-1}^2 + \sum y_{t-1}\epsilon_t + \rho \sum y_{t-1}\epsilon_{t-1}.$$

Par conséquent:

$$\hat{b} \quad = \quad b + \frac{\sum y_{t-1}\epsilon_t/n}{\sum y_{t-1}^2/n} + \rho\frac{\sum y_{t-1}\epsilon_{t-1}/n}{\sum y_{t-1}^2/n} \quad .$$

Par ailleurs, $y_t = \sum_{j=0}^{\infty} b^j(\epsilon_{t-j} + \rho\epsilon_{t-j-1})$, ce qui implique, sous les hypothèses habituelles, $\text{plim}(\frac{1}{n}\sum y_{t-1}\epsilon_t) = E(y_{t-1}\epsilon_t) = 0$, et $\text{plim}(\frac{1}{n}\sum y_{t-1}\epsilon_{t-1}) = E(y_{t-1}\epsilon_{t-1}) = \sigma^2$.

De même:

$$\text{plim}\left(\frac{1}{n}\sum y_{t-1}^2\right) = E(y_{t-1}^2) = E\left(y_t^2\right) = E\left[\sum_{j=0}^{\infty} b^{2j}\left(\epsilon_{t-j} + \rho\epsilon_{t-j-1}\right)^2\right]$$
$$+ 2E\left[\sum_{j=0}^{\infty}\sum_{k=j+1}^{\infty} b^j b^k\left(\epsilon_{t-j} + \rho\epsilon_{t-j-1}\right)\left(\epsilon_{t-k} + \rho\epsilon_{t-k-1}\right)\right]$$
$$= \left(1+\rho^2\right)\sigma^2\sum_{j=0}^{\infty} b^{2j} + 2\rho\sigma^2\sum_{j=0}^{\infty} b^j b^{j+1} = \frac{(1+\rho^2)\sigma^2}{1-b^2} + \frac{2b\rho\sigma^2}{1-b^2} = \frac{\sigma^2}{1-b^2}\left(1+\rho^2+2b\rho\right).$$

$$\text{Alors} \quad \text{plim}\,\hat{b} \quad = \quad b + \frac{\text{plim}(\frac{1}{n}\sum y_{t-1}\epsilon_t)}{\text{plim}(\frac{1}{n}\sum y_{t-1}^2)} + \rho\frac{\text{plim}(\frac{1}{n}\sum y_{t-1}\epsilon_{t-1})}{\text{plim}(\frac{1}{n}\sum y_{t-1}^2)}$$
$$= \quad b + \frac{\rho(1-b^2)}{1+\rho^2+2b\rho} \quad .$$

On remarque que $\text{plim}\,\hat{b} - b$ a le signe de ρ.

Montrons maintenant que l'on peut estimer b de façon convergente en utilisant y_{t-2} comme variable instrumentale. Il faut vérifier que:

$$\text{plim}\left(\frac{1}{n}Z'X\right) \quad = \quad \text{plim}\left(\frac{1}{n}\sum y_{t-1}y_{t-2}\right) \quad \text{est finie et non-nulle;}$$
$$\text{plim}\left(\frac{1}{n}Z'u\right) \quad = \quad \text{plim}\left(\frac{1}{n}\sum y_{t-2}u_t\right) = 0 \quad .$$

Tout d'abord:

$$\text{plim}\left(\frac{1}{n}\sum y_{t-1}y_{t-2}\right) = E\left(y_{t-1}y_{t-2}\right) = E\left(u_{t-1}u_{t-2}\right) + bE\left(y_{t-2}^2\right) =$$
$$\rho\sigma^2 + \frac{b}{1-b^2}\left[\sigma^2\left(1+\rho^2+2b\rho\right)\right] = \frac{\sigma^2}{1-b^2}(\rho+b)(1+b\rho)$$

est finie et non-nulle, sauf si $\rho = -b$ ou $\rho = -\frac{1}{b}$. Par ailleurs, $\text{plim}\left(\frac{1}{n}\sum y_{t-2}u_t\right) = E\left(y_{t-2}u_t\right) = 0$. Nous concluons que $\text{plim}\,\frac{\sum y_{t-2}y_t}{\sum y_{t-1}y_{t-2}} = \text{plim}\,\hat{b}_{VI} = b$.

Cette estimation par variables instrumentales ne résout pas le problème d'autocorrélation des erreurs, qui se pose puisque $E(u_t u_{t-1}) = \rho\sigma^2$. Ce problème peut être traité en utilisant une méthode robuste d'estimation de la variance de $\hat{b}_{VI}$, analogue à celle que nous avons introduite à la section 9.10; voir Hamilton, *Time Series Analysis*, 1994, chapitre 14.

Nous n'étudierons pas l'estimation de ce modèle par maximum de vraisemblance, car ceci relève d'un cours de matières spéciales. Il s'agit d'un cas particulier de modèle ARMA (Auto-Regressive Moving Average); ces modèles peuvent être estimés à l'aide de logiciels spécialisés.

14.7.3 Erreurs autorégressives.

Nous supposons cette fois que $u_t = \rho u_{t-1} + \epsilon_t$ avec $|\rho| < 1$, $\rho \neq \pm\frac{1}{b}$, et $E(\epsilon) = 0$, $E(\epsilon\epsilon') = \sigma^2 I$.

On a de nouveau:

$$\text{plim}\,\hat{b} = b + \frac{\text{plim}\left(\frac{1}{n}\sum y_{t-1}u_t\right)}{\text{plim}\left(\frac{1}{n}\sum y_{t-1}^2\right)}.$$

Rappelons que $E(u_t u_{t-s}) = \rho^s\sigma_u^2$. Nous avons cette fois:

$$E(y_{t-1}u_t) = E\left[(u_{t-1} + bu_{t-2} + \ldots)u_t\right] = \rho\sigma_u^2(1 + b\rho + b^2\rho^2 + \ldots) = \frac{\rho\sigma_u^2}{1-b\rho}.$$

On a aussi:

$$\begin{aligned}
E\left(y_{t-1}^2\right) = E\left(y_t^2\right) &= \sum_{j=0}^{\infty} b^{2j} E\left(u_{t-j}^2\right) + 2\sum_{j=0}^{\infty}\sum_{k=j+1}^{\infty} b^j b^k E\left(u_{t-j}u_{t-k}\right) \\
&= \frac{\sigma_u^2}{1-b^2} + 2\sigma_u^2 \sum_{j=0}^{\infty}\sum_{k=j+1}^{\infty} \rho^{k-j} b^j b^k \\
&= \frac{\sigma_u^2}{1-b^2} + 2\sigma_u^2 (b\rho) \sum_{j=0}^{\infty} b^{2j} \sum_{k=j}^{\infty} (\rho b)^{k-j} \\
&= \frac{\sigma_u^2}{1-b^2} + \frac{2b\rho\sigma_u^2}{(1-b^2)(1-b\rho)} = \frac{\sigma_u^2(1+b\rho)}{(1-b^2)(1-b\rho)} \quad .
\end{aligned}$$

Par conséquent:

$$\begin{aligned}
\text{plim}\,\hat{b} &= b + \frac{\rho\sigma_u^2/\left(1-b\rho\right)}{\sigma_u^2\left(1+b\rho\right)/\left(1-b^2\right)\left(1-b\rho\right)} \\
&= b + \frac{\rho\left(1-b^2\right)}{1+b\rho} \quad .
\end{aligned}$$

On remarque que $\operatorname{plim} \hat{b} - b$ a de nouveau le signe de ρ.

Nous allons maintenant étudier l'estimation de ce modèle par maximum de vraisemblance. En combinant les équations:

$$y_t = by_{t-1} + u_t$$

$$u_t = \rho u_{t-1} + \epsilon_t$$

on obtient:

$$y_t - by_{t-1} = \rho(y_{t-1} - by_{t-2}) + \epsilon_t$$

soit aussi:

$$y_t = (b+\rho)y_{t-1} - b\rho y_{t-2} + \epsilon_t \quad \text{pour } t = 3, \ldots, n+2 \tag{1}$$

Ce modèle est non linéaire dans les paramètres. Si nous supposons que, conditionnellement à y_{t-1} et y_{t-2}, les ϵ_t sont normales de distribution commune $N(0, \sigma^2)$, nous avons pour l'observation t:

$$f(y_t \mid y_{t-1}, y_{t-2}) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}[y_t - (b+\rho)y_{t-1} + b\rho y_{t-2}]^2\right]$$

et la densité de $(y_3, y_4, \ldots, y_{n+2})$ conditionnelle aux deux premières observations (y_1, y_2) peut donc s'écrire:

$$\begin{aligned} f(y_3, y_4, \ldots, y_{n+2} \mid y_1, y_2) &= f(y_3 \mid y_1, y_2) f(y_4 \mid y_1, y_2, y_3) \ldots f(y_{n+2} \mid y_1, y_2, \ldots, y_{n+1}) \\ &= f(y_3 \mid y_1, y_2) f(y_4 \mid y_2, y_3) \ldots f(y_{n+2} \mid y_{n+1}, y_n) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{t=3}^{n+2} [y_t - (b+\rho)y_{t-1} + b\rho y_{t-2}]^2\right] \end{aligned}$$

En prenant le logarithme de l'expression précédente et en considérant le résultat comme une fonction des paramètres inconnus (b, ρ, σ^2), on obtient la vraisemblance logarithmique:

$$\begin{aligned} \log L(b, \rho, \sigma^2) &= \text{constante} - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=3}^{n+2} [y_t - (b+\rho)y_{t-1} + b\rho y_{t-2}]^2 \\ &= \text{constante} + \sum_{t=3}^{n+2} L_t(b, \rho, \sigma^2) \end{aligned}$$

où:

$$L_t(b, \rho, \sigma^2) = -\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} [y_t - (b+\rho)y_{t-1} + b\rho y_{t-2}]^2.$$

On peut facilement vérifier que:

$$\frac{\partial L_t}{\partial b} = \frac{1}{\sigma^2}(y_{t-1} - \rho y_{t-2})\epsilon_t$$
$$\frac{\partial L_t}{\partial \rho} = \frac{1}{\sigma^2}(y_{t-1} - b y_{t-2})\epsilon_t$$
$$\frac{\partial L_t}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\epsilon_t^2$$

où:

$$\epsilon_t = y_t - (b+\rho)y_{t-1} + b\rho y_{t-2}.$$

Comme $\log L = k + \sum L_t$, ceci implique:

$$\frac{\partial \log L}{\partial b} = \frac{1}{\sigma^2}\sum_{t=3}^{n+2}(y_{t-1} - \rho y_{t-2})\epsilon_t$$
$$\frac{\partial \log L}{\partial \rho} = \frac{1}{\sigma^2}\sum_{t=3}^{n+2}(y_{t-1} - b y_{t-2})\epsilon_t$$
$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=3}^{n+2}\epsilon_t^2.$$

Pour annuler les deux premières dérivées de $\log L$, il suffit d'appliquer, de manière alternée, les moindres carrés ordinaires aux deux paramétrisations linéaires pouvant être tirées de l'équation (1), à savoir:

$$(y_t - \rho y_{t-1}) = b(y_{t-1} - \rho y_{t-2}) + \epsilon_t$$
$$(y_t - b y_{t-1}) = \rho(y_{t-1} - b y_{t-2}) + \epsilon_t$$

jusqu'à la convergence de la somme des carrés des résidus $\hat{\epsilon}_t$. On peut alors estimer σ^2 par:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=3}^{n+2}(y_t - (\hat{b}+\hat{\rho})y_{t-1} + \hat{b}\hat{\rho}y_{t-2})^2.$$

Afin de formuler les variances des estimateurs ainsi obtenus et d'énoncer un test d'autocorrélation des erreurs, nous allons tout d'abord calculer l'espérance et la matrice de covariance du vecteur:

$$\frac{\partial L_t}{\partial \theta} = \begin{pmatrix} \dfrac{\partial L_t}{\partial b} \\ \dfrac{\partial L_t}{\partial \rho} \\ \dfrac{\partial L_t}{\partial \sigma^2} \end{pmatrix}.$$

En vertu de la loi des espérances itérées, on a:

$$E\left[\frac{\partial L_t}{\partial \theta}\right] = E_{y_{t-1},y_{t-2}} E\left[\frac{\partial L_t}{\partial \theta} \;\middle|\; y_{t-1}, y_{t-2}\right] = 0$$

car l'espérance conditionnelle apparaissant dans cette expression est nulle.

De même, en utilisant la normalité conditionnelle de ϵ_t, on a $E(\epsilon_t^3 \mid y_{t-1}, y_{t-2}) = 0$ et $E(\epsilon_t^4 \mid y_{t-1}, y_{t-2}) = 3\sigma^4$; il est alors facile de vérifier que:

$$V\left[\frac{\partial L_t}{\partial \theta} \;\middle|\; y_{t-1}, y_{t-2}\right] = E\left[\frac{\partial L_t}{\partial \theta}\frac{\partial L_t}{\partial \theta'} \;\middle|\; y_{t-1}, y_{t-2}\right]$$
$$= \frac{1}{\sigma^2}\begin{pmatrix} (y_{t-1}-\rho y_{t-2})^2 & (y_{t-1}-\rho y_{t-2})(y_{t-1}-b y_{t-2}) & 0 \\ (y_{t-1}-\rho y_{t-2})(y_{t-1}-b y_{t-2}) & (y_{t-1}-b y_{t-2})^2 & 0 \\ 0 & 0 & \frac{1}{2\sigma^2} \end{pmatrix}$$

et donc, en vertu de la loi des espérances itérées:

$$V\left[\frac{\partial L_t}{\partial \theta}\right]$$
$$= \frac{1}{\sigma^2} E\left[\begin{pmatrix} (y_{t-1}-\rho y_{t-2})^2 & (y_{t-1}-\rho y_{t-2})(y_{t-1}-b y_{t-2}) & 0 \\ (y_{t-1}-\rho y_{t-2})(y_{t-1}-b y_{t-2}) & (y_{t-1}-b y_{t-2})^2 & 0 \\ 0 & 0 & \frac{1}{2\sigma^2} \end{pmatrix}\right].$$

On peut vérifier que les vecteurs $\partial L_t/\partial\theta$ ne sont pas corrélés entre eux. La moyenne de ces matrices est alors égale à $n^{-1}R(\theta)$, où $R(\theta) = V\left[\frac{\partial \log L}{\partial \theta}\right]$ est la matrice d'information introduite au chapitre X. Si une loi faible des grands nombres est applicable, on aura, par exemple:

$$\lim \frac{1}{n}\sum E(y_{t-1}-\rho y_{t-2})^2 = \operatorname{plim} \frac{1}{n}\sum (y_{t-1}-\rho y_{t-2})^2$$

et on peut alors estimer la matrice de covariance de $\sqrt{n}(\hat\theta - \theta)$ par l'inverse de:

$$\hat V_n =$$
$$\frac{1}{n\hat\sigma^2}\begin{pmatrix} \sum(y_{t-1}-\hat\rho y_{t-2})^2 & \sum(y_{t-1}-\hat\rho y_{t-2})(y_{t-1}-\hat b y_{t-2}) & 0 \\ \sum(y_{t-1}-\hat\rho y_{t-2})(y_{t-1}-\hat b y_{t-2}) & \sum(y_{t-1}-\hat b y_{t-2})^2 & 0 \\ 0 & 0 & \frac{n}{2\hat\sigma^2} \end{pmatrix}$$

puisque $\operatorname{plim} \hat{V}_n = \operatorname{plim} n^{-1} R(\theta)$, et donc $\operatorname{plim} nR^{-1}(\theta) = \operatorname{plim} \hat{V}_n^{-1}$.

De plus, la matrice $n\hat{V}_n$ est une estimation de la matrice de covariance de $\partial \log L/\partial\theta$. Ceci permet facilement d'appliquer le principe des multiplicateurs de Lagrange pour tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$.

L'emploi du critère LM est particulièrement indiqué ici. Comme nous l'avons vu, la statistique LM ne nécessite que l'estimation du modèle sous H_0. Dans le présent contexte, H_0 signifie absence d'autocorrélation; et dans ce cas, l'estimation du modèle par maximum de vraisemblance se réduit à l'emploi des moindres carrés ordinaires. En revanche, comme nous l'avons vu, l'estimation sous H_1 nécessite une procédure itérative, qui est donc plus compliquée.

Le multiplicateur de Lagrange μ associé à la contrainte H_0 lors de la maximisation de la vraisemblance est égal à $\frac{\partial \log L}{\partial \rho}$. On peut montrer (voir par exemple L.G. Godfrey, *Misspecification tests in econometrics: the Lagrange multiplier principle and other approaches*, Cambridge University Press, Cambridge 1988, pages 11 et 14) que la statistique LM prend ici la forme:

$$LM = \hat{\mu}_0' \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \left(\hat{V}_0 \left[\frac{\partial \log L}{\partial \theta} \right] \right)^{-1} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \hat{\mu}_0$$

où $\hat{\mu}_0$ est la valeur de $\frac{\partial \log L}{\partial \rho}$ évaluée aux estimations contraintes des paramètres et où $\hat{V}_0 \left[\frac{\partial \log L}{\partial \theta} \right]$ est l'estimation contrainte de la matrice de covariance de $\partial \log L/\partial\theta$. Comme l'estimation contrainte est identique à l'estimation par MCO, définissons alors:

$$\hat{u}_t = y_t - \hat{b}_{mco} y_{t-1}.$$

On vérifie aisément que:

$$\hat{\mu}_0 = \frac{1}{\hat{\sigma}_0^2} \sum \hat{u}_{t-1} \hat{u}_t$$

$$\hat{V}_0 \left[\frac{\partial \log L}{\partial \theta} \right] = \frac{1}{\hat{\sigma}_0^2} \begin{pmatrix} \sum y_{t-1}^2 & \sum y_{t-1} \hat{u}_{t-1} & 0 \\ \sum y_{t-1} \hat{u}_{t-1} & \sum \hat{u}_{t-1}^2 & 0 \\ 0 & 0 & \dfrac{n}{2\hat{\sigma}_0^2} \end{pmatrix}$$

et que, par conséquent:

$$LM = \frac{1}{\hat{\sigma}_0^2} \cdot \frac{(\sum \hat{u}_{t-1} \hat{u}_t)^2 (\sum y_{t-1}^2)}{\sum y_{t-1}^2 \sum \hat{u}_{t-1}^2 - (\sum y_{t-1} \hat{u}_{t-1})^2}.$$

Nous allons maintenant montrer que cette statistique est identique à la statistique de Breusch-Godfrey définie à la section 9.8.2. Dans le présent contexte, la statistique de Breusch-Godfrey est la statistique LM utilisée pour tester $H_0 : \rho = 0$ dans l'équation de régression auxiliaire:

$$y_t = by_{t-1} + \rho\hat{u}_{t-1} + \eta_t$$

où $\hat{u}_{t-1} = y_{t-1} - \hat{b}_{mco}y_{t-2}$.

Pour montrer ce résultat, notons que l'estimateur des coefficients de régression dans l'équation auxiliaire peut s'écrire:

$$\hat{\beta} = \begin{pmatrix} \hat{b} \\ \hat{\rho} \end{pmatrix} = \begin{pmatrix} \sum y_{t-1}^2 & \sum y_{t-1}\hat{u}_{t-1} \\ \sum y_{t-1}\hat{u}_{t-1} & \sum \hat{u}_{t-1}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_t y_{t-1} \\ \sum y_t \hat{u}_{t-1} \end{pmatrix} = (X'X)^{-1}X'y$$

et que la matrice des coefficients de la restriction $\rho = 0$ est égale à $R = (\,0 \quad 1\,)$. L'expression du multiplicateur de Lagrange démontrée à la section 6.1 prend alors la forme suivante:

$$\begin{aligned} \lambda = \hat{\lambda}_0 &= [R(X'X)^{-1}R']^{-1}(r - R\hat{\beta}) \\ &= \frac{1}{\sum y_{t-1}^2}\left[-(\sum y_{t-1}^2)(\sum y_t\hat{u}_{t-1}) + (\sum y_t y_{t-1})(\sum y_{t-1}\hat{u}_{t-1})\right] \\ &= -\sum y_t\hat{u}_{t-1} + \hat{b}_{mco}\sum y_{t-1}\hat{u}_{t-1} \\ &= -\sum (y_t - \hat{b}_{mco}y_{t-1})\hat{u}_{t-1} \\ &= -\sum \hat{u}_t\hat{u}_{t-1}. \end{aligned}$$

Par ailleurs, comme nous l'avons montré à la section 7.5:

$$\begin{aligned} \hat{V}_0^{-1}(\lambda) &= \frac{1}{\hat{\sigma}_0^2}[R(X'X)^{-1}R'] \\ &= \frac{1}{\hat{\sigma}_0^2} \cdot \frac{\sum y_{t-1}^2}{\sum y_{t-1}^2 \sum \hat{u}_{t-1}^2 - (\sum y_{t-1}\hat{u}_{t-1})^2}. \end{aligned}$$

On voit alors facilement que la statistique du test de $\rho = 0$ dans l'équation de régression auxiliaire, à savoir:

$$LM^* = \hat{\lambda}_0'\hat{V}_0^{-1}(\lambda)\hat{\lambda}_0$$

est bien égale à la statistique LM définie plus haut.

Pour terminer cette section, notons que ce modèle autorégressif à erreurs autorégressives est restrictif. En effet, l'équation (1) n'est qu'un cas particulier du modèle plus général suivant:

$$y_t = \alpha y_{t-1} + \beta y_{t-2} + \epsilon_t$$

avec $\alpha = b + \rho$ et $\beta = -b\rho$. Ces contraintes s'appellent restrictions de facteurs communs, et seront examinées au chapitre XV dans un cadre plus général. Elles sont implausibles. C'est pour cette raison que nous ne poursuivrons pas l'étude du modèle de cette section 14.7.3. La méthodologie que nous venons d'énoncer est néanmoins indispensable pour la justification du test de Breusch-Godfrey, que l'on doit employer dans ce cas-ci puisque le test de Durbin-Watson n'est pas applicable.

