CHAPITRE 3-2
L'INDUCTION STATISTIQUE APPLIQUÉE À LA RÉGRESSION MULTIPLE

Plan

CHAPITRE 3-2
L'INDUCTION STATISTIQUE APPLIQUÉE À LA RÉGRESSION MULTIPLE

Jusqu'à présent, nous avons vu comment estimer les paramètres d'un modèle théorique formalisé par une relation linéaire. La méthode d'estimation retenue, celle des moindres carrés, consiste à choisir la valeur des paramètres de façon à minimiser les erreurs de prédiction que l'on fait lorsqu'on applique le modèle aux observations qui ont servi à l'estimer.

Si l'on s'en tenait à cela, l'analyse de régression ne serait rien de plus qu'une manière de résumer les relations que l'on observe dans les données entre les variables : l'analyse de régression serait alors une technique de statistique descriptive (c'est d'ailleurs l'une des utilisations légitimes de l'analyse de régression). En général cependant, l'analyse des données a pour but de découvrir la relation sous-jacente dont les paramètres sont inconnus et qui lie entre elles les variables.

Rappelons le contexte dans lequel nous avons situé l'analyse de régression. Il est admis d'emblée que le modèle (déterministe) dont on cherche à estimer les paramètres n'est qu'une approximation de la réalité : il subsiste un écart entre les prédictions du modèle et les observations. Cette « erreur » non systématique est représentée par le terme aléatoire de la relation. L'objectif poursuivi est de connaître la valeur des paramètres (coefficients) de la relation. Mais ces paramètres ne peuvent pas s'observer directement. Ils se révèlent indirectement, à travers les valeurs observées des variables que lie la relation. Le problème, c'est que dans les observations, la relation est en quelque sorte brouillée par l'intervention du terme aléatoire, dont la valeur est tout aussi inobservable que les paramètres de la relation.

Ainsi, seules les méthodes de l'induction statistique peuvent permettre de baliser l'incertitude qui entoure l'estimation des paramètres. Et l'application de ces méthodes exige que l'on complète le modèle aléatoire en associant au modèle déterministe un modèle d'échantillonnage, qui sera un modèle du lien aléatoire entre l'échantillon et la population. Qu'est-ce qui constitue la population, et qu'est-ce qui constitue l'échantillon dans la régression multiple ? Dans ce contexte, on considère que la valeur (inobservable) du terme aléatoire associé à chaque observation est un échantillon de taille 1, tiré de l'ensemble infini des valeurs possibles que pourrait prendre le terme aléatoire dans ce cas ; cet ensemble infini de valeurs possibles constitue une population. Avec n observations, il y a donc n échantillons de taille 1, tirés de n populations. Il n'est pas exclu au départ que ces n populations soient différentes entre

elles ; il n'est pas exclu non plus qu'elles soient identiques (ce qui équivaut à dire que les valeurs des termes aléatoires sont tirées d'une même population). Les *n* populations sont-elles supposées identiques ? La réponse à cette question fait partie du modèle d'échantillonnage. Plus largement, le modèle d'échantillonnage est constitué d'hypothèses sur les distributions de probabilité des termes aléatoires. Si l'on fait l'hypothèse que ces distributions de probabilité sont identiques, cela équivaut à faire l'hypothèse que les populations sont identiques.

Nous reviendrons plus loin sur les hypothèses qui constituent le modèle d'échantillonnage de la régression linéaire classique. Pour aborder les procédures d'induction statistique, il suffit provisoirement de savoir que, dans le cadre de ce modèle d'échantillonnage, on peut estimer les paramètres des distributions d'échantillonnage des estimateurs des paramètres :

1. L'estimateur b_j du paramètre β_j est *non biaisé*. En d'autres mots, l'estimateur des moindres carrés a une distribution d'échantillonnage dont la moyenne est égale à la valeur du paramètre.
2. Il existe aussi un estimateur non biaisé de la variance d'échantillonnage $\sigma_{b_j}^2$ de chacun des coefficients estimés b_j, et de la covariance d'échantillonnage de chaque paire de coefficients estimés, $\sigma_{b_j b_h}$.

Notation :

$s_{b_j}^2 =$ valeur estimée de la variance d'échantillonnage $\sigma_{b_j}^2$.

Ces valeurs estimées sont fournies par les logiciels d'application statistique.

3-2.1 Quelques exemples de tests d'hypothèse

3-2.1.1 Test bilatéral d'une hypothèse simple sur la valeur d'un coefficient (test de Student)

Nous voulons tester une hypothèse simple du type

$H_0 : \beta_j = c$

Par exemple, Lemelin et Polèse (1995) ont estimé les paramètres du modèle suivant :

$$\ln PURB = \beta_1 + \beta_2 \ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$$

Ce modèle équivaut à

$$\ln\left(\frac{PURB}{PTOT}\right) = \ln PURB - \ln PTOT = \beta_1 + (\beta_2 - 1)\ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$$

On veut tester l'hypothèse que le coefficient β_2 est égal à un :

$H_0 : \beta_2 = 1$

Pourquoi cette hypothèse ? Parce que, si elle est vraie, cela signifie que le degré d'urbanisation $\frac{PURB}{PTOT}$ est indépendant de la population totale : en d'autres mots, peu importe la taille de la population d'un pays, la fraction de la population qui vit en milieu urbain est déterminée par le PIB per capita. Concrètement, si $\beta_2 = 1$, le modèle prédit que le degré d'urbanisation de la Chine, avec son 1,1 milliard d'habitants en 1990, est le même que celui du Kenya, qui en compte 24 millions, parce que les deux pays ont un PIB per capita de 370 $ U.S. (voir le tableau 1 de Lemelin et Polèse, 1995). Peut-on rejeter cette hypothèse ?

Rappelons les étapes à suivre pour faire un test de probabilité critique (sans seuil de signification prédéfini – *p-value test*) :

1. choisir une variable-test ;
2. vérifier que le modèle d'échantillonnage associé à cette variable-test est acceptable ;
3. calculer la valeur de la variable-test ;
4. déterminer la valeur de la probabilité critique correspondante ;
5. prendre la décision de rejeter ou non l'hypothèse, selon que l'on juge cette probabilité critique suffisamment petite ou non (plus la probabilité critique est petite, moins les observations sont compatibles avec l'hypothèse, et plus le rejet peut être catégorique).

Nous allons appliquer le test de Student, qui, dans le cas d'un test d'hypothèse simple sur un coefficient de régression, utilise la variable-test suivante :

$$t_{n-k} = \frac{b_j - c}{s_{b_j}}$$

où b_j est la valeur estimée du paramètre β_j et s_{b_j} est la valeur estimée de l'écart type d'échantillonnage de b_j. Il y a une analogie évidente entre cette variable-test et celle que l'on utilise pour le test d'une hypothèse simple sur une moyenne :

$$t_{n-1} = \frac{m_x - \gamma}{\left(s_x / \sqrt{n}\right)}$$

où le dénominateur $\left(s_x / \sqrt{n} \right)$ est l'écart type d'échantillonnage de la moyenne.

Le choix de cette variable-test se justifie parce que, *sous les conditions du modèle classique de la régression linéaire normale*, la variable $\frac{b_j - \beta_j}{s_{b_j}}$ a une distribution de Student *avec* $n - k$ *degrés de liberté*, où *n* est le nombre d'observations et *k*, le nombre de variables indépendantes du modèle (y compris la constante).

Le modèle d'échantillonnage associé au test est donc le modèle classique de régression linéaire normale. Supposons pour l'instant que ce modèle soit jugé acceptable : ce point est discuté de manière plus approfondie ci-après en 3-2.2.

Dans le cas qui nous intéresse, la valeur de la variable-test est donnée par [1]

$$t_{n-k} = \frac{0{,}971663 - 1}{0{,}0279321} = -1{,}0145$$

Le nombre d'observations *n* est égal à 64 (Lemelin et Polèse, 1995, tableau 2) ; le nombre de variables indépendantes *k* est égal à 4. On a donc 60 degrés de liberté. La probabilité critique associée à cette valeur pour un test bilatéral est de 0,314 ou 31,4 % (cette probabilité critique a été calculée à l'aide de la fonction `TDIST` du logiciel Excel [2] ; Lemelin et Polèse, 1995, p. 322, rapportent les résultats d'un test équivalent).

