Introduction à l'économétrie

(STAT-D-308)

ECTS: 5 (2, 2, 1)

Année 2013-2014

Catherine Dehon

Bâtiment R42 - bureau R42.6.204

e-mail: cdehon@ulb.ac.be

Tél.: (02) 6503858

Université libre de Bruxelles

Assistants: Alice McCathie et Alexander Wolf

3ème année - Bachelier en ingénieur de gestion

AVERTISSEMENT

Ce syllabus a été rédigé dans le but de faciliter la prise de notes pendant le cours théorique.
La mise à jour du présent syllabus sera faite via le cours théorique.
Il est bien entendu que l'examen portera sur l'ensemble de la matière vue au cours théorique (des élements pourraient être ajoutés oralement au cours) ainsi que la matière des travaux pratiques.
NB: Le texte en italique provient des slides du livre: Stock, J. H., Watson, M. W. (2007), *Introduction to Econometrics*, Addison Wesley.

A savoir ....

- **Buts du cours:**

1. Mettre en pratique des modèles économétriques dans des situations réelles.
2. Etre capable de vérifier les hypothèses posées dans un modèle.
3. Maîtriser un logiciel statistique.

- **Méthode d'enseignement et support:**

Théorie : Cours ex cathedra. Syllabus de théorie contenant la copie des transparents projetés (et commentés) au cours disponible sur le site: http://www.ulb.ac.be/soco/statrope/.

Exercices: 5 séances d'exercices sont organisées en auditoire et 6 séances en salle informatique.

- **Méthode d'évaluation:**

Un quart de la note finale est basée sur la réalisation et la présentation de 2 devoirs donnés durant le premier quadrimestre. La défense orale des devoirs est organisée durant le mois de décembre. La réalisation des devoirs est une étape obligatoire pour pouvoir présenter l'examen écrit. L'examen écrit est organisé durant la session de janvier. Aucune note personnelle n'est autorisée.

- **Report de note:**

Les règles de la Faculté SBS-EM sont strictement appliquées pour les reports de session en session et d'année en année. La seule variante se situe pour les point obtenus pour les travaux. La note des travaux est conservée pour la seconde session avec la même pondération.

• PLAN DU COURS

1. Introduction

2. Modèle de régression linéaire multiple

3. Série chronologique

4. Modèles pour variable dépendante dichotomique (modèles logit et probit, etc)

5. Modèles pour variable dépendante de comptage (modèle de Poisson)

6. Modèles de régression avec variable dépendante censurée (modèle Tobit)

• REFERENCES

- Dehon, Droesbeke & Vermandele (2008), *Eléments de statistique*, Editions de l'Université de Bruxelles
- Greene (2007), *Econometric Analysis*, Prentice Hall, London.
- Maddala & Watson (2001), *Introduction to Econometrics*, John Wiley & Sons, London.
- Mélard (2007), *Méthodes de Pévision à Court Terme*, Editions de l'Université de Bruxelles
- Stock & Watson (2007), *Introduction to Econometrics*, Addison Wesley. (source des notations en italiques).
- Verbeek (2008), *A Guide to Modern Econometrics*, Wiley-Interscience, New York.

Chapitre 1

INTRODUCTION

Economics suggests interesting relations, often with policy implications, but virtually never suggests quantitative magnitudes of causal effects:

- *What is the price elasticity of cigarettes?*
- *What is the effect of reducing class size on student achievement?*
- *What is the effect on earnings of a year of education?*
- *What is the effect on output growth of a 1 percentage point increase in interest rates by the Fed?*

But: utilisation d'outils statistiques pour

- Mesurer, quantifier les liens de causalité
- Tester la théorie économique sous-jacente
- Utiliser ces relations pour faire de la prévision
- Minimiser l'incertain, l'erreur.

Problème: dans la plupart des situations, nous ne pouvons pas réaliser une expérience (comme en chimie ou physique) où l'environnement serait sous contrôle....

Mais nous devons utiliser des données observées (nonexperimental data). Ces données posent plusieurs problèmes:

- *confounding effects (omitted factors)*
- *simultaneous causality*
- *correlation does not imply causation.*

1.1 EXEMPLE EMPIRIQUE

Class size and educational output: "Policy question: What is the effect of reducing class size by one student per class?"

What is the right output measure ("dependent variable")?

- *parent satisfaction*
- *student personal development*
- *future adult welfare and/or earnings*
- *performance on standardized tests*

Question: What do data say about the class size/test score relation?

The California Test Score Data Set.

Individuals:

All California school districts (n = 420)

Variables:

•Y: *Test scores combined math and reading (district average)*

•X: *Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teacher*

Quelques statistiques:

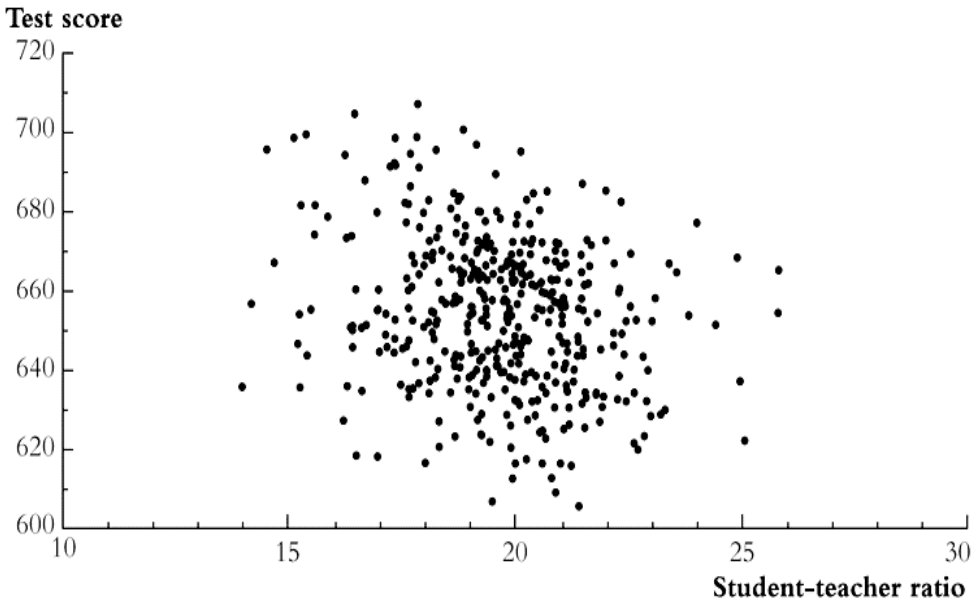
Statistics	Mean	Std	$x_{1/10}$	$x_{1/4}$	$x_{1/2}$	$x_{3/4}$	$x_{9/10}$
STR	19.6	1.9	17.3	18.6	19.7	20.9	21.9
Test score	654.2	19.1	630.4	640.0	654.5	666.7	679.1

Scatter Plot

Do districts with smaller classes have higher test scores?

FIGURE 4.2 Scatterplot of Test Score vs. Student-Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student-teacher ratio and test scores: the sample correlation is –0.23.



How can we get some numerical evidence on whether districts with low STRs have higher test scores?

- Compare average test scores in districts with low STR to those with high STR (estimation)
- Test the hypothesis that the mean test scores in the two types of districts are the same, against the alternative hypothesis that they differ (hypothesis testing)

1. ESTIMATION

Comparaison des résultats dans les districts où $STR < 20$ et dans ceux où $STR \geq 20$.

Class size	$\bar{Y}$	s_Y	n
Small	657.4	19.4	238
Large	650.0	17.9	182

2. TESTS D'HYPOTHESE

Δ: $\neq$ entre les 2 moyennes populations.

- Problème de test: H_0: $\Delta = 0$

$$H_1\text{: } \Delta \neq 0$$

- Statistique de test: $t = \dfrac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = 4.05$
- Loi asymptotique sous H_0: $t \approx N(0,1)$
- Règle de comportement: RH_0 au niveau $\alpha =$ 5% car $|t| > 1.96$
- Conclusion: les moyennes des scores dans les 2 populations sont différentes.

Original policy question: What is the effect on test scores of reducing STR by one student/class?

Have we answered this question?

- *We examined Δ the difference in means, small v. large classes*
- *But Δ doesn't really answer the policy question.*
- *Rather, the question is about $\frac{\Delta TestScore}{\Delta STR}$*
- *But this is the slope of a line relating test score and STR*
- *So somehow we need to estimate this slope ...*

Chapitre 2

MODELE DE REGRESSION LINEAIRE MULTIPLE

Relation sous forme d'équation entre une variable Y à expliquer et une ou plusieurs variables X qui seront les facteurs (variables explicatives):

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

où ε est un terme d'erreur représentant:

- des erreurs de mesures
- des effets non prévisibles
- des variables omises, ...

La fonction linéaire est la plus simple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

où les $\beta_0, \ldots, \beta_p$ sont les paramètres de régression

But: estimer les paramètres de régression inconnus sur base d'un échantillon.

ETAPES EN ANALYSE DE REGRESSION

- Formuler le problème
- Choisir les variables (réponse et explicatives)
- Collecter les données
- Spécifier le modèle (choix de la fonction f)
- Choisir la méthode d'estimation (OLS, GMM, MLE, ...)
- Estimer le modèle
- Valider le modèle
- Conclusions et/ou prévisions.

2.1 REGRESSION LINEAIRE SIMPLE

Objectif : Définir une relation de dépendance (de causalité) statistique entre 2 variables.
La variable à expliquer sera notée Y (variable réponse, variable dépendante), et la variable explicative sera noté X (variable indépendante, facteur)

Dépendance simple : relation linéaire
$\Longrightarrow$ détermination d'une droite de régression.

2.1.1 Estimateurs - Cadre théorique

Soit $\{(x_i, y_i); i = 1, \ldots, n\}$ une série statistique bivariée. Soit y la variable dépendante et x la variable explicative.

Modèle théorique à estimer:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Comment trouver la droite de regression qui "ajuste au mieux dans le nuage de points"?

Idée: Minimiser les erreurs commises entre la vraie valeur de l'observation y_i et la prévision $(\hat{y}_i)$ donnée par le droite de régression basée sur la variable explicative.

Différentes pistes: Minimiser

$$\sum_{i=1}^{n} \varepsilon_i^2 \quad \text{ou} \quad \sum_{i=1}^{n} |\varepsilon_i| \quad \text{ou} \quad \text{médiane}(\varepsilon_i) \quad \text{ou} \ \ldots$$

Les estimateurs vérifiant le critère de minimisation choisi seront notés $\hat{\beta}_0$ et $\hat{\beta}_1$:

$\hat{\beta}_0$ est une estimation du paramètre β_0

$\hat{\beta}_1$ est unes estimation du paramètre β_1.

Les propriétés des estimateurs (biais, convergence, efficacité,...) dépendront de la méthode d'estimation choisie (critère de minimisation ou autre méthode).

La droite de régression (estimation du modèle théorique) est donnée par:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad (\hat{\beta}_0, \hat{\beta}_1 \in IR).$$

Calcul **des résidus**, estimations des erreurs:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

CRITERE DES MOINDRES CARRES ORDINAIRES (MCO)

Point de vue mathématique: critére simple.

But: Trouver les valeurs de β_0 et β_1 minimisant la quantité suivante:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

$\Longrightarrow$ Système d'équations normales. Pour avoir un minimum, il faut que

$$\begin{aligned} (i) \quad & \frac{\partial}{\partial \beta_0} Q(\beta_0, \beta_1)_{|\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1} = 0 \\ (ii) \quad & \frac{\partial}{\partial \beta_1} Q(\beta_0, \beta_1)_{|\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1} = 0. \end{aligned}$$

Résolution:

Dérivons la somme des résidus carrés par rapport à β_0:

$$\frac{\partial}{\partial \beta_0} Q(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

Il s'ensuit de (i) que

$$\begin{aligned}
&\Leftrightarrow \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \\
&\Leftrightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x},
\end{aligned}$$

ce qui implique que le centre de gravité est sur la droite de régression.

Dérivons la somme des résidus carrés par rapport à β:

$$\frac{\partial}{\partial \beta_1} Q(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(x_i)$$

Il s'ensuit de (ii) que

$$\begin{aligned}
&\Leftrightarrow \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \\
&\Leftrightarrow \frac{1}{n} \sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \frac{1}{n} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} x_i^2 = 0 \\
&\Leftrightarrow \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{y}\bar{x} + \hat{\beta}_1 \bar{x}^2 - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} x_i^2 = 0 \\
&\Leftrightarrow \hat{\beta}_1 (\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{y}\bar{x} \\
&\Leftrightarrow \hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{y}\bar{x}}{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\end{aligned}$$

Estimateurs des moindres carrés ordinaires (MCO) - (Ordinary Least Squares estimators (OLS)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

Remarque: en calculant les dérivées secondes, on peut montrer que la solution est bien un minimum (exercice).

Résolution avec notation matricielle

$$Y = X\beta + \varepsilon$$

où $\beta = (\beta_0 \quad \beta_1)'$, X est la matrice de design:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \text{ et } Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \text{ et } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

Le problème de minimisation peut dés lors se réecrire de manière matricielle:

$$\begin{aligned} Q(\beta_0, \beta_1) &= \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\ &= \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - \beta' X'Y - Y'X\beta + \beta' X'X\beta \\ &= Y'Y - 2\beta' X'Y + \beta' X'X\beta \end{aligned}$$

Résolution: voir le cours de Statistique 2

Solution:$\hat{\beta} = (\hat{\beta}_0 \quad \hat{\beta}_1)' = (X'X)^{-1}X'Y$

Propriétés • Non bias de l'estimateur OLS:

$$\begin{aligned}
E(\hat{\beta}|X) &= E[(X'X)^{-1}X'Y|X] \\
&= E[(X'X)^{-1}X'(X\beta+\varepsilon)|X] \\
&= E[(X'X)^{-1}X'X\beta|X] + E[(X'X)^{-1}X'\varepsilon)|X] \\
&= \beta + (X'X)^{-1}X'E[\varepsilon|X] \\
&= \beta + (X'X)^{-1}X'E[\varepsilon] \quad \text{Hyp: X est indépendant de } \varepsilon \\
&= \beta \quad \text{Hyp : } E(\varepsilon) = 0
\end{aligned}$$

• Calcul de la variance de l'estimateur OLS:

$$\begin{aligned}
V(\hat{\beta}|X) &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] \\
&= E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X] \\
&= (X'X)^{-1}X'E(\varepsilon\varepsilon'|X)X(X'X)^{-1} \\
&= (X'X)^{-1}X'V(\varepsilon)X(X'X)^{-1} \\
&= (X'X)^{-1}\sigma^2 I X'X(X'X)^{-1} \\
&\quad \text{Hyp: X est indépendant de } \varepsilon \text{ et } E(\varepsilon) = 0 \\
&= \sigma^2(X'X)^{-1} \\
&\quad \text{Hyp : } cov(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j \text{ et var}(\varepsilon_i) = \sigma^2 \forall i
\end{aligned}$$

• Efficacité de l'estimateur OLS (dans la classe des estimateurs non biaisés) si $\varepsilon \sim N$.

Estimation du paramètre de nuisance

Le paramètre de nuisance σ intervient dans l'expression de la variance de l'estimateur $\hat{\beta}$. En pratique il faudra donc l'estimer.

Estimateur sans biais de σ^2:

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

Notons que $E(\hat{\sigma}^2) = \sigma^2$ (non biaisé).

Parfois, $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ est appelé "the estimated standard error of the regression model"

2.1.2 Tests d'hypothèse

Inference sur le pente de la droite (slope): β_1

problème de test: $H_0 : \beta_1 = 0$

$$H_1 : \beta_1 \neq 0$$

Statistique de test: $\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$

Loi sous H_0 (sous l'hypothèse de normalité des erreurs): $T = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$

Règle de comportement: Rejeter H_0 au niveau $\alpha\%$ si $|T| > \mathrm{t}_{n-2,1-\frac{\alpha}{2}}$
Si RH0: la variable X a une influence significative sur la variable Y.

Si l'hypothèse de normalité n'est pas validée, on utilisera l'approximation normale (TCL).

Inference sur la constante: β_0

Problème de test: $H_0 : \beta_0 = 0$

$$H_1 : \beta_0 \neq 0$$

Statistique de test: $\frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}$

Loi sous H_0:

$$T = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)} \sim t_{n-2}$$

Règle de comportement: Rejeter H_0 au niveau $\alpha\%$ si $|T| > \mathrm{t}_{n-2,1-\frac{\alpha}{2}}$

Interprétation: La constante du modèle β_0 est significativement différente de zéro

Si l'hypothèse de normalité n'est pas validée, on utilisera l'approximation normale (TCL).

2.1.3 Intervalles de confiance

Intervalle de confiance pour β_1

Sous l'hypothèse de normalité des erreurs, on a:

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2} \qquad \text{donc}$$

$$P(-t_{n-2,1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \leq t_{n-2,1-\frac{\alpha}{2}}) = 1-\alpha$$

ainsi, on obtient:

$$P(\hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}} SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + SE(\hat{\beta}_1) t_{n-2,1-\frac{\alpha}{2}})$$

$$= 1 - \alpha$$

Intervalle de confiance au niveau $100(1-\alpha)\%$:

$$\hat{\beta}_1 \pm t_{n-2,1-\frac{\alpha}{2}} SE(\hat{\beta}_1)$$

Interpretation: we can be 95% confident that the unknown parameter β_1 is included between the two values of the confidence interval.

2.1.4 Qualité de l'ajustement

Coefficient de détermination:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Interprétation: % de la variance de la variable Y expliquée par la variable explicative X.

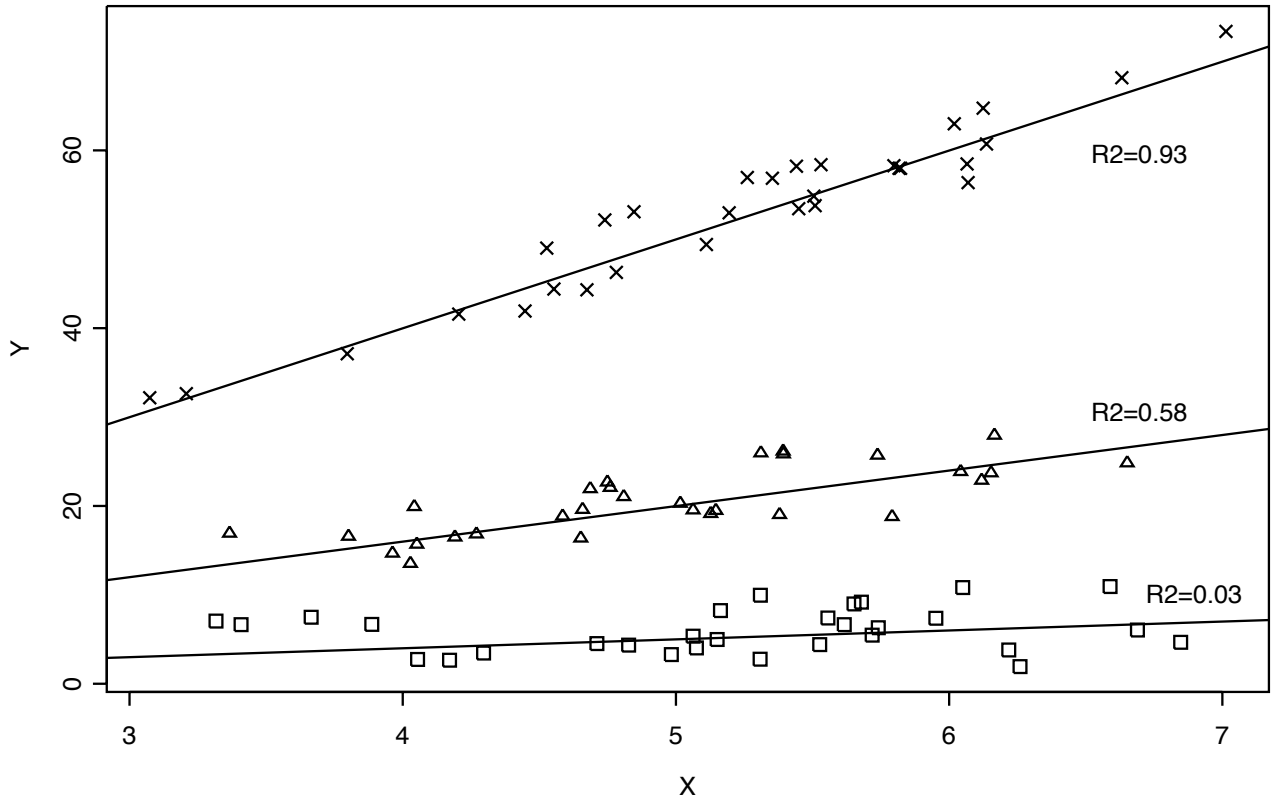
In the case of simple linear regression:

$$R^2 = r^2 = corr^2(X, Y)$$

"*These measures of goodness of fit have a fatal attraction*"

Cramer, 1987

Is R^2 a measure of the quality of the fit of a model?



The dispersion of the residuals is the same in the three data clouds. The R^2 coefficient is merely determined by the slope of the regression lines

2.1.5 Exemple avec variable X binaire

Sometimes a regressor is binary:

- $X = 1$ if female, $X = 0$ if male
- $X = 1$ if small class size, $X = 0$ if not

So far, β_1 has been called a "slope" but that doesn't make much sense if X is binary.

How do we interpret regression with a binary regressor?

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ where X is binary}$$

$x_i = 0$: $y_i = \beta_0 + \varepsilon_i \Rightarrow E[Y_i] = \beta_0$

$x_i = 1$: $y_i = \beta_0 + \beta_1 + \varepsilon_i \Rightarrow E[Y_i] = \beta_0 + \beta_1$

$$\Downarrow$$

$$\beta_1 = E(Y_i|x_i = 1) - E(Y_i|x_i = 0)$$

β_1 is the population difference in group means

Example: TestScore and STR, California data.

Let

$$D_i = \begin{cases} 1 \text{ if STR } \leq 20 \\ 0 \text{ if STR } > 20 \end{cases}$$

The OLS estimate of the regression line is:

$$\hat{TestScore_i} = 650.0 + 7.4 D_i$$

(1.3) (1.8)

Difference in means between groups: 7.4 and $SE = 1.8 \Rightarrow t = \frac{7.4}{1.8} = 4.0 > 1.96$ then the effect is significant.

Compare the regression results with the group means, computed directly:

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = 4.05$$

This is the same as in the regression!

2.1.6 Exemple

The class size/test score policy question:

- *What is the effect on test scores of reducing STR by one student/class?*

- *Object of policy interest:* $\frac{\Delta testscore}{\Delta STR}$

- *This is the slope of the line relating test score and STR*

Quelques statistiques:

Statistics	Mean	Std	$x_{1/10}$	$x_{1/4}$	$x_{1/2}$	$x_{3/4}$	$x_{9/10}$
STR	19.6	1.9	17.3	18.6	19.7	20.9	21.9
Test score	654.2	19.1	630.4	640.0	654.5	666.7	679.1

Some Notation and Terminology

- *The population regression Model:*

$$TestScor_i = \beta_0 + \beta_1 STR_i + \varepsilon_i$$

β_1 = *slope of population regression line*

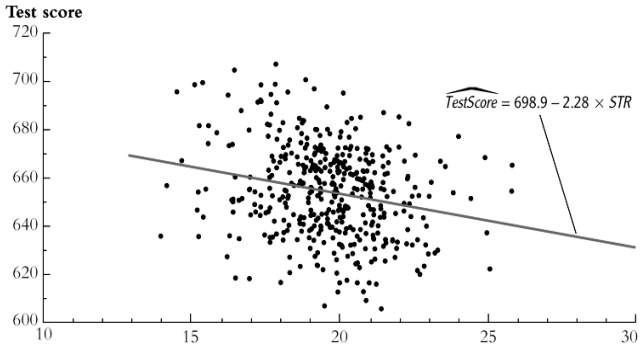
$= \frac{\Delta TestScore}{\Delta STR}$

= *change in test score for a unit change in* STR

Application to the California *Test Score – Class Size* data

FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.