CHAPITRE XV

LE MODÈLE DYNAMIQUE GÉNÉRAL

15.1 Présentation et hypothèses

Dans ce chapitre, nous allons généraliser le modèle autorégressif de la section 14.7. Une généralisation dynamique naturelle du modèle de régression multiple consiste à remplacer les variables y_t et $x_{1t}, \dots, x_{kt}$ de ce modèle par des combinaisons linéaires de leurs retards, à savoir $\phi(L)y_t$ et $\gamma_1(L)x_{1t}, \dots, \gamma_k(L)x_{kt}$. On obtient alors:

$$\phi(L)y_t = a + \gamma_1(L)x_{1t} + \cdots + \gamma_k(L)x_{kt} + \epsilon_t$$

où $\phi(L)$ est un polynôme normalisé de degré p et $\gamma_i(L)$ est un polynôme de degré q_i:

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$$
$$\gamma_i(L) = \gamma_{0i} + \gamma_{1i} L + \cdots + \gamma_{q_i i} L^{q_i}.$$

Nous supposons que, conditionnellement aux variables explicatives de ce modèle, les erreurs ϵ_t sont normales et identiquement distribuées. Comme les variables explicatives forment le vecteur $z_t = (y_{t-1}, x_{1t}, \dots, x_{kt})$ et les retards de ce vecteur, nous avons:

$$E(\epsilon_t \mid z_t, z_{t-1}, \dots) = 0$$

$$E(\epsilon_t^2 \mid z_t, z_{t-1}, \dots) = \sigma^2.$$

Comme à la section 14.7, où nous avions supposé que $-1 < b < 1$, nous faisons aussi l'hypothèse que $\phi(L)$ est inversible (ses racines doivent être toutes strictement supérieures à l'unité en valeur absolue).

On désigne ce modèle par $AD(p, q_1, \dots, q_k)$.

Exemple:

Si $p = 1$, $k = 1$, et $q_1 = 1$, le modèle s'écrit:

$$y_t = \underbrace{\phi_1 y_{t-1} + a}_{\text{partie autorégressive}} + \underbrace{\gamma_{01} x_{1t} + \gamma_{11} x_{1,t-1}}_{\text{partie retards échelonnés}} + \epsilon_t.$$

Notes:

(1) Il ne faut pas confondre ce modèle avec le modèle ARMA(p, q), qui s'énonce comme:

$$\phi(L)y_t = \gamma(L)\epsilon_t$$

où $\phi(L)$ est de degré p, $\gamma(L)$ est de degré q, et les ϵ_t sont sphériques et inobservables. Les erreurs $u_t = \gamma(L)\epsilon_t$ du modèle ARMA suivent un processus à moyenne mobile, alors que celles du modèle AD sont sphériques.

(2) Contrairement au modèle ARMA, le modèle AD peut être estimé par MCO. Les tests habituels sont asymptotiquement valides (F pour l'ordre des retards, LM pour la sphéricité des erreurs). Le modèle AD présente donc une plus grande facilité d'emploi. Pour cette raison, beaucoup d'auteurs préconisent son utilisation.

(3) Insistons sur la généralité du modèle AD, qui inclut comme cas particuliers:

—le modèle statique si $p = q_1 = \cdots = q_k = 0$;

—le modèle autorégressif pur $\phi(L)y_t = a + \epsilon_t$ si $\gamma_i(L) = 0$ pour tout i ;

—le modèle statique à erreurs autorégressives:

$$y_t = a^* + \sum_{j=1}^{k} \beta_j x_{jt} + u_t, \quad \phi(L)u_t = \epsilon_t$$

sous des restrictions dites "de facteurs communs", comme nous le verrons plus bas.

15.2 Les restrictions de facteurs communs

Ces restrictions impliquent que les polynômes de retards échelonnés $\gamma_i(L)$ ont le facteur commun $\phi(L)$. Donc:

$$\gamma_i(L) = \phi(L)\beta_i(L).$$

Une forme particulière de ces restrictions, que nous allons examiner plus en détail, est la proportionnalité des polynômes de retards échelonnés au polynôme autorégressif; cette forme particulière est donc:

$$\gamma_i(L) = \phi(L)\beta_i$$

Alors le modèle AD s'écrit:

$$\phi(L)y_t = a + \phi(L)\beta_1 x_{1t} + \cdots + \phi(L)\beta_k x_{kt} + \epsilon_t$$

ce qui implique, en multipliant les deux membres par $\phi^{-1}(L)$:

$$y_t = a^* + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + u_t$$

où $a^* = \phi^{-1}(L)a = \phi^{-1}(1)a$ et $u_t = \phi^{-1}(L)\epsilon_t$, soit aussi $\phi(L)u_t = \epsilon_t$.

Exemple:

Si $p = k = q_1 = 1$, la restriction s'écrit:

$$\gamma_1(L) = \phi(L)\beta_1$$

soit aussi:

$$\gamma_{01} + \gamma_{11}L = (1 - \phi_1 L)\beta_1 = \beta_1 - \phi_1\beta_1 L.$$

En identifiant les coefficients de même degré, on obtient:

$$\begin{aligned}\beta_1 &= \gamma_{01}\\ \gamma_{11} &= -\phi_1\beta_1\end{aligned}$$

ce qui peut s'écrire:

$$\gamma_{11} + \phi_1\gamma_{01} = 0.$$

Cette restriction est non linéaire, mais peut être testée à l'aide d'une généralisation de la statistique de Wald (on utilise une approximation linéaire de la contrainte). Le test s'appelle test de facteurs communs (test COMFAC en abrégé).

Exercice:

En substituant la restriction précédente dans le modèle:

$$y_t = a + \phi_1 y_{t-1} + \gamma_{01}x_{1t} + \gamma_{11}x_{1,t-1} + \epsilon_t$$

montrez que l'on arrive à un modèle statique à erreurs autorégressives.

15.3 Le modèle AD et la relation d'équilibre stationnaire

Le modèle AD est un modèle statistique qui ne décrit que le comportement à court terme (c'est-à-dire conditionnel au passé immédiat) de y_t. Pour obtenir une relation économique intéressante, il faut obtenir la solution statique (ou solution à long terme, ou encore: relation d'équilibre stationnaire) du modèle. Une telle solution peut être obtenue facilement si l'on suppose que les espérances de y_t et des x_{jt} sont constantes:

$$E(y_t) = E(y) \quad \text{et } E(x_{jt}) = E(x_j).$$

Alors, en égalisant les espérances des deux membres de l'équation du modèle AD, on obtient:

$$\phi(1)E(y) = a + \sum_{j=1}^{k} \gamma_j(1)E(x_j)$$

et en résolvant, il vient:

$$E(y) = a^* + \sum_{j=1}^{k} \beta_j E(x_j)$$

où $a^* = \phi^{-1}(1)a$ et $\beta_j = \phi^{-1}(1)\gamma_j(1)$. Ceci est la relation entre les niveaux d'équilibre des variables, $E(y)$ et $E(x_j)$.

Commentaires:

(1) Ceci peut être généralisé au cas où une tendance linéaire est incluse dans la liste des x_{jt}.

(2) Si l'on impose les restrictions précédentes de facteurs communs $\gamma_j(L) = \phi(L)\beta_j$, on a vu que:

$$y_t = a^* + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + u_t.$$

On a donc, à partir de cette relation:

$$\beta_j = \frac{\partial y_t}{\partial x_{jt}}$$

mais aussi, en partant du modèle AD général:

$$\beta_j = \phi^{-1}(1)\gamma_j(1) = \frac{\partial E(y_t)}{\partial E(x_{jt})}.$$

Ceci implique donc l'égalité des coefficients à long terme et à court terme, et fait apparaître que les restrictions de facteurs communs sont assez implausibles.

Exercice: On donne le modèle autorégressif $y_t = 2 + 0.8y_{t-1} + \epsilon_t$, où les erreurs ϵ_t sont indépendantes et de distribution commune $N(0, 10^{-4})$. On demande de calculer l'espérance inconditionnelle $E(y_t)$, la variance inconditionnelle $V(y_t)$, et d'expliciter la relation d'équilibre stationnaire de ce modèle. Illustrez vos résultats en simulant y_t à partir du modèle précédent (ceci peut être fait à l'aide d'EXCEL ou d'un logiciel économétrique) et en interprétant le graphique chronologique et l'histogramme des réalisations simulées.

15.4 Le modèle AD et le modèle de correction d'erreur

Nous allons maintenant reparamétriser le modèle AD en utilisant une identité algébrique. Le modèle ainsi obtenu, qui porte le nom de modèle de correction d'erreur (ECM), aura pour intérêt de faire apparaître directement les coefficients de la relation d'équilibre stationnaire, à savoir les $\phi^{-1}(1)\gamma_j(1)$. Il est important de noter que le modèle de correction d'erreur est *équivalent* au modèle AD: en particulier, les résidus $\hat{\epsilon}_t$ obtenus par moindres carrés seront identiques dans les deux modèles. Néanmoins, le modèle ECM est non linéaire dans les paramètres, tandis que le modèle AD est linéaire. L'estimation du modèle ECM nécessite donc l'emploi de la méthode des moindres carrés non linéaires, qui est présente comme option dans la plupart des logiciels économétriques.

Commençons par énoncer, sous forme de lemme, l'identité algébrique mentionnée au début de cette section.

Lemme 15.1.
Si $A(L) = A_0 + A_1 L + A_2 L^2 + \cdots + A_n L^n$ *alors:*

$$A(L) = A(1)L + A^*(L)(1 - L)$$

où:

$$A^*(L) = \sum_{j=0}^{n-1} A_j^* L^j$$

avec $A_0^* = A_0$ *et* $A_j^* = -\sum_{s=j+1}^{n} A_s$ *pour* $j = 1, \ldots, n-1$ *et* $n > 1$.

Exercice:
Vérifiez le lemme 15.1 pour $n = 1, 2, 3, 4$.

Dérivation du modèle de correction d'erreur:

- On part du modèle AD:

$$\phi(L)y_t = a + \sum_{j=1}^{k} \gamma_j(L)x_{jt} + \epsilon_t$$

- On applique le lemme aux polynômes $\phi(L)$ et $\gamma_j(L)$

$$\phi(1)y_{t-1} + \phi^*(L)\Delta y_t = a + \sum_{j=1}^{k}[\gamma_j(1)x_{j,t-1} + \gamma_j^*(L)\Delta x_{jt}] + \epsilon_t$$

$$\phi^*(L)\Delta y_t = a - \phi(1)[y_{t-1} - \sum_{j=1}^{k} \phi^{-1}(1)\gamma_j(1)x_{j,t-1}] + \sum_{j=1}^{k} \gamma_j^*(L)\Delta x_{jt} + \epsilon_t$$

$$\phi^*(L)\Delta y_t = a - \phi(1)[y_{t-1} - \sum_{j=1}^{k} \beta_j x_{j,t-1}] + \sum_{j=1}^{k} \gamma_j^*(L)\Delta x_{jt} + \epsilon_t$$

Les β_j sont les coefficients de la relation d'équilibre.

15.5 Exemple économique

Supposons que $k = 1$, et $p = q_1 = 1$. Supposons de plus que:

$$\begin{aligned} y_t = &\text{ log de la consommation par tête à prix constants} \\ x_t = &\text{ log du revenu disponible par tête à prix constants} \end{aligned}$$

Le modèle:

$$\phi(L)y_t = a + \gamma(L)x_t + \epsilon_t$$

s'écrit alors comme:

$$y_t - \phi_1 y_{t-1} = a + \gamma_0 x_t + \gamma_1 x_{t-1} + \epsilon_t$$

ou encore comme:

$$(1 - \phi_1) y_{t-1} + \Delta y_t = a + (\gamma_0 + \gamma_1) x_{t-1} + \gamma_0 \Delta x_t + \epsilon_t$$

Si l'on définit $\beta = (1 - \phi_1)^{-1}(\gamma_0 + \gamma_1) = \phi^{-1}(1)\gamma(1)$, on peut écrire:

$$\Delta y_t = a - (1 - \phi_1) y_{t-1} + (1 - \phi_1)\beta x_{t-1} + \gamma_0 \Delta x_t + \epsilon_t$$
$$\Delta y_t = a - (1 - \phi_1)[y_{t-1} - \beta x_{t-1}] + \gamma_0 \Delta x_t + \epsilon_t$$

L'interprétation de $y_t = \beta x_t + u_t$ est celle d'une fonction de consommation à long terme. Le terme entre crochets est l'erreur u_{t-1} de cette relation à long terme. Le terme $-(1 - \phi_1)u_{t-1}$ est la "correction d'erreur" qui est ajoutée à un modèle linéaire dans les différences premières des variables.

CHAPITRE XVI

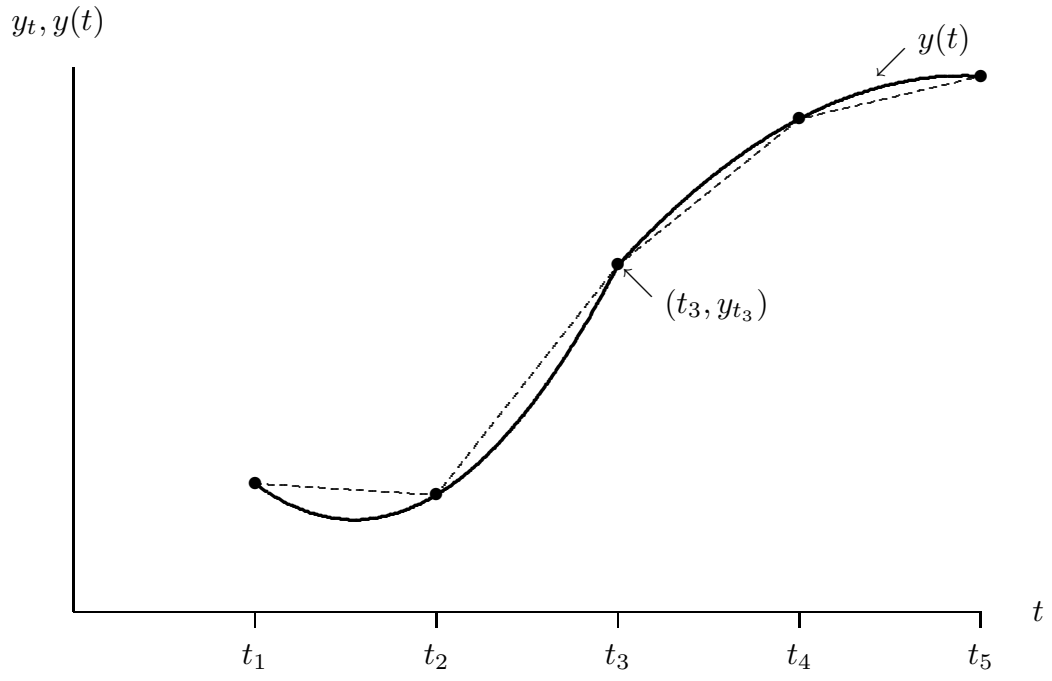
RACINES UNITAIRES ET COINTÉGRATION

16.1 Processus stochastiques

Un processus stochastique discret peut être considéré comme une suite infinie de variables aléatoires, telle que $\{Y_t\}_{t=-\infty}^{+\infty}$ ou $\{Y_t\}_{t=0}^{+\infty}$.

Un processus stochastique continu peut être considéré comme une fonction aléatoire d'une variable continue t, telle que $\{Y(t), t \in \mathbb{R}\}$ ou $\{Y(t), t \in [0,1]\}$.

En interpolant linéairement entre les points (t_i, Y_{t_i}) et $(t_{i+1}, Y_{t_{i+1}})$, on peut obtenir un processus continu à partir d'un processus discret. En posant $t_{i+1} - t_i = \frac{1}{n}$ et en faisant tendre n vers l'infini, on peut aussi obtenir la limite de ce processus, lorsque celle-ci existe. Cette technique est illustrée par le graphique suivant, où les y_{t_i} sont des réalisations des variables Y_{t_i} et où $y(t)$ est une réalisation d'un processus continu $Y(t)$, obtenu par passage à la limite.



16.2 Stationnarité faible

Un processus discret $\{Y_t\}$ est faiblement stationnaire ("covariance-stationary") si et seulement si:

$$E(Y_t) = \mu \qquad \text{pour tout } t$$
$$\text{Cov}(Y_t, Y_{t-j}) = \gamma_j \qquad \text{pour tout } j, t.$$

Les espérances et variances sont donc constantes, et la covariance entre Y_t et Y_s ne dépend que de l'intervalle séparant t et s.

Exemples:

(1) Si les variables Y_t sont $N(0,1)$, indépendantes, et identiquement distribuées pour tout t, on a:

$$\mu = 0,$$
$$\gamma_0 = 1,$$
$$\gamma_j = 0 \text{ pour tout } j \neq 0 \ .$$

Le processus est donc stationnaire.

(2) Si $Y_t = \rho Y_{t-1} + \epsilon_t$, où les ϵ_t sont $N(0,1)$ indépendantes et où $|\rho| < 1$, on a:

$$\mu = 0,$$
$$\gamma_0 = (1-\rho^2)^{-1},$$
$$\gamma_j = \rho^j (1-\rho^2)^{-1}.$$

Le processus est donc stationnaire.

(3) Un exemple de processus non stationnaire est fourni par une *marche aléatoire*:

$$Y_t = Y_{t-1} + \epsilon_t$$

où les $\epsilon_t \sim N(0, \sigma^2)$ sont indépendantes et où $Y_0 = 0$. En effet:

$$\begin{aligned} Y_t &= Y_{t-2} + \epsilon_{t-1} + \epsilon_t \\ &= Y_{t-3} + \epsilon_{t-2} + \epsilon_{t-1} + \epsilon_t \\ &= \dots \\ &= Y_0 + \epsilon_1 + \epsilon_2 + \cdots + \epsilon_t = \sum_{i=1}^{t} \epsilon_i \end{aligned}$$

On a:

$$E(Y_t) = 0, \quad V(Y_t) = t\sigma^2,$$
$$E(Y_t Y_{t-j}) = (t-j)\sigma^2 \quad \text{pour } j \geq 0.$$

La variance de Y_t dépend donc de t, de même que la covariance entre Y_t et Y_{t-j}.

16.3 Processus intégré d'ordre d

Définition:

Un processus discret $\{Y_t\}$ est $I(d)$ si et seulement si:

$$\Delta^d Y_t = \alpha + \beta t + u_t$$
$$\phi(L)u_t = \psi(L)\epsilon_t$$

où $\phi(L)$ et $\psi(L)$ sont inversibles et les ϵ_t sont sphériques.

Interprétation d'un processus I(d):

d est le nombre de fois qu'il faut différencier Y_t pour arriver à un processus stationnaire après soustraction de la tendance linéaire βt . Si $d \geq 1$, on dit que le processus est "intégré".

Cas particuliers d'un processus I(d):

(1) $d = 0 \Rightarrow Y_t$ suit un processus dit "stationnaire à tendance" ("trend-stationary").
(2) $d = 1,\ \alpha = 0,\ \beta = 0,\ \phi(L) = \psi(L) = 1 \Rightarrow Y_t$ suit une marche aléatoire ("random walk").
(3) $d = 1,\ \alpha \neq 0,\ \beta = 0,\ \phi(L) = \psi(L) = 1 \Rightarrow Y_t$ suit une marche aléatoire avec dérive ("random walk with drift").

16.4 Le test de Dickey-Fuller augmenté

Introduction

Soit $\{Y_t\}$ un processus stochastique discret. Quelle est la distribution limite de:

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Y_t$$

lorsque $n \to \infty$?

Au chapitre X, nous avons vu les cas suivants:

(a) Si les Y_t sont indépendantes et identiquement distribuées d'espérance nulle et de variance σ^2, le théorème de Lindeberg-Levy vu à la section 10.8.1 nous dit que:

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Y_t \underset{d}{\longrightarrow} N(0, \sigma^2)$$

(b) A la section 10.8.2, nous avons généralisé ce résultat à des suites de variables indépendantes, mais pas identiquement distribuées: Si les Y_t sont indépendantes d'espérance nulle et de variance σ_t^2 et si $E(Y_t^3) < \infty$, alors:

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Y_t \underset{d}{\longrightarrow} N(0, \sigma^2)$$

où $\sigma^2 = \lim \frac{1}{n} \sum_{t=1}^{n} \sigma_t^2$.

(c) A la section 10.8.3, nous avons généralisé ce résultat à des suites de variables Y_t dépendantes du type $Y_t = u_t u_{t-1}$, où les u_t sont indépendantes et identiquement distribuées d'espérance nulle. Nous avons vu que dans ce cas, sous certaines hypothèses:

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Y_t \underset{d}{\longrightarrow} N(0, \sigma^2)$$

où $\sigma^2 = \text{plim } \frac{1}{n} \sum_{t=1}^{n} Y_t^2$.

Nous devons maintenant examiner un nouveau cas, celui de l'exemple 3 de la section 16.2. On peut montrer que dans ce nouveau cas, à savoir:

$$Y_t = Y_{t-1} + \epsilon_t, \quad Y_0 = 0, \quad \epsilon_t \quad \text{i.i.d}, \quad E(\epsilon_t) = 0, \quad V(\epsilon_t) = \sigma^2,$$

nous avons les résultats suivants:

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Y_t \text{ ne converge pas}$$

$$\frac{1}{n\sqrt{n}} \sum_{t=1}^{n} Y_t \underset{d}{\longrightarrow} N(0, \frac{\sigma^2}{3}).$$

Donc, si l'on a affaire à des processus intégrés, les résultats limites habituels ne seront, en général, plus valables. D'où l'intérêt d'un test destiné à la détection de variables $I(1)$.

La régression de Dickey-Fuller

Notre point de départ sera la formulation d'un modèle suffisamment général, décrivant le comportement d'une série de réalisations y_t. Ce modèle doit permettre l'application de la définition d'un processus $I(1)$ vue à la section 16.3. On suppose donc que:

$$\phi(L) y_t = \alpha + \delta t + \epsilon_t \tag{1}$$

avec:

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p.$$

Le degré p est choisi suffisamment élevé, de façon à ce que les ϵ_t soient sphériques. Nous appliquons maintenant à $\phi(L)$ le lemme 15.1 vu au chapitre précédent. Ceci donne:

$$\phi(L) = \phi(1)L + \phi^*(L)(1-L) \tag{2}$$

avec:

$$\phi_0^* = \phi_0 = 1$$

$$\phi_j^* = -\sum_{s=j+1}^{p} \phi_s \quad \text{pour } j = 1, \ldots, p-1 \text{ et } p > 1$$

$$\phi^*(L) = \sum_{j=0}^{p-1} \phi_j^* L^j.$$

Nous substituons enfin l'équation (2) dans l'équation (1), pour obtenir:

$$\phi(1)\underbrace{y_{t-1}}_{Ly_t} + \underbrace{\Delta y_t + \sum_{j=1}^{p-1} \phi_j^* \Delta y_{t-j}}_{\phi^*(L)(1-L)y_t} = \alpha + \delta t + \epsilon_t$$

ou encore:

$$y_t = \alpha + \delta t + \rho y_{t-1} + \sum_{j=1}^{p-1} \beta_j \Delta y_{t-j} + \epsilon_t \tag{3}$$

avec $\rho = 1 - \phi(1)$ et $\beta_j = -\phi_j^*$.