À moins de choisir un seuil de signification très élevé (supérieur à 0,314), on ne peut pas rejeter l'hypothèse que $\beta_2 = 1$. Une hypothèse qui n'est pas rejetée n'est pas pour autant démontrée. Cela dit, si une hypothèse n'est pas rejetée, il est légitime de la maintenir.

3-2.1.2 TEST DE L'HYPOTHÈSE D'UN COEFFICIENT NUL

Il arrive souvent que l'on veuille tester l'hypothèse

$H_0 : \beta_j = 0$

Par exemple, dans le modèle

$\ln PURB = \beta_1 + \beta_2 \ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$

---

1 Les valeurs utilisées dans le calcul ci-dessous proviennent directement des sorties d'ordinateur reproduites à l'annexe 3-A. Elles ne se retrouvent pas telles quelles dans Lemelin et Polèse (1995).

2 Noter que la valeur de la statistique *t* qui entre comme argument dans la fonction `TDIST` ne peut pas être négative. L'utilisateur doit donc se servir du fait que la distribution de Student est symétrique.

la valeur estimée du paramètre β_4 rapporté par Lemelin et Polèse (1995, tableau 2) est de $-0{,}045$. Cette valeur semble bien petite : est-elle « significativement différente de zéro » ? En d'autres mots, peut-on rejeter l'hypothèse que ce coefficient soit nul et que l'on puisse laisser tomber la variable correspondante ?

Pour tester ce type d'hypothèse, il suffit d'appliquer le test décrit précédemment, avec $c = 0$. La variable-test prend alors la forme particulière

$$t_{n-k} = \frac{b_j - c}{s_{b_j}} = \frac{b_j}{s_{b_j}} \text{ quand } c = 0$$

Pour notre exemple [3],

$$t_{64-4} = \frac{-0{,}0453}{0{,}01345} = -3{,}368$$

La probabilité critique associée à cette valeur pour un test bilatéral avec 60 degrés de liberté est de 0,0013 ou 0,13 % (cette probabilité critique a été calculée à l'aide de la fonction `TDIST` du logiciel Excel). Avec une probabilité critique aussi faible, il serait difficile de ne pas rejeter l'hypothèse. On dira que le coefficient est significativement différent de zéro à un seuil de signification de moins de 1 %, plus exactement de 0,0013, ce qui est à peine plus de 0,1 %.

Il est si courant de tester ce type d'hypothèse que les logiciels d'application statistique le font automatiquement : c'est le « *t* » qui est rapporté par les logiciels d'application statistique. Les logiciels donnent aussi la valeur de la probabilité critique correspondante.

Les tableaux de résultats des articles scientifiques présentent aussi une évaluation du degré de signification de chaque coefficient. Lemelin et Polèse (1995) rapportent la probabilité critique, Richardson *et al.* (1990) donnent la valeur du *t* de Student et Heikkila *et al.* (1989) présentent les deux.

---

3 Les valeurs utilisées dans le calcul ci-dessous proviennent directement des sorties d'ordinateur reproduites à l'annexe 3-A. Elles ne se retrouvent pas telles quelles dans Lemelin et Polèse (1995).

Certains auteurs donnent l'écart type de chaque coefficient estimé (de toute façon, connaissant la valeur estimée du coefficient, on peut calculer son écart type à partir de sa statistique *t* et vice-versa, puisque $t_{n-k} = \frac{b_j}{s_{b_j}}$. Lorsqu'on donne seulement l'écart type ou la valeur du *t* de Student, on identifie généralement, à l'aide de renvois, les coefficients qui sont significativement différents de zéro à un seuil de signification de 1 %, de 5 % ou de 10 %.

3-2.1.3 TEST UNILATÉRAL D'UNE HYPOTHÈSE SIMPLE SUR LA VALEUR D'UN COEFFICIENT (TEST DE STUDENT)

Il est parfois plus pertinent d'appliquer un test unilatéral (voir le chapitre 2-3). Par exemple, le modèle

$$PLAR_i = K\, PURB_i^h$$

prédit que la plus grande ville du pays croît plus vite, ou moins vite que l'ensemble de la population urbaine, selon que la valeur de l'exposant, le paramètre *h*, est plus grande, ou plus petite que 1 (pour faciliter les références à l'article, nous gardons ici la même notation). En particulier, si $h < 1$, le modèle prédit que le poids relatif de la plus grande ville diminue à mesure que croît la population urbaine. Peut-on rejeter l'hypothèse que $h \geq 1$?

Lemelin et Polèse (1995, tableau 2) ont estimé les paramètres du modèle linéarisé

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

La valeur estimée de *h* est 0,636. On fait un test unilatéral de l'hypothèse

$$H_0 : h = 1$$

avec, comme hypothèse complémentaire,

$$H_A : h < 1$$

La zone de rejet se situe donc à gauche : pour rejeter H_0 et accepter H_A, il faut que l'écart $1 - h$ soit assez grand pour que l'on juge très improbable que $h \geq 1$. La variable-test est encore le *t* de Student [4] :

$$t_{64-2} = \frac{h-1}{s_h} = \frac{0{,}636-1}{0{,}0426} = -8{,}54$$

---

4 Les valeurs utilisées dans le calcul ci-dessous proviennent directement des sorties d'ordinateur reproduites à l'annexe 3-A. Elles ne se retrouvent pas telles quelles dans Lemelin et Polèse (1995).

Avec $n - k = 62$ degrés de liberté, la probabilité critique unilatérale associée à une valeur absolue aussi grande du *t* de Student est de moins de 0,0001 [5]. Nous pouvons donc décider en toute confiance de rejeter H_0 et d'accepter l'hypothèse que l'importance relative de la plus grande ville diminue à mesure que croît la population urbaine (de quoi ébranler quelques idées reçues...).

3-2.1.4 INTERVALLES DE CONFIANCE ET MARGES D'ERREUR

Il va de soi que la variable-test t_{n-k} permet aussi, comme pour la moyenne, de définir des intervalles de confiance de la forme

$$b_j - s_{bj}\, \theta_{n\text{-}k}(\alpha) < \beta_j < b_j + s_{bj}\, \theta_{n\text{-}k}(\alpha)$$

avec un niveau de confiance de $(1-\alpha)$. La démonstration est donnée plus loin, dans l'encadré. La marge d'erreur correspondante, avec le même niveau de confiance, est égale à

$$\pm\, s_{bj}\, \theta_{n\text{-}k}(\alpha)$$

Par exemple, calculons un intervalle de confiance du paramètre β_4 dans le modèle

$$\ln PURB = \beta_1 + \beta_2 \ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$$

Avec $n - k = 60$ degrés de liberté, à un niveau de confiance de 0,99 (99 %), les valeurs critiques du *t* de Student sont $-2{,}66$ et $+2{,}66$ ($\theta_{64}(0{,}01) = 2{,}66$; calculé à l'aide de la fonction `TINV` du logiciel Excel). Puisque $b_j = -0{,}0453$ et que $s_{bj} = 0{,}01345$, l'intervalle de confiance de β_4 à 99 % est donc donné par

$$-0{,}0453 - (0{,}01345 \times 2{,}66) < \beta_4 < -0{,}0453 + (0{,}01345 \times 2{,}66)$$

$$-0{,}0811 < \beta_4 < -0{,}0095$$

et la marge d'erreur, à un niveau de confiance de 99 %, est égale à

$$\pm\, 0{,}01345 \times 2{,}66 = \pm\, 0{,}358$$

Les intervalles de confiance se déduisent ici exactement comme dans le cas d'un test d'hypothèse simple sur la moyenne : l'ensemble des hypothèses qui ne seraient pas rejetées à un niveau de signification de α est donné par