Estimated slope = $\hat{\beta}_1 = -2.28$

Estimated intercept = $\hat{\beta}_0 = 698.9$

Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

Interpretation of the estimated slope and intercept

$$\hat{TestScore} = 698.9 - 2.28STR$$

- *Districts with one more student per teacher on average have test scores that are 2.28 points lower.*

- *That is:* $\frac{\Delta TestScore}{\Delta STR} = -2.28.$

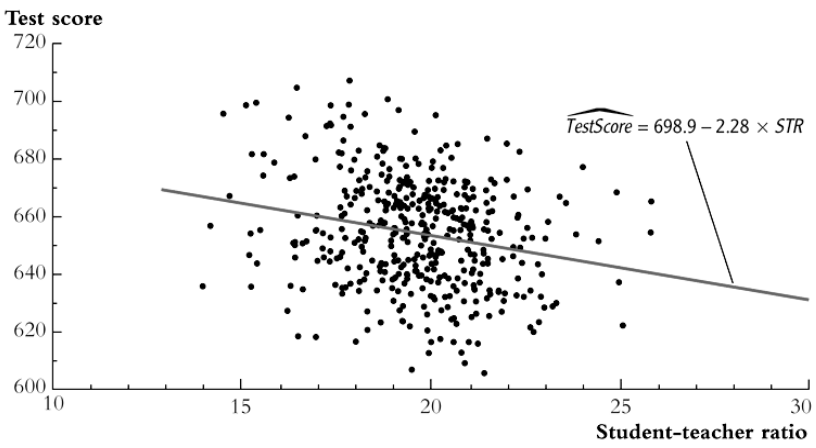
- *The intercept means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9. This interpretation of the intercept makes no sense - it extrapolates the line outside the range of the data - in this application.*

Predicted values and residuals

Predicted values & residuals:

FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.



One of the districts in the data set is Antelope, CA, for which *STR* = 19.33 and *Test Score* = 657.8

predicted value: $\hat{Y}_{Antelope} = 698.9 - 2.28 \times 19.33 = 654.8$

residual: $\hat{u}_{Antelope} = 657.8 - 654.8 = 3.0$

OLS regression: STATA output

Number of obs = 420

R-squared = 0.0512

Variable	Coef.	SE	t	$P > \|t\|$	[95% Conf	Interval]
STR	-2.28	0.52	-4.39	0.00	-3.30	-1.26
cons	698.93	10.36	67.44	0.00	678.56	719.31

Exemples:

- *t-statistic testing* $\beta_1 = 0$

$$t - stat = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-2.28}{0.52} = -4.38$$

Alternatively, we can compute the p-value

- *Intervalle de confiance à* 95%*:*

$$\begin{aligned} IC(\hat{\beta}_1) &= [\hat{\beta}_1 \pm 1.96 SE(\hat{\beta}_1)] \\ &= [-2.28 \pm 1.96 0.52] = [-3.30; -1.26] \end{aligned}$$

The confidence interval does not include zero.

Summary and Assessment

The initial policy question:

Suppose new teachers are hired so the student-teacher ratio falls by one student per class. What is the effect of this policy intervention (this "treatment") on test scores?

Does our regression analysis give a convincing answer?

Not really - districts with low STR tend to be ones with lots of other resources and higher income families, which provide kids with more learning opportunities outside school ... this suggests that

$corr(\varepsilon, STR) > 0.$

2.1.7 Focus sur les hypothèses de Gauss-Markov

• Hypothèse sur la forme du modèle

Hyp_0: Linéarité du modèle

Détection:

- *Scatter-plots* $Y \sim X_j \ \forall j$
- *Visualiser les résidus: si la structure n'est pas aléatoire, suspicion de non-linéarité*
- *Test RESET (Ramsey, 1969):*

  $y_i = \alpha + x_i\beta + \alpha_2\hat{y}_i^2 + \alpha_3\hat{y}_i^3 + \ldots + \alpha_Q\hat{y}_i^Q + \nu_i$

  Problème de test: $H_0: \alpha_2 = \ldots = \alpha_Q = 0$

Remédes:

- *Linéariser la relation en utilisant une transformation (p.e. logarithmique)*
- *Utiliser une régression "polynomiale"*
- *Utiliser des modèles non linéaire*

- **Hypothèse sur le terme d'erreur**

Hyp$_1$*:* $E(\varepsilon_i) = 0 \ \forall i = 1, \ldots, n$

Cette hypothèse est automatiquement vérifiée si une constante est inclue au modèle.

Hyp$_2$*: Normalité des erreurs* $\varepsilon \sim N$

Problème: Si petit échantillon, on ne peut pas utiliser les tests asymptotiques, (et le test de Fisher n'est plus valable.)

Détection:

- *Méthodes graphiques: histogramme des résidus, QQ-plot des résidus*
- *Test de Jarque-Bera basé sur la skewness et le kurtosis*

  $$JB = n[\frac{1}{6}(\frac{1}{n}\sum_i \frac{r_i^3}{\hat{\sigma}^3})^2 + \frac{1}{24}(\frac{1}{n}\sum_i \frac{r_i^4}{\hat{\sigma}^4} - 3)^2]$$

  sous H_0 *(=normalité):* $JB \sim_{n\to\infty} \chi^2_2$

Hyp$_3$: Non-autocorrélation $cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$

Problème: les erreurs ne sont pas indépendendantes, ce qui arrive souvent avec des données temporelles $\Rightarrow$

$$var(\hat{\beta}_{LS}) \neq \sigma^2 (X'X)^{-1}$$

SE des estimateurs de régression sont biaisées.

Détection:Automatiquement vérifiée si les individus sont sélectionnés de manière aléatoire.

Exception: données dans le temps

- *Méthode graphique: Plot de l'indice de temps versus* r_i.
- *Statistique de Durbin Watson: tester s'il y a une structure autorégressive d'ordre 1:*

$$\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$$

où $|\rho| < 1$ *et* $\nu_t \sim N(0, \sigma_\nu^2)$. *Tester* $\rho = 0$.

Remèdes:

- *trouver une nouvelle variable expliquant la dépendance dans le temps*

- *Generalized Least Squares (GLS).*

$$V(\varepsilon|X) = \sigma^2\Psi \neq \sigma^2 I$$

Posons $\Psi^{-1} = P'P$ *où* P *est une matrice carrée non singulière, alors*

$$\Psi = (P'P)^{-1} = P^{-1}(P')^{-1}$$

$$\rightarrow P\Psi P' = P(P'P)^{-1}P' = PP^{-1}(P')^{-1}P' = I$$

Donc,

$$E[P\varepsilon|X] = PE[\varepsilon|X] = 0$$

$$V[P\varepsilon|X] = PV[\varepsilon|X]P' = \sigma^2 P\Psi P' = \sigma^2 I$$

Nous allons donc transformer les données afin d'obtenir un modèle vérifiant les conditions de Gauss-Markov

Modèle: $Py = PX\beta + P\varepsilon$

Notons $y^* = Py$ *et* $X^* = PX$, *l'estimateur OLS seront donc donné par:*

$$\begin{aligned}\hat{\beta} &= (X^{*'}X^*)^{-1}X^*y^* \\ &= (X'\Psi^{-1}X)^{-1}X'\Psi^{-1}y\end{aligned}$$

Cet estimateur est l'estimateur GLS

Remarque: Souvent Ψ *est inconnu, il faut donc l'estimer, on parle alors de l'estimateur FGLS (Feasible GLS)*

Hyp$_4$: Homoscédasticité: $var(\varepsilon_i) = \sigma^2 \ \forall i = 1, \ldots, n$

problème: $var(\hat{\beta}_{LS}) \neq \sigma^2 (X'X)^{-1} \Rightarrow$ SE des estimateurs de régression sont biaisées (problème pour les tests et les IC)

Détection:

- *Méthodes graphiques: Plot $\hat{y}_i$ versus r_i, Plots des X versus les r_i.*
- *Test de White: $H_0 : \sigma_i^2 = \sigma^2 \ \forall i$ Idée: Les résidus dépendent-ils des variables explicatives?*

$$r_i \sim X_i, X_i^2, X_i X_j$$

Statistique de test: $T = nR^2 \sim_{H0} \chi^2$ (ou bien test de Fisher)

Remédes:

- *Stabiliser la variance: transformation logarithmique*

- *Autres méthodes d'estimation: Weighted least squares (WLS) Idée: minimiser la fonction:*

$$\sum_{i=1}^{n} w_i \varepsilon_i^2 = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2$$

où $w_i = \frac{1}{\sigma_i^2}$ $\forall i$, *au lieu de la fonction*

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

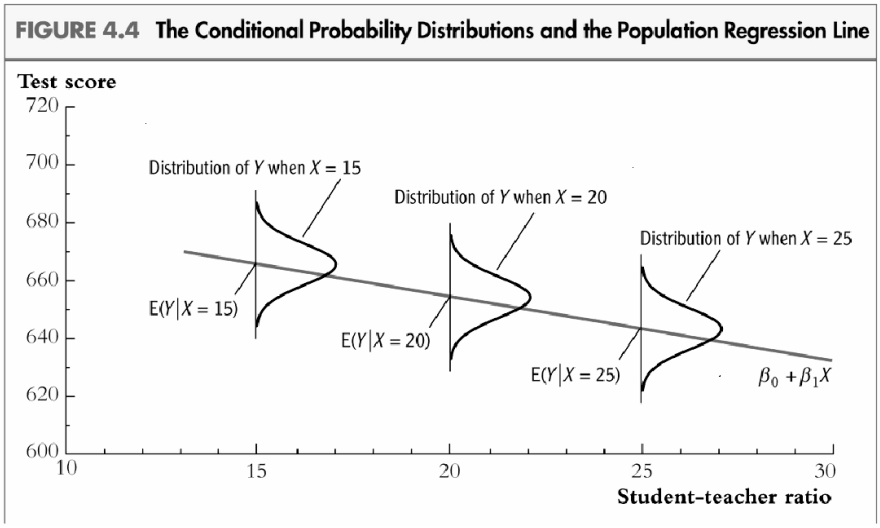
Mais les variances des erreurs ne sont pas connues

$$\Downarrow$$

Two-step procédure:

  - *Step 1: Faire une estimation par OLS afin d'estimer les poids* w_i
  - *Step 2: Utiliser l'estimateur WLS*

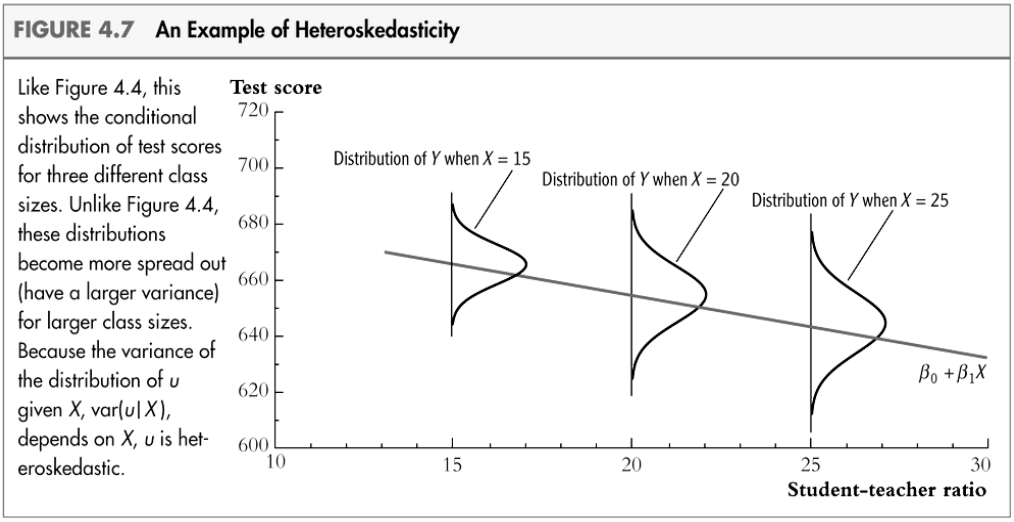
Homoskedasticity *in a picture:*



- $E(u|X=x) = 0$ (u satisfies Least Squares Assumption #1)
- The variance of u does **not** change with (depend on) x

4-27

Heteroskedasticity *in a picture:*



- $E(u|X=x) = 0$ (u satisfies Least Squares Assumption #1)
- The variance of u depends on x – so u is heteroskedastic.

4-28

- **Hypothèse sur les variables explicatives**

Hyp$_5$: Hypothèse d'exogénéité: $cov(X, \varepsilon) = 0$

Problème: *l'estimateur OLS des paramétres de régression sera biaisé et inconsistent (bias d'omission, biais de simultanéité, erreur de mesure)*

Vérification: *X doit être mesurée sans erreurs, pas de biais d'omission, pas d'interdépendance.*

Remédes:

- *Si biais d'omission: ajouter la variable manquante au modèle*

- *Si variable explicative endogène utiliser la méthode des variables instrumentales (IV):*

  Z est une variable instrumentale valide pour X dans l'équation:

  $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

  $$si\ corr(Z, X) \neq 0\ et\ corr(Z, \varepsilon) = 0$$

  Two-step procedure:
  Step 1: $x_i = \gamma_0 + \gamma_1 z_i + \varepsilon_i \ \rightarrow \hat{x}_i$ *par OLS*
  Step 2: Estimer par OLS l'équation:

  $$y_i = \beta_0 + \beta_1 \hat{x}_i + \tilde{\varepsilon}_i$$

- *Si équation simultanée (exemple offre $\Leftrightarrow$ demande): utiliser des estimateurs DMC (double moindres carrés, TMC (triple moindres carrés).*

Back to class size:

- *What is an ideal randomized controlled experiment for measuring the effect on Test Score of reducing STR?*
- *How does our regression analysis of observational data differ from this ideal?*

1. *The treatment is not randomly assigned*
2. *In the US - in our observational data - districts with higher family incomes are likely to have both smaller classes and higher test scores.*
3. *As a result it is plausible that*

$$corr(\varepsilon, STR) \neq 0$$

4. *If so, $\hat{\beta}_1$ is biased: does an omitted factor make class size seem more important than it really is?*

2.2 REGRESSION LINEAIRE MULTIPLE

2.2.1 Biais d'omission

Dans l'exemple, nous avons obtenu:

$$\hat{TestScore} = 698.9 - 2.28STR \quad R^2 = 0.05$$
$$(10.4) \quad (0.52)$$

Is this a credible estimate of the causal effect on test scores of a change in the student-teacher ratio? No: there are omitted confounding factors (family income; whether the students are native English speakers) that bias the OLS estimator.

Pour avoir un biais de variable omise (Z), il faut:

- que Z soit un determinant de Y
- que Z soit corrélé avec la variable X.

In the test score example:

- *English language ability (whether the student has English as a second language) plausibly affects standardized test scores: Z is a determinant of Y.*
- *Immigrant communities tend to be less affluent and thus have smaller school budgets - and higher STR: Z is correlated with X.*

TABLE 5.1 Differences in Test Scores for California School Districts with Low and High Student Teacher Ratios, by the Percentage of English Learners in the District

	Student-Teacher Ratio < 20		Student-Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All Districts	657.4	238	650.0	182	7.4	4.04
Percent of English Learners						
< 2.2%	664.1	78	665.4	27	−1.3	−0.44
2.2–8.8%	666.1	61	661.8	44	4.3	1.44
8.8–23.0%	654.6	55	649.7	50	4.9	1.64
> 23.0%	636.7	44	634.8	61	1.9	0.68

- Districts with fewer English Learners have higher test scores
- Districts with lower percent *EL* (*PctEL*) have smaller classes
- Among districts with comparable *PctEL*, the effect of class size is small (recall overall "test score gap" = 7.4)

Conséquence de l'omission d'une variable explicative pertinente

Exemple : Le vrai modèle est donné par:

$$M_0 : y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 + \varepsilon_i.$$

En réalité, on travaille avec le modèle mal spécifié:

$$M : y_i = \gamma_2 x_{i2} + \gamma_3 + v_i,$$

L'estimateur $\hat{\gamma}_2$ dans le modèle M est donné par:

$$\begin{aligned}
\hat{\gamma}_2 &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_{i2} - \bar{x}_2)}{\sum_{i=1}^n (x_{i2} - \bar{x}_2)^2} \\
&= \frac{\sum_{i=1}^n (\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 + \varepsilon_i - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 - \beta_3 - \bar{\varepsilon})(x_{i2} - \bar{x}_2)}{\sum_{i=1}^n (x_{i2} - \bar{x}_2)^2} \\
&= \frac{\sum_{i=1}^n (\beta_1 (x_{i1} - \bar{x}_1) + \beta_2 (x_{i2} - \bar{x}_2) + (\varepsilon_i - \bar{\varepsilon}))(x_{i2} - \bar{x}_2)}{\sum_{i=1}^n (x_{i2} - \bar{x}_2)^2}.
\end{aligned}$$

En utilisant l'hypothése d'exogénéité, on a:

$$\begin{aligned}
E[\hat{\gamma}_2 | x] &= \beta_2 + \beta_1 \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum_{i=1}^n (x_{i2} - \bar{x}_2)^2} \\
&= \beta_2 + \beta_1 \frac{\mathrm{cov}(x_1, x_2)}{\mathrm{var}(x_2)}.
\end{aligned}$$

$\Rightarrow$ si x_1 et x_2 sont corrélés, $E[\hat{\gamma}_2] \neq \beta_2$.

Conséquence de l'inclusion d'une variable non pertinente

L'estimateur MCO $\hat{\beta}$ restera non biaisé, mais sera moins précis.

Exemple : Le vrai modèle est donné par :

$$M_0 : y_i = \gamma_2 + \varepsilon_i.$$

En réalité, on travaille avec le modèle mal spécifié:

$$M : y_i = \beta_1 x_{i1} + \beta_2 + v_i.$$

On déduit de la relation:

$$\hat{\beta}_2 = \bar{y} - \hat{\beta}_1 \bar{x} = \gamma_2 + \bar{\varepsilon} - \hat{\beta}_1 \bar{x}$$

$$\Rightarrow E[\hat{\beta}_2|x] = \gamma_2 + 0 - \bar{x} E[\hat{\beta}_1|x]$$

$$\begin{aligned} \text{Or } E[\hat{\beta}_1|x] &= \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)E[y_i - \bar{y}|x]}{\sum_{i=1}^{n}(x_{i1} - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)E[\gamma_2 + \varepsilon_i - \gamma_2 - \bar{\varepsilon}|x]}{\sum_{i=1}^{n}(x_{i1} - \bar{x})^2} \\ &= 0 \end{aligned}$$

Donc, on obtient

$$E[\hat{\beta}_2|x] = \gamma_2.$$

Pour la précision, dans le vrai modèle M_0 on estimerait γ_2 par, $\hat{\gamma}_2 = \bar{y}$ où l'on sait que

$$\text{var}(\hat{\gamma}_2) = \frac{\sigma_\varepsilon^2}{n}$$

(où $\sigma_\varepsilon^2 = E[\varepsilon^2]$) tandis que dans le modèle mal spécifié, on a:

$$\begin{aligned}
\text{var}(\hat{\beta}_2|x) &= \text{var}(\gamma_2 + \bar{\varepsilon} - \hat{\beta}_1\bar{x}|x) \\
&= \text{var}(\bar{\varepsilon}|x) + \bar{x}^2\text{var}(\hat{\beta}_1|x) - 2\text{cov}(\bar{\epsilon}, \hat{\beta}_1|x)\bar{x} \\
&= \frac{\sigma_\epsilon^2}{n} + \bar{x}^2\text{var}(\hat{\beta}_1|x) - 0
\end{aligned}$$

Donc, on a que:

$$\text{var}(\hat{\beta}_2|x) > \text{var}(\hat{\gamma}_2)$$

La variance de $\hat{\beta}_2$ est donc plus grande que la variance de $\hat{\gamma}_2$. L'ajout de "bruit" dans la modèle va rendre plus imprécis les estimations.

Il reste à montrer que $\bar{\epsilon}$ et $\hat{\beta}_1$ sont non corrélés:

$$\begin{aligned}\text{cov}(\bar{\epsilon}, \hat{\beta}_1|x) &= E[(\bar{\epsilon} - E(\bar{\epsilon}))(\hat{\beta}_1 - E(\hat{\beta}_1))|x] = E[\bar{\epsilon}\hat{\beta}_1|x] \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) E[\bar{\epsilon}(\epsilon_i - \bar{\epsilon})|x]}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

puisque

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\gamma_2 + \epsilon_i - \gamma_2 - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2}.\end{aligned}$$

Or les résidus sont iid,

$$\begin{aligned}E[\bar{\epsilon}(\epsilon_i - \bar{\epsilon})|x] &= E[\frac{1}{n}\sum_{j=1}^n \epsilon_j \epsilon_i - \frac{1}{n^2}\sum_{j,l} \epsilon_j \epsilon_l] \\ &= \frac{1}{n} E[\epsilon_i(\epsilon_1 + \cdots + \epsilon_n)] \\ &\quad - \frac{1}{n^2}\sum_{j=1}^n E[\epsilon_j(\epsilon_1 + \cdots + \epsilon_n)] \\ &= \frac{1}{n} E[\epsilon_i^2] - \frac{1}{n^2} n E[\epsilon_j^2] = \frac{\sigma_\epsilon^2}{n} - \frac{\sigma_\epsilon^2}{n} = 0,\end{aligned}$$

il vient $cov(\bar{\epsilon}, \hat{\beta}_1) = 0$.

2.2.2 Estimations, problèmes de test, Intervalles de confiance

- Estimateurs OLS: $\hat{\beta} = (X'X)^{-1}X'Y$

- Statistique de test pour $H_0 : \beta_i = 0$

$$T = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \sim t_{n-p}$$

où $i = 0, 1, 2$ et $p = 3$ est le nombre de paramètres de régression à estimer.

- Intervalle de confiance pour β_i:

$$\hat{\beta}_i \pm t_{n-p,1-\frac{\alpha}{2}} SE(\hat{\beta}_i)$$

Attention aux hypothèses !!

2.2.3 Hypothèse supplémentaire

Hyp$_6$: Pas de multicolinéarité parfaite

Aucune variable explicative ne peut être une combinaison linéaire parfaite des autres régresseurs

-Quid si violation: il est impossible d'obtenir des estimateurs (near singular matrix).

-Remèdes: enlever la variable redondante.

Rappel mathématique: une matrice carrée ne peut être inversée si son déterminant est nul. Or

$$\hat{\beta} = (X'X)^{-1}X'Y$$

et $\det(X'X) = 0$ si multicolinéarité parfaite.

$\text{Hyp}_6 bis$: Presque multicolinéarité

Quid si violation:

- les variables pertinentes sont non significatives
- signe des estimations contre-intuitif
- estimateurs instables

Détection:

- vérifier la matrice des corrélations des variables explicatives et scatter plot deux à deux des variables explicatives
- $SE(\hat{\beta})$ trés grand, signe des estimations contre intuitif
- Variables non significatives avec le test de student mais globalement significatives avec le test de Fisher.

Remède:

- Enlever des variables
- Analyse en composante principale (combinaisons linéaires des X) qui sont non corrélées
- Revenir à la définition des variables et essayer de comprendre pourquoi certaines variables sont fortement corrélées (et peut-être les transformées)

2.2.4 Interprétation

Soit la régression linéaire multiple:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, i = \ldots, n$$

où

- X_1, X_2 sont 2 variables explicatives

- β_0 est le paramètre constant

- β_1 est l'effet sur Y d'un changement unitaire de X_1 en tenant constant X_2

- β_2 est l'effet sur Y d'un changement unitaire de X_2 en tenant constant X_1

- ϵ_i est le terme d'erreur

Quel est la changement sur Y si on ajoute à la variable X_1 une quantité ΔX_1 tout en tenant X_2 constant?

