

E C O N O M E T R I E

Gérard GRELLET

L'économétrie peut être définie comme l'application des méthodes statistiques à l'étude des phénomènes économiques. Plus précisément la démarche économétrique comporte trois étapes :

- 1) Construire un modèle testable qui soit justifié par la théorie économique et qui puisse être vérifié statistiquement ;
- 2) Estimer les paramètres du modèle ;
- 3) Vérifier que les écarts entre les observations et les résultats théoriques du modèle ne sont pas systématiques .

Historique

Premiers développements

Les tentatives de modélisation à partir de données empiriques ont une longue histoire que l'on peut faire remonter aux " Political Arithmeticians " anglais du XVII^{ème} siècle et auxquels sont attachés les noms de William Petty , Gregory King et Charles Devenant . Gregory King chercha par exemple à établir une loi entre d'une part les déficits des récoltes de blé et d'autre part les variations du prix du blé . A partir du XVIII^{ème} et surtout du XIX^{ème} siècle les économistes tentèrent d' établir des lois économiques à l'instar des lois de la physique newtonnienne . Ce projet fut mené en termes scientifiques par Moore puis par Schultz , Lenoir , Tinbergen et Frisch entre 1914 et 1938 . Les deux grands axes de recherche furent alors l'estimation d'une loi de demande (ce qui conduisit au problème de l'identification) et celle des cycles économiques . Clément Juglar (1819 - 1905) fut le premier à utiliser les séries temporelles pour analyser les cycles et fut suivi par Kuznets et Kondratieff . Toutefois les théoriciens du cycle se limitèrent à l'étude de la périodicité du cycle et ne s'attachèrent guère à celle de la quantification des relations causales sous jacentes. Leur apport à l'économétrie est donc resté marginal.

La naissance de l'économétrie moderne

L'économétrie moderne est née à la fin des années 30 et pendant les années 40. Elle est la résultante de trois phénomènes : le développement de la théorie de l'inférence statistique à la fin du XIX^{ème} siècle ; la théorie macroéconomique et la comptabilité nationale qui offrent des agrégats objectivement mesurables (contrairement à la microéconomie fondée sur l'utilité subjective) ; enfin, et surtout , la forte demande de travaux économétriques, soit de la part d'organismes publics de prévision et de planification , soit de la part d'entreprises qui ont de plus en plus besoin de modéliser la demande et leur environnement économique général. A partir des années 60 l'introduction de l'informatique et des logiciels standardisés va rendre presque routinière l'utilisation de l'économétrie .

En simplifiant de façon sans doute abusive l'on peut distinguer deux grandes périodes de la recherche économétrique moderne . Jusqu'à la fin des années 70 l'économétrie va étudier la spécification et la solvabilité de modèles macroéconomiques à équations simultanées . Puis à la suite de ce que l'on a appelé la révolution des anticipations rationnelles et de la critique de Lucas, la recherche se tournera davantage vers la microéconomie et l'analyse des séries temporelles .

Les modèles économétriques d'équations simultanées

La plus grande partie de la recherche économétrique américaine (effectuée pour une large part au sein de la Cowles Commission) entre 1944 et 1960 porta sur les conditions d'estimation des modèles macroéconomiques d'équations simultanées comportant un élément aléatoire . En 1939 Tinbergen construisait un modèle des cycles économiques comportant 31 équations de comportement de 17 identités . Chacune des équations était estimée au moyen de la méthode des moindres carrés , ce qui , nous le verrons ne pouvait conduire qu'à des estimations inconsistantes . En 1944 Haavelmo posait les conditions générales de solvabilité . Entre 1945 et 1950 Klein présentait ses premiers modèles dont la solution était obtenue par la méthode du maximum de vraisemblance . En 1949 Koopmans déterminait les conditions de solvabilité dans le cas d'un modèle linéaire . En 1954 Theil introduisait la méthode des doubles moindres carrés permettant des calculs effectifs . Toutefois la généralisation des modèles économétriques à équations simultanées utilisée pour des modèles prévisionnels se heurta pendant longtemps au manque de moyens informatiques . Le premier modèle utilisé à des fins prévisionnelles fut celui de Klein - Goldberger en 1955. D'autres modèles suivirent à la fin des années 50 , en particulier celui de la Brookings Institution. Avec l'avancée des techniques informatiques les années 60 et le début des années 70 virent une éclosion de modèles macroéconomiques jouant un rôle important dans la prévision . Le modèle dit de Brookings comprenait ainsi 400 équations . Après 1970 furent commercialisés des modèles standards comme celui dit de Wharton . La stabilité relative de l'environnement économique jusqu'en 1974 leur assura un certain succès .

L'analyse de la régression

L'importance des moyens consacrés à la résolution des problèmes d'identification laissa quelque peu dans l'ombre la recherche sur la corrélation . Le principal obstacle théorique était le traitement de l'autocorrélation des résidus aléatoires . En 1950 Durbin et Watson élaboraient leur célèbre test . Les années 50 virent d'autre part l'apparition de modèles à retards échelonnés avec les travaux de Koyck , d'Almon , de Cagan et de Friedman .

La révolution des anticipations rationnelles et la remise en cause des modèles macroéconométriques

Les années 70 furent celles de la remise en cause radicale des modèles macroéconométriques élaborés pendant les années 60 . Une des raisons vient de ce que l'abandon du système de Bretton Woods puis le quadruplement du prix du pétrole conduisirent à des bouleversements qui ne pouvaient être anticipés par les modèles économétriques . Au niveau théorique il apparut rapidement que les modèles macroéconométriques ne possédaient pas de fondations microéconomiques suffisamment solides . En particulier Lucas montra dès 1972 que si les agents forment leurs anticipations

sur une base endogène à partir de leur expérience il n'est plus possible de considérer que les coefficients structurels des modèles macroéconométriques restent inchangés . Ainsi toute mesure de politique économique doit conduire à un changement dans le comportement des agents tant au niveau de la consommation que de l'investissement . Ceci remet bien évidemment en cause les modèles macroéconométriques traditionnels qui ne distinguaient pas les paramètres expliqués par des causes structurelles de ceux expliqués par la réponse aux mesures de politiques économiques . Une estimation de ces deux types de paramètres a été effectuée par Lucas et Sargent qui les obtinrent directement comme solutions de modèles d'optimisation dynamique . Sur cette base la recherche économétrique des années 80 porta sur les problèmes d'agrégation des préférences des agents , d'inégalité dans la répartition de l'information et sur le processus d'apprentissage .

Vers une économétrie sans théorie ?

La critique de Lucas a ouvert la voie à des critiques plus radicales et a conduit certains économètres comme Sims à dénier à la théorie toute pertinence dans l'estimation des modèles . L'approche même en termes d'anticipations rationnelles est alors rejetée dans la mesure où elle nécessite une connaissance à priori des délais. Plus fondamentalement les modèles macroéconométriques reposaient sur une distinction entre variables "endogènes" et "exogènes" . Cette distinction qui suppose une connaissance théorique à priori est rejetée.

Cette critique a conduit à retenir des modèles autorégressifs où n'existe pas à priori une classification entre variables endogènes et exogènes . La question de l'utilité de tels modèles reste toutefois controversée dans la mesure où ils ne fournissent pas une explication structurelle de l'activité économique .

CHAPITRE I

LE MODELE CLASSIQUE DE REGRESSION LINEAIRE

I - 1) La droite de régression

Dans ce chapitre nous chercherons à déterminer une relation linéaire entre une ou plusieurs variables explicatives (ou exogènes) et une variable déterminée (ou endogène) à partir d'un ensemble de n observations temporelles .

Nos observations sont partielles et constituent un échantillon (que nous supposerons ici représentatif) de l'ensemble des observations.

Il s'agit donc d'estimer des coefficients à partir d'un échantillon aléatoire.

Par exemple si nous avons une seule variable expliquée $Y(t)$ et une seule variable explicative, $X(t)$, nous chercherons à déterminer si la relation entre ces deux variables peut être estimée par une relation linéaire de type:

$$(1-0) \quad Y(t) = c(1) + c(2)* X(t) , t= 1, \dots , n$$

où le signe * est signe de multiplication.

les coefficients $c(1)$ et $c(2)$ devant être estimés à partir de n observations (voir graphique 1-1). Cette droite est dite " droite de régression". Remarquons que sur le graphique 1-1 les 10 observations peuvent être considérées comme un échantillon représentatif du nombre infini des observations possibles dans la mesure où nos observations sont effectuées à intervalles réguliers dans le temps.

Les coefficients $c(1)$ et $c(2)$ sont déterminés de façon à ce que la relation (1-1) minimise la somme des écarts entre les valeurs observées $Y(t)$ et les valeurs $Y(t)$ données par la droite de régression. La droite de régression passe donc par le "milieu " des points observés (voir annexe 1).

Les écarts entre les observations et la relation (1-1) peuvent être expliqués par :

- des erreurs d'observation
- des variables explicatives qui ne sont pas incluses dans la relation (1-1) . Par exemple si $Y(t)$ représente la consommation et $X(t)$ le revenu, l'on sait que d'autres variables peuvent expliquer la consommation comme le temps qu'il fait , les modes , les possibilités d'emprunt. Il faut supposer que ces autres variables explicatives ont chacune une influence très faible car sinon il faudrait les inclure explicitement dans la relation (1-1). Nous verrons dans le chapitre suivant comment déterminer si des variables explicatives importantes ont été oubliées dans une fonction .
- des erreurs qui viennent de ce que la vraie relation n'est pas en fait linéaire .

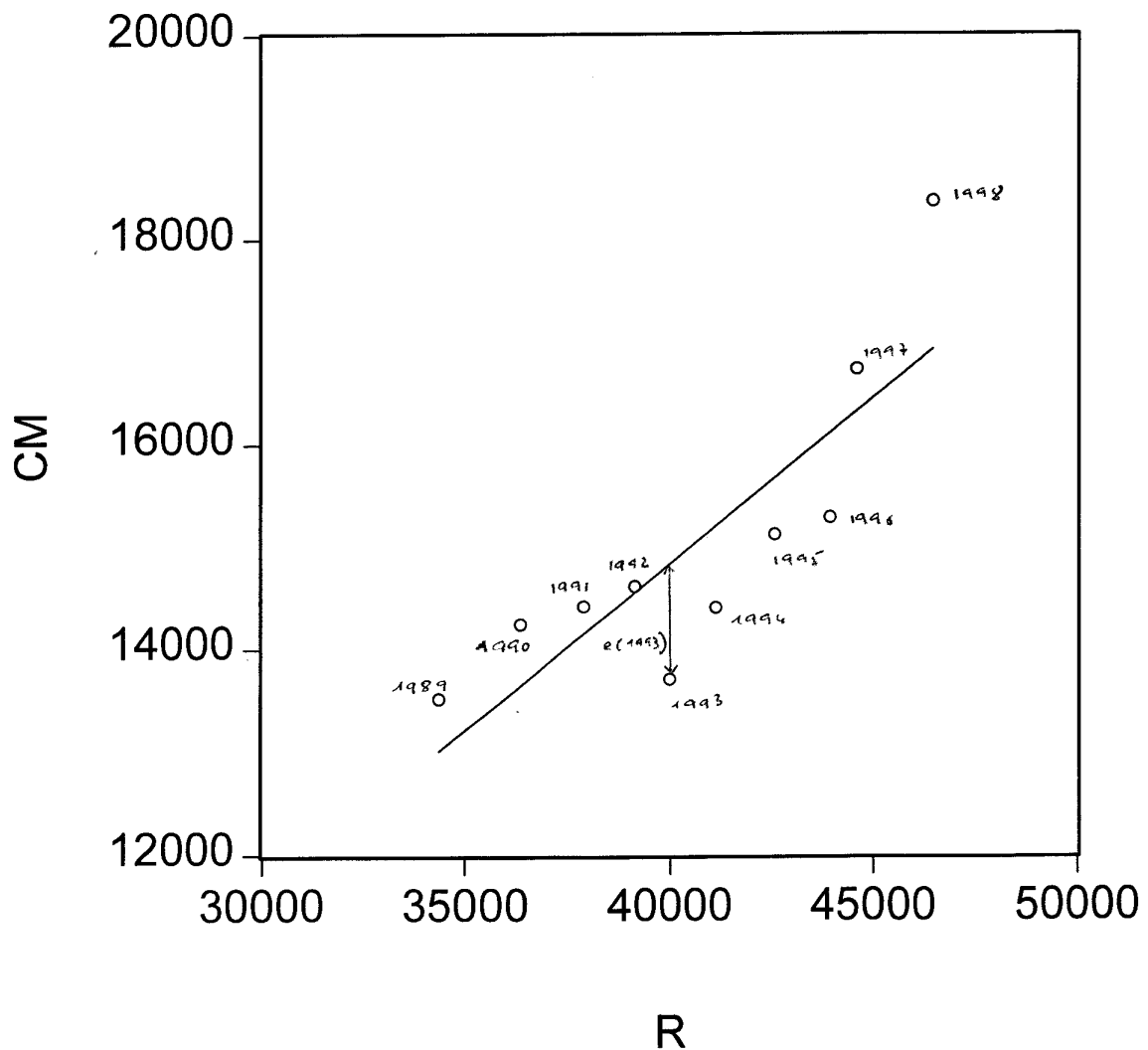
Dans ces deux derniers cas l'on dit que la relation a été mal spécifiée.

Ces trois sources d'écart sont considérées ici comme aléatoires (cette hypothèse sera examinée au chapitre suivant). Nous désignerons leur somme par $e(t)$ au temps t .

Nos observations seront donc considérées comme la somme de $c(1) + c(2)*X(t)$ et de l'élément aléatoire $e(t)$.

L'analyse de la régression peut être facilement généralisée à n variables.

Graphique 1 – 1



Relation entre la consommation des ménages français et le revenu national

I - 2) Le problème des observations aberrantes

L'on observe quelquefois des observations "aberrantes" dues soit à une erreur grossière (par exemple une erreur de virgule) soit à l'incorporation dans les données de cas exceptionnels .