---

5 Pour *t*=4, la fonction `TDIST` du logiciel Excel donne une probabilité critique unilatérale de 0,0000857. En deçà de cette probabilité, la fonction `TINV` commence à donner des résultats aberrants. Rien ne garantit que la fonction `TDIST` donne des résultats valides pour des valeurs de *t* supérieures à 4. Aussi vaut-il mieux, et cela suffit, constater que la probabilité critique associés à *t*=8,54 est inférieure à 0,0001.

$$-\theta_{n-k}(\alpha) < \frac{b_j - c}{s_{b_j}} < +\theta_{n-k}(\alpha)$$

$$-\theta_{n-k}(\alpha)\, s_{b_j} < (b_j - c) < +\theta_{n-k}(\alpha)\, s_{b_j}$$

$$-b_j - \theta_{n-k}(\alpha)\, s_{b_j} < -c < -b_j + \theta_{n-k}(\alpha)\, s_{b_j}$$

$$+b_j + \theta_{n-k}(\alpha)\, s_{b_j} > +c > +b_j - \theta_{n-k}(\alpha)\, s_{b_j}$$

$$b_j - \theta_{n-k}(\alpha)\, s_{b_j} < c < b_j + \theta_{n-k}(\alpha)\, s_{b_j}$$

3-2.1.5 TEST D'UNE OU DE PLUSIEURS RELATIONS LINÉAIRES ENTRE DES COEFFICIENTS (TEST *F* DE FISHER)

Le test de Student permet d'examiner une seule hypothèse à la fois, et cette hypothèse ne peut porter que sur un seul coefficient. Le test de Fisher est beaucoup plus polyvalent : il permet d'examiner plusieurs hypothèses à la fois et il permet d'examiner des hypothèses qui associent plus d'un coefficient. Comme le test de Student, celui de Fisher exige que soient respectées les conditions du modèle classique de la régression linéaire normale. Nous n'entrerons pas ici dans la mécanique du test de Fisher et nous nous contenterons de donner quelques exemples de son utilité.

Rappelons néanmoins que la variable-test de Fisher se calcule au moyen de la formule suivante :

$$F_{p,n-k} = \frac{\left(\frac{SSR_H}{p}\right)}{\left(\frac{SSR}{n-k}\right)}$$

où p est le nombre des contraintes linéaires simultanées qui constituent l'hypothèse, et SSR_H est la somme des carrés des résidus obtenue sous contrainte (c'est-à-dire lorsqu'on estime les paramètres du modèle en le forçant à respecter l'hypothèse).

Ainsi, Lemelin et Polèse (1995) ont estimé les paramètres des modèles suivants (pour faciliter les références à l'article, nous reprenons ici la même notation) :

$$\ln PLAR = p' + q' \ln PTOT + r' \ln GNPC + t' (\ln GNPC)^2 + s \ln PURB$$

$$\ln PLAR = \ln K + h \ln PURB$$

L'hypothèse à tester est qu'il est acceptable (c'est-à-dire non rejeté) de laisser tomber les variables qui sont absentes de la seconde équation. En fait, cette hypothèse est constituée de *trois* hypothèses simples :

$H_1 : q' = 0$

$H_2 : r' = 0$

$H_3 : t' = 0$

Les probabilités critiques données au tableau 2 de Lemelin et Polèse (1995) permettent de conclure que, prise séparément, aucune de ces hypothèses ne peut être rejetée : pour H_1, la probabilité critique est de 0,484 ; pour H_2, elle est de 0,189 ; et pour H_3, elle est de 0,173. Mais chacun de ces trois tests d'hypothèse repose sur un modèle d'échantillonnage où les deux autres variables sont présentes : dans le premier cas, par exemple, *étant donné que le modèle contient les variables ln GNPC et (ln GNPC)*^{2}, on ne peut pas rejeter l'hypothèse que $q' = 0$. Qu'en est-il alors de l'hypothèse que les *trois* coefficients sont nuls *simultanément* ? Voilà le type d'hypothèse que permet d'examiner le test de Fisher. Lemelin et Polèse (1995, p. 323) rapportent que l'application de ce test donne une probabilité critique de 0,53 pour l'hypothèse que *q'*, *r'*, et *t'* sont nuls simultanément : on ne peut pas rejeter cette hypothèse.

Considérons maintenant la fonction de production Cobb-Douglas :

$Y = A\, K^B\, T^C$

où

Y est la quantité produite

K est la quantité de capital utilisée

T est la quantité de main-d'oeuvre utilisée

A, *B* et *C* sont des paramètres.

Quand on applique la transformation logarithmique, le modèle devient linéaire :

$\log Y = \log A + B \log K + C \log T$

L'une des hypothèses que l'on voudra examiner dans ce modèle est la suivante :

$H_0 : B + C = 1$

L'intérêt de cette hypothèse est que, si $B + C = 1$, il s'agit d'une fonction de production à rendements constants à l'échelle. Cela veut dire que, si l'on augmente (ou diminue) tous les facteurs de production proportionnellement, alors la production augmente (ou diminue) dans la même proportion.

Cette propriété est assez simple à démontrer. Soit K_0, T_0 et Y_0 les valeurs initiales de K, T et Y, et soit λ la proportion selon laquelle on augmente les quantités de facteurs. Alors

$$Y = A\,(\lambda\,K_0)^B\,(\lambda\,T_0)^C = \lambda^{B+C}\,A\,K_0^B\,T_0^C = \lambda^{B+C}\,Y_0$$

Et si $B + C = 1$,

$$Y = \lambda\,Y_0$$

On ne peut pas appliquer le test de Student à l'hypothèse « $H_0 : B + C = 1$ », parce qu'il s'agit, non pas d'une hypothèse sur un paramètre, mais d'une hypothèse sur une relation entre deux paramètres. Mais le test de Fisher permet d'examiner ce type d'hypothèse.

Plus généralement et formellement, le test de Fisher permet de tester toute hypothèse que l'on peut exprimer par une ou plusieurs contraintes linéaires sur les coefficients. Une contrainte linéaire sur les coefficients β_1, β_2, β_3, etc. s'écrit sous la forme :

$$\sum_{j=1}^{k} w_j\,\beta_j = w_1\,\beta_1 + w_2\,\beta_2 + \cdots + w_k\,\beta_k = c$$

où c et les w_j sont des constantes définies par l'utilisateur, selon la contrainte qu'il veut représenter.

L'hypothèse la plus fréquemment testée avec le test de Fisher (mais pas nécessairement la plus intéressante) est :

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

C'est l'hypothèse que tous les coefficients de la régression, sauf la constante β_1, sont nuls : cette hypothèse est donc constituée de $(k - 1)$ hypothèses simples. Autrement dit, c'est l'hypothèse que la « vraie » valeur du coefficient de détermination multiple R^2 est zéro et que le R^2 calculé ne représente rien d'autre que la corrélation fortuite due aux termes aléatoires. Les logiciels d'application statistique fournissent automatiquement la valeur de la variable-test associée à cette hypothèse.

3-2.2 Spécification d'un modèle d'échantillonnage : les conditions du modèle classique de régression linéaire normale

Nous l'avons dit, la validité des tests d'hypothèse qui viennent d'être décrits dépend de la validité du modèle d'échantillonnage sur lequel ils reposent. Nous allons donc jeter un coup d'oeil sur ce modèle.

3-2.2.1 LE MODÈLE CLASSIQUE DE LA RÉGRESSION LINÉAIRE

Réf. : Kennedy (1992, 43-45)

Le modèle d'échantillonnage classique est constitué de quatre hypothèses. Ces hypothèses sont passablement générales en ce qu'elles imposent peu de restrictions à la forme générale de la distribution de probabilité du terme aléatoire.