• Droite de régression (théorique) avant le changement:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

• Droite de régression (théorique) après le changement:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$$

La différence entre les 2 régressions est donc donnée par:

$$\Delta Y = \beta_1 \Delta X_1$$

$$\Rightarrow \beta_1 = \frac{\Delta Y}{\Delta X_1} \text{ ,} X_2 \text{ étant constant}$$

$$\Rightarrow \beta_2 = \frac{\Delta Y}{\Delta X_2} \text{ ,} X_1 \text{ étant constant}$$

Exemple: Qualité et réputation des écoles à Bruxelles

E. Arias, C. Dehon et N. Gothelf

Question: Peut-on mesurer un impact de la qualité d'une école (mesuré par le taux de réussite à l'université) sur sa réputation ?

$$SR = \gamma_0 + \gamma_1 R + \gamma_2 Standing + \gamma_3 Nord + \gamma_4 Centre + \varepsilon,$$

où $SR = \frac{\sharp \text{inscriptions}}{\sharp \text{places disponibles}}$

Variable	Coefficient	Ecart-Type
Taux de réussite	0.02^{**}	0.01
Standing	0.15^{***}	0.05
Nord	1.12^{***}	0.26
Centre	0.44	0.34
Constante	0.13	0.42
R^2=0.52	N=58	
* $p < 0.1$,	** $p < 0.05$,	*** $p < 0.01$

2.2.5 Interactions between independent variables

Interactions Between continuous and binary Independent Variables

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

- D_i is binary, X is continuous
- As specified above, the effect on Y of X (holding constant D) = β_2, which does not depend on D
- To allow the effect of X to depend on D, include the "interaction term" $D_i \times X_i$ as a regressor:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \times X_i) + u_i$$

Interpreting the coefficients

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \times X_i) + u_i$$

General rule: compare the various cases

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) \qquad \text{(b)}$$

Now change X:

$$Y + \Delta Y = \beta_0 + \beta_1 D + \beta_2 (X + \Delta X) + \beta_3 [D \times (X + \Delta X)] \qquad \text{(a)}$$

subtract (a) – (b):

$$\Delta Y = \beta_2 \Delta X + \beta_3 D \Delta X \text{ or } \frac{\Delta Y}{\Delta X} = \beta_2 + \beta_3 D$$

- The effect of X depends on D (what we wanted)
- β_3 = increment to the effect of X, when $D = 1$

Example: *TestScore, STR, HiEL* (=1 if *PctEL* $\geq$ 20)

$$\widehat{TestScore} = 682.2 - 0.97STR + 5.6HiEL - 1.28(STR \times HiEL)$$
$$(11.9)\ (0.59) \qquad (19.5) \qquad (0.97)$$

- When $HiEL = 0$:

$$\widehat{TestScore} = 682.2 - 0.97STR$$

- When $HiEL = 1$,

$$\widehat{TestScore} = 682.2 - 0.97STR + 5.6 - 1.28STR$$
$$= 687.8 - 2.25STR$$

- Two regression lines: one for each *HiSTR* group.
- Class size reduction is estimated to have a larger effect when the percent of English learners is large.

Binary-continuous interactions: the two regression lines

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \times X_i) + u_i$$

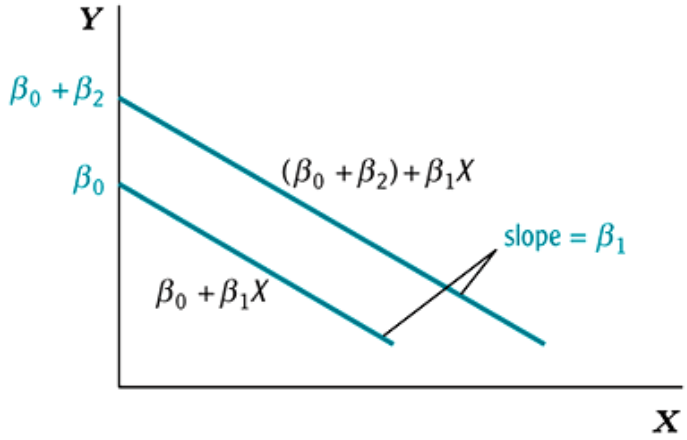
Observations with $D_i = 0$ (the "$D = 0$" group):

$$Y_i = \beta_0 + \beta_2 X_i + u_i$$

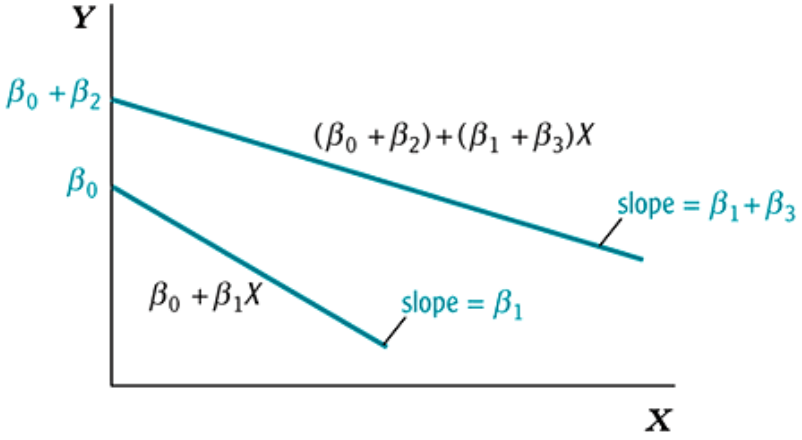
Observations with $D_i = 1$ (the "$D = 1$" group):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 + \beta_2 X_i + \beta_3 X_i + u_i \\ &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_i + u_i \end{aligned}$$

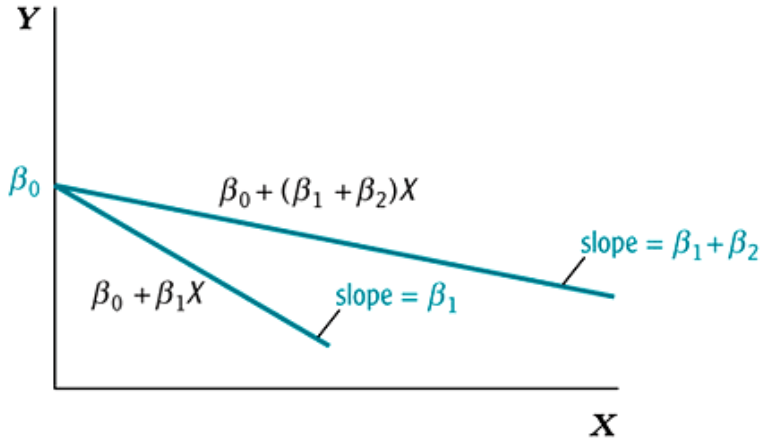
FIGURE 6.8 Regression Functions Using Binary and Continuous Variables



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

Interactions of binary variables and continuous variables can produce three different population regression functions: (a) $\beta_0 + \beta_1 X + \beta_2 D$ allows for different intercepts but has the same slope; (b) $\beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D)$ allows for different intercepts and different slopes; and (c) $\beta_0 + \beta_1 X + \beta_2 (X \times D)$ has the same intercept but allows for different slopes.

2.2.6 Modèles log-log

Logarithmic transforms permit modeling relations in "percentage" terms (like elasticities).

Here's why:

$$\ln(x + \Delta x) - \ln(x) = \ln(1 + \frac{\Delta x}{x}) \approx \frac{\Delta x}{x}$$

Numerically:

$\ln(1.01) = 0.00995 \approx 0.01;$

$\ln(1.10) = 0.0953 \approx 0.10$

Three cases:

- *linear-log* :$y_i = \beta_0 + \beta_1 \ln(x_i) + \epsilon_i$
- *log-linear* :$\ln(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$
- *log-log* :$\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + \epsilon_i$

The interpretation of the slope coefficient differs in each case.

Lin - log regression: $Y = \beta_0 + \beta_1 \ln(X) + \epsilon$

Change X: $Y + \Delta Y = \beta_0 + \beta_1 \ln(X + \Delta X) + \epsilon$

Subtraction: $\Delta Y = \beta_1[ln(X + \Delta X) - ln(X)]$

$$\Downarrow$$

As $[ln(X + \Delta X) - ln(X)] \approx \frac{\Delta x}{x}$, *we have:*

$$\Delta Y \approx \beta_1 \frac{\Delta X}{X} \text{ small } \Delta X$$

Then the percentage change in X is given by $100\frac{\Delta X}{X}$, *so a* 1% *increase in X (multiplying X by 1.01) is associated with a* $\frac{\beta_1}{100}$ *change in Y.*

Example: TestScore vs. ln(Income)

$$\hat{TestScore}_i = 557.8 + 36.42 \ln(Income_i)$$

so a 1% *increase in Income is associated with an increase in TestScore of* 0.36 *points.*

Log-linear regression:$\ln(Y) = \beta_0 + \beta_1 X + \epsilon$

Now change: $\ln(Y + \Delta Y) = \beta_0 + \beta_1(X + \Delta X) + \epsilon$

Subtraction: $\ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X$

$$\Downarrow$$

$$\frac{\Delta Y}{Y} \approx \beta_1 \Delta X$$

Now $100\frac{\Delta Y}{Y}$ *= percentage change in* Y*, so a change in* X *by one unit* ($\Delta X = 1$) *is associated with a* $100\beta_1\%$ *change in* Y.

Log-log: $\ln(Y) = \beta_0 + \beta_1 \ln(X) + \epsilon$

Change: $\ln(Y+\Delta Y) = \beta_0+\beta_1 \ln(X+\Delta X)+\epsilon$

Subtraction:

$$\ln(Y+\Delta Y)-\ln(Y) = \beta_1[ln(X+\Delta X)-ln(X)]$$

$$\Downarrow$$

$$\frac{\Delta Y}{Y} \approx \beta_1 \frac{\Delta X}{X} \; small \; X$$

So a 1% *change in* X *is associated with a* $\beta_1\%$ *change in* Y. *In the log-log specification,* β_1 *has the interpretation of an elasticity.*

Example: ln(TestScore) vs. ln(Income)

$$\ln(\hat{TestScore_i}) = 6.336 + 0.0554 \ln(Income_i)$$

An 1% *increase in Income is associated with an increase of* 0.0554% *in TestScore*

Summary: Logarithmic transformations

Three cases, differing in whether Y and/or X is transformed by taking logarithms.

After creating the new variable(s), the regression is linear in the new variables and the coefficients can be estimated by OLS.

Hypothesis tests and confidence intervals are now standard.

The interpretation of β_1 differs from case to case.

Choice of specification should be guided by judgment (which interpretation makes the most sense?), plotting predicted values

2.2.7 Qualité de l'ajustement

Coefficient de détermination:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Interprétation: % de la variance de la variable Y expliquée par la variable explicative X.

PROBLEME: Le R^2 augmente toujours lorqu'on ajoute une variable explicative au modèle

$\Rightarrow$ **Solution en comparant 2 modèles**

$$M0 : y_i = \alpha + \epsilon_i$$

$$M1 : y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i(p-1)} + \epsilon_i$$

Estimations non biaisées des variances des erreurs:

$$\hat{\sigma}_0^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2$$

$$\hat{\sigma}_1^2 = \frac{1}{n-p} \sum_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{p-1} x_{i(p-1)})^2$$

$$\Downarrow$$

Mesure de la qualité globale du modèle M1 par rapport à M0:

$$\begin{aligned}\tilde{R}^2 &= 1 - \frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2} = 1 - \frac{\frac{1}{n-p}\sum_i e_i^2}{\frac{1}{n-1}\sum_i (y_i - \bar{y})^2} \\ &= 1 - \frac{n-1}{n-p}\frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2} \\ &= 1 - \frac{n-1}{n-p}(1 - R^2)\end{aligned}$$

- $\tilde{R}^2 \approx 0$: les variables exogénes n'apportent presque rien au modèle
- $\tilde{R}^2 \approx 1$: les variables exogénes sont pertinantes et expliquent un grand % de la variation de Y.

Remarque: $\tilde{R}^2$ peut être négatif pour de très mauvais modèle.

2.2.8 Intervalle de confiance "multiple"

Modèle:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i(p-1)} + \epsilon_i$$

What is a joint confidence set for β_1 and β_2?

A 95% confidence set is:

- *A set-valued function of the data that contains the true parameter(s) in 95% of hypothetical repeated samples.*
- *The set of parameter values that cannot be rejected at the 5% significance level when taken as the null hypothesis.*

The coverage rate of a confidence set is the probability that the confidence set contains the true parameter values

A "common sense" confidence set is the union of the 95% *confidence intervals for* β_1 *and* β_2*, that is, the rectangle:*

$$\{\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1), \hat{\beta}_2 \pm 1.96SE(\hat{\beta}_2)\}$$

What is the coverage rate of this confidence set?

Does its coverage rate equal the desired confidence level of 95

$$
\begin{aligned}
&P[(\beta_1, \beta_2) \in \{\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1), \hat{\beta}_2 \pm 1.96SE(\hat{\beta}_2)\}] \\
&= P[\hat{\beta}_1 - 1.96SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 1.96SE(\hat{\beta}_1), \\
&\hat{\beta}_2 - 1.96SE(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + 1.96SE(\hat{\beta}_2)] \\
&= P[-1.96 \leq \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \leq 1.96, \\
&-1.96 \leq \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \leq 1.96] \\
&= P[|t1| \leq 1.96 \ \textit{and} \ |t2| \leq 1.96] \neq 95\%
\end{aligned}
$$

Why? Recall: the probability of incorrectly rejecting the null

$$
\begin{aligned}
&= P_{H_0}[|t1| > 1.96 \ \textit{and/or} \ |t2| > 1.96] \\
&= P_{H_0}[|t1| > 1.96, |t2| > 1.96] \\
&+ P_{H_0}[|t1| > 1.96, |t2| \leq 1.96] \\
&+ P_{H_0}[|t1| \leq 1.96, |t2| > 1.96] \\
&= P_{H_0}[|t1| > 1.96] \times P_{H_0}[|t2| > 1.96] \\
&+ P_{H_0}[|t1| > 1.96] \times P_{H_0}[|t2| \leq 1.96] \\
&+ P_{H_0}[|t1| \leq 1.96] \times P_{H_0}[|t2| > 1.96] \ \textit{(if t1, t2 are independent)} \\
&= 0.05 \times 0.05 + 0.05 \times 0.95 + 0.95 \times 0.05 = 0.0975
\end{aligned}
$$

Which is not the desired 5% !!

2.2.9 Tests multiples

Test de validité global du modèle

FM: $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$

Problème de test: $H_0 : \beta 1 = \beta_2 \ldots = \beta_p = 0$

H_1: au moins un paramètre $\neq 0$

$\rightarrow$ RM: $y_i = \beta_0 + \epsilon_i$, Il y a donc p restrictions.

Somme des carrés des erreurs

$$SSE(FM) = \sum_i (y_i - \hat{y}_i)^2$$

$$SSE(RM) = \sum_i (y_i - \bar{y})^2$$

Rappel de la décomposition de la variance:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

Et par définition du coefficient de détermination:

$$\begin{aligned} R^2 &= 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \\ &= 1 - \frac{SSE(FM)}{SSE(RM)} \end{aligned}$$

ou encore

$$\begin{aligned} R^2 &= 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \\ &= \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \\ &= \frac{SSE(RM) - SSE(FM)}{SSE(RM)} \end{aligned}$$

Statistique de test:

$$F = \frac{\frac{SSE(RM) - SSE(FM)}{p}}{\frac{SSE(FM)}{n-p-1}} = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

Loi sous H_0: $F \sim F_{p;n-p-1}$

RH_0 au niveau $\alpha = 5\%$ si $F > F_{p;n-p-1;0,95}$

Interprétation: Si on rejette H_0, au moins une des variables explicatives est significative mais attention cela ne vaut pas dire que le modèle est adéquat.

Remarque: On peut détecter un problème de quasi-multicollinéarite si

- les tests (simple) de student ne rejette pas H_0 pour toutes les variables explicatives
- le test de Fisher rejette H_0, et donc au moins une variable est significative

$\Rightarrow$ Contradiction dûe au problème de quasi-multicollinéarité

Test sur un ensemble de restrictions linéaires sur les paramètres de régression

But: simplifier le modèle en enlevant en une seule étape un groupe de $m < p$ variables explicatives (par exemple variable qualitative).

$$\text{FM: } y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Problème de test: $H_0 : \beta 1 = \beta_2 \ldots = \beta_m = 0$

H_1: au moins un paramètres $\beta 1, \ldots, \beta_m \neq 0$

$$\downarrow$$

$$RM : y_i = \beta_0 + \beta_{m+1} x_{im+1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Statistique de test:

$$F = \frac{\frac{SSE(RM) - SSE(FM)}{m}}{\frac{SSE(FM)}{n-p-1}} = \frac{\frac{R^2_{FM} - R^2_{RM}}{m}}{\frac{1-R^2_{FM}}{n-p-1}}$$

Ne pas rejeter H_0 implique que RM est aussi bon que FM et donc par soucis de parcimonie on garde RM

Test des contraintes sur sous-ensemble de paramètres de régression

Le test de Fisher pourra être appliqué de la même manière pour n'importe quelle contrainte sur un sous-ensemble de paramètres.
Exemples d'hypothèse nulle:

$H_0 : \beta_1 = \beta_2$ (1 restriction)

$H_0 : \beta_1 = \beta_2 = \beta_3$ (2 restrictions)

$H_0 : \beta_1 = \beta_2$ et $\beta_3 = \beta_4$ (2 restrictions)

$H_0 : 2\beta_1 = \beta_2$ et $\beta_3 = \beta_4$ (2 restrictions)

2.2.10 Méthode simple pour sélectionner des variables explicatives

Il existe une infinité de variables $X_1, X_2, \cdots$, qui peuvent expliquer la variable Y. Pour faciliter l'interprétation, on préfère un modèle plus parcimonieux $\Rightarrow$:

1. Ne pas inclure les variables qui sont non pertinentes d'un point de vue intuitif.
2. Estimez le modèle et calculez la p-value de chaque variable. Supprimez les variables avec une valeur > 0.2 ou > 0.3.
3. Estimez comme modèle final et vérifier la perte de qualité (R^2) par rapport au modèle initial

Méthode de sélection pas à pas : lméthode de sélection progressive et rétrograde.

2.2.11 Problème d'endogénéité: estimation par IV

Exemple:

Effet des études supérieures sur le revenu?

Modèle:

$$R_i = x_i'\beta + \delta S_i + \varepsilon_i$$

où S_i est le nombre d'années d'études.

Question: Peut-on utiliser les MCO pour estimer cette régression ?

Réponse: Non car il est raisonnable de penser que la variable S est endogène ($\text{corr}(S, \varepsilon) \neq 0$).

Méthodologie: 1) Tester le problème d'endogénité 2) Si ce problème existe, utiliser une autre méthode d'estimation.

Estimation par la méthode des variables instrumentales

• Besoin d'une variable instrumentale z:

$$\text{corr}(z, \varepsilon) = 0$$
$$\text{corr}(z, S) \neq 0$$

Par exemple pour S (la scolarité), z pourrait être le niveau d'éducation des parents ($Educ$).

L'estimateur OLS est basé sur les équations des moments:

$$E\{(R_i - (x_i'\beta + \delta S_i))x_i'\} = 0$$
$$E\{(R_i - (x_i'\beta + \delta S_i))S_i\} = 0$$

L'endogénéité de S implique que la deuxième contrainte doit être remplacée par:

$$E\{(R_i - (x_i'\beta + \delta S_i))z_i\} = 0$$

L'estimateur IV (variable instrumentale) est obtenu comme solution des équations:

$$E\{(R_i - (x_i'\hat{\beta}_{IV} + \hat{\delta}_{IV}S_i))x_i'\} = 0$$
$$E\{(R_i - (x_i'\hat{\beta}_{IV} + \hat{\delta}_{IV}S_i))z_i\} = 0$$

$$\Downarrow$$

$$\hat{\theta}_{IV} = (\hat{\beta}_{IV}\ \hat{\delta}_{IV}) = (Z'X)^{-1}Z'Y$$

où X est la matrice de design reprenant en colonnes les variables x et S, et Z est celle reprenant en colonnes les variables x et z

Théorème:

$$\hat{\theta}_{IV} \overset{a}{\sim} N(\theta, \sigma^2(Z'X)^{-1}(Z'Z)(X'Z)^{-1})$$

Question: Que faire si plusieurs variables endogènes et plusieurs variables instrumentales? (réponse plus tard)

Estimation par la méthode des doubles moindres carrés

Par exemple pour S (la scolarité) on peut prendre comme variables instrumentales le niveau d'éducation des parents (EM et EP).

X est la matrice de design reprenant en colonnes les variables x et S
Z est celle reprenant en colonnes les variables x, EM et EP.

- Etape 1:Projeter X sur Z:

$$\hat{X} = Z(Z'Z)^{-1}Z'X$$

- Etape 2: Régresser R sur $\hat{X}$:

$$R_i = x_i'\beta + \delta\hat{S}_i + \nu_i$$

L'estimateur TSLS est donné par:

$$\hat{\theta}_{TSLS} = (\hat{X}'X)^{-1}\hat{X}'Y = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

Test d'Hausman - Test d'endogénéité

Problem of test: H_0: no endogeneity

H_1: problem of endogeneity

Under H_0: $\hat{\theta}_{LS}$ is consistent and efficient (under normality) and $\hat{\theta}_{IV}$ is consistent but inefficient

Under H_1: $\hat{\theta}_{IV}$ is still consistent but not $\hat{\theta}_{LS}$

Construction of the statistics of test:

Known results:

$$\hat{\theta}_{LS} \overset{a}{\sim} N(\theta, \sigma^2(X'X)^{-1})$$

$$\hat{\theta}_{IV} \overset{a}{\sim} N(\theta, \sigma^2(Z'X)^{-1}(Z'Z)(X'Z)^{-1})$$

Define the statistics:

$$\hat{q} = \hat{\theta}_{IV} - \hat{\theta}_{LS}$$

and its covariance matrix under normality:

$$V(\hat{q}) = V(\hat{\theta}_{IV}) - V(\hat{\theta}_{LS})$$

The Hausman test statistic is defined as

$$H = \hat{q}' \left[\hat{V}(\hat{q})\right]^{-1} \hat{q}$$

where $\hat{V}(\hat{q})$ is a consistent estimator of $V(\hat{q})$.

Hausman (1978) shows that under the null, H is distributed asymptotically as a central χ^2_p where p is the number of unknown parameters.

Chapitre 3

SERIES CHRONOLOGIQUES

A Basic Course in Time Series

Session 1: Basic concepts and testing for stationarity

I Definitions

A *Stochastic Process* is a sequence of stochastic variables: $\ldots, Y_1, Y_2, Y_3, \ldots, Y_T \ldots$. We observe the process from $t = 1$ to $t = T$, yielding a sequence of numbers

$$y_1, y_2, y_3, \ldots, y_T$$

which we call a *time series*.

We only treat regularly spaced, discrete time series. Note that the observations in a time series are not independent! We need to rely on the concept of stationarity.

We say that a stochastic process is stationary if

1. $E[Y_t]$ is the same for all t
2. $\text{Var}[Y_t]$ is the same for all t
3. $\text{Cov}(Y_t, Y_{t-k})$ is the same for all t

Then we define the autocorrelation of order k as

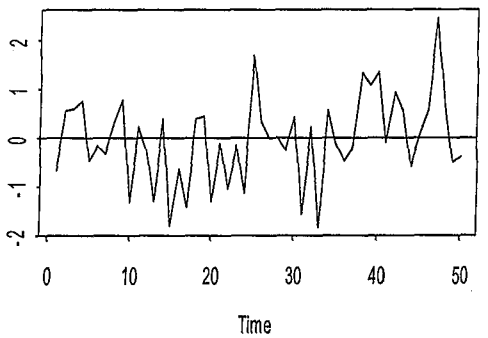
$$\rho_k = \text{Corr}(Y_t, Y_{t-k}) = \frac{\text{Cov}(Y_t, Y_{t-k})}{\text{Var}[Y_t]} = \frac{\gamma_k}{\gamma_0}$$

The autocorrelations give insight in the dependency structure of the process.

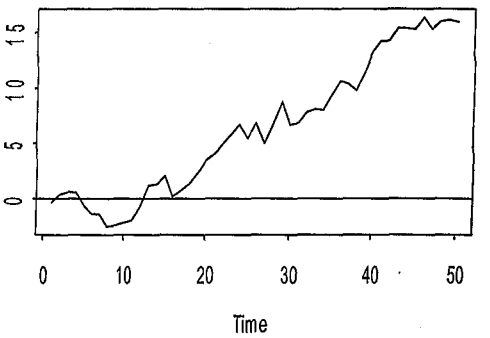
Exercice:

Which of them could be generated from a stationary proces? By means of which transformations can you make the other series stationary?

