VARIATIONS SUR LA RÉGRESSION LINÉAIRE MULTIPLE

Ce chapitre expose la pratique élémentaire de la régression multiple. Une très grande variété de méthodes et de tests ont été proposés par les économètres, des plus simples aux plus complexes, on présente ici les plus usuels.

Des méthodes plus avancées, faisant appel par exemple à la théorie des séries chronologiques, seront étudiées ultérieurement ; elles pourront donner un éclairage nouveau et quelque peu différent à ce qui suit.

PRINCIPES GÉNÉRAUX

La pratique correcte de la régression demande bon sens, prudence et circonspection.

L'économètre se doit de connaître les idées économiques comme les données disponibles concernant le domaine qu'il étudie. Il doit également connaître ses séries, leur origine, leur définition précise, et avoir examiné leur évolution, leurs particularités, avant de les employer dans des régressions ou d'autres méthodes élaborées.

Par ailleurs, il faut toujours demander le calcul des résidus et leur représentation graphique. On va voir qu'ils permettent en effet de tester *a posteriori* les hypothèses faites sur l'aléa et plus généralement sur le modèle.

VALEURS ATYPIQUES

Il arrive qu'une, ou plusieurs observations paraissent ne pas se conformer au modèle, bien que l'équation estimée semble satisfaisante. Pour une telle observation, la valeur ajustée reste très éloignée de la valeur réelle, ou ce qui revient au même, le résidu est très important.

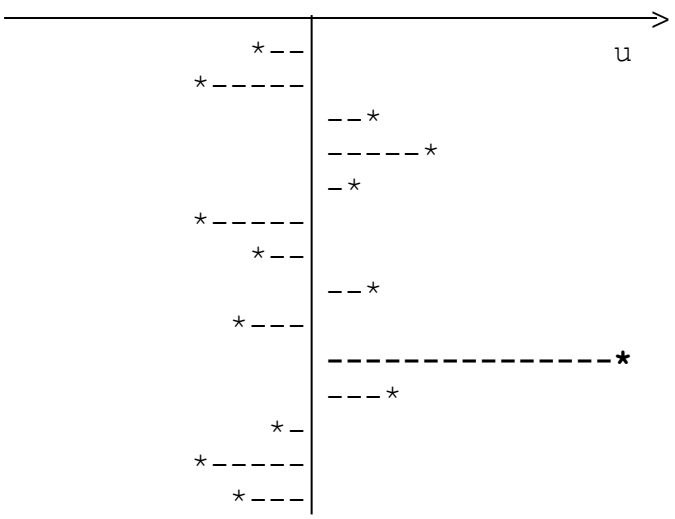
Diagnostic

Une observation *atypique* se signale par un écart important sur le diagramme des résidus.

Si on désire un critère quantifié, on peut retenir, par exemple, les résidus qui dépassent deux fois l'écart-type estimé de l'aléa en valeur absolue.

L'emploi de données ne suivant plus le modèle dans les procédures d'estimation met naturellement à mal les qualités des estimateurs des mco.

Exemple :



Solutions

Lorsque des valeurs atypiques ont été trouvées, il faut commencer par vérifier les données correspondantes, des erreurs ayant pu se glisser dans les chiffres. Si ce n'est pas le cas, il s'agit en général d'observations particulières pour lesquelles on pouvait s'attendre à des singularités (par exemple en octobre 1929, en mai 1968...)

On peut réestimer l'équation en supprimant la, ou les observations contestées. Et on examine en particulier comment sont modifiées les différentes estimations.

Cependant, une certaine dispersion des résidus est la conséquence normale de la présence d'un aléa; il serait absurde de vouloir supprimer systématiquement toutes les observations un peu marginales !

LES DONNÉES

Données temporelles

Les séries statistiques utilisées pour estimer les équations économétriques sont la plupart du temps des ***séries temporelles*** (annuelles, trimestrielles, etc.). La structure particulière de telles séries peut perturber les hypothèses des mco et altérer la qualité des régressions.

Il faut d'autre part veiller à bien employer des données cohérentes, et choisir par exemple si l'on prend les séries en valeurs dites ***nominales*** (ou encore ***courantes***) ou, ce qui est généralement préférable, en valeurs convenablement ***déflatées***, encore dites ***constantes***.

Dans le même esprit il faut choisir avec attention entre les données en échelle et les données *per capita*, ou rapportées à une autre grandeur.

Des données mesurées à une période unique sont dites ***en coupe instantanée*** ou ***transversales***, par opposition aux séries temporelles, encore dites ***longitudinales***. Certaines études utilisent des

données, dites ***croisées***, ou encore ***de panel***, qui relèvent une série d'observations individuelles pour des périodes successives, ainsi le ratio bénéfice/CA annuel mesuré pour 85 entreprises pendant 20 ans. L'usage de telles séries requiert des précautions, voire des méthodes, particulières.

Variables retardées

On a déjà noté la possibilité de valeurs ***retardées*** de variables exogènes dans une équation économétrique. Ainsi dans la fonction de salaire :

$$W = a + b.R + c.R_{-1} + d.t + \varepsilon$$

où R est le revenu national, t, le temps et ε l'aléa.

De telles variables retardées ne présentent pas de problème théorique.

Différent est le cas où des valeurs retardées de la variable expliquée elle-même apparaissent comme explicatives. Ainsi dans la fonction de consommation :

$$C = a + b.C_{-1} + c.R + \varepsilon$$

De tels modèles sont dits ***autorégressifs***. Le problème est que cette variable décalée dépend elle-même du modèle (par les périodes précédentes), et que strictement on ne devrait pas l'employer sans précautions dans l'estimation de l'équation.

