Econométrie I

Bernard Lejeune

HEC-Université de Liège

Notes à l'usage des étudiants de 3ème année de
bachelier en sciences économiques et de gestion

Année académique 2010 - 2011

Préambule

En parallèle des présentes notes de cours, les étudiants sont invités à lire :

Hill R.C., Griffiths W.E. et Lim G.C. (2008), *Principles of Econometrics*, Third Edition, John Wiley & Sons.

D'autres ouvrages peuvent aussi utilement être consultés :

Griffiths W.E., Hill R.C., Judge G.G. (1993), *Learning and Practicing Econometrics*, John Wiley & Sons.

Wooldridge J.M. (2006), *Introductory Econometrics : A Modern Approach*, Fourth Edition, South-Western.

Johnston J. et DiNardo J. (1997), *Econometrics Methods*, Fourth Edition, Mc Graw-Hill.

Les étudiants qui souhaitent en savoir plus pourront consulter :

Goldberger A.S. (1991), *A Course in Econometrics*, Harvard University Press.

Wooldridge J.M. (2010), *Econometric Analysis of Cross-Section and Panel Data*, Second Edition, MIT Press.

Cameron A.C. et Trivedi P.K. (2005), *Microeconometrics : Methods and Applications*, Cambridge University Press.

Hamilton J.D. (1994), *Time Series Analysis*, Princeton University Press.

Table des matières

Chapitre 1

Introduction

1.1. Economie et économétrie

L'objet de la théorie économique est d'expliquer les comportements économiques au travers de modèles décrivant des relations entre des variables économiques : consommation, épargne, revenu, salaire, production, prix, emploi, investissement, taux d'intérêt, etc...

L'économétrie est un ensemble de méthodes statistiques conçues pour évaluer des relations *empiriques* — càd. que l'on peut observer dans des données — entre des variables, en particulier des relations suggérées par la théorie économique.

Un outil central de l'économétrie est le *modèle de régression*. La plus grande partie du présent cours est consacrée à l'étude des différentes facettes de ce modèle.

1.2. Le modèle de régression

En bref, le modèle de régression s'efforce de décrire la façon dont la *valeur moyenne* prise par une variable, appelée variable *dépendante*, *expliquée* ou encore *endogène*, varie en fonction des valeurs prises par une ou plusieurs autre(s) variable(s), appelée(s) variable(s) *conditionnante(s)*, *indépendante(s)*, *explicative(s)* ou encore *exogène(s)*.

Il est important de noter que le choix du sens de la causalité entre ces variables est un choix à priori. Il n'est pas déterminé par l'outil statistique. Au demeurant, il peut très bien exister une relation entre des variables sans qu'il y ait pour autant une quelconque causalité.

Le modèle de régression permet :

1- de représenter et de quantifier des relations entre variables,

2- de faire des prévisions, en particulier de *l'effet marginal* d'une variable, les autres variables étant maintenues constantes (effet d'une variable *ceteris paribus*, càd.

toutes autres choses étant égales),

3- de tester des théories.

Ces différents aspects ne sont évidemment pas sans liens.

Le modèle de régression, et de façon plus générale l'économétrie, s'appuie sur :

1- l'économie mathématique, en tant que pourvoyeur des relations formalisées entre variables,

2- un ensemble d'outils issus de la théorie des probabilités et de l'inférence statistique,

3- des données. Parmi les données, on distingue :

a- les *données en coupe* ou données individuelles. Ces données sont constituées d'un ensemble de variables (de stock et/ou de flux) mesurées au cours d'une même période de temps pour des unités statistiques distinctes (des individus, des ménages, des firmes, des pays, etc...). Elles sont généralement notées :

$$x_i,\ i = 1, ..., n$$

Il s'agit typiquement de données de nature microéconomique, issues d'enquêtes.

b- les *données chronologiques* ou séries temporelles. Ces données sont constituées d'un ensemble de variables mesurées pour une même unité statistique au cours du temps. Elles sont généralement notées :

$$x_t,\ t = 1, ..., T$$

Il s'agit typiquement de données de nature macroéconomique, issues des comptes nationaux.

c- les *données en panel*. Ces données combinent les deux types de données précédents. Elles sont constituées d'un ensemble de variables mesurées pour des unités statistiques distinctes au cours de plusieurs périodes successives. Elles sont généralement notées :

$$x_{it},\ i = 1, ..., n\ ,\ t = 1, ..., T$$

Il s'agit souvent de données de nature microéconomique, issues d'enquêtes.

On notera que les données en panel, et encore davantage les séries temporelles, soulèvent des problèmes, et donc appellent à des développements techniques spécifiques, qui seront ignorés dans le cadre de ce cours d'introduction.

On remarquera également que, dans presque tous les cas, les données à disposition des économètres sont *non-expérimentales* : il est impossible de modifier de façon expérimentale le revenu d'un ménage pour vérifier si il ajuste ou non, et de combien, son niveau de consommation. Il en est évidemment de même pour des variables de type macroéconomique. Pour étudier une relation, l'économètre doit presque

toujours se contenter de données observables (dans l'exemple ci-avant, les couples revenu - consommation de différents ménages) sur lesquelles il n'a aucun contrôle de type expérimental.

1.3. L'approche économétrique

Une analyse économétrique débute toujours par l'identification d'une (ou plusieurs) relation(s) entre variables, suggérée(s) par la théorie économique, et dont la connaissance quantitative apporterait des éléments de réponse à la question que l'on se pose. Par exemple, si l'on s'interroge sur la propension marginale à consommer des biens culturels des ménages, on s'intéressera naturellement à la relation entre consommation de biens culturels (les dépenses des ménages en la matière) et revenu d'un ménage :

$$Cons = f(revenu)$$

Une fois la relation d'intérêt identifiée, l'approche économétrique consiste en la construction d'un *modèle probabiliste* de cette relation, comprenant comme ingrédients essentiels des variables aléatoires (v.a.) et des paramètres (par., inconnus à priori), pour l'exemple ci-dessus[1] :

$$\underset{\text{v.a.}}{E(Cons}|\underset{\text{v.a.}}{revenu}) = \underset{\text{par.}}{\beta_1} + \underset{\text{par.}}{\beta_2}\ \underset{\text{v.a.}}{revenu} \Leftrightarrow \underset{\text{v.a.}}{Cons} = \underset{\text{par.}}{\beta_1} + \underset{\text{par.}}{\beta_2}\ \underset{\text{v.a.}}{revenu} + \underset{\text{v.a.}}{e}\ ,$$

ce modèle probabiliste étant tel que les données (les observations) dont on dispose peuvent être considérées (pour des raisons d'échantillonnage et/ou de modélisation) comme des *réalisations particulières* des variables aléatoires du modèle, pour une certaine valeur des paramètres inconnus. En somme, cela revient à regarder les observations dont on dispose comme le résultat d'une loterie, les règles de la loterie étant définie par la structure du modèle et la valeur de ses paramètres.

Sur cette base, en utilisant des outils statistiques appropriés, on pourra :

1- estimer les paramètres inconnus du modèle et évaluer la précision de ces estimations (quantification de la relation d'intérêt),

2- tester des hypothèses économiques liées aux paramètres du modèle (dans l'exemple ci-dessus, pour tester si la propension marginale à consommer est inférieur à 1, on testera si $\beta_2 < 1$),

3- faire des prévisions et évaluer la précision de ces prévisions,

4- tester l'adéquation du modèle probabiliste (ses hypothèses statistiques) aux données.

[1] sous la forme générique d'un modèle de régression.

1.4. Rappel de théorie des probabilités

Avant d'entrer dans le vif du sujet, les étudiants sont invités à rafraîchir leurs connaissances relatives à une série de concepts de base de la théorie des probabilités :

- Variables aléatoires et distributions de probabilité (cas discret et cas continu).
- Espérance et variance d'une variable aléatoire, propriétés de l'espérance et de la variance.
- Couples de variables aléatoires :
  - loi jointe, marginale et conditionnelle.
  - espérance et variance conditionnelle.
  - indépendance, covariance et corrélation.
  - espérance et variance de combinaisons linéaires.
- Lois usuelles :
  - loi de Bernouilli.
  - loi normale.
  - loi du khi-carré (χ^2).
  - loi de Student (t).
  - loi de Fisher(-Snedecor) (F).

Un résumé de ces concepts est fourni dans l'annexe B de l'ouvrage de Hill, Griffiths et Lim (2008), dont la référence est donnée dans le Préambule de ces notes.

Chapitre 2

Le modèle de régression linéaire simple

2.1. Du modèle économique au modèle économétrique

2.1.1. Un modèle économique

Supposons qu'on s'intéresse à la relation entre consommation et revenu. De Keynes (1936)[2] : "en moyenne et la plupart du temps, les hommes tendent à accroître leur consommation à mesure que leur revenu croît, mais non d'une quantité aussi grande que l'accroissement du revenu". De façon formelle, cette assertion peut être décrite par la relation théorique :

$$y = f(x)\,, \text{ avec } 0 < \frac{dy}{dx} < 1,$$

où y = consommation et x = revenu.

2.1.2. Construction du modèle économétrique I : la droite de régression

On cherche une *contrepartie empirique* de la relation théorique $y = f(x)$, une contrepartie empirique prenant la forme d'un *modèle probabiliste paramétré*.

L'essence de l'approche économétrique, et de façon plus générale de toute la statistique inférentielle, est de regarder les données dont on dispose comme des *réalisations particulières* de variables aléatoires. Construire un modèle économétrique de la relation d'intérêt implique donc de s'interroger sur la façon dont les données sont obtenues, générées.

---

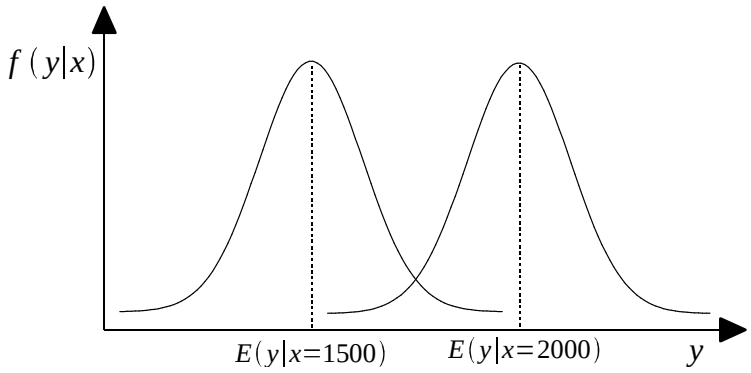
[2] Keynes, J. M. (1936), *The General Theory of Employment, Interest and Money*, Palgrave Macmillan.

A. Données en coupe

Les données en coupe sont généralement obtenues par tirages aléatoires d'individus au sein d'une population, ou peuvent à tout le moins généralement être considérées comme telles. Dans notre exemple, de telles données seraient constituées par les valeurs du couple (x, y) du revenu et de la consommation d'un échantillon de ménages tirés au hasard dans une population.

Au travers de l'épreuve aléatoire 'tirer un individu au hasard dans la population et noter la valeur de son revenu x et de sa consommation y', on peut représenter la population par une *distribution de probabilité jointe* $f(x, y)$, inconnue et à priori complexe, qui correspond à la distribution de fréquence des couples de variables (x, y) dans la population.

Lorsqu'on cherche à expliquer y en fonction de x, l'information pertinente est concentrée dans la *distribution conditionnelle* $f(y|x)$ qui, pour chaque valeur de x, correspond à la distribution de fréquence des différentes valeurs de y dans la population. Typiquement :



Graphique 1 : Distributions conditionnelles

La distribution conditionnelle $f(y|x)$ peut être résumée par l'*espérance conditionnelle* de y sachant x — aussi appelée *courbe de régression* de y en x — qui, pour chaque valeur de x, correspond à la valeur moyenne de y dans la population. De manière générale, on a :

$$E(y|x) = g(x) \qquad \text{(i.e., une fonction de } x\text{)}$$

L'espérance conditionnelle de y sachant x constitue, dans le modèle de régression, la contrepartie empirique de la relation théorique d'intérêt $y = f(x)$.

Avant de poursuivre, illustrons ces différents concepts pour une population hypothétique dont la distribution (discrète) jointe du revenu ($= x$) et de la consom-

mation de biens culturels ($= y$) est donnée par :

$f(x,y)$	50	100	150	200	250	300	$f(x)$
1500	0,28	0,08	0,04	0	0	0	0,40
2000	0,03	0,15	0,06	0,06	0	0	0,30
2500	0	0,03	0,06	0,15	0,03	0,03	0,30
$f(y)$	0,31	0,26	0,16	0,21	0,03	0,03	1

De la distribution jointe $f(x, y)$, on peut déduire les distributions marginales de x et de y. Elles sont données[3], respectivement, par :

$$f(x) = \sum_y f(x, y) \quad \text{et} \quad f(y) = \sum_x f(x, y)$$

De la distribution jointe $f(x, y)$ et de la distribution marginale de x, on peut par ailleurs déduire la distribution conditionnelle et l'espérance conditionnelle de y sachant x. Elles sont données[4], respectivement, par :

$$f(y|x) = \frac{f(x, y)}{f(x)} \quad \text{et} \quad E(y|x) = \sum_y y f(y|x)$$

On obtient :

$f(y\|x)$	50	100	150	200	250	300	$E(y\|x)$
1500	0,7	0,2	0,1	0	0	0	70
2000	0,1	0,5	0,2	0,2	0	0	125
2500	0	0,1	0,2	0,5	0,1	0,1	195

L'espérance conditionnelle $E(y|x) = g(x)$ définit un modèle probabiliste de la relation théorique d'intérêt $y = f(x)$, dont les variables aléatoires[5] x et y ont des probabilités de réalisation décrites par la distribution jointe inconnue $f(x, y)$. On obtient un *modèle probabiliste paramétré* de la relation théorique d'intérêt si on suppose une forme fonctionnelle, dépendant de paramètres, pour $g(x)$. Le modèle de régression linéaire simple suppose :

$$E(y|x) = \beta_1 + \beta_2 x \qquad \text{(i.e., une fonction linéaire de } x\text{)}$$

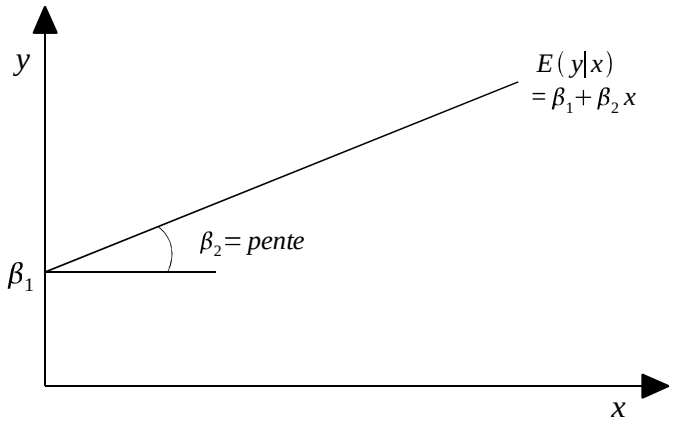
Graphiquement :

---

[3] Dans le cas continu, $f(x) = \int_{-\infty}^{\infty} f(x, y)dy$ et $f(y) = \int_{-\infty}^{\infty} f(x, y)dx$.

[4] Dans le cas continu, $E(y|x) = \int_{-\infty}^{\infty} y f(y|x)dy$, $f(y|x)$ ayant la même définition.

[5] Dans la théorie des probabilités, on distingue dans la notation les variables aléatoires (notées en majuscule) et leurs réalisations (notées en minuscule). Pour alléger la notation, et comme il est usuel de le faire, dans le cadre de ces notes, nous ne ferons pas cette distinction : x, y, X ou Y désigneront toujours des variables (ou vecteurs) aléatoires lorsqu'on raisonne dans le cadre d'un modèle probabiliste *avant observation* (avant de les observer, leurs valeurs sont inconnues : ce sont des variables aléatoires), et des valeurs prises par ces variables aléatoires lorsqu'elles représentent des observations dans un *échantillon particulier*.



Graphique 2 : La droite de régression

Si le modèle de régression linéaire simple est correct, chaque observation (x_i, y_i) satisfait le modèle probabiliste :

$$E(y_i|x_i) = \beta_1 + \beta_2 x_i \,, \qquad i = 1, ..., n,$$

où β_1 et β_2 sont des paramètres inconnus à estimer et, avant observation, y_i et x_i sont des variables aléatoires.

B. Données chronologiques

Dans notre exemple, de telles données pourraient être constituées soit d'observations du revenu x et de la consommation y d'un ménage au cours du temps, soit de données macroéconomiques agrégées (revenu et consommation nationales ; le plus probable).

Pour ce type de données, il n'est plus possible de s'appuyer sur l'idée d'un échantillonnage au sens strict (physique) du terme. On peut néanmoins continuer à regarder les observations dont on dispose comme le résultat de tirages aléatoires (x_t, y_t) dans des distributions telles que le modèle de régression linéaire simple est satisfait, càd. telles que, pour tout t :

$$E(y_t|x_t) = \beta_1 + \beta_2 x_t \,, \qquad t = 1, ..., T$$

Ainsi, avec les séries temporelles, on passe d'une approche d'échantillonnage au sens strict (physique) du terme à une approche purement probabiliste de *modélisation*, qui contient la première comme cas particulier. Ce n'est toutefois qu'une question d'interprétation : l'outil statistique reste inchangé.

2.1.3. Construction du modèle économétrique II : hypothèses complémentaires

Outre l'hypothèse centrale que les observations sont telles que

$$E(y_i|x_i) = \beta_1 + \beta_2 x_i\,, \qquad i = 1, ..., n,$$

le modèle de régression linéaire simple s'appuie sur un ensemble d'hypothèses statistiques complémentaires qui, pour l'essentiel, peuvent être relâchées si nécessaire.