Les deux premières hypothèses portent sur les paramètres des distributions de probabilité des termes aléatoires :

H1.	Pour chaque observation, la valeur du terme aléatoire est tirée d'une population théorique de moyenne nulle : en conséquence, $E(u_i) = 0$ pour tous les i.
H2a)	Pour toutes les observations, les populations théoriques d'où sont tirées les valeurs des termes aléatoires ont la même variance [6] : $\sigma_i^2 = \sigma^2$ pour tous les i.
H2b)	Pour chaque observation, la valeur du terme aléatoire est statistiquement indépendante des valeurs des termes aléatoires des autres observations [7] : $\sigma_{ij} = 0$ pour toutes les combinaisons i,j où $i \neq j$.

Nous verrons plus concrètement ce que peuvent signifier ces conditions lorsque nous examinerons ce qui se passe quand elles ne sont pas réalisées. La troisième hypothèse circonscrit le rôle de l'aléatoire dans le modèle :

H3.	Les variables indépendantes x_{ij} sont non aléatoires.

L'hypothèse H3 exige en particulier que les valeurs des variables indépendantes soient mesurées sans erreur. Parmi les autres situations incompatibles avec cette condition, mentionnons la présence de valeurs retardées de la variable dépendante du côté des variables indépendantes, comme, par exemple, dans un modèle où le taux de chômage à chaque mois dépend, entre autres, du taux de chômage du mois précédent (un modèle de la forme $C_t = a + bC_{t-1} + \cdots$). Mentionnons aussi les systèmes d'équations simultanées, où la variable dépendante d'une équation apparaît parmi les variables indépendantes d'une autre équation (par exemble, un modèle où le PIB dépend de la consommation et des autres composantes de

---

6 Cette propriété s'appelle « homoscédasticité » (des mots grecs ομος, égal, et σκεδασις, dispersion). Son contraire est l'« hétéroscédasticité » (du mot grec ετερος, autre).

7 Cette propriété s'appelle l'absence d'autocorrélation.

la demande, tandis que la consommation dépend du PIB : $Y = C + X$ et $C = a + bY$). Dans ces circonstances, les méthodes doivent être adaptées pour tenir compte du non-respect de H3.

La quatrième hypothèse enfin se rapporte aux relations entre les variables indépendantes et au nombre d'observations.

H4.	Il y a moins de paramètres à estimer qu'il y a d'observations et il n'y a pas de redondance parmi les variables indépendantes [8].

« Pas de redondance » signifie ici qu'aucune variable indépendante ne peut être représentée comme une combinaison linéaire des autres : les variables indépendantes sont *linéairement indépendantes* entre elles. H4 est moins une hypothèse qu'une condition d'application, puisque l'on peut déterminer par analyse des données si cette condition est respectée.

Les hypothèses H1 à H4, combinées au modèle linéaire général exposé à la section 1, définissent un modèle aléatoire que l'on désigne habituellement comme le « modèle classique de la régression linéaire ». La spécification de ce modèle d'échantillonnage est toutefois incomplète, puisque la forme de la distribution de probabilité des termes aléatoires n'est pas définie. Néanmoins, lorsque les hypothèses H1 à H4 sont respectées, l'estimateur des moindres carrés ordinaires a plusieurs propriétés désirables. Ces propriétés sont démontrées dans le *théorème de Gauss-Markov*.

3-2.2.2 PROPRIÉTÉS DE L'ESTIMATEUR DES MOINDRES CARRÉS DANS LE MODÈLE CLASSIQUE DE LA RÉGRESSION LINÉAIRE : LE THÉORÈME DE GAUSS-MARKOV

Nous nous contenterons ici d'énoncer, sans les démontrer, les principales conclusions du théorème de Gauss-Markov [9]. Ces conditions se rapportent notamment aux paramètres de la distribution d'échantillonnage des estimés b_j. Ce théorème établit donc les fondements des tests d'hypothèses applicables aux estimés b_j.

Lorsque les hypothèses H1 à H4 sont respectées, alors les résultats de la méthode des moindres carrés ont les propriétés suivantes :

---

8 Techniquement, cela se traduit par la condition que le rang de la matrice $\boldsymbol{X}$, d'ordre $n \times k$, soit égal à $k < n$.

9 Le mathématicien Carl Friedrich Gauss (1777-1855) est l'inventeur de la distribution normale et de la méthode des moindres carrés (1794 ; méthode appliquée pour la première fois par Gauss à l'estimation en 1801 de la trajectoire de l'astéroïde Cérès) ; Andreï Andreïevitch Markov (1856-1922) est notamment l'auteur d'un théorème limite central.

1. La méthode des moindres carrés produit des estimateurs *non biaisés* des paramètres β_j.

   En d'autres mots, chacun des estimés b_j a une distribution d'échantillonnage dont la moyenne (l'espérance mathématique) est égale à la « vraie » valeur du coefficient, β_j.

2. La méthode des moindres carrés produit aussi un estimateur non biaisé de σ^2, la variance commune des termes aléatoires. La variable aléatoire

   $$\frac{1}{n-k}\sum_i (y_i - \hat{y}_i)^2 = \frac{SSR}{n-k}$$

   est un estimateur non biaisé de σ^2.

3. La méthode des moindres carrés produit aussi un estimateur *non biaisé* de la variance d'échantillonnage $\sigma_{b_j}^2$ de chacun des coefficients estimés b_j [10], et de la covariance d'échantillonnage de chaque paire de coefficients estimés, $\sigma_{b_j b_h}$ [11].

4. Dans l'ensemble de tous les estimateurs des β_j qui sont linéaires et qui ne sont pas biaisés, l'estimateur des moindres carrés b_j est le « meilleur », ou celui qui a la plus grande *efficacité relative* (à propos des propriétés désirables des estimateurs, voir le chapitre 2-2), c'est-à-dire qu'il est celui dont la distribution d'échantillonnage a la plus petite variance : on dit que l'estimateur des moindres carrés est « BLUE » (« Best Linear Unbiased Estimate »).

3-2.2.3 LE MODÈLE CLASSIQUE DE LA RÉGRESSION LINÉAIRE *NORMALE*

Comme nous l'avons signalé, le modèle d'échantillonnage défini en combinant le modèle linéaire général avec les hypothèses H1 à H4 est incomplet. Avant de pouvoir procéder à des tests d'hypothèses sur les paramètres, il faut compléter la spécification du modèle d'échantillonnage. Plus précisément, il faut spécifier la forme de la distribution des termes aléatoires u_i. On pense naturellement à

H5. Chacun des termes aléatoires a une distribution normale.

Puisque la distribution normale ne comporte que deux paramètres (la moyenne et la variance), on obtient en combinant H1, H2 et H5 une spécification quasi complète de la distribution des termes aléatoires :

---

[10] Ces valeurs estimées sont données automatiquement par les logiciels d'application statistique.

[11] La matrice estimée des variances-covariances est donnée par $(\boldsymbol{X'X})^{-1}\, SSR\, /\, (n-k)$.

Les termes aléatoires ont des distributions normales identiques, de moyenne nulle, et sont indépendants entre eux ; le seul paramètre inconnu est leur variance commune σ^2.

En vertu des hypothèses H1 à H5, les valeurs des termes aléatoires sont donc tirées d'une même population.

Bien sûr, le choix de la distribution normale est commode. D'abord, puisqu'une combinaison linéaire de variables normales est aussi une variable normale, et puisque les estimateurs des moindres carrés sont linéaires par rapport aux y_i (et donc par rapport aux termes aléatoires u_i), les estimateurs ont eux aussi une distribution normale [12]. Ensuite, la distribution normale est relativement simple à manier, notamment parce qu'elle ne comporte que deux paramètres.

Mais a-t-on des raisons valables de croire que la distribution des termes aléatoires soit réellement normale ? Cela est parfois une implication du modèle théorique ou de la nature du phénomène étudié. Mais la justification la plus rigoureuse fait appel au *Théorème limite central*. En vertu de certaines variantes de ce théorème, si le terme aléatoire de la régression représente l'influence combinée d'un grand nombre de variables manquantes (dans le modèle), alors, on peut considérer que la distribution normale est une approximation raisonnable de la distribution de l'influence combinée des variables manquantes [13].