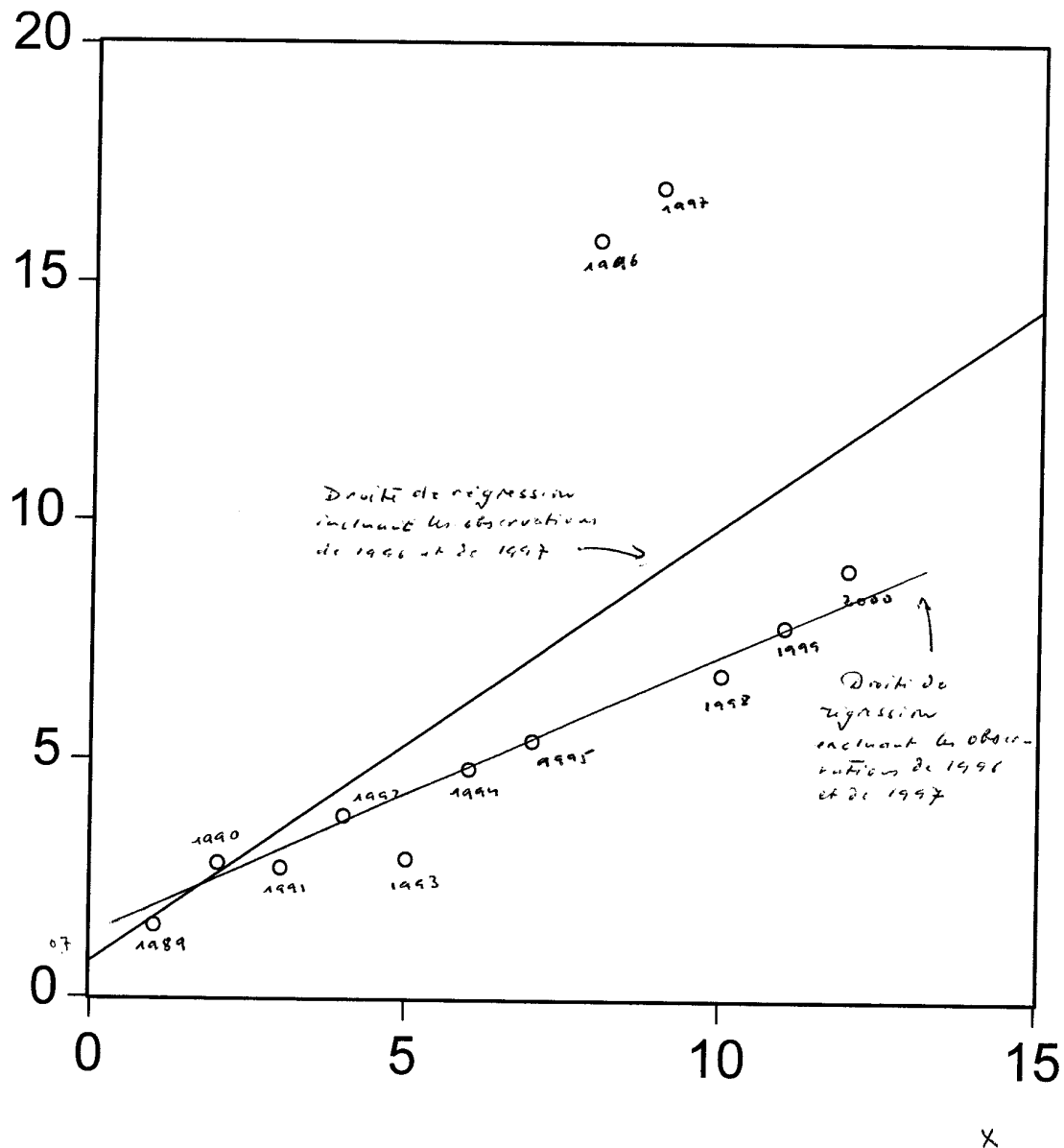
Considérons par exemple le cas du graphique 1 – 2 . Sur ce graphique l'on constate deux observations « aberrantes » , celles des années 1996 et 1997. Si nous les incluons dans nos observations nous obtenons une droite de régression :

$$Y = 0.7 + 0.62 X$$

Par contre si nous ne les incluons pas notre droite de régression devient :

$$Y = 1.13 + 0.46 X$$

L'inclusion ou la non inclusion de certaines variables peut conduire à des résultats significativement différents. Dans ce cas il convient de s'assurer que les observations "aberrantes " ne proviennent pas d'une erreur de mesure . Si tel n'est pas le cas l'on peut chercher à comprendre les raisons de l'apparition de ces cas exceptionnels et faire deux régressions, l'une sans , l'autre avec ces observations.



GRAPHIQUE 1-2

I - 3) La prise en compte des variables qualitatives explicatives

Par données qualitatives il faut entendre les données qui n'ont pas d'échelle de mesure mais que l'on peut classer en catégories.

Certaines de ces données sont sans ambiguïté comme le lieu de naissance ou le sexe mais d'autres doivent être considérées avec précaution comme les critères sociaux souvent arbitraires. Nous supposons ici que toutes les variables qualitatives peuvent être classées en deux catégories comme par exemple le fait de voter oui ou non à une élection, d'être français ou non, d'appartenir ou non aux moins de vingt ans, etc..

L'on peut alors à tester si l'appartenance à un groupe "qualitatif" entraîne une différence objective sur la variable déterminée.

Une variable qualitative qui peut prendre deux valeurs est appelée variable binaire (ou quelquefois variable muette ou encore variable auxiliaire). Par exemple l'on peut chercher à savoir si l'appartenance à l'un ou l'autre sexe entraîne une différence significative de salaire . L'on crée alors une variable binaire qui prend la valeur 1 pour les salariés masculins et 0 pour les salariés féminins .

L'on aura donc le modèle suivant :

$$Y(i) = c(1) + c(2)*X(i) + e(i)$$

où $Y(i)$ est le revenu du i ème salarié et où $X(i)$ est une variable binaire .

Dans cet exemple simple $c(1)$ mesure l'espérance mathématique du revenu féminin et $c(2)$ l'espérance mathématique de la différence de salaire entre hommes et femmes . En effet :

$$E (Y (i)) = c(1) \text{ si } X(i) = 0 \text{ et } c(1) + c(2) \text{ si } X(i) = 1$$

Pour savoir s'il existe une différence significative entre salariés masculins et féminins il suffit donc de tester si $c(2)$ est significativement différent de 0. En fait les estimations de $c(1)$ et de $c(2)$ obtenues par régression sont respectivement les revenus moyens des salariés féminins et la différence entre les revenus moyens des salariés masculins et féminins .

Une autre application est d'étudier s'il existe une différence significative des paramètres d'une fonction économique entre différentes périodes .Par exemple si nous considérons le graphique (1-3) l'on peut :

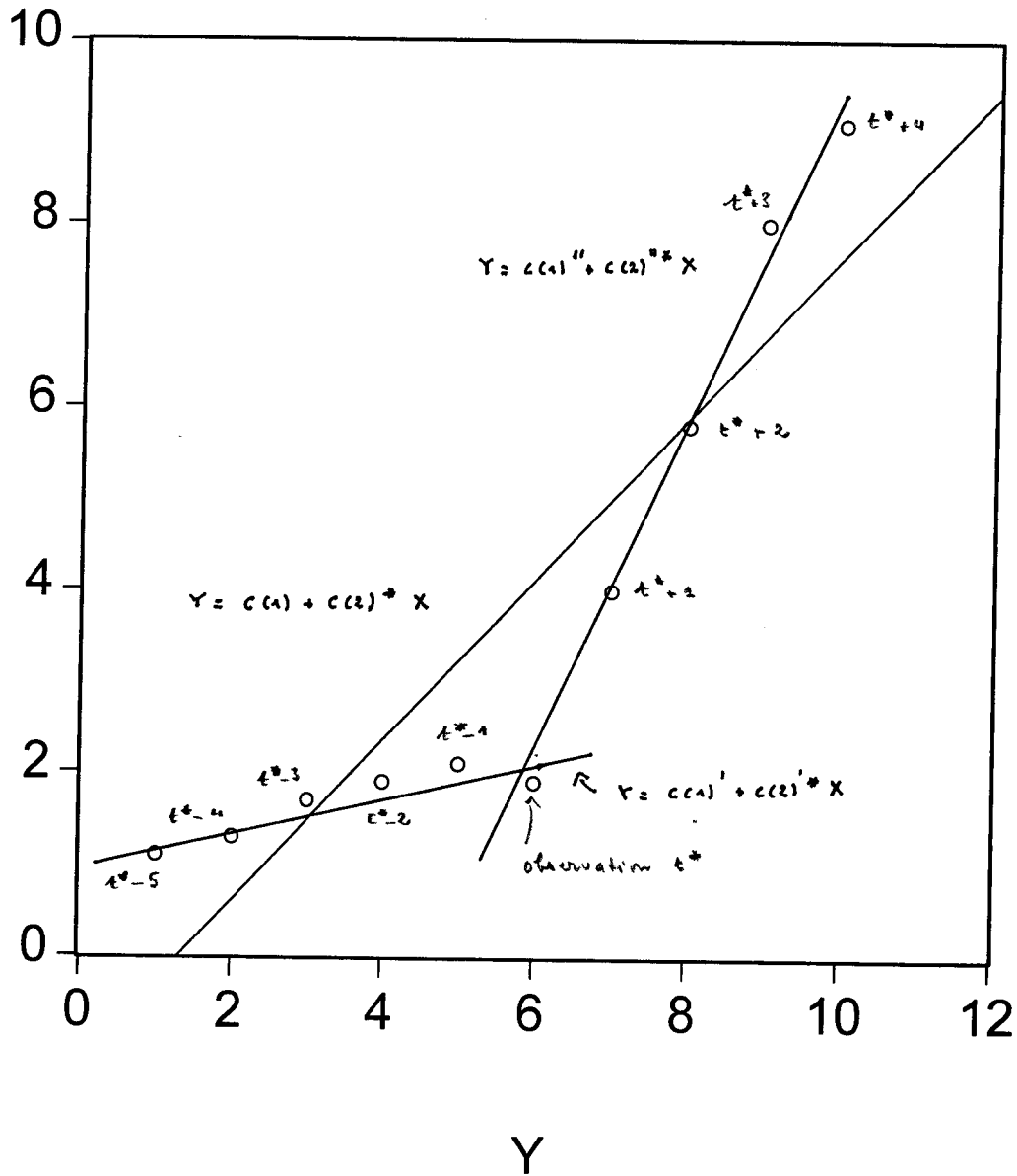
- soit effectuer une régression générale sur les dix observations et l'on obtiendra une estimation :

$$Y = c(1) + c(2)* X \quad , t = 1, \dots, 10$$

- soit considérer que les paramètres des cinq premières observations sont différents des paramètres des cinq dernières observations .Dans ce cas l'on procédera à deux estimations :

-

$$Y = c(1)' + c(2)'* X \quad , t = 1, \dots, 5$$



GRAPHIQUE 1-3

$$Y = c(1)'' + c(2)'' * X, t = 6, \dots, 10$$

et l'on testera si $c(1)'$ est significativement différent de $c(1)''$ et si $c(2)'$ est significativement différent de $c(2)''$

Si tel était le cas il ne serait pas légitime d'effectuer une régression sur l'ensemble des variables.

I - 4) La régression log log

La régression log log présente deux avantages:

- elle permet de rendre linéaire une relation non linéaire;
- elle permet de faire apparaître les coefficients de la droite de régression comme des coefficients d'élasticité puisque : $d \log y / d \log x = dy/dx / x/y$

I - 5) L'intervalle de confiance des coefficients

Dans la relation 1 - 1 nous avons estimé un coefficient $c(2)$ sur la base de n observations. La question peut être posée si ce coefficient est significativement différent de 0. Si tel n'était pas le cas il n'existerait pas de relation certaine entre les $y(t)$ et les $x(t)$.

Pour dire si $c(2)$ est significativement différent de 0 il faut que la valeur 0 ne soit pas comprise à l'intérieur de l'intervalle de confiance construit autour de $c(2)$ avec une probabilité d'erreur donnée.

Il n'est toutefois pas possible de calculer directement l'intervalle de confiance autour de $c(2)$ puisque nous ne connaissons pas directement l'écart type de $c(2)$. L'on peut toutefois estimer celui-ci mais il faut alors utiliser une loi de Student à $n-2$ degrés de liberté pour les valeurs tabulaires t associées à des probabilités données. L'on peut ainsi montrer que l'intervalle de confiance autour de $c(2)$ est :

$$c(2) \pm t_{\alpha} \left[\frac{\sum (y - y_i)/(n-2)}{\sum (x_i - \bar{x})^2} \right]^{\frac{1}{2}} \quad \text{ou} \quad c(2) \pm t_{\alpha} s$$

où t est la valeur donnée par la table de Student avec $n-2$ degrés de liberté et une probabilité d'erreur donnée.

Si la valeur 0 n'est pas comprise dans l'intervalle de confiance construit avec une probabilité donnée, la relation entre les $x(t)$ et les $y(t)$ peut être considérée comme significative, ce qui ne signifie toutefois pas que les $x(t)$ "déterminent" les $y(t)$ (voir p.)

L'on peut de façon équivalente calculer le t à partir duquel la valeur 0 serait incluse dans l'intervalle de confiance soit $c(2) + |t^*| s = 0$ ou $c(2) / s = t$

Si $t > t$ alors la relation entre les $x(t)$ et les $y(t)$ peut être considérée comme significative à une probabilité donnée.

En pratique l'on peut considérer que $t = 2$ pour $n > 20$.

I - 6) La régression dans E-view

1) Créer un fichier pour rentrer les données

File / New / Workfile

Quick / Empty group

Ou

*Importer les données d'un tableur Excel (auquel l'on a donné un nom):
Procs / Import / Read Text / Excel / nom du tableur*

Ou

Faire un copier coller dans le workfile avec le même nombre de lignes et de colonnes que dans le tableur

2) Estimer les coefficients de la régression

Quick / estimate equation

Ou Object/ equation

L'on écrit le nom de la variable dépendante suivie de c (la constante) et du nom des variables explicatives.

Les variables suivies de (-1) , (-2) , (-3) etc ... sont les variables retards. Exemple : fbcf(-1)

*Pour écrire les logarithmes des données l'on fait précéder log du nom de la variable .
Exemple : log(fbcf) .*

Rappelons que les coefficients d'une équation logarithmique sont les coefficients d'élasticité

Cliquer sur OK dans la boîte de dialogue .

L'estimation se fait en ajustant l'échantillon de façon à ne pas tenir compte des données manquantes .

E-views donne le choix entre plusieurs techniques d'estimation . Dans une première approche l'on peut se contenter de la méthode des moindres carrés ("LS")

Résultats

L'on se bornera à ne considérer que les t de Student des différents coefficients (significatifs si $t > 2$) et le Durbin Watson (qui doit être proche de 2)

Exemple :

L'on cherche à estimer la relation , que l'on suppose linéaire , entre la consommation des ménages français (CM) et le revenu national français (R) entre 1989 et 1998.

Ces données sont présentées dans le tableau suivant construit à partir de :

*new / workfile (indiquer que le tableau va de 1989 à 1998)
quick / empty group*

obs	CM	R
1989	13520.00	34377.00
1990	14244.00	36329.00
1991	14418.00	37902.00
1992	14615.00	39127.00
1993	13708.00	39986.00
1994	14405.00	41132.00
1995	15122.00	42577.00
1996	15287.00	43939.00
1997	16738.00	44599.00
1998	18373.00	46453.00

Pour estimer la droite de régression l'on clique sur :

quick/estimate equation

*L'on demande alors d'estimer la relation $CM=c(1)+c(2)*R$*

L'on obtient alors :

Dependent Variable: CM				
Method: Least Squares				
Date: 12/22/02 Time: 14:33				
Sample: 1989 1998				
Included observations: 10				
CM=C(1)+C(2)*R				
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	1316.162	2966.452	0.443682	0.6690
C(2)	0.337085	0.072699	4.636725	0.0017
R-squared	0.728806	Mean dependent var	15016.00	
Adjusted R-squared	0.694907	S.D. dependent var	1514.256	
S.E. of regression	836.4024	Akaike info criterion	16.47295	
Sum squared resid	5596552.	Schwarz criterion	16.53347	
Log likelihood	-80.36477	Durbin-Watson stat	0.801124	

La droite de régression est donc :

$$CM = 1316 + 0.33R$$

Remarquons que le Durbin Watson est très faible (0.8) ce qui signifie ici que la spécification linéaire de la fonction n'est pas bonne (voir chapitre suivant).

Annexe 1

La droite de régression

Soit un ensemble de n observations $(X(i), Y(i))$, $i = 1, \dots, n$

L'analyse de la régression consiste à estimer les coefficients $c(0)$ et $c(1)$ d'une fonction linéaire :

$$Y(i) = c(1) + c(2) * X(i) + e(i)$$

qui passe par le "milieu" des observations .

La méthode la plus couramment utilisée est la méthode dite des moindres carrés . Les coefficients $c(1)$ et $c(2)$ sont définis en minimisant la somme des écarts $e(i)$. L'on obtient alors :

$$c'(2) = \frac{E (x(i) - \bar{x}) (y(i) - \bar{y})}{E (x(i) - \bar{x})^2}$$

$$E (x(i) - \bar{x})^2$$

$$c'(1) = \bar{Y} - c'(2) \bar{X}$$

La droite de régression s'écrira donc :

$$Y = c'(1) + c'(2) X$$

CHAPITRE II

LA REMISE EN CAUSE DES HYPOTHESES

DU MODELE LINEAIRE CLASSIQUE

II – 1) Les hypothèses du modèle classique de régression linéaire

Le modèle classique de régression linéaire est fondé sur les quatre hypothèses suivantes :

1) La relation entre la variable endogène et la variable exogène est linéaire ;

2) Les variables endogènes observées possèdent un élément aléatoire e . Cet élément aléatoire peut provenir par exemple de l'existence d'erreurs non systématiques d'observations ou du fait de la non prise en compte de certaines variables, en particulier de variables dont l'influence sur la variable endogène n'est pas systématique .