Ces hypothèses sont les suivantes :

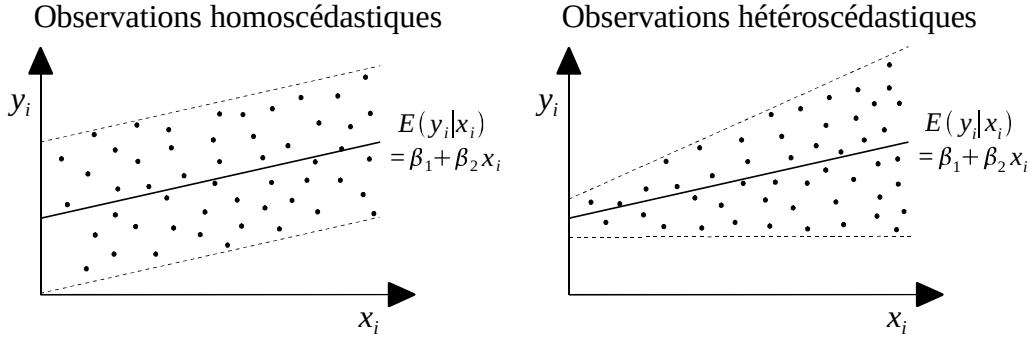
1- De manière générale, on peut avoir :

$$Var(y_i|x_i) = h(x_i) \qquad \text{(i.e., une fonction de } x_i)$$

Le modèle de base suppose que :

$$Var(y_i|x_i) = \sigma^2 \qquad \text{(i.e., une constante, ne dépend pas de } x_i)$$

Lorsque la variance (conditionnelle) est constante, on parle d'*homoscédasticité*. Lorsque la variance (conditionnelle) n'est pas constante, on parle d'*hétéroscédasticité*. Graphiquement :



Graphique 3 : Homoscédasticité et hétéroscédasticité

2- Les variables explicatives x_i sont *fixes*, *non-stochastiques*, et prennent au moins deux valeurs distinctes.
Si les x_i prenaient tous la même valeur, il serait impossible de mener une analyse de régression, càd. de regarder la façon dont y_i varie en fonction de x_i, puisque x_i ne varierait pas.
L'hypothèse que les x_i sont non-stochastiques (non-aléatoires) est faite pour des raisons de commodité technique. Elle équivaut à raisonner conditionnellement aux valeurs de x_i observées dans l'échantillon. Au sens strict, cette hypothèse correspond au cas d'un *échantillonnage stratifié,* où les x_i sont choisis à l'avance, puis les y_i correspondants tirés aléatoirement dans les sous-populations d'individus caractérisés par les x_i choisis. Ainsi, pour chaque x_i choisi, un y_i est tiré aléatoirement dans la distribution conditionnelle $f(y|x_i)$.
Sous cette hypothèse de régresseurs (variables explicatives) non-stochastiques,

on peut réécrire les hypothèses :

$$\begin{array}{l} E(y_i|x_i) = \beta_1 + \beta_2 x_i \\ Var(y_i|x_i) = \sigma^2 \end{array} \qquad i = 1, ..., n,$$

sous la forme plus simple[6] :

$$\begin{array}{l} E(y_i) = \beta_1 + \beta_2 x_i \\ Var(y_i) = \sigma^2 \end{array} \qquad i = 1, ..., n$$

3- Les observations $y_1, ..., y_n$ sont statistiquement indépendamment distribuées, ou de façon moins restrictive, sont toutes 2 à 2 non corrélées (pour rappel, l'indépendance statistique implique la non-corrélation) :

$$Cov(y_i, y_j) = 0\,, \quad \forall i \neq j$$

Cette hypothèse est automatiquement satisfaite dans le cas de tirages avec remise (ou de tirages sans remise si la population est infinie[7]).

4- De façon *optionnelle*, on fait parfois l'hypothèse que la distribution conditionnelle $f(y_i|x_i)$ est normale, auquel cas on a :

$$y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2) \qquad i = 1, ..., n$$

Sous cette hypothèse optionnelle, on pourra obtenir des résultats d'inférence valables en *échantillon fini*. Sans cette hypothèse, les mêmes résultats ne seront valables qu'en grand échantillon (on dit *asymptotiquement*).

En résumé, le modèle de régression linéaire simple considère chaque observation (x_i, y_i) comme la réalisation d'un processus aléatoire satisfaisant les hypothèses suivantes $(i = 1, ..., n)$:

(1) $E(y_i) = \beta_1 + \beta_2 x_i$

(2) $Var(y_i) = \sigma^2$

(3) $Cov(y_i, y_j) = 0, \forall i \neq j$

(4) les x_i sont non-stochastiques et prennent au moins 2 valeurs distinctes

(5) (optionnel) $y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$

2.1.4. Introduction d'un terme d'erreur

Le modèle de régression est le plus souvent exprimé en faisant apparaître un terme d'erreur.

---

[6] Le conditionnement explicite par rapport à x_i est redondant lorsque x_i est traité comme non-stochastique : à chaque observation i est associée une valeur x_i qui est censée avoir été choisie à l'avance. En pratique, ce n'est pas le cas. On peut néanmoins, sans grande conséquence pour ce qui suit, faire comme si c'était bien le cas. Nous reviendrons sur ce point au Chapitre 7. Cette écriture simplifiée ne doit cependant pas nous faire perdre de vue que l'on raisonne toujours conditionnellement aux x_i observés.

[7] En pratique, si la population est bien plus grande que l'échantillon tiré.

Par définition, le terme d'erreur e_i est donné par :

$$e_i = y_i - E(y_i) = y_i - \beta_1 - \beta_2 x_i \,,$$

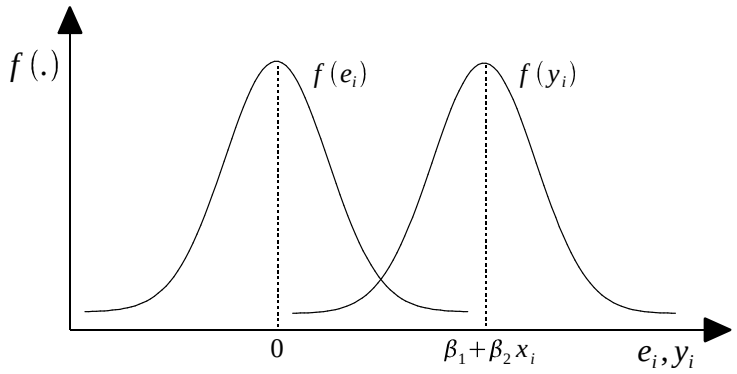
de sorte qu'on peut réécrire le modèle comme :

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

La variable y_i étant (avant observation) une variable aléatoire, e_i est aussi une variable aléatoire, dont les propriétés sont :

$$\begin{aligned}
E(e_i) &= E\left(y_i - E(y_i)\right) = E(y_i) - E(y_i) = 0 \\
Var(e_i) &= E\left[\left(e_i - E(e_i)\right)^2\right] = E(e_i^2) \\
&= E\left[\left(y_i - E(y_i)\right)^2\right] \\
&= Var(y_i) = \sigma^2 \\
Cov(e_i, e_j) &= E\left[\left(e_i - E(e_i)\right)\left(e_j - E(e_j)\right)\right] = E(e_i e_j) \\
&= E\left[\left(y_i - E(y_i)\right)\left(y_j - E(y_j)\right)\right] \\
&= Cov(y_i, y_j) = 0 \,, \qquad \forall i \neq j
\end{aligned}$$

Par ailleurs, e_i étant une fonction linéaire de y_i, si $y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$, alors e_i suit aussi une loi normale[8] : $e_i \sim N\left(E(e_i), Var(e_i)\right)$, soit $e_i \sim N\left(0, \sigma^2\right)$. Graphiquement :



Graphique 4 : Distribution de y_i et de e_i sous l'hypothèse de normalité

Etant donné ces propriétés, on peut finalement réécrire le modèle de régression linéaire simple et ses hypothèses sous la forme :

A1 $\quad y_i = \beta_1 + \beta_2 x_i + e_i \,, \qquad i = 1, ..., n$

A2 $\quad E(e_i) = 0 \;\Leftrightarrow\; E(y_i) = \beta_1 + \beta_2 x_i \,, \qquad i = 1, ..., n$

A3 $\quad Var(e_i) = \sigma^2 = Var(y_i) \,, \qquad i = 1, ..., n$

A4 $\quad Cov(e_i, e_j) = 0 = Cov(y_i, y_j) \,, \qquad \forall i \neq j$

A5 $\quad$ les x_i sont non-stochastiques et prennent au moins 2 valeurs distinctes

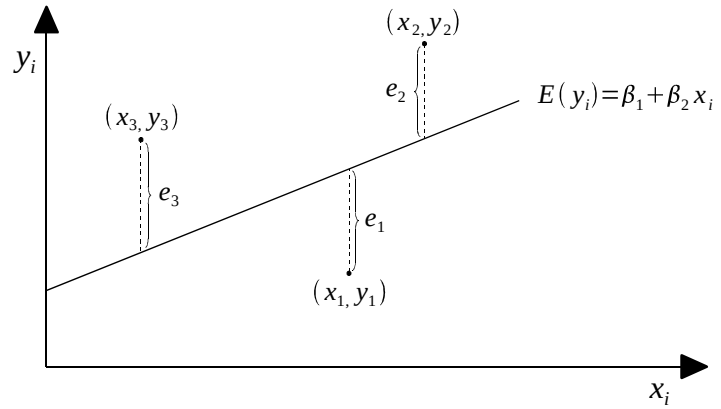
A6 $\quad$ (optionnel) $e_i \sim N\left(0, \sigma^2\right) \;\Leftrightarrow\; y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2) \,, \qquad i = 1, ..., n$

[8] Pour rappel, toute fonction linéaire d'une variance aléatoire normale suit aussi une loi normale.

Quelques points méritent d'être épinglés :

1- La formulation de A1 pourrait donner à penser que le terme d'erreur aléatoire e_i a une 'vie propre'. Ce n'est pas le cas. Le terme d'erreur aléatoire e_i reflète l'écart entre y_i et son espérance *pour une valeur donnée de* x_i. Si on s'intéressait à la relation entre y_i et une autre variable explicative z_i (par exemple, la consommation y_i en fonction du patrimoine z_i dans une population), e_i serait redéfini comme l'écart entre y_i et son espérance *pour une valeur donnée de* z_i, les paramètres β_1 et β_2, voire le caractère linéaire de la relation, étant eux-mêmes redéfinis.

2- Contrairement à y_i qui est une variable aléatoire *observable*, e_i est une variable aléatoire *non observable* puisqu'elle dépend des paramètres inconnus β_1 et β_2 qui *peuvent seulement être estimés*. De la même façon, $E(y_i) = \beta_1 + \beta_2 x_i$ est non observable et peut seulement être estimée.



Graphique 5 : Liens entre (x_i, y_i), e_i et la droite de régression $E(y_i) = \beta_1 + \beta_2 x_i$

3- La dispersion de y_i autour de son espérance pour une valeur donnée de x_i, en d'autres termes l'erreur aléatoire e_i, peut être attribuée :

   a- tout d'abord à l'effet de toutes les variables qui affectent de façon systématique y_i mais non prises en compte dans le modèle,

   b- et au delà, à la variabilité naturelle, intrinsèque, de y_i, qui subsisterait même si toutes les variables qui affectent de façon systématique y_i étaient prises en compte.

   Notons que e_i n'est pas censée refléter une erreur de spécification due à une non-linéarité de $E(y_i|x_i)$. La forme linéaire de $E(y_i|x_i)$ est censée être correcte, même si en pratique il y a fort à parier qu'elle ne l'est éventuellement qu'approximativement.

4- Le modèle de régression linéaire simple peut être utilisé pour évaluer une relation entre deux variables chaque fois que les hypothèses sur lesquelles il s'appuie (linéarité, homoscédasticité, non-corrélation) sont à priori crédibles, ce qui en première approximation peut être jaugé en faisant un graphique des observations (à tout le moins en ce qui concerne les hypothèses de linéarité et d'homoscédasticité).

5- L'hypothèse de linéarité du modèle peut, à première vue, apparaître comme très restrictive. Cette impression est réduite si on note que rien n'empêche les

variables x et y qui interviennent dans le modèle d'être des transformations (le logarithme, le carré, le cube,...) d'autres variables. En fait, l'hypothèse de linéarité requiert seulement que le modèle soit linéaire dans les *paramètres*, pas dans les *variables*. Ainsi, un modèle très utilisé en pratique est le modèle log-log :

$$\begin{aligned} & \ln y_i = \beta_1 + \beta_2 \ln x_i + e_i \\ \Leftrightarrow \quad & y_i^* = \beta_1 + \beta_2 x_i^* + e_i \end{aligned}$$

L'un des attraits de ce modèle est que le paramètre β_2 s'interprète comme l'élasticité de y par rapport à x, ce qui n'est pas le cas du modèle linéaire avec les variables originales. Nous reviendrons en détail sur ce point dans la suite.

2.2. Estimation des paramètres du modèle

On suppose que les observations disponibles sont des réalisations de variables aléatoires satisfaisant les hypothèses A1 - A5 (plus éventuellement A6) du modèle de régression linéaire simple. On cherche à *estimer* les paramètres inconnus β_1 et β_2 de la droite de régression $E(y_i) = \beta_1 + \beta_2 x_i$, qui représente la relation d'intérêt dans la population.

Un *estimateur* est une *règle de décision* établie à priori (avant observation) qui décrit, à l'aide d'une recette ou d'une formule, comment utiliser les observations d'un échantillon pour estimer les paramètres inconnus d'un modèle. Les observations étant des réalisations de variables aléatoires, un estimateur est lui-même une variable aléatoire (sa valeur varie d'un échantillon à l'autre). Une *estimation* est l'application de la règle de décision définissant l'estimateur à un *échantillon particulier*.

2.2.1. L'estimateur des moindres carrés ordinaires

L'estimateur standard du modèle de régression linéaire simple est l'estimateur des moindres carrés ordinaires (MCO). Il est défini par la *droite des moindres carrés* :

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \,,$$

où $\hat{\beta}_1$ et $\hat{\beta}_2$ sont choisis de façon à minimiser la sommes des carrés des *résidus* $\hat{e}_i = y_i - \hat{y}_i$, soit tels que :

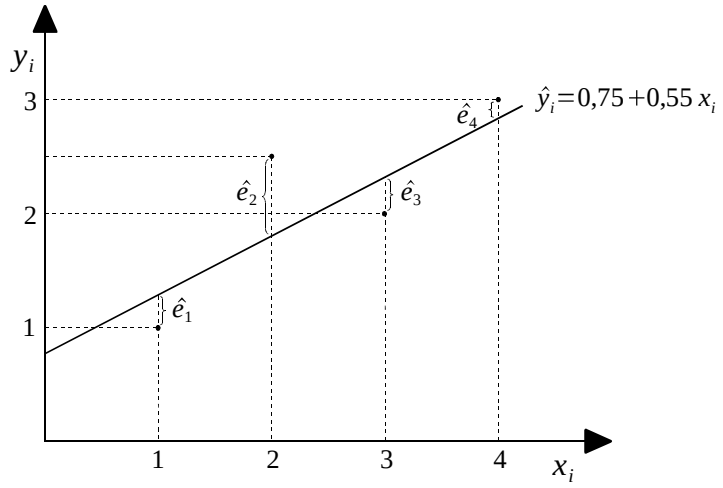
$$\left(\hat{\beta}_1, \hat{\beta}_2\right) = \text{Argmin}_{\beta_1,\beta_2} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

Par exemple, pour les données suivantes :

x_i	1	2	3	4
y_i	1	2,5	2	3

,

cela donne graphiquement :



Graphique 6 : La droite des moindres carrés

On peut obtenir analytiquement les estimateurs MCO $\hat{\beta}_1$ et $\hat{\beta}_2$ en recherchant le minimum de la fonction[9] :

$$SCR(\beta_1, \beta_2) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

Les dérivées partielles de $SCR(\beta_1, \beta_2)$ par rapport à β_1 et β_2 sont données par :

$$\begin{aligned} \frac{\partial SCR(\beta_1, \beta_2)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i) = -2 \sum_{i=1}^{n} e_i \\ \frac{\partial SCR(\beta_1, \beta_2)}{\partial \beta_2} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_1 - \beta_2 x_i) = -2 \sum_{i=1}^{n} x_i e_i \end{aligned}$$

de sorte que les *conditions de premier ordre* définissant $\hat{\beta}_1$ et $\hat{\beta}_2$ s'écrivent :

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = \sum_{i=1}^{n} \hat{e}_i = 0 \tag{2.1}$$

$$\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = \sum_{i=1}^{n} x_i \hat{e}_i = 0 \tag{2.2}$$

ou encore, en réarrangeant :

$$n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{2.3}$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{2.4}$$

[9] SCR(.) désigne la Somme des Carrés des Résidus.

Ces équations sont connues sous le nom d'équations normales.