L'ajout de l'hypothèse H5 complète la spécification du modèle d'échantillonnage connu sous le nom de « modèle classique de régression linéaire *normale* » (c'est-à-dire le modèle classique de régression linéaire, *plus* l'hypothèse de normalité des termes aléatoires). Quand les hypothèses H1 à H5 sont respectées, la statistique met à la disposition du chercheur toute une panoplie de variables-tests qui permettent de réaliser divers types de tests d'hypothèse. Nous avons déjà vu comment on peut appliquer notamment le test *t* de Student, ainsi que le test *F* de Fisher.

Il revient cependant au chercheur de décider si les conditions H1 à H5 qui constituent le modèle d'échantillonnage sont acceptables et si, par conséquent, les tests qui en dépendent sont valides [14]. Car dans le cadre des tests classiques, comme les tests de Student dont nous avons donné quelques exemples, le modèle d'échantillonnage lui-même n'est pas remis en question. La décision d'accepter ou non les hypothèses H1 à H5 peut s'appuyer sur des considérations *a*

---

12 C'est cette propriété qui permet d'appliquer notamment le test *t* de Student.

13 Gujarati (1992, p. 93) ; Theil (1971, p. 368-370) ; Freund (1962, p. 185-188) ; Malinvaud (1969, p. 268-271).

14 C'est la raison d'être de la remarque qu'on trouve au bas de la p. 19 de Lemelin et Polèse (1995) : « Strictly speaking, however, the classical hypotheses under which the tests are exact are not fully realized [...] ».

priori. Mais elle peut être validée *a posteriori*, d'abord par un examen visuel des résidus de la régression, puis, plus rigoureusement, par l'application de *tests diagnostiques*. Ces tests d'hypothèse sont, pour ainsi dire, des tests « de niveau supérieur », qui portent justement sur certains aspects du modèle d'échantillonnage [15].

[15] Par exemple, Heikkila *et al.* (1989) examinent à l'aide de tests diagnostiques la question de la multicollinéarité spatiale entre les diverses variables de distance (p.228-228).

SPÉCIFICATION D'UN MODÈLE ALÉATOIRE : RÉSUMÉ

CONDITIONS DU MODÈLE CLASSIQUE DE RÉGRESSION LINÉAIRE

H1. Pour chaque observation, la valeur du terme aléatoire est tirée d'une population théorique de moyenne nulle :

$E(u_i) = 0$ pour tous les i

H2a) Pour toutes les observations, les populations théoriques d'où sont tirées les valeurs des termes aléatoires ont la même variance :

$\sigma_i^2 = \sigma^2$ pour tous les i

H2b) Pour chaque observation, la valeur du terme aléatoire est statistiquement indépendante des valeurs des termes aléatoires des autres observations :

$\sigma_{ij} = 0$ pour toutes les combinaisons i,j où $i \neq j$

H3. Les variables indépendantes x_{ij} sont non aléatoires (en particulier mesurées sans erreur).

H4. Il y a moins de paramètres à estimer qu'il y a d'observations et il n'y a pas de redondance parmi les variables indépendantes.

PROPRIÉTÉS DE L'ESTIMATEUR DES MOINDRES CARRÉS DANS LE MODÈLE CLASSIQUE DE LA RÉGRESSION LINÉAIRE : LE THÉORÈME DE GAUSS-MARKOV

1. L'estimateur des moindres carrés de β_j est non biaisé :
   la moyenne de la distribution d'échantillonnage de b_j est égale à β_j.
2. $\dfrac{1}{n-k}\sum_i (y_i - \hat{y}_i)^2 = \dfrac{SSR}{n-k}$ est un estimateur non biaisé de σ^2.
3. La méthode des moindres carrés produit aussi un estimateur *non biaisé* de la variance d'échantillonnage $\sigma_{b_j}^2$ de chacun des coefficients estimés b_j, et de la covariance d'échantillonnage de chaque paire de coefficients estimés, $\sigma_{b_j b_h}$.
4. L'estimateur des moindres carrés est l'estimateur linéaire qui a la plus grande efficacité relative (la plus petite variance d'échantillonnage) : il est *BLUE* (« Best Linear Unbiased Estimate »).

CONDITION SUPPLÉMENTAIRE DU MODÈLE CLASSIQUE DE RÉGRESSION LINÉAIRE *NORMALE*

H5. Chacun des termes aléatoires a une distribution normale.

3-2.3 Les hypothèses du modèle d'échantillonnage sont-elles respectées ? Et sinon, qu'arrive-t-il ?

Réf. : Wonnacott et Wonnacott (1992, p. 524-527)

Lorsqu'on effectue des tests d'hypothèse du genre de ceux qui font appel aux variables-tests comme le *t* de Student ou le *F* de Fisher, le modèle d'échantillonnage lui-même n'est pas remis en question (voir 3-2.1). L'analyste doit cependant se demander si le modèle d'échantillonnage ne doit pas être remis en question.

Il existe des procédures statistiques formelles pour tester certains aspects du modèle d'échantillonnage. Mais on peut souvent établir un diagnostic préliminaire par un simple examen visuel du graphique des résidus. En général, tout motif géométrique, toute apparence d'organisation devraient nous mettre la puce à l'oreille.

Les principaux problèmes que peut révéler un examen des résidus sont :

5. une mauvaise spécification du modèle théorique ;
6. l'autocorrélation des termes aléatoires ;
7. l'hétéroscédasticité ;
8. des observations excentriques.

Voyons concrètement en quoi consiste chacun de ces quatre problèmes. Nous dirons ensuite quelques mots à propos de la multicollinéarité.

3-2.3.1 Erreur de spécification du modèle

Spécifier un modèle, c'est dresser la liste des variables indépendantes et définir la forme de la relation entre celles-ci et la variable dépendante. L'erreur de spécification la plus fréquente consiste à omettre l'une des variables indépendantes qui devraient faire partie du modèle. On fait aussi une erreur de spécification, quant à la forme, quand on pose une relation linéaire entre les variables, alors qu'il faudrait une relation linéaire entre leurs logarithmes. Il est clair que la présence d'une erreur de spécification remet en question l'ensemble des hypothèses du modèle classique de régression linéaire, à commencer par la relation linéaire elle-même.

Il y a mille et une possibilités d'erreur de spécification d'un modèle. Contentons-nous d'en donner une illustration. Supposons que le « vrai » modèle soit donné par

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + u_i$$

La variable dépendante y_i est fonction de x_i et du carré de x_i. Avec la constante, cela fait trois variables indépendantes. Bien qu'il ne soit pas linéaire à proprement parler, ce modèle est « linéarisable » : il suffit pour cela de considérer que x_i et x_i^2 sont deux variables différentes.

Supposons maintenant que l'on estime le modèle incomplet

$$y_i = \alpha_0 + \alpha_1 x_i + u_i$$

Le terme manquant, $\alpha_2 x_i^2$, se manifestera alors dans les résidus. Si l'on examine le graphique de la relation entre les résidus et la variable indépendante x_i, on détectera entre eux une relation systématique, qui aura la forme d'une courbe. Cette situation est illustrée aux figures 6 à 8.