Ceci est la régression de Dickey-Fuller. Si y_t est $I(1)$, $\sum_j \beta_j \Delta y_{t-j} + \epsilon_t$ est $I(0)$. La comparaison avec la définition d'un processus $I(1)$ montre que $\rho = 1$. Le test est celui de

$$\begin{aligned} H_0 &: \rho = 1 \qquad \text{contre} \\ H_1 &: \rho < 1. \end{aligned}$$

La statistique de Dickey-Fuller est alors la statistique t pour le test de cette hypothèse, à savoir:

$$\text{TDF} = \frac{\hat{\rho}_{mco} - 1}{\hat{\sigma}_{\hat{\rho}_{mco}}}$$

Mais cette statistique n'a pas une distribution limite normale car ρ est le coefficient d'un régresseur $I(1)$. Les valeurs critiques de la statistique TDF sont fournies par Hamilton, *Time Series Analysis*, 1994, Table B6, Case 4, p. 763. Pour prendre un exemple, si $n = 100$ et $\alpha = 0.05$, on va rejeter $H_0 : \rho = 1$ si TDF< -3.45, alors que la valeur critique normale est égale à -1.645.

Pour le test de la nullité d'un ou de plusieurs β_j (coefficients de Δy_{t-j}), on peut utiliser les tests habituels (t ou F, tables Student et Fisher).

Limite en distribution de TDF sous $\mathbf{H_0}$.

Le résultat suivant est démontré par Hamilton, *Time Series Analysis*, 1994, pp. 499–500.

Sous $H_0 : \rho = 1$, TDF converge en distribution vers la variable aléatoire suivante:

$$\frac{[\,0 \quad 1 \quad 0\,]\, A^{-1} \begin{bmatrix} W(1) \\ \frac{1}{2}[W^2(1) - 1] \\ W(1) - \int_0^1 W(r)\, dr \end{bmatrix}}{\left([\,0 \quad 1 \quad 0\,]\, A^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right)^{\frac{1}{2}}}$$

où:

$$A = \begin{bmatrix} 1 & \int_0^1 W(r)\, dr & \frac{1}{2} \\ \int_0^1 W(r)\, dr & \int_0^1 W^2(r)\, dr & \int_0^1 rW(r)\, dr \\ \frac{1}{2} & \int_0^1 rW(r)\, dr & \frac{1}{3} \end{bmatrix}$$

et où $W(r)$ est un *mouvement Brownien standard*, qui est le processus stochastique continu obtenu comme limite de:

$$Z_t = Z_{t-1} + \epsilon_t, \quad Z_0 = 0, \quad \epsilon_t \sim N(0, \frac{1}{n}) \quad \text{indépendantes,}$$

lorsque $t = 1, \ldots, n$ et $n \to \infty$.

Afin d'expliquer la nature de ce processus continu, nous allons en donner une interprétation constructive, qui permettra notamment de simuler les distributions des intégrales apparaissant dans la variable limite précédente. Ces intégrales sont des variables aléatoires: le processus $W(r)$ peut en effet être considéré come une fonction aléatoire de r (voir la section 16.1) et l'intégrale d'une fonction est un nombre.

Considérons alors la suite des variables précédentes, qui peuvent s'écrire:

$$Z_t = \sum_{s=1}^{t} \epsilon_s \quad \text{pour } t = 1, \ldots, n.$$

Z_t a la distribution $N(0, \frac{t}{n})$. Soit $r \equiv \frac{t}{n}$; comme une variable normale centrée est entièrement caractérisée par sa variance, r caractérise entièrement Z_t. Notre définition implique donc que si $n \to \infty$, $\{Z_t\}$ converge en distribution vers:

$$\{W(r), 0 \leq r \leq 1\}.$$

Généralisons maintenant ceci au cas où l'on a une suite de variables Y_t caractérisées par:

$$Y_t = Y_{t-1} + u_t, \quad Y_0 = 0, \quad u_t \sim N(0,1) \quad \text{indépendantes.}$$

On peut se ramener au cas précédent en divisant les deux membres de l'égalité précédente par $\sqrt{n}$, et en définissant $Z_t = Y_t/\sqrt{n}$, $\epsilon_t = u_t/\sqrt{n}$. On a alors:

$$\{\frac{Y_t}{\sqrt{n}}\} \underset{d}{\longrightarrow} \{W(r), 0 \leq r \leq 1\}.$$

On peut donc approcher une réalisation de $W(r)$ en engendrant un grand nombre de réalisations u_t des innovations, et en engendrant par récurrence des réalisations $y_t/\sqrt{n}$ pour $t = 1, \ldots, n$.

Les variables $W(1)$ et $W^2(1)$ qui apparaissent dans la variable limite sont faciles à comprendre: $W(1)$ est la valeur de $W(r)$ au point $r = 1$, c'est donc la variable normale réduite Z_n. $W^2(1)$ est le carré d'une normale réduite, c'est-à-dire une χ^2 à un degré de liberté.

Intéressons-nous maintenant aux intégrales apparaissant dans la variable limite. On peut approcher les intégrales par des sommes de surfaces de rectangles dont les bases sont de longueur $1/n$ et les hauteurs $Y_t/\sqrt{n}$, donc:

$$\begin{aligned}
\int_0^1 W(r)\,dr &\approx \frac{\sum Y_t}{n\sqrt{n}} \\
\int_0^1 W^2(r)\,dr &\approx \frac{1}{n}\frac{\sum Y_t^2}{n} = \frac{\sum Y_t^2}{n^2} \\
\int_0^1 rW(r)\,dr &\approx \frac{1}{n}\sum \frac{t}{n}\frac{Y_t}{\sqrt{n}} = \frac{1}{n^2\sqrt{n}}\sum tY_t
\end{aligned}$$

Pour simuler, par exemple, $\int_0^1 W(r)\,dr$, on peut:

(1) engendrer $n = 1000$ réalisations de variables u_t normales réduites indépendantes;
(2) calculer par récurrence $n = 1000$ réalisations y_t;
(3) calculer:

$$\frac{\sum_{t=1}^n y_t}{n\sqrt{n}}.$$

On a alors une réalisation simulée d'une approximation de $\int_0^1 W(r)\,dr$.

Si l'on refait cet exercice 10000 fois, on a alors 10000 réalisations simulées de cette variable aléatoire. L'histogramme de ces 10000 réalisations est une bonne approximation de la densité de $\int_0^1 W(r)\,dr$.

En fait, Hamilton (*Time Series Analysis*, 1994, p.485) montre que $\int_0^1 W(r)\,dr$ a la distribution $N(0, 1/3)$. Dans des cas plus compliqués, tels que la simulation de la distribution limite de la statistique TDF, la méthode de simulation est la seule possible. Il faut bien noter que les variables aléatoires apparaissant dans la variable limite sont fonction d'un *même* processus $W(r)$.

Notes sur le test TDF:

(1) Si l'on n'inclut pas la constante ou la tendance linéaire dans la régression de Dickey-Fuller, la distribution limite change (les tables à employer sont différentes !). Voir Hamilton, pp.528–529, pour les détails.
(2) L'inclusion d'une constante et d'une tendance linéaire dans la régression de Dickey-Fuller est conseillée dans l'intérêt de la robustesse (il est plus grave d'omettre à tort des régresseurs que de faire l'erreur inverse).
(3) La variable limite précédente a été obtenue sous l'hypothèse auxiliaire que $\delta = 0$ (pas de tendance linéaire dans l'équation (3) de cette section lorsque $\rho = 1$, c'est-à-dire dans le modèle en différences premières). Le test précédent n'est donc approprié que si les y_t ne présentent pas de tendance quadratique manifeste. La meilleure stratégie à adopter dans le cas contraire reste une question ouverte.
(4) La technique de calcul des valeurs critiques illustre la puissance de la méthodologie de simulation stochastique.
(5) La variable limite reste inchangée si les erreurs de la régression de Dickey-Fuller ne sont pas normales, pour autant qu'un théorème central limite fonctionnel soit applicable (voir Hamilton, p.479).

16.5 Variables cointégrées

On peut obtenir un processus $I(0)$ à partir d'un processus $I(1)$ en prenant les différences premières du processus $I(1)$. Malheureusement, ceci supprime toutes les informations à long terme. Pour cette raison, on a défini une autre approche permettant d'obtenir un processus $I(0)$, celle de la cointégration.

Définition:

Soit $Y_{1t}, Y_{2t}, \ldots, Y_{kt}$ des processus stochastiques $I(1)$. Ces processus sont dits cointégrés s'il existe un vecteur $a \neq 0$ tel que :

$$a'Y_t = \sum_{i=1}^{k} a_i Y_{it}$$

soit un processus $I(0)$.

Exemple:

Soit y_{1t} une série d'observations sur le logarithme de la consommation par tête à prix constants, et soit y_{2t} une série d'observations sur le logarithme du revenu disponible par

tête à prix constants. On fait l'hypothèse que ces deux séries sont des réalisations de processus $I(1)$:

$$y_{1t} = \mu_1 + y_{1,t-1} + \epsilon_{1t}$$
$$y_{2t} = \mu_2 + y_{2,t-1} + \epsilon_{2t}$$

On aura cointégration si la série $y_{1t} - \alpha y_{2t} = u_t$ est une réalisation d'un processus $I(0)$.

Interprétation:

Le vecteur cointégrant est ici $a = (1, -\alpha)$. On a une "relation de cointégration":

$$y_{1t} = \alpha y_{2t} + u_t$$

où u_t est $I(0)$. On peut interpréter cette relation comme une fonction de consommation à long terme, mais l'interprétation est différente de celle que l'on avait dans le cas où y_{1t} et y_{2t} étaient stationnaires. En effet, les "niveaux d'équilibre" de y_{1t} et y_{2t} n'existent pas, car:

$$\begin{aligned} y_{it} &= \mu_i + y_{i,t-1} + \epsilon_{it} \\ &= \mu_i + \mu_i + y_{i,t-2} + \epsilon_{i,t-1} + \epsilon_{it} \\ &= \dots \\ &= t\mu_i + \sum_{s=1}^{t} \epsilon_{is} + y_{i0}; \end{aligned}$$

donc $E(y_{it})$ n'est pas bornée.

On ne peut donc pas avoir une relation entre les niveaux d'équilibre des variables, mais $y_{1t} = \alpha y_{2t}$ peut être considérée comme l'équation d'un "attracteur".

Test de l'hypothèse de cointégration.

L'idée de base est la suivante. On va faire un test de racines unitaires sur les résidus de la relation de cointégration obtenus par la méthode des moindres carrés ordinaires (cette méthodologie est la plus ancienne et la plus simple).

Il faut néanmoins prendre garde au fait que les distributions limites sont différentes de celles des tests de Dickey-Fuller précédents, car l'estimation par moindres carrés repose sur l'hypothèse de cointégration. La mise en oeuvre se déroule comme suit:

(1) On teste si $y_t, x_{t1}, \dots, x_{tk}$ sont $I(1)$, à l'aide du test TDF précédent appliqué à chacune de ces variables.

(2) On estime par moindres carrés ordinaires la relation de cointégration:

$$y_t = \alpha + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t$$

Ceci donne des résidus $\hat{u}_t$.

(3) On teste $\rho = 1$ contre $\rho < 1$ dans la régression:

$$\hat{u}_t = \rho \hat{u}_{t-1} + \sum_{j=1}^{p} \Delta \hat{u}_{t-j} + \epsilon_t.$$

La statistique TCO $= (\hat{\rho} - 1)/\hat{\sigma}_{\hat{\rho}}$ est à comparer avec les valeurs critiques fournies par Hamilton, Table B9, Case 3, p.766. Ces valeurs critiques sont valables dans le cas où au moins l'une des variables y_t, $x_{1t}, \ldots, x_{kt}$ possède une dérive non nulle.

16.6 Régressions de cointégration

Quelles sont les propriétés des estimateurs par moindres carrés ordinaires des coefficients de la relation:

$$y_t = \alpha + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$$

où *toutes* les variables $y_t, x_{t1}, \ldots, x_{tk}$ sont $I(1)$ mais où u_t est $I(0)$? Stock (Econometrica 55, 1987, pp.1035–1056) montre que si $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)$, alors:

$$\sqrt{n}(\hat{\beta} - \beta) \underset{p}{\longrightarrow} 0 \quad \text{(on dit que l'on a "superconvergence");}$$
$$n(\hat{\beta} - \beta) \underset{d}{\longrightarrow} \text{ vecteur non standard.}$$

Le problème ne se pose donc pas au niveau de l'estimation ponctuelle, mais au niveau des tests. L'étude de ces derniers ne sera pas faite ici. Plusieurs méthodologies possibles sont décrites dans Hamilton, chap. 19 et 20.

On peut substituer dans un modèle de correction d'erreur les résidus d'une relation de cointégration estimée par moindres carrés ordinaires. Pour reprendre l'exemple de la section 15.5, on peut estimer β par moindres carrés ordinaires dans la relation $y_t = \beta x_t + u_t$, puis estimer, toujours par moindres carrés ordinaires, a, ϕ_1, et γ_0 dans le modèle:

$$\Delta y_t = a - (1 - \phi_1)[y_{t-1} - \hat{\beta} x_{t-1}] + \gamma_0 \Delta x_t + \epsilon_t.$$

16.7 Régressions factices ("spurious regresssions")

Que se passe-t-il si l'on estime par moindres carrés la relation:

$$y_t = \alpha + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$$

où *toutes* les variables $y_t, x_{t1}, \ldots, x_{tk}$, et u_t sont $I(1)$? Dans ce cas, on n'a pas de cointégration.

Phillips (Journal of Econometrics 33, 1986, pp.311–340) montre que:

(1) $(\frac{\hat{\alpha}}{\sqrt{n}}, \hat{\beta}_1, \ldots, \hat{\beta}_k) \underset{d}{\longrightarrow}$ vecteur non standard

(2) Pour le test $\beta = 0$ contre $\beta \neq 0$:

$$n^{-1} F_{obs} \underset{d}{\longrightarrow} \text{ variable non standard.}$$

Donc $\hat{\alpha}$ et F_{obs} *divergent* et les $\hat{\beta}_i$ ne convergent pas en probabilité! Ceci même si les $k+1$ variables $y_t, x_{t1}, \ldots, x_{tk}$ sont indépendantes entre elles. Pour tout c, on a que:

$$\lim_{n\to\infty} P[F_{obs} > c] = 1,$$

donc on rejettera *toujours* $\beta = 0$ si n est assez grand.

16.8 Conclusions

(1) La modélisation économétrique des variables $I(1)$ est un problème difficile. Le domaine manque de maturité (plusieurs questions restent ouvertes).

(2) La notion de cointégration est récente et reste contestée. Elle présente notamment deux difficultés:

—L'équivalence observationnelle, en petit échantillon, d'un processus $I(1)$ et d'un processus "presque non stationnaire", par exemple le suivant:

$$Y_t = 0.9999 Y_{t-1} + \epsilon_t.$$

—Le manque de puissance des tests de racines unitaires couramment utilisés.

Donc la classification d'une variable entre $I(0)$ et $I(1)$ reste un peu un jugement de valeurs, or l'étude de la relation entre les variables dépend crucialement d'une telle classification.

(3) Les distributions limites des statistiques de test et des estimateurs dépendent crucialement des hypothèses faites sur le modèle vrai. On peut tester ces hypothèses, mais ceci n'élimine pas le risque d'une inférence incorrecte.

(4) La cointégration est donc une hypothèse de travail, qui donne de bons résultats dans certains cas, pas dans d'autres. Ce n'est pas une panacée.

(5) Il faut connaître les concepts de base car les problèmes posés sont importants. Le but de cette introduction était précisément de rendre familiers ces concepts de base (qui peuvent être déroutants lorsqu'on les rencontre pour la première fois).

TROISIÈME PARTIE

SYSTÈMES D'ÉQUATIONS SIMULTANÉES

CHAPITRE I.

INTRODUCTION

1.1 Explication intuitive du biais dû à la simultanéité

Il arrive souvent qu'un modèle économique comprenne plusieurs équations simultanées. Comme nous allons le voir, si l'on ne tient pas compte de cette situation lors de l'estimation des paramètres du modèle, les estimateurs obtenus pourront présenter un *biais de simultanéité*, qui ne disparaîtra pas lorsque la taille de l'échantillon tend vers l'infini (défaut de convergence). En effet, certains régresseurs seront stochastiques, et seront corrélés avec le terme d'erreur contemporain.

Nous illustrerons ce phénomène au moyen de deux exemples.

Exemple 1

Le modèle suivant, dont l'origine remonte à Haavelmo, comporte deux équations: une équation stochastique de comportement, et une définition (identité comptable):

$$\begin{aligned} C_t &= a + bY_t + u_{1t} \\ Y_t &= C_t + I_t \end{aligned}$$

où C_t est la consommation, Y_t le revenu national, I_t l'investissement, et u_{1t} est un terme d'erreur formant un vecteur u_1 avec $E(u_1) = 0, E(u_1 u_1^{'}) = \sigma^2 I$.

En substituant la première équation dans la seconde, on obtient:

$$Y_t = a + bY_t + u_{1t} + I_t,$$

soit aussi:

$$Y_t = \frac{a}{1-b} + \frac{1}{1-b} I_t + \frac{u_{1t}}{1-b}.$$

Donc si $E(I_t u_{1t}) = 0$, on a:

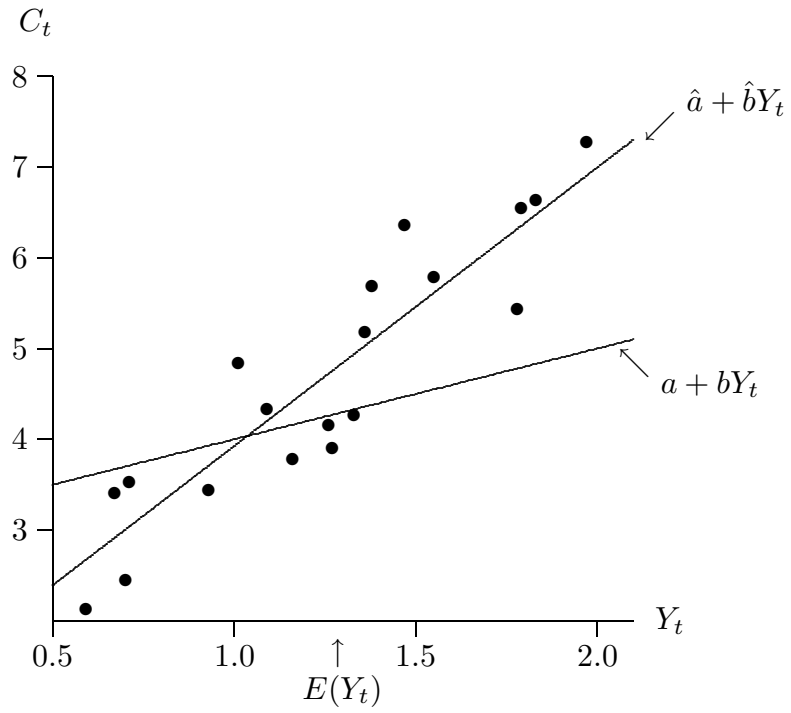
$$E(Y_t u_{1t}) = E(\frac{u_{1t}^2}{1-b}) = \frac{\sigma^2}{1-b} \neq 0,$$

et l'application des moindres carrés ordinaires à la première équation ne donne pas des estimateurs convergents.

Si $E(Y_t u_{1t}) > 0$, nous aurons, avec une probabilité relativement forte:

$$\begin{aligned} u_{1t} > E(u_{1t}) = 0 \quad &\text{lorsque} \quad Y_t > E(Y_t) \\ u_{1t} < E(u_{1t}) = 0 \quad &\text{lorsque} \quad Y_t < E(Y_t) \quad . \end{aligned}$$

Si l'on représente alors les deux droites $C_t = a + bY_t$ et $\hat{C}_t = \hat{a} + \hat{b}Y_t$, la pente de cette dernière droite est la plus forte, car $\hat{a}$ et $\hat{b}$ minimisent la somme des carrés des résidus:



Exemple 2

Nous avons ici deux équations de comportement, une loi d'offre et une loi de demande. Les quantités demandées (q_t) dépendent du prix (p_t) et du revenu (r_t). Le prix (p_t) dépend des quantités offertes (q_t) et du coût de production (c_t). Le système s'écrit:

$$\text{(i)} \qquad q_t = a_1 + b_1 r_t + c_1 p_t + u_{1t}$$

$$\text{(ii)} \qquad p_t = a_2 + b_2 c_t + c_2 q_t + u_{2t}$$

Donc p_t dépend de q_t dans (ii), qui dépend de u_{1t} dans (i): nous concluons que p_t est corrélée avec u_{1t}. Mais p_t apparaît comme régresseur dans (i): nous avons donc un problème de simultanéité comme auparavant.

1.2 Variables endogènes et prédéterminées

Les variables p_t et q_t de l'exemple précédent sont dites endogènes: elles sont déterminées par le modèle, et dépendent des termes d'erreur de chacune des équations. Les variables c_t et r_t sont dites prédéterminées: par hypothèse, elles ne sont corrélées avec aucun des termes d'erreurs contemporains.

Comme on le verra par la suite, il est important de faire une distinction entre variables exogènes et variables prédéterminées. Les variables exogènes sont déterminées par des relations n'appartenant pas au modèle: elles ne sont donc corrélées, ni avec les termes d'erreurs contemporains, ni avec les autres termes d'erreur. En revanche, les variables prédéterminées comprennent, non seulement les variables exogènes, mais aussi les variables endogènes retardées, pour autant que les erreurs ne soient pas corrélées dans le temps.