Dans le modèle classique de régression linéaire l'on suppose de plus que :

i) l'espérance mathématique des éléments aléatoires est nulle :

$$E(e) = 0$$

ii) la variance de l'élément aléatoire est constante (hypothèse d'homoscédasticité)

iii) les éléments aléatoires sont statistiquement indépendants soit

$$E(e(i).e(j)) = 0, \forall i \text{ et } j$$

iv) les éléments aléatoires sont distribués suivant une loi normale . Cette dernière hypothèse est justifiée par le théorème central limite .

Ce caractère aléatoire des erreurs constitue une hypothèse fondamentale du modèle classique . Elle est justifiée si les erreurs n'ont pas un caractère systématique . Ceci suppose que le modèle de régression n'ait pas oublié une variable explicative importante . C'est cette hypothèse d'une loi de distribution statistique autour des vraies valeurs estimées qui va permettre de faire des estimations des paramètres du modèle d'ajustement .

3) Les variables exogènes sont certaines . Elles ne comportent donc pas d'élément aléatoire.

4) Les variables exogènes sont non corrélées entre elles .

Nous examinerons dans ce chapitre la remise en cause de cinq de ces hypothèses du modèle classique de régression linéaire :

1) l'indépendance dans le temps des éléments aléatoires ;

2) la constance dans le temps de la variance des éléments aléatoires ;

3) le caractère non stochastique des variables exogènes ;

4) la non corrélation des différentes variables exogènes;

5) la constance des coefficients du modèle

II - 2) La remise en cause de la première hypothèse : Les résidus sont auto - corrélés

Quand les éléments aléatoires sont corrélés entre eux l'on a :

$e(t) = z e(t-1) + u(t)$ où $u(t)$ est un élément aléatoire d'espérance statistique nulle mais où z est significativement différent de 0 .

Graphiquement le problème de l'auto-corrélation des éléments aléatoires peut être représenté sur le graphique (2-1) .

Si les observations sont représentées par des croix le modèle de régression peut conduire à l'estimation de la fonction $Y = c(1) + c(2)* X$. Or cette relation peut être fausse si les e suivent la loi suivante :

pour $X < a$, $e > 0$

$a < X < b$, $e < 0$

$X > b$, $e > 0$

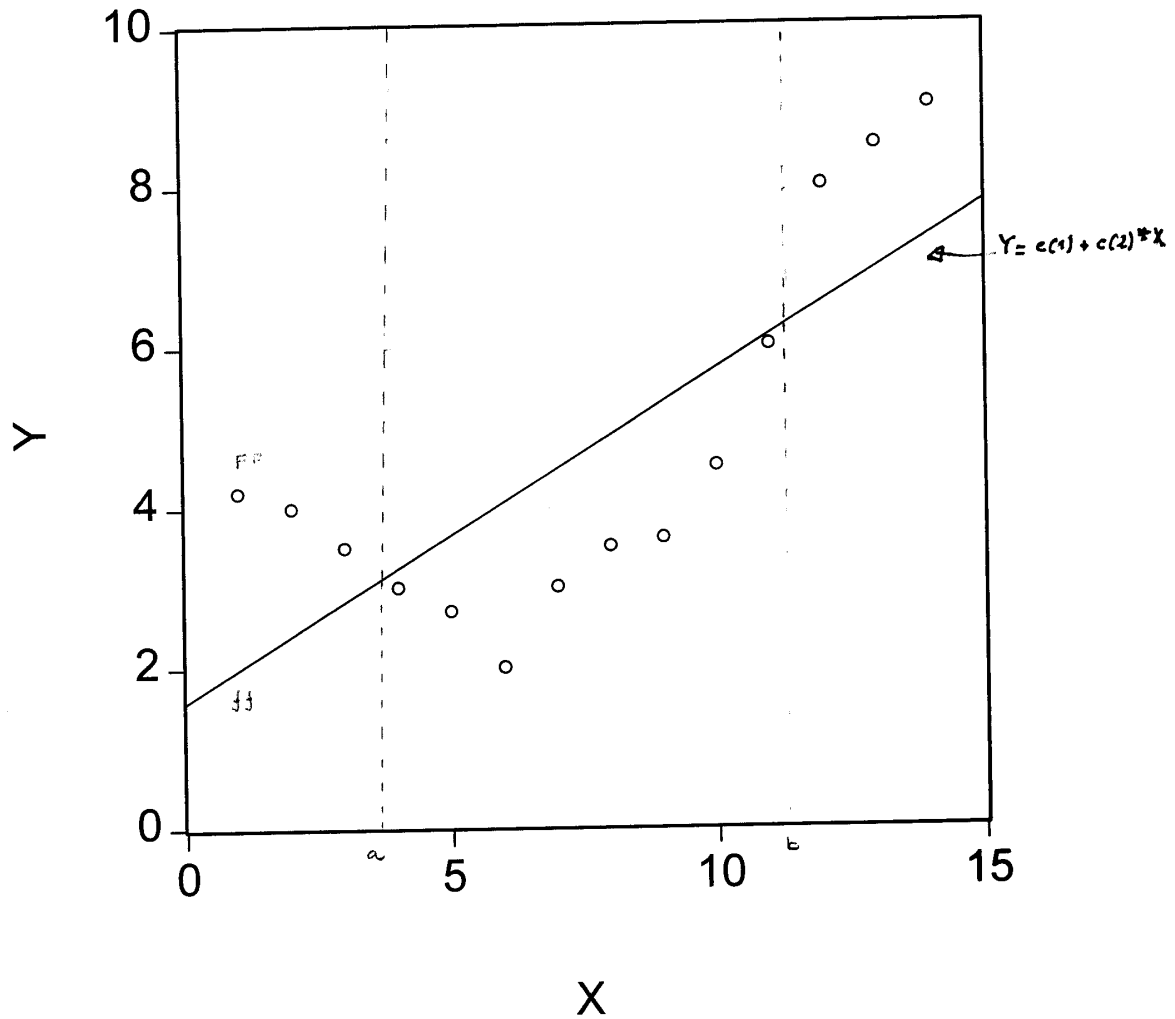
Pour mesurer l'auto - corrélation des éléments aléatoires l'on peut effectuer un test visuel soit utiliser le test de Durbin Watson

Pour une taille d'échantillon donnée et pour un nombre donné de variables explicatives la table des d du test de Durbin Watson donne deux valeurs critiques $d(l)$ et $d(v)$.

- si $d(l) < d < d(v)$ le test n'est pas concluant
- si $d < d(l)$ il existe une auto - corrélation positive entre les éléments aléatoires
- si $d > d(v)$ l'on peut considérer que n'existe pas d'auto - corrélation positive

Ce test est inapplicable au modèle à retards échelonnés.

En pratique l'on peut considérer que n'existe pas d'auto-corrélation positive entre les résidus si $d > 2$. En cas de doute l'on peut chercher à visualiser le comportement des $e(t)$.



GRAPHIQUE 2-1

D'où vient l'autocorrélation entre les éléments résiduels ? Elle peut naître de deux causes :

- la mauvaise spécification de la fonction
- l'omission d'une variable significative .

a) La mauvaise spécification de la fonction .

Supposons comme sur le graphique (2-1) que nous ayons estimé une relation FF non linéaire par une relation ff linéaire . L'on constate alors que les écarts sont distribués systématiquement au dessus puis en dessous puis à nouveau au dessus de la relation linéaire

b) L'omission d'une variable significative ou l'inclusion d'une variable non significative

Supposons que le "bon" modèle soit :

$$Y' = c(0) + c(1)*X + c(2)*Z + u$$

mais qu'il soit estimé par une fonction de type:

$$Y = c(0) + c(1)*X + v$$

Dans ce cas les résidus ne sont plus distribués de façon aléatoire autour de Y mais reflètent le comportement de Z. Les conséquences de l'omission de Z sont alors :

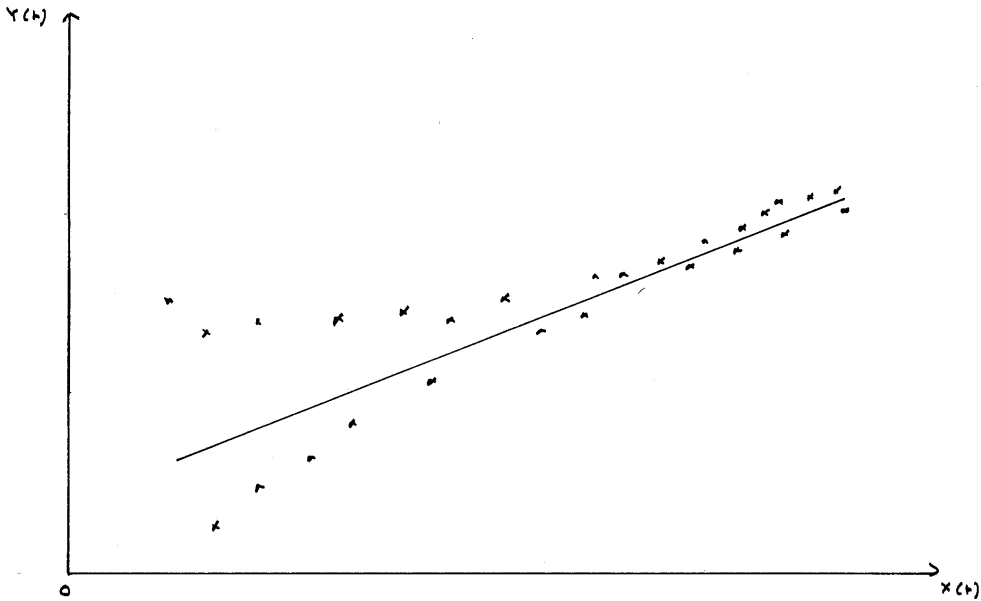
- 1) un biais dans l'estimation des paramètres
- 2) des erreurs dans l'estimation de la variance de X

Comment détecte-t-on l'omission d'une variable significative ?

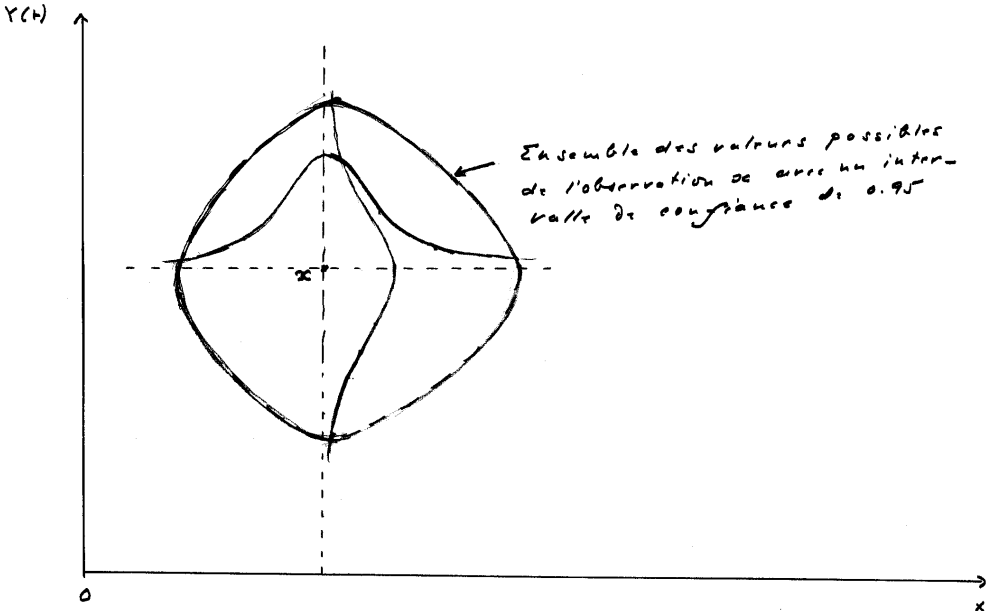
L'auto - corrélation des résidus n'est pas une preuve suffisante puisque nous venons de voir qu'elle peut provenir de la mauvaise spécification de la fonction . C'est pourquoi l'on l'on comparera les taux de régression avant et après l'inclusion d'une nouvelle variable soit $R_a(\text{ancien})$ et $R_n(\text{nouveau})$. L'on calculera alors :

$$F = \frac{(R_n - R_a) / \text{nombre de paramètres de l'ancien modèle (m)}}{(1 - R_n) / \text{nbr d'observations (N) - nbr de paramètres dans le nouveau modèle (k)}}$$

F est donné par la table de Fisher Snédécour . Si la valeur de F est supérieure à la valeur de F tabulaire au niveau de signification de 5% avec pour degrés de liberté m au numérateur et (N - k) au dénominateur l'on peut retenir l'hypothèse que le modèle originel est mal spécifié.



Graphique 2-3



Graphique 2-4

II - 2 La remise en cause de la seconde hypothèse : l'hétéroscédasticité

L'on dira que l'élément aléatoire connaît une hétéroscédasticité quand sa variance présente de larges écarts dans le temps. Il existe alors un biais important dans l'estimation des paramètres puisque toutes les observations n'ont pas le même poids dans l'estimation de Y. Les observations dans les périodes où la variance est élevée ont davantage de poids que les observations dans les périodes où la variance est faible.

L'hétéroscédasticité peut être forte dans le cas de phénomènes d'apprentissage .

Pour déceler l'existence de l'hétéroscédasticité l'on utilisera soit un test visuel soit le test de Goldfield - Quandt .

a) Le test visuel

Sur le graphique (2-2) l'on constate une variance décroissante avec le temps . Ce test doit être utilisé avec précaution dans la mesure où la série est courte .

b) Le test de Goldfield - Quandt

Dans le cas d'une régression de Y par rapport à X on calcule les régressions pour le premier et le dernier quart de la série soit :

$$Y = c(1) + c(1)*X \text{ pour le premier quart}$$

$$Y = c(2) + c(2)*X \text{ pour le dernier quart}$$

On calcule ensuite la somme des carrés des résidus pour les deux régressions soit :

$$S C R (1) = \sum e_1^2 \text{ et } S C R (2) = \sum e_2^2$$

or $S C R (1) / S C R (2)$ obéit à une loi de Fisher à $n - 2$ degrés de liberté .