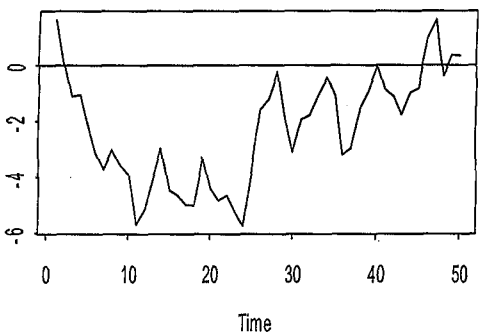
White Noise



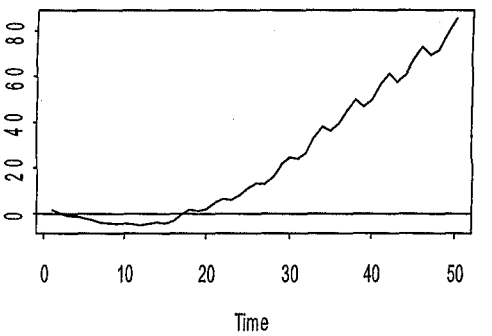
Trend



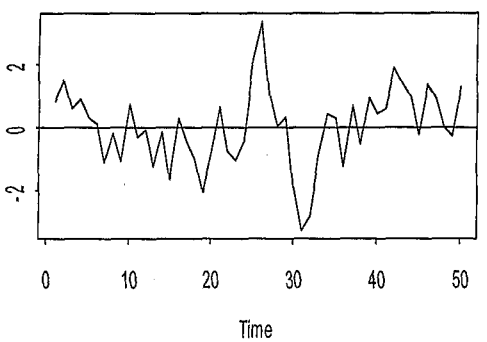
AR(1)



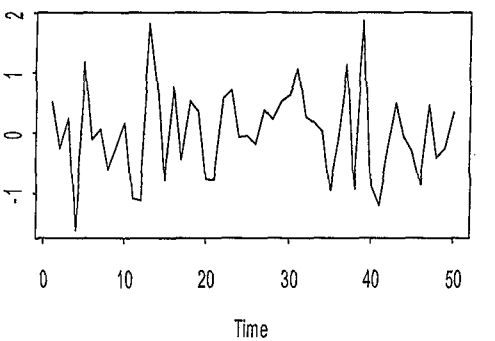
Trend+Saisonality



AR(1)



ARCH



Bruit Blanc - White Noise

<u>Definition</u> : Un processus stochastique $\{\varepsilon_t \mid t \in \mathbb{Z}\}$ est un bruit blanc (gaussien) quand :

* $\varepsilon_t \sim N(0, \sigma^2)$
* $\text{cov}(\varepsilon_t, \varepsilon_{t-k}) = 0 \quad \forall k = 1, 2, \dots$

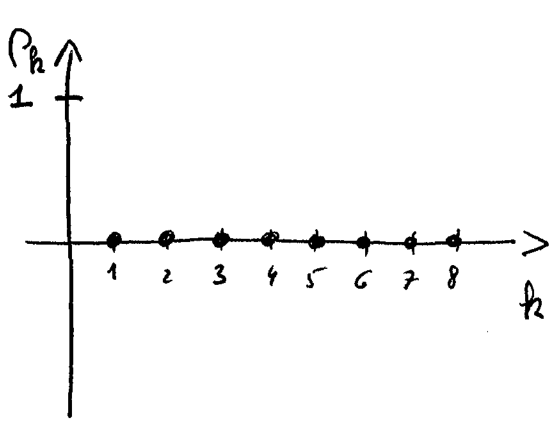
$\Rightarrow$ il n'y a pas de structure dans le temps

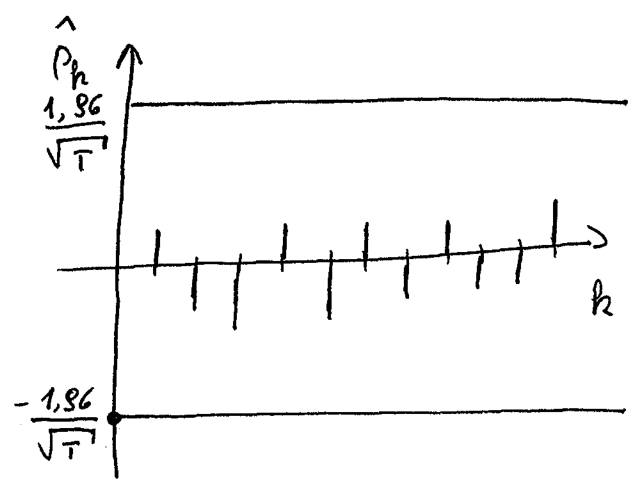
<u>Fonction d'autocorrélation</u> :

$\gamma_0 = \text{Var}(\varepsilon_t) = \sigma^2$

$\gamma_k = \text{Cov}(\varepsilon_t, \varepsilon_{t-k}) = 0 \qquad \forall k > 0$

$\Rightarrow \rho_0 = 1 \quad , \quad \rho_k = 0 \quad \forall k > 0$





estimation sur base

CORRELOGRAMME.

II The Correlogram

The autocorrelations can be estimated by

$$r_k = \hat{\rho}_k = \frac{\Sigma_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\Sigma_{t=1}^{T}(y_t - \bar{y})^2}.$$

A plot of $\hat{\rho}_k$ versus k is called a *correlogram*. On a correlogram, we often see 2 lines, corresponding to the critical values of the test statistic $\sqrt{T}\hat{\rho}_k$ for testing $H_0 : \rho_k = 0$ for a specific value of k.

The Q-Statistic, or Ljung-Box statistic, is testing

$$H_0 : \rho_1 = \rho_2 = \ldots = \rho_{kmax} = 0,$$

for a specified value of $kmax$. The Q-statistic is often used to test whether the series is coming from a *White Noise* stochastic proces, which is a sequence of iid stochastic variables. The choice of $kmax$ is not obvious.

Tests de bruit blanc

* Test individuel de bruit blanc

Problème de test : $\begin{cases} H_0 : \rho_k = 0 \\ H_1 : \rho_k \neq 0 \end{cases}$

* Statistique de test : $\sqrt{T} \cdot r_k$

Loi asymptotique sous H_0 : $\sqrt{T}\, r_k \underset{T \to \infty}{\overset{TCL}{\sim}} N(0,1)$

Règle de décision : $R H_0$ au niveau $\alpha = 5\%$ si

$$\left|\sqrt{T}\, r_k\right| \geq 1{,}96 \iff |r_k| > \frac{1{,}96}{\sqrt{T}}$$

* Remarques

- $r_k \xrightarrow[\text{proba}]{} \rho_k$
- $DW = 2(1 - r_1)$
- SAS utilise d'autres lois asymptotiques pour

* Test global du bruit blanc

Tests individuels appliqués à plusieurs retards k font apparaître des autocorrélations significatives pour des retards inexplicables.

P_q ? A cause du risque de 1ère espèce :

$$\alpha = P(R H_0 / H_0)$$

Donc si $K = 24$ et $\alpha = 5\%$ => on s'attend à $RH_0 \pm 24 \cdot \frac{5}{100} \approx 1$ fois même si le processus est vraiment 1 bruit blanc

=> Test global :

$$\begin{cases} H_0 : \rho_1 = \dots = \rho_K = 0 \\ H_1 : \exists \text{ au moins 1 } j \in \{1, \dots K\} \text{ t.q } \rho_j \neq 0 \end{cases}$$

<u>statistique de test</u> : $Q = T \sum_{k=1}^{K} \pi_k^2 \underset{T \to \infty}{\sim} \chi^2_K$

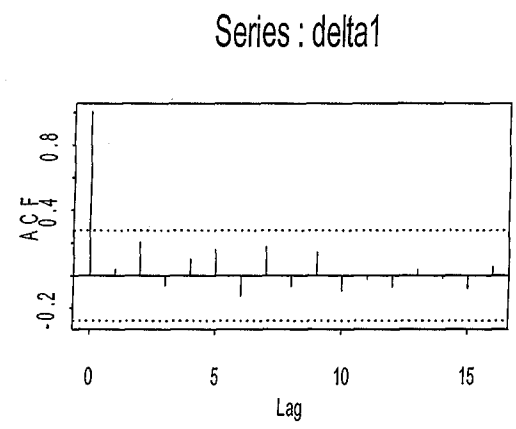
<u>Problèmes</u>

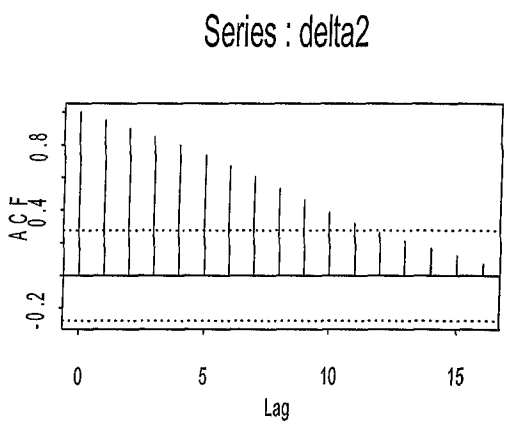
- Test peu puissant

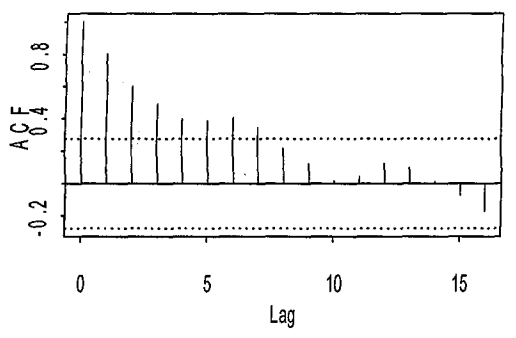
It will also turn out to be useful to look at the *Partial Correlations*

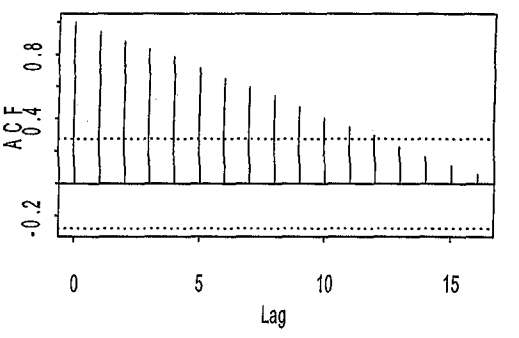
$$\pi_k = \text{Corr}(Y_t, Y_{t-k} | Y_{t-1}, \ldots, Y_{t-k+1})$$

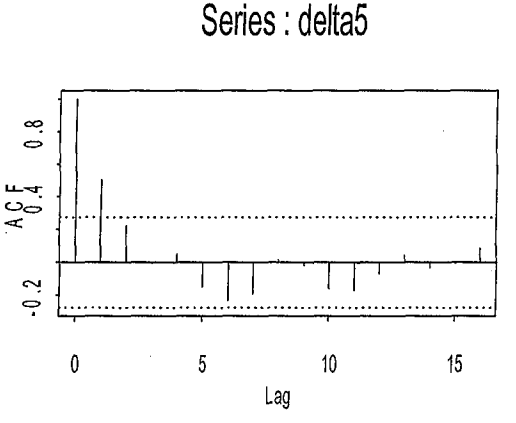
for $k = 0, 1, 2, \ldots$. This equals the correlation between Y_{t-k} and the residuals from a regression of Y_t on the variables $Y_{t-1}, \ldots, Y_{t-k+1}$.

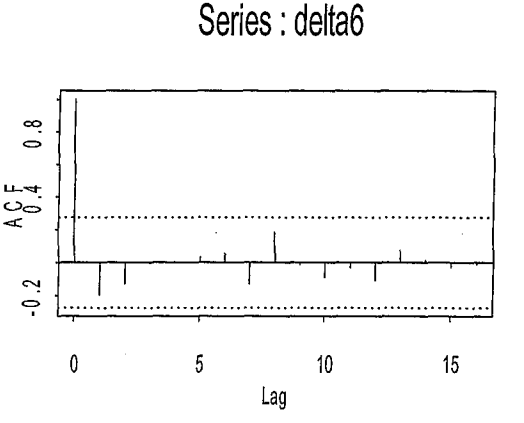

Series : delta1
ACF
0.8
0.4
-0.2
0
5
10
15
Lag


Series : delta2
ACF
0.8
0.4
-0.2
0
5
10
15
Lag


Series : delta3
ACF
0.8
0.4
-0.2
0
5
10
15
Lag


Series : delta4
ACF
0.8
0.4
-0.2
0
5
10
15
Lag


Series : delta5
ACF
0.8
0.4
-0.2
0
5
10
15
Lag


Series : delta6
ACF
0.8
0.4
-0.2
0
5
10
15
Lag

III Deterministic and stochastic trend

Consider the model

$$Y_t = a + \rho Y_{t-1} + \varepsilon_t, \qquad (1)$$

with ε_t a white noise stochastic process. Interpret ε_t as the innovation term or "shock."

Using (1) we can write

$$Y_t = a(1+\rho+\rho^2+\ldots)+\varepsilon_t+\rho\varepsilon_{t-1}+\rho^2\varepsilon_{t-2}+\rho^3\varepsilon_{t-3}+\ldots$$

$a = 0$

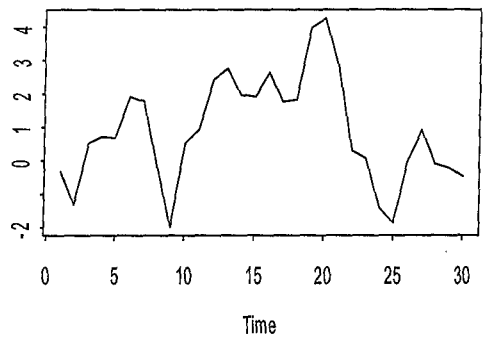
- $|\rho| > 1$, then past innovations may have an enormous influence. The series exploses.
- $|\rho| < 1 \rightarrow$ stationary proces. We call this process an autoregressive process of order 1, AR(1).
- $|\rho| = 1 \rightarrow$ Random Walk (variance increasing with t). The series is not stationary but $\Delta Y_t = Y_t - Y_{t-1}$ is.

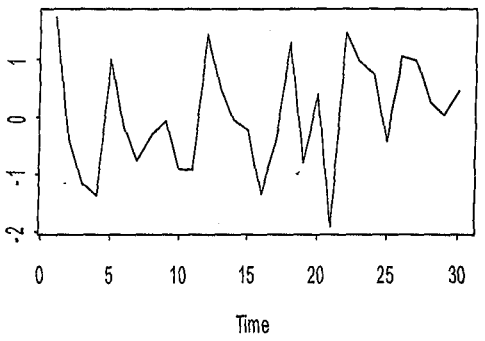
$\boxed{a \neq 0}$

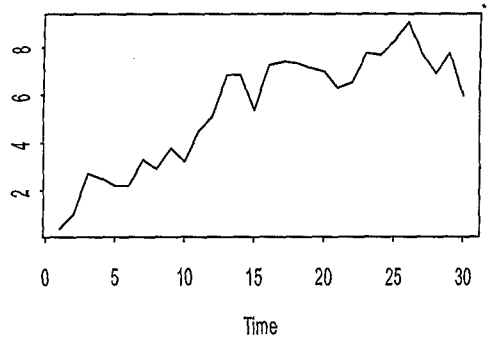
- $|\rho| > 1$, then past innovations may have an enormous influence. The series exploses.
- $|\rho| < 1$ the series tends to fluctuates around a constant with bounded variance around a constant. Series is stationary.
- $|\rho| = 1 \rightarrow$ Random Walk with drift. The series fluctuates with increasing variance around a straight line. There is a stochastic trend (as opposed to deterministic trend). The series is not stationary but $\Delta Y_t = Y_t - Y_{t-1}$ is.

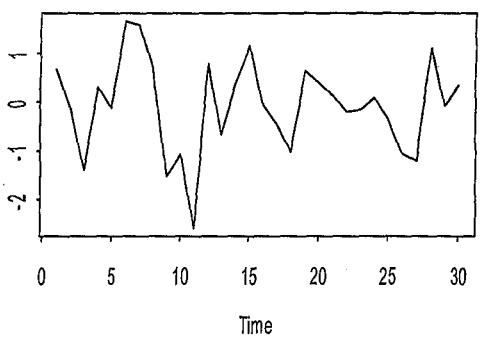
In case that $\rho = 1$, we say that the series has a *Unit Root.*

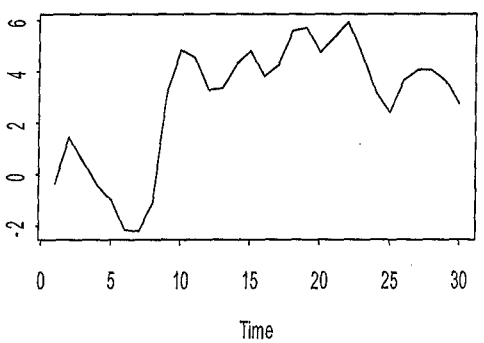
Exercice: Match the plots with (a) deterministic trend (b) stochastic trend (c) AR, $\rho = 0.3$ (d) AR, $\rho = 0.8$ (e) random walk (f) white noise

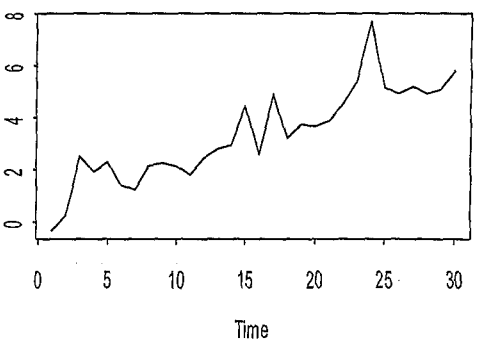

-2
0
1
2
3
4
0
5
10
15
20
25
30
Time


-2
-1
0
1
0
5
10
15
20
25
30
Time


2
4
6
8
0
5
10
15
20
25
30
Time


-2
-1
0
1
0
5
10
15
20
25
30
Time


-2
0
2
4
6
0
5
10
15
20
25
30
Time


0
2
4
6
8
0
5
10
15
20
25
30
Time

IV Testing for Unit roots

The classical test for $H_0 : \rho = 1$ in (1) is the Dickey-Fuller test. If H_0 is not rejected, econometricians will work with ΔY_t. Otherwise, they conclude that Y_t is stationary. It is important to notice that it is not allowed to perform the usual "t-test" here. One needs no use other critical values.

Nowadays, the Dickey-Fuller test is not recommended anymore, since it rejects H_0 not often enough (test is too conservative.)

The Augmented Dickey-Fuller (ADF) Test

Consider the model

$$\Delta Y_t = \mu + \gamma Y_{t-1} + \delta_1 \Delta Y_{t-1} + \ldots + \delta_q \Delta Y_{t-q} + \varepsilon_t$$

with ε_t white noise. We test

$$H_0 : \gamma = 0.$$

Again, we are not allowed to use simple "t-stats." The maximum "lag" needs to be specified.

The Phillips-Perron (PP) Test

The test regression for the PP test is:

$$\Delta Y_t = \mu + \gamma Y_{t-1} + \varepsilon_t$$

with ε_t any stationary proces (therefore PP is called a nonparametric test). We test

$$H_0 : \gamma = 0.$$

Session 2: ARIMA models

I Notations

The "Lag" operator L is defined as

$$LY_t = Y_{t-1}.$$

Note that $L^s Y_t = Y_{t-s}$.

The difference operator Δ is defined as

$$\Delta Y_t = (I - L)Y_t = Y_t - Y_{t-1}.$$

Trends can be eliminated by applying Δ once (for linear trend) or twice (for quadratric trend). If a stationary process is then obtained, we say that Y_t is integrated of order 1 or 2.

Seasonal effects of order s can be eliminated by applying the difference operator of order s:

$$\Delta_s Y_t = (I - L^s)Y_t = Y_t - Y_{t-s}$$

II Moving Average processes

A stationary stochastic process Y_t is a moving average of order 1, MA(1), if it satisfies

$$Y_t = a + u_t - \theta u_{t-1},$$

where a, and θ are unknown parameters.

The autocorrelations of an MA(1) are given by

- $\rho_0 = 1$
- $\rho_1 = \text{Corr}(Y_t, Y_{t-1}) = -\frac{\theta}{(1+\theta^2)}$
- $\rho_2 = 0$
- $\rho_3 = 0$
- ...

The correlogram can therefore be used to help us to *specify* an MA(1) proces:

MA(1) : Processus stationnaire ?

* $E[Y_t] = E[\ \ + u_t - \theta u_{t-1}] = a + E[u_t] - \theta E[u_{t-2}] = a.$

* $Var(Y_t) = Var(a + u_t - \theta u_{t-1}) = Var(u_t) + \theta^2 Var(u_{t-1})$

$= \sigma^2 + \theta^2\sigma^2 = \sigma^2(1+\theta^2) = \gamma_0$

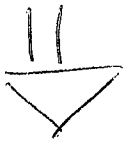
* $Cov(Y_t, Y_{t-2}) = Cov(a + u_t - \theta u_{t-1}, a + u_{t-1} - \theta u_{t-2})$

$= -\theta\ Cov(u_{t-1}, u_{t-2}) = -\theta\sigma^2 = \gamma_1$

: $Cov(Y_t, Y_{t-2}) = Cov(a + u_t - \theta u_{t-1}, a + u_{t-2} - \theta u_{t-3})$

$= \theta = \gamma_2$

* $Cov(Y_t, Y_{t-k}) = 0 \quad \forall k \geq 2$



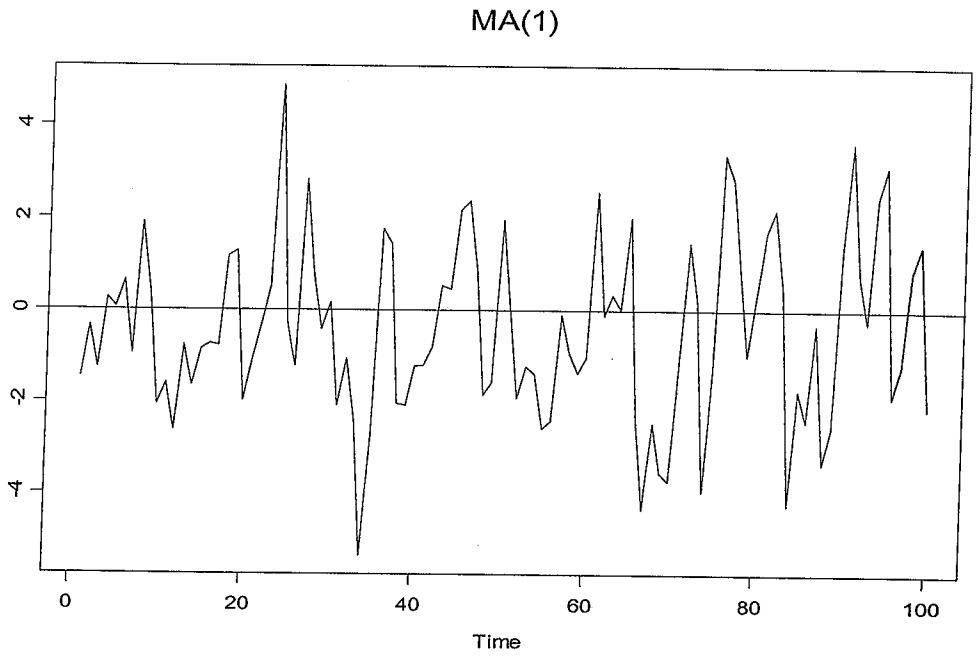
Autocorrélogramme :

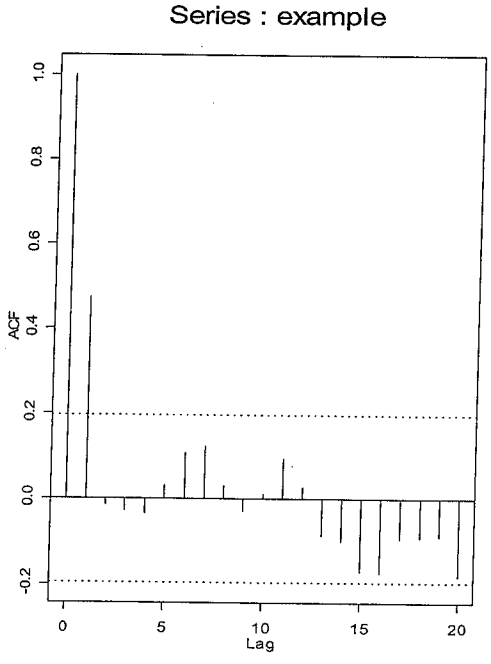
$\rho_0 = 1 = corr(Y_t, Y_t)$

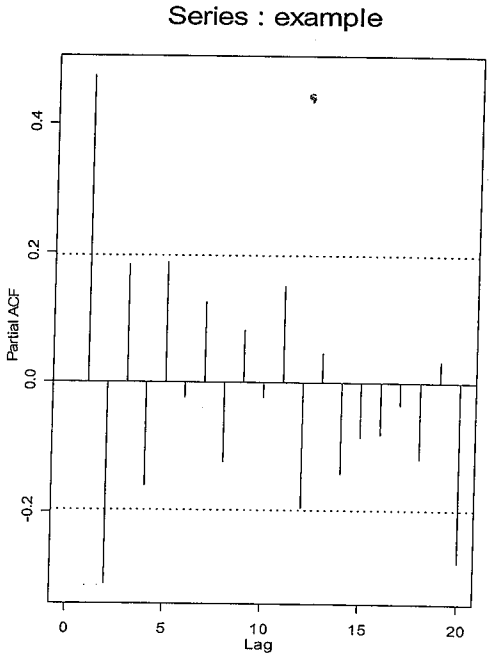
$\rho_1 = Corr(Y_t, Y_{t-1}) = \dfrac{Cov(Y_t, Y_{t-1})}{Var(Y_t)} = \dfrac{\gamma_1}{\gamma_0} = \dfrac{-\theta\sigma^2}{\sigma^2(1+\theta^2)}$

$= -\dfrac{\theta}{1+\theta^2}$

$\rho_2 = \rho_3 = \ldots = 0$


MA(1)
4
2
0
-2
-4
0
20
40
60
80
100
Time


Series : example
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20
Lag


Series : example
Partial ACF
0.4
0.2
0.0
-0.2
0
5
10
15
20
Lag

A stationary stochastic process Y_t is a moving average of order q, MA(q), if it satisfies

$$Y_t = a + u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \ldots - \theta_q u_{t-q},$$

where a, and $\theta_1, \ldots, \theta_q$ are unknown parameters.