1.3 Présentation matricielle et hypothèses

Nous pouvons écrire le système d'équations précédent sous la forme canonique suivante:

$$\begin{aligned} q_t - c_1 p_t - a_1 - b_1 r_t - 0 c_t &= u_{1t} \\ -c_2 q_t + p_t - a_2 - 0 r_t - b_2 c_t &= u_{2t} \quad , \end{aligned}$$

ou, sous forme matricielle:

$$\begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} \begin{pmatrix} q_t \\ p_t \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \end{pmatrix} \begin{pmatrix} 1 \\ r_t \\ c_t \end{pmatrix} = \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

avec les restrictions $\beta_{11} = 1$, $\beta_{22} = 1$, $\gamma_{13} = 0$, $\gamma_{22} = 0$. En général donc, nous avons le format suivant pour un système de g équations, comportant g variables endogènes et k variables prédéterminées:

$$By_t + \Gamma x_t = u_t$$

où B est une matrice $g \times g$ de coefficients des variables endogènes;
Γ est une matrice $g \times k$ de coefficients des variables prédéterminées;
y_t est un vecteur $g \times 1$ de variables endogènes;
x_t est un vecteur $k \times 1$ de variables prédéterminées;
u_t est un vecteur $g \times 1$ d'erreurs inobservables.

Les hypothèses de ce modèle sont les suivantes:

(H_1) $E(u_t) = 0$ pour tout $t = 1, \ldots, n$

(H_2) $E(u_t u_t^{'}) = \Sigma$

(H_3) $E(u_t u_s^{'}) = O_{g \times g}$ $(t \neq s)$

(H_4) B est régulière

(H_5) rang $(X) = k < n$

(H_6) plim $\left(\frac{1}{n} X^{'} U\right) = O_{k \times g}$

(H_7) plim $\left(\frac{1}{n} X^{'} X\right) = \Sigma_{XX}$ est définie positive

$$\text{où} \quad X = \begin{pmatrix} x_1^{'} \\ x_2^{'} \\ \vdots \\ x_n^{'} \end{pmatrix} \quad \text{est} \quad n \times k \quad \text{et}$$

$$U = \begin{pmatrix} u_1^{'} \\ u_2^{'} \\ \vdots \\ u_n^{'} \end{pmatrix} \quad \text{est} \quad n \times g \quad .$$

En réunissant toutes les observations t sur $By_t + \Gamma x_t = u_t$, on peut aussi s'écrire $YB^{'} + X\Gamma^{'} = U$, où Y est $n \times g$.

1.4 Forme structurelle et forme réduite

Le système $By_t + \Gamma x_t = u_t$ s'appelle la forme structurelle du modèle: c'est la représentation formelle d'un modèle économique et ce sont donc les paramètres de ce système que nous voulons estimer. Néanmoins, comme nous l'avons vu, nous ne pouvons estimer ces paramètres par la méthode des moindres carrés ordinaires appliquée à chaque équation.

Nous allons donc transformer la forme structurelle en un système dérivé, dit forme réduite, qui exprime chaque variable endogène en fonction de toutes les variables prédéterminées du modèle, et des erreurs.

Prémultiplions les deux membres de $By_t + \Gamma x_t = u_t$ par B^{-1}. Il vient:

$$y_t = \Pi x_t + v_t \quad \text{avec} \quad \Pi = -B^{-1}\Gamma \quad \text{et} \quad v_t = B^{-1}u_t \quad .$$

Comme nous le verrons, les g équations de ce nouveau système peuvent être estimées par moindres carrés ordinaires, sans problème de simultanéité.

La forme réduite peut aussi s'écrire:

$$Y = X\Pi^{'} + V \quad , \text{où} \quad V = U(B^{'})^{-1} \quad .$$

Comme cas particuliers de la forme réduite, nous pouvons mentionner:

(1) Le modèle MANOVA (multivariate analysis of variance) où les variables exogènes ne prennent que les valeurs 0 et 1.

(2) Le modèle autorégressif vectoriel (VAR). Ce modèle peut s'écrire:

$$\Phi(L)y_t = \pi_0 + v_t$$

où $\Phi(L)$ est une matrice de polynômes:

$$\Phi(L) = I - \Pi_1 L - \cdots - \Pi_p L^p.$$

On a alors:

$$y_t = \pi_0 + \Pi_1 y_{t-1} + \cdots + \Pi_p y_{t-p} + v_t$$

ce qui correspond bien à l'équation $y_t = \Pi x_t + v_t$, si l'on définit:

$$x_t = \begin{pmatrix} 1 \\ y_{t-1} \\ \vdots \\ y_{t-p} \end{pmatrix}$$

$$\Pi = (\, \pi_0 \quad \Pi_1 \quad \Pi_2 \quad \ldots \quad \Pi_p \,).$$

(3) Le modèle autorégressif à retards échelonnés vectoriel, où l'on a un nombre arbitraire de variables exogènes formant un vecteur z_t et un nombre arbitraire de retards de ces variables. Il s'agit d'une généralisation du modèle VAR précédent, qui peut s'écrire comme:

$$\Phi(L)y_t = \Gamma(L)z_t + v_t.$$

Un cas particulier de ce type de modèle sera étudié en détail à la section 1.7.

1.5 Propriétés statistiques de la forme réduite

Il est facile de vérifier que:

$$
\begin{aligned}
&E(v_t) = 0 \\
&E(v_t v_t') = B^{-1}\Sigma(B')^{-1} \\
&E(v_t v_s') = O_{g\times g} \quad \text{pour } t \neq s \\
&\operatorname{plim}\left(\frac{1}{n}X'V\right) = O_{k\times g}.
\end{aligned}
$$

Donc les erreurs de la forme réduite sont d'espérance nulle, homoscédastiques, non corrélées dans le temps, et non corrélées avec les régresseurs contemporains.

On peut par conséquent estimer les équations de la forme réduite par moindres carrés ordinaires. La colonne i de l'égalité matricielle $Y = X\Pi' + V$ peut s'écrire:

$$y^i = X\beta^i + v^i$$

où β^i est la colonne i de la matrice Π'. Ceci est une équation de régression du type habituel, et par conséquent:

$$\hat{\beta}^i = (X'X)^{-1}X'y^i$$

$$\hat{\Pi}' = (X'X)^{-1}X'Y.$$

On montrera plus loin (section 5.1) que cet estimateur est aussi l'estimateur par maximum de vraisemblance lorsque les erreurs sont normales. En revanche, comme nous l'avons indiqué, la forme structurelle ne peut pas être estimée par MCO.

1.6 Interprétation économique de la forme réduite

Reprenons le modèle de la section 1.1:

$$
\begin{aligned}
C_t &= a + bY_t + u_{1t} \\
Y_t &= C_t + I_t
\end{aligned}
$$

L'estimation des paramètres de cette forme structurelle ne fournit que les propensions marginales et moyennes à consommer. On pourrait aussi se demander quel est l'impact sur la consommation d'une augmentation des dépenses d'investissement. Cet impact est bien entendu mesuré par le multiplicateur.

Nous allons voir que ce multiplicateur n'est autre que l'un des coefficients de la forme réduite. Ces coefficients mesurent donc l'effet sur les variables endogènes d'un changement des variables prédéterminées, lorsque l'on tient compte de la simultanéité du système.

La forme structurelle s'écrit $By_t + \Gamma x_t = u_t$, avec

$$B = \begin{pmatrix} 1 & -b \\ -1 & 1 \end{pmatrix}, \Gamma = \begin{pmatrix} -a & 0 \\ 0 & -1 \end{pmatrix},$$

$$y_t = \begin{pmatrix} C_t \\ Y_t \end{pmatrix}, x_t = \begin{pmatrix} 1 \\ I_t \end{pmatrix}, \quad \text{et} \quad u_t = \begin{pmatrix} u_{1t} \\ 0 \end{pmatrix}.$$

Donc:

$$\begin{aligned} \Pi &= -B^{-1}\Gamma = -\frac{1}{1-b}\begin{pmatrix} 1 & b \\ 1 & 1 \end{pmatrix}\begin{pmatrix} -a & 0 \\ 0 & -1 \end{pmatrix} \\ &= -\frac{1}{1-b}\begin{pmatrix} -a & -b \\ -a & -1 \end{pmatrix}, \end{aligned}$$

et la forme réduite s'écrit:

$$\begin{aligned} C_t &= \frac{a}{1-b} + \frac{b}{1-b} I_t + v_{1t} \\ Y_t &= \frac{a}{1-b} + \frac{1}{1-b} I_t + v_{2t} \quad . \end{aligned}$$

On obtient donc directement $\dfrac{dC_t}{dI_t} = \dfrac{b}{1-b}$ et $\dfrac{dY_t}{dI_t} = \dfrac{1}{1-b}$.

1.7 Forme réduite dynamique, forme finale, multiplicateurs

Certaines variables prédéterminées sont ici des variables endogènes retardées. Dans le cas particulier d'un seul retard, nous pouvons écrire la forme réduite comme:

$$y_t = \Pi_1 y_{t-1} + \Pi_2 z_t + v_t$$

où y_t est le vecteur des variables endogènes contemporaines, y_{t-1} est le vecteur des variables endogènes retardées, z_t est le vecteur des variables exogènes et Π_1, Π_2 sont des sous-matrices de Π.

Nous allons, au moyen de substitutions successives, exprimer y_t en fonction des seules variables exogènes et des erreurs.

$$\begin{aligned} \text{On a} \quad y_t &= \Pi_1 \left(\Pi_1 y_{t-2} + \Pi_2 z_{t-1} + v_{t-1}\right) + \Pi_2 z_t + v_t \\ &= \Pi_1^2 y_{t-2} + \Pi_1 \Pi_2 z_{t-1} + \Pi_2 z_t + \Pi_1 v_{t-1} + v_t \end{aligned}$$

et, après s substitutions:

$$y_t = \Pi_1^{s+1} y_{t-s-1} + \sum_{j=0}^{s} \Pi_1^j \Pi_2 z_{t-j} + \sum_{j=0}^{s} \Pi_1^j v_{t-j} \quad .$$

On fait alors l'hypothèse que $\lim_{s \longrightarrow \infty} \Pi_1^s = O$, et l'on obtient en passant à la limite:

$$y_t = \sum_{j=0}^{\infty} C_j z_{t-j} + \sum_{j=0}^{\infty} \Pi_1^j v_{t-j},$$

avec:

$$C_j \underset{\text{def}}{=} \Pi_1^j \Pi_2.$$

Cette dernière équation s'appelle la *forme finale* du modèle. Elle permet d'obtenir, par simple lecture, les multiplicateurs dynamiques. On distingue:

(1) Les multiplicateurs d'impact: ce sont les composantes de $C_0 = \Pi_2$.
(2) Les multiplicateurs de délai j: ce sont les composantes de C_j. Ils mesurent l'effet sur les y_t d'une variation *temporaire* des variables exogènes à la période $t - j$.
(3) Les multiplicateurs cumulés: ce sont les composantes de la matrice $D_\tau = \sum_{j=0}^{\tau} C_j$. Ils mesurent l'effet sur les y_t d'une variation prolongée des variables exogènes durant les $\tau + 1$ périodes $t - \tau$, $t - \tau + 1, \ldots, t$.
(4) Les multiplicateurs d'équilibre: ce sont les composantes de la matrice:

$$D_\infty = \sum_{j=0}^{\infty} C_j = \left(I + \Pi_1 + \Pi_1^2 + \ldots\right) \Pi_2 = \left(I - \Pi_1\right)^{-1} \Pi_2.$$

Ils mesurent l'effet d'une variation des z_t soutenue pendant une infinité de périodes. Le niveau d'équilibre des variables endogènes est alors donné par $E(\bar{y}) = D_\infty \bar{z}$, où $\bar{z}$ est le nouveau niveau des variables exogènes.

A titre d'exemple, considérons la forme structurelle suivante:

$$\begin{aligned} C_t &= 0.25 + 0.5 Y_t + u_{1t} \\ I_t &= 0.15 + 0.1 Y_t + 0.3 Y_{t-1} + u_{2t} \\ Y_t &= C_t + I_t + G_t \quad . \end{aligned}$$

Supposons qu'à partir d'une situation d'équilibre, le niveau G des dépenses gouvernementales augmente d'une unité à la période $t - 1$, et revienne à la période suivante à son

niveau initial. On demande les effets de cette augmentation temporaire sur C, Y et I à la période t et à la période $t+1$.

Nous avons ici:

$$y_t = \begin{pmatrix} C_t \\ Y_t \\ I_t \end{pmatrix}; \quad x_t = \begin{pmatrix} y_{t-1} \\ \\ z_t \end{pmatrix}; \quad z_t = \begin{pmatrix} 1 \\ \\ G_t \end{pmatrix}$$

et la forme structurelle $By_t + \Gamma x_t = u_t$ s'écrit:

$$\begin{pmatrix} 1 & -0.5 & 0 \\ 0 & -0.1 & 1 \\ -1 & 1 & -1 \end{pmatrix} \begin{pmatrix} C_t \\ Y_t \\ I_t \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & -0.25 & 0 \\ 0 & -0.3 & 0 & -0.15 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} C_{t-1} \\ Y_{t-1} \\ I_{t-1} \\ 1 \\ G_t \end{pmatrix} = \begin{pmatrix} u_{1t} \\ u_{2t} \\ 0 \end{pmatrix} \quad .$$

On vérifie aisément que

$$\Pi = -B^{-1}\Gamma = \begin{pmatrix} 0 & 0.375 & 0 & 0.75 & 1.25 \\ 0 & 0.75 & 0 & 1 & 2.5 \\ 0 & 0.375 & 0 & 0.25 & 0.25 \end{pmatrix}$$

et la forme réduite s'écrit $y_t = \Pi_1 y_{t-1} + \Pi_2 z_t + v_t$, avec:

$$\Pi_1 = \begin{pmatrix} 0 & 0.375 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.375 & 0 \end{pmatrix} \text{ et } \quad \Pi_2 = \begin{pmatrix} 0.75 & 1.25 \\ 1 & 2.5 \\ 0.25 & 0.25 \end{pmatrix} \quad .$$

Les réponses aux questions posées sont données par les multiplicateurs de délai 1, et de délai 2. On vérifie que:

$$C_1 = \Pi_1\Pi_2 = \begin{pmatrix} 0.375 & 0.9375 \\ 0.75 & 1.875 \\ 0.375 & 0.9375 \end{pmatrix}$$

$$C_2 = \Pi_1^2\Pi_2 = \begin{pmatrix} 0.28125 & 0.703125 \\ 0.5625 & 1.40625 \\ 0.28125 & 0.703125 \end{pmatrix} \quad .$$

Donc, si une situation d'équilibre prévaut à la période $t-2$ (soit si $G_{t-2} = \bar{G}$) et si $G_{t-1} - \bar{G} = 1$ tandis que $G_s - \bar{G} = 0$ pour $s \neq t-1$, on a, à un terme d'erreur près:

$$\begin{array}{rr} C_t - \bar{C} = 0.9375 & C_{t+1} - \bar{C} = 0.703125 \\ Y_t - \bar{Y} = 1.875 & Y_{t+1} - \bar{Y} = 1.40625 \\ I_t - \bar{I} = 0.9375 & I_{t+1} - \bar{I} = 0.703125 \end{array}$$

En effet:

$$y_t - \bar{y} = C_0(z_t - \bar{z}) + C_1(z_{t-1} - \bar{z}) + C_2(z_{t-2} - \bar{z}) + \ldots + \epsilon_t$$

$$y_{t+1} - \bar{y} = C_0(z_{t+1} - \bar{z}) + C_1(z_t - \bar{z}) + C_2(z_{t-1} - \bar{z}) + \ldots + \epsilon_{t+1} \quad .$$

Si maintenant l'augmentation des dépenses gouvernementales se maintient pour un nombre infini de périodes, la consommation augmentera, à l'équilibre, de 5 unités; le revenu national, de 10 unités; l'investissement, de 4 unités. En effet:

$$D_\infty = (I - \Pi_1)^{-1}\Pi_2 = \begin{pmatrix} 2.25 & 5 \\ 4 & 10 \\ 1.75 & 4 \end{pmatrix} \quad .$$

1.8 Relation entre la forme réduite dynamique et le modèle AD

Le modèle de la section précédente peut aussi s'écrire:

$$\Phi(L)y_t = \Gamma(L)z_t + v_t$$

où $\Phi(L) = I - \Pi_1 L$ et $\Gamma(L) = \Pi_2$. On s'aperçoit que la matrice D_∞ des multiplicateurs d'équilibre n'est autre que $[\Phi(1)]^{-1}\Gamma(1)$. De manière plus générale, tous les résultats du chapitre XV de la seconde partie ont une généralisation vectorielle dans le présent contexte.

CHAPITRE II.

LE PROBLÈME DE L'IDENTIFICATION

2.1 Structures observationnellement équivalentes

Lorsque nous estimons les paramètres de la forme réduite par la méthode des moindres carrés ordinaires, le problème suivant se pose. Comme nous l'avons signalé à la section 1.4, ce sont les composantes des matrices B et Γ qui nous intéressent en premier lieu. Peut-on, alors, trouver des estimations convergentes *uniques* de ces composantes à partir d'estimations convergentes des composantes de Π? Ce problème est celui de l'identification de B et de Γ.

Pour que B et Γ puissent être identifiées, il faut qu'il existe une correspondance *bijective* entre Π d'une part, B et Γ d'autre part. Donc, il faut qu'à toute forme réduite corresponde une et une seule forme structurelle et réciproquement. Il est facile de voir que sans restrictions sur les coefficients, ceci ne sera jamais le cas. A une forme réduite donnée correspondrait une *infinité* de formes structurelles; ces dernières sont dites observationnellement équivalentes (elles impliquent la même forme réduite).

Considérons en effet les deux formes structurelles suivantes:

$$By_t + \Gamma x_t = u_t \quad , \quad \text{et} \quad (FB)y_t + (F\Gamma)x_t = Fu_t$$

où F est une matrice $g \times g$ régulière, différente de la matrice unité. A la seconde forme structurelle correspond la forme réduite $y_t = -B^{-1}\Gamma x_t + B^{-1}u_t$, comme on le voit facilement si l'on prémultiplie les deux membres par $(FB)^{-1} = B^{-1}F^{-1}$. Cette forme réduite est identique à la première. Les deux formes structurelles sont donc observationnellement équivalentes. Or, il existe une infinité de matrices F régulières.

On vérifie que les deux formes structurelles conduisent à la même fonction de vraisemblance. Le problème du maximum de vraisemblance n'a donc pas de solution unique.

Comment, alors, estimer B et Γ? Nous ne pouvons le faire que grâce aux restrictions à priori que nous fournit la théorie économique sur les composantes de ces matrices. Le problème d'identification est donc conceptuellement fort semblable au problème de multicolinéarité étudié à la section 5.7.1 de la deuxième partie.

En particulier, certaines des composantes seront nulles: les variables correspondantes apparaîtront dans certaines équations, mais pas dans les autres (voir la section 1.1 de cette troisième partie). Ces restrictions impliqueront alors des restrictions sur la matrice F, car

les matrices de coefficients FB et $F\Gamma$ de la structure transformée doivent obéir aux mêmes restrictions que la structure d'origine (dans le cas contraire, nous changerions le modèle!) Si ces restrictions impliquent une matrice de transformation *unique*, il y a correspondance bijective entre forme structurelle et forme réduite: B et Γ sont alors identifiables.

2.2 Systèmes récursifs

Un système récursif est caractérisé par une matrice B triangulaire et une matrice $\Sigma = E(u_t u_t^{'})$ diagonale. Un exemple d'un tel système est donné par:

$$\begin{pmatrix} \beta_{11} & \beta_{12} \\ \\ \beta_{21} & \beta_{22} \end{pmatrix} \begin{pmatrix} y_{1t} \\ \\ y_{2t} \end{pmatrix} + \begin{pmatrix} \gamma_{11} \\ \\ \gamma_{21} \end{pmatrix} x_{1t} = \begin{pmatrix} u_{1t} \\ \\ u_{2t} \end{pmatrix}$$

avec les restrictions $\beta_{11} = \beta_{22} = 1$, $\beta_{12} = 0$, et $E(u_{1t}u_{2t}) = \sigma_{12} = \sigma_{21} = 0$. On peut alors écrire:

$$\begin{aligned} y_{1t} &= -\gamma_{11}x_{1t} + u_{1t} \\ \\ y_{2t} &= -\beta_{21}y_{1t} - \gamma_{21}x_{1t} + u_{2t} \quad . \end{aligned}$$

L'application des moindres carrés ordinaires à chaque équation donne des estimateurs convergents. La propriété est évidente pour la première équation. En ce qui concerne la seconde, il est immédiat que $E(y_{1t}u_{2t}) = 0$, puisque $E(x_{1t}u_{2t}) = 0$ et $E(u_{1t}u_{2t}) = 0$.

Nous allons illustrer la section précédente en vérifiant, par le biais de la matrice de transformation F, que les deux équations du système sont identifiables.

Les matrices de la forme structurelle transformée:

$$\begin{aligned} FB &= \begin{pmatrix} f_{11} & f_{12} \\ \\ f_{21} & f_{22} \end{pmatrix} \begin{pmatrix} \beta_{11} & \beta_{12} \\ \\ \beta_{21} & \beta_{22} \end{pmatrix} = \begin{pmatrix} f_{11}\beta_{11} + f_{12}\beta_{21} & f_{11}\beta_{12} + f_{12}\beta_{22} \\ \\ f_{21}\beta_{11} + f_{22}\beta_{21} & f_{21}\beta_{12} + f_{22}\beta_{22} \end{pmatrix} \\ \\ F\Gamma &= \begin{pmatrix} f_{11}\gamma_{11} + f_{12}\gamma_{21} \\ \\ f_{21}\gamma_{11} + f_{22}\gamma_{21} \end{pmatrix} \end{aligned}$$

doivent obéir aux trois mêmes restrictions que les matrices B et Γ. De même, la matrice de covariance de la forme structurelle transformée doit être diagonale. Nous avons donc les quatre restrictions suivantes (il faut bien noter que ce sont les seules):

$$
\begin{aligned}
f_{11}\beta_{11} + f_{12}\beta_{21} &= 1 \\
f_{11}\beta_{12} + f_{12}\beta_{22} &= 0 \\
f_{21}\beta_{12} + f_{22}\beta_{22} &= 1 \\
f_{11}(\sigma_{11}f_{21} + \sigma_{12}f_{22}) + f_{12}(\sigma_{21}f_{21} + \sigma_{22}f_{22}) &= 0
\end{aligned}
$$

ou, en substituant les quatre restrictions sur les paramètres de la forme structurelle d'origine:

$$
\begin{aligned}
f_{11} + f_{12}\beta_{21} &= 1 \\
f_{12} &= 0 \\
f_{22} &= 1 \\
f_{11}\sigma_{11}f_{21} + f_{12}\sigma_{22}f_{22} &= 0 \quad .
\end{aligned}
$$

Comme $\sigma_{11} \neq 0$, ces quatre équations ont comme solution unique $f_{11} = 1, f_{12} = 0, f_{21} = 0, f_{22} = 1$.