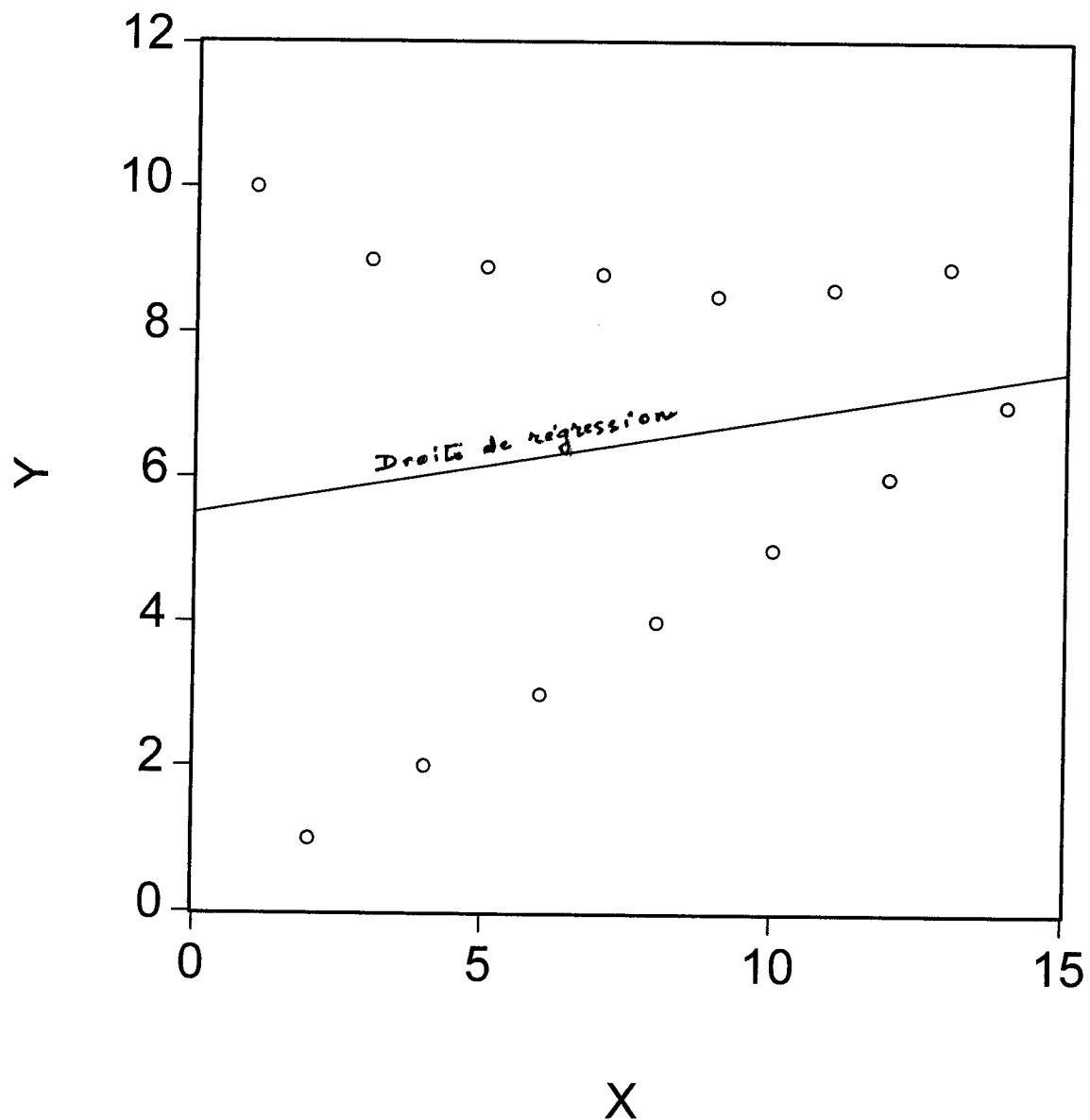
$n =$ nombre total d'observations - nombre d'observations omises - 2 (nombre de paramètres estimés)

Si F dépasse le F tabulaire au niveau de signification de 5% alors l'hypothèse d'hétéroscédasticité peut être retenue.'

Ce test suppose toutefois que nous puissions séparer la série en deux parties : l'une avec une variance faible et l'autre avec une variance élevée .

c) Que faire en cas d'hétéroscédasticité ?

L'on peut diviser les observations de Y par leur écart type de sorte que quand l'écart type est élevé la pondération est faible. Cette procédure est disponible sous E - view .



Graphique 2-2

II - 4 : Le non respect de la troisième hypothèse : la multicollinéarité

La multicollinéarité peut être définie comme l'existence d'une relation linéaire entre les variables indépendantes. Elle suppose donc l'existence d'une relation linéaire telle que :

$$c(1) * X(1) + c(2) * X(2) + \dots + c(n) * X(n) = 0$$

Quand n'existent que deux variables exogènes la multicolinéarité peut être décelée en calculant directement les coefficients de corrélation entre les deux variables. L'on peut ainsi considérer qu'un coefficient de corrélation supérieur à 0.9 ou 0.8 indique une très forte association linéaire et donc une colinéarité. Dans le cas où le modèle comporte plus de deux variables, par exemple si :

$$Y = c(0) + c(1)X(1) + c(2)X(2) + \dots + c(n)X(n)$$

l'on testera n équations :

$$X(1) = -c(2)/c(1)X(2) - \dots - c(n)/c(1)X(n)$$

$$X(2) = -c(1)/c(2)X(1) - \dots - c(n)/c(2)X(n)$$

.....

$$X(n) = -c(1)/c(n)X(1) - \dots - c(n-1)/c(n)X(n-1)$$

Si les variables sont fortement corrélées il devient très difficile de déterminer leur influence respective sur la variable dépendante. En d'autres termes pour que celle-ci soit décelable il faut que dans l'échantillon chaque variable indépendante ait un comportement indépendant.

Statistiquement la multicolinéarité conduit à une très large indétermination des paramètres et donc à de très grands intervalles de confiance. Elle se traduit par la non robustesse des coefficients estimés – en ajoutant ou en retirant une variable l'on modifie considérablement les coefficients des autres variables explicatives. Toutefois la multicolinéarité n'entraîne en elle-même aucun biais dans l'estimation des coefficients. Elle a les mêmes conséquences qu'un petit nombre d'observations. Elle n'est donc pas très gênante pour effectuer des prédictions globales.

Pour éviter la multicolinéarité il est nécessaire de ne travailler qu'avec un nombre réduit de variables explicatives que nous savons indépendantes les unes par rapport aux autres ce qui, malheureusement, appauvrit le modèle.

II - 5 Le non respect de la quatrième hypothèse : Les variables indépendantes sont elles-mêmes des variables stochastiques

Les variables explicatives peuvent être elles-mêmes des variables stochastiques soit parce que la mesure de la variable est sujette à erreur soit parce que la variable est elle-même dépendante d'une variable aléatoire.

Par exemple dans le cas de la relation :

$$Y = c(0) + c(1)X + e$$

X peut être stochastique parce que sa mesure repose sur une estimation statistique sujette à des erreurs non systématiques. Deux estimations différentes effectuées simultanément peuvent alors donner des résultats différents. Mais X peut également être stochastique parce

qu'il dépend d'un élément aléatoire :

$$X = c'(0) + c'(1) * Y + e'$$

Dans ces différents cas l'on ne peut savoir dans quelle mesure la différence entre la valeur observée de la variable endogène et la valeur donnée de cette variable par la régression est due à une erreur sur la variable explicative ou au caractère aléatoire de la variable dépendante.

L'on constate sur le graphique (2- 3) que les valeurs possibles de Y sont comprises dans une ellipse et non sur une droite.

Il est facile de comprendre que si une variable explicative est stochastique, les estimations de $c(0)$ et de $c(1)$ vont souffrir d'une large indétermination.

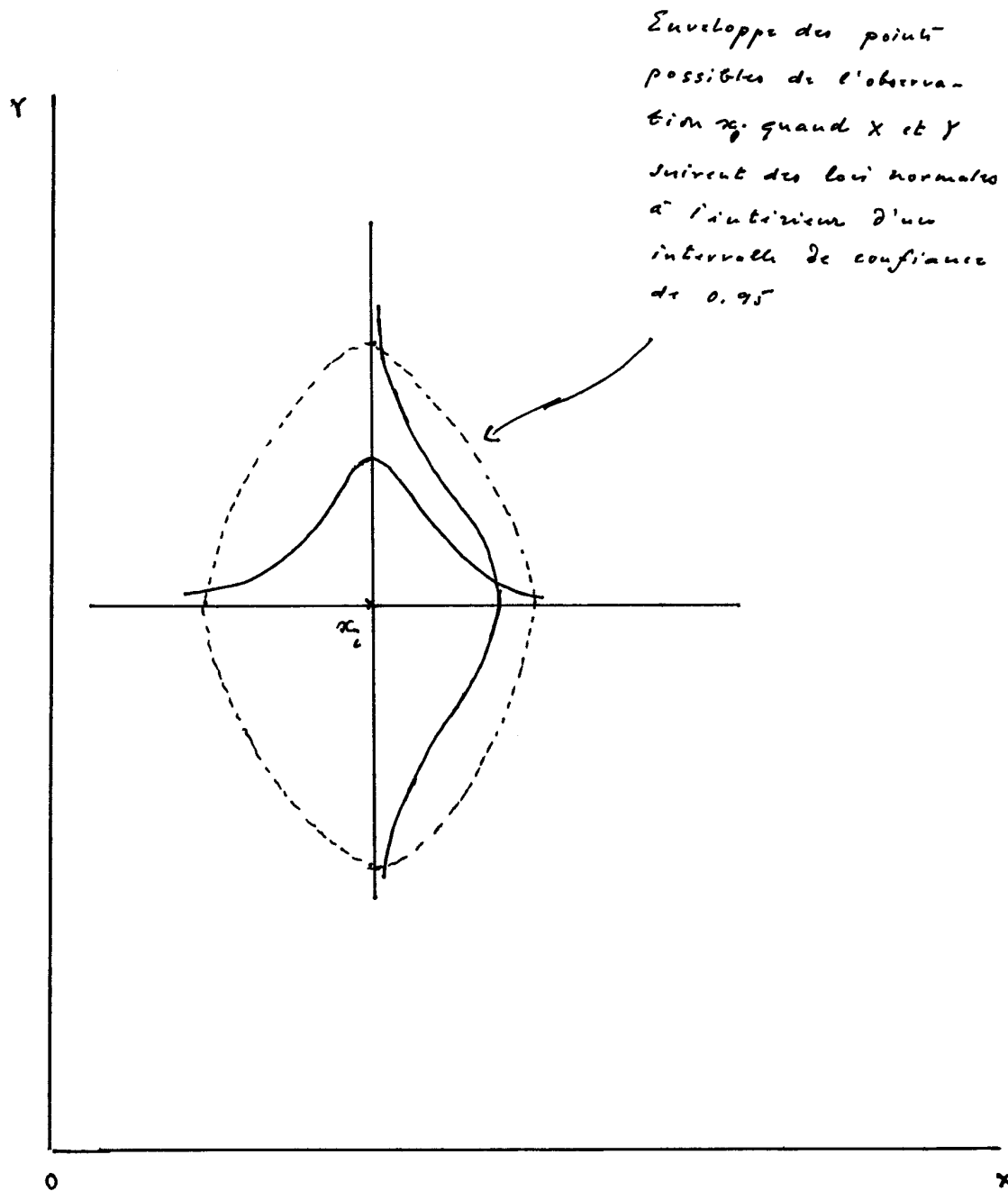
Pour aussi fréquent que soit le problème des variables explicatives stochastiques l'économètre reste mal armé pour le résoudre de sorte qu'il a souvent tendance à passer le problème sous silence . Ce n'est que dans les cas les plus graves qu'il utilise la méthode des variables instrumentales .

La méthode des variables instrumentales est simple dans son principe et difficile dans son application .

Le principe est de remplacer une variable stochastique par une variable non stochastique dite "instrumentale" fortement corrélée avec la variable . Par exemple si la récolte de blé peut être considérée comme une variable stochastique l'on peut chercher à la remplacer par un indicateur pluviométrique si l'on sait par ailleurs que la corrélation est forte entre les fluctuations de la production céréalière et celles de la météo.

Plus précisément l'on dira qu'une variable Z est instrumentale si :

- 1) la corrélation entre Z et e est nulle quand la taille de l'échantillon augmente ;
- 1) la corrélation entre Z et X n'est pas nulle quand la taille de l'échantillon augmente .



Graphique 2-3

Une telle procédure pose toutefois deux problèmes .

En premier lieu il n'est pas toujours possible de trouver une telle variable instrumentale ;

En second lieu l'utilisation des variables instrumentales conduit à des estimations consistantes mais ne garantit pas des estimations non biaisées.

Nous reprendrons le problème des variables instrumentales au chapitre III.

II-6 Le non respect de la cinquième hypothèse : les coefficients du modèle ne sont pas invariants dans le temps

Sur le graphique 1-3 l'on peut penser que la relation $Y = c(1) + c(2)*X(t)$ recouvre en fait deux relations différentes:

une relation : $Y = c(1)' + c(2)''*X(t)$ pour $t < t^*$

et une relation $Y = c(1)'' + c(2)'''*X(t)$ pour $t > t^*$.

Afin de tester une telle possibilité l'on utilisera soit des variables muettes soit le test de Chow . Ce dernier est effectué en calculant :

$$F = \frac{SRC(5) / k}{SRC(4) / (N(1) + N(2) - 2k)}$$

Si SRC(1) est la somme des résidus au carré de la régression sur l'ensemble de la période ; SRC(2) celle sur la première période ; SRC(3) celle sur la seconde période , alors SRC(4) est défini comme la somme de SRC(2) et SRC(3) et SRC(5) comme la différence entre SRC(1) et SRC(4) ; k est le nombre de paramètres à estimer ; N(1) est le nombre d'observations de la première période ; N (2) le nombre d'observations de la seconde période . L'on compare alors le F ainsi obtenu avec le F tabulaire , calculé avec un degré de confiance donné , un degré de liberté égal à k au numérateur et égal à N(1) + N(2) - 2k au dénominateur . Si le F calculé excède le F tabulaire l'on peut rejeter l'hypothèse que les deux régressions sont identiques .

II – 6) La pêche miraculeuse à la bonne régression

La pratique économétrique se borne trop souvent à tester différents modèles à partir d'un ensemble de données et à sélectionner celui qui possède le meilleur t de Student , le meilleur R et le meilleur coefficient de Durbin Watson . Nous allons tenter de montrer pourquoi une telle pratique est en fait très insatisfaisante .

Prenons par exemple la recherche d'un modèle permettant d'expliquer la consommation macroéconomique . Plusieurs modèles théoriques s'offrent à l'économètre comme :

- une explication par le revenu courant et une constante soit :

$$(1) \quad C(t) = c(1) + c(2) * Y(t) + e(t)$$

- une explication par le revenu courant et la valeur de la consommation de la période précédente soit :

$$(2) \quad C(t) = c(1) + c(2)*Y(t) + c(3)*C(t-1) + e(t)$$

- une explication par la valeur moyenne du revenu courant sur les quatre dernières périodes d'observation soit :

$$(3) \quad C(t) = c(1) + c(2) * ((Y(t) + Y(t-1) + Y(t-2) + Y(t-3)) / 4) + e(t)$$

- une explication par le revenu courant et la valeur des actifs financiers soit :

$$(4) \quad C(t) = c(1) + c(2) * Y(t) + c(3) * A(t) + e(t)$$

Supposons que nous testions ces quatre modèles et que nous trouvions :

$$(1) \quad C(t) = 10.5 + 0.82 * Y(t) \quad R_{\text{aj}} = 0.71$$

(3.2) (2.2)

$$(2) \quad C(t) = 12.2 + 0.78 * Y(t) + 0.31 * C(t-1) \quad R_{\text{aj}} = 0.81$$

(2.3) (3.1) (1.8)

$$(3) \quad C(t) = 12.8 + 0.68 * (Y(t) + Y(t-1) + Y(t-2) + Y(t-3)) / 4 \quad R_{\text{aj}} = 0.85$$

(4.2) (4.1)

$$(4) \quad C(t) = 14 + 0.61 * Y(t) + 0.2 * A(t) \quad R_{\text{aj}} = 0.66$$

(3.1) (3.05) (2.1)

Nous serions tentés de choisir le troisième modèle comme le meilleur. En fait une telle procédure reste très ambiguë.

Considérons tout d'abord le critère du R_{aj} . Nous avons vu que la méthode des moindres carrés permet de fournir des estimations des paramètres en minimisant la somme des résidus au carré. Il en résulte que l'introduction d'une nouvelle variable explicative réduit la somme des résidus au carré même si cette nouvelle variable est faiblement corrélée avec la variable expliquée. L'on peut corriger partiellement cette difficulté en calculant le coefficient de corrélation ajusté (CCA) avec :

$$CCA^2 = \frac{1 - N - 1 (1 - R^2)}{N - n}$$

où N est le nombre d'observations et n le nombre de variables.

Toutefois même CCA^2 ne peut être utilisé pour effectuer une sélection entre différentes régressions. En effet supposons que nous voulions tester si l'addition d'une variable supplémentaire $X(k+1)$ donne un $CCA^2(k+1)$ significativement meilleur que $CCA^2(k)$. Ceci revient à tester si :

$$F = \frac{CCA^2(k+1) - CCA^2(k) \cdot (N - n - 1)}{1 - CCA^2(k+1)}$$

est significativement supérieur au F tabulaire. L'on constate qu'en manipulant N et n l'on peut toujours obtenir un F significatif même si l'augmentation de CCA^2 n'est pas elle-même

significative . D'autre part le test F suppose ici que le coefficient de corrélation est une variable stochastique. Or nous ne connaissons pas à priori la distribution des R^2 qui de toute façon ne peuvent être supposés identiques pour des modèles différents .