De (2.3), on obtient :

$$\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_2 \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y} - \hat{\beta}_2 \bar{x}\,, \tag{2.5}$$

ce qui implique que *la droite des moindres carrés passe par le point moyen* $(\bar{x}, \bar{y})$ *de l'échantillon.* On notera au passage que, comme indiqué par (2.1), *la somme des résidus* $\hat{e}_i$ *est nulle.*

De (2.4), en utilisant (2.5), on obtient :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_2 \bar{x}) \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

$$\Leftrightarrow \qquad \hat{\beta}_2 - \hat{\beta}_2 \left(\frac{\bar{x} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \right) = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

soit, comme $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$:

$$\hat{\beta}_2 \left(1 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n \sum_{i=1}^{n} x_i^2} \right) = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2}$$

$$\Leftrightarrow \quad \hat{\beta}_2 \left(\frac{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}{n \sum_{i=1}^{n} x_i^2} \right) = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2}$$

et donc :

$$\hat{\beta}_2 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \tag{2.6}$$

Finalement, on notera que :

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}\bar{y}\\ &= \sum_{i=1}^{n}x_iy_i-\bar{x}(n\bar{y})-\bar{y}(n\bar{x})+n\bar{x}\bar{y}\\ &= \sum_{i=1}^{n}x_iy_i-\frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i \qquad (2.7)\end{aligned}$$

et

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2\\ &= \sum_{i=1}^{n}x_i^2-2\bar{x}(n\bar{x})+n\bar{x}^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2\\ &= \sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 \qquad (2.8)\end{aligned}$$

de sorte que (2.6) peut s'écrire :

$$\hat{\beta}_2=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{Cov_e(x_i,y_i)}{Var_e(x_i)}\,, \qquad (2.9)$$

où $Cov_e(x_i, y_i)$ désigne la covariance empirique entre x_i et y_i, et $Var_e(x_i)$ la variance empirique des x_i.

De (2.9), on peut voir que :

1- $\hat{\beta}_2$ est nul si x_i et y_i sont non corrélés (i.e., si $Cov_e(x_i, y_i) = 0$),

2- $\hat{\beta}_2$ n'est pas défini si il n'y a aucune variabilité dans les x_i (i.e., si $Var_e(x_i) = 0$).

2.2.2. L'estimateur du maximum de vraisemblance

Si, outre les hypothèses A1 - A5, on suppose aussi l'hypothèse A6 remplie, càd. que les y_i sont distribués de façon normale, on peut dériver l'estimateur du maximum de vraisemblance des paramètres inconnus β_1, β_2 et σ^2 du modèle.

Sous la normalité, il y a équivalence entre non-corrélation et indépendance statistique[10]. La densité jointe des observations $(y_1, ..., y_n)$, appelée *vraisemblance*, peut

[10] Cf. infra p. 24.

donc être décomposée comme suit[11] :

$$\begin{aligned} & f(y_1, ..., y_n | x_1, ..., x_n; \beta_1, \beta_2, \sigma^2) \\ = \; & f(y_1 | x_1; \beta_1, \beta_2, \sigma^2) \times ... \times f(y_n | x_n; \beta_1, \beta_2, \sigma^2) \\ = \; & \prod_{i=1}^{n} f(y_i | x_i; \beta_1, \beta_2, \sigma^2) \,, \end{aligned}$$

où[12]

$$f(y_i | x_i; \beta_1, \beta_2, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - \beta_1 - \beta_2 x_i}{\sigma}\right)^2}$$

Il est plus commode de manipuler le logarithme de la densité jointe que la densité jointe elle-même. En prenant le logarithme de la densité jointe des observations, on obtient la *fonction de log-vraisemblance* de l'échantillon :

$$\begin{aligned} L(\beta_1, \beta_2, \sigma^2) &= \ln f(y_1, ..., y_n | x_1, ..., x_n; \beta_1, \beta_2, \sigma^2) \\ &= \sum_{i=1}^{n} \left(-\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y_i - \beta_1 - \beta_2 x_i)^2 \right) \end{aligned}$$

Les estimateurs du maximum de vraisemblance (MV) $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{\sigma}^2$ sont définis par les valeurs de β_1, β_2 et σ^2 qui maximisent la vraisemblance[13], ou ce qui revient au même[14], la log-vraisemblance de l'échantillon :

$$\left(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2 \right) = \text{Argmax}_{\beta_1, \beta_2, \sigma^2} L(\beta_1, \beta_2, \sigma^2)$$

Les dérivées partielles de $L(\beta_1, \beta_2, \sigma^2)$ par rapport à β_1, β_2 et σ^2 sont données par :

$$\begin{aligned} \frac{\partial L(\beta_1, \beta_2, \sigma^2)}{\partial \beta_1} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i) \\ \frac{\partial L(\beta_1, \beta_2, \sigma^2)}{\partial \beta_2} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i (y_i - \beta_1 - \beta_2 x_i) \\ \frac{\partial L(\beta_1, \beta_2, \sigma^2)}{\partial \sigma^2} &= \sum_{i=1}^{n} -\frac{1}{2\sigma^2} + \sum_{i=1}^{n} \frac{1}{2\sigma^4} (y_i - \beta_1 - \beta_2 x_i)^2 \end{aligned}$$

[11] Bien que redondant lorsque les x_i sont non-stochastiques, le conditionnnement par rapport aux x_i est ici explicitement indiqué pour rappeller que l'on raisonne bien conditionnellement aux x_i observés.

[12] Pour rappel, la fonction de densité de la loi normale $N(m, \sigma^2)$ est donnée par $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$.

[13] Càd. les valeurs de β_1, β_2 et σ^2 qui rendent la plus élevée la probabilité d'observation de l'échantillon dont on dispose. Autrement dit, les valeurs de β_1, β_2 et σ^2 pour lesquelles l'échantillon dont on dispose est le plus probable d'être observé.

[14] Le logarithme étant une fonction strictement croissante, la vraisemblance et la log-vraisemblance ont par construction le même maximum par rapport à β_1, β_2 et σ^2.

de sorte que les *conditions de premier ordre* définissant $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{\sigma}^2$ s'écrivent :

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \tag{2.10}$$

$$\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \tag{2.11}$$

$$\sum_{i=1}^{n} -\frac{1}{2\hat{\sigma}^2} + \sum_{i=1}^{n} \frac{1}{2\hat{\sigma}^4}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 = 0 \tag{2.12}$$

Les conditions (2.10) et (2.11) sont identiques aux conditions (2.1) et (2.2) définissant les estimateurs MCO. On en conclut que, sous l'hypothèse de normalité, les estimateurs MV de β_1, β_2 sont identiques aux estimateurs MCO.

De (2.12), on tire :

$$-n + \sum_{i=1}^{n} \frac{1}{\hat{\sigma}^2}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$$

$$\Leftrightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2 \tag{2.13}$$

Ainsi, l'estimateur MV $\hat{\sigma}^2$ de σ^2 est simplement donné par la variance empirique des résidus.

Deux points méritent d'être épinglés :

1- Si on supposait une autre loi que la loi normale pour les y_i, les estimateurs MV et MCO ne correspondraient plus. Ils seraient différents.

2- La formulation du modèle de régression linéaire simple, et au delà du modèle de régression linéaire multiple que nous étudierons ensuite, est fortement lié à la normalité. On peut en effet montrer que si 2 variables aléatoires x et y sont distribuées selon une loi normale jointe[15] :

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N(m, \Sigma)\,,$$

on a toujours :

$$E(y|x) = a + bx \qquad \text{(i.e., une fonction linéaire de } x\text{)}$$
$$Var(y|x) = \sigma^2 \qquad \text{(i.e., une constante)}$$

Plus généralement, si k variables aléatoires sont distribuées selon une loi normale

[15] Pour un rappel concernant la loi normale multivariée, cf. infra p. 24.

jointe :

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \sim N(m, \Sigma)\,,$$

on a encore de façon semblable :

$$E(x_1|x_2, ..., x_k) = a + b_2x_2 + ... + b_kx_k$$
$$Var(x_1|x_2, ..., x_k) = \sigma^2$$

Il en est de même pour tout conditionnement par rapport à un sous-ensemble de $(x_2, ..., x_k)$. A chaque fois, l'*espérance conditionnelle* est une *fonction linéaire* (dont les paramètres varient selon l'ensemble conditionnant) et la *variance conditionnelle* est *constante*, comme dans le modèle de régression linéaire (simple ou multiple).

2.2.3. Exemple : estimation d'une fonction de consommation

Hill, Griffiths et Lim (2008) considèrent[16] un ensemble de données en coupe (x_i, y_i), où x_i désigne le revenu d'un ménage (en centaines de \$) et y_i les dépenses alimentaires de ce ménage (en \$).

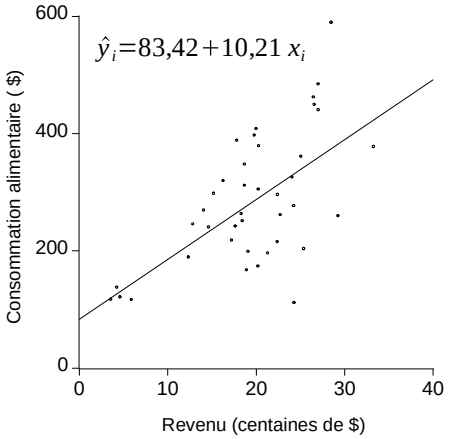
L'estimation par MCO, sur un échantillon de 40 ménages, du modèle :

$$y_i = \beta_1 + \beta_2 x_i + e_i\,,$$

donne :

$$\hat{y}_i = 83,42 + 10,21x_i$$

Graphiquement :

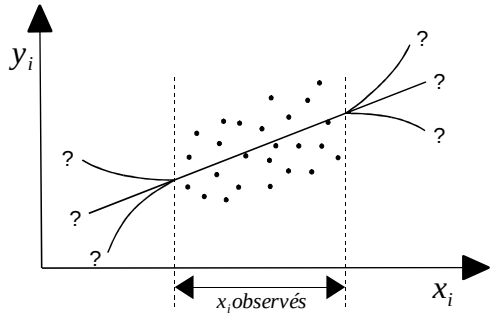


Graphique 7 : La fonction de consommation estimée

[16] Voir p. 18 et suivantes.

A. Interprétation des coefficients estimés

- $\hat{\beta}_2 =$ la *pente* : elle représente ici la propension marginale à consommer. Dans cet exemple, il est estimé qu'une augmentation du revenu de 100 \$ accroît la consommation alimentaire moyenne d'un ménage de $\dfrac{d\hat{y}_i}{dx_i} = \dfrac{d\hat{E}(y_i)}{dx_i} = 10,21\,\$$ (attention aux unités de mesure !). Pour une augmentation de 1 \$ du revenu, cela donne une augmentation de $0,1021\,\$$ de la consommation alimentaire moyenne, soit une propension marginale à consommer des biens alimentaires de $0,1021$.
- $\hat{\beta}_1 =$ l'*intercept* (ordonnée à l'origine) : il représente ici le niveau moyen *théorique* de la consommation alimentaire pour un revenu nul. Ce niveau théorique est estimé à $83,42\,\$$.
  Il est important de noter que l'intercept doit le plus souvent être interprété avec prudence car il n'y a généralement aucune observation au voisinage de $x_i = 0$. Lorsque c'est le cas (comme ici), l'intercept est généralement peu ou pas interprétable.
  De façon plus générale, il est toujours hasardeux d'utiliser la relation estimée pour évaluer (prédire) les $\hat{y}_i$ correspondants à des valeurs de x_i éloignées de celles observées dans l'échantillon. Graphiquement :



Graphique 8 : Prévisions hasardeuses

B. Prévision et élasticité

Si l'on reste dans le voisinage des x_i observés dans l'échantillon, on peut utiliser la relation estimée pour faire de la prévision. Dans notre exemple, pour un revenu de 2000 \$, soit $x_i = 20$, on peut prédire le niveau de dépense alimentaire par :

$$\hat{y}_i = 83,42 + 10,21\,(20) = 287,62\,\$$$

Les économistes sont souvent intéressés par des élasticités. La relation estimée étant linéaire, l'élasticité η de y par rapport à x :

$$\eta = E_{y,x} = \frac{\dfrac{dy}{y}}{\dfrac{dx}{x}} = \frac{dy}{dx}\frac{x}{y},$$

n'est pas constante, mais variable, en fonction de x.

Sur base de la relation estimée $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$, une estimation de l'élasticité $E_{y,x}$ pour une valeur donnée de x_i est fournie par :

$$\hat{\eta}_i = \hat{E}_{y,x} = \frac{d\hat{y}_i}{dx_i}\frac{x_i}{\hat{y}_i} = \hat{\beta}_2 \frac{x_i}{\hat{y}_i}$$

La valeur de cette élasticité $\hat{\eta}_i$ varie fonction de x_i.

Pour résumer les $\hat{\eta}_i$, il est courant de calculer une élasticité au point moyen de l'échantillon[17] $(\bar{x}, \bar{y})$ au travers de l'expression[18] :

$$\hat{\eta} = \hat{\beta}_2 \frac{\bar{x}}{\bar{y}}$$

Dans notre exemple, le point moyen de l'échantillon étant $(19,60; 283,57)$, on obtient :

$$\hat{\eta} = 10,21 \frac{19,60}{283,57} = 0,71$$

L'élasticité estimée $\hat{\eta}$ étant inférieure à 1, les dépenses alimentaires apparaissent comme un bien de nécessité (par opposition à un bien de luxe), ce qui est conforme à l'intuition.

2.3. Ecriture matricielle du modèle et de l'estimateur MCO

Pour faciliter l'examen de ses propriétés et son extension au cas de plusieurs variables explicatives, il est utile de réécrire le modèle, ses hypothèses et l'estimateur MCO sous forme matricielle.

2.3.1. Vecteurs aléatoires : notations et propriétés

On commence par établir quelques conventions de notation et propriétés relatives aux vecteurs aléatoires.

A. Cas bivarié

Soit le vecteur aléatoire bivarié :

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

[17] Rappelons que la droite des moindres carrés passe par le point moyen $(\bar{x}, \bar{y})$ de l'échantillon.

[18] Une alternative est de calculer $\hat{\eta}_i$ pour tous les points x_i de l'échantillon, puis d'en prendre la moyenne.

Par définition, on note[19] :

$$E(X) = E\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} E(x_1) \\ E(x_2) \end{bmatrix}$$

et

$$\begin{aligned} V(X) &= E\left[(X - E(X))(X - E(X))'\right] \\ &= E\left[\begin{bmatrix} x_1 - E(x_1) \\ x_2 - E(x_2) \end{bmatrix}\begin{bmatrix} x_1 - E(x_1) & x_2 - E(x_2) \end{bmatrix}\right] \\ &= E\begin{bmatrix} (x_1 - E(x_1))^2 & (x_1 - E(x_1))(x_2 - E(x_2)) \\ (x_2 - E(x_2))(x_1 - E(x_1)) & (x_2 - E(x_2))^2 \end{bmatrix} \\ &= \begin{bmatrix} Var(x_1) & Cov(x_1, x_2) \\ Cov(x_2, x_1) & Var(x_2) \end{bmatrix} \end{aligned}$$

$V(X)$ est appelée la *matrice de variance - covariance* de X. On notera que si dans l'expression de l'espérance $E(X)$, X peut être aussi bien un vecteur qu'une matrice, dans l'expression de la matrice de variance - covariance $V(X)$, X ne peut être qu'un vecteur (colonne). On remarquera encore que $V(X)$ est nécessairement une matrice symétrique.

Les opérateurs $E(X)$ et $V(X)$ ont des propriétés très intéressantes. Si :

$$\begin{aligned} A &= \text{un vecteur } k \times 1 \text{ de constantes,} \\ B &= \text{une matrice } k \times 2 \text{ de constantes,} \end{aligned}$$

alors :

$$E(A + BX) = A + BE(X) \tag{2.14}$$
$$V(A + BX) = BV(X)B' \tag{2.15}$$

La propriété (2.14) est évidente[20]. La propriété (2.15) se vérifie pour sa part aisément. Si on pose :

$$\begin{aligned} Z &= A + BX - E(A + BX) \\ &= B(X - E(X)), \end{aligned}$$

on obtient :

$$\begin{aligned} V(A + BX) &= E(ZZ') \\ &= E\left[B(X - E(X))(X - E(X))'B'\right] \\ &= BE\left[(X - E(X))(X - E(X))'\right]B' \\ &= BV(X)B' \end{aligned}$$

Illustrons ces propriétés pour le cas où $A = 0$ et $B = \begin{bmatrix} b_1 & b_2 \end{bmatrix}$. Pour ce cas,

[19] X' désigne la transposée de la matrice X, parfois aussi notée ${}^tX$.

[20] Elle découle des propriétés de base de l'espérance.

on a :

$$A + BX = \begin{bmatrix} b_1 & b_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = b_1 x_1 + b_2 x_2$$

et

$$\begin{aligned} E(A+BX) &= \begin{bmatrix} b_1 & b_2 \end{bmatrix} \begin{bmatrix} E(x_1) \\ E(x_2) \end{bmatrix} = b_1 E(x_1) + b_2 E(x_2) \\ V(A+BX) &= \begin{bmatrix} b_1 & b_2 \end{bmatrix} \begin{bmatrix} Var(x_1) & Cov(x_1,x_2) \\ Cov(x_2,x_1) & Var(x_2) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\ &= b_1^2 Var(x_1) + b_2^2 Var(x_2) + b_1 b_2 Cov(x_1,x_2) + b_2 b_1 Cov(x_2,x_1) \\ &= b_1^2 Var(x_1) + b_2^2 Var(x_2) + 2 b_1 b_2 Cov(x_1,x_2) \end{aligned}$$

On retrouve simplement les propriétés habituelles de l'espérance et de la variance d'une fonction linéaire de variables aléatoires.