En fait, cette situation est commune, en particulier dans le cas des modèles économétriques à plusieurs équations, et relève du formalisme des *séries temporelles*, Le plus souvent, on utilise cependant les mco, du moins dans un premier temps, puis on met en œuvre, si l'on dispose des outils appropriés, les techniques plus perfectionnées d'étude des séries temporelles.

MULTICOLINÉARITÉ

Il ne s'agit pas là d'un problème lié à l'aléa, mais aux données disponibles pour l'estimation.

Multicolinéarité stricte

Deux colonnes de nombres (ou *vecteurs*, par exemple les observations de la variable X et celles de Z) sont dites ***colinéaires*** si elles sont proportionnelles. Plus généralement, plusieurs vecteurs sont colinéaires s'il est possible d'exprimer l'un d'entre eux comme combinaison linéaire des autres (ainsi si Z = 2.X - 0,5.W).

Exemple : soit le modèle à trois explicatives:

$$Y = a + b.X + c.Z + d.T + \varepsilon$$

On suppose que sur les observations disponibles, X et Z sont exactement colinéaires, et dans un rapport 3 (Z = 3.X).

On suppose que l'équation estimée:

$$Y = A + B.X + C.Z + D.T$$

satisfait le critère des moindres carrés, les valeurs retenues : A, B, C et D, minimisant la somme des carrés des résidus.

La nouvelle équation estimée :

$$Y = A + (B+3).X + (C-1).Z + D.T$$

obtenue en remplaçant « une unité » de Z par trois de X, donne les mêmes résidus et convient donc aussi bien; en fait il est impossible dans l'estimation de distinguer l'influence de X de celle de Z.

Plus généralement, la multicolinéarité stricte entre les variables explicatives rend la régression impossible (certains logiciels identifient et signalent cette situation). Cette situation est néanmoins facilement évitable, car elle résulte le plus souvent d'une relation évidente entre explicatives envisagée , par exemple vouloir utiliser le budget civil, le budget militaire et le budget total pour un ensemble de pays !

Multicolinéarité approchée

Si la multicolinéarité exacte est très improbable (ou alors due à la naïveté du choix des explicatives), la multicolinéarité approchée des observations des variables explicatives envisagées est possible, en particulier dans le domaine des séries temporelles. On conçoit que certaines explicatives étant très "proches", la précision des estimations les concernant en soit affectée.

Diagnostic

La multicolinéarité est l'une des explications possibles à de mauvais tests de significativité dans l'estimation d'un modèle a priori pertinent, à des valeurs, voire des signes, déconcertants pour des paramètres estimés, ou encore à une très forte sensibilité de ceux-ci à l'ajout ou à la suppression d'observations.

L'examen de la matrice des corrélations entre les explicatives permet de repérer la corrélation éventuelle de couples de variables explicatives. Les coefficients de corrélation multiple : R^2, associés aux régressions de chaque explicative sur l'ensemble des autres, calculés par les bons logiciels économétriques, permettent d'identifier des multicolinéarité impliquant plus de deux variables (une valeur très élevée d'un tel R^2 dénote une forte et fâcheuse colinéarité entre les variables concernées).

Solutions

Le problème étant davantage lié aux données qu'au modèle, il n'y a pas de solution économétrique vraiment satisfaisante. Résultant d'un conflit entre la taille des données et celle du modèle, il accepte deux types de solutions.

On peut tenter de réduire le modèle, en éliminant une variable mineure qu'on avait retenu dans le but de n'omettre aucun facteur explicatif, si du moins elle peut être retirée sans altérer l'esprit du modèle.

Dans les modèles *à **retards échelonnés***, dont les variables ***décalées*** sont souvent très corrélées, on peut imposer une structure aux retards (méthode d'Almon).

On peut tenter d'élargir l'ensemble des données, en utilisant d'autres observations, éventuellement pour estimer séparément certains coefficients du modèle.

Lorsque la colinéarité est due à la présence d'un *trend* commun, on peut tenter d'éliminer cette tendance commune en passant au modèle sur les accroissements (ou *différences*).

Exemple: le modèle:

$$Y = a + b.X + c.Z + \varepsilon$$

devient par passage aux différences:

$$(Y_t - Y_{t-1}) = b.(X_t - X_{t-1}) + c.(Z_t - Z_{t-1}) + (\varepsilon_t - \varepsilon_{t-1})$$

plus simplement noté:

$$\Delta Y = b.\Delta X + c.\Delta Z + \Delta\varepsilon$$

On note que la constante a disparu des variables explicatives. D'autre part, la forme de l'aléa est modifiée, le nouvel aléa: $\Delta\varepsilon$, ne vérifiant a priori pas les mêmes condition que l'aléa initial: ε.

Remarque

En dépit de ce qui précède, la régression est en fait une technique relativement robuste qui distingue généralement bien les variables, il ne faut pas voir systématiquement la multicolinéarité derrière de mauvais résultats, une explication plus prosaïque et plus fréquente est simplement que le modèle retenu était mauvais…

DUMMY VARIABLES (VARIABLES INDICATRICES)

Une ***dummy variable*** est une variable ***indicatrice***, ou encore ***logique***, c'est à dire prenant les valeurs 0 ou 1 (pour indiquer que l'observation présente une certaine caractéristique, appartient à une certaine sous-période, etc.) ; l'usage a consacré l'emploi de la désignation anglaise, y compris le plus souvent dans les textes en français.