Considérons maintenant le critère du t de Student . Supposons que nous fassions la régression d'une variable $Y(t)$ sur des séries de n valeurs arbitraires d'une variable $X(t)$, soit

$$Y(t) = c(1) + c(2) * X(t)$$

L'on peut s'attendre à ce que le t de Student associé à $c(2)$ ne soit pas significatif . Toutefois si nous régressons $Y(t)$ sur un grand nombre de séries $X(t)$ prises au hasard il existe une probabilité qu'un certain pourcentage de ces séries présente un coefficient $c(2)$ avec un t significatif , par exemple 5 sur 100 séries . Or il serait parfaitement absurde de sélectionner parmi ces cinq résultats le coefficient qui a le t de Student le plus élevé . En d'autres termes plus le nombre de variables potentielles rejetées est élevé par rapport au nombre de variables préalablement sélectionnées , moins le résultat est significatif . Ce biais est quelquefois appelé " biais de Lovell " . .

Le t de Student n'a donc pas la même signification selon qu'il résulte du test d'un seul modèle ou de la sélection d'un modèle parmi beaucoup d'autres ... ce qu'omettent généralement de mentionner les auteurs des modèles économétriques .

CHAPITRE III

L' ESTIMATION ECONOMETRIQUE DES MODELES STRUCTURAUX

Nous définirons ici les modèles structuraux comme les modèles à équations simultanées dans lesquels une variable peut être exogène dans une équation et endogène dans une autre.

Par exemple le modèle dit de Phillips peut être décrit par les deux équations suivantes

$$3-1 \quad dw/dt = c(1) + c(2) dp/dt + c(3) dCh/dt + u(t)$$

$$3-2 \quad dp/dt = c(4) + c(5) dw/dt + c(6) dr/dt + c(7) dM/dt + e(t)$$

avec Ch : niveau de chômage

r : coût du capital

M : prix des matières premières importées

p : niveau des prix

w : niveau du taux de salaire

Un autre exemple est celui du modèle utilisé par le professeur Lawrence Klein au début des années 50 pour étudier les fluctuations de l'activité économique américaine :

$$3-3 \quad C(t) = c(1) + c(2) P(t) + c(3) (W(t) + W'(t)) + c(4) P(t-1) + u(1)$$

$$3-4 \quad I(t) = c(5) + c(6) P(t) + c(7) P(t-1) + c(8) K(t-1) + u(2)$$

$$3-5 \quad W(t) = c(9) + c(10) (Y(t) + T(t) - W'(t)) + c(11) (Y(t-1) + T(t-1) - W'(t-1)) + c(12) + u(3)$$

$$3-6 \quad Y(t) + T(t) = C(t) + I(t) + G(t)$$

$$3-7 \quad Y(t) = W'(t) + W(t) + P(t)$$

$$3-8 \quad K(t) = K(t-1) + I(t)$$

où C représente la consommation, I l'investissement, G les dépenses publiques, P les profits, W la masse salariale du secteur privé, W' la masse salariale du secteur public, K le stock de capital, T le montant des impôts, Y le revenu net, t le temps et où u(1), u(2) et u(3) sont des éléments stochastiques. c(1) ...c(12) sont des coefficients à déterminer à partir des régressions.

Dans ce modèle existent six variables endogènes : C, I, W, W', P et K et trois variables prédéterminées : P(t-1), K(t-1) et Y(t-1).

L'on remarque que des variables qui apparaissent exogènes dans une équation apparaissent endogènes dans une autre équation et sont donc soumises à des variations stochastiques. Par exemple dp/dt, exogène dans l'équation (3-1), est en fait une variable déterminée par un élément stochastique dans l'équation (3-2). De même dans l'équation (3-3), W apparaît être une variable exogène alors que d'après l'équation (3-4) elle est en fait une variable stochastique. Dans chacun de ces exemples est violée une des conditions fondamentales du modèle classique, à savoir que les variables explicatives ne doivent pas être aléatoires.

Pour résoudre ce problème nous utiliserons la méthode dite des « doubles moindres carrés »

III – 1) La méthode des doubles moindres carrés

Dans cette procédure l'on remplace les variables explicatives stochastiques par leur régression sur des variables non stochastiques dites « variables instrumentales ». Par exemple dans notre exemple précédent l'on peut chercher à remplacer dp/dt dans l'équation 3-1 par sa régression sur dr/dt et dM/dt qui joueront alors le rôle de variables instrumentales . L'on aura alors :

$$dw/dt = c(1)' + c(2)'(c(4)' + c(5)'(dr/dt) + c(6)'(dM/dt)) + c(3)'dCh / dt + u'(t)$$

(n.b. : remarquez que les coefficients $c(1)$, $c(2)$, $c(3)$, $c(5)$, $c(6)$ sont différents de $c(1)'$, $c(2)'$, $c(3)'$, $c(5)'$ et $c(6)'$).

Les variables retards peuvent être utilisées comme variables instrumentales.

Résoudre un système d'équations simultanées avec la méthode des doubles moindres carrés suppose de plus que soient respectées certaines règles :

- le système ne doit pas inclure d'identités : dans le modèle de Klein l'on ne considérera que les équations 3-3, 3-4 et 3-5 .
- les équations doivent être indépendantes
- pour chaque équation la somme des variables instrumentales et des constantes doit être au moins égale au nombre de variables endogènes figurant à droite de l'équation .

L'estimation simultanée si elle résout le problème des variables explicatives stochastiques n'est pas sans créer de nouveaux problèmes. Tout d'abord il n'existe pas toujours de variables instrumentales disponibles qui répondent à la double exigence d'être à la fois non stochastiques et fortement corrélées avec la variable explicative stochastique. L'on peut même dire qu'il s'agit d'exigences contradictoires. En second lieu toute erreur d'estimation dans une équation se transmet aux autres équations puisque celles ci sont interdépendantes. Dans ces conditions l'on peut être en droit d'effectuer des estimations équation par équation ou tout au moins comparer les résultats de celles ci avec celles de la méthode des doubles moindres carrés.

III-2) Estimer un système d'équations par la méthode des doubles moindres carrées avec E-view .

Object/ new object : system

Ecrire les équations .

Si les variables instrumentales sont les mêmes pour toutes les équations l'on ajoute une ligne débutant par « inst » suivie du nom des variables instrumentales :

Exemple : inst dM/dt dr/dt

Si les variables instrumentales sont différentes pour chaque équation l'on fait suivre chaque équation de @ et du nom des variables instrumentales pour l'équation .

Donner un nom au système en cliquant sur « name »

Estimer en cliquant sur « estimate equation ». L'on choisira ici la méthode des doubles moindres carrés (« TSLS »)

Exemple

L'on cherche à estimer le système suivant afin de modéliser la conjoncture algérienne :

$$ctr = c(1) + c(2) * rr$$

$$rr = c(3) + c(4) * fbcfr(-1) + c(5) * xr$$

$$xr = c(6) + c(7) * pp$$

$$fbcfr = c(8) + c(9) * (rr - rr(-1))$$

où

ctr représente la consommation totale réelle

rr le revenu total réel

fbcfr la formation brute de capital fixe réelle

xr les exportations réelles

pp le prix du pétrole sur le marché mondial

à partir des données suivantes (source : FMI)

obs	CTR	RR	FBCFR	PP	XR
1992	15.93878	27.04082	5.673469	19.00000	5.428571
1993	15.74576	25.50847	5.491525	16.00000	4.271186
1994	15.28571	25.01299	5.285714	16.00000	4.441558
1995	15.54000	26.28000	5.410000	17.00000	5.330000
1996	16.16949	28.20339	5.415254	20.00000	6.618644
1997	15.75781	27.28125	4.984375	19.00000	6.539062
1998	16.70992	27.24427	5.557252	13.00000	4.977099
1999	17.37037	30.01481	5.844444	18.00000	6.800000
2000	17.23571	35.92857	6.207143	28.00000	12.48571

Pour résoudre ce système dans Eviews à partir de ces données et en considérant que pp , fbcfr(-1) et rr(-1) sont variables instrumentales l'on ouvre :

objects/ new object/ system

puis l'on écrit le système

$inst\ pp\ fbcfr(-1)\ rr(-1)$

$ctr=c(1)+c(2)*rr$

$rr=c(3)+c(4)*xr$

$xr=c(5)+c(6)*pp$

$fbcfr=c(7)+c(8)*(rr-rr(-1))$

(l'on vérifiera que les conditions d'estimation de ce système sont remplies)

l'on clique alors sur « estimate » puis sur « two stages least squares » et l'on obtient :

(voir tableau page suivante)

Si les coefficients sont tous significatifs les statistiques de Durbin Watson ne sont pas bonnes...ce qui n'est guère surprenant du fait du faible nombre de nos variables explicatives.

Une fois le système estimé il faut lui donner un nom (ici » sysalgeria »).

L'on clique sur Procs/Make Model/Solve/OK

puis sur View / endogeneous table

et l'on obtient :

obs	CTRF	RRF	XRF	FBCFRF
1992	15.93878	27.04082	5.428571	5.673469
1993	15.99179	26.51434	5.094575	5.372811
1994	15.99179	26.51434	5.094575	5.421571
1995	16.09076	27.21745	5.658083	5.486689
1996	16.38768	29.32677	7.348610	5.616927
1997	16.28871	28.62366	6.785101	5.356452
1998	15.69488	24.40501	3.404048	5.030858
1999	16.18973	27.92055	6.221592	5.747164
2000	17.17944	34.95164	11.85668	6.072757

System: UNTITLED				
Estimation Method: Two-Stage Least Squares				
Date: 12/24/02 Time: 09:47				
Sample: 1993 2000				
Included observations: 8				
Total system (balanced) observations 32				
Instruments: PP FBCFR(-1) RR(-1) C				
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	12.25960	1.766320	6.940757	0.0000
C(2)	0.140761	0.062309	2.259078	0.0332
C(3)	20.15767	0.874110	23.06081	0.0000
C(4)	1.247732	0.127159	9.812374	0.0000
C(5)	-3.921567	1.466036	-2.674945	0.0132
C(6)	0.563509	0.077823	7.240920	0.0000
C(7)	5.421571	0.095906	56.52995	0.0000
C(8)	0.092615	0.043195	2.144100	0.0424
Determinant residual covariance		1.88E-05		
Equation: CTR=C(1)+C(2)*RR				
Observations: 8				
R-squared	0.606129	Mean dependent var	16.22685	
Adjusted R-squared	0.540484	S.D. dependent var	0.790211	
S.E. of regression	0.535665	Sum squared resid	1.721625	
Durbin-Watson stat	1.330788			
Equation: RR=C(3)+C(4)*XR				
Observations: 8				
R-squared	0.946929	Mean dependent var	28.18422	
Adjusted R-squared	0.938084	S.D. dependent var	3.502400	
S.E. of regression	0.871501	Sum squared resid	4.557089	
Durbin-Watson stat	1.447902			
Equation: XR=C(5)+C(6)*PP				
Observations: 8				
R-squared	0.897315	Mean dependent var	6.432908	
Adjusted R-squared	0.880200	S.D. dependent var	2.640111	
S.E. of regression	0.913797	Sum squared resid	5.010151	
Durbin-Watson stat	0.964296			
Equation: FBCFR=C(7)+C(8)*(RR-RR(-1))				
Observations: 8				
R-squared	0.649524	Mean dependent var	5.524464	
Adjusted R-squared	0.591111	S.D. dependent var	0.367292	
S.E. of regression	0.234863	Sum squared resid	0.330963	
Durbin-Watson stat	1.119624			

Remarquons ici que ce tableau fait suivre chaque variable de « F » afin de ne pas confondre la valeur de ces variables, solutions du modèle avec les valeurs historiques observées. L'on comparera alors les variables solutions du modèle avec les variables observées afin d'estimer la pertinence du modèle.

III - 3) Le modèle VAR

Dans un modèle VAR l'on ne se donne pas de modèle théorique à priori . Tout modèle peut en effet être arbitraire si les variables apparaissent à la fois à la droite et à gauche des équations et si nous n'avons pas de causalité .

Dans ce cas l'on peut chercher à régresser chaque variable endogène sur l'ensemble des autres variables , endogènes et exogènes . Le modèle sera déterminé en ne retenant que les coefficients significatifs .

Dans un tel modèle nous n'avons donc à effectuer aucune hypothèse théorique . Nous devons seulement :

- choisir les variables endogènes et exogènes ;
- choisir le nombre de délais à considérer .

Par exemple si nous avons deux variables endogènes X et Y et une variable exogène Z et deux délais notre modèle pourra s'écrire :

$$\begin{aligned} X &= c(0) + c(1)X(-1)+c(2)X(-2)+c(3)Y+c(4)Y(-1)+c(5)Y(-2)+c(6)Z+c(7)Z(-1)+c(8)Z(-2) \\ Y &= c(9)+c(10)Y(-1)+c(11)Y(-2)+c(12)X+c(13)X(-1)+c(14)X(-2)+c(15)Z+c(16)Z(-1)+c(17)Z(-2) \end{aligned}$$

Estimer un système VAR avec E-views

Quick/ Estimate Var

Pour créer un model à partir des coefficients estimés

Procs / make model

Exemple

A partir de nos données algériennes l'on obtient le modèle suivant en considérant que les variables endogènes sont fbcfr et xr , que les variables exogènes sont pp et une constante et qu'il existe deux retards :

III-5) La simulation

Une fois estimés les coefficients structuraux du système d'équations l'on obtient un modèle .

Dans la mesure où ce modèle comporte n équations indépendantes il permet de déterminer n variables endogènes à partir des variables prédéterminées .

L'on peut alors comparer les n variables endogènes ainsi obtenues à leurs valeurs historiques .

L'estimation des coefficients d'un modèle structural permet également d'effectuer des exercices de simulation .

III-6) La simulation dans E-view

- 1) *System / Procs / Make model (enregistrer sous un nom « Name*
- 1) *Ecrire le modèle . Ajouter une ligne « ASSIGN @ all F ». Cette commande assure que toutes les solutions du modèle apparaissent sous leur nom suivi de F afin de ne pas les confondre avec les valeurs historiques .*
- 1) *Solve / choisir « dynamic solution » et ne pas cocher « stop on missing data »*
- 1) *View / endogeneous table : les solutions du modèle apparaissent avec un F suivant leur nom .*
- 1) *View / endogeneous graphs*

Comparer les solutions du modèle aux valeurs historiques

NB . Le message « Near singular matrix » signifie que les équations du modèle ne sont pas indépendantes.

CHAPITRE IV

LA CAUSALITE

Un des problèmes soulevés par les modèles structurels est qu'ils ne nous donnent aucune indication sur les relations causales entre les variables. Or le sens de la causalité économique est un élément essentiel pour élaborer une politique économique ou pour effectuer des prévisions. Ce chapitre sera consacré à l'estimation économétrique de la causalité.

Il faut commencer par définir le terme de causalité. Littré définit la causalité comme " la vertu par laquelle une cause produit un effet " mais en ce sens il est impossible de mesurer la causalité. Les économètres vont utiliser une autre définition plus spécifique de la causalité due à Granger. L'on dira qu'une variable $X(t)$ exerce une causalité au sens de Granger sur une variable $Y(t)$ si l'ensemble des variables passées de X soit $X(t-1), X(t-2) \dots$ permet d'améliorer la prédiction de Y par rapport à une régression sur les valeurs passées de $Y(t-1), Y(t-2), \dots$. En d'autres termes l'on peut conclure que $X(t)$ exerce une causalité au sens de Granger sur $Y(t)$ si dans la régression :

$$4-1 \quad Y(t) = c(1)Y(t-1) + \dots + c(n)Y(t-n) + c(1)'X(t-1) + \dots + c(n)'X(t-n) + e(t)$$

les coefficients $c'(i)$, $i = 1, \dots, n$, sont significativement différents de 0, c'est à dire si 4-1 améliore la régression par rapport à 4-2

$$4-2 \quad Y(t) = c(1)Y(t-1) + \dots + c(n)Y(t-n) + e'(t)$$

Le test de Granger peut être utilisé pour étudier des relations entre deux variables $X(t)$ et $Y(t)$ par exemple la masse monétaire et le P.N.B. nominal, l'investissement et le P.N.B. réels etc...) quand on s'interroge sur le sens de leur relation causale. L'on estime alors les deux régressions suivantes :

4-3 :

$$Y(t) = c(1)Y(t-1) + \dots + c(n)Y(t-n) + c(1)'X(t-1) + \dots + c(m)'X(t-m) + u(t)$$

4-4 :

$$X(t) = c(1)X(t-1) + \dots + c(m)X(t-m) + c(1)''Y(t-1) + c(n)''Y(t-n) + u'(t)$$

L'on choisit les valeurs de n et de m de façon à ce que les résidus aléatoires $u(t)$ et $u'(t)$ ne soient pas corrélés.