Donc les restrictions impliquent $F = I$, et nous ne pouvons avoir deux formes structurelles différentes impliquant la même forme réduite. Les deux équations sont identifiables.

Exercice: Calculez la forme réduite du système précédent. Pourquoi ne peut-on pas identifier les paramètres de la seconde équation structurelle lorsque $E(u_{1t}u_{2t}) \neq 0$?

2.3 La condition de rang

Lorsque les seules restrictions sont des restrictions linéaires homogènes portant sur les β_{ij} et γ_{ij}, jointes à des restrictions de normalisation ($\beta_{ij} = 1$ pour un seul j dans l'équation i), nous allons voir qu'il n'est pas nécessaire de passer par l'approche de la section précédente. Une condition nécessaire et suffisante pour l'identifiabilité d'une équation peut en effet être énoncée en fonction du rang d'une certaine matrice.

2.3.1 Formulation en fonction des coefficients de la forme réduite.

Comme $\Pi = -B^{-1}\Gamma$, nous pouvons énoncer la relation suivante, qui lie les paramètres de la forme structurelle à ceux de la forme réduite:

$$B\Pi + \Gamma = O_{g\times k}$$

soit aussi:

$$AW = O_{g\times k}$$

où:

$$A = (\, B \quad \Gamma \,) \quad \text{est} \quad g \times (g+k)$$

$$W = \begin{pmatrix} \Pi \\ \\ I_k \end{pmatrix} \quad \text{est} \quad (g+k) \times k \quad .$$

Soit alors α_i la i-ième ligne de A. Il s'agit du vecteur des coefficients de la i-ième équation structurelle. Le rang de W est égal à k. En effet, comme rang $(I_k) = k$, rang $(W) \geq k$; mais W n'a que k colonnes, donc rang $(W) \leq k$. Donc $\alpha_i W = O_{1\times k}$ est un système homogène de k équations indépendantes avec $g+k$ inconnues. L'ensemble des solutions est donc un espace vectoriel de dimension $(g+k) - k = g$.

Les restrictions homogènes devront ramener cette dimension à l'unité pour que l'équation i soit identifiable. Le vecteur α_i sera alors déterminé à un facteur de proportionnalité près et la restriction de normalisation permettra de le déterminer de façon unique.

Ces restrictions homogènes, au nombre de R_i, sont regroupées dans le système $\alpha_i \Phi_i = O_{1\times R_i}$. La matrice Φ_i a $g+k$ lignes et R_i colonnes. Au total, le système d'équations qui devrait nous permettre de retrouver les paramètres de la i-ième équation structurelle à partir des restrictions et des paramètres de la forme réduite est le suivant:

$$\alpha_i \, (\, W \quad \Phi_i \,) = O_{1\times(k+R_i)}$$

et le rang de $(\, W \quad \Phi_i \,)$ doit être égal à $g+k-1$ pour que toutes les solutions soient proportionnelles.

2.3.2 Formulation équivalente en fonction des coefficients de la forme structurelle.

Cette formulation est plus facile à utiliser que la précédente, car elle n'implique pas le calcul de Π.

Théorème.

Le rang de $(W \quad \Phi_i)$ *est égal à* $g + k - 1$ *si et seulement si le rang de* $A\Phi_i$ *est égal à* $g - 1$.

Démonstration:

Voir Judge et al., *The Theory and Practice of Econometrics*, 1985, p.577.

2.4 La condition d'ordre

Supposons maintenant que les seules restrictions homogènes soient des restrictions *d'exclusion* (du type $\beta_{ij} = 0$ ou $\gamma_{ij} = 0$). Nous pouvons alors énoncer un critère encore plus simple que le précédent. Il faut néanmoins insister sur le fait que ce critère est une condition nécessaire, mais pas suffisante, pour l'identification d'une équation. Si la condition d'ordre n'est pas vérifiée, l'équation n'est pas identifiable; si la condition d'ordre est satisfaite, il faut néanmoins vérifier la condition de rang.

Repartons de l'équation rang $(W \quad \Phi_i) = g+k-1$. Comme $(W \quad \Phi_i)$ a $k+R_i$ colonnes et $g+k$ lignes, cette condition ne sera certainement *pas* vérifiée si $R_i < g-1$; en effet, dans ce cas, rang $(W \quad \Phi_i) \leq k+R_i < k+g-1$. Une condition *nécessaire* pour l'identification de l'équation i est donc $R_i \geq g-1$. Comme les R_i restrictions sont des restrictions d'exclusion, on a:

$$R_i = g - g_i + k - k_i$$

où g_i et k_i sont les nombres de variables respectivement endogènes et prédéterminées *incluses* dans l'équation i. Il faut donc que:

$$R_i = g - g_i + k - k_i \geq g - 1$$

$$\text{soit} \qquad k - k_i \geq g_i - 1 \quad .$$

Cette dernière inégalité est la condition d'ordre.

Le nombre de variables prédéterminées exclues ne peut être inférieur au nombre de variables endogènes incluses moins 1.

Si $k - k_i = g_i - 1$, l'équation est dite juste-identifiée.

Si $k - k_i > g_i - 1$, l'équation est dite sur-identifiée.

2.5 Exemple

Reprenons le système récursif de la section 2.2. Nous allons voir que sans la restriction $\sigma_{12} = 0$, la première équation reste identifiable, mais la seconde ne l'est pas.

La matrice A s'écrit, en tenant compte des restrictions:

$$A = \begin{pmatrix} 1 & 0 & \gamma_{11} \\ \\ \beta_{21} & 1 & \gamma_{21} \end{pmatrix} \quad .$$

Pour la première équation, $\Phi_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$. Donc $A\Phi_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, qui est de rang $1 = g - 1$. La première équation est donc identifiable. Comme $k - k_1 = 0 = g_1 - 1 = 0$, elle est juste-identifiée.

Pour la seconde équation, $k - k_2 = 0 < g_2 - 1 = 1$. Cette équation n'est pas identifiable.

Exercice: Discutez l'identification des deux équations de l'exemple 2 de la section 1.1

CHAPITRE III.

MÉTHODES D'ESTIMATION À INFORMATION LIMITÉE

3.1 Introduction

Nous verrons dans ce chapitre la méthode des moindres carrés indirects, qui n'est applicable qu'à une équation juste-identifiée ($k - k_i = g_i - 1$); la méthode des moindres carrés doubles, qui est applicable à toute équation identifiable ($k - k_i \geq g_i - 1$); et l'estimateur de classe k, qui généralise celui des moindres carrés doubles et qui inclut aussi, comme cas particulier, l'estimateur par maximum de vraisemblance à information limitée. Le terme *information limitée* signifie que l'on ne tient compte, lors de l'estimation des coefficients de la i-ième équation structurelle, que des restrictions a priori sur cette équation (indépendamment de la formulation des autres équations). Les méthodes de cette classe ont donc l'avantage de la simplicité et de la robustesse. En revanche, les méthodes à information complète, que nous verrons au chapitre IV, sont potentiellement plus efficaces car elles utilisent les restrictions a priori sur toutes les équations du système.

L'estimateur de moindres carrés doubles, que nous verrons à la section 3.3, est l'estimateur à information limitée le plus couramment utilisé. C'est un estimateur par variables instrumentales, qui est asymptotiquement équivalent à celui du maximum de vraisemblance à information limitée.

3.2 Moindres carrés indirects

3.2.1 Présentation de la méthode.

Nous avons mentionné plus haut que les équations de la forme réduite $y_t = \Pi x_t + v_t$ pouvaient être estimées par moindres carrés ordinaires: on régresse chaque variable endogène sur toutes les variables prédéterminées présentes dans le modèle. Ceci fournit une estimation convergente de la matrice Π, soit $\hat{\Pi}$.

Si l'équation i est juste-identifiée, on peut en déduire des estimations convergentes des composantes de α_i en résolvant le système

$$\alpha_i \, (\, \hat{W} \quad \Phi_i \,) = O_{1\times(k+R_i)} \quad , \quad \text{où} \quad \hat{W} = \begin{pmatrix} \hat{\Pi} \\ \\ I_k \end{pmatrix} ,$$

et en imposant la condition de normalisation.

3.2.2 Limitations.

Montrons que cette procédure n'est pas applicable lorsque $R_i \neq g-1$. La matrice $(\hat{W} \quad \Phi_i)$ est de dimensions $(g+k) \times (k+R_i)$.

Si $R_i > g-1$, son rang sera de $g+k$ en général, même si rang $(W \quad \Phi_i) = g+k-1$. Nous avons donc $g+k$ équations indépendantes en $g+k$ variables. La solution unique est le vecteur nul, et cette solution est donc incompatible avec la condition de normalisation!

Si $R_i < g-1$, le rang de $(\hat{W} \quad \Phi_i)$ sera strictement inférieur à $k+g-1$, et nous aurons une infinité de solutions.

Illustrons ce qui précède au moyen de l'exemple suivant:

$$\begin{aligned} S_t &= a_0 + a_1 p_t + a_2 E_t + u_{1t} \\ p_t &= b_0 + b_1 S_t + b_2 r_t + b_3 p_{t-1} + u_{2t} \end{aligned}$$

où S_t est le taux de variation des salaires; p_t est le taux d'inflation; E_t est le taux de chômage; r_t est le taux d'intérêt.

Les deux variables endogènes sont p_t et S_t; les quatre variables prédéterminées sont la constante, E_t, r_t et p_{t-1}.

La matrice A a la forme suivante:

$$A = \begin{pmatrix} 1 & -a_1 & -a_0 & -a_2 & 0 & 0 \\ -b_1 & 1 & -b_0 & 0 & -b_2 & -b_3 \end{pmatrix} .$$

Les deux matrices Φ_1 et Φ_2 sont

$$\Phi_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \Phi_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} .$$

$$\text{Donc} \quad A\Phi_1 = \begin{pmatrix} 0 & 0 \\ -b_2 & -b_3 \end{pmatrix} \quad \text{et} \quad A\Phi_2 = \begin{pmatrix} -a_2 \\ 0 \end{pmatrix} .$$

Les deux matrices sont de rang 1, donc les deux équations sont identifiables. Pour la première équation, $k-k_1 = 2 > g_1 - 1 = 1$. Pour la seconde, $k-k_2 = 1 = g_2 - 1 = 1$. Donc la première équation est sur-identifiée, la seconde est juste-identifiée.

Nous résumons les données de l'échantillon dans la matrice des sommes de carrés et de produits suivante:

	S_t	p_t	Constante	E_t	r_t	p_{t-1}
S_t	361	100	10	20	80	80
p_t	100	279	80	10	60	40
Constante	10	80	100	0	0	0
E_t	20	10	0	20	0	0
r_t	80	60	0	0	40	0
p_{t-1}	80	40	0	0	0	80

Les paramètres de la forme réduite sont estimés par moindres carrés ordinaires. Donc:

$$\hat{\Pi} = \begin{pmatrix} 10 & 20 & 80 & 80 \\ 80 & 10 & 60 & 40 \end{pmatrix} \begin{pmatrix} \frac{1}{100} & 0 & 0 & 0 \\ 0 & \frac{1}{20} & 0 & 0 \\ 0 & 0 & \frac{1}{40} & 0 \\ 0 & 0 & 0 & \frac{1}{80} \end{pmatrix}$$

$$= \begin{pmatrix} 0.1 & 1 & 2 & 1 \\ 0.8 & 0.5 & 1.5 & 0.5 \end{pmatrix} .$$

Estimons les paramètres de la *seconde* équation structurelle par la méthode des moindres carrés indirects. Ces estimations sont obtenues en résolvant:

$$\begin{pmatrix} -b_1 & 1 & -b_0 & 0 & -b_2 & -b_3 \end{pmatrix} \begin{pmatrix} 0.1 & 1 & 2 & 1 \\ 0.8 & 0.5 & 1.5 & 0.5 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$$

ce qui implique $\hat{b}_0 = 0.75, \hat{b}_1 = 0.5, \hat{b}_2 = 0.5, \hat{b}_3 = 0$.

Si nous tentons de faire la même démarche pour la première équation, nous obtenons:

$$(\,1 \quad -a_1 \quad -a_0 \quad -a_2 \quad 0 \quad 0\,) \begin{pmatrix} 0.1 & 1 & 2 & 1 \\ 0.8 & 0.5 & 1.5 & 0.5 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = (\,0 \quad 0 \quad 0 \quad 0\,) \quad .$$

La troisième équation de ce système s'énonce comme $2 - 1.5\, a_1 = 0$, la quatrième comme $1 - 0.5\, a_1 = 0$. Ces deux équations sont incompatibles.

3.3 Moindres carrés doubles

Contrairement à la précédente, cette méthode peut être appliquée à toute équation identifiée. Nous fournirons deux interprétations de l'estimateur par moindres carrés doubles:

(1) une interprétation heuristique;
(2) une interprétation en termes de variables instrumentales;

3.3.1 Notation.

Supposons que nous voulions estimer les paramètres de la i-ième équation structurelle. Celle-ci peut s'écrire:

$$y_i = Y_i\beta_i + X_i\gamma_i + u_i$$

$$\text{ou} \quad y_i = T_i\delta_i + u_i \quad \text{avec} \quad T_i = (\,Y_i \quad X_i\,) \quad \text{et} \quad \delta_i = \begin{pmatrix} \beta_i \\ \\ \gamma_i \end{pmatrix} \quad .$$

y_i est le vecteur $n \times 1$ des observations sur la variable endogène dont le coefficient est normalisé à l'unité dans l'équation i;

Y_i est la matrice $n \times (g_i - 1)$ des observations sur les variables endogènes qui sont incluses comme régresseurs dans l'équation i;

X_i est la matrice $n \times k_i$ des observations sur les variables prédéterminées incluses dans l'équation i.

3.3.2 Premier exemple d'application.

Pour la forme structurelle de la section 1.1:

$$C_t = a + bY_t + u_{1t}$$

$$Y_t = C_t + I_t$$

nous avons calculé la forme réduite:

$$C_t = \Pi_{11} + \Pi_{12} I_t + v_{1t}$$

$$Y_t = \Pi_{21} + \Pi_{22} I_t + v_{2t}.$$

Si la matrice Π était connue, on pourrait calculer:

$$\tilde{Y}_t = \Pi_{21} + \Pi_{22} I_t.$$

Si I_t est non stochastique, $\tilde{Y}_t$ est non stochastique. On pourrait alors imaginer d'estimer par MCO les paramètres a et b dans l'équation modifiée:

$$C_t = a + b\tilde{Y}_t + w_t.$$

En fait, Π est inconnue. Mais on peut l'estimer de façon convergente par MCO, et calculer:

$$\hat{Y}_t = \hat{\Pi}_{21} + \hat{\Pi}_{22} I_t.$$

L'estimateur de a et b par moindres carrés doubles se calcule en appliquant les MCO à l'équation structurelle modifiée:

$$C_t = a + b\hat{Y}_t + e_t.$$

3.3.3 Présentation heuristique générale.

Cette présentation conduit aisément aux équations normales. Nous définirons l'estimateur de δ_i par moindres carrés doubles comme le vecteur obtenu en:

- régressant, par moindres carrés ordinaires, chacune des variables de Y_i sur toutes les variables prédéterminées du modèle, afin d'obtenir une matrice de valeurs calculées $\hat{Y}_i$;
- puis en remplaçant Y_i par $\hat{Y}_i$ dans l'équation $y_i = Y_i\beta_i + X_i\gamma_i + u_i$ et en appliquant une nouvelle fois les moindres carrés ordinaires à l'équation ainsi obtenue.

L'idée est donc la suivante:

Nous avons, en vertu de la forme réduite, l'égalité $Y = X\Pi^{'} + V$. Si Π était une matrice connue, le fait de remplacer la matrice Y par la matrice $X\Pi^{'}$ "purgerait" donc les variables endogènes de leur partie aléatoire. On pourrait alors appliquer les moindres carrés ordinaires à une équation structurelle où l'on aurait remplacé les composantes de Y_i par ces valeurs purgées, puisque ce sont ces parties aléatoires qui sont responsables du biais de simultanéité.

En pratique, bien sûr, Π est une matrice inconnue. Mais nous pouvons l'estimer de façon convergente, en appliquant les moindres carrés ordinaires à chaque équation de la forme réduite. Soit $\hat{\Pi}$ l'estimation obtenue.

Supposons, sans perte de généralité, que Y_i forme les premières colonnes de Y, et partageons la matrice $\hat{\Pi}^{'}$ de la façon suivante:

$$\hat{\Pi}^{'} = (\,\hat{\Pi}_i^{'} \quad \hat{\Pi}_0^{'}\,)$$

où $\hat{\Pi}_i^{'}$ est $k \times (g_i - 1)$ et $\hat{\Pi}_0^{'}$ est $k \times (g - (g_i - 1))$.

On voit directement que $\hat{Y}_i = X\hat{\Pi}_i^{'}$. Par ailleurs, $\hat{\Pi}_i^{'}$, étant obtenue par régression des colonnes de Y_i sur celles de la matrice X, est égale à $\hat{\Pi}_i^{'} = (X^{'}X)^{-1}X^{'}Y_i$. Donc $\hat{Y}_i = X(X^{'}X)^{-1}X^{'}Y_i$ est la matrice obtenue lors de la première étape de la méthode des moindres carrés doubles.

Pour la seconde étape, nous avons l'équation de régression $y_i = \hat{Y}_i\beta_i + X_i\gamma_i + \epsilon_i$, que nous pouvons aussi écrire $y_i = Z_i\delta_i + \epsilon_i$ avec $Z_i = (\,\hat{Y}_i \quad X_i\,)$. Les équations normales s'écrivent alors $(Z_i^{'}Z_i)\hat{\delta}_i = Z_i^{'}y_i$, soit:

$$(E.N.1) \qquad \begin{pmatrix} \hat{Y}_i^{'}\hat{Y}_i & \hat{Y}_i^{'}X_i \\ X_i^{'}\hat{Y}_i & X_i^{'}X_i \end{pmatrix} \begin{pmatrix} \hat{\beta}_i \\ \hat{\gamma}_i \end{pmatrix} = \begin{pmatrix} \hat{Y}_i^{'}y_i \\ X_i^{'}y_i \end{pmatrix} \quad .$$

3.3.4 Justification par les variables instrumentales.

Supposons, sans perte de généralité, que la matrice X_i forme les premières colonnes de X, et définissons $P_X = X(X'X)^{-1}X'$. On a $P_X X_i = X_i$, car $(X'X)^{-1}X'X_i$ forme les k_i premières colonnes d'une matrice unité d'ordre k. D'autre part $P_X Y_i = \hat{Y}_i$. On a alors:

$$Z_i = (\,\hat{Y}_i \quad X_i\,) = P_X\,(\,Y_i \quad X_i\,) = P_X T_i$$

et par conséquent:

$$\begin{aligned} \hat{\delta}_i &= (Z_i'Z_i)^{-1}Z_i'y_i \\ &= [(P_X T_i)'(P_X T_i)]^{-1}(P_X T_i)'y_i \\ &= [T_i'P_X T_i]^{-1}T_i'P_X y_i \\ &= [Z_i'T_i]^{-1}Z_i'y_i \end{aligned}$$

ou encore:

$$(E.N.2) \quad \begin{pmatrix} Y_i' X(X'X)^{-1}X'Y_i & Y_i' X_i \\ X_i' Y_i & X_i' X_i \end{pmatrix} \begin{pmatrix} \hat{\beta}_i \\ \hat{\gamma}_i \end{pmatrix} = \begin{pmatrix} Y_i' X(X'X)^{-1}X' y_i \\ X_i' y_i \end{pmatrix} \quad .$$

L'expression $[T_i' P_X T_i]^{-1} T_i' P_X y_i$ montre que l'on a bien un estimateur par variables instrumentales: les observations sur ces variables forment la matrice X. La convergence en probabilité de $\hat{\delta}_i$ vers δ_i est garantie par l'hypothèse H_6 de la section 1.3.

Il est intéressant de noter que $T_i' P_X T_i$ est d'ordre $k_i + g_i - 1$ et de rang inférieur ou égal à k. Donc si la condition d'ordre n'est pas vérifiée ($k - k_i < g_i - 1$), la matrice des coefficients des équations normales sera singulière.

3.3.5 Distribution asymptotique.

Puisque l'estimateur des moindres carrés doubles est un estimateur par variables instrumentales, le théorème 13.8 de la seconde partie lui est immédiatement applicable. Nous avons donc le résultat suivant.

Théorème.

Soit $\hat{\delta}_i$ l'estimateur de δ_i par moindres carrés doubles. Sous les hypothèses d'un théorème central limite:

(1) $\operatorname{dlim}\left(\sqrt{n}(\hat{\delta}_i - \delta_i)\right) \sim N(0, \sigma_{ii}\Sigma_{ZZ}^{-1})$ *où* $\Sigma_{ZZ} = \operatorname{plim}\left(\frac{1}{n} Z_i' Z_i\right)$.

(2) *Si* $\hat{\sigma}_{ii} = \frac{1}{n}(y_i - T_i\hat{\delta}_i)'(y_i - T_i\hat{\delta}_i)$, *alors* $\operatorname{plim}\hat{\sigma}_{ii} = \sigma_{ii}$.