The autocorrelations of an MA(q) process are equal to zero for lags larger than q. If the correlogram showns a strong decline and becomes non significant after lag q, then there is evidence that the series was generated by an MA(q) process

The obtained residuals should be close to a white noise. It is good practice to make a correlogram of the residuals, in order to *validate* an MA(q) model.

III Autoregressive processes

A stationary stochastic process Y_t is an autoregressive of order 1, AR(1), if it satisfies

$$Y_t = a + \phi Y_{t-1} + u_t, \qquad , |\phi| < 1$$

where a, and ϕ are unknown parameters.

The autocorrelations of an AR(1) are given by

- $\rho_0 = 1$
- $\rho_1 = \text{Corr}(Y_t, Y_{t-1}) = \phi$
- $\rho_2 = \phi^2$
- $\rho_3 = \phi^3$
- ...

On the other hand, the partial autocorrelations are given by

- $\pi_0 = 1$

Autocorrélogramme AR(1).

① Posons $a = 0$.

- $\gamma_0 = \mathrm{Var}(Y_t) = \mathrm{Var}(\phi Y_{t-1} + u_t) = \phi^2 \mathrm{Var}(Y_{t-1}) + \mathrm{Var}(u_t)$
$= \phi^2 \gamma_0 + \sigma^2$

$$\Rightarrow \boxed{\gamma_0 = \frac{\sigma^2}{1-\phi^2}} > 0 \quad \text{car} \quad |\phi| < 1$$

$$\gamma_k = \mathrm{cov}(Y_t, Y_{t-k}) = E[Y_t Y_{t-k}] = E[\phi Y_{t-1} Y_{t-k} + u_t Y_{t-k}]$$
$$= \phi \gamma_{k-1} + 0$$

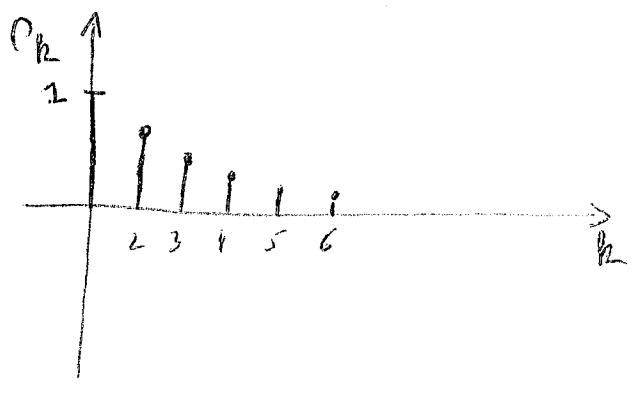
$$\Rightarrow \gamma_k = \phi \gamma_{k-1} = \phi^2 \gamma_{k-2} = \dots = \phi^k \gamma_0$$

$$\Downarrow$$

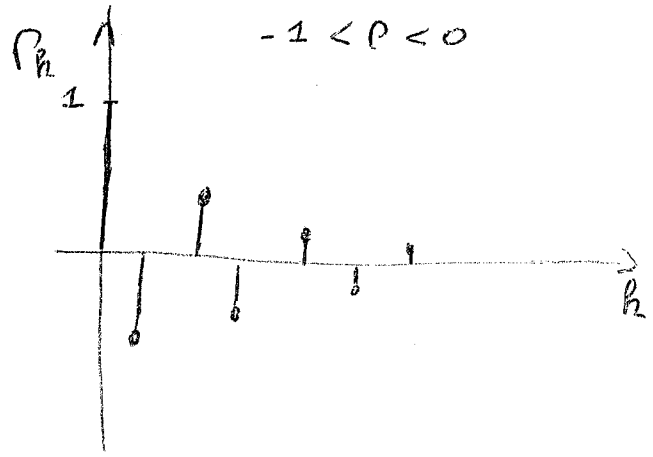
$$\boxed{\gamma_k = \phi^k \cdot \frac{\sigma^2}{1-\phi^2}}$$

Ce qui implique : $\boxed{\rho_k = \mathrm{cor}(Y_t, Y_{t-k}) = \frac{\gamma_k}{\gamma_0} = \phi^k}$.

$0 < \rho < 1$



$-1 < \rho < 0$



Posons $a \neq 0$: exercice

9.5.4 *Autocorrélations partielles*

L'autocorrélation partielle de retard k est définie à partir de la notion de *coefficient de corrélation partielle* entrevue dans le chapitre sur la régression multiple. Nous y avons mentionné que le coefficient de corrélation partielle entre X_1 et X_2, en éliminant l'influence de X_3, ..., X_r, peut être calculé comme un rapport entre deux déterminants.

On appelle autocorrélation partielle de retard k le coefficient de corrélation partielle entre Y_t et Y_{t-k} en éliminant l'influence de Y_{t-1}, ..., Y_{t-k+1}. On le note ici π_k. On peut montrer que les π_k peuvent s'obtenir en fonction des ρ_k à l'aide des formules suivantes :

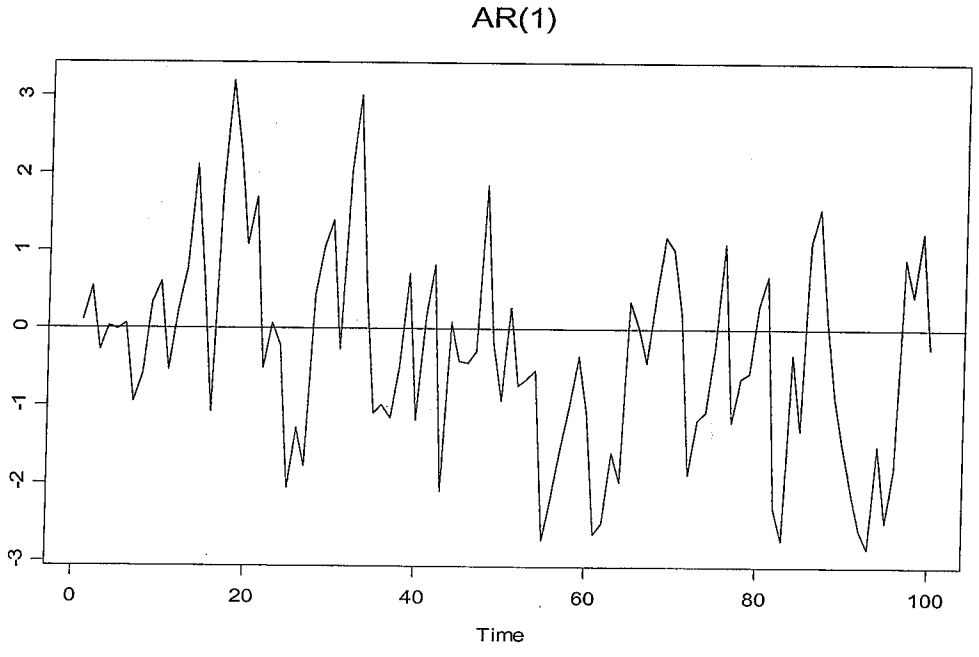
$$\pi_k = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & & & & & \vdots \\ \vdots & & & & & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & & & & & \vdots \\ \vdots & & & & & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}}$$

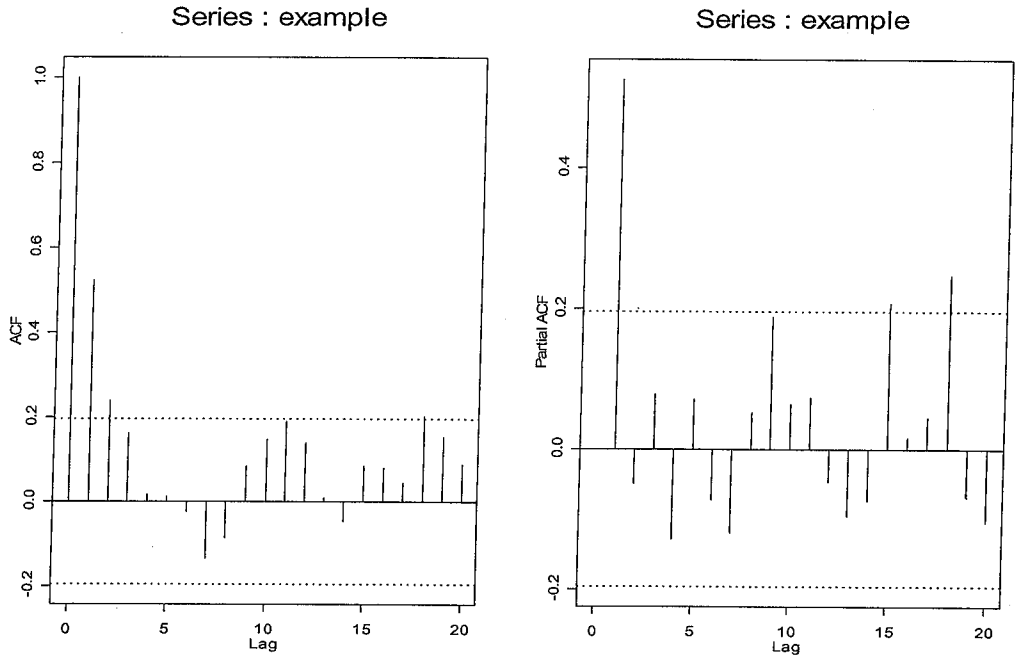
On a toujours $\pi_1 = \rho_1$. Pour un processus AR(1),

$$\pi_2 = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} = 0.$$

- $\pi_1 = \text{Corr}(Y_t, Y_{t-1}) = \phi$
- $\pi_2 = 0$
- $\pi_3 = 0$
- ...

The partial correlogram can therefore be used to *specify* an AR(1) proces:





A stationary stochastic process Y_t is an autoregressive of order p, AR(p), if it satisfies

$$Y_t = a + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} - \ldots + \theta_p Y_{t-p} + u_t$$

where a, and $\phi_1, \ldots, \phi_q$ are unknown parameters.

NB : Le processus $AR(p)_{20}$ sera stationnaire ssi les racines du polynôme : $1 - \phi_1 x - \ldots \phi_p x^p$ sont telles que $|x_i| > 1$.

The partial correlations of lag larger than p are equal to zero for an AR(p) process. The correlations tend more slowly to zero, and sometimes have a sinusoidal form.

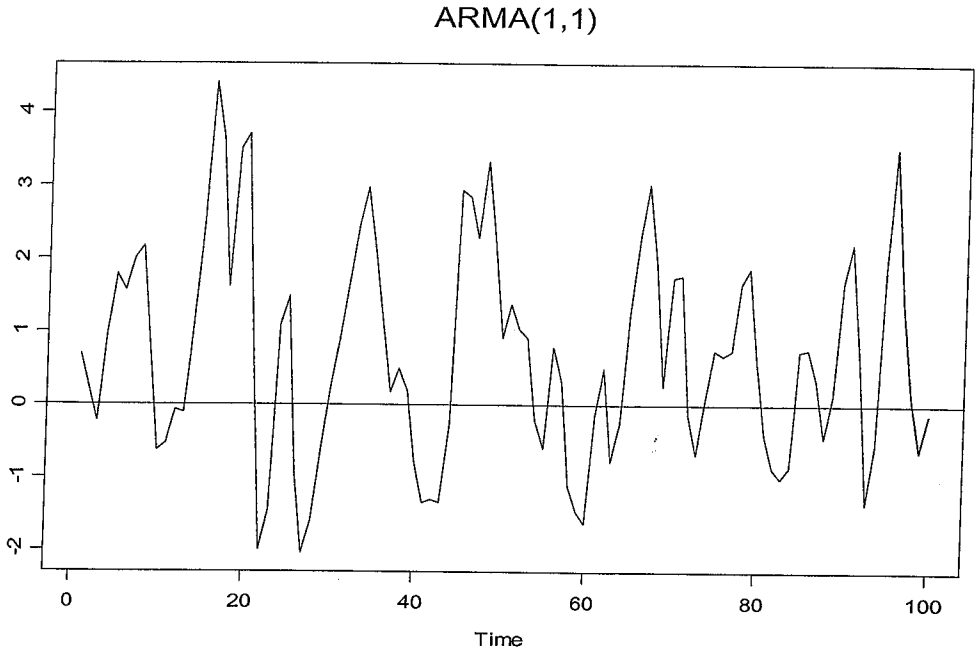
IV ARMA processes

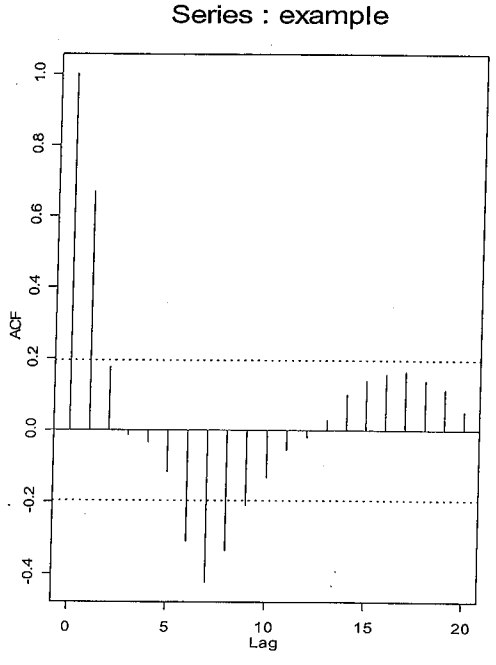
If neither the correlogram nor the partial correlogram "implode" to zero after a certain lag, then an ARMA specification may be appropriate.

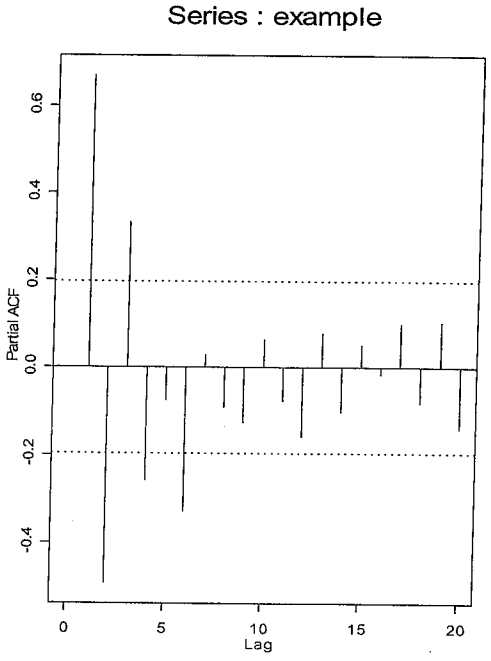
A stationary stochastic process Y_t is an $ARMA(p, q)$ if it satisfies

$$Y_t = a + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} - \ldots + \theta_p Y_{t-p} + u_t - \theta_1 u_{t-1} - \ldots - \theta_q u_{t-q}.$$

! Stick to simple models. Start with ARMA(1,1), then with ARMA(1,2) or ARMA(2,1). Validate the model by looking at the residuals.


ARMA(1,1)
4
3
2
1
0
-1
-2
0
20
40
60
80
100
Time


Series : example
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
5
10
15
20
Lag


Series : example
Partial ACF
0.6
0.4
0.2
0.0
-0.2
-0.4
0
5
10
15
20
Lag

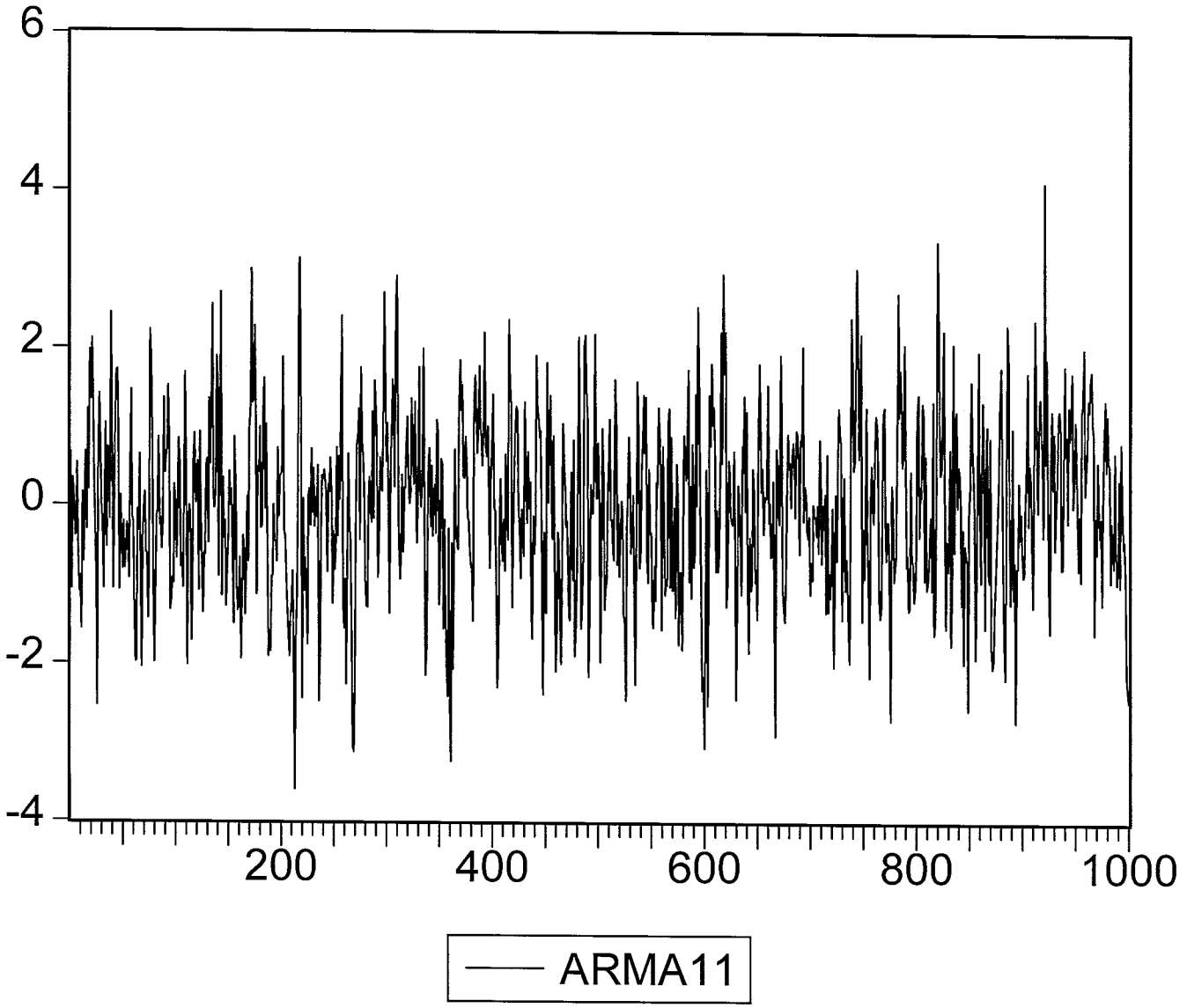
IV ARIMA processes

ARMA processes are stationary. Sometimes a series is not stationary, but it may be integrated of a certain order.

Definition: Y_t is an ARIMA(p,d,q) process if $\Delta^d Y_t$ is an ARMA(p,q) process.

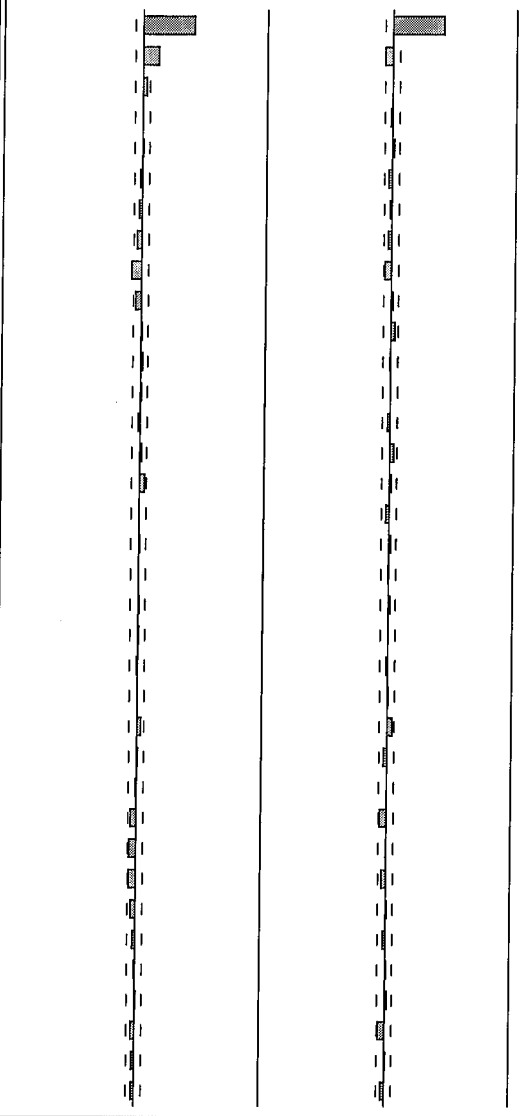
Recall that we prefer simple models, so low values for p and q. The order of integration d is often 0 or 1.

Exercice 1: On the next page you see the correlogram/partial correlogram of 3 time series of length 1000. Try to specify an appropriate ARMA model.