B. Cas général

Soit le vecteur aléatoire $n \times 1$:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Par définition, on a de façon semblable :

$$E(X) = E \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_n) \end{bmatrix}$$

et

$$\begin{aligned} V(X) &= E\left[(X - E(X))(X - E(X))'\right] \\ &= \begin{bmatrix} Var(x_1) & Cov(x_1,x_2) & \cdots & Cov(x_1,x_n) \\ Cov(x_2,x_1) & Var(x_2) & \cdots & Cov(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_n,x_1) & Cov(x_n,x_2) & \cdots & Var(x_n) \end{bmatrix} \end{aligned}$$

On a bien entendu toujours les mêmes propriétés. Si :

$$\begin{aligned} A &= \text{un vecteur } k \times 1 \text{ de constantes,} \\ B &= \text{une matrice } k \times n \text{ de constantes,} \end{aligned}$$

alors :

$$E(A+BX) = A+BE(X) \tag{2.16}$$
$$V(A+BX) = BV(X)B' \tag{2.17}$$

C. La loi normale multivariée

Par définition, un vecteur aléatoire X de dimension $n \times 1$ suit une loi normale multivariée d'espérance m (vecteur $n \times 1$) et de matrice de variance - covariance Σ (matrice symétrique $n \times n$), notée $X \sim N(m, \Sigma)$, si sa densité jointe est donnée par[21] :

$$f(x_1, x_2, ..., x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(X-m)'\Sigma^{-1}(X-m)\right]}$$

Cette fonction de densité contient comme cas particulier la fonction de densité (univariée) d'une variable aléatoire normale $x \sim N(m, \sigma^2)$: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$.

La loi normale multivariée possède plusieurs propriétés remarquables :

1- Si $X \sim N(m, \Sigma)$ et si :

$$A = \text{un vecteur } k \times 1 \text{ de constantes,}$$
$$B = \text{une matrice } k \times n \text{ de constantes,}$$

alors :

$$Z = A + BX \sim N\left(E(Z), V(Z)\right)$$

où :

$$E(Z) = A + Bm$$
$$V(Z) = B\Sigma B'$$

En d'autres termes, une combinaison linéaire $Z = A+BX$ d'un vecteur aléatoire normal suit aussi une loi normale, l'espérance $E(Z)$ et la matrice de variance - covariance $V(Z)$ de cette loi étant simplement obtenues par application des propriétés (2.16) et (2.17).
Illustrons cette propriété par deux exemples :

a- Pour le cas où $A = 0$ et $B = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$, on a :

$$Z = x_1$$

et

$$E(Z) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix} = m_1$$

[21] $\det(\Sigma)$ désigne le déterminant de Σ.

$$V(Z) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} Var(x_1) & \cdots & Cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ Cov(x_n, x_1) & \cdots & Var(x_n) \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$= Var(x_1)$$

b- Pour le cas où $A = 0$ et $B = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$, on a :

$$Z = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

et

$$E(Z) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

$$V(Z) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} Var(x_1) & \cdots & Cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ Cov(x_n, x_1) & \cdots & Var(x_n) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} Var(x_1) & Cov(x_1, x_2) \\ Cov(x_2, x_1) & Var(x_2) \end{bmatrix}$$

2- Si $X \sim N(m, \Sigma)$ avec[22] $\Sigma = \sigma^2 I$, alors $(x_1, x_2, ..., x_n)$ sont statistiquement indépendants. En d'autres termes, *sous la normalité*, indépendance et non-corrélation sont équivalents.

3- Si $X \sim N(0, \Sigma)$, alors :

$$X'\Sigma^{-1}X \sim \chi^2(n) \tag{2.18}$$

En particulier, si $\Sigma = I$, alors :

$$X'X = \sum_{i=1}^{n} x_i^2 \sim \chi^2(n)$$

Ce cas particulier correspond à la définition standard de la loi du khi-carré[23].

4- Enfin, si $X \sim N(0, \sigma^2 I)$ et que A désigne une matrice $n \times n$ *symétrique idempotente* (càd. telle que $A' = A$ et $AA = A$) de *rang* r, alors :

$$\frac{1}{\sigma^2} X'AX \sim \chi^2(r) \tag{2.19}$$

---

[22] I désigne une *matrice identité*, càd. une matrice carrée composée de 1 sur la diagonale principale, et de 0 partout ailleurs.

[23] Définie comme la loi que suit la somme des carrés de variables aléatoires $N(0, 1)$ indépendantes (cf. l'annexe B de Hill, Griffiths et Lim (2008)).

2.3.2. Le modèle et ses hypothèses sous forme matricielle

On note :

$$X_i = \begin{bmatrix} 1 & x_i \end{bmatrix} \quad \text{et} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

Par définition, on a :

$$\begin{aligned} & y_i = \beta_1 + \beta_2 x_i + e_i \\ \Leftrightarrow \quad & y_i = X_i \beta + e_i \,, \qquad\qquad i = 1, ..., n \end{aligned}$$

En empilant les n observations de l'échantillon, on peut écrire :

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \,, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{et} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

de sorte que :

$$\begin{cases} y_1 = \beta_1 + \beta_2 x_1 + e_1 \\ y_2 = \beta_1 + \beta_2 x_2 + e_2 \\ \qquad\qquad \vdots \\ y_n = \beta_n + \beta_2 x_n + e_n \end{cases}$$

$$\Leftrightarrow \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

soit, de façon compacte :

$$Y = X\beta + e$$

Sur base de cette notation matricielle, les hypothèses A1 - A6 du modèle de régression linéaire simple peuvent s'écrire[24] :

A1	$Y = X\beta + e$
A2	$E(e) = 0 \;\Leftrightarrow\; E(Y) = X\beta$
A3 - A4	$V(e) = \sigma^2 I = V(Y)$
A5	X est non-stochastique et $\text{rg}(X) = 2$
A6	(optionnel) $e \sim N(0, \sigma^2 I) \;\Leftrightarrow\; Y \sim N(X\beta, \sigma^2 I)$

On note que, sous forme matricielle, les hypothèses A3 (concernant les variances) et A4 (concernant les covariances) sont regroupées sous la forme d'une hypothèse sur la matrice de variance - covariance de e, ou de façon équivalente de Y.

[24] $\text{rg}(X)$ désigne le rang de la matrice X.

L'hypothèse $\text{rg}(X) = 2$ requiert que les 2 colonnes de X soient linéairement indépendantes, ce qui est le cas si il n'existe pas de constantes non nulles c_1, c_2 telles que :

$$c_1 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = 0$$

Cette hypothèse est violée si il n'y a aucune variabilité dans les x_i (i.e., si $x_i =$ une constante, $\forall i$), et est bien entendu satisfaite si les x_i prennent au moins 2 valeurs distinctes.

2.3.3. L'estimateur MCO sous forme matricielle

L'estimateur MCO est défini par :

$$\begin{aligned} \left(\hat{\beta}_1, \hat{\beta}_2\right) &= \text{Argmin}_\beta \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2 \\ &= \text{Argmin}_\beta \sum_{i=1}^{n} e_i^2 \end{aligned}$$

Sous forme matricielle :

$$\begin{aligned} \hat{\beta} &= \text{Argmin}_\beta (Y - X\beta)'(Y - X\beta) \\ &= \text{Argmin}_\beta e'e \end{aligned}$$

L'estimateur MCO $\hat{\beta}$ est obtenu en recherchant le minimum de la fonction :

$$\begin{aligned} SCR(\beta) &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta \\ &= Y'Y - 2\beta'X'Y + \beta'X'X\beta \qquad (\text{car } Y'X\beta = \beta'X'Y) \end{aligned}$$

La dérivée partielle de $SCR(\beta)$ par rapport au vecteur β :

$$\frac{\partial SCR(\beta)}{\partial \beta} = \begin{bmatrix} \frac{\partial SCR(\beta)}{\partial \beta_1} \\ \frac{\partial SCR(\beta)}{\partial \beta_2} \end{bmatrix}$$

peut être obtenue en appliquant les règles de dérivation matricielle suivantes :

1- Si a est un vecteur $k \times 1$ et β aussi un vecteur $k \times 1$, alors :

$$\frac{\partial(\beta'a)}{\partial \beta} = a \tag{2.20}$$

Illustrons cette propriété pour $k = 2$, avec $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ et $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$. Pour ce

cas, on a :

$$\beta' a = \beta_1 a_1 + \beta_2 a_2$$

de sorte que :

$$\frac{\partial(\beta' a)}{\partial \beta} = \begin{bmatrix} \frac{\partial(\beta' a)}{\partial \beta_1} \\ \frac{\partial(\beta' a)}{\partial \beta_2} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = a$$

2- Si A est une matrice $k \times k$ symétrique et β encore un vecteur $k \times 1$, alors :

$$\frac{\partial(\beta' A \beta)}{\partial \beta} = 2A\beta \tag{2.21}$$

Illustrons à nouveau cette propriété pour $k = 2$, avec $a = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$ et $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$. Pour ce cas, on a :

$$\begin{aligned} \beta' A \beta &= \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \\ &= a_{11}\beta_1^2 + 2a_{12}\beta_1\beta_2 + a_{22}\beta_2^2 \end{aligned}$$

de sorte que :

$$\begin{aligned} \frac{\partial(\beta' A \beta)}{\partial \beta} &= \begin{bmatrix} \frac{\partial(\beta' A \beta)}{\partial \beta_1} \\ \frac{\partial(\beta' A \beta)}{\partial \beta_2} \end{bmatrix} = \begin{bmatrix} 2a_{11}\beta_1 + 2a_{12}\beta_2 \\ 2a_{12}\beta_1 + 2a_{22}\beta_2 \end{bmatrix} \\ &= 2 \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = 2A\beta \end{aligned}$$

La matrice $X'X$ étant une matrice symétrique :

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = (X'X)' ,$$

par application des règles de calcul (2.20) et (2.21), on obtient :

$$\begin{aligned} \frac{\partial SCR(\beta)}{\partial \beta} &= \frac{\partial\left(-2\beta' X'Y\right)}{\partial \beta} + \frac{\partial\left(\beta' X'X\beta\right)}{\partial \beta} \\ &= -2X'Y + 2X'X\beta \end{aligned}$$

de sorte que la *condition de premier ordre* définissant $\hat{\beta}$ s'écrit :

$$X'(Y - X\hat{\beta}) = X'\hat{e} = 0 \quad \Leftrightarrow \quad X'X\hat{\beta} = X'Y , \tag{2.22}$$

soit, sous forme détaillée :

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix},$$

ce qui n'est rien d'autre que les *équations normales* (2.3) et (2.4) obtenues à la Section 2.2.1.

Finalement, de (2.22), on obtient la forme matricielle de l'estimateur MCO :

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{2.23}$$

Deux remarques méritent d'être faites :

1- L'hypothèse $\text{rg}(X) = 2$ du modèle assure l'existence de l'estimateur MCO $\hat{\beta}$. En effet, on peut montrer que, pour toute matrice A, $\text{rg}(A) = \text{rg}(A') = \text{rg}(AA') = \text{rg}(A'A)$. On a donc $\text{rg}(X'X) = 2$ (rang plein), ce qui implique que $X'X$ est non singulière, et donc inversible.

2- On peut pareillement dériver sous forme matricielle les estimateurs MV de β et de σ^2. L'estimateur MV de β est évidemment identique à l'estimateur MCO (2.23). L'estimateur MV de σ^2 peut pour sa part être écrit sous forme matricielle comme :

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i^2 = \frac{\hat{e}'\hat{e}}{n}, \tag{2.24}$$

où $\hat{e} = Y - X\hat{\beta}$.

2.3.4. Résultats complémentaires

On détaille ci-dessous quelques résultats complémentaires qui seront utiles dans la suite :

1- Les *valeurs estimées* (prédites) de Y par le modèle sont données par :

$$\hat{Y} = X\hat{\beta} = \underbrace{X(X'X)^{-1}X'}_{P_X} Y = P_X Y \tag{2.25}$$

Lorsqu'on prémultiplie par P_X un *vecteur* a quelconque de dimension $n \times 1$, on obtient un vecteur $n \times 1$ donnant les valeurs estimées $\hat{a}$ de la régression de a sur X. De façon plus générale, lorsqu'on prémultiplie par P_X une *matrice* A quelconque de dimension $n \times l$, on obtient une matrice $\hat{A}$ de dimension $n \times l$ dont les colonnes donnent les valeurs estimées des régressions des différentes colonnes de A sur X.
Si $A = X$, on a simplement :

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

Autrement dit, les valeurs estimées des régressions des différentes colonnes de X sur X sont tout simplement égales à X.
La matrice P_X possède des propriétés remarquables :

a- P_X est symétrique : $P_X = P_X'$

b- P_X est idempotente : $P_X P_X = X(X'X)^{-1}X'X(X'X)^{-1}X' = P_X$

2- Les *résidus* du modèle sont donnés par :

$$\hat{e} = Y - X\hat{\beta} = Y - P_X Y = \underbrace{(I - P_X)}_{M_X} Y = M_X Y \tag{2.26}$$

Lorsqu'on prémultiplie par M_X un *vecteur* a quelconque de dimension $n \times 1$, on obtient un vecteur $n \times 1$ donnant les résidus $\hat{e}$ de la régression de a sur X. De façon plus générale, lorsqu'on prémultiplie par M_X une *matrice* A quelconque de dimension $n \times l$, on obtient une matrice $\hat{E}$ de dimension $n \times l$ dont les colonnes donnent les résidus des régressions des différentes colonnes de A sur X.
Si $A = X$, on a :

$$\hat{E} = M_X X = (I - P_X)\, X = X - X = 0$$

Autrement dit, les résidus des régressions des différentes colonnes de X sur X sont tout simplement nuls.
La matrice M_X possède également des propriétés remarquables :

a- M_X est symétrique : $M_X = (I - P_X) = (I - P_X)' = M_X'$

b- M_X est idempotente : $M_X M_X = (I - P_X)\,(I - P_X) = I - P_X - P_X + P_X P_X = I - P_X = M_X$

3- On a d'une part :

$$\hat{e} = M_X Y\,,$$

et d'autre part :

$$Y = X\beta + e\,,$$

de sorte que :

$$\hat{e} = M_X(X\beta + e) = M_X X\beta + M_X e\,,$$

et comme $M_X X = 0$:

$$\hat{e} = M_X e$$

Les résidus et l'erreur aléatoire (non observable) sont reliés par la matrice M_X.

Chapitre 3

Propriétés de l'estimateur MCO

L'estimateur MCO est donné par $\hat{\beta} = (X'X)^{-1} X'Y$. Si X est (par commodité) supposé fixe, Y est un vecteur aléatoire, dont les observations sont regardées comme une réalisation particulière, pour un échantillon particulier.

$\hat{\beta}$ étant une fonction de Y, c'est aussi une variable aléatoire : sa valeur varie d'un échantillon à l'autre, càd. d'une réalisation à l'autre du vecteur aléatoire Y.

Notons que l'*estimation* obtenue de l'application de la formule $\hat{\beta} = (X'X)^{-1} X'Y$ à un échantillon particulier ne possède en tant que telle aucune propriété statistique. C'est l'*estimateur*, en tant que règle de décision, qui possède des propriétés statistiques.

3.1. La distribution d'échantillonnage de l'estimateur MCO

L'estimateur MCO étant une variable aléatoire, il possède une distribution dont les caractéristiques peuvent être étudiées. Ces caractéristiques (espérance, variance, ...) nous renseignent sur la *qualité* de la règle de décision qu'est l'estimateur MCO.

La distribution jointe $f(\hat{\beta}_1, \hat{\beta}_2)$ de $\hat{\beta}$ est appelée la *distribution d'échantillonnage* de l'estimateur MCO.

De façon générale, la distribution d'échantillonnage *exacte* de $\hat{\beta}$ dépend :

1- des x_i,

2- des paramètres du modèle : $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ et σ^2,

3- de la taille d'échantillon n,

4- de la loi des y_i (au delà de leur deux premiers moments).

Sauf cas particulier (cf. infra), le calcul de la distribution d'échantillonnage exacte de $\hat{\beta}$ est très malaisé. Par contre, ses deux premiers moments (espérance et matrice de variance - covariance), peuvent aisément être obtenus.

3.1.1. L'espérance de $\hat{\beta}$

Soit l'estimateur MCO :

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

De l'hypothèse A1 $Y = X\beta + e$, on obtient :

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} X'(X\beta + e) \\ &= (X'X)^{-1} X'X\beta + (X'X)^{-1} X'e \end{aligned}$$

soit :

$$\hat{\beta} = \beta + (X'X)^{-1} X'e \tag{3.1}$$

Par ailleurs, de l'hypothèse A5 qui assure que X est non-stochastique, on a :

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

de sorte que, de l'hypothèse A2 $E(e) = 0$, on obtient finalement :

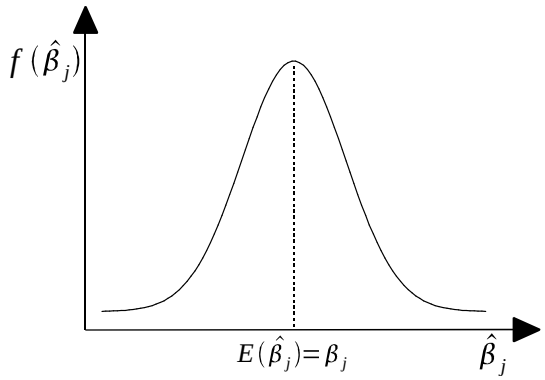
$$E(\hat{\beta}) = \begin{bmatrix} E(\hat{\beta}_1) \\ E(\hat{\beta}_2) \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \beta$$

On dit que $\hat{\beta}$ est un estimateur *non biaisé* de β.

Ainsi, sous la condition que les hypothèses A1, A2 et A5 sont correctes[25], la *tendance centrale* de la distribution d'échantillonnage de l'estimateur $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$ est bien 'calée' sur la vraie valeur β du vecteur de paramètres que l'on cherche à estimer, ce qui est évidemment une bonne nouvelle.

Graphiquement, en termes de distributions marginales $f(\hat{\beta}_1)$ et $f(\hat{\beta}_2)$ impliquées par la distribution jointe $f(\hat{\beta}_1, \hat{\beta}_2)$, on a donc :

[25] Notez que ni l'hypothèse A3 - A4, ni l'hypothèse A6, ne sont invoquées.