Bien que de telles variables n'aient pas le même caractère quantitatif qu'une variable traditionnelle, tels le PNB ou le revenu imposable, rien n'interdit d'en faire usage dans la régression, dans le but d'affiner un modèle en prenant ainsi en compte des caractéristiques particulières d'une fraction des observations.

Exemple : un modèle microéconomique veut expliquer la demande de bière chez les adultes de moins de soixante ans par le revenu et l'âge ; on estime l'équation :

$$BIERE = a + b.REV + c.AGE + \varepsilon$$

Cela donne, compte tenu des unités choisies et de l'échantillon utilisé :

Variable	DF	Coeff.	Std err.	t Value	Pr>|t|
Intercept	1	342.88483	72.34342	4.74	<.0001
REV	1	0.02859	0.00724	3.95	0.0003
AGE	1	-7.57556	2.31699	-3.27	0.0023

Pensant que ces facteurs ont peut-être quelque influence, on ajoute les dummy variables SEX, égale à 1 pour les femmes, et UNIV, indiquant le statut étudiant ; on obtient :

Variable	DF	Coeff.	Std err.	t Value	Pr>|t|
Intercept	1	459.48391	51.95339	8.84	<.0001
REV	1	0.02840	0.00493	5.76	<.0001
AGE	1	-7.81976	1.55801	-5.02	<.0001
SEX	1	-186.24968	28.07179	-6.63	<.0001
UNIV	1	-71.04265	54.58473	-1.30	0.2016

En fait, l'emploi de dummy variables revient à autoriser la constante à varier selon l'appartenance des observations à différentes catégories ; il modifie néanmoins généralement les estimations de tous les coefficients.

Il est également loisible de créer de nouvelles variables par combinaison de variables initiales et de dummy variables, comme il a été signalé lors des tests d'hypothèse linéaire. Ainsi, après avoir renoncé à la variable UNIV d'apparence peu significative, on dédouble la variable REV en GREV et FREV pour distinguer l'influence du revenu sur l'achat de bière selon le sexe ; on trouve :

Variable	DF	Coeff.	Std err.	t Value	Pr>|t|
Intercept	1	422.63935	56.19019	7.52	<.0001
GREV	1	0.03986	0.00849	4.69	<.0001
FREV	1	0.02290	0.00535	4.28	0.0001
AGE	1	-8.13598	1.52280	-5.34	<.0001
SEX	1	-126.84390	44.47949	-2.85	0.0072

Bien que les coefficients de FREV et GREV soient visiblement différents, on pourrait tester la significativité de cette différence par le test de Fisher approprié.

Remarques

Il convient, dans l'emploi de dummy variables ou de variables dérivées, d'être attentif à ne pas créer de multicolinéarités. Ce serait le cas dans l'exemple précédent si l'on voulait employer simultanément les variables REV, FREV et GREV, les deux dernières ayant évidemment pour somme la première.

Il est par ailleurs d'un intérêt modeste de singulariser une observation unique par une dummy variable, cela revient en effet au même que de l'ôter de l'ensemble des observations utilisées pour l'estimation et de lui permettre une autre valeur pour la constante.

HYPOTHÈSES DES MCO (RAPPEL)

Hypothèses

Soit une équation économétrique expliquant une variable: Y, comme combinaison linéaire de variables (X, Z,... et généralement la constante) pour lesquelles on dispose de séries d'observations, et d'une perturbation aléatoire : ε , non observable :

$$Y = a + b.X + ... + c.Z + \varepsilon$$

On rappelle les hypothèses des moindres carrés (mco) :

- h_0: l'équation, c'est à dire le modèle, est correcte
- h_1: les aléas sont centrés (i.e. d'espérance nulle: $E(\varepsilon_i) = 0$)
- h_2: les aléas ont même dispersion ($var(\varepsilon_i) = \sigma^2$)
- h_3: les aléas sont indépendants
- h_4: les aléas sont normaux (ou gaussiens, et de loi: $N(0,\sigma)$ du fait de h_1 et h_2)

L'estimation numérique des coefficients du modèle demande en outre que le nombre: N, d'observations dépasse, si possible largement, le nombre de variables explicatives.

Propriétés

- Sous les hypothèses: h_0 et h_1, les estimations des mco des coefficients sont sans biais.

- Sous les hypothèses: h_0, h_1, h_2 et h_3, les estimations des mco des coefficients sont optimales au sens déjà indiqué (ie efficaces parmi les estimateurs sans biais linéaires par rapport à la variable à expliquer), et la somme des carrés des résidus : SCR, permet d'estimer la variance et l'écart-type de l'aléa et des estimations des coefficients.

- Sous les hypothèses: h_0, h_1, h_2, h_3 et h_4, les estimations des mco des coefficients suivent des lois normales, et permettent les tests de significativité de Student, ainsi que les tests de Fisher d'hypothèses linéaires.

Une, ou plusieurs, de ces hypothèses peuvent être mises en défaut, l'économètre doit pouvoir déceler ces situations, évaluer leur incidence sur la régression par les mco, et si nécessaire mettre en œuvre des méthodes alternatives d'estimation.