Quatre cas sont alors possibles :

- si les coefficients $c(i)$, $c(i)'$, $c(i)''$, $c(i)'''$ sont tous significativement différents de 0 la causalité entre Y et X est bilatérale ;

- si les coefficients $c(i)'$ et $c(i)'''$ sont tous non significativement différents de 0 il n'existe pas de lien de causalité entre X et Y ;

- si les coefficients $c(i)'$ sont significativement différents de 0 alors que les coefficients $c(i)'''$ ne sont pas significativement différents de 0 la relation de causalité est de X à Y ;

- si les coefficients $c(i)'$ ne sont pas significativement différents de 0 alors que les coefficients $c(i)''$ sont significativement différents de 0 la relation de causalité est de Y à X.

L'on peut bien évidemment prolonger l'analyse en distinguant la causalité instantanée de la causalité retardée.

Le test de Granger n'est pas sans poser trois problèmes :

Il n'est d'abord pratiquement utilisable que dans le cas de deux variables même si théoriquement l'on peut envisager le cas de plusieurs variables .

Il est possible que X et Y soient déterminés toutes deux par une troisième variable Z qui agit avec des délais différents . L'observateur en conclura que X exerce une causalité au sens de Granger sur Y. Par exemple la vente de glaces en août exerce une causalité au sens de Granger sur la date des vendanges en octobre. Ces deux phénomènes ne sont bien évidemment pas liés entre eux par une relation de cause à effet mais dépendent tout deux du degré d'ensoleillement au mois d'août .

Enfin le test de Granger ne peut nous donner que des indications sur la causalité ex post et non sur celle ex ante. En fait nous ne pouvons guère utiliser l'estimation de relations causales dans le passé comme guide par la politique future dans la mesure où un changement de politique peut produire un changement structurel et donc influencer sur la nature des causalités .

La causalité dans E-views

Workfile/show/mettre le nom des variables

Group/view/Granger causality

CHAPITRE V

LES SERIES TEMPORELLES

Les séries temporelles peuvent être définies comme un ensemble d'observations ordonnées dans le temps. En ajoutant une dimension temporelle il est possible d'expliquer les variables endogènes non seulement par les variables explicatives présentes mais également par des variables explicatives passées, en particulier par des valeurs passées de la variable endogène.

Il existe deux grandes classes de séries temporelles : les séries auto-régressives et les séries à moyenne mobile. Dans le cas des séries auto-régressives (AR) les seules variables explicatives considérées sont les valeurs passées de la variable endogène. Les séries à moyenne mobile (MA) sont fondées sur une moyenne pondérée des résidus antérieurs. L'on peut bien évidemment combiner ces deux approches (séries ARMA).

Les séries temporelles peuvent être utilisées pour résoudre deux types de problèmes : celui de l'auto-corrélation des résidus et celui de la construction d'un modèle empirique sans modèle théorique explicatif, soit parce que nous ne connaissons ou ne pouvons mesurer les variables explicatives, soit parce que les coefficients estimés sont entachés d'un degré d'incertitude beaucoup trop élevé pour être utilisables à des fins prévisionnelles.

V- 1 L'utilisation d'une série autorégressive quand existe une auto-corrélation des résidus

Nous avons vu que quand existe une auto-corrélation des résidus ($DW \ll 2$) les estimations des coefficients sont biaisées.

Si nous ne parvenons pas à déterminer les raisons d'une telle auto-corrélation l'on peut chercher à décomposer le résidu entre d'une part la relation d'auto-corrélation et d'autre part le résidu véritablement aléatoire soit :

$$e = c(1) e(-1) + e'$$

où $e = c(1)e(-1)$ représente la relation d'auto-corrélation des résidus et où e' est un résidu aléatoire d'espérance mathématique nulle et tel que : $E(e', e'(-k)) = 0$ pour tout k .

L'on vérifiera que $c(1)$ est significativement différent de 0 et que les e' ne sont pas auto-corrélés en calculant le DW.

S'il n'en est pas ainsi l'on calculera une fonction auto-régressive des résidus d'un ordre supérieur par exemple :

$$e = c(1)e(-1) + c(2)e(-2) + e' \quad (\text{ordre 2})$$

$$e = c(1)e(-1) + c(2)e(-2) + c(3)e(-3) + e' \quad (\text{ordre 3})$$

et plus généralement

$$e = c(1)e(-1) + \dots + c(n)e(-n) + e' \quad (\text{ordre } n)$$

Pour calculer l'ordre n l'on peut recourir à la fonction d'auto-corrélation partielle qui montre le

gain obtenu sur le coefficient d'auto-corrélation en incluant la variable retard d'un ordre supplémentaire . n est obtenu quand cette fonction n'apporte pas de gain significatif pour n+1.

Procédure sous E-view

View/ residual tests /correlogram

Le corrélogramme nous permet de déterminer l'ordre n (voir V-2)

L'on ajoute alors AR(n) dans les variables explicatives de la régression .

V-2 La modélisation empirique des séries

Il existe trois procédures courantes qui permettent de modéliser à partir d'un échantillon de taille T des séries dont on ne connaît pas les variables explicatives mais dont on peut penser qu'elles suivent des lois temporelles : l'auto-régression ("autoregression" en abrégé AR) , l'intégration ("integration" en abrégé I) et la moyenne mobile ("moving average" en abrégé MA). Nous ne considérerons ici que des séries stationnaires c'est à dire des séries qui ont une moyenne constante et une covariance constante entre deux éléments séparés par k périodes quelque soit k . Le cas des séries non stationnaires sera examiné au § V-2 -4

Bien évidemment toutes ces procédures peuvent être utilisées pour la modélisation des résidus qui présentent une auto-corrélation.

V-2-1 : L'autoregression AR

Considérons par exemple le modèle :

$$(V-1) \quad Y = c(1) + c(2)*X + c(3)*Z$$

La prévision de Y va dépendre des estimations de $c(0)$, $c(1)$, $c(2)$ et des valeurs données aux variables exogènes X et Z ce qui peut conduire à une multiplication des erreurs . Par contre avec une série dite auto - régressive l'on utilisera une régression de type :

$$(V-2) \quad Y = c(1) + c(2)*Y(-1) + c(3)*Y(-2) + \dots + c(n-1)*Y(-n)$$

Dans l'équation (V-2) , $Y(-1) \dots Y(-n)$ sont connus ce qui réduit considérablement les erreurs possibles sur les données.

Ceci explique que si une équation comme celle (V-2) ne repose sur aucune explication économique, elle peut souvent conduire à de meilleures prédictions que celles de l'équation (V-1). L'on peut être ainsi conduit à utiliser un modèle auto - régressif quand :

- i) Il existe une réelle incertitude sur les variables causales en particulier quand la théorie économique ne permet pas de spécifier de façon suffisamment simple le modèle structurel sous-jacent
- ii) L'on peut soupçonner l'existence d'un cycle économique

iii) L'objet du modèle est d'effectuer une prévision à court terme ;

iv) Il existe une série longue d'observations sur la variable à prédire .

L'on peut par exemple utiliser les modèles auto-régressifs pour l'étude du mouvement des prix réels des matières premières dans la mesure où les séries disponibles sont longues , où le nombre de variables explicatives est élevé et où l'on cherche à effectuer des prévisions à court terme. Il serait par contre absurde d'utiliser de telles séries pour l'étude de l'inflation pour laquelle le nombre de variables explicative est faible et pour laquelle l'on sait qu' il n'existe pas de cycle .

Pour estimer une fonction d'auto – corrélation sous E – view il suffit de taper AR(n) comme seule variable explicative d'une régression . Il est toutefois nécessaire de connaître au préalable l'ordre n de la fonction AR(n).Pour cela l'on utilisera une fonction d'auto – corrélation partielle (ou corrélogramme) . Une telle fonction est construite à partir de la séquence des coefficients d'auto - corrélation partielle qui mesurent la corrélation entre y et y(-p) une fois pris en compte les effets de y(-1), y(-2)...y(-p+1) .

Par exemple dans la régression :

$$y = c(0) + c(1) y(-1) + e$$

le paramètre c(1) est le premier coefficient d'auto-corrélation partielle

Dans la régression :

$$y = c(0) + c(1)y(-1) + c(2) y(-2) + e$$

Le paramètre c(2) est le second coefficient d'auto-corrélation partielle . Il mesure l'association entre y(-2) et y une fois "enlevé" l'effet de y(-1) .

Pour choisir l'ordre p d'une fonction AR l'on choisira donc c(p) tel que c(p) soit statistiquement différent de 0 mais ne soit pas statistiquement différent de 0 pour c(p+1) .

Comme nos observations peuvent être considérées comme un échantillon de la série nous ne pouvons faire que des estimations des coefficients d'auto-corrélation partielle . Nous pouvons toutefois faire l'hypothèse que ces coefficients sont distribués suivant une loi normale de moyenne nulle et de variance $1/T$ si l'échantillon est suffisamment large . L'on peut alors comparer la valeur du test statistique $t = c(n) / (1/ (T) \exp 1/2)$ avec la valeur critique 1.96 (approximée par 2) .L'on peut alors tester si le coefficient d'auto-corrélation l'hypothèse que le coefficient d'auto-corrélation partielle est en fait nul. En pratique l'on retiendra cette hypothèse si t est inférieur à 2.

Par exemple si nous estimons les trois modèles AR suivants à partir d'un échantillon de 100 observations .

$$AR(1) : y = -0.166 + 0.464 y (-1)$$

$$AR(2) : y = - 0.114 + 0.212 y (-1) + 0.493 y (-2)$$

$$AR(3) : y = -0.115 + 0.212 y(-1) + 0.498 y(-2) - 0.017 y(-3)$$

Les trois coefficients estimés d'auto-corrélation partielle sont : 0.464 , 0.493 , -0.017 ce qui donne des valeurs du t statistique de 4.64 , 4.93 et -0.17 Seuls les deux premiers coefficients d'auto-corrélation partielle sont supérieurs à 2 de sorte que l'ordre de notre fonction est 2 .

Exemple :

L'on cherche à modéliser l'évolution de l'indice du prix mensuel du café (variable SERCAF) de janvier 1992 à mai 2001. Notre série de 113 observations est présentée dans le tableau V-1(source : FMI).

Il est difficile de construire un modèle théorique de cette série car l'évolution des cours du café dépend d'un grand nombre de facteurs (apparition de nouveaux producteurs , variations climatiques, phénomènes spéculatifs, accords entre producteurs...) . Par contre l'on sait que les cours du café obéissent à des cycles de sorte que l'on peut tester l'hypothèse que les prix sont auto-corrélés.

Pour tester cette hypothèse l'on estime tout d'abord le modèle AR(1) soit :

$$SERCAF = c(1) + c(2)*SERCAF(-1)$$

L'on obtient :

Dependent Variable: SERCAF				
Method: Least Squares				
Date: 12/25/02 Time: 18:19				
Sample(adjusted): 2 113				
Included observations: 112 after adjusting endpoints				
SERCAF=C(1)+C(2)*SERCAF(-1)				
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	1831.913	500.4480	3.660546	0.0004
C(2)	0.731927	0.064979	11.26404	0.0000
R-squared	0.535627	Mean dependent var	6834.679	
Adjusted R-squared	0.531406	S.D. dependent var	3565.590	
S.E. of regression	2440.787	Akaike info criterion	18.45572	
Sum squared resid	6.55E+08	Schwarz criterion	18.50427	
Log likelihood	-1031.521	Durbin-Watson stat	2.392569	

Le t de c(2) est excellent : il existe donc une forte auto-corrélation d'un mois sur l'autre du prix du café. Celle-ci explique l'existence d'une forte spéculation : quand un mouvement à la hausse ou à la baisse se déclenche il existe une très forte probabilité statistique qu'elle se prolonge le mois suivant .

—
—

obs	SERCAF
1	4107.000
2	3775.000
3	3925.000
4	3775.000
5	3565.000
6	2802.000
7	3082.000
8	3638.000
9	2934.000
10	2825.000
11	3711.000
12	3263.000
13	3661.000
14	3879.000
15	3847.000
16	3822.000
17	2688.000
18	3374.000
19	3389.000
20	3747.000
21	4363.000
22	4391.000
23	469.0000
24	4949.000
25	4942.000
26	4983.000
27	5119.000
28	5617.000
29	6195.000
30	7334.000
31	8552.000
32	11425.00
33	12993.00
34	13265.00
35	12537.00
36	11589.00
37	10684.00
38	10336.00
39	10885.00
40	11125.00
41	11087.00
42	10531.00
43	9689.000
44	9358.000
45	9346.000
46	915.0000
47	879.0000
48	8445.000
49	7703.000
50	8996.000
51	9538.000
52	9774.000
53	9429.000
54	8787.000
55	8054.000
56	7408.000
57	7472.000
58	7699.000
59	7852.000
60	7741.000

obs	SERCAF
61	792.0000
62	9571.000
63	11559.00
64	12524.00
65	12754.00
66	13193.00
67	12563.00
68	12236.00
69	11995.00
70	11892.00
71	11287.00
72	11848.00
73	12405.00
74	13025.00
75	12679.00
76	11574.00
77	979.0000
78	8683.000
79	7705.000
80	7528.000
81	7289.000
82	6912.000
83	7065.000
84	7241.000
85	7297.000
86	6723.000
87	6612.000
88	662.0000
89	6423.000
90	674.0000
91	6463.000
92	6158.000
93	5719.000
94	5577.000
95	6033.000
96	7016.000
97	746.0000
98	7095.000
99	6667.000
100	6419.000
101	6223.000
102	5705.000
103	5774.000
104	5656.000
105	5356.000
106	526.0000
107	516.0000
108	4783.000
109	4392.000
110	4286.000
111	4315.000
112	4147.000
113	4064.000

L'on peut tester une auto-corrélation d'ordre 2 suivant le modèle :

$$SERCAF = c(1) + c(2)*SERCAF(-1) + c(3)*SERCAF(-2)$$

et l'on obtient :

Dependent Variable: SERCAF				
Method: Least Squares				
Date: 12/25/02 Time: 18:22				
Sample(adjusted): 3 113				
Included observations: 111 after adjusting endpoints				
SERCAF=C(1)+C(2)*SERCAF(-1)+C(3)*SERCAF(-2)				
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	1361.477	518.2210	2.627214	0.0099
C(2)	0.534489	0.092682	5.766894	0.0000
C(3)	0.267428	0.092675	2.885659	0.0047
R-squared	0.566722	Mean dependent var	6862.243	
Adjusted R-squared	0.558698	S.D. dependent var	3569.753	
S.E. of regression	2371.406	Akaike info criterion	18.40701	
Sum squared resid	6.07E+08	Schwarz criterion	18.48024	
Log likelihood	-1018.589	Durbin-Watson stat	2.045628	

L'on constate que $c(2)$ et $c(3)$ sont très significativement différents de 0.

Plus généralement l'on peut effectuer le corrélogramme de notre série (voir page suivante). L'on constate que l'auto-corrélation est significative jusqu'à l'ordre 11 (ce qui peut être visualisé par le dépassement des traits en pointillé). En termes économiques les mouvements de prix sont systématiquement orientés dans le sens de la hausse ou de la baisse sur des périodes de 11 mois – celles qui séparent deux récoltes : il n'existe en effet qu'une récolte de café par an et toutes les récoltes se font dans la zone intertropicale.