Graphique 9 : Distribution d'échantillonnage de $\hat{\beta}_j$ $(j = 1, 2)$

3.1.2. La matrice de variance - covariance de $\hat{\beta}$

Sous les hypothèses A1, A2 et A5, on a :

$$\hat{\beta} = \beta + (X'X)^{-1} X'e \quad \text{et} \quad E(\hat{\beta}) = \beta$$

En y ajoutant l'hypothèse A3 - A4 $V(e) = E(ee') = \sigma^2 I$, on obtient :

$$\begin{aligned}
V(\hat{\beta}) &= \begin{bmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) \\ Cov(\hat{\beta}_2, \hat{\beta}_1) & Var(\hat{\beta}_2) \end{bmatrix} \\
&= E\left[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'\right] \\
&= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] && (\text{car } E(\hat{\beta}) = \beta) \\
&= E\left[(X'X)^{-1} X'ee'X (X'X)^{-1}\right] && (\text{car } \hat{\beta} - \beta = (X'X)^{-1} X'e)) \\
&= (X'X)^{-1} X'E(ee')X (X'X)^{-1} && (\text{car } X \text{ fixe}) \\
&= \sigma^2 (X'X)^{-1} X'X (X'X)^{-1} && (\text{car } E(ee') = \sigma^2 I)
\end{aligned}$$

soit, sous la condition que les hypothèses A1 à A5 sont correctes[26], finalement :

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \tag{3.2}$$

On peut montrer que, sous forme détaillée, cela donne :

$$Var(\hat{\beta}_1) = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \tag{3.3}$$

[26] Notez que l'hypothèse A6 n'est pas invoquée.

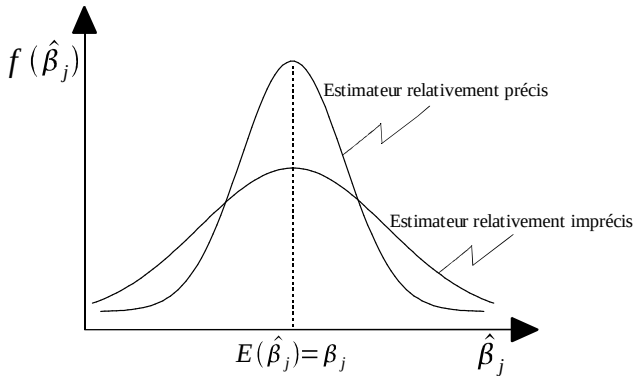
$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.4}$$

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = Cov(\hat{\beta}_2, \hat{\beta}_1) = \sigma^2 \left[\frac{-\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \tag{3.5}$$

Ces expressions peuvent être vérifiées en utilisant la propriété (2.8) établie à la Section 2.2.1, et le fait que, pour une matrice 2×2, on a :

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

Les variances d'échantillonnage $Var(\hat{\beta}_1)$ et $Var(\hat{\beta}_2)$ indiquent la *dispersion* des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ autour de leur espérance $E(\hat{\beta}_1)$ et $E(\hat{\beta}_2)$, soit comme $E(\hat{\beta}_1) = \beta_1$ et $E(\hat{\beta}_2) = \beta_2$, autour des vraies valeurs β_1 et β_2 que l'on cherche à estimer. Plus ces variances sont faibles, plus ces estimateurs sont précis. Graphiquement :



Graphique 10 : Précision de $\hat{\beta}_j$ $(j = 1, 2)$

Il est important de noter que l'absence de biais et des variances d'échantillonnage faibles ne garantissent pas que dans un *échantillon particulier*, les $\hat{\beta}_j$ estimés seront nécessairement proches de leur vraie valeur β_j que l'on cherche à estimer. Cependant, plus les variances d'échantillonnage sont faibles, plus la probabilité qu'il en soit ainsi est grande.

La covariance d'échantillonnage $Cov(\hat{\beta}_1, \hat{\beta}_2)$ indique la mesure dans laquelle les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ tendent à s'écarter *de concert ou non* de leur espérance $E(\hat{\beta}_1)$ et $E(\hat{\beta}_2)$, soit comme $E(\hat{\beta}_1) = \beta_1$ et $E(\hat{\beta}_2) = \beta_2$, la mesure dans laquelle ils tendent à s'écarter de concert ou non des vraies valeurs β_1 et β_2 que l'on cherche à estimer.

3.1.3. Les facteurs déterminant $V(\hat{\beta})$

Sur base des expressions détaillées (3.3), (3.4) et (3.5), on peut voir que les

facteurs déterminant $V(\hat{\beta})$ sont :

1- La variance σ^2 du terme d'erreur :

$$\text{si } \sigma^2 \nearrow \text{ , alors } Var(\hat{\beta}_1),\ Var(\hat{\beta}_2) \text{ et } |Cov(\hat{\beta}_1, \hat{\beta}_2)| \nearrow$$

Autrement dit, plus la dispersion des y_i autour de la droite de régression $E(y_i) = \beta_1 + \beta_2 x_i$ est grande, moins la précision d'estimation est grande.

2- La dispersion de la variable explicative x_i :

$$\text{si } \sum_{i=1}^{n}(x_i - \bar{x})^2 \nearrow \text{ , alors } Var(\hat{\beta}_1),\ Var(\hat{\beta}_2) \text{ et } |Cov(\hat{\beta}_1, \hat{\beta}_2)| \searrow$$

Autrement dit, plus la dispersion des x_i est grande, plus la précision d'estimation est grande.

3- La taille n de l'échantillon :

$$\text{si } n \nearrow \text{ , } \sum_{i=1}^{n}(x_i - \bar{x})^2 \nearrow \text{ , alors } Var(\hat{\beta}_1),\ Var(\hat{\beta}_2) \text{ et } |Cov(\hat{\beta}_1, \hat{\beta}_2)| \searrow$$

Autrement dit, plus la taille d'échantillon est grande, plus la précision d'estimation est grande.

4- La moyenne $\bar{x}$ des x_i (son éloignement par rapport à 0) :

$$\begin{aligned} \text{si } |\bar{x}| \nearrow \text{ , } \quad & \text{alors } Var(\hat{\beta}_1) \text{ et } |Cov(\hat{\beta}_1, \hat{\beta}_2)| \nearrow \\ & \text{mais } Var(\hat{\beta}_2) \text{ reste inchangée} \end{aligned}$$

Autrement dit, plus la moyenne des x_i est éloignée de 0, moins la précision d'estimation de β_1 est grande, la précision d'estimation de β_2 restant inchangée. On notera par ailleurs que :

$$\begin{aligned} Cov(\hat{\beta}_1, \hat{\beta}_2) &> 0 \quad \text{si } \bar{x} < 0 \\ \text{et } Cov(\hat{\beta}_1, \hat{\beta}_2) &< 0 \quad \text{si } \bar{x} > 0 \end{aligned}$$

3.2. Le théorème Gauss - Markov

Un bon estimateur est un estimateur qui délivre, avec une probabilité élevée, des valeurs proches de la valeur que l'on cherche à estimer. Autrement dit, un bon estimateur est un estimateur dont la distribution d'échantillonnage est, d'une part, *centrée* sur la valeur que l'on cherche à estimer, et d'autre part, *peu dispersée* autour de cette valeur.

L'estimateur MCO étant non biaisé (i.e., $E(\hat{\beta}) = \beta$), il est bien centré sur la valeur que l'on cherche à estimer. Par ailleurs, sa dispersion est donnée par sa matrice de variance - covariance $V(\hat{\beta})$.

Est-il possible de trouver un autre estimateur non biaisé de β, dont la dispersion,

càd. la matrice de variance - covariance, serait plus petite que celle de l'estimateur MCO ? Autrement dit, est-il possible de trouver un meilleur estimateur de β que l'estimateur MCO ?

Le théorème Gauss - Markov indique que non, à tout le moins si on se restreint à considérer la classe des estimateurs linéaires (et non biaisés) de β.

3.2.1. Estimateurs linéaires de β

L'estimateur MCO de β peut s'écrire :

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} X'Y \\ &= WY \end{aligned}$$

où $W = (X'X)^{-1} X'$ est une matrice $2 \times n$. Sous forme détaillée :

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} w_{1i} y_i \\ \sum_{i=1}^{n} w_{2i} y_i \end{bmatrix}$$

On voit ainsi que $\hat{\beta}_1$ et $\hat{\beta}_2$ ne sont rien d'autre que des combinaisons linéaires des y_i, les éléments w_{1i} et w_{2i} de ces combinaisons linéaires étant fonction de X, qui est supposé non-stochastique. On dit que $\hat{\beta}$ est un estimateur *linéaire* (et par ailleurs non biaisé) de β.

De façon générale, tout estimateur *linéaire* de β s'écrit :

$$\hat{\beta}^* = AY$$

où A est une matrice $2 \times n$ dont les éléments ne dépendent pas des y_i (stochastiques), mais qui peuvent dépendre des x_i (non-stochastiques). On obtient l'estimateur MCO en prenant $A = (X'X)^{-1} X'$.

Un estimateur linéaire $\hat{\beta}^*$ n'est pas nécessairement non biaisé. En effet, sous les hypothèses A1, A2 et A5, on a :

$$\begin{aligned} E(\hat{\beta}^*) &= E(AY) \\ &= E\left[A(X\beta + e)\right] && (\text{car } Y = X\beta + e) \\ &= AX\beta + AE(e) && (\text{car } A \text{ et } X \text{ fixes}) \\ &= AX\beta && (\text{car } E(e) = 0) \end{aligned}$$

On peut cependant voir que cet estimateur sera non biaisé si A est tel que :

$$AX = I$$

Notez que, dans le cas de l'estimateur MCO, on a bien $AX = (X'X)^{-1} X'X = I$.

3.2.2. Le meilleur estimateur linéaire sans biais de β

Le théorème Gauss - Markov peut s'énoncer comme suit :

Sous les hypothèses A1, A2, A3 - A4 et A5, l'estimateur MCO de β est l'estimateur qui possède la plus petite (au sens matriciel) matrice de variance - covariance parmi tous les estimateurs linéaires et sans biais de β. C'est le meilleur estimateur linéaire sans biais de β.

Voici la preuve de ce résultat. Un estimateur linéaire de β s'écrit :

$$\hat{\beta}^* = AY$$

En posant $C = A - (X'X)^{-1}X'$, on peut réécrire $\hat{\beta}^*$ comme :

$$\begin{aligned}
\hat{\beta}^* &= \left(A - (X'X)^{-1}X' + (X'X)^{-1}X'\right) Y \\
&= \left(C + (X'X)^{-1}X'\right) Y \\
&= \left(C + (X'X)^{-1}X'\right) (X\beta + e) \qquad (\text{car } Y = X\beta + e) \\
&= CX\beta + (X'X)^{-1}X'X\beta + \left(C + (X'X)^{-1}X'\right) e \\
&= (CX + I)\,\beta + \left(C + (X'X)^{-1}X'\right) e
\end{aligned}$$

On sait de la section précédente que l'estimateur $\hat{\beta}^*$ est non biaisé si $AX = I$, soit, puisque $A = C + (X'X)^{-1}X'$, si :

$$\left(C + (X'X)^{-1}X'\right) X = I \quad \Leftrightarrow \quad CX = 0$$

Sous la restriction $CX = 0$, on a :

$$\hat{\beta}^* = \beta + \left(C + (X'X)^{-1}X'\right) e\,,$$

de sorte qu'on a bien :

$$\begin{aligned}
E(\hat{\beta}^*) &= E\left[\beta + \left(C + (X'X)^{-1}X'\right) e\right] \\
&= \beta + \left(C + (X'X)^{-1}X'\right) E(e) \qquad (\text{car } C \text{ et } X \text{ fixes}) \\
&= \beta \qquad (\text{car } E(e) = 0)\,,
\end{aligned}$$

et que la matrice de matrice de variance - covariance de $\hat{\beta}^*$ est donnée par :

$$\begin{aligned}
V(\hat{\beta}^*) &= E\left[(\hat{\beta}^* - E(\hat{\beta}^*))(\hat{\beta}^* - E(\hat{\beta}^*))'\right] \\
&= E\left[(\hat{\beta}^* - \beta)(\hat{\beta}^* - \beta)'\right] \qquad (\text{car } E(\hat{\beta}^*) = \beta)\,,
\end{aligned}$$

soit, puisque $\hat{\beta}^* - \beta = (C + (X'X)^{-1}X')\, e$:

$$\begin{aligned}
V(\hat{\beta}^*) &= E\left[\left(C + (X'X)^{-1}X'\right) ee' \left(C' + X(X'X)^{-1}\right)\right] \\
&= \left(C + (X'X)^{-1}X'\right) E(ee') \left(C' + X(X'X)^{-1}\right) \qquad (\text{car } C \text{ et } X \text{ fixes}) \\
&= \sigma^2 \left[\left(C + (X'X)^{-1}X'\right) \left(C' + X(X'X)^{-1}\right)\right] \qquad (\text{car } E(ee') = \sigma^2 I) \\
&= \sigma^2 \left[CC' + (X'X)^{-1}X'C' + CX(X'X)^{-1} + (X'X)^{-1}X'X(X'X)^{-1}\right] \\
&= \sigma^2 \left[CC' + (X'X)^{-1}\right] \qquad (\text{car } CX = X'C' = 0)\,,
\end{aligned}$$

et donc, comme la matrice de variance - covariance de l'estimateur MCO $\hat{\beta}$ est égale à $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$:

$$V(\hat{\beta}^*) = V(\hat{\beta}) + \sigma^2 CC' \tag{3.6}$$

La matrice CC' est nécessairement *semi-définie positive*[27], càd. telle que pour tout vecteur a de dimension 2×1, $a'CC'a \geq 0$. En effet, $a'C$ est un vecteur $1 \times n$ et $a'CC'a$ n'est rien d'autre que la somme des carrés des éléments de ce vecteur, qui est nécessairement supérieure ou égale à 0.

La matrice CC' étant nécessairement semi-définie positive, on a l'*inégalité matricielle*[28] :

$$V(\hat{\beta}^*) \geq V(\hat{\beta}) \tag{3.7}$$

Autrement dit, sous les hypothèses A1, A2, A3 - A4 et A5, la matrice de variance - covariance de tout autre estimateur linéaire non biaisé $\hat{\beta}^*$ de β excède (au sens matriciel) la matrice de variance - covariance de l'estimateur MCO $\hat{\beta}$.

Les relations (3.6) et (3.7) impliquent que, pour tout $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$, on a :

$$\begin{aligned}
Var(a'\hat{\beta}^*) &= Var(a_1\hat{\beta}_1^* + a_2\hat{\beta}_2^*) \\
&= a'V(\hat{\beta}^*)a \\
&= a'V(\hat{\beta})a + \underbrace{\sigma^2 a'CC'a}_{\geq 0} \\
&\geq a'V(\hat{\beta})a = Var(a'\hat{\beta})
\end{aligned} \tag{3.8}$$

Ainsi, la variance de toute combinaison linéaire de $\hat{\beta}^*$ est toujours supérieure ou égale à la variance de la même combinaison linéaire de $\hat{\beta}$. Pour estimer une telle combinaison linéaire de β, il vaut donc mieux utiliser l'estimateur MCO $\hat{\beta}$ qu'un autre estimateur linéaire non biaisé $\hat{\beta}^*$. Notons que (3.8) implique en particulier que :

$$Var(\hat{\beta}_1^*) \geq Var(\hat{\beta}_1) \quad \text{et} \quad Var(\hat{\beta}_2^*) \geq Var(\hat{\beta}_2)$$

Il est important de remarquer que la théorème Gauss - Markov assure que $\hat{\beta}$ est

[27] Pour rappel, une matrice M est semi-définie positive si la forme quadratique $x'Mx \geq 0$, pour tout x.

[28] Pour rappel, au sens matriciel, $M_1 \geq M_2$ si et seulement si $M_1 - M_2$ est une matrice semi-définie positive.

le meilleur estimateur (variance minimale) parmi seulement les estimateurs *linéaires* et *non biaisés* de β, pas parmi tous les estimateurs possibles. Cependant, si aux hypothèses A1, A2, A3 - A4 et A5, on ajoute l'hypothèse optionnelle de normalité A6, on peut montrer que $\hat{\beta}$ est alors le meilleur estimateur (variance minimale) parmi tous les estimateurs *non biaisés*, qu'ils soient *linéaires ou non*. Sous cette hypothèse supplémentaire, $\hat{\beta}$ est le meilleur estimateur sans biais de β.

On notera finalement que le théorème Gauss - Markov s'applique à la règle de décision que constitue l'estimateur MCO $\hat{\beta}$, pas à une estimation obtenue pour un échantillon particulier.

3.3. La distribution d'échantillonnage de $\hat{\beta}$ sous l'hypothèse de normalité

En s'appuyant sur les hypothèses A1, A2, A3 - A4 et A5, on a pu obtenir l'espérance et la matrice de variance - covariance de $\hat{\beta}$, càd. les deux premiers moments de la distribution d'échantillonnage jointe $f(\hat{\beta}_1, \hat{\beta}_2)$ de $\hat{\beta}$.

Sous l'hypothèse additionnelle de normalité A6, la distribution d'échantillonnage de $\hat{\beta}$ est entièrement déterminée : c'est une loi normale (bivarié). En effet, $\hat{\beta}$ est une combinaison linéaire de Y, et on sait qu'une combinaison linéaire d'un vecteur distribué de façon normale suit également une loi normale (cf. Section 2.3.1). On a donc, sous les hypothèses A1 à A6 :

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1}) \,,$$

et en particulier :

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 q_{jj}), \qquad j = 1, 2,$$

où $q_{jj} = [(X'X)^{-1}]_{jj}$ désigne l'élément (j, j) de la matrice $(X'X)^{-1}$.