Remarque

La présentation élémentaire suppose les variables explicatives non aléatoires; une formulation plus réaliste des mco autorise cette possibilité, *à la condition impérative* qu'elles soient indépendantes de la perturbation aléatoire (leur incidence sur la variable à expliquer passant uniquement par la partie fonctionnelle explicite de l'équation retenue). Les estimations des mco sont alors simplement conditionnées par les valeurs observées des explicatives, et les propriétés énoncées sont conservées.

ERREUR DE SPÉCIFICATION

Cas général

Il y a une erreur dans la ***spécification*** d'un modèle, quand l'hypothèse h_0 est mise en défaut, c'est à dire quand l'équation proposée n'est pas "la bonne", ainsi, dans le cas d'un modèle linéaire, quand l'ensemble des variables explicatives retenues n'est pas "le bon".

Strictement, compte tenu de leur caractère simplificateur les modèles économétriques sont toujours mal spécifiés; aussi est-il important d'étudier ce type d'erreur. On expose deux cas simples.

Cas d'une variable explicative excédentaire

On suppose que le modèle véritable est:

$$Y = a + b.X + c.Z + \varepsilon$$

et que l'on estime à tort le modèle:

$$Y = a + b.X + c.Z + d.W + \varepsilon$$

incluant la variable explicative superflue: W.

La régression par les mco donne l'équation estimée:

$$Y = A + B.X + C.Z + D.W$$

En fait le modèle estimé n'est pas faux, mais le coefficient: d, inconnu est nul. Les estimations: A, B et C de a, b et c sont sans biais mais elles sont moins précises que celles du modèle vrai, elles ne sont plus efficaces, l'estimation étant "brouillée" par la présence de variables inutiles.

La présence de variables explicatives superflues est donc un moindre mal, les estimations des coefficients restant sans biais.

Cas d'une variable omise

On prend l'exemple micro-économique d'une fonction de consommation des ménages pour une période donnée, qu'on suppose de la forme:

$$C = a.R + b.T + \varepsilon \qquad (\text{avec } a > 0 \text{ et } b > 0)$$

où R représente le revenu du ménage, et T le nombre de personnes qui le composent, ε étant l'aléa.

Négligeant l'effet de taille, on estime à tort le modèle:

$$C = a.R + \varepsilon$$

L'estimation: A, du coefficient: a, du modèle initial est biaisée. On peut toutefois préciser le sens du biais selon le signe de la corrélation entre R et T.

Si R et T sont corrélés positivement, l'estimation A surestime l'influence du revenu (R "se substituant" pour partie à T, absent du modèle estimé...).

Si R et T sont corrélés négativement (ménages dans une même phase du cycle de vie), A sous-estime l'influence du revenu.

En général l'omission de variables explicatives introduit des biais dans les estimations des coefficients des variables restantes, malheureusement, il est rarement possible de prévoir le sens des biais, comme c'était le cas dans l'exemple didactique précédent.

HÉTÉROSCÉDASTICITÉ DES ALÉAS

On met à présent en cause l'hypothèse h_2, de même variance (ou ***homoscédasticité***) des aléas.

Un exemple classique

Soit une relation économétrique valable au niveau individuel (par exemple une équation de consommation):

$$Y = a + b.X + \varepsilon$$

où l'aléa: ε, est supposé vérifier les hypothèses des mco, en particulier h_2, σ^2 notant la variance commune. Si on agrège k observations, les totaux: X_T et Y_T, pour X et Y vérifient:

$$Y_T = k.a + b.X_T + \varepsilon_T$$

où l'aléa: ε_T, est de variance $k.\sigma^2$.

Les moyennes: X_m et Y_m, quant à elles, vérifient:

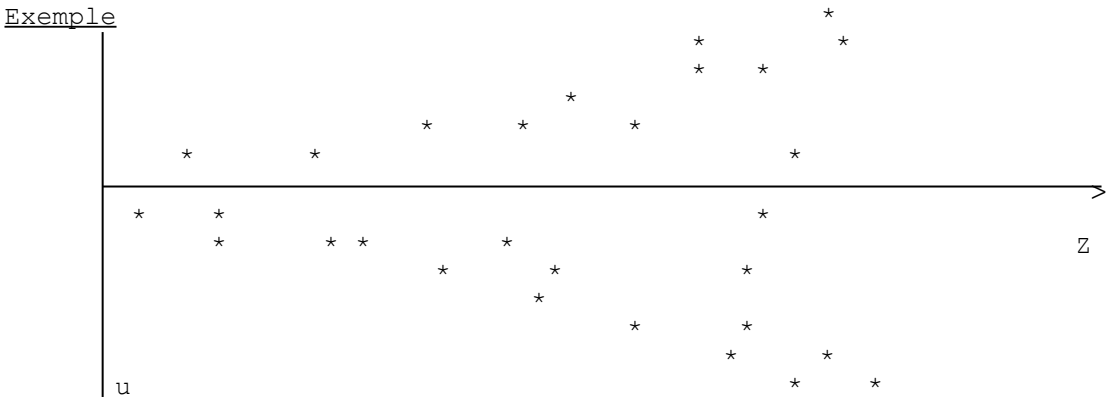
$$Y_m = a + b.X_m + \varepsilon_m$$

où l'aléa: ε_m, est de variance σ^2/k.

Qu'on emploie des totaux ou des moyennes, on voit que les aléas du modèle agrégé sur des groupes d'effectifs différents sont ***hétéroscédastiques***.

Les estimations par les mco d'un modèle satisfaisant les hypothèses: h_0 et h_1, mais dont l'aléa est hétéroscédastique, restent sans biais, mais elles ne sont plus efficaces, l'estimation de l'écart-type des aléas n'a plus de sens, celles des écarts-types des coefficients estimés ne sont plus correctes, ni par suite les tests qui en dérivent. Sous des hypothèses usuelles concernant les données, les estimations sont convergentes.