Notons qu'il n'est pas nécessaire de calculer chaque résidu pour calculer $\hat{\sigma}_{ii}$. On vérifie en effet par simple substitution que:

$$\hat{\sigma}_{ii} = \frac{1}{n}\left\{ y_i' y_i - 2\hat{\delta}_i' \begin{pmatrix} Y_i' y_i \\ X_i' y_i \end{pmatrix} + \hat{\delta}_i' \begin{pmatrix} Y_i' Y_i & Y_i' X_i \\ X_i' Y_i & X_i' X_i \end{pmatrix} \hat{\delta}_i \right\} \quad .$$

3.3.6 Exemple numérique.

Reprenons maintenant l'exemple de la section 3.2.2. Pour la première équation, les observations sur la variable p_t forment la matrice Y_1; celles sur la constante et sur la variable E_t forment la matrice X_1. Le vecteur y_1 n'est autre que (S_t).

Construisons les équations normales à partir de (E.N.2). On obtient par simple lecture:

$$X^{'}Y_1 = \begin{pmatrix} 80 \\ 10 \\ 60 \\ 40 \end{pmatrix} \quad X_1^{'}X_1 = \begin{pmatrix} 100 & 0 \\ 0 & 20 \end{pmatrix}$$

$$X_1^{'}y_1 = \begin{pmatrix} 10 \\ 20 \end{pmatrix} \qquad X^{'}y_1 = \begin{pmatrix} 10 \\ 20 \\ 80 \\ 80 \end{pmatrix}$$

$$X^{'}X = \begin{pmatrix} 100 & 0 & 0 & 0 \\ 0 & 20 & 0 & 0 \\ 0 & 0 & 40 & 0 \\ 0 & 0 & 0 & 80 \end{pmatrix} \quad X_1^{'}Y_1 = \begin{pmatrix} 80 \\ 10 \end{pmatrix} \quad .$$

Par conséquent, $\hat{\beta}_1$ et $\hat{\gamma}_1$ sont la solution du système:

$$\begin{pmatrix} 179 & 80 & 10 \\ 80 & 100 & 0 \\ 10 & 0 & 20 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\gamma}_1 \end{pmatrix} = \begin{pmatrix} 178 \\ 10 \\ 20 \end{pmatrix} \quad .$$

Nous obtenons comme solution:

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\gamma}_1 \end{pmatrix} = \frac{1}{22000} \begin{pmatrix} 200 & -160 & -100 \\ -160 & 348 & 80 \\ -100 & 80 & 1150 \end{pmatrix} \begin{pmatrix} 178 \\ 10 \\ 20 \end{pmatrix}$$

$$= \begin{pmatrix} 32/22 \\ -234/220 \\ 6/22 \end{pmatrix} = \begin{pmatrix} \hat{a}_1 \\ \hat{a}_0 \\ \hat{a}_2 \end{pmatrix} \quad .$$

En ce qui concerne maintenant la seconde équation, les observations sur S_t forment la matrice Y_2; celles sur la constante, r_t et p_{t-1}, forment la matrice X_2; celles sur p_t forment le vecteur y_2. Nous avons alors:

$$X^{'}Y_2 = \begin{pmatrix} 10 \\ 20 \\ 80 \\ 80 \end{pmatrix} \quad X_2^{'}X_2 = \begin{pmatrix} 100 & 0 & 0 \\ 0 & 40 & 0 \\ 0 & 0 & 80 \end{pmatrix}$$

$$X_2^{'}y_2 = \begin{pmatrix} 80 \\ 60 \\ 40 \end{pmatrix} \quad X^{'}y_2 = \begin{pmatrix} 80 \\ 10 \\ 60 \\ 40 \end{pmatrix}$$

$$X_2^{'}Y_2 = \begin{pmatrix} 10 \\ 80 \\ 80 \end{pmatrix}$$

et les équations normales sont:

$$\begin{pmatrix} 261 & 10 & 80 & 80 \\ 10 & 100 & 0 & 0 \\ 80 & 0 & 40 & 0 \\ 80 & 0 & 0 & 80 \end{pmatrix} \begin{pmatrix} \hat{b}_1 \\ \hat{b}_0 \\ \hat{b}_2 \\ \hat{b}_3 \end{pmatrix} = \begin{pmatrix} 178 \\ 80 \\ 60 \\ 40 \end{pmatrix}$$

système dont la solution est $\hat{b}_0 = 0.75$, $\hat{b}_1 = 0.5$, $\hat{b}_2 = 0.5$, $\hat{b}_3 = 0$. Nous retombons sur les mêmes résultats que ceux obtenus par moindres carrés indirects! Ceci est dû au fait que l'équation 2 soit juste-identifiée. Cette propriété est générale, comme on peut le démontrer.

Estimons maintenant les variances asymptotiques des estimateurs $\hat{a}_0$, $\hat{a}_1$, $\hat{a}_2$. On a:

$$\hat{\sigma}_{11} = \frac{1}{100}\Bigg\{361 - 2\,(\,1.45 \quad -1.06 \quad 0.27\,) \begin{pmatrix} 100 \\ 10 \\ 20 \end{pmatrix} + (\,1.45 \quad -1.06 \quad 0.27\,) \begin{pmatrix} 279 & 80 & 10 \\ 80 & 100 & 0 \\ 10 & 0 & 20 \end{pmatrix} \begin{pmatrix} 1.45 \\ -1.06 \\ 0.27 \end{pmatrix} \Bigg\} = 5.4575$$

et les estimations des variances asymptotiques sont:

$$\hat{\sigma}^2_{\hat{a}_0} = 5.4575\left(\frac{348}{22000}\right) = 0.0863$$

$$\hat{\sigma}^2_{\hat{a}_1} = 0.0496$$

$$\hat{\sigma}^2_{\hat{a}_2} = 0.2853.$$

Comme:

$$\frac{\hat{a}_2}{\hat{\sigma}_{\hat{a}_2}} = \frac{6/22}{\sqrt{0.2853}} = 0.5106 < 1.96,$$

$\hat{a}_2$ n'est pas significativement différent de zéro.

3.4 L'estimateur de classe k

Il fut défini par H. Theil comme la solution $\begin{pmatrix} \hat{\beta}_i^k \\ \hat{\gamma}_i^k \end{pmatrix}$ des équations normales suivantes:

$$\begin{pmatrix} Y_i'Y_i - k\hat{V}_i'\hat{V}_i & Y_i'X_i \\ X_i'Y_i & X_i'X_i \end{pmatrix} \begin{pmatrix} \hat{\beta}_i^k \\ \hat{\gamma}_i^k \end{pmatrix} = \begin{pmatrix} \left(Y_i - k\hat{V}_i\right)' y_i \\ X_i'y_i \end{pmatrix} .$$

où $\hat{V}_i$ est une matrice de résidus de la forme réduite, définie comme:

$$\hat{V}_i = (I - X(X'X)^{-1}X')Y_i = MY_i$$

Si $k = 0$, nous avons l'estimateur obtenu par moindres carrés ordinaires appliqués à la i-ième équation structurelle.

Si $k = 1$, nous avons l'estimateur de moindres carrés doubles, comme on peut le voir facilement à partir des équations normales (E.N.2) puisque $P_X Y_i = Y_i - \hat{V}_i$ et puisque $Y_i'\hat{V}_i = \hat{V}_i'\hat{V}_i$.

Si k est aléatoire et $\operatorname{plim} k = 1$, nous avons un estimateur convergent. Si, en particulier, k est égal à la plus petite racine $\hat{\ell}$ d'une certaine équation déterminantale, on obtient l'estimateur de maximum de vraisemblance à information limitée; on peut prouver que $\operatorname{plim}\sqrt{n}(\hat{\ell} - 1) = 0$ (voir Judge et al., *The Theory and Practice of Econometrics*, p. 602).

CHAPITRE IV.

MÉTHODES D'ESTIMATION À INFORMATION COMPLÈTE

Nous estimons ici, globalement, les paramètres d'un système entier. Nous supposons que toute équation non identifiable, et toute identité, a été supprimée du système (les identités sont éliminées par substitution). Les méthodes de ce chapitre permettent, dans certains cas, un gain d'efficacité asymptotique.

4.1 Le produit de Kronecker et certaines de ses propriétés

Cette opération permet, dans le cadre des systèmes d'équations, l'élaboration d'une notation très compacte.

Si A est une matrice $m \times n$ et B est une matrice $p \times q$, $A \otimes B$ est la matrice $mp \times nq$ suivante:

$$A \otimes B \quad = \quad \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \dots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix} \quad .$$

Mentionnons quelques-unes des propriétés de ce produit.

4.1.1 Si $A = \begin{pmatrix} B & C \\ D & E \end{pmatrix}$, alors $A \otimes F = \begin{pmatrix} B \otimes F & C \otimes F \\ D \otimes F & E \otimes F \end{pmatrix}$.

Il n'y a pas de propriété analogue lorsque c'est la matrice F qui est partagée.

4.1.2 $(A \otimes B)' = A' \otimes B'$

4.1.3 $A \otimes (B + C) = A \otimes B + A \otimes C$

4.1.4 $(B + C) \otimes A = B \otimes A + C \otimes A$

4.1.5 $(A \otimes B) \otimes C = A \otimes (B \otimes C)$

4.1.6 $tr(A \otimes B) = (trA)(trB)$ si A et B sont carrées.

4.1.7 Si A est $m \times m$ et B est $n \times n$:
$\det(A \otimes B) = (\det A)^n (\det B)^m$

4.1.8 Si A et B sont régulières:
$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$

4.1.9 Si les produits AC et BD sont définis:
$(A \otimes B)(C \otimes D) = AC \otimes BD$.

4.2 L'opérateur de vectorisation et certaines de ses propriétés

Soit A une matrice $m \times n$ dont les colonnes sont les vecteurs a^i:

$$A = (\, a^1 \quad a^2 \quad \dots \quad a^n \,)$$

on définit:

$$\operatorname{vec} A = \begin{pmatrix} a^1 \\ a^2 \\ \vdots \\ a^n \end{pmatrix}$$

Le vecteur $\operatorname{vec} A$ est donc $mn \times 1$.

Les propriétés les plus importantes de cet opérateur sont les suivantes:

4.2.1 Si les matrices A, B, C sont conformes pour la multiplication, alors $\operatorname{vec}(ABC) = (C' \otimes A) \operatorname{vec} B$;

4.2.2 Si les matrices A et B sont conformes pour la multiplication et si AB est carrée, la trace de (AB) est égale à $(\operatorname{vec} A')' \operatorname{vec} B$.

Pour une étude approfondie des opérateurs $\otimes$ et vec et d'autres opérations matricielles avancées, on peut consulter Magnus et Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 1988.

4.3 Moindres carrés généralisés et forme réduite

Comme premier exemple d'application des deux opérateurs précédents, nous allons montrer que dans le cas d'une forme réduite, l'emploi des moindres carrés généralisés est équivalent à l'estimation par MCO de chaque équation individuelle.

Nous avons vu, à la section 1.4, que la forme réduite pouvait s'écrire:

$$Y = X\Pi' + V.$$

Comme $X\Pi' = X\Pi' I_g$, l'application de la règle 4.2.1 donne:

$$\operatorname{vec} Y = (I_g \otimes X) \operatorname{vec} \Pi' + \operatorname{vec} V.$$

Cette équation peut aussi s'écrire comme:

$$\mathcal{Y} = \mathcal{X}\beta + \mathcal{U}$$

avec:

$$\mathcal{Y} = \operatorname{vec} Y$$

$$\mathcal{X} = I_g \otimes X$$

$$\mathcal{U} = \operatorname{vec} V$$

$$\beta = \operatorname{vec} \Pi'.$$

On vérifie aisément que $E(\mathcal{U}) = O_{ng\times 1}$, et que la matrice de covariance $E(\mathcal{U}\mathcal{U}')$ est égale à $\Omega = \Sigma_v \otimes I_n$, où $\Sigma_v = B^{-1}\Sigma(B')^{-1}$ est la matrice de covariance contemporaine des erreurs de la forme réduite.

Mais $\Sigma_v \otimes I_n$ n'est pas diagonale. Nous avons un cas particulier du modèle traité à la section 8.2.3 de la seconde partie. Pourquoi, alors, peut-on estimer les équations de ce modèle par moindres carrés ordinaires et non par moindres carrés généralisés? Ceci vient du fait que les régresseurs soient les mêmes dans chaque équation ($\mathcal{X} = I_g \otimes X$). Nous allons vérifier, à l'aide des propriétés des deux sections précédentes, que la formule des MCG se simplifie:

$$\begin{aligned}
\operatorname{vec}\hat{\Pi}' = \hat{\beta} &= (\mathcal{X}'\Omega^{-1}\mathcal{X})^{-1}\mathcal{X}'\Omega^{-1}\mathcal{Y} \\
&= [(I_g \otimes X)'(\Sigma_v \otimes I_n)^{-1}(I_g \otimes X)]^{-1}[I_g \otimes X]'(\Sigma_v \otimes I_n)^{-1}\mathcal{Y} \\
&= [(I_g \otimes X)'(\Sigma_v^{-1} \otimes I_n)(I_g \otimes X)]^{-1}[I_g \otimes X]'(\Sigma_v^{-1} \otimes I_n)\mathcal{Y} \\
&= [\Sigma_v^{-1} \otimes (X'X)]^{-1}[\Sigma_v^{-1} \otimes X']\mathcal{Y} \\
&= [\Sigma_v \otimes (X'X)^{-1}][\Sigma_v^{-1} \otimes X']\mathcal{Y} \\
&= [I_g \otimes (X'X)^{-1}X']\mathcal{Y} \\
&= \begin{pmatrix} (X'X)^{-1}X' & O & \dots & O \\ O & (X'X)^{-1}X' & \dots & O \\ \vdots & \vdots & \vdots & \vdots \\ O & O & \dots & (X'X)^{-1}X' \end{pmatrix} \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^g \end{pmatrix}
\end{aligned}$$

4.4 Moindres carrés triples

4.4.1 Présentation heuristique.

La méthode des moindres carrés doubles revient à estimer δ_i dans l'équation $X^{'}y_i = (X^{'}T_i)\delta_i + X^{'}u_i$ par moindres carrés généralisés. Si nous regroupons les g équations de ce type, nous obtenons:

$$\begin{pmatrix} X^{'}y_1 \\ X^{'}y_2 \\ \vdots \\ X^{'}y_g \end{pmatrix} = \begin{pmatrix} X^{'}T_1 & O & \dots & O \\ O & X^{'}T_2 & \dots & O \\ \vdots & \vdots & \vdots & \vdots \\ O & O & \dots & X^{'}T_g \end{pmatrix} \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_g \end{pmatrix} + \begin{pmatrix} X^{'}u_1 \\ \vdots \\ X^{'}u_g \end{pmatrix} \quad .$$

soit aussi:

$$\mathcal{Y} = \mathcal{X}\delta + \mathcal{U}$$

où $\mathcal{Y}$ est $gk \times 1$, et $\mathcal{X}$ est $gk \times \sum_{i=1}^{g}(k_i + g_i - 1)$.

En ce qui concerne les erreurs $\mathcal{U}$, on a, sous l'hypothèse simplificatrice que X est non stochastique, $E(\mathcal{U}) = 0$, et:

$$\begin{aligned} E(\mathcal{U}\mathcal{U}^{'}) &= \Sigma \otimes (X^{'}X) \\ &= \begin{pmatrix} \sigma_{11}(X^{'}X) & \sigma_{12}(X^{'}X) & \dots & \sigma_{1g}(X^{'}X) \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{g1}(X^{'}X) & \sigma_{g2}(X^{'}X) & \dots & \sigma_{gg}(X^{'}X) \end{pmatrix} \quad . \end{aligned}$$

La méthode des moindres carrés triples s'énonce alors comme suit:

(1) On applique les moindres carrés doubles à chaque équation individuelle. Ceci donne, pour l'équation i, un vecteur de résidus $\hat{u}_i = y_i - T_i\hat{\delta}_i$.
(2) Soit $\hat{U} = (\,\hat{u}_1 \quad \dots\dots \quad \hat{u}_g\,)$. La matrice Σ est estimée par $S = \frac{1}{n}\hat{U}^{'}\hat{U}$.
(3) On applique enfin la formule de Aitken au système précédent pour obtenir $\hat{\delta}$. Ceci donne:

$$\hat{\delta} = \{\mathcal{X}^{'}[S^{-1} \otimes (X^{'}X)^{-1}]\mathcal{X}\}^{-1}\mathcal{X}^{'}[S^{-1} \otimes (X^{'}X)^{-1}]\mathcal{Y} \quad .$$

Si l'élément (i,j) de S^{-1} est noté s^{ij}, on vérifie facilement que:

$$\hat{\delta} = \begin{pmatrix} s^{11}A_{11} & \dots & s^{1g}A_{1g} \\ \vdots & \vdots & \vdots \\ s^{g1}A_{g1} & \dots & s^{gg}A_{gg} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^{g} s^{1j}T_1^{'}X(X^{'}X)^{-1}X^{'}y_j \\ \vdots \\ \sum_{j=1}^{g} s^{gj}T_g^{'}X(X^{'}X)^{-1}X^{'}y_j \end{pmatrix}$$

où $A_{ij} = T_i^{'} X(X^{'}X)^{-1}X^{'}T_j$.

4.4.2 Justification par les variables instrumentales.

Définissons:

$$\mathcal{T} = \begin{pmatrix} T_1 & O & \dots & O \\ O & T_2 & \dots & O \\ \vdots & \vdots & \vdots & \vdots \\ O & O & \dots & T_g \end{pmatrix}$$

$$z = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_g \end{pmatrix}$$

$$\delta = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_g \end{pmatrix}$$

$$u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_g \end{pmatrix}.$$

Le système des g équations structurelles peut alors s'écrire:

$$z = \mathcal{T}\delta + u.$$

On vérifie aisément que la matrice $\mathcal{X}$ et le vecteur $\mathcal{Y}$ de la section 4.4.1 peuvent s'écrire:

$$\mathcal{X} = (I_g \otimes X')\mathcal{T}$$

$$\mathcal{Y} = (I_g \otimes X')z$$

En substituant ces expressions dans:

$$\hat{\delta} = \{\mathcal{X}^{'}[S^{-1} \otimes (X^{'}X)^{-1}]\mathcal{X}\}^{-1}\mathcal{X}^{'}[S^{-1} \otimes (X^{'}X)^{-1}]\mathcal{Y}$$

on obtient après simplification:

$$\hat{\delta} = [\mathcal{T}'(S^{-1} \otimes P_X)\mathcal{T}]^{-1}\mathcal{T}'(S^{-1} \otimes P_X)z$$

avec $P_X = X(X'X)^{-1}X'$.

Nous avons donc bien un estimateur par variables instrumentales; les instruments forment la matrice $(S^{-1} \otimes P_X)\mathcal{T}$.

Vérifions que ces instruments vérifient bien la propriété du lemme 13.6 de la seconde partie. Le vecteur $\operatorname{plim} \frac{1}{n} Z'u$ prend ici la forme:

$$\operatorname{plim} \frac{1}{n} \mathcal{T}'(S^{-1} \otimes P_X)u$$

vecteur dont les sous-vecteurs prennent la forme:

$$\operatorname{plim} \frac{1}{n} \sum_j s^{ij} T_i' X (X'X)^{-1} X' u_j =$$

$$\operatorname{plim} \sum_j s^{ij} \left(\frac{1}{n} T_i' X\right) \left(\frac{1}{n} X'X\right)^{-1} \frac{1}{n} X' u_j =$$

$$\sum_j s^{ij} \left(\operatorname{plim} \frac{1}{n} T_i' X\right) \left(\operatorname{plim} \frac{1}{n} X'X\right)^{-1} \operatorname{plim} \frac{1}{n} X' u_j = 0$$

en vertu de l'hypothèse H_6 de la section 1.3.

4.4.3 Comparaison avec les moindres carrés doubles.

Il est facile de vérifier que si l'on applique les moindres carrés doubles à chaque équation du système, on obtient l'estimateur:

$$\hat{\delta}^0 = [\mathcal{T}'(I_g \otimes P_X)\mathcal{T}]^{-1} \mathcal{T}'(I_g \otimes P_X)z$$

Donc, dans ce cas, les instruments forment la matrice $(I_g \otimes P_X)\mathcal{T}$, au lieu de $(S^{-1} \otimes P_X)\mathcal{T}$ dans le cas des moindres carrés triples. Si Σ^{-1} n'est pas diagonale, les moindres carrés triples utilisent plus d'information que les moindres carrés doubles, et sont donc potentiellement plus efficaces.

Trois remarques peuvent être faites:

(1) Si l'on impose la contrainte $\sigma_{ij} = 0$, $i \neq j$, S et S^{-1} sont diagonales. $\hat{\delta}$ est alors identique à l'estimateur obtenu en appliquant les moindres carrés doubles à chaque équation du système: il n'y a aucun gain d'efficacité.

(2) Si *chaque* équation du système est juste-identifiée, $\hat{\delta}$ est identique à l'estimateur obtenu en appliquant les moindres carrés indirects à chaque équation. On obtiendra aussi des résultats identiques en appliquant les moindres carrés doubles à chaque équation. Il n'y a donc gain d'efficacité que lorsque l'une, au moins, des équations est suridentifiée.

(3) Enfin, si le système ne comprend qu'une seule équation de comportement, les moindres carrés triples sont bien entendu équivalents aux moindres carrés doubles.

4.4.4 Distribution asymptotique.

L'estimateur par moindres carrés triples, nous l'avons montré, est un estimateur par variables instrumentales. Il est donc convergent, asymptotiquement sans biais, et asymptotiquement normal. A l'encontre de l'estimateur par moindres carrés doubles, il est de plus asymptotiquement efficace.