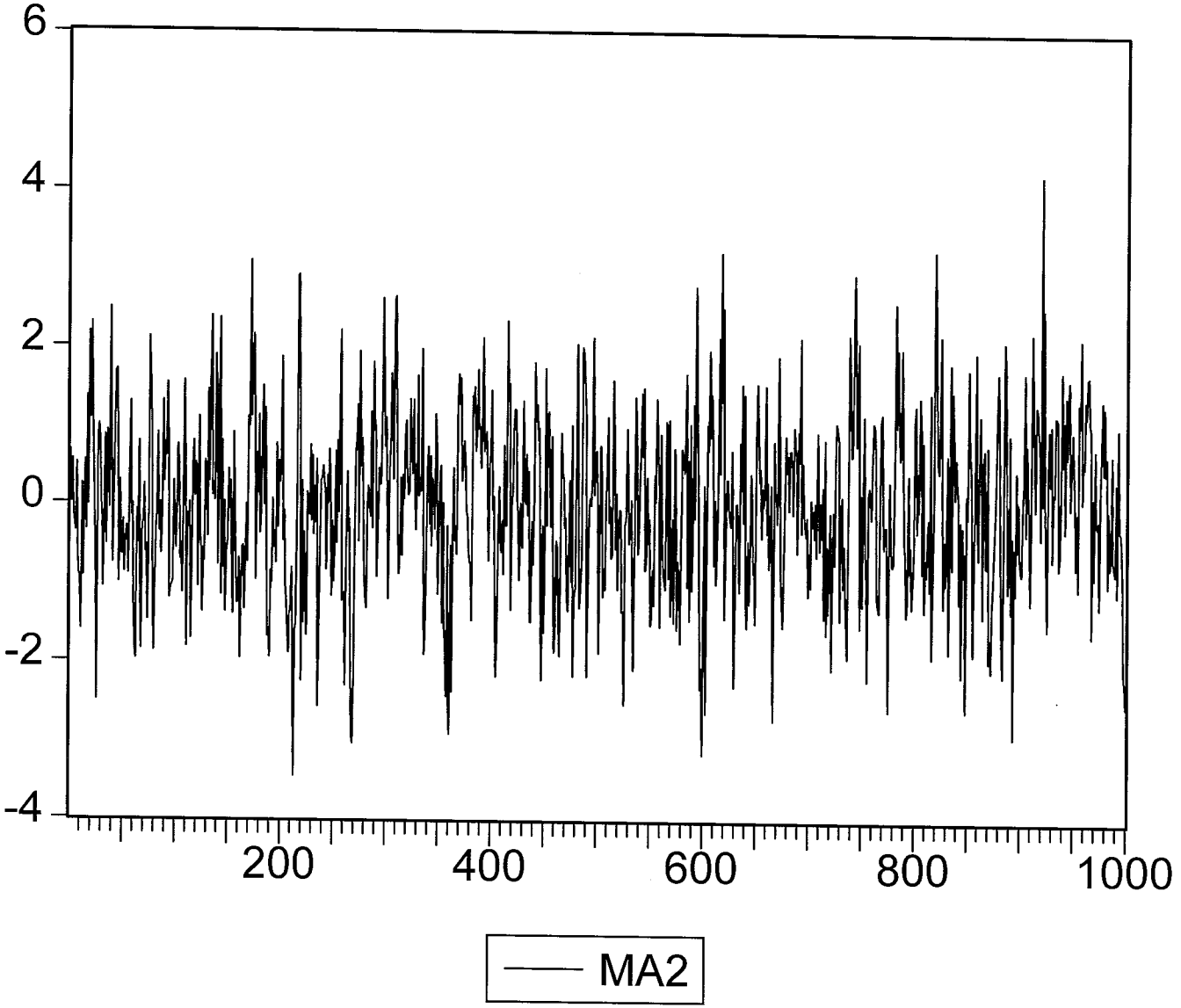

6
4
2
0
-2
-4
200
400
600
800
1000
ARMA11

Correlogram of ARMA11

Date: 11/08/04 Time: 14:25
Sample: 1 1000
Included observations: 1000

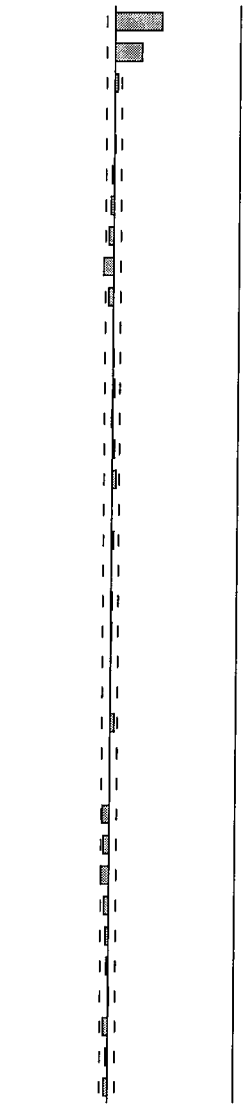
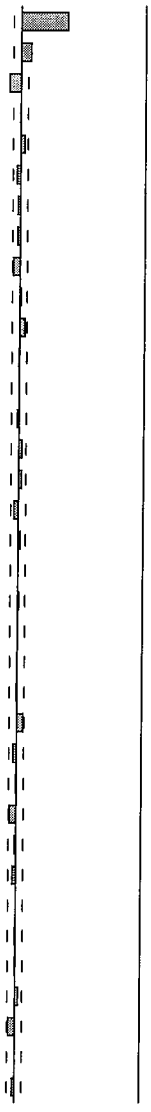


	AC	PAC	Q-Stat	Prob
1	0.432	0.432	187.56	0.000
2	0.135	-0.064	205.75	0.000
3	0.040	0.007	207.35	0.000
4	0.004	-0.010	207.37	0.000
5	0.013	0.021	207.55	0.000
6	-0.010	-0.028	207.66	0.000
7	-0.021	-0.009	208.10	0.000
8	-0.039	-0.030	209.64	0.000
9	-0.076	-0.056	215.45	0.000
10	-0.042	0.018	217.23	0.000
11	0.011	0.034	217.36	0.000
12	0.020	0.000	217.76	0.000
13	0.015	0.002	217.98	0.000
14	-0.009	-0.019	218.06	0.000
15	0.022	0.040	218.54	0.000
16	0.041	0.019	220.27	0.000
17	0.006	-0.031	220.31	0.000
18	0.016	0.024	220.56	0.000
19	0.008	-0.006	220.63	0.000
20	0.014	0.018	220.82	0.000
21	0.013	0.002	220.99	0.000
22	-0.006	-0.014	221.02	0.000
23	0.004	0.011	221.04	0.000
24	0.040	0.047	222.66	0.000
25	0.009	-0.028	222.73	0.000
26	0.000	0.002	222.73	0.000
27	-0.046	-0.057	224.94	0.000
28	-0.049	-0.007	227.45	0.000
29	-0.057	-0.035	230.80	0.000
30	-0.033	0.013	231.94	0.000
31	-0.021	-0.017	232.38	0.000
32	-0.008	0.010	232.44	0.000
33	0.010	0.020	232.55	0.000
34	-0.032	-0.056	233.64	0.000
35	-0.023	0.009	234.17	0.000
36	-0.021	-0.025	234.61	0.000

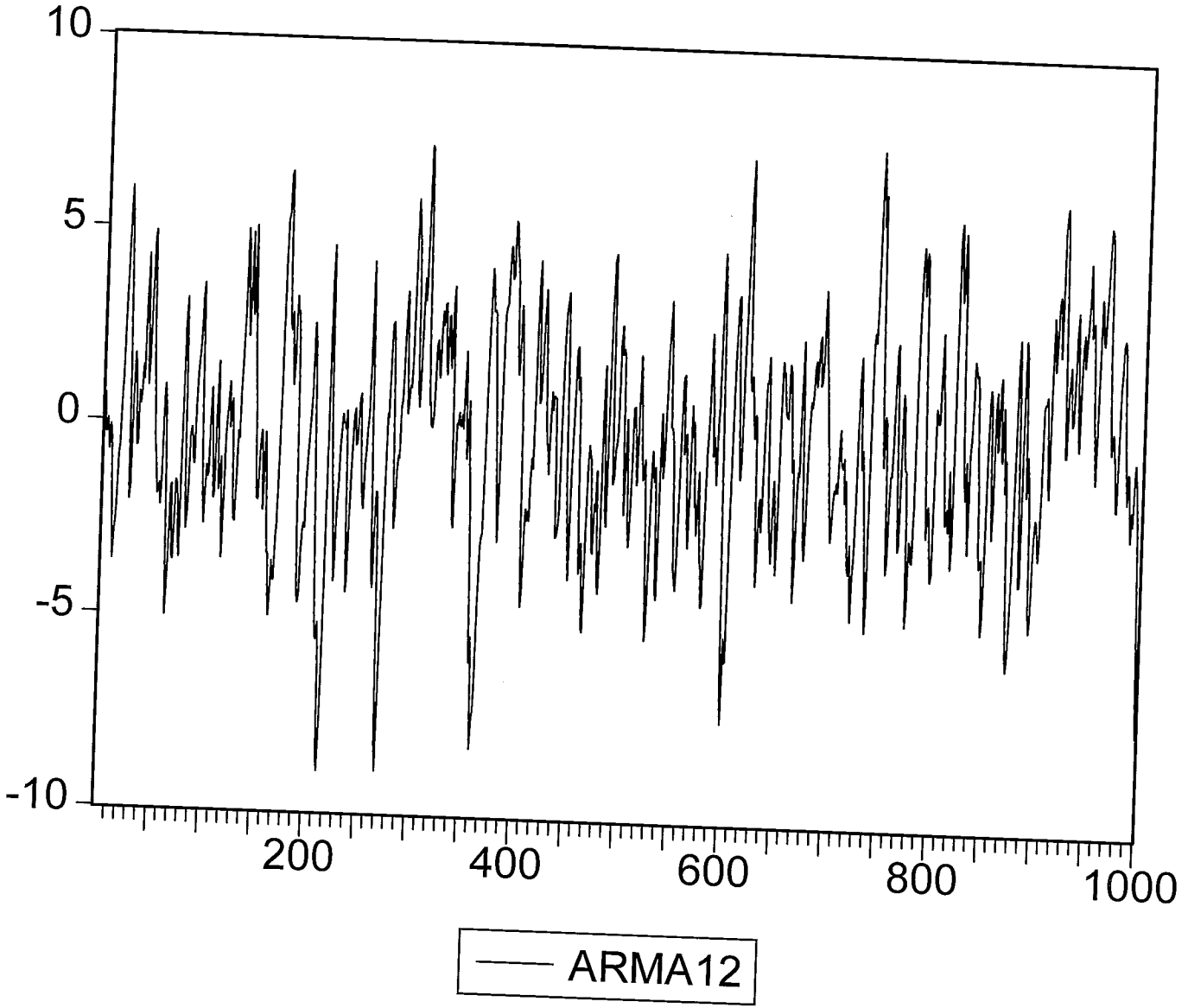

6
4
2
0
-2
-4
200
400
600
800
1000
MA2

Correlogram of MA2

Date: 11/08/04 Time: 14:26
Sample: 1 1000
Included observations: 1000

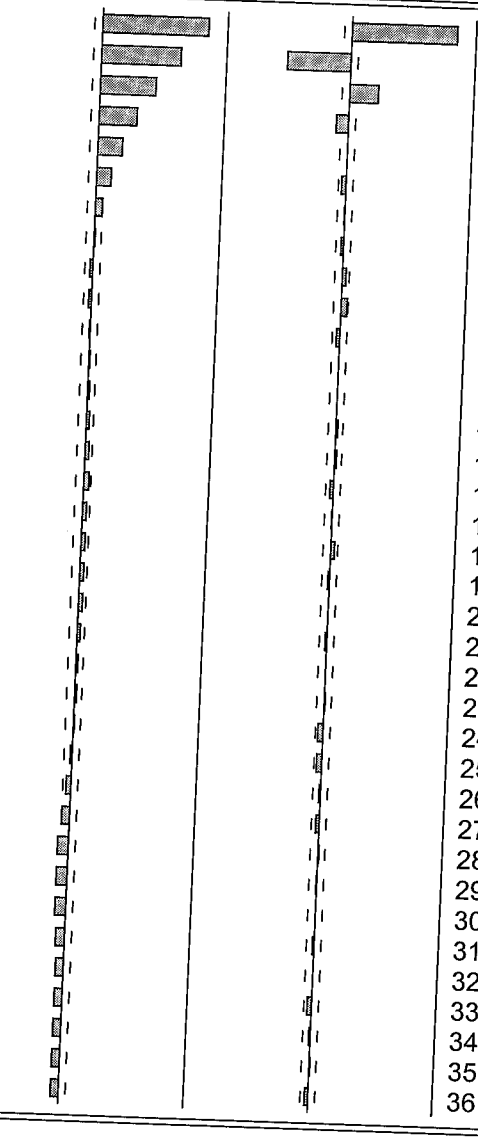



	AC	PAC	Q-Stat	Prob
1	0.388	0.388	151.27	0.000
2	0.225	0.087	202.14	0.000
3	0.033	-0.096	203.21	0.000
4	0.001	-0.001	203.21	0.000
5	0.015	0.039	203.42	0.000
6	-0.013	-0.032	203.60	0.000
7	-0.025	-0.024	204.21	0.000
8	-0.041	-0.019	205.89	0.000
9	-0.074	-0.055	211.41	0.000
10	-0.040	0.012	213.03	0.000
11	0.008	0.044	213.08	0.000
12	0.013	-0.007	213.25	0.000
13	0.020	0.003	213.65	0.000
14	-0.006	-0.014	213.68	0.000
15	0.022	0.030	214.16	0.000
16	0.040	0.030	215.82	0.000
17	0.006	-0.036	215.85	0.000
18	0.023	0.021	216.39	0.000
19	0.007	0.004	216.45	0.000
20	0.016	0.010	216.70	0.000
21	0.011	0.003	216.83	0.000
22	0.000	-0.006	216.83	0.000
23	0.001	-0.001	216.83	0.000
24	0.040	0.054	218.48	0.000
25	0.001	-0.030	218.48	0.000
26	0.007	-0.002	218.54	0.000
27	-0.052	-0.055	221.31	0.000
28	-0.044	-0.008	223.27	0.000
29	-0.061	-0.032	227.16	0.000
30	-0.034	0.007	228.36	0.000
31	-0.019	-0.004	228.74	0.000
32	-0.013	-0.005	228.90	0.000
33	0.012	0.028	229.05	0.000
34	-0.036	-0.057	230.37	0.000
35	-0.015	0.004	230.60	0.000
36	-0.027	-0.019	231.36	0.000


10
5
0
-5
-10
200
400
600
800
1000
ARMA12

Correlogram of ARMA12

Date: 11/08/04 Time: 14:26
Sample: 1 1000
Included observations: 1000



	AC	PAC	Q-Stat	Prob
1	0.883	0.883	781.47	0.000
2	0.662	-0.530	1221.5	0.000
3	0.464	0.238	1437.8	0.000
4	0.316	-0.101	1538.5	0.000
5	0.208	0.004	1581.9	0.000
6	0.124	-0.034	1597.3	0.000
7	0.059	-0.006	1600.8	0.000
8	0.011	-0.019	1600.9	0.000
9	-0.017	0.034	1601.2	0.000
10	-0.017	0.055	1601.5	0.000
11	-0.002	-0.026	1601.5	0.000
12	0.012	0.004	1601.7	0.000
13	0.020	0.003	1602.1	0.000
14	0.026	0.017	1602.8	0.000
15	0.035	0.018	1604.1	0.000
16	0.042	-0.025	1605.8	0.000
17	0.040	-0.001	1607.4	0.000
18	0.037	0.034	1608.9	0.000
19	0.036	-0.013	1610.2	0.000
20	0.033	0.007	1611.3	0.000
21	0.028	-0.011	1612.1	0.000
22	0.022	0.003	1612.6	0.000
23	0.018	0.015	1612.9	0.000
24	0.010	-0.044	1613.0	0.000
25	-0.009	-0.047	1613.1	0.000
26	-0.037	-0.010	1614.5	0.000
27	-0.064	-0.028	1618.8	0.000
28	-0.083	0.016	1625.8	0.000
29	-0.088	-0.005	1633.8	0.000
30	-0.083	0.008	1640.9	0.000
31	-0.072	-0.016	1646.3	0.000
32	-0.062	0.007	1650.3	0.000
33	-0.058	-0.040	1653.7	0.000
34	-0.059	-0.016	1657.3	0.000
35	-0.059	0.010	1661.0	0.000
36	-0.058	-0.026	1664.5	0.000

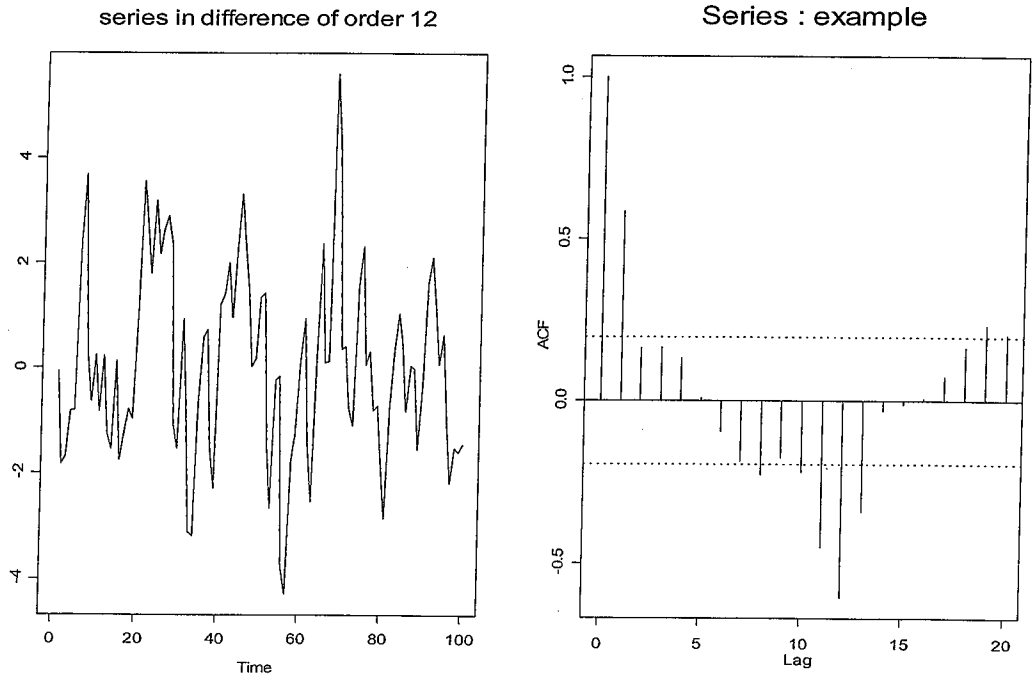
PROCEDURE

Try to specify an appropriate ARIMA(p,d,q) model. The next scheme could be followed

1. Make a time plot. Is the series stationay or do you need to take differences? In case of doubt, apply a unit root test.
2. Look at the correlogram ($\rightarrow$ MA) and the partial correlogram ($\rightarrow$ AR). If no clear structure, try ARMA.
3. Validate the model by examining the residuals. If necessary, respecify the model.

IV SARIMA processes

If there is a seasonal effect of order s (often $s = 12$ or $s = 4$) present in the series, then we need to apply the difference operator Δ_s to try to make the series stationary. Often, however, there will still remain some seasonality in the correlation structure of the series:



A parsimoneous way to model such correlation structures is by using SARMA models.

Example: SARMA(0,1)(0,1) model for $s = 12$

$$\begin{aligned} Y_t &= c + (I - \theta_1 L)(I - \Theta_1 L^{12})u_t \\ &= c + (I - \theta_1 L)(u_t - \Theta_1 u_{t-12}) \\ = c + u_t &- \Theta_1 u_{t-12} - \theta_1 u_{t-1} + \theta_1 \Theta_1 u_{t-13} \end{aligned}$$

where u_t is a white noise process:

Example: SARMA(1,0)(0,1) model for $s = 12$

$$\begin{aligned} (I - \phi_1 L)Y_t &= c + (I - \Theta_1 L^{12})u_t \\ \Rightarrow Y_t - \phi_1 Y_{t-1} &= c + (u_t - \Theta_1 u_{t-12} \\ \Rightarrow Y_t &= c + \phi_1 Y_{t-1} u_t - \Theta_1 u_{t-12} \end{aligned}$$

where u_t is a white noise process.

General definition: A stationary stochastic process Y_t is a SARMA(p,q)(P,Q) of order s if

$$(I - \phi_1 L - \ldots - \phi_p L^p)(I - \Phi_1 L^s - \ldots - \Phi_P L^{sP}) Y_t$$

$$= c + (I - \theta_1 L - \ldots - \theta_q L^q)(I - \Theta_1 L^s - \ldots - \Theta_Q L^{sQ}) u_t$$

where u_t is a white noise process.

Some guidelines for choosing the appropriate Arima-model:

Denote s the order of the seasonality. First try to see whether the beginning of the correlogram (or partial correlogram) indicates an MA(q) (or AR(p)) structure. If there are further significant (partial)correlations around s, then try a SARMA(0,q)(0,1) (or an SARMA(p,0)(1,0)). If there are also significant (partial) correlations around $2s$, then try a SARMA(0,q)(0,2) (or an SARMA(p,0)(2,0)).
If non of this seems to work, then try simple models combining AR and MA terms, as SARMA(1,0)(0,1), SARMA(1,1)(1,0), ...

Definition: Y_t is an SARIMA(p,d,q)(P,D,Q) process of order s if
$\Delta^d \Delta_s^D Y_t$ is a SARMA(p,q)(P,Q) process.

In practice, $d = 0, 1, 2$ and $D = 0, 1$.

Sesssion 3: Forecasting and the Box-Jenkins Approach

I Principles of forecasting

Using the series $y_1, \ldots, y_T$ we want to predict the values that $Y_{T+1}, \ldots, Y_{T+h}$ will take. We call h the horizon of the prediction. Prediction are made using

$$\hat{y}_s^{(T)} = E[Y_s | y_1, y_2, \ldots, y_T] \qquad (2)$$

for $s = T+1, \ldots, T+h$. To compute $\hat{y}_s$ we need to specify a *model* for the underlying stochastic process.

If y_s is observed, then the forecast error $y_s - \hat{y}_s^{(T)}$ can be computed.
In particular, the *one-step ahead forecast errors* are defined as

$$\hat{u}_s = \hat{y}_s^{(s-1)} - y_s.$$

They form a good approximation of the innovation process. Future innovations are predicted by their expected value 0.

Example: the AR(1) model, h=2

Since $Y_t = c + \phi Y_{t-1} + u_t$ we have

$$\hat{y}_{T+1}^{(T)} = c + \phi Y_T$$

and

$$\hat{y}_{T+2}^{(T)} = c + \phi \hat{y}_{T+1}^{(T)}.$$

Example: the MA(1) model, h=2

Since $Y_t = c + u_t + \theta u_{t-1}$ we have

$$\hat{y}_{T+1}^{(T)} = c + \theta \hat{u}_T$$

and

$$\hat{y}_{T+2}^{(T)} = c.$$

Example: the ARIMA(1,1,1) model, h=2

Denote $Z_t = \Delta Y_t$.

Since $Z_t = c + \phi Z_{t-1} + u_t + \theta u_{t-1}$ we have

$$\hat{Z}_{T+1}^{(T)} = c + \phi Z_T + \theta \hat{u}_T$$

and

$$\hat{Z}_{T+2}^{(T)} = c + \phi \hat{Z}_{T+1}^{(T)}.$$

Finally

$\hat{y}_{T+1}^{(T)} = y_T + \hat{Z}_{T+1}^{(T)}$ and $\hat{y}_{T+2}^{(T)} = \hat{y}_{T+1}^{(T)} + \hat{Z}_{T+2}^{(T)}$

Exercises:

1. How to predict values from a random walk?
2. What is the predicted value for an AR(1) model at horizon infinity? And for an MA(1) at horizon infinity?
3. How to predict from an SARIMA(1,1,0)(0,1,1) of order 12, upto horizon $h = 2$.

Remarks:

- In all the above prediction formulas, there are unknown parameters which still need to be estimated. This will be done computing the Least Squares estimator (only for AR-models) or the *Maximum Likelihood* estimator from the available data.
- There are existing formulas for the standard error (SE) around a forecast. The interval [prediction - 2 SE, prediction + 2 SE] yields then a 95% prediction interval.

III Model comparison

Sometimes the correlograms are not conclusive enough for the specification of a model. For discriminating between models, we can use:

A Information criteria

The *Mean Squared Error* equals

$$MSE = \frac{1}{T} \sum_{t=1}^{T} \hat{u}_t,$$

with $\hat{u}_t$ the one step ahead forecast errors.

Akaike info criterion (AIC)

$$AIC = \ln(MSE) + 2\frac{p}{T}$$

where p is the number of parameters in the model. The AIC penalizes for the complexicity of the model.

Schwarz criterion (SC)

$$AIC = \ln(MSE) + 2\frac{p\ln(T)}{T}$$

penalizes even more for the complexity of the model.

- Select the model with the smaller information criterion.
- Often used to determine the *order* of an AR(p) model.
- The criteria depend on the unit of measurement of Y_t

B Forecast Evaluation

A good model succeeds in making accurate forecasts. Forecast evaluation can be used to discriminate between models. Herefore, divide the sample $y_1, \ldots, y_T$ into two parts:

1. $y_1, \ldots, y_S$ will be used to make forecasts

2. $y_{S+1}, \ldots, y_T$ will be used as a validation sample. Forecast errors can now be computed as

$$\text{error}_t = \hat{y}_t^{(S)} - y_t$$

for $t = S+1, \ldots, T$

Root Mean Squared Error:

$$RMSE = \sqrt{\frac{1}{T-S} \sum_{t=S+1}^{T} \text{error}_t^2}.$$

Mean Absolute Error

$$MAE = \frac{1}{T-S} \sum_{t=S+1}^{T} |\text{error}_t|$$

Less sensible to huge errors.

Mean Absolute Percentage Error

$$MAPE = \frac{1}{T-S} \sum_{t=S+1}^{T} \frac{|\text{error}_t|}{y_t}$$

Has the advantage of begin independent of the scale of the series.

Theil Inequality Coefficient

$$\text{Theil} = \frac{RMSE}{\sqrt{\frac{1}{T-S}\Sigma_{t=S+1}^{T}\hat{y}_t^2} + \frac{1}{T-S}\Sigma_{t=S+1}^{T} y_t^2}$$

Always between 0 and 1. The smaller, the better.