Diagnostic



L'hétéroscédasticité, qui est une variation de l'échelle des aléas s'observe sur les résidus de la régression par les mco.

L'examen de graphiques croisant les résidus avec une explicative associée à la taille, ou avec le temps pour des séries temporelles, permet de repérer les structures "en trompette", caractéristiques d'un effet de taille, où d'une inflation de l'échelle.

Différents tests quantitatifs ont été proposés pour tester la présence d'une éventuelle hétéroscédasticité des aléas.

Test de ***Goldfeld et Quandt***

On opère comme suit : on classe l'ensemble des observations par valeurs croissante d'une variable supposée représentative de l'effet de taille, on partage par moitié l'échantillon (certains éliminent la partie centrale), puis on procède à la régression par les mco sur les deux groupes séparément.

Sous les hypothèses de normalité, et dans l'hypothèse H_0 d'homoscédasticité la quantité :

$$\frac{SCR_2 / (N_2-k)}{SCR_1 / (N_1-k)}$$ suit une loi de Fisher $F(N_2-k,N_1-k)$

(avec des notations évidentes, les groupes 1 et 2 désignant respectivement les petites et les grandes tailles).

Exemple: la fonction micro-économique de dépense de logement des ménages selon le revenu a été estimée pour deux groupes, le premier gagnant de 5000$ à 10000$ et le second de 15000$ à 20000$. On a obtenu les régressions :

$C = 0,600 + 0,276.R$ $\quad N_1 = 10$ $\quad R^2 = 0,94$ $\quad SCR_1 = 0,3000$
(3,1) (11,3)

$C = 1,54 + 0,200.R$ $\quad N_2 = 10$ $\quad R^2 = 0,55$ $\quad SCR_2 = 2,024$
(1,4) (3,1)

on calcule donc (2,024/8) / (0,300/8) = 6,7 qui dépasse le seuil le 3,44 au risque 5% lu dans la table voulue, ce qui conduit à rejeter l'homoscédasticité du modèle.

Les résultats, comme la théorie, peuvent inciter également à douter de la stabilité des paramètres et à mettre en œuvre un test de Chow...

Test de ***White***

Les carrés des résidus d'une première régression par les mco sont régressés sur l'ensemble des explicatives, de leurs carrés et de leur produits croisés. White a montré que sous l'hypothèse d'homoscédasticité, la quantité $W = N.R^2$ suit asymptotiquement une loi de chi-deux à N-k+1 degrés de liberté (N désignant le nombre d'observations et k le nombre d'explicatives de cette régression auxiliaire) ; cela permet le test : si la quantité W calculée dépasse le seuil donné par la table du chi-deux au niveau de risque choisi, on rejette l'homoscédasticité.

Ces deux tests sont disponibles dans les logiciels économétriques usuels et donnent directement la quantité calculée et le résultat sans qu'il soit nécessaire de consulter la table correspondante.

<u>Solutions</u>

Si on a une idée de la forme de l'hétéroscédasticité , la méthode consiste à corriger les séries par un déflateur convenable ramenant à un aléa homoscédastique. Cette méthode qui revient à donner un « poids » différent à chaque observation est souvent appelée ***moindres carrés pondérés***.

<u>Exemple</u>: Soit le modèle:

$$Y = a + b.X + c.T + \varepsilon$$

on suppose la série des variances des aléas: ε_i, proportionnelles à la série: l_i.

Le modèle transformé par division par les $\sqrt{l_i}$ est:

$$\frac{y_i}{\sqrt{l_i}} = a.\frac{1}{\sqrt{l_i}} + b.\frac{x_i}{\sqrt{l_i}} + c.\frac{t_i}{\sqrt{l_i}} + \frac{\varepsilon_i}{\sqrt{l_i}}$$

de variables: $Y/\sqrt{L}$, $1/\sqrt{L}$, $X/\sqrt{L}$ et $T/\sqrt{L}$, et d'aléa: $\varepsilon/\sqrt{L}$, vérifie l'hypothèse h_2.

On remarque que le modèle transformé ne contient plus la constante comme explicative, il devrait donc être estimé sans celle-ci si la formulation initiale était correcte; on préférera souvent garder la constante dans le modèle transformé.

En fait et sauf exception, on a rarement une idée aussi précise que précédemment de la forme exacte de l'hétérogénéité d'échelle responsable de cette hétéroscédasticité de l'aléa, mais en cas de données

de cette nature, on améliore généralement la situation en éliminant le facteur *taille* en passant à des variables *per capita* ou équivalentes.

White et d'autres ont proposé des méthodes fondées sur une estimation préalable et asymptotiquement convergente de l'hétéroscédasticité par les mco, certaines reprennent les estimations des coefficients tandis que d'autres corrigent seulement les écart-type estimés (pour rendre valides les tests de Student), elles sont disponibles comme options des mco dans les logiciels économétriques usuels sous des intitulés tels que *correction de l'hétéroscédasticité*.

L'hétéroscédasticité enfin est parfois la conséquence de l'emploi de séries non déflatées sur une période longue. Elle peut aussi être la conséquence apparente d'une erreur de spécification, et disparaître lorsque celle-ci est corrigée (par exemple par passage aux logarithmes).