3.4. Propriétés de $\hat{\beta}$ en grand échantillon : convergence et normalité asymptotique

Les propriétés statistiques de $\hat{\beta}$ obtenues ci-dessus (espérance, matrice de variance - covariance, distribution sous l'hypothèse de normalité) sont des propriétés valables en *échantillon fini*, càd. quelle que soit la taille n de l'échantillon. On s'intéresse maintenant aux propriétés asymptotiques de $\hat{\beta}$, càd. lorsque $n \to \infty$.

3.4.1. Convergence

On a vu que, sous les hypothèses A1 à A5 et quelle que soit la taille n d'échantillon,

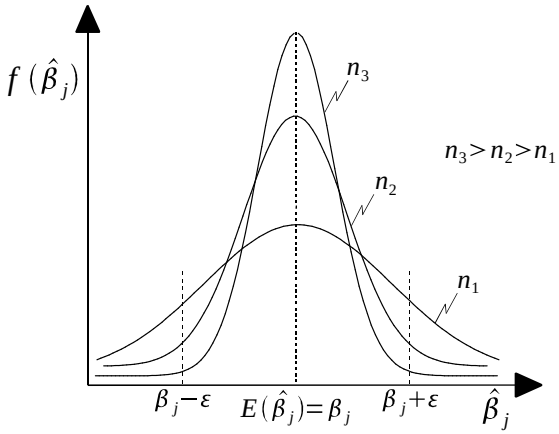
on a :

$$E(\hat{\beta}_1) = \beta_1 \quad \text{et} \quad E(\hat{\beta}_2) = \beta_2 \,,$$

et que, lorsque n augmente :

$$Var(\hat{\beta}_1),\ Var(\hat{\beta}_2) \text{ et } |Cov(\hat{\beta}_1, \hat{\beta}_2)| \text{ diminuent}$$

On en déduit qu'à mesure que n augmente, la distribution d'échantillonnage jointe $f(\hat{\beta}_1, \hat{\beta}_2)$, et par voie de conséquence les distributions marginales associées $f(\hat{\beta}_1)$ et $f(\hat{\beta}_2)$, sont de plus en plus concentrées autour de leur vraie valeur β_1 et β_2. Graphiquement :



Graphique 11 : Distribution de $\hat{\beta}_j$ $(j = 1, 2)$ lorsque $n \nearrow$

De façon générale, on dit qu'un estimateur $\hat{\theta}$ (pas nécessairement non biaisé) *converge en probabilité* vers θ, ce qu'on note $\hat{\theta} \xrightarrow{p} \theta$ ou plim $\hat{\theta} = \theta$, si :

$$\lim_{n\to\infty} I\!P\left(|\hat{\theta} - \theta| < \varepsilon\right) = 1 \,,$$

où ε est un nombre positif arbitrairement petit.

Autrement dit, un estimateur $\hat{\theta}$ converge en probabilité vers une certaine valeur θ si la probabilité que $\hat{\theta}$ prenne une valeur aussi proche que l'on veut de θ tend vers 1 lorsque $n \to \infty$.

Des *conditions suffisantes* pour que $\hat{\theta} \xrightarrow{p} \theta$ sont :

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta \quad \text{et} \quad \lim_{n\to\infty} Var(\hat{\theta}) = 0$$

Sous les hypothèses A1 à A5, ces conditions sont manifestement remplies pour $\hat{\beta}_1$ et $\hat{\beta}_2$. Sous les hypothèses A1 à A5, on a donc :

$$\hat{\beta} \xrightarrow{p} \beta$$

Ainsi, sous les hypothèses A1 à A5, la probabilité que $\hat{\beta}$ soit aussi proche que

l'on veut de la vraie valeur β, càd. d'obtenir pour un échantillon particulier une estimation aussi proche que l'on veut de β, tend vers 1 lorsque la taille d'échantillon n tend vers l'infini.

3.4.2. Distribution asymptotique

On a vu que sous les hypothèses A1 à A6, càd. en particulier sous l'hypothèse de normalité, on a en *échantillon fini* (quel que soit n) :

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

On peut montrer que le même résultat tient *asymptotiquement*, càd. en grand échantillon, lorsque $n \to \infty$, sous les seules hypothèses A1 à A5 (sans invoquer A6 donc).

En d'autres termes, quelle que soit la loi des y_i (au-delà de leurs deux premiers moments), donc même lorsque les y_i n'ont *pas* une distribution normale, $\hat{\beta}$ a une distribution d'échantillonnage jointe $f(\hat{\beta}_1, \hat{\beta}_2)$ qui, à mesure que n augmente, est non seulement de plus en plus concentrée autour de la vraie valeur β, mais encore qui s'approche de plus en plus de la forme en cloche typique d'une loi normale (bivariée).

Formellement, sous les hypothèses A1 à A5, on a :

$$\left[V(\hat{\beta})\right]^{-\frac{1}{2}} \left(\hat{\beta} - \beta\right) \xrightarrow{d} N(0, I)\,,$$

où $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ et '$\xrightarrow{d}$' indique une convergence en distribution[29], soit, exprimé sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand :

$$\hat{\beta} \approx N(\beta, \sigma^2(X'X)^{-1})$$

Cette approximation est souvent, mais pas toujours, raisonnable dès $n > 30$. Elle est évidemment d'autant plus raisonnable que n est grand.

3.5. Estimation de σ^2 et de $V(\hat{\beta})$

Les résultats de distribution d'échantillonnage de $\hat{\beta}$ obtenus ci-dessus nous permettront, dès le chapitre suivant, d'établir des procédures d'inférence statistique : intervalle de confiance et tests d'hypothèses relatifs à β, et ensuite intervalles de prévision.

Au préalable, il nous faut encore obtenir un *estimateur* de la matrice de variance-covariance $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ de l'estimateur MCO $\hat{\beta}$, ce qui nécessite de trouver un *estimateur* de la variance σ^2 du terme d'erreur du modèle.

[29] On dit aussi convergence en loi.

3.5.1. Estimateur de σ^2

Etant donné que :

$$\sigma^2 = E(e_i^2) \quad \text{où } e_i = y_i - \beta_1 - \beta_2 x_i\,, \qquad i = 1, ..., n$$

il semble, par analogie, naturel de considérer comme estimateur $\hat{\sigma}^2$ de σ^2 :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{\hat{e}'\hat{e}}{n} \quad \text{où } \hat{e}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\,,$$

càd. de remplacer l'espérance $E(.)$ par sa contrepartie empirique $\frac{1}{n}\sum_{i=1}^{n}(.)$ et e_i, qui est non observable, par son estimateur $\hat{e}_i$. Cet estimateur est l'estimateur MV obtenu à la Section 2.2.2.

Bien qu'on puisse, sous les hypothèses A1 à A5, montrer qu'il est *convergent*, cet estimateur $\hat{\sigma}^2$ est un estimateur *biaisé* de σ^2. En effet, on a[30] :

$$\begin{aligned}
E(\hat{\sigma}^2) &= \frac{1}{n}E(\hat{e}'\hat{e}) \\
&= \frac{1}{n}E(e'M_X'M_Xe) && (\text{car } \hat{e} = M_Xe, \text{ cf. Section 2.3.4}) \\
&= \frac{1}{n}E(e'M_Xe) && (\text{car } M_X' = M_X \text{ et } M_XM_X = M_X) \\
&= \frac{1}{n}E\left[\text{tr}\left(e'M_Xe\right)\right] && (\text{car } \text{tr}(a) = a, \text{ pour } a \text{ scalaire}) \\
&= \frac{1}{n}E\left[\text{tr}\left(M_Xee'\right)\right] && (\text{car } \text{tr}(AB) = \text{tr}(BA), \text{ si } AB \text{ et } BA \text{ existent}) \\
&= \frac{1}{n}\text{tr}\left[E\left(M_Xee'\right)\right] && (\text{car } E\left[\text{tr}(.)\right] = \text{tr}\left[E(.)\right]) \\
&= \frac{1}{n}\text{tr}\left[M_XE(ee')\right] && (\text{car } M_X \text{ fixe}) \\
&= \frac{\sigma^2}{n}\text{tr}\left[M_X\right] && (\text{car } E(ee') = \sigma^2 I) \\
&= \frac{\sigma^2}{n}\text{tr}\left[I - X(X'X)^{-1}X'\right] && (\text{car } M_X = I - X(X'X)^{-1}X') \\
&= \frac{\sigma^2}{n}\left[\text{tr}(I) - \text{tr}(X(X'X)^{-1}X')\right] && (\text{car } \text{tr}(A+B) = \text{tr}(A) + \text{tr}(B)) \\
&= \frac{\sigma^2}{n}\left[n - \text{tr}((X'X)^{-1}X'X)\right] && (\text{car } \text{tr}(AB) = \text{tr}(BA)) \\
&= \frac{\sigma^2}{n}\left[n-2\right] && (\text{car } \text{tr}((X'X)^{-1}X'X) = \text{tr}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 2)
\end{aligned}$$

[30] Ci-dessous, $\text{tr}(M)$ désigne la trace de la matrice (carrée) M, càd. la somme de ses éléments diagonaux.

On voit ainsi que $\hat{\sigma}^2$ sous-estime systématiquement σ^2 :

$$E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2 < \sigma^2 \tag{3.9}$$

De (3.9), on peut aisément déduire un estimateur *convergent* et *non biaisé* de σ^2 :

$$\hat{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{\hat{e}'\hat{e}}{n-2} \tag{3.10}$$

Pour cet estimateur $\hat{s}^2$, on a en effet :

$$E(\hat{s}^2) = E(\frac{n}{n-2}\hat{\sigma}^2) = \frac{n}{n-2}E(\hat{\sigma}^2) = \frac{n}{n-2}\frac{n-2}{n}\sigma^2 = \sigma^2$$

Trois points méritent d'être soulignés :

1- Le facteur $(n-2)$ est généralement appelé *nombre de degrés de liberté*, 2 étant le nombre de paramètres préalablement estimés pour obtenir les $\hat{e}_i$.

2- $\hat{s}^2$ est convergent et non biaisé sous les hypothèses A1 à A5.

3- Lorsque $n \to \infty$, $\hat{s}^2$ et $\hat{\sigma}^2$ deviennent identiques.

3.5.2. Estimateur de $V(\hat{\beta})$

Sur base de l'estimateur $\hat{s}^2$, sous les hypothèses A1 à A5, un estimateur *convergent* et *non biaisé* de $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ est donné par :

$$\hat{V}(\hat{\beta}) = \hat{s}^2(X'X)^{-1} \tag{3.11}$$

soit, sous forme détaillée :

$$\begin{aligned}
\hat{Var}(\hat{\beta}_1) &= \hat{s}^2\frac{\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} = \hat{s}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] \\
\hat{Var}(\hat{\beta}_2) &= \frac{\hat{s}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
\hat{Cov}(\hat{\beta}_1,\hat{\beta}_2) &= \hat{Cov}(\hat{\beta}_2,\hat{\beta}_1) = \hat{s}^2\left[\frac{-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]
\end{aligned}$$

A partir des estimateurs $\hat{Var}(\hat{\beta}_1)$ et de $\hat{Var}(\hat{\beta}_2)$, des estimateurs *convergents*, mais *pas non biaisés* des écarts-types $s.e.(\hat{\beta}_1)$ et $s.e.(\hat{\beta}_2)$ de $\hat{\beta}_1$ et $\hat{\beta}_2$ sont donnés

par :

$$s.\hat{e}.(\hat{\beta}_1) = \sqrt{V\hat{a}r(\hat{\beta}_1)} \quad \text{et} \quad s.\hat{e}.(\hat{\beta}_2) = \sqrt{V\hat{a}r(\hat{\beta}_2)}$$

3.5.3. Exemple : la fonction de consommation de HGL (2008)

Pour les données de Hill, Griffiths et Lim (2008) considérée à la Section 2.2.3, on a vu que la fonction de consommation estimée était :

$$\hat{y}_i = \underbrace{83,42}_{\hat{\beta}_1} + \underbrace{10,21}_{\hat{\beta}_2} x_i$$

où x_i désigne le revenu d'un ménage (en centaines de \$) et y_i les dépenses alimentaires de ce ménage (en \$).

Pour ces données, on obtient comme estimation de σ^2 et de $V(\hat{\beta})$:

$$\hat{s}^2 = 8013,29 \quad \text{et} \quad \hat{V}(\hat{\beta}) = \begin{bmatrix} 1884,44 & -85,90 \\ -85,90 & 4,38 \end{bmatrix}$$

soit, sous forme détaillée :

$$V\hat{a}r(\hat{\beta}_1) = 1884,44 \qquad \Longrightarrow s.\hat{e}.(\hat{\beta}_1) = 43,41$$

$$V\hat{a}r(\hat{\beta}_2) = 4,38 \qquad \Longrightarrow s.\hat{e}.(\hat{\beta}_2) = 2,09$$

$$C\hat{o}v(\hat{\beta}_1, \hat{\beta}_2) = C\hat{o}v(\hat{\beta}_2, \hat{\beta}_1) = -85,90$$

Chapitre 4

Intervalle de confiance et test d'hypothèse

On a vu à la Section 3.3 que, sous les hypothèses A1 à A6, on a de façon exacte :

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}),$$

et en particulier :

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 q_{jj}), \qquad j = 1, 2, \tag{4.1}$$

où $q_{jj} = [(X'X)^{-1}]_{jj}$ désigne l'élément (j, j) de la matrice $(X'X)^{-1}$.

En s'appuyant sur ce résultat de distribution d'échantillonnage de $\hat{\beta}$, on peut construire des intervalles de confiance et des tests d'hypothèses relatifs à β.

Dans un premier temps, nous supposerons toujours que, outre les hypothèses A1 à A5, l'hypothèse optionnelle de normalité A6 est satisfaite. Nous verrons en fin de chapitre ce qu'il en est si on renonce à cette hypothèse.

4.1. Intervalles de confiance pour β_1 et β_2

L'estimateur MCO $\hat{\beta}$ délivre une estimation ponctuelle de $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$. Sur base du résultat de distribution d'échantillonnage (4.1), on peut construire des *intervalles de confiance*, aussi appelés *estimateurs par intervalle*, de β_1 et de β_2 qui, plutôt que de délivrer une valeur ponctuelle, fournissent des intervalles de valeurs plausibles pour β_1 et β_2, et par là même rendent compte de la *précision d'estimation* de β_1 et β_2.

Notons que comme l'estimateur MCO, les intervalles de confiance sont des *règles de décision*.

4.1.1. Cas où σ^2 est connu

Pour simplifier, on commence par considérer le cas où σ^2 est connu.

On sait que, sous les hypothèses A1 à A6, on a pour $\hat{\beta}_j$ $(j = 1, 2)$:

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j)) \,,$$

où $Var(\hat{\beta}_j) = \sigma^2 q_{jj}$, avec $q_{jj} = [(X'X)^{-1}]_{jj}$, de sorte que :

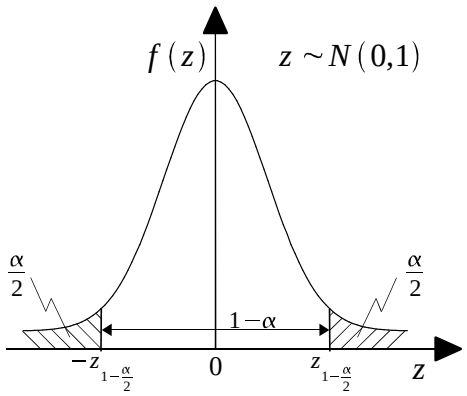
$$\hat{z} = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim N(0, 1) \,, \tag{4.2}$$

où $s.e.(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)}$.

$\hat{z}$ ayant une distribution normale standardisée, en utilisant une table[31] ou l'ordinateur, on peut trouver la *valeur critique* $z_{1-\frac{\alpha}{2}}$ qui est telle que :

$$\begin{aligned} & I\!P(z \leq -z_{1-\frac{\alpha}{2}}) = I\!P(z \geq z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2} \\ \Leftrightarrow \quad & I\!P(-z_{1-\frac{\alpha}{2}} \leq z \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha \,, \end{aligned}$$

où $z \sim N(0, 1)$. $z_{1-\frac{\alpha}{2}}$ est le *quantile d'ordre* $1 - \frac{\alpha}{2}$ de la loi $N(0, 1)$. Graphiquement :



Graphique 12 : Quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $N(0, 1)$

Par exemple, pour $\alpha = 0,05$, et donc $1 - \alpha = 0,95$, on a $z_{1-\frac{\alpha}{2}} = 1,96$.

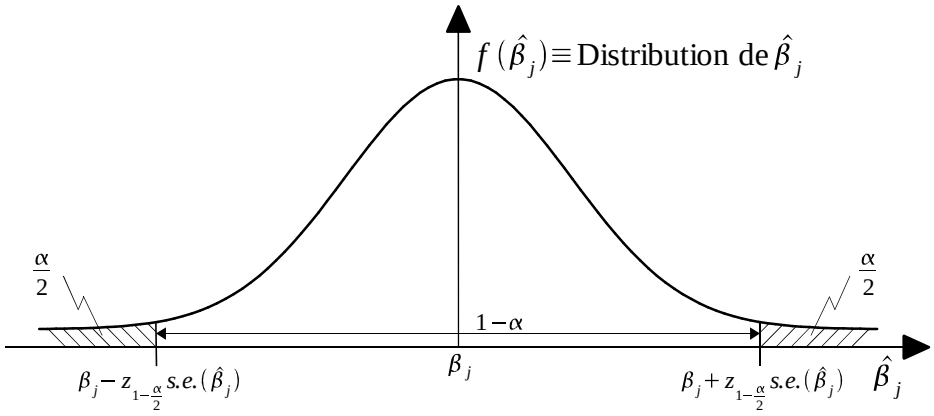
Pour α donné, et donc la valeur critique $z_{1-\frac{\alpha}{2}}$, on a ainsi :

$$I\!P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \tag{4.3}$$

$$\Leftrightarrow \quad I\!P\left(\beta_j - z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j) \leq \hat{\beta}_j \leq \beta_j + z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j)\right) = 1 - \alpha \tag{4.4}$$

[31] Voir l'annexe E de Hill, Griffiths et Lim (2008).