V-2 -3 Les modèles à moyenne mobile (MA)

Dans une série à moyenne mobile d'ordre q chaque élément y est généré par une moyenne pondérée de l'ensemble des résidus aléatoires passés :

$$y = c(1) + e + c(2)*e(-1) + c(3)*e(-2) + \dots + c(q+1)*e(-q)$$

A titre d'exemple l'on peut supposer que y représente la variation de prix journalière sur le marché d'une matière première soumis à des chocs exogènes dénotés e soit :

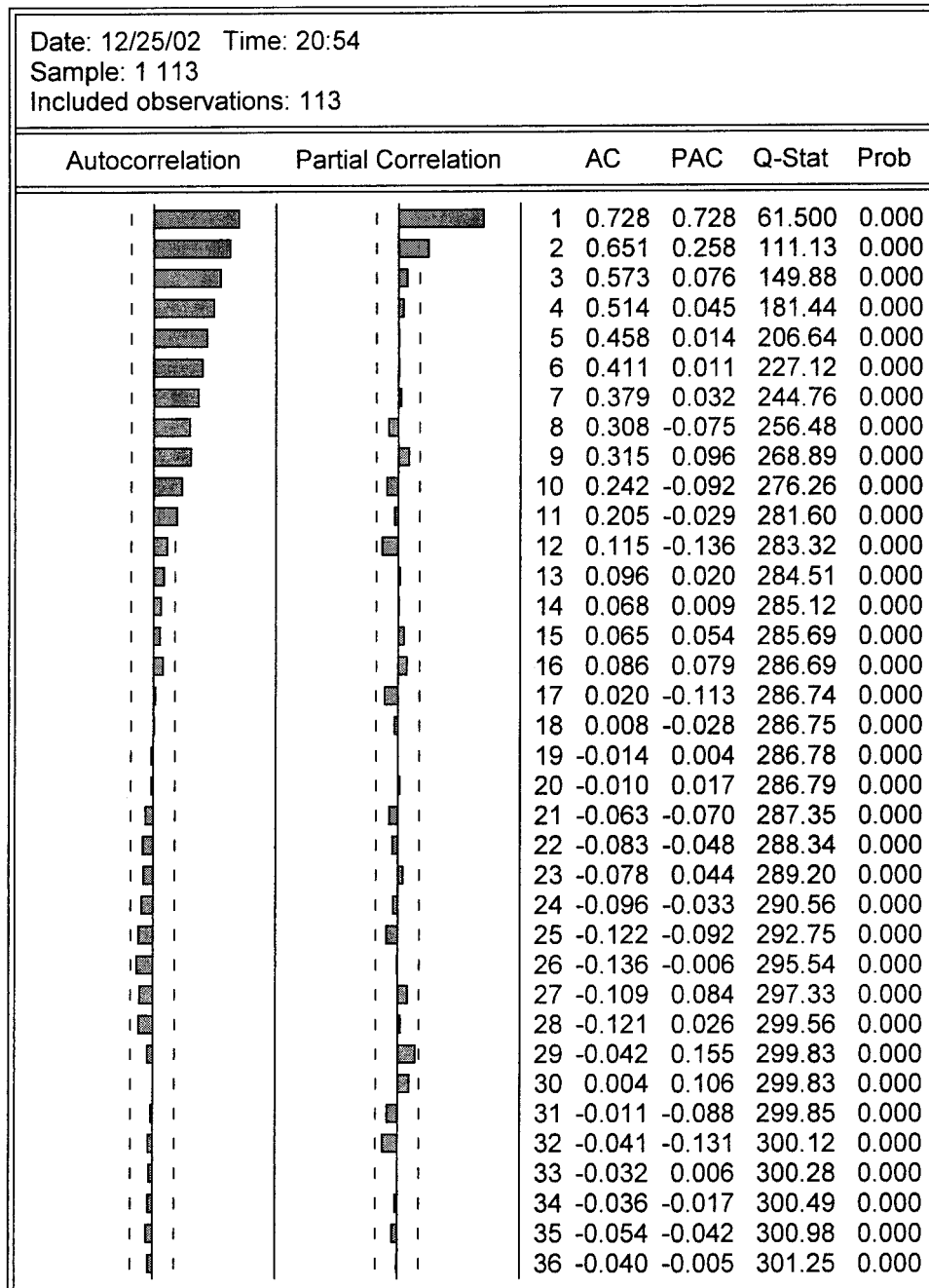
$$y(t) = c(1) + e(t)$$

L'on peut toutefois penser que l'impact de ce choc exogène ne sera pas complètement absorbé par le marché en une seule journée de sorte que l'on aura :

$$y(t+1) = c(1) + e(t+1) + c(2)*e(t)$$

où $e(t+1)$ représente l'impact du choc exogène $e(t)$ au temps $t+1$ et $c(2) e(t)$ l'impact au temps $t+1$ du choc exogène de la période t .

Correlogram of SERCAF



— Nous supposons que chaque élément aléatoire $e(t)$, ..., $e(t-q)$ a une espérance mathématique nulle et une même variance. Par exemple un modèle à moyenne mobile d'ordre 2 que l'on écrira MA (2) a pour équation :

$$y = e + c'(1) e(-1) + c'(2) e(-2)$$

De même que pour les modèles auto - régressifs il convient d'estimer l'ordre de la série .

L'ordre de la série

L'on peut montrer que le coefficient d'auto - corrélation d'un processus MA(q) est nul après q délais . Il est donc possible d'identifier l'ordre q d'un processus MA en déterminant le délai à partir duquel les auto - corrélations ne diffèrent plus significativement de 0 . Toutefois nous ne connaissons pas la véritable fonction d'auto- corrélation et celle ci doit être estimée à partir d'un échantillon de T observations . Si le nombre T est suffisamment large les auto - corrélations de l'échantillon sont distribuées suivant une loi Normale avec une moyenne nulle et un écart type moyen égal à $1 / (T) \exp 1/2$. L'on peut donc tester à un niveau de confiance donné , l'hypothèse que l'auto - corrélation d'un ordre q + 1 est nul en comparant l'auto - corrélation obtenue dans l'échantillon avec les limites de l'intervalle de confiance . Si cet intervalle de confiance ne comprend pas la valeur 0 alors l'on peut rejeter l'hypothèse que ce coefficient d'auto - corrélation est nul .

Estimation des paramètres d'un processus MA (q) sous E - views

L'estimation des coefficients d'une fonction MA se fera en introduisant MA (q) comme seule variable explicative dans une régression .

Exemple :

L'on cherche à modéliser un indice du prix du coton sur les dix principaux marchés américains (voir tableau page suivante ; source FMI)

L'on estime l'équation :

$$\text{coton} = c(1) + c(2)ma(1) + c(3)ma(2)$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6933.156	180.6719	38.37429	0.0000
MA(1)	1.109444	0.066406	16.70690	0.0000
MA(2)	0.571130	0.074980	7.617058	0.0000
R-squared	0.775839	Mean dependent var		7002.737
Adjusted R-squared	0.771800	S.D. dependent var		1515.976
S.E. of regression	724.1856	Akaike info criterion		16.03394
Sum squared resid	58213370	Schwarz criterion		16.10594
Log likelihood	-910.9343	F-statistic		192.0902
Durbin-Watson stat	1.205664	Prob(F-statistic)		0.000000
Inverted MA Roots	-.55 -.51i	-.55+.51i		

$c(2)$ et $c(3)$ sont significativement différents de 0

Remarque :

Le corrélogramme du coton montre qu'existe un cycle du coton comme il existe un cycle du café. Toutefois il est beaucoup moins marqué car :

- le coton est plus facilement stockable que le café

- l'élasticité prix du coton est plus forte que pour le café
- le coton est produit dans les deux hémisphères.

V – 2 – 4) Les modèles à la fois autorégressifs et à moyenne mobile (ARMA)

Il est possible de combiner les deux modèles précédents . Le modèle est dit alors ARMA et a pour ordre (p , q). Il peut être décrit par l'équation :

$$y = c(1) + c(2)*y(-1) + \dots + c(p+1)*y(-p) + e + c'(1)*e(-1) + \dots + c'(q)*e(-q)$$

Par exemple un modèle ARMA (1,1) sera décrit par :

$$y = c(1)+c(2)*y(-1) + e + c(3)*e(-1)$$

Comme le processus ARMA est une combinaison des processus AR et MA , la fonction d'auto- corrélation reflète à la fois les caractéristiques des processus AR et MA.

E-views propose systématiquement d'inclure ARMA dans les régressions

obs	SER01	obs	SER01
1	5475.000	61	7466.000
2	5507.000	62	7473.000
3	5404.000	63	7844.000
4	5693.000	64	7345.000
5	5974.000	65	7396.000
6	6154.000	66	7479.000
7	6544.000	67	7646.000
8	6192.000	68	7716.000
9	6006.000	69	753.0000
10	5296.000	70	7428.000
11	5334.000	71	7338.000
12	5603.000	72	6871.000
13	575.0000	73	6615.000
14	5757.000	74	6916.000
15	607.0000	75	705.0000
16	5904.000	76	648.0000
17	6013.000	77	6916.000
18	5827.000	78	7601.000
19	5853.000	79	7799.000
20	5699.000	80	7466.000
21	5763.000	81	7646.000
22	593.0000	82	73.00000
23	5718.000	83	7191.000
24	6327.000	84	632.0000
25	6909.000	85	6045.000
26	7684.000	86	5872.000
27	7787.000	87	6237.000
28	7947.000	88	6115.000
29	8318.000	89	6077.000
30	8331.000	90	5782.000
31	7351.000	91	5123.000
32	7633.000	92	534.0000
33	7844.000	93	5168.000
34	7133.000	94	5232.000
35	7735.000	95	5142.000
36	8536.000	96	5007.000
37	9247.000	97	5443.000
38	9618.000	98	5725.000
39	11616.00	99	6352.000
40	11283.00	100	5584.000
41	10655.00	101	6122.000
42	1152.000	102	5885.000
43	10527.00	103	5616.000
44	8946.000	104	6282.000
45	9496.000	105	6448.000
46	9183.000	106	6596.000
47	9016.000	107	6583.000
48	8894.000	108	6519.000
49	8773.000	109	6026.000
50	8709.000	110	5981.000
51	8517.000	111	52.00000
52	9189.000	112	461.0000
53	885.0000	113	445.0000
54	8741.000	114	4053.000
55	8248.000	115	4002.000
56	8254.000		
57	7889.000		
58	7703.000		
59	7255.000		
60	7735.000		

Correlogram of Residuals

Date: 12/26/02 Time: 11:56						
Sample: 1992:02 2001:07						
Included observations: 114						
Q-statistic probabilities adjusted for 2 ARMA term(s)						
Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
		1	0.377	0.377	16.597	
		2	0.548	0.474	52.093	
		3	0.638	0.528	100.66	0.000
		4	0.347	-0.044	115.13	0.000
		5	0.411	-0.202	135.63	0.000
		6	0.406	-0.068	155.78	0.000
		7	0.361	0.231	171.90	0.000
		8	0.313	0.106	184.12	0.000
		9	0.350	-0.012	199.53	0.000
		10	0.351	0.012	215.17	0.000
		11	0.288	0.039	225.81	0.000
		12	0.299	-0.029	237.43	0.000
		13	0.304	-0.026	249.55	0.000
		14	0.169	-0.187	253.30	0.000
		15	0.255	0.023	262.02	0.000
		16	0.137	-0.052	264.55	0.000
		17	0.088	-0.101	265.60	0.000
		18	0.194	0.087	270.80	0.000
		19	-0.007	-0.083	270.81	0.000
		20	0.063	-0.103	271.37	0.000
		21	0.100	0.038	272.80	0.000
		22	-0.089	-0.106	273.94	0.000
		23	0.049	0.029	274.29	0.000
		24	-0.034	-0.037	274.46	0.000
		25	-0.065	0.021	275.08	0.000
		26	-0.039	0.019	275.31	0.000
		27	-0.082	-0.012	276.34	0.000
		28	-0.080	-0.046	277.32	0.000
		29	-0.124	-0.020	279.71	0.000
		30	-0.116	-0.037	281.82	0.000
		31	-0.171	-0.110	286.50	0.000
		32	-0.169	0.012	291.13	0.000
		33	-0.156	0.060	295.12	0.000
		34	-0.231	-0.034	303.90	0.000
		35	-0.183	-0.019	309.49	0.000
		36	-0.109	0.143	311.52	0.000

CHAPITRE VI

LA COINTEGRATION

Dans une économie en croissance ou soumise à l'inflation la plupart des séries macroéconomiques possèdent un trend temporel. Elles sont dites « non stationnaires » car leur moyenne n'est pas constante dans le temps.

La corrélation entre deux séries non stationnaires s'avère souvent excellente sans qu'il existe pour autant une justification économique de cette corrélation. Leur corrélation ne naît que parce qu'elles augmentent toutes deux dans le temps. Par exemple le PNB nominal de l'Inde et celui de la France ont sur une longue période une très bonne corrélation sans que l'on puisse dire que l'un dépend de l'autre.

Dans ce chapitre nous commencerons par définir et par tester la non stationnarité des séries temporelles. Nous étudierons ensuite l'existence de relations stables entre les séries non stationnaires.

VI – 1) Le test de stationnarité

Une série autorégressive :

$$y = c(1) + c(2)*y(-1) + e$$

est dite stationnaire si $-1 < c(2) < 1$

Si $c(2) = 1$ la série est dite non stationnaire

Si $c(2) > 1$ la série est dite explosive. Ce type de série n'est pratiquement jamais rencontré par l'économiste.

Pour savoir si une série est stationnaire il faut donc tester si $c(2)$ est significativement différent de 1 ou , en posant $c'(2) = 1 - c(2)$, si, dans :

$$dy = c'(1) + c'(2) *y (-1) + e$$

$c'(2)$ est significativement différent de 0 .

Malheureusement l'on peut montrer que la table de distribution du t de Student ne peut être ici utilisée. Il faut utiliser pour ce test une table dite de Dickey-Fuller.

Le test de stationnarité peut être généralisé à une série autorégressive d'ordre n . L'on utilise alors une table dite table de Dickey-Fuller augmentée.