Théorème. *Soit $\hat{\delta}$ l'estimateur de δ par moindres carrés triples, et soit $\hat{\delta}^0$ l'estimateur de δ obtenu en appliquant les moindres carrés doubles à chaque équation.*

Sous les hypothèse d'un théorème central limite:

(1) $\operatorname{plim} \hat{\delta} = \delta$
(2) $\operatorname{dlim} \sqrt{n}(\hat{\delta} - \delta) \sim N(0, Q)$ *où:*

$$Q = \operatorname{plim} n[\mathcal{T}'(\Sigma^{-1} \otimes P_X)\mathcal{T}]^{-1}$$

(3) $\operatorname{plim} S^{-1} = \Sigma^{-1}$, *où S a été précédemment définie.*
(4) *Si Q^0 est la matrice de covariance asymptotique de $\sqrt{n}(\hat{\delta}^0 - \delta)$, alors:*
$Q^0 = Q + B$, *où B est définie non négative.*

Nous allons justifier ce théorème au moyen d'un argument par analogie. A la section 13.3.3 de la seconde partie, nous avions trouvé la matrice de covariance asymptotique:

$$V = \operatorname{plim} n\hat{\sigma}^2 (Z'X)^{-1} Z'Z (X'Z)^{-1}.$$

Cette matrice peut aussi s'écrire:

$$V = \operatorname{plim} n(Z'X)^{-1} V(Z'u \mid Z)(X'Z)^{-1}.$$

Dans le cas qui nous occupe, Z doit être remplacé par $(\Sigma^{-1} \otimes P_X)\mathcal{T}$, et X doit être remplacé par $\mathcal{T}$. De plus, nous avons $E(uu' \mid Z) = \Sigma \otimes I_n$ au lieu de $E(uu' \mid Z) = \sigma^2 I$. Par conséquent, $V(Z'u \mid Z)$ devient:

$$\begin{aligned} E[\mathcal{T}'(\Sigma^{-1} \otimes P_X)uu'(\Sigma^{-1} \otimes P_X)\mathcal{T}) \mid Z] &= \mathcal{T}'(\Sigma^{-1} \otimes P_X)(\Sigma \otimes I)(\Sigma^{-1} \otimes P_X)\mathcal{T} \\ &= \mathcal{T}'(\Sigma^{-1} \otimes P_X)\mathcal{T} \end{aligned}$$

En faisant ces remplacements dans l'expression de V et en simplifiant, on obtient:

$$Q = \operatorname{plim} n[\mathcal{T}'(\Sigma^{-1} \otimes P_X)\mathcal{T}]^{-1}$$

qui est identique à la matrice de covariance de l'énoncé.

4.4.5 Exemple numérique.

Appliquons la méthode précédente au modèle de la section 3.2. Il nous faut d'abord calculer

$$S = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_{22} \end{pmatrix}$$

La variance $\hat{\sigma}_{11}$ a été calculée à la section 3.3.6 ($\hat{\sigma}_{11} = 5.4575$). On obtient de même:

$$\hat{\sigma}_{22} = \frac{1}{100}\Bigg\{279 - 2\,(0.5 \quad 0.75 \quad 0.5 \quad 0)\begin{pmatrix} 100 \\ 80 \\ 60 \\ 40 \end{pmatrix} + (0.5 \quad 0.75 \quad 0.5 \quad 0)\begin{pmatrix} 361 & 10 & 80 & 80 \\ 10 & 100 & 0 & 0 \\ 80 & 0 & 40 & 0 \\ 80 & 0 & 0 & 80 \end{pmatrix}\begin{pmatrix} 0.5 \\ 0.75 \\ 0.5 \\ 0 \end{pmatrix}\Bigg\} = 2.03$$

$$\hat{\sigma}_{12} = \frac{1}{100}\Bigg\{100 - (1.45 \quad -1.06 \quad 0.27)\begin{pmatrix} 279 \\ 80 \\ 10 \end{pmatrix} - (0.5 \quad 0.75 \quad 0.5 \quad 0)\begin{pmatrix} 361 \\ 10 \\ 80 \\ 80 \end{pmatrix} + (1.45 \quad -1.06 \quad 0.27)\begin{pmatrix} 100 & 80 & 60 & 40 \\ 10 & 100 & 0 & 0 \\ 20 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 0.5 \\ 0.75 \\ 0.5 \\ 0 \end{pmatrix}\Bigg\} = -3.3018.$$

Les blocs:

$$T_1^{'}X(X^{'}X)^{-1}X^{'}T_1, \quad T_2^{'}X(X^{'}X)^{-1}X^{'}T_2, \quad T_1^{'}X(X^{'}X)^{-1}X^{'}y_1 \quad T_2^{'}X(X^{'}X)^{-1}X^{'}y_2$$

ont également été calculés à la section 3.3.6. Il reste à trouver:

$$T_2^{'}X(X^{'}X)^{-1}X^{'}T_1, \quad T_1^{'}X(X^{'}X)^{-1}X^{'}y_2, \quad T_2^{'}X(X^{'}X)^{-1}X^{'}y_1.$$

Nous avons:

$$T_1^{'}X = \begin{pmatrix} Y_1^{'}X \\ X_1^{'}X \end{pmatrix} = \begin{pmatrix} 80 & 10 & 60 & 40 \\ 100 & 0 & 0 & 0 \\ 0 & 20 & 0 & 0 \end{pmatrix}$$

$$T_2^{'}X = \begin{pmatrix} Y_2^{'}X \\ X_2^{'}X \end{pmatrix} = \begin{pmatrix} 10 & 20 & 80 & 80 \\ 100 & 0 & 0 & 0 \\ 0 & 0 & 40 & 0 \\ 0 & 0 & 0 & 80 \end{pmatrix}$$

$$X^{'}y_1 = \begin{pmatrix} 10 \\ 20 \\ 80 \\ 80 \end{pmatrix} \quad X^{'}y_2 = \begin{pmatrix} 80 \\ 10 \\ 60 \\ 40 \end{pmatrix}$$

Il est facile alors de vérifier que:

$$T_2^{'}X(X^{'}X)^{-1}X^{'}T_1 = \begin{pmatrix} 178 & 10 & 20 \\ 80 & 100 & 0 \\ 60 & 0 & 0 \\ 40 & 0 & 0 \end{pmatrix}$$

$$T_2^{'}X(X^{'}X)^{-1}X^{'}y_1 = \begin{pmatrix} 261 \\ 10 \\ 80 \\ 80 \end{pmatrix}$$

$$T_1^{'}X(X^{'}X)^{-1}X^{'}y_2 = \begin{pmatrix} 179 \\ 80 \\ 10 \end{pmatrix}$$

Les équations normales des moindres carrés triples s'écrivent alors:

$$\begin{pmatrix} 11.484 \begin{pmatrix} 179 & 80 & 10 \\ 80 & 100 & 0 \\ 10 & 0 & 20 \end{pmatrix} & 18.679 \begin{pmatrix} 178 & 80 & 60 & 40 \\ 10 & 100 & 0 & 0 \\ 20 & 0 & 0 & 0 \end{pmatrix} \\ 18.679 \begin{pmatrix} 178 & 10 & 20 \\ 80 & 100 & 0 \\ 60 & 0 & 0 \\ 40 & 0 & 0 \end{pmatrix} & 30.875 \begin{pmatrix} 261 & 10 & 80 & 80 \\ 10 & 100 & 0 & 0 \\ 80 & 0 & 40 & 0 \\ 80 & 0 & 0 & 80 \end{pmatrix} \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_0 \\ \hat{a}_2 \\ \hat{b}_1 \\ \hat{b}_0 \\ \hat{b}_2 \\ \hat{b}_3 \end{pmatrix} =$$

$$\begin{pmatrix} 11.484 \begin{pmatrix} 178 \\ 10 \\ 20 \end{pmatrix} + 18.679 \begin{pmatrix} 179 \\ 80 \\ 10 \end{pmatrix} \\ 18.679 \begin{pmatrix} 261 \\ 10 \\ 80 \\ 80 \end{pmatrix} + 30.875 \begin{pmatrix} 178 \\ 80 \\ 60 \\ 40 \end{pmatrix} \end{pmatrix}$$

La solution de ce système, conduit au vecteur de paramètres suivant:

$$\hat{\delta} = \begin{pmatrix} 1.4545 \\ -1.0636 \\ 0.2727 \\ 0.5 \\ 0.75 \\ 0.39 \\ 0.165 \end{pmatrix}$$

et à la matrice de covariance asymptotique estimée:

$$\begin{pmatrix} 0.0496 & -0.0397 & -0.0248 & 0 & 0 & -0.045 & -0.015 \\ -0.0397 & 0.0863 & 0.0198 & 0 & -0.033 & 0.036 & 0.012 \\ -0.0248 & 0.0198 & 0.2853 & -0.1651 & 0.0165 & 0.3527 & 0.1726 \\ 0 & 0 & -0.1651 & 0.1015 & -0.0101 & -0.203 & -0.1015 \\ 0 & -0.033 & 0.0165 & -0.0101 & 0.0213 & 0.0203 & 0.0101 \\ -0.045 & 0.036 & 0.3527 & -0.203 & 0.0203 & 0.4477 & 0.2166 \\ -0.015 & 0.012 & 0.1726 & -0.1015 & 0.0101 & 0.2166 & 0.1064 \end{pmatrix}$$

4.5 Maximum de vraisemblance à information complète

Cette méthode est la première en date de toutes celles que nous avons vues. C'est aussi la plus coûteuse à appliquer, et, pour cette raison, la moins employée. Son intérêt théorique est néanmoins très grand: en vertu des propriétés des estimateurs par maximum de vraisemblance, les estimateurs obtenus sont convergents, asymptotiquement sans biais, et asymptotiquement efficaces. En fait, en vertu d'un théorème d'équivalence asymptotique, nous pourrons justifier rigoureusement l'emploi de la méthode des moindres carrés triples par le biais du maximum de vraisemblance.

4.5.1 La vraisemblance logarithmique.

La forme structurelle s'écrit:

$$YB^{'} + X\Gamma^{'} = U$$

et la t-ième ligne $u_t^{'}$ de U est un vecteur aléatoire satisfaisant $u_t^{'} \sim N(0,\, \Sigma)$. Les autres hypothèses de ce chapitre restent inchangées.

La densité jointe de l'un des vecteurs u_t s'écrit:

$$f_u\left(u_t\right) = (2\pi)^{-g/2} \left(\det \Sigma\right)^{-1/2} \exp\left(-\frac{1}{2} u_t^{'} \Sigma^{-1} u_t\right)$$

Les y_t et les u_t sont liés par la relation $By_t + \Gamma x_t = u_t$. Donc la matrice jacobienne $\frac{\partial u_t}{\partial y_t} = B^{'}$, et en vertu du théorème de la section 2.2 de la première partie, nous pouvons écrire la densité de y_t conditionnelle à x_t comme:

$$f_t(y_t) = f_u(By_t + \Gamma x_t) \mid \det B^{'} \mid = \; f_u(By_t + \Gamma x_t) \mid \det B \mid \quad .$$

Par conséquent, la densité des variables endogènes conditionnelle aux variables exogènes s'écrit comme:

$$f_Y(y_1,\ldots,y_n) = \prod_{t=1}^{n} f_t(y_t) =$$
$$(2\pi)^{-ng/2}(\det\Sigma)^{-n/2} \mid \det B \mid^n \exp\left[-\frac{1}{2}\sum_{t=1}^{n}(By_t+\Gamma x_t)^{'}\Sigma^{-1}(By_t+\Gamma x_t)\right]$$

ou, puisque:

$$\sum_{t=1}^{n} u_t^{'}\Sigma^{-1}u_t = tr\, U\Sigma^{-1}U^{'} = tr\,\Sigma^{-1}U^{'}U :$$

$$f_Y(y_1,\ldots,y_n) =$$
$$(2\pi)^{-ng/2}(\det\Sigma)^{-n/2} \mid \det B \mid^n \exp\left[-\frac{1}{2}tr\,\Sigma^{-1}\left(YB^{'}+X\Gamma^{'}\right)^{'}\left(YB^{'}+X\Gamma^{'}\right)\right].$$

Pour obtenir la vraisemblance logarithmique, on prend le logarithme de cette expression considérée comme fonction de B, Γ, et Σ:

$$\log L(B,\Gamma,\Sigma) =$$
$$k - \frac{n}{2}\log(\det\Sigma) + n\log(\mid \det B \mid) - \frac{1}{2}tr\,\Sigma^{-1}\left(YB^{'}+X\Gamma^{'}\right)^{'}\left(YB^{'}+X\Gamma^{'}\right)$$

ou encore:

$$\log L = k + \frac{n}{2}\log\left(\det\Sigma^{-1}\right) + n\log(\mid \det B \mid)$$
$$- \frac{1}{2}tr\,\Sigma^{-1}BY^{'}YB^{'} - \frac{1}{2}tr\,\Sigma^{-1}\Gamma X^{'}YB^{'} - \frac{1}{2}tr\,\Sigma^{-1}BY^{'}X\Gamma^{'} - \frac{1}{2}tr\,\Sigma^{-1}\Gamma X^{'}X\Gamma^{'}.$$

4.5.2 Les conditions de premier ordre.

Pour trouver les dérivées, nous notons que:

$$\frac{1}{2}tr\,\Sigma^{-1}\Gamma X^{'}YB^{'} + \frac{1}{2}tr\,\Sigma^{-1}BY^{'}X\Gamma^{'} = tr\,BY^{'}X\Gamma^{'}\Sigma^{-1} = tr\,\Gamma X^{'}YB^{'}\Sigma^{-1} \quad .$$

et nous utilisons les formules suivantes (voir Magnus et Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 1988):

$$\frac{\partial \log\left(\mid \det A \mid\right)}{\partial A} = \left(A^{'}\right)^{-1}$$

$$\frac{\partial}{\partial A} tr\, AC = C^{'}$$

$$\frac{\partial}{\partial A} tr\, DACA^{'} = 2DAC \quad \text{si } D \text{ et } C \text{ sont symétriques.}$$

Par conséquent:

$$\begin{aligned}
\frac{\partial \log L}{\partial \Sigma^{-1}} &= \frac{n}{2}\Sigma - \frac{1}{2}\left(YB^{'} + X\Gamma^{'}\right)^{'}\left(YB^{'} + X\Gamma^{'}\right) = O \\
\frac{\partial \log L}{\partial B} &= n\left(B^{'}\right)^{-1} - \Sigma^{-1}BY^{'}Y - \Sigma^{-1}\Gamma X^{'}Y = O \\
\frac{\partial \log L}{\partial \Gamma} &= -\Sigma^{-1}BY^{'}X - \Sigma^{-1}\Gamma X^{'}X = O \quad .
\end{aligned}$$

On peut écrire ces expressions de manière plus condensée comme:

$$\begin{aligned}
\hat{\Sigma} &= \frac{1}{n}\hat{U}^{'}\hat{U} \\
\left(\hat{B}^{'}\right)^{-1} &= \frac{1}{n}\hat{\Sigma}^{-1}\hat{U}^{'}Y \\
\hat{\Sigma}^{-1}\hat{U}^{'}X &= O \\
\text{avec} \quad \hat{U} &= Y\hat{B}^{'} + X\hat{\Gamma}^{'} \quad .
\end{aligned}$$

Ce système est non linéaire, et doit être résolu par des méthodes numériques. Pour qu'il ait une solution unique, on doit lui ajouter les restrictions d'identification. Il faut noter que la formule de $\hat{\Sigma}$ est précisément celle que nous avons employée en moindres carrés triples. D'autre part, la troisième équation est impliquée par $\hat{U}^{'}X = O$, équation que nous pouvons mettre en parallèle avec les équations normales du modèle de régression classique, qui peuvent s'écrire $X^{'}\hat{u} = 0$.

CHAPITRE V.

ANALYSE STATISTIQUE DE LA FORME RÉDUITE (RÉGRESSION MULTIVARIÉE)

5.1 Estimation par maximum de vraisemblance

Il est facile, à partir des résultats de la section 4.5, de trouver les estimateurs par maximum de vraisemblance des paramètres de la forme réduite. En effet, la forme réduite est un cas particulier de la forme structurelle lorsque l'on impose $\hat{B} = I_g$, et qu'il n'y a pas de restrictions a priori sur la matrice Γ.

Les conditions de premier ordre de la section 4.5.2 s'écrivent alors:

$$\hat{\Sigma}^{-1}\hat{U}'X = O_{g\times k}$$

$$\hat{\Sigma} = \frac{1}{n}\hat{U}'\hat{U}$$

Il est facile de vérifier que les estimateurs:

$$\hat{\Gamma} = -\hat{\Pi} = -Y'X(X'X)^{-1}$$

$$\hat{\Sigma} = \frac{1}{n}(Y'[I - X(X'X)^{-1}X']Y)$$

satisfont bien à ces conditions.

En effet, si nous définissons $M = [I - X(X'X)^{-1}X']$, nous avons, en utilisant les estimateurs de B et de Γ, la matrice de résidus suivante:

$$\hat{U} = YI_g + X\hat{\Gamma}' = Y - X(X'X)^{-1}X'Y = MY.$$

La matrice M est symétrique et idempotente, et vérifie $M'X = O$. Il s'ensuit donc que $\hat{U}'X = O$ et que $\hat{U}'\hat{U} = Y'MY$, ce qui implique bien les conditions de premier ordre.

Nous allons maintenant estimer les variances des coefficients de régression de la forme réduite. Nous pouvons écrire:

$$\hat{\Pi}' = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\Pi' + V) = \Pi' + (X'X)^{-1}X'V.$$

Par conséquent:

$$\operatorname{vec}(\hat{\Pi}' - \Pi') = \operatorname{vec}[(X'X)^{-1}X'V] = [I_g \otimes (X'X)^{-1}X'] \operatorname{vec} V.$$

Si nous supposons, pour simplifier l'argument, que X est non stochastique, la matrice de covariance de $\operatorname{vec} \hat{\Pi}'$ s'écrit:

$$\begin{aligned} E\{(\operatorname{vec}[\hat{\Pi}' - \Pi'])(\operatorname{vec}[\hat{\Pi}' - \Pi'])'\} &= [I_g \otimes (X'X)^{-1}X'] E(\operatorname{vec} V \operatorname{vec}' V)[I_g \otimes X(X'X)^{-1}] \\ &= [I_g \otimes (X'X)^{-1}X'][\Sigma \otimes I_n][I_g \otimes X(X'X)^{-1}] \\ &= [\Sigma \otimes (X'X)^{-1}(X'X)(X'X)^{-1}] \\ &= [\Sigma \otimes (X'X)^{-1}] \end{aligned}$$

et l'on peut donc estimer la matrice de covariance par:

$$\hat{V}(\operatorname{vec} \hat{\Pi}') = \hat{\Sigma} \otimes (X'X)^{-1}.$$

Si X est stochastique, on peut utiliser la même règle d'estimation mais son interprétation est asymptotique. La justification utilise les mêmes arguments qu'aux chapitres XIII et XIV de la seconde partie.

Exercice: Soit la forme réduite suivante, où l'on a 2 équations et 3 variables prédéterminées:

$$\begin{aligned} y_{1t} &= \Pi_{11} + \Pi_{12}x_{1t} + \Pi_{13}x_{2t} + v_{1t} \\ y_{2t} &= \Pi_{21} + \Pi_{22}x_{1t} + \Pi_{23}x_{2t} + v_{2t}. \end{aligned}$$

Formulez la statistique de Wald pour le test de $H_0 : \Pi_{13} = \Pi_{22}$ contre $H_1 : \Pi_{13} \neq \Pi_{22}$.

Note:

Pour le calcul du rapport des vraisemblances, nous devrons, à la section suivante, diviser par $\det \hat{\Sigma}$. Il est donc intéressant de connaître des conditions nécessaires pour la régularité de $\hat{\Sigma}$.

On a vu que $\hat{\Sigma} = Y'MY/n$ avec $M = I - X(X'X)^{-1}X'$. $\hat{\Sigma}$ est d'ordre g et M est de rang $n - k$. Donc si $n - k < g$, $\hat{\Sigma}$ est singulière. Le nombre d'observations doit être supérieur à la somme du nombre de régresseurs par équation et du nombre d'équations.

5.2 Tests d'hypothèses sur les coefficients par le rapport des vraisemblances

Comme à la section précédente, nous pouvons formuler la vraisemblance de la forme réduite comme un cas particulier de celle de la forme structurelle; cette dernière vraisemblance a été vue à la section 4.5. Si nous posons $\Gamma' = -\Pi'$, $B = I$, et $U = V$, nous obtenons:

$$L(\Pi, \Sigma) = (2\pi)^{-ng/2}(\det \Sigma)^{-n/2} \exp[-\frac{1}{2} tr\Sigma^{-1}(Y - X\Pi')'(Y - X\Pi')].$$

Si nous définissons $\hat{V} = Y - X\hat{\Pi}'$, la vraisemblance maximisée s'écrit:

$$L(\hat{\Pi}, \hat{\Sigma}) = (2\pi)^{-ng/2}(\det \hat{\Sigma})^{-n/2} \exp[-\frac{1}{2} tr\hat{\Sigma}^{-1}\hat{V}'\hat{V}].$$

On peut simplifier cette expression en notant que $\hat{V}'\hat{V} = n\hat{\Sigma}$, et que donc:

$$tr\hat{\Sigma}^{-1}\hat{V}'\hat{V} = tr\hat{\Sigma}^{-1}(n\hat{\Sigma}) = ng.$$

Par conséquent:

$$L(\hat{\Pi}, \hat{\Sigma}) = (2\pi)^{-ng/2}(\det \hat{\Sigma})^{-n/2} \exp\left(-\frac{ng}{2}\right).$$

Considérons alors la partition suivante des colonnes de Π:

$$\Pi = (\ \Pi_* \quad \Pi_{**}\ )$$

et le test de l'hypothèse:

$$H_0 : \Pi_* = \Pi_*^0 \quad \text{contre} \quad H_1 : \Pi_* \neq \Pi_*^0.$$

Un exemple de ce test est celui où $\Pi_*^0 = O$: dans ce cas, on teste l'omission des premières variables explicatives de la forme réduite. Si nous désignons par $\hat{\Pi}_0$ et $\hat{\Sigma}_0$ les estimations contraintes de Π et de Σ, le rapport des vraisemblances peut s'écrire:

$$\begin{aligned}
\lambda &= \frac{L(\hat{\Pi}_0, \hat{\Sigma}_0)}{L(\hat{\Pi}, \hat{\Sigma})} \\
&= \frac{(2\pi)^{-ng/2}(\det \hat{\Sigma}_0)^{-n/2} \exp\left(-\frac{ng}{2}\right)}{(2\pi)^{-ng/2}(\det \hat{\Sigma})^{-n/2} \exp\left(-\frac{ng}{2}\right)} \\
&= \left(\frac{\det \hat{\Sigma}_0}{\det \hat{\Sigma}}\right)^{-n/2}.
\end{aligned}$$

Nous obtenons donc une généralisation de l'expression démontrée à la section 7.2 de la seconde partie: au lieu d'avoir des variances estimées, on a des déterminants de matrices de covariances (qui portent aussi le nom de variances généralisées).