Graphiquement :



Graphique 13 : Intervalle non-stochastique pour $\hat{\beta}_j$

On voit que, connaissant la vraie valeur β_j et l'écart-type $s.e.(\hat{\beta}_j)$ — qui dépend de n, de σ^2 et des x_i, cf. Section 3.1.3 —, on obtient aisément un intervalle *non-stochastique*[32] $\left[\beta_j - z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j)\,;\, \beta_j + z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j)\right]$ au sein duquel l'estimateur MCO $\hat{\beta}_j$ a, sous les hypothèses A1 à A6, une probabilité $(1-\alpha)$ de prendre sa valeur.

Un intervalle de confiance pour β_j $(j = 1, 2)$ est obtenu en suivant la même logique que ci-dessus, mais en 'l'inversant'. De (4.3), on peut en effet également obtenir[33] :

$$\mathbb{P}\left(\hat{\beta}_j - z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j)\right) = 1 - \alpha\,, \tag{4.5}$$

soit un *intervalle de confiance à* $(1-\alpha)\times 100\%$ pour β_j :

$$\left[\hat{\beta}_j - z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j)\,;\, \hat{\beta}_j + z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j)\right] \tag{4.6}$$

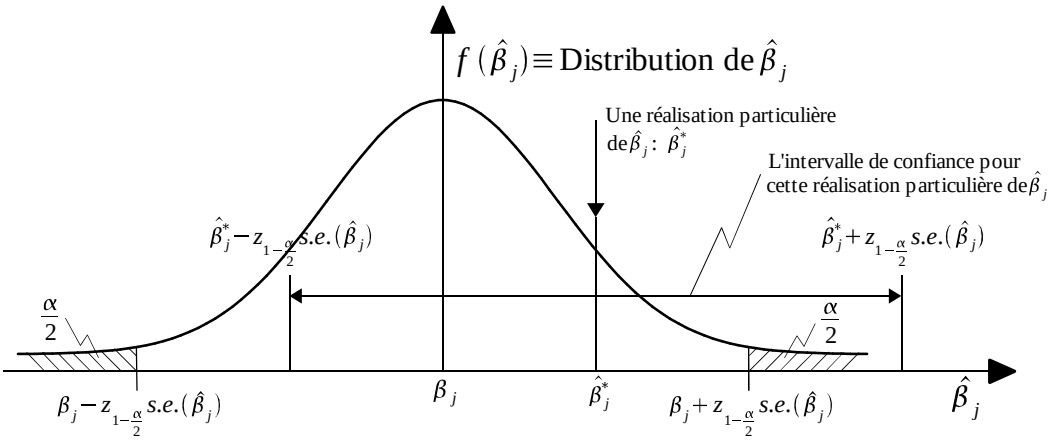
Le Graphique 14 ci-dessous illustre le lien entre l'intervalle de confiance (4.6) obtenu de (4.5), dont les bornes sont *stochastiques*[34], et l'intervalle *non-stochastique* associé à (4.4).

---

[32] i.e., qui ne varie pas d'un échantillon à l'autre.

[33] En isolant β_j plutôt que $\hat{\beta}_j$ au centre des inégalités (vérifiez-le !).

[34] i.e., elles varient d'un échantillon à l'autre, comme conséquence de la variation de $\hat{\beta}_j$ d'un échantillon à l'autre.



Graphique 14 : Intervalle de confiance pour β_j

Etant donné (4.5), sous les hypothèses A1 à A6, il y a une probabilité $(1-\alpha)$ que l'intervalle (stochastique) de confiance (4.6) recouvre la vraie valeur (inconnue) β_j. Notons cependant que pour un échantillon particulier, rien ne garantit que ce soit effectivement le cas. Simplement, étant donné la règle de décision adoptée, il y a de fortes chances (si on choisit α petit) qu'il en soit bien ainsi.

4.1.2. Cas où σ^2 est inconnu

En pratique, on ne peut pas appliquer le résultat de la section précédente car σ^2 est inconnu.

Pour contourner ce problème, il semble assez logique de remplacer la valeur inconnue de σ^2 par son estimateur convergent et non biaisé (sous les hypothèses A1 à A5) :

$$\hat{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{\hat{e}'\hat{e}}{n-2}$$

Quel est l'impact de ce remplacement ? Pour le savoir, on cherche ce que devient la distribution d'échantillonnage de :

$$\hat{z} = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 q_{jj}}},$$

où $q_{jj} = [(X'X)^{-1}]_{jj}$, lorsque σ^2 est remplacé par son estimateur $\hat{s}^2$.

Des calculs effectués à la Section 3.5.1, on sait que :

$$\hat{e}'\hat{e} = e' M_X e\,,$$

où M_X est une matrice symétrique idempotente dont la trace $\mathrm{tr}(M_X) = n-2$.

On peut montrer que le rang et la trace d'une matrice symétrique sont égaux. On a donc que $\hat{e}'\hat{e}$ est égal à une forme quadratique $e'M_Xe$ où :

$$e \sim N(0, \sigma^2 I) \qquad \text{(hypothèse A6)},$$

et M_X est une matrice symétrique idempotente de rang $(n-2)$.

D'après la propriété (2.19) de la loi normale multivariée[35], on a ainsi :

$$\frac{\hat{e}'\hat{e}}{\sigma^2} = \frac{e'M_Xe}{\sigma^2} \sim \chi^2(n-2)$$

$$\Leftrightarrow \quad \hat{v} = \frac{(n-2)\hat{s}^2}{\sigma^2} \sim \chi^2(n-2) \qquad (\text{car } \hat{e}'\hat{e} = (n-2)\hat{s}^2)$$

On peut encore montrer que $\hat{z} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 q_{jj}}}$ et $\hat{v} = \frac{(n-2)\hat{s}^2}{\sigma^2}$ sont indépendamment distribués, de sorte que de la définition de la loi de Student[36], on a :

$$\hat{t} = \frac{\hat{z}}{\sqrt{\frac{\hat{v}}{n-2}}} = \frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 q_{jj}}}}{\sqrt{\frac{\hat{s}^2}{\sigma^2}}} \sim t(n-2)\,,$$

soit, en simplifiant :

$$\hat{t} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{s}^2 q_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)} \sim t(n-2) \tag{4.7}$$

On constate que le remplacement de σ^2 par son estimateur $\hat{s}^2$ fait passer d'une loi normale standardisée à une loi de Student à $(n-2)$ degrés de liberté, qui est plus dispersée que la loi normale, mais qui tend vers elle lorsque $n \to \infty$ (Cf. l'annexe B de Hill, Griffiths et Lim (2008)).

Sur base du résultat (4.7), en suivant la même démarche qu'à la section précédente, on obtient facilement un intervalle de confiance pour β_j.

En utilisant une table[37] ou l'ordinateur, on peut trouver la *valeur critique* $t_{n-2;1-\frac{\alpha}{2}}$ qui est telle que :

$$I\!P(t \leq -t_{n-2;1-\frac{\alpha}{2}}) = I\!P(t \geq t_{n-2;1-\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

$$\Leftrightarrow \quad I\!P(-t_{n-2;1-\frac{\alpha}{2}} \leq t \leq t_{n-2;1-\frac{\alpha}{2}}) = 1-\alpha\,,$$

où $t \sim t(n-2)$. $t_{n-2;1-\frac{\alpha}{2}}$ est le *quantile d'ordre* $1-\frac{\alpha}{2}$ de la loi $t(n-2)$. Graphique-

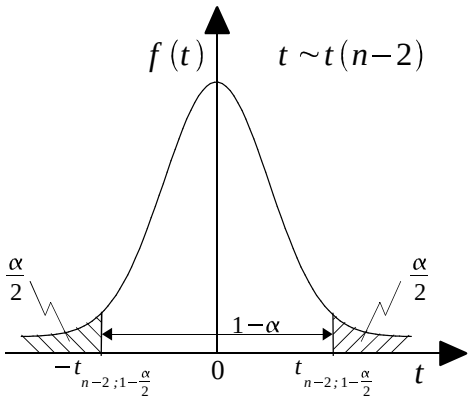
---

[35] Cf. Section 2.3.1.

[36] Si $z \sim N(0,1)$, $v \sim \chi^2(m)$ et que z et v sont indépendamment distribués, alors : $t = \frac{z}{\sqrt{\frac{v}{m}}} \sim t(m)$. Cf. l'annexe B de Hill, Griffiths et Lim (2008).

[37] Voir l'annexe E de Hill, Griffiths et Lim (2008).

ment :



Graphique 15 : Quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $t(n-2)$

Par exemple, pour $\alpha = 0,05$ et $n = 20$, , on a $t_{n-2;1-\frac{\alpha}{2}} = 2,101$.

Pour α et n donné, et donc la valeur critique $t_{n-2;1-\frac{\alpha}{2}}$, on a ainsi :

$$I\!P\left(-t_{n-2;1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)} \leq t_{n-2;1-\frac{\alpha}{2}}\right) = 1 - \alpha\,,$$

d'où on peut tirer :

$$I\!P\left(\hat{\beta}_j - t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j)\right) = 1 - \alpha\,, \tag{4.8}$$

soit un *intervalle de confiance à* $(1-\alpha)\times 100\%$ pour β_j :

$$\left[\hat{\beta}_j - t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j)\,;\, \hat{\beta}_j + t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j)\right] \tag{4.9}$$

Etant donné (4.8), sous les hypothèses A1 à A6, comme dans le cas où σ^2 est connu, il y a une probabilité $(1-\alpha)$ que l'intervalle (*stochastique*[38]) de confiance (4.9) recouvre la vraie valeur (inconnue) β_j.

Appliqué à un échantillon particulier, l'intervalle de confiance (4.9) à $(1-\alpha)\times 100\%$ pour β_j $(j = 1, 2)$ synthétise de façon très parlante l'information disponible tant sur le niveau (estimation ponctuelle) que sur la variabilité d'échantillonnage, et donc la précision, de l'estimation obtenue : le centre de l'intervalle de confiance donne l'estimation ponctuelle, et ses bornes, qui dépendent de α et de l'écart-type[39] *estimé* $s.\hat{e}.(\hat{\beta}_j)$ de l'estimateur $\hat{\beta}_j$, indique (pour α donné) l'ampleur *estimée* de sa variabilité.

[38] Il varie à nouveau d'un échantillon à l'autre.

[39] Notons que contrairement à la variance, l'écart-type est lui exprimé dans les mêmes unités de mesure que le paramètre.

4.1.3. Exemple : la fonction de consommation de HGL (2008)

Pour les données de Hill, Griffiths et Lim (2008) considérée à la Section 2.2.3, on vu que la fonction de consommation estimée pour un échantillon de 40 ménages était :

$$\hat{y}_i = \underbrace{83,42}_{\hat{\beta}_1} + \underbrace{10,21}_{\hat{\beta}_2} x_i$$

où x_i désigne le revenu d'un ménage (en centaines de \$) et y_i les dépenses alimentaires de ce ménage (en \$).

On a par ailleurs vu à la Section 3.5.3 qu'on avait pour ces données :

$$\begin{aligned} \hat{Var}(\hat{\beta}_1) = 1884,44 \qquad &\Longrightarrow s.\hat{e}.(\hat{\beta}_1) = 43,41 \\ \hat{Var}(\hat{\beta}_2) = 4,38 \qquad &\Longrightarrow s.\hat{e}.(\hat{\beta}_2) = 2,09 \end{aligned}$$

Pour $\alpha = 0,05$ et $(n-2) = 40-2 = 38$, on a $t_{n-2;1-\frac{\alpha}{2}} = t_{38;0,975} = 2,024$, de sorte qu'un intervalle de confiance à 95% pour β_1 est donné par :

$$\begin{aligned} \hat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_1) &= 83,42 \pm 2,024 \times 43,41 \\ &= [-4,44\,;\, 171,28]\ , \end{aligned}$$

et un intervalle de confiance à 95% pour β_2 est donné par :

$$\begin{aligned} \hat{\beta}_2 \pm t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_2) &= 10,21 \pm 2,024 \times 2,09 \\ &= [5,98\,;\, 14,44] \end{aligned}$$

De l'intervalle de confiance à 95% pour β_2, on peut affirmer avec un *niveau de confiance de 95%* qu'une augmentation du revenu de 100 \$ accroît la consommation alimentaire moyenne d'un ménage d'un montant compris entre 5,98 \$ et 14,44 \$ (attention aux unités de mesure !). Pour une augmentation de 1 \$ du revenu, cela donne une augmentation de la consommation alimentaire moyenne comprise entre 0,0598 \$ et 0,1444 \$. On constate que l'estimation de β_2 obtenue est assez précise.

4.2. Tests d'hypothèses de β_1 et β_2

Les tests d'hypothèses sont des *règles de décision* statistiques permettant d'évaluer si une hypothèse ou une conjecture théorique est ou non compatible avec les observations dont on dispose.

En s'appuyant sur les propriétés d'échantillonnage de $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$, on peut construire des tests d'hypothèses *bilatéraux* et *unilatéraux* concernant les vraies valeurs de β_1 et β_2.

4.2.1. Cas où σ^2 est connu

Comme pour les intervalles de confiance, on commence, pour simplifier, par considérer le cas où σ^2 est connu.

4.2.1.1. Statistique de test

On sait que, sous les hypothèses A1 à A6, on a pour $\hat{\beta}_j$ $(j = 1, 2)$:

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j)) \,,$$

où $Var(\hat{\beta}_j) = \sigma^2 q_{jj}$, avec $q_{jj} = [(X'X)^{-1}]_{jj}$.

Ainsi, si la vraie valeur de β_j *est égale à* β_j^o, on a :

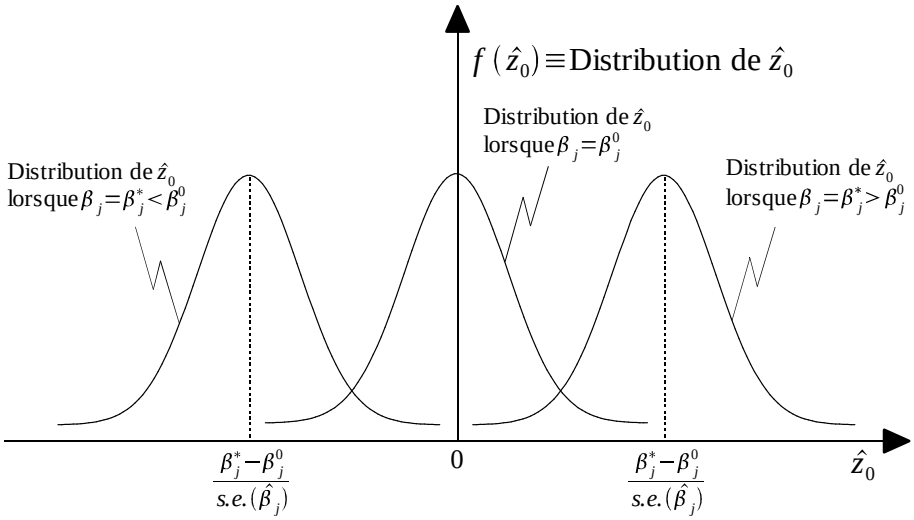
$$\begin{aligned} & \hat{\beta}_j \sim N(\beta_j^o, Var(\hat{\beta}_j)) \\ \Leftrightarrow \qquad & \hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \sim N(0,1) \,, \end{aligned} \tag{4.10}$$

où $s.e.(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)}$, tandis que si la vraie valeur de β_j *est différente de* β_j^o *et par exemple égale à* β_j^* $(\beta_j^* \neq \beta_j^o)$, on a :

$$\begin{aligned} & \hat{\beta}_j \sim N(\beta_j^*, Var(\hat{\beta}_j)) \\ \Leftrightarrow \qquad & (\hat{\beta}_j - \beta_j^o) \sim N(\beta_j^* - \beta_j^o, Var(\hat{\beta}_j)) \\ \Leftrightarrow \qquad & \hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \sim N\left(\frac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}, 1\right) \end{aligned} \tag{4.11}$$

En d'autres termes, si $\beta_j = \beta_j^o$, $\hat{z}_o$ suit une loi normale standardisée, tandis que si $\beta_j = \beta_j^*$ $(\neq \beta_j^o)$, le même $\hat{z}_o$ suit une loi normale, toujours de variance unitaire, mais

d'espérance différente de 0. Graphiquement :



Graphique 16 : Distribution de $\hat{z}_o$

Etant donné ses propriétés, on peut utiliser $\hat{z}_o$ comme *statistique de test* pour tester des hypothèses telles que $\mathrm{H}_0 : \beta_j = \beta_j^o$ contre $\mathrm{H}_1 : \beta_j \neq \beta_j^o$ (test *bilatéral*) ou $\mathrm{H}_0 : \beta_j \leq \beta_j^o$ (resp. $\beta_j \geq \beta_j^o$) contre $\mathrm{H}_1 : \beta_j > \beta_j^o$ (resp. $\beta_j < \beta_j^o$) (tests *unilatéraux*).