Illustration géométrique d'une erreur de spécification

Figure 6 - Observations et régressions
Une relation quadratique

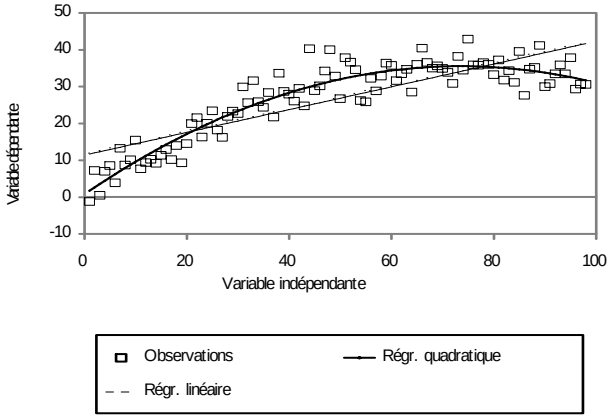


Figure 7 - Résidus régress. quadratiqu
Résidus sans erreur de spécification

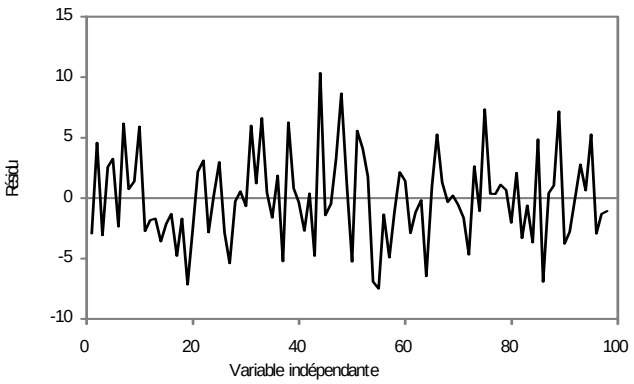
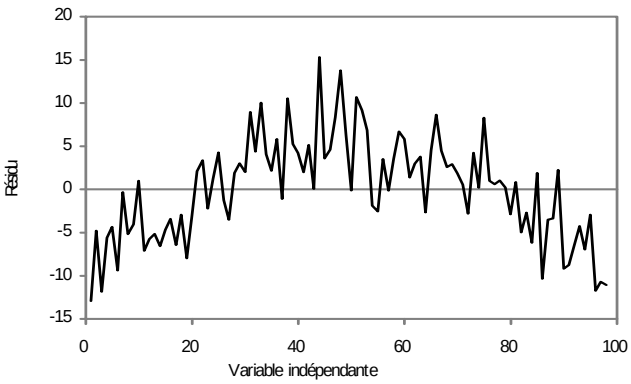


Figure 8 - Résidus régression linéaire
Résidus avec erreur de spécification



3-2.3.2 AUTOCORRÉLATION DES TERMES ALÉATOIRES

L'autocorrélation des termes aléatoires entre en contradiction avec l'hypothèse H2b du modèle classique de la régression linéaire, selon laquelle les termes aléatoires des différentes observations sont indépendants les uns des autres ($\sigma_{ij} = 0$ pour toutes les combinaisons i,j où $i \neq j$).

L'autocorrélation est fréquente lorsque les observations se rapportent à des moments successifs dans le temps. Ce type de données s'appelle séries temporelles ou séries chronologiques. Or, avec les séries temporelles, il arrive souvent qu'à cause d'une certaine inertie, les déviations aléatoires prennent un certain temps à se résorber [16]. Ainsi, si u_t, la valeur du terme aléatoire à la période t, est positive, la moyenne de u_{t+1} *étant donné* $u_t > 0$ (l'espérance mathématique conditionnelle) ne sera pas nulle, mais positive. Dans ces conditions, on a

$$\sigma_{t,t-1} \neq 0$$

ce qui entre en contradiction avec l'hypothèse H2b du modèle classique de la régression linéaire.

On peut souvent détecter la présence d'autocorrélation dans les séries temporelles en examinant le graphique des résidus en fonction du temps. On observera alors que les erreurs successives, au lieu de faire des sauts désordonnés, ont l'air de s'enchaîner les unes aux autres. Les figures 9 et 10, ci-après, illustrent ce phénomène.

Les données sous-jacentes aux résidus de régressions présentés aux figures 9 et 10 ont été générées au moyen de l'équation

$$y_t = x_t + 10\ \eta_t$$

où η_t a été généré au moyen d'un processus autorégressif de la forme

$$\eta_t = (1 - \alpha)\ \varepsilon_t + \alpha\ \eta_{t-1}$$

16 Considérons par exemple le phénomène de l'évolution des prix. Il y a dans le fonctionnement de l'économie des « rigidités institutionnelles » (en particulier les contrats de longue durée, comme les conventions collectives) qui font que les prix ne réagissent pas immédiatement aux changements dans les facteurs fondamentaux. Dans la relation entre les prix et les facteurs fondamentaux, cela se traduit par de l'autocorrélation.

ε_t ayant une distribution proche de la normale [17]. Les figures sont répétées trois fois, avec trois valeurs différentes du paramètre α : 0,9, 0,6 et 0.

Il peut aussi y avoir de l'autocorrélation dans des données spatiales. La détection est alors plus difficile : en effet, alors que le temps n'a qu'une seule dimension, l'espace en a deux, de sorte qu'on ne peut pas tracer le type de graphique évoqué plus haut.

Conséquences

Les estimateurs des moindres carrés demeurent non biaisés, mais leur variance est forte (les estimateurs sont moins précis). De plus, les formules données précédemment pour estimer la variance des estimateurs sous-estiment la vraie variance (c'est-à-dire qu'elles donnent l'illusion de plus de précision) ; il s'ensuit aussi que les tests statistiques ne sont plus valides.

Test de détection

Il existe un test de détection de l'autocorrélation temporelle, le test de Durbin-Watson (voir aussi Kennedy, 1992, p. 128).

Remèdes

Le remède à appliquer lorsqu'il y a autocorrélation est de compléter le modèle en posant des hypothèses sur le mécanisme d'autocorrélation, ce qui permet d'utiliser une méthode appelée méthode des *moindres carrés généralisés*.

---

17 Plus exactement, il s'agit de la transformation logistique d'une variable ayant une distribution uniforme entre zéro et un.

Illustration géométrique de l'autocorrélation (autocorrélation forte)

Figure 9 - Observations et régression
Autocorrélation des termes aléatoires

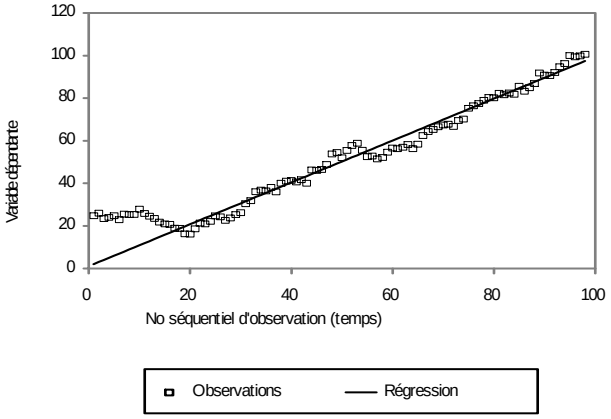
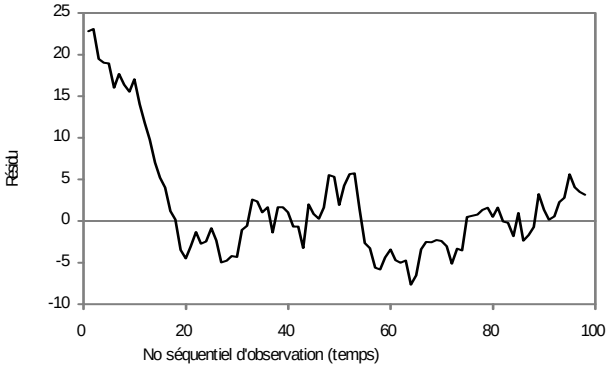


Figure 10 - Résidus de la régression
Autocorrélation des termes aléatoires



Termes aléatoires η_t générés au moyen de la formule

$$\eta_t = (1 - \alpha)\ \varepsilon_t + \alpha\ \eta_{t-1}, \text{ avec } \alpha = 0{,}9$$

Illustration géométrique de l'autocorrélation (autocorrélation moyenne)