AUTOCORRÉLATION DES ALÉAS

On met maintenant en doute l'hypothèse h_3, d'indépendance des aléas. **Les méthodes exposées concernent exclusivement les modèles et séries temporels**.

Les aléas affectant un modèle économétrique ne sont pas le résultat d'un tirage aléatoire, mais l'effet de variables secondaires, non explicitement prises en compte par le modèle. On conçoit que celles-ci, du fait de leurs propres évolutions, leur donnent une "mémoire", et qu'ils risquent de ne pas être réellement indépendants d'une période à la suivante.

Autocorrélation

Une forme simple de liaison entre deux quantités aléatoires est la corrélation linéaire. S'agissant de la corrélation entre une série (les aléas), et la série *retardée*, on parle d'***autocorrélation***. C'est cette forme de dépendance qui est testée, et qui justifiera les solutions proposées.

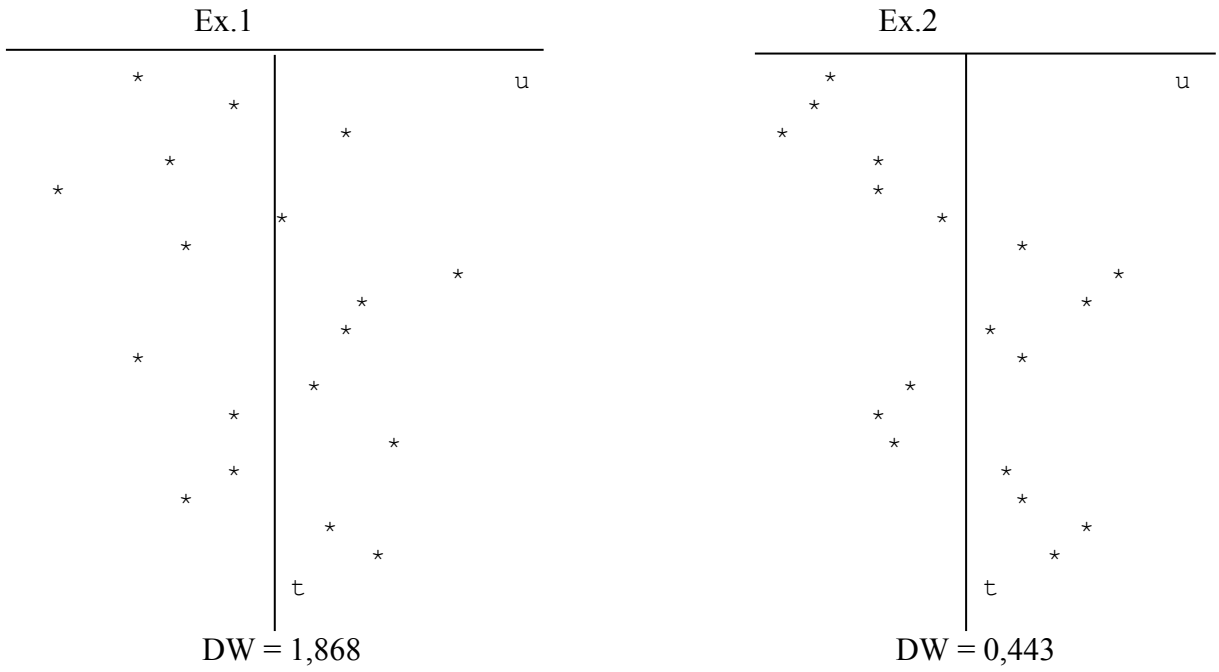
Une formalisation simple, et fréquemment employée, prête à l'aléa une comportement *autorégressif du premier ordre* :

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

où w_t est un aléa indépendant homoscédastique, et ρ le coefficient d'autocorrélation de ε.

Les estimations par les mco d'un modèle satisfaisant les hypothèses h_0, h_1 et h_2, mais dont l'aléa est autocorrelé restent sans biais, mais elles ne sont plus efficaces et les estimations des écarts-types ne sont plus correctes, comme les tests qui en découlent. Sous des hypothèses usuelles concernant les données, les estimations sont convergentes.

Diagnostic



L'examen du diagramme des résidus de la régression donne une première idée. Des aléas indépendants se traduisent par des résidus se succédant "au hasard" (Ex.1), tandis que des structures en séquences de signes opposés indiquent souvent une autocorrélation positive - la plus fréquente (Ex.2).

Test de Durbin-Watson

Un test a été proposé par ***Durbin* et *Watson***, à partir de la quantité:

$$DW = \frac{\Sigma (u_t - u_{t-1})^2}{\Sigma u_t^2}$$

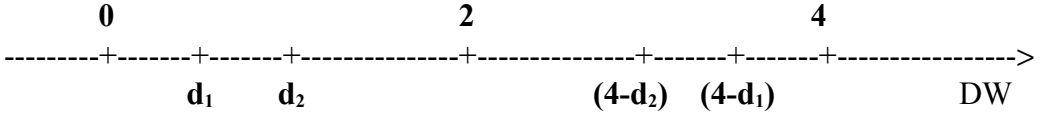
Cette quantité est calculée par les logiciels économétriques, elle est comprise entre 0 et 4.

DW est relié au coefficient d'autocorrélation: ρ, estimé sur les résidus par la relation:

$$\rho \approx 1 - DW/2$$

Proche de 0, DW indique une autocorrélation positive, et proche de 4, une autocorrélation négative (plus rare, mais qui peut provenir d'un prétraitement des variables).