Exemple :

L'on cherche à savoir si la production industrielle des pays de l'OCDE est ou non stationnaire à partir de la série mensuelle suivante de janvier 1992 à mai 2001 (source : FMI)

obs	SERINDPRO
1992:01	9440.000
1992:02	9471.000
1992:03	9430.000
1992:04	9454.000
1992:05	9390.000
1992:06	9403.000
1992:07	9413.000
1992:08	9307.000
1992:09	9381.000
1992:10	9358.000
1992:11	9298.000
1992:12	9192.000
1993:01	9253.000
1993:02	9296.000
1993:03	9364.000
1993:04	9299.000
1993:05	9244.000
1993:06	9255.000
1993:07	9231.000
1993:08	9252.000
1993:09	9318.000
1993:10	9252.000
1993:11	9297.000
1993:12	9313.000
1994:01	9353.000
1994:02	9379.000
1994:03	9502.000
1994:04	9533.000
1994:05	9558.000
1994:06	9642.000
1994:07	9621.000
1994:08	9737.000
1994:09	9728.000
1994:10	9777.000
1994:11	9876.000
1994:12	10004.000
1995:01	9924.000
1995:02	9987.000
1995:03	9996.000
1995:04	9970.000
1995:05	10031.000
1995:06	9997.000
1995:07	9912.000
1995:08	10027.000
1995:09	10005.000
1995:10	9995.000
1995:11	10060.000
1995:12	10095.000
1996:01	10082.000
1996:02	10124.000
1996:03	10038.000
1996:04	10181.000
1996:05	10232.000
1996:06	10162.000
1996:07	10286.000
1996:08	10277.000
1996:09	10342.000
1996:10	10382.000
1996:11	10408.000
1996:12	10416.000

obs	SERINDPRO
1997:01	10544.000
1997:02	10555.000
1997:03	10551.000
1997:04	10628.000
1997:05	10677.000
1997:06	10697.000
1997:07	10811.000
1997:08	10740.000
1997:09	10826.000
1997:10	10869.000
1997:11	10746.000
1997:12	10863.000
1998:01	10868.000
1998:02	10800.000
1998:03	10805.000
1998:04	10754.000
1998:05	10705.000
1998:06	10756.000
1998:07	10757.000
1998:08	10765.000
1998:09	10809.000
1998:10	10794.000
1998:11	10757.000
1998:12	10725.000
1999:01	10806.000
1999:02	10808.000
1999:03	10938.000
1999:04	10830.000
1999:05	10850.000
1999:06	10968.000
1999:07	11028.000
1999:08	11188.000
1999:09	11168.000
1999:10	11135.000
1999:11	11316.000
1999:12	11295.000
2000:01	11306.000
2000:02	11385.000
2000:03	11459.000
2000:04	11514.000
2000:05	11620.000
2000:06	11613.000
2000:07	11635.000
2000:08	11805.000
2000:09	11662.000
2000:10	11699.000
2000:11	11715.000
2000:12	11754.000
2001:01	11550.000
2001:02	11576.000
2001:03	11459.000
2001:04	11333.000
2001:05	11256.000
2001:06	11178.000

Dans E-views l'on rentre la série puis l'on clique sur :

View/Unit root test/

L'on choisit alors comme test : Augmented Dickey – Fuller

Le tableau suivant apparaît :

Augmented Dickey-Fuller Unit Root Test on SERINDPRO

ADF Test Statistic	-1.373637	1% Critical Value*	-3.4911	
		5% Critical Value	-2.8879	
		10% Critical Value	-2.5807	
*MacKinnon critical values for rejection of hypothesis of a unit root.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(SERINDPRO)				
Method: Least Squares				
Date: 12/27/02 Time: 11:08				
Sample(adjusted): 1992:06 2001:06				
Included observations: 109 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
SERINDPRO(-1)	-0.011766	0.008566	-1.373637	0.1725
D(SERINDPRO(-1))	-0.205615	0.096408	-2.132753	0.0353
D(SERINDPRO(-2))	0.138282	0.093627	1.476944	0.1427
D(SERINDPRO(-3))	0.417186	0.097906	4.261084	0.0000
D(SERINDPRO(-4))	0.158363	0.102954	1.538188	0.1271
C	128.7869	88.61824	1.453278	0.1492
R-squared	0.180154	Mean dependent var	16.40367	
Adjusted R-squared	0.140355	S.D. dependent var	73.12966	
S.E. of regression	67.80365	Akaike info criterion	11.32458	
Sum squared resid	473525.5	Schwarz criterion	11.47273	
Log likelihood	-611.1897	F-statistic	4.526657	
Durbin-Watson stat	2.041518	Prob(F-statistic)	0.000903	

L'ADF de la première ligne est de -1.373637. Il est donc inférieur à toutes les valeurs critiques. L'on ne peut donc rejeter l'hypothèse que les $c(i)$ soient en fait nuls et donc que les $c(i)$ soient en fait égaux à 1, c'est-à-dire que la série soit en fait non stationnaire. La production industrielle des pays de l'OCDE n'a donc pas stagné entre 1992 et 2001.

VI – 2) La cointégration

L'étude de la cointégration permet de tester l'existence d'une relation stable de long terme entre deux variables non stationnaires , en incluant des variables retards et des variables exogènes .

Il existe plusieurs tests de la cointégration le plus général étant celui de Johanssen qui est celui utilisé par E-views . Quelque soit le test retenu il n'a de signification que sur des séries non stationnaires longues , supérieures à 100 observations.

Nous ne présenterons pas ici l'interprétation du test dans la mesure où E-views présente directement les résultats sous forme normalisée.

Exemple :

Nous cherchons à savoir s'il existe une relation stable de long terme entre les variations mensuelles de prix aux Etats-Unis et les variations mensuelles de la masse monétaire américaine M1 entre janvier 1992 et mai 2001 (voir tableau page suivante . Source FMI)

Nous commençons par tester si les deux séries sont effectivement non stationnaires.

Le test de Johanssen s'effectue sous :

View/cointegration test

Le logiciel nous demande s'il existe un trend dans chaque série (nous répondons oui car nous ne travaillons pas sur une série de la masse monétaire réelle : il existe donc une dérive inflationniste), d'éventuelles variables exogènes (non) et le nombre de délais à prendre en considération (nous prenons six délais car l'on peut considérer que l'impact de l'augmentation de la masse monétaire sur les prix s'effectue dans les six mois).

Les résultats apparaissent sous la forme suivante :

obs	P	M
1992:01	9164.000	9063.000
1992:02	9151.000	9095.000
1992:03	9296.000	9141.000
1992:04	9537.000	9155.000
1992:05	9429.000	9168.000
1992:06	9510.000	9200.000
1992:07	9620.000	9220.000
1992:08	9701.000	9246.000
1992:09	9827.000	9273.000
1992:10	10008.00	9305.000
1992:11	10214.00	9319.000
1992:12	10456.00	9312.000
1993:01	10398.00	9358.000
1993:02	10219.00	9391.000
1993:03	10305.00	9424.000
1993:04	10574.00	945.0000
1993:05	10568.00	9463.000
1993:06	10718.00	9476.000
1993:07	10831.00	9476.000
1993:08	10881.00	9502.000
1993:09	10988.00	9522.000
1993:10	11116.00	9561.000
1993:11	1129.000	9568.000
1993:12	11533.00	9568.000
1994:01	11417.00	9594.000
1994:02	11238.00	9627.000
1994:03	11311.00	966.0000
1994:04	11522.00	9673.000
1994:05	11320.00	9680.000
1994:06	11416.00	9712.000
1994:07	11503.00	9739.000
1994:08	11437.00	9778.000
1994:09	11461.00	9804.000
1994:10	11477.00	9811.000
1994:11	11559.00	9824.000
1994:12	11742.00	9824.000
1995:01	11590.00	9863.000
1995:02	11349.00	9903.000
1995:03	11389.00	9935.000
1995:04	11599.00	9968.000
1995:05	11335.00	9988.000
1995:06	11404.00	10008.00
1995:07	11452.00	10008.00
1995:08	11389.00	10034.00
1995:09	11380.00	10054.00
1995:10	11324.00	10086.00
1995:11	11380.00	10080.00
1995:12	11521.00	10073.00
1996:01	11294.00	10132.00
1996:02	11051.00	10165.00
1996:03	11173.00	10218.00
1996:04	11311.00	10257.00
1996:05	11051.00	10277.00
1996:06	11142.00	10283.00
1996:07	11100.00	10303.00
1996:08	19970.00	10323.00
1996:09	10912.00	10356.00
1996:10	10776.00	10388.00
1996:11	10864.00	10408.00
1996:12	11047.00	10408.00

obs	P	M
1997:01	10863.00	10441.00
1997:02	10652.00	10474.00
1997:03	10675.00	10500.00
1997:04	10731.00	10513.00
1997:05	10536.00	10506.00
1997:06	10641.00	10520.00
1997:07	10653.00	10533.00
1997:08	10687.00	10552.00
1997:09	10591.00	10579.00
1997:10	10571.00	10605.00
1997:11	10736.00	10598.00
1997:12	10969.00	10585.00
1998:01	10794.00	10605.00
1998:02	10655.00	10625.00
1998:03	10752.00	10644.00
1998:04	10869.00	10664.00
1998:05	10700.00	10684.00
1998:06	10743.00	10697.00
1998:07	10735.00	10710.00
1998:08	10686.00	10723.00
1998:09	10700.00	10736.00
1998:10	10768.00	10762.00
1998:11	10973.00	10762.00
1998:12	11204.00	10756.00
1999:01	11030.00	10782.00
1999:02	10843.00	10795.00
1999:03	10966.00	10828.00
1999:04	11126.00	10907.00
1999:05	10954.00	10907.00
1999:06	10972.00	10907.00
1999:07	10967.00	10940.00
1999:08	10927.00	10966.00
1999:09	10863.00	11018.00
1999:10	10953.00	11038.00
1999:11	11133.00	11045.00
1999:12	11483.00	11045.00
2000:01	11264.00	11077.00
2000:02	10968.00	11143.00
2000:03	11081.00	11235.00
2000:04	11249.00	11241.00
2000:05	11004.00	11255.00
2000:06	11026.00	11314.00
2000:07	11039.00	11340.00
2000:08	10958.00	11340.00
2000:09	10905.00	11399.00
2000:10	10935.00	11419.00
2000:11	10933.00	11425.00
2000:12	11124.00	11419.00
2001:01	10998.00	11491.00
2001:02	10878.00	11537.00
2001:03	11078.00	11563.00
2001:04	11232.00	11609.00
2001:05	11115.00	11661.00
2001:06	11230.00	11681.00

Johansen Cointegration Test

Date: 12/28/02 Time: 08:12				
Sample: 1992:01 2001:06				
Included observations: 107				
Test assumption: Linear deterministic trend in the data				
Series: P M				
Lags interval: 1 to 6				
Eigenvalue	Likelihood Ratio	5 Percent Critical Value	1 Percent Critical Value	Hypothesized No. of CE(s)
0.120722	15.62388	15.41	20.04	None *
0.017213	1.857852	3.76	6.65	At most 1
*(**) denotes rejection of the hypothesis at 5%(1%) significance level				
L.R. test indicates 1 cointegrating equation(s) at 5% significance level				
Unnormalized Cointegrating Coefficients:				
	P	M		
	0.000164	-5.53E-05		
	2.51E-05	0.000101		
Normalized Cointegrating Coefficients: 1 Cointegrating Equation(s)				
	P	M	C	
	1.000000	-0.336693	-7528.458	
		(0.17354)		
	Log likelihood	-1804.788		

Remarque : Les résultats sont présentés en différences , non en valeurs absolues.

Ces résultats s'interprètent de la façon suivante :

- *il existe une seule relation de cointégration dans l'intervalle de confiance de 5% du test de vraisemblance*
- *la relation normalisée est :*

$$dp = 0.33 dm + 7528$$

(remarque : pour que la variable dépendante p soit normalisée il faut qu'elle soit dans la première colonne de gauche de notre tableau de données)

En d'autres termes une hausse de 10% de la masse monétaire M1 conduit dans les six mois à une hausse des prix de 3,3%.

CONCLUSION

La théorie économétrique a été élaborée dans les années 40 en supposant d'une part que la théorie économique est capable de fournir des modèles directement testables et d'autre part que la confrontation avec les données nous permet de rejeter ou d'accepter sans ambiguïté une théorie.

Il s'agit, nous l'avons vu, d'une double illusion.

Tout d'abord les modèles théoriques sont loin d'être toujours utilisables par l'économètre. En effet certaines variables ne sont ni directement observables ni directement mesurables. D'autre part de nombreux modèles théoriques restent insuffisamment spécifiés. Par exemple ils n'indiquent pas les délais à prendre en compte. Enfin les théories économiques sont formalisées de façon trop générale pour être testables (comme la théorie walrasienne) soit sont construites sous l'hypothèse "Ceteris Paribus". Dans ce dernier cas nous ne pouvons savoir si la non correspondance entre le modèle théorique et les observations proviennent de la mauvaise spécification de la fonction ou bien d'un changement dans l'environnement du modèle.

Les difficultés proprement statistiques ne sont pas moins réelles. Trop souvent les données utilisées par l'économètre restent très approximatives - pour ne pas dire fausses - ce qui rendent douteux les résultats obtenus ; D'autre part les méthodes statistiques ne sont souvent justifiées que pour des séries très longues alors que l'économiste ne dispose généralement que de séries courtes. Au reste pour des séries très longues l'homogénéité des variables pose souvent problème.

Ces difficultés ne doivent pas cacher l'importance de l'économétrie dans l'analyse économique contemporaine. S'il n'est souvent pas possible de rejeter une théorie économique sur la seule base des résultats économétriques, les tests d'acceptabilité statistique sont devenus des procédures routinières, d'autant que les programmes informatiques sont aujourd'hui facilement disponibles. L'économétrie a du reste contribué à modifier la théorie économique dans la mesure l'exigence de modèles testables devient une condition d'acceptabilité dans les publications scientifiques.

BIBLIOGRAPHIE

- niveau facile ** niveau moyen *** niveau difficile

- CHAREMZA W.W. et DEADMAN D.F. « New Directions in Econometric Practice » Edward Elgar , 1992 (***)
- CUTHBERTSON K. , HALL S.G. et TAYLOR M.P. « Applied Econometric Techniques » Philip Allan 1992 (***)
- DESAI M. « Testing Monetarism » Frances Pinter , Londres 1981(**)
- DODGE Y. « Statistique . Dictionnaire Encyclopédique » Dunod , 1993 (*)
- DOUGHERTY « Introduction to Econometrics » Oxford University Press 1992 (*)
- ENGLE R.F. et GRANGER C.W.J. (eds) « Long Run Economic Relations : Readings in Cointegration » Oxford University Press 1991 (***)
- GHOSH S.K. « Econometrics » Prentice Hall 1991 (**)
- GRANGER , C.W. J. et NEWBOLD P. « Forecasting Economic Time Series » Academic Press , 1986 (***)
- GRIFFITHS W.E. , HILL R.C. et JUDGE G.G. « Learning and Practicing Econometrics » John Wiley & Sons 1993 (**)
- GUJARATI D.N. « Basic Econometrics » Mac Graw Hill 1988 (*)
- HARVEY « The Econometric Analysis of Time Series » Philip Allan 1990 (**)
- JUDGE, G.G. , HILL, R.C. , GRIFFITHS W.E. , LUTKEPOHL H. et LEE , T.C. « Introduction to the Theory and Practice of Econometrics » John Wiley & Sons , 1988 (**)
- LUCAS R.E. « Economic Policy Evaluation : A Critique » in Brunner K. et Metzler A.M. (eds) « The Phillips Curve and Labor Markets » Carnegie Rochester Conference Series on Public Policy , vol. 1 , North Holland (1976)
- MACKINNON , J.G. « Critical Values for Cointegration Tests » in R.F. Engle et C.W.J Granger (eds) 1991.
- PINDYCK S.R. et RUBINFELD D.L. « Econometric Models & Econometric Forecasts » Mc Graw Hill 1991 (**)
- SALVATORE « Econométrie et Statistique Appliquées » Mc Graw Hill 1985 (*).
- SIMS « Money Income and Causality » American Economic Review 1972 (**)
- SIMS C.A. « Macroeconomics and Reality » Econometrica , vol .48, 1980 (***)
- WONNACOTT T.H et R. J « Statistique » Economica 1988 (*)
- WONNACOTT T.H et R.J « Econometrics » John Wiley 1979 (*)

TABLE DES MATIERES

Chapitre 1 Le modèle classique de régression linéaire

I – 1) La droite de régression

I – 2) Le problème des observations aberrantes

I – 3) La prise en compte des variables qualitatives explicatives

I – 4) La régression dans E – views

Annexe 1

Chapitre II La remise en cause des hypothèses du modèle linéaire classique

II – 1) Les hypothèses du modèle classique de régression linéaire

II – 2) La remise en cause de la première hypothèse : les résidus sont auto-corrélés

II – 3) La remise en cause de la seconde hypothèse : l'hétéroscédasticité

II – 4) Le non respect de la troisième hypothèse : la multicollinéarité

II – 5) Le non respect de la quatrième hypothèse : les variables explicatives sont elles mêmes des variables stochastiques

II – 6) Le non respect de la cinquième hypothèse : les coefficients du modèle ne sont pas invariants dans le temps

II – 7) La pêche miraculeuse à la bonne régression

Chapitre III L'estimation économétrique des modèles structuraux

III – 1) La méthode des doubles moindres carrés

III – 2) Estimer un système d'équations par la méthode des doubles moindres carrés avec E – views

III – 3) Le modèle VAR

Chapitre IV La causalité

Chapitre V Les séries temporelles

V – 1) L'utilisation d'une série auto-régressive quand existe une auto-corrélation des résidus

V – 2) La modélisation empirique des séries

Chapitre VI La cointégration

VI – 1) Le test de stationnarité

VI – 2) La cointégration

Conclusion

Bibliographie