En vertu du théorème de la section 10.12 de la seconde partie, la distribution limite sous H_0 de $-2\log\lambda$ est une $\chi^2_{(p)}$, où p est le nombre d'éléments de Π_*. Mais dans ce cas-ci, on n'a pas, *en général*, une transformation monotone de λ ayant une distribution F sous H_0 en petit échantillon. La situation est donc différente de celle que nous avons rencontrée au chapitre VII de la seconde partie.

On a constaté, notamment à l'aide d'études de simulation, que l'emploi des valeurs critiques asymptotiques (celles de la χ^2) conduit, en petit échantillon, à un rejet trop fréquent de l'hypothèse nulle, même si celle-ci est vraie. Ceci signifie que les valeurs critiques exactes de $-2\log\lambda$ sont supérieures à celles de la χ^2 si n est faible.

Anderson (*An Introduction to Multivariate Statistical Analysis*, 1984) propose la correction suivante, qui n'est basée sur une argumentation théorique rigoureuse que lorsque X est non stochastique. Mais des études de simulation ont montré que cette correction donnait de bons résultats en général, même lorsque le modèle comporte des variables endogènes retardées. Au lieu de $-2\log\lambda$, on utilise $\gamma(-2\log\lambda)$, où le facteur de correction γ est défini comme:

$$\gamma = \frac{n - q_2 - \frac{1}{2}(g + q_1 + 1)}{n}$$

où q_1 est le nombre de colonnes de Π_* et où $q_2 = k - q_1$. On compare cette statistique à la valeur critique d'une χ^2 ayant $p = gq_1$ degrés de liberté. Si X est non stochastique, l'erreur d'approximation est d'ordre n^{-2}.

Il est possible de montrer que cette correction est analogue à celle qui consiste à utiliser, dans la définition de la statistique t, l'estimateur sans biais de la variance des erreurs au lieu de l'estimateur par maximum de vraisemblance.

5.3 Forme réduite dérivée

Si, au lieu d'estimer Π par $\hat{\Pi} = Y'X(X'X)^{-1}$, on utilise:

$$\tilde{\Pi} = -\hat{B}^{-1}\hat{\Gamma}$$

où $\hat{B}$ et $\hat{\Gamma}$ ont été calculées par l'une des méthodes d'estimation de la forme structurelle (MCD, MCT, MVIL, ou MVIC), on parle de *forme réduite dérivée*. Si chaque équation est juste-identifiée, $\tilde{\Pi} = \hat{\Pi}$; mais si tel n'est pas le cas, $\tilde{\Pi}$ est potentiellement plus efficace que $\hat{\Pi}$ car il tient compte de plus de restrictions.

Les méthodes d'estimation de la forme structurelle permettent d'estimer les variances asymptotiques des éléments de $\hat{B}$ et $\hat{\Gamma}$, mais $\tilde{\Pi}$ est une fonction non linéaire de ces matrices. Dans cette section, nous allons donc énoncer un théorème permettant d'estimer les variances des éléments de $\tilde{\Pi}$. Des versions de ce théorème sont énoncées dans Monfort, *Cours de Probabilité*, p. 166 et dans Hamilton, *Time Series Analysis*, p. 186. Il peut bien

sûr aussi servir dans d'autres contextes, chaque fois que l'on veut faire un test d'hypothèses sur une fonction non linéaire de paramètres; une application courante est le test des restrictions de facteurs communs, que nous avons rencontrées au chapitre XV de la seconde partie.

Théorème. *Soit θ un vecteur de paramètres inconnus et soit $\hat{\theta}$ son estimateur. Supposons que:*

(1)
$$\operatorname{dlim} \sqrt{n}(\hat{\theta} - \theta_0) \sim N(0, \Omega)$$

(2) *La fonction $g(\theta) \in \mathbb{R}^m$ ait toutes ses dérivées partielles continues*

(3) *La matrice jacobienne:*

$$\nabla g = \begin{pmatrix} \frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_1}{\partial \theta_k} \\ \vdots & \vdots & \vdots \\ \frac{\partial g_m}{\partial \theta_1} & \cdots & \frac{\partial g_m}{\partial \theta_k} \end{pmatrix}_{\theta=\theta_0} \quad \textit{soit de rang } m$$

alors:

$$\operatorname{dlim} \sqrt{n}(g(\hat{\theta}) - g(\theta_0)) \sim N(0, (\nabla g)\Omega(\nabla g)')$$

Comme exemple, nous allons estimer la variance asymptotique de l'un des coefficients de la forme réduite du modèle de Haavelmo. Nous avons vu à la section 1.6 que la première équation de cette forme réduite pouvait s'écrire comme $C_t = \Pi_{11} + \Pi_{12} I_t + v_{1t}$, avec $\Pi_{11} = a/(1-b)$. Supposons que a et b aient été estimés par $\hat{a}$ et $\hat{b}$, et que leurs variances et leur covariance asymptotiques aient été estimées par $\hat{\sigma}^2_{\hat{a}}$, $\hat{\sigma}^2_{\hat{b}}$, et $\hat{\sigma}_{\hat{a}\hat{b}}$. L'application du théorème précédent à $\tilde{\Pi}_{11} = \hat{a}/(1-\hat{b})$ donne alors:

$$\hat{V}(\tilde{\Pi}_{11}) = \frac{1}{(1-\hat{b})^2}\hat{\sigma}^2_{\hat{a}} + \frac{\hat{a}^2}{(1-\hat{b})^4}\hat{\sigma}^2_{\hat{b}} + 2\frac{\hat{a}}{(1-\hat{b})^3}\hat{\sigma}_{\hat{a}\hat{b}}.$$

Exercice. Reprenez l'exemple de la section 15.2 de la seconde partie, portant sur les restrictions de facteurs communs. Comment testeriez-vous l'hypothèse $H_0 : \gamma_{11} + \phi_1\gamma_{01} = 0$ contre $H_1 : \gamma_{11} + \phi_1\gamma_{01} \neq 0$?

CHAPITRE VI.

COMPARAISON DES MOINDRES CARRÉS TRIPLES ET DU MAXIMUM DE VRAISEMBLANCE À INFORMATION COMPLÈTE

Nous allons montrer dans ce chapitre que les estimateurs MCT et MVIC ont la même distribution limite normale, et sont par conséquent asymptotiquement équivalents. L'estimateur MCT hérite donc des propriétés d'efficacité asymptotique de la méthode du maximum de vraisemblance.

En fait, comme nous le verrons, l'estimateur MVIC peut être considéré comme un estimateur par variables instrumentales, mais ces variables sont construites à l'aide de la forme réduite dérivée au lieu de l'être par la forme réduite directe.

Les développements de ce chapitre sont dus à Hausman ("An instrumental variable approach to full information estimators for linear and certain nonlinear econometric models", *Econometrica* 43, 1975, pp. 727–738).

6.1 Reformulation des équations normales des moindres carrés triples

Nous avons vu, à la section 4.4.2, que si l'on réunissait les n observations sur les g équations de la forme structurelle, on pouvait écrire, en tenant compte des restrictions de normalisation et d'exclusion:

$$z = \mathcal{T}\delta + u$$

où $\mathcal{T}$ était une matrice diagonale par blocs, avec des blocs diagonaux donnés par les matrices $T_i = (\, Y_i \quad X_i \,)$ définies à la section 3.3.1.

L'estimateur MCT pouvait s'écrire comme:

$$\hat{\delta} = (Z'\mathcal{T})^{-1} Z'z$$

avec $Z = (S^{-1} \otimes P_X)\mathcal{T}$. P_X était égale à $X(X'X)^{-1}X'$ et S était l'estimateur de Σ obtenu en appliquant les moindres carrés doubles à chaque équation séparément.

La matrice Z peut être obtenue en supprimant de la matrice suivante:

$$\begin{aligned} Z_* &= (S^{-1} \otimes P_X)[I_g \otimes (\, Y \quad X \,)] \\ &= S^{-1} \otimes P_X \,(\, Y \quad X \,) \end{aligned}$$

les colonnes qui correspondent aux restrictions d'exclusion et de normalisation.

Considérons alors le système suivant:

$$(Z_*'\mathcal{T})\hat{\delta} = Z_*'z. \tag{1}$$

On peut écrire ce système sous la forme:

$$W'\hat{U}S^{-1} = O_{(k+g)\times g} \tag{2}$$

où:

$$W = P_X \begin{pmatrix} Y & X \end{pmatrix}$$

et où:

$$\operatorname{vec}\hat{U} = z - \mathcal{T}\hat{\delta}.$$

En effet, l'égalité (2) implique:

$$\operatorname{vec}(W'\hat{U}S^{-1}) = (S^{-1} \otimes W')\operatorname{vec}\hat{U} = 0$$

ce qui est bien équivalent à l'égalité (1), en vertu de la définition de Z_*.

On peut obtenir l'estimateur MCT en supprimant, dans le système (1), les équations qui correspondent aux restrictions de normalisation et d'exclusion (puisque les équations de ce système correspondent à des colonnes de Z_*). De même, on peut obtenir l'estimateur MCT en sélectionnant, dans l'égalité matricielle (2), les éléments qui correspondent aux éléments non contraints de la matrice $\begin{pmatrix} B' \\ \Gamma' \end{pmatrix}$.

6.2 Reformulation des conditions de premier ordre du maximum de vraisemblance à information complète

La contribution fondamentale de Hausman a été de noter que les conditions de premier ordre du maximum de vraisemblance, que nous avons vues à la section 4.5.2, pouvaient s'écrire sous une forme analogue à l'équation (2) de la section précédente, à savoir:

$$\tilde{W}'\hat{U}\hat{\Sigma}^{-1} = O_{(k+g)\times g}$$

ce qui permet la comparaison des deux méthodes d'estimation. Nous allons démontrer ce résultat.

Tout d'abord, la condition de premier ordre sur Σ peut s'écrire:

$$nI_g = \hat{U}'\hat{U}\hat{\Sigma}^{-1}. \tag{a}$$

Ensuite, la condition de premier ordre sur B peut s'écrire:

$$\hat{B}^{-1}(nI_g) = Y'\hat{U}\hat{\Sigma}^{-1}. \tag{b}$$

En combinant (a) et (b), il vient:

$$\hat{B}^{-1}\hat{U}'\hat{U}\hat{\Sigma}^{-1} - Y'\hat{U}\hat{\Sigma}^{-1} = O$$

ce qui implique, puisque $\hat{U}' = \hat{B}Y' + \hat{\Gamma}X'$:

$$\hat{B}^{-1}(\hat{B}Y' + \hat{\Gamma}X')\hat{U}\hat{\Sigma}^{-1} - Y'\hat{U}\hat{\Sigma}^{-1} = O$$

soit aussi, en développant:

$$\hat{B}^{-1}\hat{B}Y'\hat{U}\hat{\Sigma}^{-1} + \hat{B}^{-1}\hat{\Gamma}X'\hat{U}\hat{\Sigma}^{-1} - Y'\hat{U}\hat{\Sigma}^{-1} = O$$

et en simplifiant:

(c) $$\hat{B}^{-1}\hat{\Gamma}X'\hat{U}\hat{\Sigma}^{-1} = O.$$

Enfin, la condition de premier ordre sur Γ implique:

(d) $$-X'\hat{U}\hat{\Sigma}^{-1} = O.$$

En regroupant (c) et (d) et en changeant de signe, il vient:

$$\begin{pmatrix} -\hat{B}^{-1}\hat{\Gamma}X' \\ X' \end{pmatrix} \hat{U}\hat{\Sigma}^{-1} = O$$

ce qui montre que l'on a bien $\tilde{W}'\hat{U}\hat{\Sigma}^{-1} = O$, avec:

$$\tilde{W} = (-X(\hat{B}^{-1}\hat{\Gamma})' \quad X)$$

.

6.3 Comparaison des deux nouvelles formulations

La comparaison avec les MCT est alors immédiate, si l'on note que la matrice W de la section 6.1 pouvait s'écrire comme:

$$W = P_X (Y \quad X) = (P_X Y \quad X) = (X\hat{\Pi}' \quad X)$$

avec $\hat{\Pi}' = (X'X)^{-1}X'Y$, tandis que la matrice $\tilde{W}$ de la section 6.2 peut s'écrire:

$$\tilde{W} = (X\tilde{\Pi}' \qquad X)$$

avec $\tilde{\Pi} = -\hat{B}^{-1}\hat{\Gamma}$. Pour former les instruments, les MCT utilisent la forme réduite directe, tandis que le MVIC utilise la forme réduite dérivée.

En d'autres termes, les MCT utilisent les instruments:

$$(S^{-1} \otimes I_n) \begin{pmatrix} P_X T_1 & O & \dots & O \\ O & P_X T_2 & \dots & O \\ \vdots & \vdots & \vdots & \vdots \\ O & O & \dots & P_X T_g \end{pmatrix}$$

avec $P_X T_i = (\, X\hat{\Pi}_i' \quad X_i \,)$; tandis que le MVIC utilise les instruments:

$$(\hat{\Sigma}^{-1} \otimes I_n) \begin{pmatrix} \tilde{T}_1 & O & \dots & O \\ O & \tilde{T}_2 & \dots & O \\ \vdots & \vdots & \vdots & \vdots \\ O & O & \dots & \tilde{T}_g \end{pmatrix}$$

avec $\tilde{T}_i = (\, X\tilde{\Pi}_i' \quad X_i \,)$.

6.4 Conséquences

On peut déduire facilement de ce qui précède l'équivalence asymptotique des MCT et des MVIC. En effet, comme les estimateurs sont convergents:

$$\operatorname{plim} \hat{\Pi}_i = \operatorname{plim} \tilde{\Pi}_i = \Pi_i$$

$$\operatorname{plim} S = \operatorname{plim} \hat{\Sigma} = \Sigma$$

et les matrices de covariance asymptotiques sont donc les mêmes en vertu du théorème de Slutsky.

Or, sous l'hypothèse d'un théorème central limite, les distributions limites des estimateurs MCT et MVIC sont normales multivariées. Elles sont donc entièrement caractérisées par leurs espérances et leurs matrices de covariance.

Donc les distributions limites sont les mêmes; ceci constitue la meilleure justification théorique possible de la méthode des MCT, qui est plus facile à mettre en oeuvre que celle du MVIC.

CHAPITRE VII.

MÉTHODES NUMÉRIQUES DE MAXIMISATION DE LA VRAISEMBLANCE

Pour une excellente présentation de ces méthodes, le lecteur pourra consulter l'article de synthèse de R. Quandt, "Computational problems and methods", dans: *Handbook of Econometrics vol. I* (1983), édité par Griliches et Intriligator, pp. 699–764. Nous nous bornerons ici à parler des méthodes les plus courantes.

7.1 Méthode de Newton-Raphson

L'idée de base de cette méthode est de définir une suite d'approximations quadratiques de la vraisemblance. En maximisant successivement chacune de ces approximations, on espère converger vers un maximum de la vraisemblance. L'approximation quadratique à l'itération k se fait autour du maximum de l'approximation utilisée à l'itération $k-1$.

Soit donc θ un vecteur $k \times 1$ de paramètres à estimer et soit θ_0 une valeur de θ. Soit $\mathcal{L}(\theta) = \log L(\theta)$ la vraisemblance logarithmique. Nous écrivons le gradient de $\mathcal{L}$ comme:

$$g(\theta) = \frac{\partial \mathcal{L}}{\partial \theta}$$

et la matrice Hessienne de $\mathcal{L}$ comme:

$$H(\theta) = \frac{\partial^2 \mathcal{L}}{\partial \theta \partial \theta'}$$

.

Une approximation quadratique de $\mathcal{L}(\theta)$ autour de θ_0 est donnée par:

$$L^*_{\theta_0}(\theta) = \mathcal{L}(\theta_0) + g'(\theta_0)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)' H(\theta_0)(\theta - \theta_0)$$

En vertu des règles de la section 3.4 de la seconde partie, les conditions de premier ordre pour la maximisation de cette approximation sont données par:

$$\frac{\partial L^*}{\partial \theta} = g(\theta_0) + H(\theta_0)(\theta - \theta_0) = 0$$

ce qui implique:

$$\hat{\theta} = \theta_0 - H^{-1}(\theta_0) g(\theta_0).$$

La méthode de Newton-Raphson est une application récurrente de cette règle, à savoir:

$$\theta_{k+1} = \theta_k - H^{-1}(\theta_k) g(\theta_k)$$

7.2 Méthodes quasi-Newton

La méthode précédente a plusieurs limitations. La matrice Hessienne $H(\theta_k)$ peut ne pas être définie négative pour certaines valeurs des paramètres. Elle est souvent difficile à calculer. Enfin, la règle de la fin de la section précédente implique souvent un déplacement trop important, surtout lorsque l'on est proche du maximum.

Pour ces raisons, il est utile de généraliser cette règle. Si l'on définit A_k comme une *approximation* de $H^{-1}(\theta_k)$, g_k comme $g(\theta_k)$, est d_k comme $-A_k g_k$, une telle généralisation est la suivante:

$$\theta_{k+1} = \theta_k + \lambda_k d_k$$

où λ_k est un scalaire positif qui maximise la fonction *d'une seule variable* suivante:

$$F(\lambda_k) = \mathcal{L}(\theta_k + \lambda_k d_k)$$

Le vecteur d_k définit donc la direction dans laquelle on se déplace et λ_k est l'amplitude du déplacement dans la direction d_k.

On peut noter que $g_k' d_k$ est la dérivée de $\mathcal{L}(\theta_k + \lambda_k d_k)$ par rapport à λ_k. Comme $g_k' d_k = -g_k' A_k g_k$, cette dérivée sera positive si A_k est définie négative. Si A_k est l'inverse de la Hessienne et si $\mathcal{L}$ est concave, un accroissement marginal de λ_k aura donc pour effet d'augmenter la vraisemblance.

De nombreuses méthodes empiriques ont été proposées pour choisir A_k. Dans les sections suivantes, nous passerons en revue celle du score et celle de Davidon-Fletcher-Powell, qui sont parmi les plus employées.

7.3 Méthode du score

On remplace ici la matrice Hessienne par son espérance, et définit donc:

$$A_k = \left[E\left(\frac{\partial^2 \mathcal{L}}{\partial\theta\partial\theta'} \right)_{\theta=\theta_k} \right]^{-1}.$$

A_k est donc l'opposée de l'inverse de la matrice d'information, que nous avions définie à la section 10.10 de la seconde partie comme:

$$R(\theta) = -E\left(\frac{\partial^2 \mathcal{L}}{\partial\theta\partial\theta'} \right) = E\left(\frac{\partial \mathcal{L}}{\partial\theta} \frac{\partial \mathcal{L}}{\partial\theta'} \right)$$

Les avantages de cette méthode sont les suivants:

(1) La matrice d'information est d'ordinaire d'expression plus simple que la Hessienne;
(2) Une matrice d'information régulière est définie positive, même si la vraisemblance n'est pas localement concave; A_k est alors définie négative, ce qui est nécessaire pour la convergence de l'algorithme comme nous l'avons vu;
(3) Au point stationnaire, la Hessienne de $\mathcal{L}$ est en général égale à $-R(\theta)$ (voir la dérivation de $R(\theta)$ dans le modèle de régression multiple, vue à la section 10.10 de la seconde partie); lorsque l'on s'approche de l'optimum, la méthode du score devient donc pratiquement équivalente à celle de Newton-Raphson;
(4) A la convergence de l'algorithme, la matrice $-A_k$ est une estimation de la matrice de covariance asymptotique de $\hat{\theta}$ (voir la section 10.11 de la seconde partie).

7.4 Méthode de Davidon, Fletcher, Powell

On utilise ici la règle de récurrence suivante:

$$A_{k+1} = A_k + \frac{(\Delta\theta_k)(\Delta\theta_k)'}{(\Delta\theta_k)'(\Delta g_k)} - \frac{1}{(\Delta g_k)' A_k (\Delta g_k)} [A_k(\Delta g_k)(\Delta g_k)' A_k]$$

avec la condition initiale $A_0 = -I$ et où g_k est le gradient de $\mathcal{L}$ évalué à l'itération précédente.

On démontre que sous certaines conditions, la suite de matrices définie par cette règle converge vers l'inverse de la Hessienne de $\mathcal{L}$.

Cette méthode ne nécessite que le calcul des dérivées premières de $\mathcal{L}$, et est donc commode lorsque la matrice d'information est difficile à calculer.

7.5 Choix de l'amplitude du déplacement

On peut calculer λ_k par balayage, mais la procédure est coûteuse. Une solution plus opérationnelle est la suivante:

(1) On choisit un nombre $\epsilon \in]0, \frac{1}{2}[$.
(2) On choisit $\lambda_k > 0$ tel que:

$$\epsilon \leq \frac{\mathcal{L}(\theta_k + \lambda_k d_k) - \mathcal{L}(\theta_k)}{\lambda_k g_k' d_k} \leq 1 - \epsilon.$$

En d'autres termes, on choisit une solution approchée de l'équation:

$$f(\lambda_k) = \frac{\mathcal{L}(\theta_k + \lambda_k d_k) - \mathcal{L}(\theta_k)}{\lambda_k g_k' d_k} = \frac{1}{2}.$$

Cette solution existe toujours, pour autant que $g_k' d_k$ soit strictement positif et que $\mathcal{L}$ soit bornée supérieurement. Il est en effet facile de montrer que:

$$\lim_{\lambda_k \to \infty} f(\lambda_k) \leq 0$$

et, à l'aide de la règle de L'Hôpital, que:

$$\lim_{\lambda_k \to 0} f(\lambda_k) = 1.$$

La procédure que nous venons de décrire a deux avantages:

(1) L'inégalité de gauche, qui implique $f(\lambda_k) > 0$, garantit un accroissement de $\mathcal{L}$ à chaque itération, car $\lambda_k g_k' d_k > 0$;
(2) L'inégalité de droite, qui implique $f(\lambda_k) < 1$, empêche λ_k de tendre vers 0, ce qui impliquerait $\theta_{k+1} = \theta_k$.