Le test de Durbin-Watson de l'autocorrélation des aléas revient à examiner si la valeur DW calculée s'écarte suffisamment vers 0 ou vers 4, pour renoncer à l'hypothèse d'indépendance des aléas au niveau de risque choisi.



Exactement, on lit dans la table de Durbin-Watson, pour le nombre de variables voulu et le niveau de risque retenu (e.g. 5%), les seuils d_1 et d_2.

- Si $DW < d_1$ on rejette l'indépendance et on conclut à l'autocorrélation (positive) des aléas.
- Si $d_1 < DW < d_2$ on est dans une zone d'indétermination.
- Si $d_2 < DW$ on ne rejette pas l'indépendance des aléas.

La présence d'une zone d'indétermination provient du fait que le test dépend aussi des données, aussi Durbin et Watson ont-ils tabulé le seuil le plus sévère et le moins sévère.

Le test s'opère de même, avec les seuils: $4-d_1$ et $4-d_2$, pour une autocorrélation négative.

Ce test de Durbin-Watson n'est établi que pour un modèle avec constante. Il est par ailleurs invalide lorsque apparaissent des valeurs retardées de la variable à expliquer comme explicatives (ce qui met même à mal l'absence de biais dans les mco !), il est alors trop "laxiste" et risque de ne pas révéler l'autocorrélation. D'autres tests, tel le test de Durbin doivent lui être alors préférés.

Test de Durbin

En cas de présence de l'endogène retardée parmi les explicatives, on calcule la quantité:

$$h = \rho.\{N / [1 - N.var(b)]\}^{1/2}$$

où ρ et var (b) désignent respectivement les estimations du coefficient d'autocorrélation des aléas et de la variance du coefficient estimée de l'endogène retardée; cette quantité, dite *de Durbin*, suit asymptotiquement une loi normale: N(0,1), en l'absence d'autocorrélation, ce qu'elle permet donc de tester.

Solutions

En présence d'une autocorrélation positive des aléas, on peut essayer d'estimer le modèle sur les différences. L'autocorrélation devient en général négative, mais on voit surtout si les estimations ont beaucoup varié.

Une méthode plus élaborée consiste à estimer le coefficient d'autocorrélation: ρ, par exemple à partir de DW par la relation indiquée plus haut, puis à estimer le modèle sur les ***semi-différences***:

$$Y_t - \rho.Y_{t-1} = a.(1-\rho) + b.(X_t - \rho.X_{t-1}) + ... + (\varepsilon_t - \rho.\varepsilon_{t-1})$$

dont les aléas approchent en principe l'indépendance.

Diverses autres méthodes ***autorégressives*** sont proposées Parmi celles-ci la méthode **de *Cochrane-Orcutt*** itère le procédé précédent jusqu'à stabilisation des estimations du coefficient d'autocorrélation, tandis que la méthode du maximum de vraisemblance ou les moindres carrés non linéaires (voir chapitre suivant), mathématiquement et numériquement plus élaborés, postulant une formalisation de l'aléa telle celle autorégressive du premier ordre indiquée au début estiment simultanément l'autocorrélation et les coefficients du modèle.

Prenant mieux en compte la structure véritable des aléas, les estimations résultant des méthodes précédentes sont a priori meilleures que celles des mco.

Il faut ici encore prendre garde au fait que certaines des méthodes précédentes modifient a priori la constante.

Exemples : on reprend la fonction de consommation du modèle de Klein.

L'estimation par les mco donnait :

$$C = 0{,}1929.P + 0{,}0899.P_{-1} + 0{,}7962.W + 16{,}237 \quad R^2 = 0{,}9810$$
$$(2{,}12) \quad (0{,}099) \quad (19{,}9) \quad (12{,}5) \quad DW = 1{,}367$$

L'estimation sur les différences (sans la constante) donne :

$$DC = 0{,}4174.DP + 0{,}1631.DP_{-1} + 0{,}4978.DW \quad R^2 = 0{,}9059$$
$$(3{,}44) \quad (1{,}55) \quad (4{,}22) \quad DW = 1{,}9561$$

L'estimation sur les semi-différences, avec un ρ estimé de 0,3165, donne :

$$C = 0{,}2256.P + 0{,}0748.P_{-1} + 0{,}7485.W + 12{,}263 \quad R^2 = 0{,}9638$$
$$(2{,}27) \quad (0{,}77) \quad (13{,}3) \quad (9{,}68) \quad DW = 1{,}544$$

L'estimation par la méthode de Cochrane-Orcutt donne :

$$C = 0{,}4637.P + 0{,}2080.P_{-1} + 0{,}4330.W + 0{,}7856 \quad R^2 = 0{,}87$$
$$(3{,}11) \quad (1{,}62) \quad (2{,}67) \quad (2{,}26) \quad DW = 1{,}42$$

Et celle par la méthode du maximum de vraisemblance avec aléa autorégressif du premier ordre :

$$C = 0{,}4282.P + 0{,}1715.P_{-1} + 0{,}4645.W + 27{,}111 \quad R^2 = 0{,}9824$$
$$(3{,}62) \quad (1{,}66) \quad (3{,}94) \quad (6{,}26) \quad DW = 2{,}043 \text{ avec } \rho = 0{,}8869$$

Ces différentes régressions modifient notablement les résultats de la première, effectuée par les mco sur le modèle initial. A priori, la dernière, plus perfectionnée dans son principe, est sans doute celle à retenir.