Figure 9 - Observations et régression
Autocorrélation des termes aléatoires

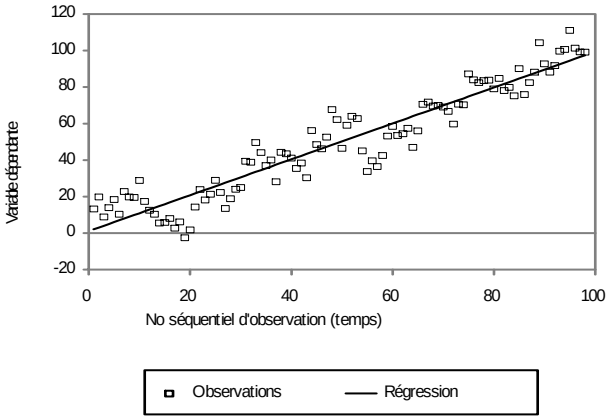
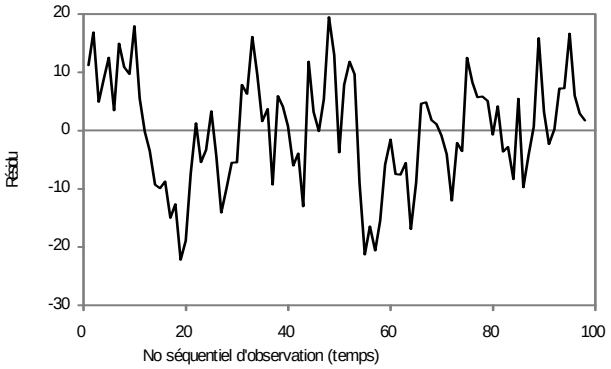


Figure 10 - Résidus de la régression
Autocorrélation des termes aléatoires



Termes aléatoires η_t générés au moyen de la formule

$$\eta_t = (1 - \alpha)\ \varepsilon_t + \alpha\ \eta_{t-1}, \text{ avec } \alpha = 0{,}6$$

Illustration géométrique de l'autocorrélation (pas d'autocorrélation)

Figure 9 - Observations et régression
Autocorrélation des termes aléatoires

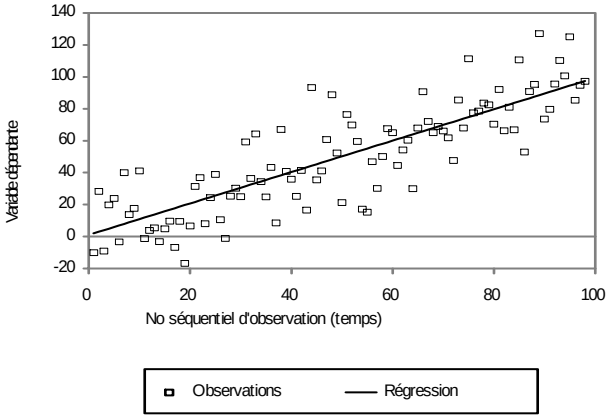
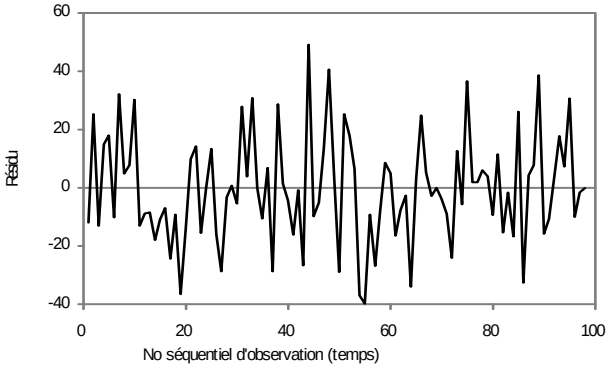


Figure 10 - Résidus de la régression
Autocorrélation des termes aléatoires



Termes aléatoires η_t générés au moyen de la formule

$$\eta_t = (1 - \alpha)\ \varepsilon_t + \alpha\ \eta_{t-1}, \text{ avec } \alpha = 0$$

3-2.3.3 HÉTÉROSCÉDASTICITÉ

L'hétéroscédasticité est le contraire de l'homoscédasticité. L'homoscédasticité est l'hypothèse H2a du modèle classique de la régression linéaire, selon laquelle la variance de l'erreur est la même pour toutes les observations ($\sigma_i^2 = \sigma^2$ pour tous les *i*).

L'hétéroscédasticité se rencontre souvent en particulier dans les études en coupe transversale, où il arrive que le terme aléatoire soit proportionnel à la « taille » du sujet observé.

> Par exemple, dans une étude sur les dépenses de logement des ménages, il se pourrait que la variabilité des dépenses de logement augmente avec le revenu : les ménages à faible revenu sont forcément obligés de limiter leurs dépenses de logement ; parmi les ménages à revenu élevé cependant, certains choisiront de consacrer une grande partie de leur revenu à un logement luxueux, alors que d'autres préféreront se loger confortablement mais sans luxe, pour dépenser leur argent autrement. Dans ce cas particulier, les variations aléatoires dues aux différences de goût [18] seront plus grandes pour les ménages plus aisés.

Sur un graphique des résidus, l'hétéroscédasticité pourra apparaître comme un motif en forme de trompette ou de cornet lorsque les observations sont rangées par ordre croissant de la

variable dépendante ou de l'une des variables indépendantes. Au lieu de n'avoir en abscisse que les numéros d'ordre des observations, on peut aussi construire un graphique où l'on représente les résidus en fonction des valeurs correspondantes de la variable dépendante ou de l'une des variables indépendantes en abscisse.

Les données sous-jacentes aux résidus de régressions présentés aux figures 11 et 12 ont été générées au moyen de l'équation

$$y_i = x_i + 100\ \eta_i$$

où $x_i = i$ et où η_i a été généré au moyen de l'équation

$$\eta_i = 0{,}1\left(\varepsilon_i \sqrt{x_i}\right)$$

ε_i ayant une distribution proche de la normale [19].

Il est à noter que $x_i = i$, de sorte que les observations sont automatiquement rangées par ordre croissant de la variable indépendante x_i.

Conséquences

La précision de l'estimateur est moins bonne et les tests d'hypothèse ne sont pas valides.

Tests de détection

Test de Goldfeld et Quandt (Theil, 1971, p. 196-199 ; voir aussi Kennedy, 1992, p. 126).

Remèdes

Correction du modèle d'échantillonnage par la transformation des données. Par exemple, si on trouve que la variance est liée à l'une des variables indépendantes, disons x_{ik}, et que cette relation peut être représentée approximativement par

$$\sigma_i^2 = x_{ik}\,\sigma^2$$

alors on peut recréer l'homoscédasticité et les conditions de Gauss-Markov en appliquant aux variables la transformation qui suit :

$$y_i' = \frac{y_i}{\sqrt{x_{ik}}} \text{ et } x_{ij}' = \frac{x_{ij}}{\sqrt{x_{ik}}}$$

---

[18] Nous avons ici un exemple de modèle auquel il manque certaines variables inobservables (les goûts), dont l'effet est représenté par le terme aléatoire.

Illustration géométrique de l'hétéroscédasticité

Figure 11 - Observations et régression
Hétéroscédasticité

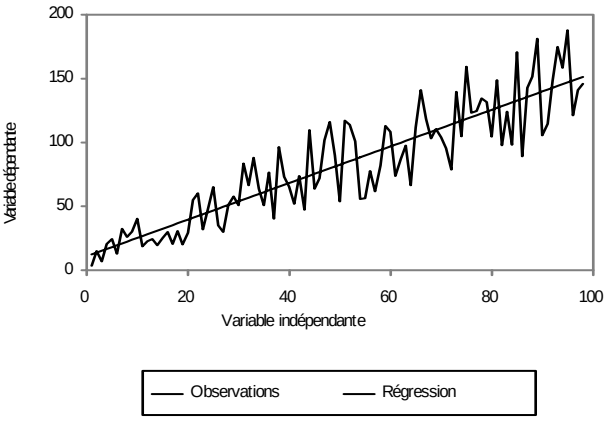
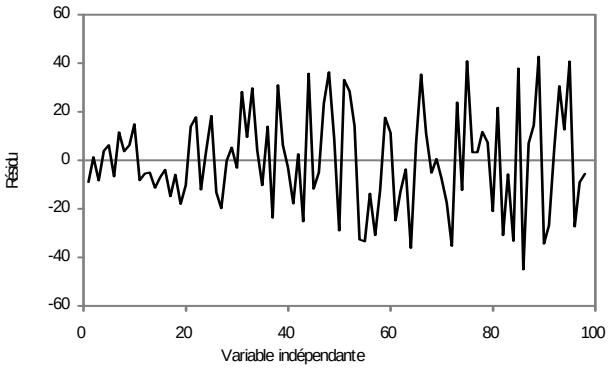