Remark: The above Forecast Evaluations can also be computed using one step ahead forecasts (or *static* forecasts) instead of the multi-step

forecasts (or *dynamic forecasts*). Note, however, that as soon as y_s is unknown, the static forecast $y_{s+1}^{(s)}$ cannot be computed. For predicting future events, static forecasts are therefore not very useful.

IV Box-Jenkins Method

The Box-Jenkins Method for analyzing a time series consists of

1. Specification of a SARIMA model (using ...)
2. Estimation of the model
3. Validation of the model (using .....)

If the model is validated, then it can be used for prediction. If not, we need to specify another model.

Chapitre 4

MODELES POUR VARIABLE DEPENDANTE DICHOTOMIQUE

4.1 Codage de variables qualitatives

Problèmes relatifs à des caractères qualitatifs:

- le choix d'un parti aux élections
- réussir ou non son année d'étude
- être au chômage ou non
- acheter ou ne pas acheter un bien.

$\Longrightarrow$ Il faut introduire un codage.

Exemple: Le choix d'un parti où il y a 3 modalités: la gauche, le centre, la droite:

$$Y = \begin{cases} 1 & \text{si vote pour le parti de la gauche} \\ 2 & \text{si vote pour le parti du centre} \\ 3 & \text{si vote pour le parti de la droite} \end{cases}$$

ou encore

$Z = (Z_1, Z_2)$ où

$$Z_1 = \begin{cases} 1 & \text{si vote pour le parti de la gauche} \\ 0 & \text{sinon} \end{cases}$$

et

$$Z_2 = \begin{cases} 1 & \text{si vote pour le parti de la droite} \\ 0 & \text{sinon} \end{cases}$$

Definition: Une variable est dichotomique (binaire) si elle ne prend que deux modalités disjointes. Dans le cas général, elle est dite polytomique.

4.2 Exemple de prévision d'une variable binaire en utilisant une autre variable binaire

Soit l'évènement

$$Y = \begin{cases} 0 & \text{si j'achète une voiture d'occasion} \\ 1 & \text{si j'achète une voiture neuve} \end{cases}$$

La variable aléatoire Y obéit à une loi de Bernouilli.

BUT: étudier ce comportement d'achat par rapport au sexe de l'acheteur:

$$X = \begin{cases} 0 & \text{si acheteur féminin} \\ 1 & \text{si acheteur masculin} \end{cases}$$

Soit un échantillon de taille $n = 100$ résumé dans un tableau de contingence:

	$X=1$	$X=0$	$\sum$
$Y=1$	40	20	60
$Y=0$	20	20	40
$\sum$	60	40	100

Estimation des probabilités conditionnelles théoriques de la variable Y sachant X:

$$\hat{P}(Y=1|X=0) = \frac{\hat{P}(Y=1 \text{ et } X=0)}{\hat{P}(X=0)} = \frac{20/100}{40/100} = \frac{1}{2}$$
$$\hat{P}(Y=1|X=1) = \frac{\hat{P}(Y=1 \text{ et } X=1)}{\hat{P}(X=1)} = \frac{40/100}{60/100} = \frac{2}{3}$$
$$\hat{P}(Y=0|X=0) = 1 - \hat{P}(Y=1|X=0) = 1 - \frac{1}{2} = \frac{1}{2}$$
$$\hat{P}(Y=0|X=1) = 1 - \hat{P}(Y=1|X=1) = 1 - \frac{2}{3} = \frac{1}{3}$$

Les rapport des probabilités (odds-ratio) résument

- le comportement des femmes:

$$\frac{P(Y=1|X=0)}{P(Y=0|X=0)} = \frac{1/2}{1/2} = 1$$

$\Rightarrow$ une femme a autant d'attirance pour les voitures neuves que pour celles d'occasion.

- le comportement des hommes:

$$\frac{P(Y=1|X=1)}{P(Y=0|X=1)} = \frac{2/3}{1/3} = 2$$

$\Rightarrow$ un homme aura deux fois plus tendance à acheter une neuve qu'une occasion.

Cet exemple était très simple car:

- une seule variable explicative
- variable explicative dichotomique.

MAIS quid si plusieurs variables explicatives (continues, dichotomiques, ordinales) ??

Exemple: Le choix des collèges aux USA classés en deux catégories: collège privé ($Y = 1$) ou public ($Y = 0$). Le choix de l'étudiant va dépendre de plusieurs facteurs tels que:

- le revenu (quantitative)
- la catégorie socio-professionnelle des parents (ordinale)
- les convictions idéologiques (nominale)
- le genre (dichotomique)

$$\Downarrow$$

Modéliser

$$P(Y = 1|X_1, \ldots, X_p)$$

en fonction des variables explicatives sur base d'un échantillon $(y_i, x_{1i}, ..., x_{ip})$ où $1 \leq i \leq n$.

L'absence de continuité (et souvent aussi l'absence d'ordre naturel) entre les modalités que peut prendre la variable dépendante nécessite une autre approche que le modèle de régression linéaire:

$$y_i = x_i'\beta + \varepsilon_i \qquad\qquad i = 1, \ldots, n$$

où β est un vecteur de p paramètres inconnus et où ε_i est la perturbation associée à la ième observation.

Pourquoi ne pas utiliser un tel modèle?

- Le nuage des points se trouve sur deux espaces parallèles $y = 0$ et $y = 1$ et ne peut donc pas être approché par un seul hyperplan de régression.
- y_i est qualitatif, et $x_i'\beta + \varepsilon_i$ est quantitatif.
- Comme y_i ne prend que deux valeurs, il en est de même pour la perturbation ε_i, ce qui contredit l'hypothèse de normalité.
- $E(y|x) = P(y = 1|x) * 1 + P(y = 0|x) * 0 = P(y = 1|x) \in [0, 1]$ or le modèle impose $E(y|x) = x'\beta \in IR$.
- $Var(y|x) = E(y^2|x) - E(y|x)^2 = P(y = 1|x) * 1^2 + P(y = 0|x) * 0^2 - p(x)^2 = p(x) - p(x)^2 = p(x)(1 - p(x))$, il y a donc hétéroscédasticité.

Le modèle linéaire n'est donc pas approprié pour des variables dépendantes dichotomiques !!!

4.3 Modèle dichotomique simple

Soit une enquête sur n individus étudiant la propension à acheter une voiture:

$$y = \begin{cases} 1 & \text{si achat de voiture} \\ 0 & \text{si pas d'achat de voiture.} \end{cases}$$

On suppose que la personne a un comportement rationnel, c'est-à-dire qu'il va acheter une voiture si son utilité est plus grande que son coût:

$$U_i \geq C_i \Rightarrow y_i = 1$$
$$U_i < C_i \Rightarrow y_i = 0$$

ou autrement dit

$$y_i* = U_i - C_i \geq 0 \Rightarrow y_i = 1$$
$$y_i* = U_i - C_i < 0 \Rightarrow y_i = 0$$

L'utilité ainsi que y_i* sont des quantités continues mais non observables: ce sont des variables dite LATENTES.

Modélisation des probabilités conditionnelles d'acheter une voiture en imposant un lien linéaire entre les variables explicatives et la variable latente:

$$y_i* = \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i = \beta' x_i + \varepsilon_i$$

On déduit:

$$P(y_i = 1|x_i) = P(y_i* \geq 0|x_i) = P(\beta' x_i + \varepsilon_i \geq 0|x_i) = P(\varepsilon_i \geq -\beta' x_i|x_i)$$

En posant la symétrie pour la loi des erreurs:

$$P(y_i = 1|x_i) = P(\varepsilon_i \leq \beta' x_i|x_i) = F(\beta' x_i).$$

Le modèle est donc:

$$P(y_i = 1|x_i) = F(\beta' x_i) \in [0, 1]$$

où $F(t) = P(\varepsilon_i \leq t)$ est la fonction de répartition des erreurs.

Remarques:

• La distribution conditionnelle $y_i|x_i$ est complètement connue si on connait β. Le but sera donc d'estimer β. On utilisera la méthode de maximum de vraisemblance.

• Fonction F de répartition: on l'a supposée symétrique donc $F(0) = \frac{1}{2}$.

Donc si $x_i'\beta \geq 0 \Longrightarrow P(y_i = 1|x_i) \geq 0.5 \Longrightarrow y_i = 1$

Et si $x_i'\beta < 0 \Longrightarrow P(y_i = 1|x_i) < 0.5 \Longrightarrow y_i = 0$

• Dernière information nécessaire pour complèter le modèle: choix de la forme de F:

$F(t) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{t} \exp(-t^2/2)dt \Longrightarrow$ Modèle PROBIT

$F(t) = \frac{1}{1+\exp(-t)} \equiv \frac{\exp(t)}{1+\exp(t)} \Longrightarrow$ Modèle LOGIT

$F(t) = 1 - \exp(-\exp(t)) \Longrightarrow$ Modèle extrême value

(aussi appelé *complementary log-log*)

4.4 Estimation par la méthode du maximum de vraisemblance

Probabilité d'apparition d'un individu de l'échantillon:

$$P(y_i = d|x_i) = P(y_i = 1|x_i)^{y_i} P(y_i = 0|x_i)^{1-y_i}$$

où $d = \{0, 1\}$.

La vraisemblance du vecteur $y = (y_1, \ldots, y_n)'$ (c'est-à-dire la probabilité d'apparition de l'échantillon):

$$\begin{aligned} L(y, \beta) &= \prod_{i=1}^{n} P(y_i = 1|x_i)^{y_i} P(y_i = 0|x_i)^{1-y_i} \\ &= \prod_{i=1}^{n} F(x_i'\beta)^{y_i} (1 - F(x_i'\beta))^{1-y_i}. \end{aligned}$$

La fonction de log-vraisemblance est donnée par:

$$\begin{aligned} \log L(y, \beta) &= \sum_{i=1}^{n} \{y_i \log F(x_i'\beta) + (1 - y_i) \log(1 - F(x_i'\beta))\} \\ &= \sum_{i=1|y_i=1}^{n} \log F(x_i'\beta) + \sum_{i=1|y_i=0}^{n} \log(1 - F(x_i'\beta)). \end{aligned}$$

L'estimateur du maximum de vraisemblance est défini par:

$$\hat{\beta}_{MV} = \underset{\beta}{\operatorname{argmax}} \log L(\beta).$$

Pour trouver le maximum, dérivons la fonction de log-vraisemblance par rapport au vecteur β des paramètres inconnus et l'annulation de cette dérivée donnera l'équation de vraisemblance:

$$0 = \frac{\partial}{\partial \beta} \log L(\beta)|_{\hat{\beta}_{MV}}$$

$$0 = \sum_{i=1|y_i=1}^{n} \frac{f(x_i'\beta)}{F(x_i'\beta)} \frac{\partial}{\partial \beta}(x_i'\beta) - \sum_{i=1|y_i=0}^{n} \frac{f(x_i'\beta)}{1 - F(x_i'\beta)} \frac{\partial}{\partial \beta}(x_i'\beta)|_{\hat{\beta}_{MV}}$$

où

$$\frac{\partial}{\partial \beta}(x_i'\beta) = \begin{array}{l} \frac{\partial}{\partial \beta_1}(x_i'\beta) = x_{i1} \\ \frac{\partial}{\partial \beta_2}(x_i'\beta) = x_{i2} = x_i \in IR^p \\ \ldots \\ \frac{\partial}{\partial \beta_p}(x_i'\beta) = x_{ip} \end{array}$$

La solution est donc donnée par le système:

$$\sum_{i=1}^{n}\{\frac{y_i - F(x_i'\beta)}{F(x_i'\beta)(1-F(x_i'\beta))}f(x_i'\beta)\}x_i|_{\hat{\beta}_{MV}} = 0 \in IR^p$$

où $f(t)$ est la densité associée à $F(t)$.

Remarques

• Le terme entre accolade est souvent appelé résidu généralisé du modèle.

• Les conditions de 1er ordre sont nécessaires mais non suffisantes (pour Probit et Logit, la fonction $\log L(\beta)$ est strictement concave).

• La solution de l'équation de vraisemblance n'existe pas toujours. En effet, les estimateurs MV ne seront pas définis si un hyperplan sépare l'espace parfaitement en 2 parties.

• Système de p équations non linéaires à résoudre $\Rightarrow$ résolution numérique (Newton Raphson).

4.5 Interprétation des paramètres

Modèle dichotomique:

$$P(y_i = 1|x_i) = F(\beta' x_i) = F(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + \beta_p x_{ip})$$

BUT: Pouvoir expliquer un changement marginal d'une variable. Si on augmente x_j d'une unité, la probabilité d'une réussite augmente de:

$$\frac{\partial}{\partial x_j} F(\beta' x_i) = f(\beta' x_i)\beta_j \neq \beta_j$$

Mais cette augmentation dépend de toutes les caractéristiques de l'individu i au travers de x_i

$$\Downarrow$$

Le paramètre β_j (valeur numérique) ne possède pas une interprétation naturelle.

Néanmoins le signe est interprétable car $f(\beta' x_i)$ est toujours positif, donc:

$$signe(\frac{\partial}{\partial x_j} F(\beta' x_i)) = signe(\beta_j).$$

Exemple. Supposons que $y_i = 1$ si achat d'une voiture et x_i est le revenu de l'individu i. Soit le modèle estimé

$$\hat{P}(y_i = 1|x_i) = F(-0.5 + 0.39x_i)$$

On peut dire que si le revenu augmente alors la probabilité d'acheter une voiture augmente.

NB: Dans le modèle Logit, on peut avoir des interprétations plus fines , on pourra tout de même interpréter les valeurs des β.

En effet, en inversant la relation

$$P(y_i = 1|x_i) = F(x_i'\beta),$$

on obtient

$$x_i'\beta = F^{-1}(P(y_i = 1|x_i)) = \log(\frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)})$$

donc,

$$\beta_j = \frac{\partial}{\partial x_j}(x_i'\beta) = \frac{\partial}{\partial x_j}(\log(\frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)})$$

Dans ce modèle on peut dire que β_j est le changement marginal du 'log odds ratio' par rapport à x_j, et donc que $\exp(\beta_j)$ est le changement marginal du odds-ratio.

Exemple. Reprenons l'exemple de la voiture:

$$\hat{P}(y_i = 1|x_i) = F(-0.5 + 0.39x_i)$$

$\exp(0.39) = 1.48$ et donc si le revenu augmente de 1 unité, alors le odds ratio augmente de 48%.

4.6 Mesure de la qualité de l'ajustement

- Pourcentage d'observations bien classées: Après l'étape d'estimation, on classe les n individus:

$$\hat{P}(y_i = 1|x_i) \geq 0.5 \Longrightarrow \hat{y}_i = 1$$
$$\hat{P}(y_i = 1|x_i) < 0.5 \Longrightarrow \hat{y}_i = 0$$

	$y_i = 1$	$y_i = 0$
$\hat{y}_i = 1$	n_{11}	n_{10}
$\hat{y}_i = 0$	n_{01}	n_{00}

Pourcentage des observations bien classées:

$$\frac{n_{00} + n_{11}}{n}.$$

Si ce pourcentage est inférieur à 50% c'est très mauvais car pire que le hasard.

Simple mais cette mesure a le défaut de considérer de la même manière une 'petite erreur' (pe $y_i = 1$ et $\hat{P}(y_i = 1|x_i) = 0.49$) et 'une grande erreur' (pe $y_i = 1$ et $\hat{P}(y_i = 1|x_i) = 0.01$).

- R^2 de Mc Fadden: Mesure de qualité comparant 2 modèles:
  - - Modèle complet :$P(y_i = 1|x_i) = F(\beta' x_i) \Longrightarrow \hat{\beta}_{MV}$
  - - Modèle nul :$P(y_i = 1) = F(\beta_1) \Longrightarrow \hat{\beta}_R$

  R^2 basé sur les fonctions de log vraisemblances:

$$R^2 = 1 - \frac{\log L(\hat{\beta}_{MV})}{\log L(\hat{\beta}_R)}.$$

  Si R^2 est proche de zéro, cela veut dire que l'apport des variables explicatives est presque nul et donc que le modèle est mauvais, par contre si R^2 est proche de 1 alors le modèle ajuste bien les données.

4.7 Test du rapport de vraisemblance - Test sur plusieurs paramètres

Il existe trois tests qui exploitent le principe du maximum de vraisemblance:

- le test de Wald
- le test du score (du multiplicateur de Lagrange)
- le test du rapport de vraisemblance.

Ces trois tests sont asymptotiquement équivalents. Et dans la pratique c'est le test du rapport de vraisemblance qui est préféré.

• Problème de test: Le but est de tester un ensemble de restrictions définies par la fonction

$$c : IR^p \to IR^k : \beta \to c(\beta)$$

où k est le nombre de restrictions.

Exemple 1: $H_0 : c(\beta) = 0$ avec

$$c(\beta) = (\beta_2 \;\; \beta_3 \;\; \ldots \;\; \beta_p)',$$

on a donc $p - 1$ restriction:

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_p = 0$$

Exemple 2: $H_0 : c(\beta) = 0$ avec

$$c(\beta) = (\beta_2 + \beta_3 \quad \beta_p)',$$

on a donc 2 restrictions:

$$H_0 : \beta_2 + \beta_3 = 0 \text{ et } \beta_p = 0$$

Idée: Comparer la vraisemblance du modèle complet:

$$P(y_i = 1|x_i) = F(\beta_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$$

avec la vraisemblance du modèle sous H_0, c'est-à-dire lorqu'on a appliqué les restrictions imposées par H_0.

Puisque la vraisemblance de l'échantillon est plus grande dans le modèle complet (plus libre) que dans le modèle restreint, on aura:

$$L(\hat{\beta}_{MV}) \geq L(\hat{\beta}_R)$$

et donc

$$|\log L(\hat{\beta}_{MV})| \leq |\log L(\hat{\beta}_R)|$$

- Statistique de test:

$$\xi_R = -2(\log L(\hat{\beta}_R) - \log L(\hat{\beta}_{MV}))$$

- Loi sous H_0: $\xi_R \approx \chi^2_k$

- Règle de décision: RH_0 au niveau α si

$$\xi_R > \chi^2_{k,1-\alpha}$$

4.8 Extensions

• Régression logistique polytomique - Modèle logit multinomial (p.e. Variable dépendante: mode de transport pour aller au travail (voiture, bicyclette, bus, train).

• Modèle logit polytomique ordonné (p.e. Variable dépendante: note globale pour avis pédagogiques (Très Favorable, Favorable, Défavorable, Très Défavorable).

• Modèle dichotomique emboité (p.e. Variable dépendante: Fumeur ou non fumeur - et dans le cas de fumeur Light-Non Light).

4.9 Références

- Albert, A. & Anderson, J. A. (1984), "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models", *Biometrika*, 71, 1-10.
- Andersen, E. (1997), *Introduction to the Statistical Analysis of Categorical Data*, Springer-Verlag, Berlin.
- Agresti, A. (1990), *Categorical Data Analysis*, Wiley-Interscience, New York.
- Hosmer, D. & Lemeshow S. (1989), *Applied Logistic Regression*, Wiley-Interscience, New York.
- Lloyd, C. (1999), *Statistical Analysis of Categorical Data*, Wiley-Interscience, New York.

Chapitre 5

MODELES POUR VARIABLE DEPENDANTE DE COMPTAGE

5.1 Introduction

Nombreuses caractéristiques prennent comme valeurs des nombres entiers positifs:

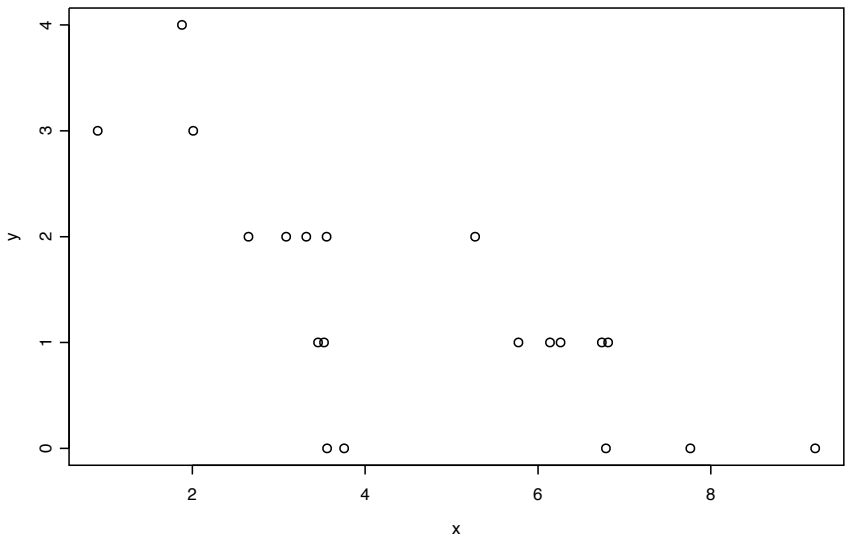
- le nombre d'arrivées journalières dans un aéroport
- le nombre d'accidents du travail
- le nombre de faillites dans un secteur industriel
- le nombre de brevets déposés par des firmes
- etc

But: Expliquer la relation entre une variable réponse discrète et des variables explicatives.

Exemple Soit un échantillon de taille n où $\{(x_i, y_i)|x_i \in IR, y_i \in IN\}$ avec

- $y_i =$ le nombre d'accidents du travail de l'entreprise i sur une année
- $x_i =$ un indice de la qualité du social dans l'entreprise.

Question: "Quel est le nombre attendu d'accidents du travail pour une firme ayant un indice de qualité du social x_i ?"



5.2 Modélisation linéaire ?

Modèle linéaire:

$$y_i = x_i'\beta + \varepsilon_i$$

Ce modèle est inadéquat pour différentes raisons:

- le nuage des points n'a pas une structure linéaire
- $x_i'\beta \in IR$ ce qui est en contradiction avec $E[Y|X] \geq 0$
- pour x_i fixe, $\varepsilon_i = y_i - x_i'\beta$ ne prend qu'un ensemble discret de valeurs $\Rightarrow$ contradiction avec l'hypothèse de normalité des erreurs.

$$\Downarrow$$

Si y décrit le nombre de fois qu'un évenement s'est produit pendant une certaine période , on utilisera le modèle de Poisson.

5.3 Modèle de Poisson simple

Soient y_i $(i = 1, \ldots, n)$ les n observations de la variable discrète à valeurs dans IN.

Hypothèse 1: Les variables Y_i sont supposées indépendantes avec comme loi des lois de Poisson de paramètre λ_i.

Hypothèse 2: Les paramètres λ_i sont liés aux valeurs prises par p variables explicatives de la manière suivante:

$$\lambda_i = \exp(x_{i1}\beta_1 + \ldots + x_{ip}\beta_p) = \exp(x_i'\beta)$$

où $x_i = (x_{i1} \ldots x_{ip})'$ et β est le vecteur des paramètres inconnus.

$$\Downarrow$$

La moyenne conditionnelle de y_i sachant x_i est donnée par:

$$E[Y_i|x_i] = \exp(x_i'\beta).$$

Remarque: $E[Y_i|x_i] \geq 0$ quelque soit β.

La méthode d'estimation classique pour estimer le vecteur β est la méthode de maximum de vraisemblance.

5.4 Estimation par maximum de vraisemblance

Echantillon: $\{(x_i, y_i)|1 \leq i \leq n\}$, $y_i \in IN$ et $x_i = (x_{i1}, \ldots, x_{ip})'$

Soit Y_i la variable aléatoire associée à l'observation y_i: $Y_i \sim P(\lambda_i)$ où $\lambda_i = \exp(x_i^t \beta)$.