Il est enfin fréquent qu'une apparente autocorrélation des aléa provienne d'une erreur de spécification et disparaisse avec l'introduction d'une variable oubliée ou la correction de la forme fonctionnelle retenue (par exemple après passage en logarithme).

STEPWISE REGRESSION

Il existe un certain nombre de procédures "automatiques" qui ont l'ambition de trouver "la meilleure" régression expliquant une variable parmi un ensemble de variables explicatives éventuelles. Elles procèdent en général par sélection *pas-à-pas* des variables explicatives avec une règle d'arrêt.

L'une des plus astucieuses, connue sous le nom anglais de ***stepwise regression***, est disponible sur les logiciels d'économétrie.

Le principe est le suivant : disposant d'observations de la variable à expliquer et d'un ensemble d'explicatives éventuelles, on impose tout d'abord la constante et l'on travaille désormais sur les variables centrées. On retient ensuite la variable la plus corrélée avec la variable à expliquer et on opère les mco; on introduit ensuite si nécessaire la variable la plus corrélée avec le résidu de la régression précédente (ce résidu représentant la part non encore *expliquée*), et ainsi de suite... À chaque étape, on examine d'autre part s'il convient d'abandonner certaines des variables précédemment retenues et devenues inutiles. Ces introductions et rejets se font grâce à des tests de significativité de Student à des niveaux fixés par l'utilisateur. L'algorithme s'arrête rapidement sur "la meilleure" régression au sens de cette méthode...

L'usage de telles procédures, que nous ne recommandons pas, ne saurait dépasser la phase exploratoire d'une étude économétrique. De simples règles "mécaniques" ne peuvent en effet encore remplacer la compréhension économique des phénomènes que l'on cherche à modéliser.

PRÉSENTATION D'UN ENSEMBLE DE RÉGRESSIONS

Il est commode de rendre compte des différents modèles estimés dans une étude économétrique par des tableaux rectangulaire de la forme suivante, dans laquelle les variables explicatives envisagées sont en colonne et les relations en ligne (certains auteurs adoptant le choix contraire).

Équation	SAL	TINT	IR	t	Const.	R^2	DW	SCR
I	1,062				0,656	0,989	0,389	56694
(mco)	(40,8)				(4,69)			
II	0,693		3,736		1,047	0,993	1,006	35655
(mco)	(5,87)		(3,17)		(6,23)			
III	1,324		-0,768		-0,486	0,832	1,216	
(aut.)	(7,42)		(-0,56)					
IV	0,733	-0,00140	3,653		1,005	0,993	0,903	34709
(mco)	(5,43)	(-0,07)	(3,03)		(5,52)			
V	1,268	-0,0347	-0,527		0,0254	0,812	1,172	
(aut.)	(28,2)	(-2,28)	(-0,43)					
VI	0,532			0,1126	2,262	0,999	1,806	6074
(mco)	(11,8)			(11,9)	(15,8)			
VII	0,600	-0,01762	-0,690	0,1212	3,142	0,999	1,892	4349
(mco)	(11,8)	(-2,27)	(-1,13)	(6,45)	(11,2)			

Il s'agit ici d'expliquer la consommation des ménages en France en fonction des salaires versés: SAL, du taux d'intérêt: TINT, de l'impôt sur le revenu: IR, du temps: t, et de la constante.

Les différentes équations présentées ont été estimées soit par les mco soit par la méthode autorégressive de Cochrane-Orcutt, sur les données annuelles déflatées de 1966 à 1985.

APPENDICE MATHÉMATIQUE

Moindres carrés généralisés

Les cas précédemment exposés du modèle à aléas hétéroscédastiques et du modèle à aléa autorégressif peuvent être présentés dans un cadre commun, plus général que celui des mco, celui des ***moindres carrés généralisés*** (ou ***mcg***, et *gls* pour les anglo-saxons).

Soit le modèle linéaire écrit sous forme matricielle :

$$Y = X.a + \varepsilon$$

avec les mêmes notations que dans le chapitre précédent, où a est le vecteur à k composantes des coefficients, et ε le vecteur des aléas, de dimension N.

L'aléa ε est toujours centré, mais de matrice de variance-covariance : Ω, de dimension N_xN, supposée inversible.

On peut trouver une matrice carrée N_xN inversible : H, telle que le vecteur aléatoire $H.\varepsilon$ ait une matrice de variance-covariance *scalaire* :

$$V(H.\varepsilon) = H.\Omega.H' = \sigma^2.I$$

c'est à dire telle que :

$$\sigma^2.H^{-1}.H'^{-1} = \Omega$$

et le modèle transformé :

$$(H.Y) = (H.X).a + (H.\varepsilon)$$

satisfait les hypothèses des mco et admet le même vecteur : a, de coefficients que le modèle initial; son estimateur des mco étant :

$$A = (X'.\Omega^{-1}.X)^{-1}.X'.\Omega^{-1}.Y$$

c'est là l'*estimateur des moindres carrés généralisés* du modèle de départ, de matrice de variance-covariance :

$$V(A) = (X'.\Omega^{-1}.X)^{-1}$$

Dans le cas du modèle à aléas hétéroscédastiques indépendants, la matrice Ω est diagonale mais non scalaire, et le modèle sur donnée "déflatées" en est le modèle transformé; dans le cas du modèle à aléa autorégressif, c'est le modèle aux semi-différences, à la première ligne près, et en supposant l'autocorrélation ρ connue exactement.

Le problème est qu'en général, la matrice Ω et même son type sont inconnus...

----===**O**===-----

(28.05.2009)