Figure 12 - Résidus de la régression
Hétéroscédasticité



Termes aléatoires η_i générés au moyen de la formule

$$\eta_i = 0{,}1\left(\varepsilon_i \sqrt{x_i}\right)$$

3-2.3.4 OBSERVATIONS EXCENTRIQUES

Les observations excentriques (*outliers*) sont parfois dues à des situations où sont intervenus des facteurs exceptionnels qui ne sont pas pris en compte dans le modèle. Bien que les observations excentriques n'entrent pas en conflit avec les hypothèses du modèle classique de la régression linéaire, elles peuvent fausser les résultats de la régression. Un examen des résidus permet de repérer les observations excentriques : on peut ensuite se demander si elles s'expliquent par des facteurs particuliers et si on doit les écarter des données. Il faut cependant se garder d'éliminer des observations *ad hoc*, par commodité ! Les figures qui suivent offrent un exemple visuel de résidus en présence d'observations excentriques.

---

19 Plus exactement, il s'agit de la transformation logistique d'une variable ayant une distribution uniforme entre zéro et un.

Illustration géométrique de la présence d'observations excentriques

Figure 13 - Observations et régression
Observations excentriques (Outliers)

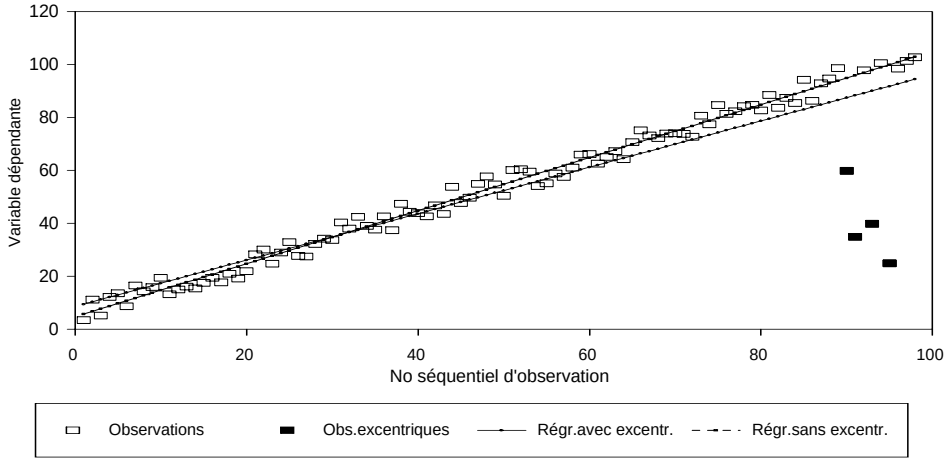


Figure 14 - Résidus des régressions
avec observ excentriques (Outliers)

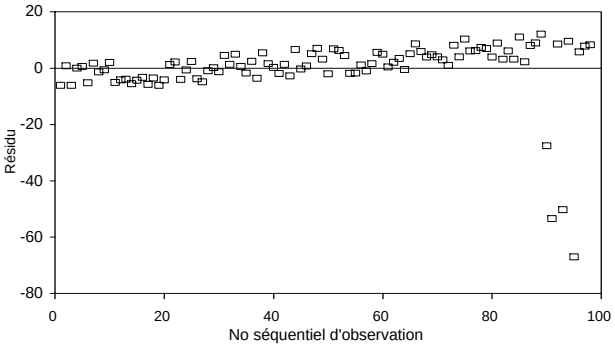
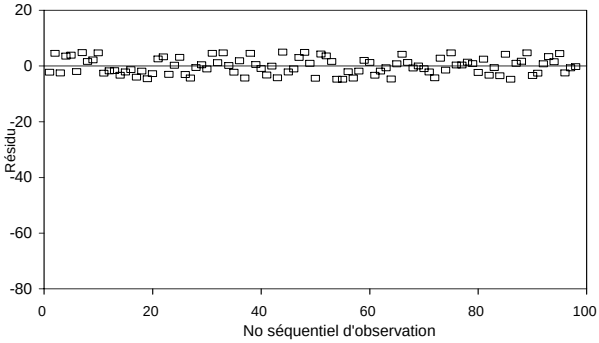


Figure 15 - Résidus des régressions
sans observ. excentriques (Outliers)



3-2.3.5 MULTICOLLINÉARITÉ

Réf. : Wonnacott et Wonnacott (1992, p. 568-572)

Définition

On distingue la multicollinéarité stricte et la multicollinéarité approximative.

a) Multicollinéarité stricte

La **multicollinéarité stricte** entre en contradiction avec la partie de l'hypothèse H4 du modèle classique de la régression linéaire selon laquelle il ne doit pas y avoir de redondance parmi les variables indépendantes [20]. La multicollinéarité stricte est extrêmement rare avec des données réelles. Elle peut cependant résulter d'une erreur de spécification lorsque le modèle contient des variables muettes (*dummy variables*). Nous reviendrons sur cette question lorsqu'il sera question d'analyse de variance au moyen de la régression multiple (chapitre 4-2).

La multicollinéarité stricte ne pose aucun problème de détection, puisque sa présence est diagnostiquée par les logiciels d'application statistique (à cause de l'impossibilité des calculs d'estimation).

b) Multicollinéarité approximative

La **multicollinéarité approximative** est beaucoup plus fréquente. Elle se produit lorsque l'une des variables indépendantes est fortement correliée à une autre ou à une combinaison linéaire des autres. Cette variable, sans être strictement redondante, peut être « presque » redondante : l'information supplémentaire qu'elle apporte n'ajoute pas grand'chose à celle que contiennent déjà les autres variables.

Conséquences

La précision des estimateurs est faible, c'est-à-dire que leurs variances d'échantillonnage s_{bj}^2 sont grandes. On ne peut pas bien séparer l'influence des variables qui sont correliées entre elles.

---

[20] Techniquement, lorsqu'il y a multicollinéarité stricte, le rang de la matrice $\boldsymbol{X}$ est inférieur à k. Alors l'inverse $(\boldsymbol{X'X})^{-1}$ n'existe pas, et l'estimateur non plus.

Concrètement, dans le cas de deux variables, cela peut se manifester de la façon suivante : aucune des deux variables n'a un coefficient significativement différent de zéro ; mais si l'on retire les deux variables, le modèle contraint ainsi défini est rejeté par le test *F*.

Tests de détection

Pour la corrélation entre variables deux à deux, on peut examiner les coefficients de corrélation simple des variables indépendantes entre elles. Mais la multicollinéarité est souvent plus complexe et il faut recourir à des analyses plus raffinées (voir Kennedy, 1992, p. 180, à propos du *condition index*).

Remèdes

Dans certains cas, il y a peut-être une ou plusieurs variables indépendantes de trop dans le modèle. Mais très souvent, il faut accepter de « vivre avec » et renoncer à séparer l'influence des variables correliées. Dans certains cas, ce serait même une erreur grave d'éliminer une variable pour cause de multicollinéarité.

Le schéma suivant illustre une telle situation. Les facteurs inobservables *A*, *B* et *C* influencent la variable dépendante *Y*. Les facteurs inobservables ne peuvent pas figurer dans le modèle. À leur place, le modèle contient deux variables indépendantes observables, X_1 et X_2 : la première est influencée par *A* et *B* et la seconde, par *B* et *C*. À cause de l'influence commune de *B*, X_1 et X_2 seront correliées. Mais si l'on écarte l'une des deux, on élimine du même coup le facteur sous-jacent *A* ou *C*.