La probabilité associée à la valeur observée est:

$$L(y_i) = P(Y_i = y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

La fonction de vraisemblance est donc:

$$L(\beta, y_1, \ldots, y_n) = \prod_{i=1}^{n} L(y_i) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \frac{e^{-\sum_{i=1}^{n} \lambda_i} \prod_{i=1}^{n} \lambda_i^{y_i}}{\prod_{i=1}^{n} y_i!}.$$

La fonction de log vraisemblance:

$$\begin{aligned} \log L(\beta, y_1, \ldots, y_n) &= -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} y_i \log \lambda_i - \sum_{i=1}^{n} \log(y_i!) \\ &= -\sum_{i=1}^{n} \exp(x_i^t \beta) + \sum_{i=1}^{n} y_i (x_i^t \beta) - \sum_{i=1}^{n} \log(y_i!) \end{aligned}$$

L'estimateur de maximum de vraisemblance est:

$$\hat{\beta}_{MV} = \underset{\beta}{\operatorname{argmax}} \log L(\beta, y_1, \ldots, y_n).$$

Condition du 1er ordre:

$$\frac{\partial}{\partial \beta} \log L(\beta, y_1, \ldots, y_n)_{|\hat{\beta}_{MV}} = 0 \in I\!R^p.$$

Dérivons par rapport à β_j $(1 \leq j \leq p)$:

$$\begin{aligned}
&\frac{\partial}{\partial \beta_j} \log L(\beta, y_1, \ldots, y_n) \\
= & -\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \exp(x_i^t \beta) + \frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} y_i (x_i^t \beta) \\
= & -\sum_{i=1}^{n} \exp(x_i^t \beta) \frac{\partial}{\partial \beta_j} (x_i^t \beta) + \sum_{i=1}^{n} y_i \frac{\partial}{\partial \beta_j} (x_i^t \beta) \\
= & -\sum_{i=1}^{n} \exp(x_i^t \beta) x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \\
= & \sum_{i=1}^{n} x_{ij} [y_i - \exp(x_i^t \beta)].
\end{aligned}$$

On obtient donc le système de p équations non linéaires à résoudre:

$$\sum_{i=1}^{n} x_i[y_i - \exp(x_i^t\beta)]|_{\hat{\beta}_{MV}} = 0.$$

Pour la condition du 2ème, il faudrait calculer la matrice des dérivées secondes (matrice Hessienne), et prouver la concavité en montrant que celle-ci est définie négative

La concavité de cette fonction a 2 implications:

• la condition du 1er ordre est une condition nécessaire et suffisante (un seul maximum)

• pas de problème de convergence pour les algorithmes itératifs (p.e. Newton Raphson).

5.5 Inférence statistique: Tests

5.5.1 Problème de test sur 1 paramètre

L'estimateur de maximum de vraisemblance des paramètres inconnus a comme loi asymptotique une normale p-variée:

$$\hat{\beta} \approx N(\beta, V(\hat{\beta}))$$

où $V(\hat{\beta}) = I(\beta)^{-1}$ avec $I(\beta)$ la matrice d'information de Fisher. La matrice d'information de Fisher est donnée par

$$\begin{aligned} I(\beta) &= -E[\frac{\partial}{\partial\beta}\frac{\partial}{\partial\beta}\log L(\beta)] = -E[H(\beta)] \\ &= -E[-\sum_{i=1}^{n}(x_i x_i^t)\exp(x_i^t\beta)] \\ &= \sum_{i=1}^{n}(x_i x_i^t)\exp(x_i^t\beta). \end{aligned}$$

Nous pouvons donc obtenir une estimation de la matrice variance covariance des estimateurs:

$$\hat{V}(\hat{\beta}) = [\sum_{i=1}^{n}(x_i x_i^t)\exp(x_i^t\hat{\beta})]^{-1} \in IR^{p\times p}.$$

Pour l'estimateur de β_j, on obtient donc:

$$SE(\hat{\beta}_j) = \sqrt{\hat{V}(\hat{\beta})_{jj}} = \sqrt{[\sum_{i=1}^{n}(x_i x_i^t)\exp(x_i^t\hat{\beta})]^{-1}_{jj}}.$$

• Ceci nous permet de construire un intervalle de confiance approximatif à 95% pour β_j:

$$IC(\beta_j) = \hat{\beta}_j \pm 2SE(\hat{\beta}_j)$$

• Problème de test est $H_0 : \beta_j = 0$ avec comme alternative $H_1 : \beta_j \neq 0$ (la variable explicative est-elle significative ?)

On rejette l'hypothèse nulle à α% si

$$|t| = |\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}| > z_{1-\frac{\alpha}{2}}$$

5.5.2 Problème de test sur plusieurs paramètres - Test du rapport de maximum de vraisemblance

$Y_i \sim P(\lambda_i)$ où $\lambda_i = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3)$

Testons $H_0 : \beta_1 = \beta_2 = 0$. Notons que si H_0 n'est pas rejetté, aucune des variables explicatives n'est pertinente pour expliquer la moyenne conditionnelle de Y_i

Pour ce test, nous comparons les log vraisemblances du modèle complet et du modèle restreint (modèle où H_0 est appliqué) avec l'aide de la statistique de test:

$$\xi = -2\{\log L(\hat{\beta}_R) - \log L(\hat{\beta}_{MV})\}.$$

où β_R est le vecteur des paramètres inconnus du modèle restreint

Sous H_0 $\xi \sim \chi^2_{r=2}$, donc rejet de l'hypothèse nulle au niveau $\alpha\%$ si $\xi > \chi^2_{2,1-\alpha}$.

5.6 Références

- Cameron, A. C. (1998), *Regression analysis of count data*, Cambridge University Press, NY.
- Greene, W. H. (1997), *Econometric Analysis*, Prentice Hall, London.
- Winkelmann, R. (1997), *Econometric analysis of count data*, Springer, Berlin.

Chapitre 6

MODELES DE REGRESSION CENSURES

6.1 Introduction

Quelques exemples dans la littérature économique étudiant des données censurées:

• Achat par les ménages de biens durables (Tobin, 1958)

• Nombre d'heures de travail des femmes (Quester & Greene, 1982)

• Dépenses en vacances

La variable dépendante est nulle pour une fraction non négligeable des observations.

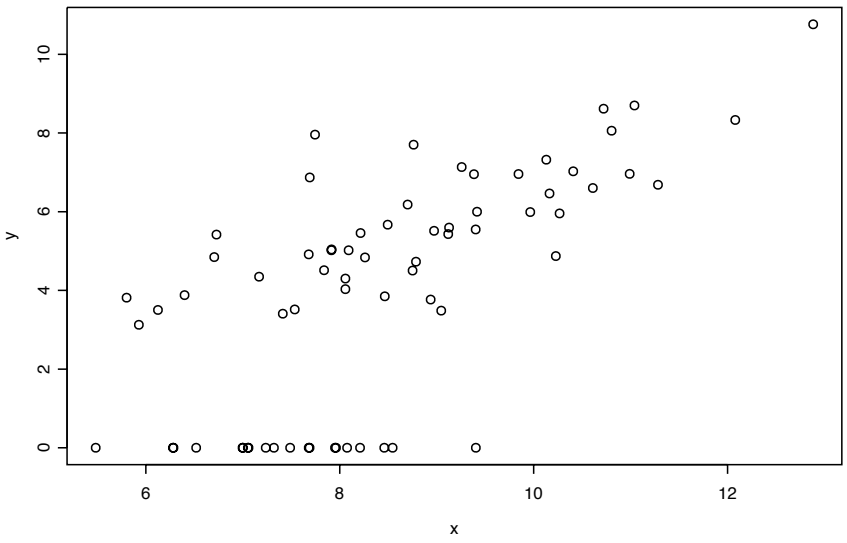
Achat par les ménages de biens durables.

Soit

- D_t la somme consacrée par ménage pendant une certaine période de temps $[t-1, t]$ à l'achat d'un bien durable (une voiture, TV, etc)
- R_t le revenu du ménage à l'instant t

$$\Downarrow$$

D_t a des valeurs sur les réels positifs, mais sa valeur est nulle pour beaucoup de ménages.



D_t en fonction du revenu des ménages.

L'utilisation du modèle de régression simple

$$D_t = \beta R_t + \alpha + \varepsilon$$

sur ce type de données est inadéquat essentiellement pour deux raisons:

• la structure linéaire est cassée puisque le nuage de points contient deux parties différentes

• les erreurs ne suivent pas une distribution continue, puisque l'on peut observer $D = 0$ avec une probabilité plus grande que 0.

La forme particulière du nuage provient du fait que les observations ne portent pas sur la consommation totale du bien durable, mais uniquement sur la modification de celui-ci.

Modélisation du problème

- S_{t-1} le stock du bien du ménage en $t-1$
- S_t^* le stock désiré pour l'instant t
- de_t est la dépréciation de S_{t-1}

$$\Downarrow$$

Le ménage devra pour satisfaire sa consommation désirée modifier son stock de:

$$D_t^* = S_t^* - (S_{t-1} - de_t).$$

Les achats du bien D_t durant la période sont:

$$D_t = \begin{cases} D_t^* & \text{si } D_t^* > 0 \\ 0 & \text{si } D_t^* \leq 0 \end{cases}$$

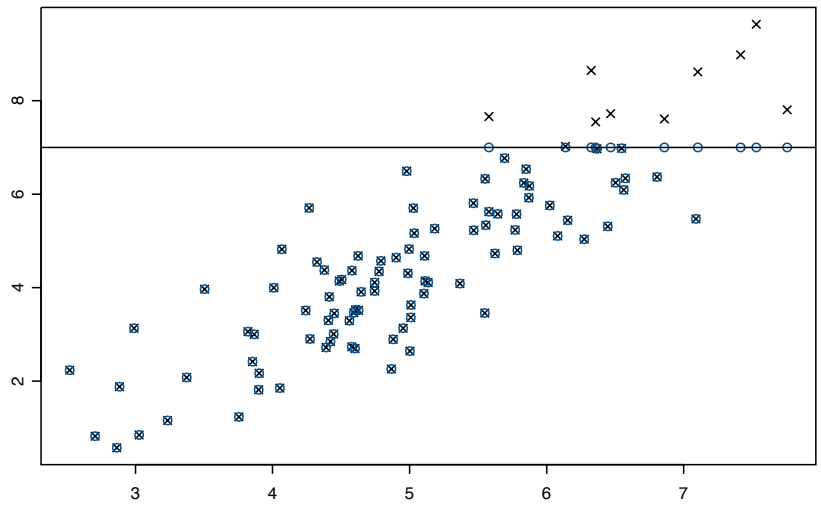
On modélise alors la variable latente D_t^* comme une fonction linéaire des variables explicatives:

$$D_t^* = \beta R_t + \alpha + \varepsilon_t \qquad \Longrightarrow$$

$$D_t = \begin{cases} \beta R_t + \alpha + \varepsilon_t & \text{si } \beta R_t + \alpha + \varepsilon_t > 0 \\ 0 & \text{si } \beta R_t + \alpha + \varepsilon_t \leq 0 \end{cases}$$

Exemple 2: Censure à droite

Supposons que le gouvernement accorde des aides aux nouvelles PME crées dans l'année. Ces aides sont fonction du secteur d'activité, du nombre d'employés, etc, mais l'aide ne peut dépasser un certain niveau de L euros.



(y_i, y_i^*) *en fonction du nombre d'employés.*

Notons y_i^* l'aide versée à l'entreprise i sans la clause de seuil et y_i la variable contrainte:

$$y_i = \begin{cases} y_i^* & \text{si } y_i^* < L \\ L & \text{si } y_i^* \geq L \end{cases}$$

6.2 Modèle de Tobit simple (censure à gauche)

La variable dépendante est définie $\forall i$ par:

$$y_i = \begin{cases} y_i^* & \text{si } y_i^* > 0 \\ 0 & \text{si } y_i^* \leq 0 \end{cases}$$

où

$$y_i^* = x_i'\beta + \varepsilon_i$$

avec $\beta' = (\beta_1 \ldots \beta_p)$ est le vecteur des paramètres inconnus et $x_i' = (x_{i1} \ldots x_{ip}) \in IR^p$ le vecteur des variables explicatives.

Les hypothèses sur les erreurs sont:

- ε_i sont indépendantes
- $\varepsilon_i \sim N(0, \sigma^2)$.

Le modèle Tobit comporte un aspect qualitatif (dans la séparation qui est faite des observations selon le signe de y_i^*) et un aspect quantitatif.

La variable y_i^* est dite variable latente, elle n'est pas observable pour l'ensemble de l'échantillon.

Elle est nécessaire pour répondre à des questions comme:

- "Quelle augmentation de consommation du bien durable d'un ménage dont la dépense observée est nulle impliquera une dépense ?"
- "Quel sera l'accroissement de l'aide versée à une entreprise si le plafond est relevé ?"

6.3 Non fondé de la méthode des moindres carrés

6.3.1 Traitement de l'échantillon complet

Une condition essentielle pour appliquer MCO est l'existence de la relation $E[y_i|x_i] = x_i'\beta$, or on peut démontrer que ce n'est pas le cas ici.

Rappel: Soit X une v.a. avec une fonction de densité f_X. La fonction de répartition F satisfait $F(x) = \int_{-\infty}^{x} f_X(t)dt$ et $\frac{dF(x)}{dx} = f_X(x)$.

L'espérance de la variable $g(X)$ est donnée par

$$E[g(X)] = \int_{-\infty}^{\infty} g(t)f_X(t)dt.$$

L'espérance conditionnelle de la variable $g(X)$ étant donné l'évenement B est donnée par

$$E[g(X)|B] = \frac{\int_B g(t)f_X(t)dt}{P(B)}.$$

Notons aussi que $\phi'(x) = -x\phi(x)$

où $\phi(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}$.

Lemme: Dans le modèle Tobit simple, on a

$$P(y_i > 0|x_i) = \Phi(\frac{x_i'\beta}{\sigma})$$

.

Démonstration:

$$\begin{aligned} P(y_i > 0|x_i) &= P(y_i^* > 0|x_i) \\ &= P(x_i'\beta + \varepsilon_i > 0|x_i) \\ &= P(\varepsilon_i > -x_i'\beta) \\ &= P(\frac{\varepsilon_i}{\sigma} > \frac{-x_i'\beta}{\sigma}) \\ &= P(Z > \frac{-x_i'\beta}{\sigma}) \\ &= P(Z < \frac{x_i'\beta}{\sigma}) \\ &= \Phi(\frac{x_i'\beta}{\sigma}) \end{aligned}$$

où Z est une variable aléatoire de loi $N(0,1)$.

Décomposons la variable dépendante en deux parties:

$$\begin{aligned} y_i &= y_i I(y_i > 0) + y_i I(y_i \leq 0) \\ &= y_i I(y_i > 0) + y_i I(y_i = 0) \\ &= y_i^* I(y_i^* > 0) + 0 \\ &= (x_i'\beta + \varepsilon_i) I(x_i'\beta + \varepsilon_i > 0) \\ &= x_i'\beta I(\frac{\varepsilon_i}{\sigma} > \frac{-x_i'\beta}{\sigma}) + \sigma\frac{\varepsilon_i}{\sigma} I(\frac{\varepsilon_i}{\sigma} > \frac{-x_i'\beta}{\sigma}) \end{aligned}$$

où $I(condition)$ est la fonction indicatrice égale à 1 si condition vraie et à 0 sinon.

Montrons maintenant que la condition

$$E[y_i|x_i] = x_i'\beta$$

n'est pas satisfaite.

$$
\begin{aligned}
E[y_i|x_i] &= x_i'\beta E[I(Z > \frac{-x_i'\beta}{\sigma})] + \sigma E[ZI(Z > \frac{-x_i'\beta}{\sigma})] \\
&= x_i'\beta \int I(t > \frac{-x_i'\beta}{\sigma})\phi(t)dt \\
&+ \sigma \int I(t > \frac{-x_i'\beta}{\sigma})t\phi(t)dt \\
&= x_i'\beta \int_{\frac{-x_i'\beta}{\sigma}}^{\infty} \phi(t)dt + \sigma \int_{\frac{-x_i'\beta}{\sigma}}^{\infty} t\phi(t)dt \\
&= x_i'\beta \int_{-\infty}^{\frac{x_i'\beta}{\sigma}} \phi(t)dt - \sigma \int_{\frac{-x_i'\beta}{\sigma}}^{\infty} d\phi(t) \\
&= x_i'\beta\Phi(\frac{x_i'\beta}{\sigma}) - \sigma[\phi(\infty) - \phi(\frac{-x_i'\beta}{\sigma})] \\
&= x_i'\beta\Phi(\frac{x_i'\beta}{\sigma}) + \sigma\phi(\frac{x_i'\beta}{\sigma}).
\end{aligned}
$$

Conclusion: MCO n'est pas applicable (l'espérance de y_i n'est plus une combinaison linéaire des β).

6.3.2 Traitement de l'échantillon tronqué ($y_i > 0$)

L'échantillon n'est plus représentatif de la population et on parle de modèle tronqué.

Calculons l'espérane conditionnelle:

$$\begin{aligned} E[y_i|x_i \text{ et } y_i > 0] &= \frac{\int_{t>0} t f_{y_i}(t)dt}{P(y_i > 0|x_i)} \\ &= \frac{\int_0^\infty t f_{y_i}(t)dt}{P(y_i > 0|x_i)} \\ &= \frac{\int_{-\infty}^\infty t f_{y_i}(t)dt}{P(y_i > 0|x_i)} \\ &= \frac{E[y_i|x_i]}{P(y_i > 0|x_i)} \\ &= x_i'\beta + \sigma \frac{\phi(\frac{x_i'\beta}{\sigma})}{\Phi(\frac{x_i'\beta}{\sigma})}. \end{aligned}$$

Conclusion: Si on utilise le modèle de régression $y_i = x_i'\beta + \varepsilon_i$, on aura des estimations biaisées (c'est comme si on omettait une variable). Ce biais s'appelle le biais de sélection.

6.4 Méthode d'estimation de Heckman en deux étapes

Cette méthode va exploiter successivement le caractère qualitatif et quantitatif du modèle.

- Première étape

  Modèle qualitatif associé au modèle Tobit:

$$z_i = \begin{cases} 1 \text{ si } y_i > 0 \\ 0 \text{ sinon} \end{cases}$$

  Il s'agit d'un modèle dichotomique (Probit car loi normale) où

$$P(z_i = 1|x_i) = P(y_i > 0|x_i) = \Phi(\frac{x_i'\beta}{\sigma}) = \Phi(x_i'\gamma),$$

  où $\gamma = \frac{\beta}{\sigma}$. Le vecteur γ des paramètres inconnus peut être estimé en utilisant la méthode du maximum de vraisemblance (voir modèle dichotomique).

- Deuxième étape

  La partie quantitative du modèle Tobit correspond aux observations y_i non censurées:

$$y_i = x_i'\beta + \sigma\frac{\phi(x_i'\gamma)}{\Phi(x_i'\gamma)} + \epsilon_i$$

  En pratique, on doit remplacer γ par son estimation trouvée à la première étape:

$$y_i = x_i'\beta + \sigma\frac{\phi(x_i'\hat{\gamma})}{\Phi(x_i'\hat{\gamma})} + \eta_i$$

  Cette équation est en réalité un modèle linéaire en β et σ. La méthode MCO donne les estimateurs $\hat{\beta}$ et $\hat{\sigma}$ qui seront:

  – asymptotiquement non biaisés

  – relativement robustes

  – asymptotiquement non efficaces.

Une autre méthode d'estimation est la méthode du maximum de vraisemblance (plus complexe).

6.5 Extensions: Modèle de type 2

Une classification en 5 classes a été réalisée en se basant sur les fonctions de vraisemblances dans l'article de

Amemiya, T. (1984), Tobit Models: A Survey, *Journal of Econometrics*, **24**, 3–61.

Focus sur le type 2

$$y_{1i} = \begin{cases} y_{1i}^* & \text{si } y_{2i}^* > 0 \\ 0 & \text{si } y_{2i}^* \leq 0 \end{cases}$$

avec comme équation de sélection (p.e. participation au marché du travail):

$$y_{2i}^* = z_i'\gamma + \varepsilon_{2i}$$

et comme équation d'intérêt (p.e. salaire)

$$y_{1i}^* = x_i'\beta + \varepsilon_{1i}$$

6.5.1 Méthode d'estimation en 2 étapes de Heckman

• **Théorème:** Si ε_1 et ε_2 ont une distribution normale bivariée de moyenne μ_1 et μ_2, d'écarts-types σ_1 et σ_2, et de corrélation ρ, alors:

$$E[\varepsilon_1|\varepsilon_2 > a] = \mu_1 + \rho\sigma_1 \frac{\phi(\frac{a-\mu_2}{\sigma_2})}{1 - \Phi(\frac{a-\mu_2}{\sigma_2})}$$

$$E[\varepsilon_1|\varepsilon_2 < a] = \mu_1 - \rho\sigma_1 \frac{\phi(\frac{a-\mu_2}{\sigma_2})}{\Phi(\frac{a-\mu_2}{\sigma_2})}$$

Remarque: si $\rho = 0$, alors on retrouve que

$$E[\varepsilon_1|\varepsilon_2 > a] = \mu_1$$

• **Corollaire:** Si ε_1 et ε_2 ont une distribution normale bivariée de moyennes nulles, d'écarts-types σ_1 et $\sigma_2 = 1$, et de corrélation ρ, alors:

$$E[\varepsilon_1|\varepsilon_2 > a] = \rho\sigma_1 \frac{\phi(a)}{1 - \Phi(a)}$$

En se basant sur l'échantillon tronqué:

$$
\begin{aligned}
E[y_{1i}|x_i, z_i \quad \text{et} \quad & y_{1i} \text{ est observé}] \\
&= E[y_{1i}|x_i, z_i \text{ et } y_{2i}^* > 0] \\
&= E[y_{1i}|x_i, z_i \text{ et } \varepsilon_{2i} > -z_i'\gamma] \\
&= x_i'\beta + E[\varepsilon_{1i}|\frac{\varepsilon_{2i}}{\sigma_2} > \frac{-z_i'\gamma}{\sigma_2}] \\
&= x_i'\beta + \rho\sigma_1 \frac{\phi(\frac{-z_i'\gamma}{\sigma_2})}{1 - \Phi(\frac{-z_i'\gamma}{\sigma_2})} \\
&= x_i'\beta + \rho\sigma_1 \frac{\phi(\frac{z_i'\gamma}{\sigma_2})}{\Phi(\frac{z_i'\gamma}{\sigma_2})} \\
&= x_i'\beta + \beta_\lambda \lambda_i
\end{aligned}
$$

Estimation en deux étapes d'Heckman:

• Estimer l'équation probit de sélection par maximum de vraisemblance $\Rightarrow \hat{\gamma} \Rightarrow \hat{\lambda}_i$

• Estimer β et $\beta_\lambda = \rho\sigma_1$ par MCO

6.5.2 L'effet d'un traitement

Problème:

Effet des études supérieures sur le revenu?

Modèle:

$$R_i = x_i'\beta + \delta S_i + \varepsilon_{1i}$$

où S_i est une variable dichotomique égale à 1 si l'individu i a fait des études supérieures.

Question: Est-ce que δ mesure la valeur des études supérieures ?

Réponse: Non s'il y a un effet d'auto-sélection. Dans ce cas, l'estimation de δ par MCO va surestimer l'effet du traitement.

- Participation au traitement (au cursus):

$$S_i = 1 \text{ si } S_i^* = z_i'\gamma + \varepsilon_{2i} > 0;$$
$$S_i = 0 \text{ sinon}$$

- Equation de revenus:

$$E[R_i|S_i = 1, x_i, z_i] = x_i'\beta + \delta + E[\varepsilon_{1i}|\varepsilon_{2i} > -z_i'\gamma]$$

Pour les non-participants, la contre-partie est donnée par:

$$E[R_i|S_i = 1, x_i, z_i] = x_i'\beta + E[\varepsilon_{1i}|\varepsilon_{2i} < -z_i'\gamma]$$

La différence entre les revenus espérés des participants et des non-participants est donc:

$$E[R_i|S_i = 1, x_i, z_i] - E[R_i|S_i = 0, x_i, z_i] =$$
$$\delta + E[\varepsilon_{1i}|\varepsilon_{2i} > -z_i'\gamma] - E[\varepsilon_{1i}|\varepsilon_{2i} < -z_i'\gamma]$$

6.6 Références

- Amemiya, T. (1984), Tobit Models: A Survey, *Journal of Econometrics*, **24**, 3–61.
- Breen, R. (1996), *Regression models censored, sample selected, or truncated data*, SAGE Publications.
- Greene, W. H. (1997), *Econometric Analysis*, Prentice Hall, London.
- Maddala, G. S. (1986) *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press, London.

Chapitre 7

PANEL DATA

Données mésurées dans deux dimensions

- cross-section (pays, firmes, individus, ...)
- plusieurs périodes d'observations dans le temps

Micro-panels (N grand, T petit)

N individus interrogés sur T années :

National Longitudinal Surveys (NLS) : plus de 3000 articles basés sur ces enquêtes

Macro-panels (N petit, T grand)

Variables macro pour N pays et T années :

Indicateurs de développement (WDI) de la World Bank

7.1 Modèle

Modèle théorique:

$$y_{it} = x'_{it}\beta + \varepsilon_{it} \quad i = 1, \ldots, n; t = 1, \ldots, T$$

où t représente la dimension temporelle, et i la cross-section.

Exemples:

- Le revenu de 1000 ménages sur 10 ans
- La croissance du PNB des pays européens sur les 40 dernières années