4.2.1.2. Test bilatéral

Un test bilatéral au *seuil* ou *niveau* (de signification) α de l'*hypothèse nulle* $\mathrm{H}_0 : \beta_j = \beta_j^o$ contre l'*hypothèse alternative* $\mathrm{H}_1 : \beta_j \neq \beta_j^o$ est donné par la règle de décision :

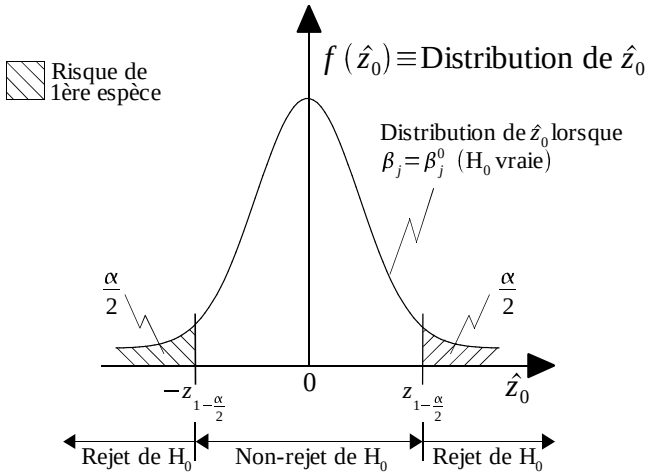
$$\begin{cases} \text{- Rejet de } \mathrm{H}_0 \text{ si } |\hat{z}_o| = \left| \dfrac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \right| > z_{1-\frac{\alpha}{2}} \\ \text{- Non-rejet de } \mathrm{H}_0 \text{ sinon} \end{cases}$$

où la *valeur critique* $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $N(0,1)$ (cf. le Graphique 12 de la Section 4.1.1).

Le seuil α du test est le *risque de première espèce* (ou *probabilité d'erreur de type I*) du test, càd. la probabilité de rejeter H_0 lorsque H_0 est vraie :

$$I\!P(\mathrm{RH}_0 \,|\, \mathrm{H}_0 \text{ est vraie}) = \alpha$$

Graphiquement :



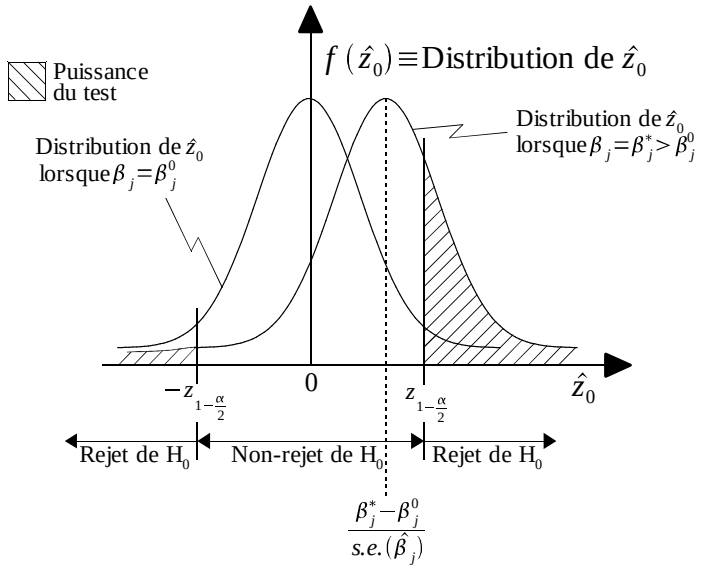
Graphique 17 : Risque de première espèce du test bilatéral

Plus α est choisi petit, plus on peut être confiant dans le fait que, lorsqu'on rejette H_0, cela est effectivement dû au fait que H_0 est fausse[40].

La *puissance* du test est la probabilité de rejeter H_0 lorsque H_0 est fausse :

$$I\!P(\mathrm{RH}_0 \,|\, \mathrm{H}_0 \text{ est fausse}),$$

probabilité qui est égale à 1 moins le *risque de deuxième espèce*[41] (ou *probabilité d'erreur de type II*). Graphiquement :



Graphique 18 : Puissance du test bilatéral

La puissance du test dépend :

1- du seuil du test (si $\alpha \searrow$, la puissance $\searrow$),

[40] En effet, si H_0 était vraie, il n'y aurait qu'une petite probabilité α de la rejeter.

[41] càd. la probabilité de ne pas rejeter H_0 lorsque H_0 est fausse : $I\!P(\mathrm{NRH}_0 \,|\, \mathrm{H}_0 \text{ est fausse})$.

2- de la fausseté de H_0 (si $|\beta_j^* - \beta_j^o| \nearrow$, la puissance $\nearrow$),

3- de la précision d'estimation (si $s.e.(\hat{\beta}_j) \searrow$, la puissance $\nearrow$).

Au contraire du risque de première espèce α, la puissance du test *n'est pas sous contrôle*. C'est pourquoi, sauf si on a de bonnes raisons de penser que la puissance du test est élevée (par exemple parce que la taille d'échantillon est très élevée, de sorte que le précision d'estimation est grande), il faut se garder d'interpréter un non-rejet de H_0 comme une preuve convaincante que H_0 est vraie[42]. Il s'agit 'seulement' d'une absence de preuve que H_0 est fausse (ce qui n'est pas si mal !).

4.2.1.3. Relation entre test bilatéral et intervalle de confiance

On peut établir un lien entre le test bilatéral de $H_0 : \beta_j = \beta_j^o$ contre $H_1 : \beta_j \neq \beta_j^o$ et l'intervalle de confiance pour β_j $(j = 1, 2)$.

Dans le test bilatéral au *seuil* α de $H_0 : \beta_j = \beta_j^o$ contre $H_1 : \beta_j \neq \beta_j^o$ décrit à la section précédente, on *ne rejette pas* H_0 lorsque :

$$|\hat{z}_o| = \left| \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \right| \leq z_{1-\frac{\alpha}{2}} ,$$

soit, lorsque :

$$-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \leq z_{1-\frac{\alpha}{2}}$$

$$\Leftrightarrow \qquad \beta_j^o - z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j) \leq \hat{\beta}_j \leq \beta_j^o + z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j)$$

$$\Leftrightarrow \qquad \hat{\beta}_j - z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j) \leq \beta_j^o \leq \hat{\beta}_j + z_{1-\frac{\alpha}{2}} s.e.(\hat{\beta}_j)$$

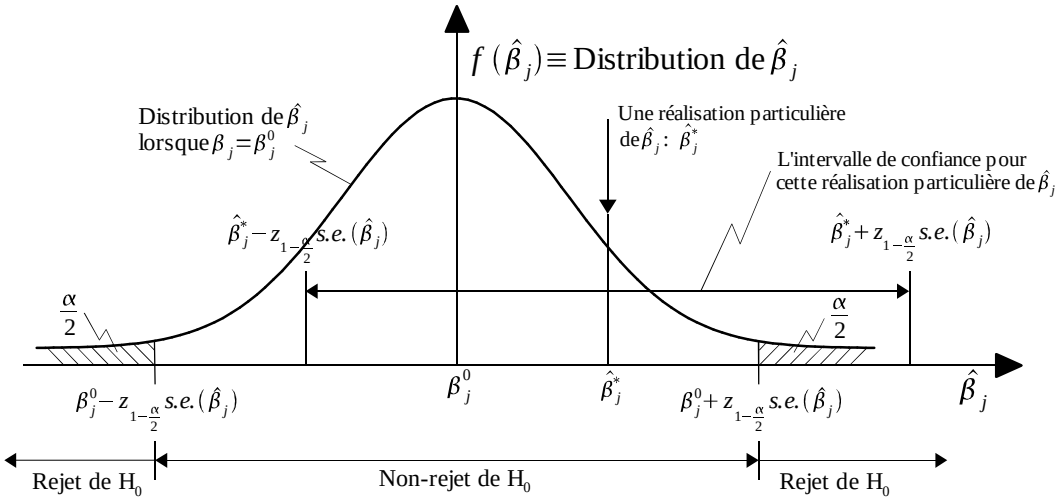
Les bornes de ce dernier intervalle ne sont rien d'autre que les bornes de l'intervalle de confiance à $(1-\alpha) \times 100\%$ pour β_j obtenu à la Section 4.1.1 (cf. équation (4.6)).

En d'autres termes, on peut réaliser de façon totalement équivalente un test au seuil α de $H_0 : \beta_j = \beta_j^o$ contre $H_1 : \beta_j \neq \beta_j^o$ sur base de l'intervalle de confiance à $(1-\alpha) \times 100\%$ pour β_j en appliquant la règle de décision :

$$\begin{cases} \text{- Rejet de } H_0 \text{ si } & \beta_j^o \text{ n'appartient pas à l'intervalle de} \\ & \text{confiance (4.6) à } (1-\alpha) \times 100\% \text{ pour } \beta_j \\ \text{- Non-rejet de } H_0 \text{ sinon} & \end{cases}$$

[42] Pour pouvoir interpréter un non-rejet de H_0 comme une preuve convaincante que H_0 est vraie, il faut être assuré que la puissance du test est grande, ou autrement dit, que le risque de deuxième espèce $I\!P(\text{NR}H_0 \,|\, H_0 \text{ est fausse})$ est petit. Dans ce cas, lorsqu'on ne rejette pas H_0, on peut être confiant dans le fait que H_0 est effectivement vraie (puisque si H_0 était fausse, il n'y aurait qu'une petite probabilité — le risque de deuxième espèce — de ne pas la rejeter).

Graphiquement :



Graphique 19 : Test bilatéral et intervalle de confiance

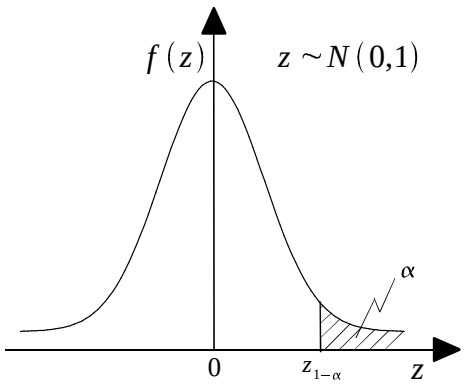
On remarquera incidemment que lorsque la précision d'estimation de β_j est faible, et donc son intervalle de confiance est large, on ne pourra pas rejeter H_0 : $\beta_j = \beta_j^o$ contre H_1 : $\beta_j \neq \beta_j^o$ pour un tout aussi large éventail (que l'intervalle de confiance) de valeurs de β_j^o.

4.2.1.4. Test unilatéraux

Un test *unilatéral à droite* au *seuil* α de H_0 : $\beta_j \leq \beta_j^o$ contre H_1 : $\beta_j > \beta_j^o$ est donné par la règle de décision :

$$\begin{cases} \text{- Rejet de } H_0 \text{ si } \hat{z}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} > z_{1-\alpha} \\ \text{- Non-rejet de } H_0 \text{ sinon} \end{cases}$$

où la *valeur critique* $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $N(0,1)$, càd. la valeur $z_{1-\alpha}$ telle que $I\!P(z \leq z_{1-\alpha}) = 1-\alpha$, où $z \sim N(0,1)$. Graphiquement :



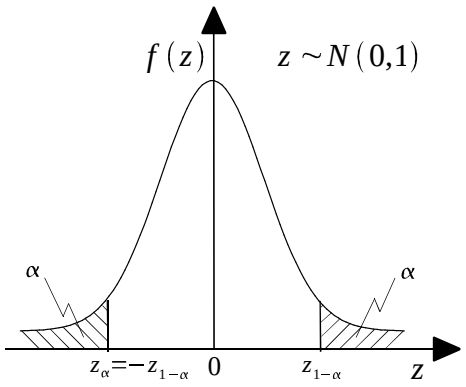
Graphique 20 : Quantile d'ordre $1-\alpha$ de la loi $N(0,1)$

Par exemple, pour $\alpha = 0,05$, et donc $1-\alpha = 0,95$, on a $z_{1-\alpha} = 1,6449$.

De façon symétrique, un test *unilatéral à gauche* au *seuil* α de H_0 : $\beta_j \geq \beta_j^o$ contre H_1 : $\beta_j < \beta_j^o$ est donné par la règle de décision :

$$\begin{cases} \text{- Rejet de } \mathrm{H}_0 \text{ si } \hat{z}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} < z_\alpha (= -z_{1-\alpha}) \\ \text{- Non-rejet de } \mathrm{H}_0 \text{ sinon} \end{cases}$$

où la *valeur critique* z_α est le quantile d'ordre α de la loi $N(0,1)$, càd. la valeur z_α telle que $I\!P(z \leq z_\alpha) = 1-\alpha$, où $z \sim N(0,1)$. Graphiquement :



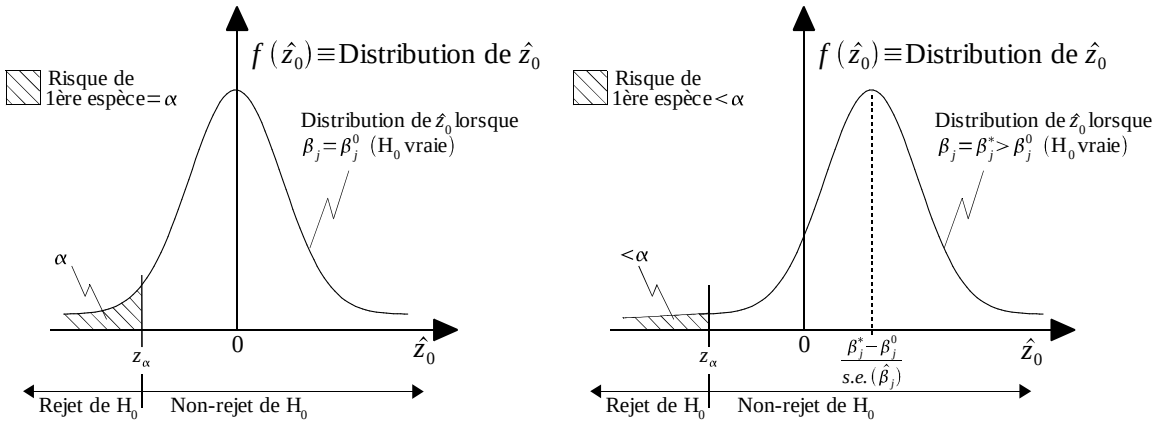
Graphique 21 : Quantile d'ordre α de la loi $N(0,1)$

Par exemple, pour $\alpha = 0,05$, et donc $1-\alpha = 0,95$, on a $z_\alpha = -z_{1-\alpha} = -1,6449$.

Le seuil α des tests unilatéraux (à droite ou à gauche) est la *valeur maximum* du *risque de première espèce* de ces tests. On a toujours :

$$I\!P(\mathrm{RH}_0 \,|\, \mathrm{H}_0 \text{ est vraie}) \leq \alpha$$

l'égalité se réalisant pour H_0 vraie avec $\beta_j = \beta_j^o$. Graphiquement (cas du test unilatéral à gauche) :



Graphique 22 : Risque de première espèce du test unilatéral à gauche

Comme dans le cas du test bilatéral, plus α est choisi petit, plus on peut être

confiant dans le fait que, lorsqu'on rejette H_0, cela est effectivement dû au fait que H_0 est fausse.

A nouveau comme dans le cas du test bilatéral, la *puissance* des tests unilatéraux (à droite ou à gauche) :

$$I\!P(\mathrm{RH}_0 \,|\, \mathrm{H}_0 \text{ est fausse}),$$

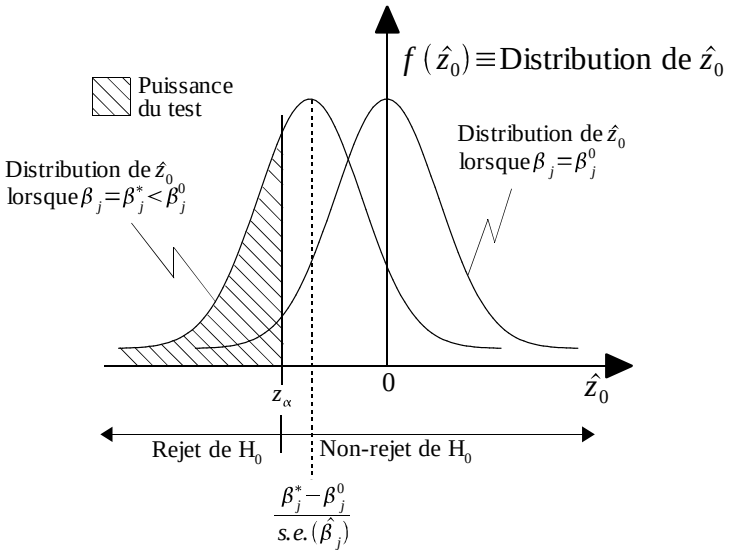
dépend :

1- du seuil du test (si $\alpha \searrow$, la puissance $\searrow$),

2- de la fausseté de H_0 (si $|\beta_j^* - \beta_j^o| \nearrow$, la puissance $\nearrow$),

3- de la précision d'estimation (si $s.e.(\hat{\beta}_j) \searrow$, la puissance $\nearrow$).

Graphiquement (cas du test unilatéral à gauche) :



Graphique 23 : Puissance du test unilatéral à gauche

Toujours comme dans le cas du test bilatéral, au contraire du risque de première espèce qui est toujours inférieur ou égal à α, la puissance des tests unilatéraux *n'est pas sous contrôle*, de sorte qu'on se gardera d'interpréter (sauf si on a de bonnes raisons de le faire) un non-rejet de H_0 comme une preuve convaincante que H_0 est vraie.

4.2.1.5. *P*-valeur d'un test bilatéral et unilatéral

La mise en oeuvre des procédures de test décrites ci-dessus délivre un résultat binaire (on rejette ou on ne rejette pas), et ce résultat peut être différent selon le choix du seuil α du test : on peut par exemple rejeter H_0 au seuil de 5%, mais pas au seuil de 1%.

Ayant calculé la valeur de la statistique $\hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)}$ pour un *échantillon parti-*

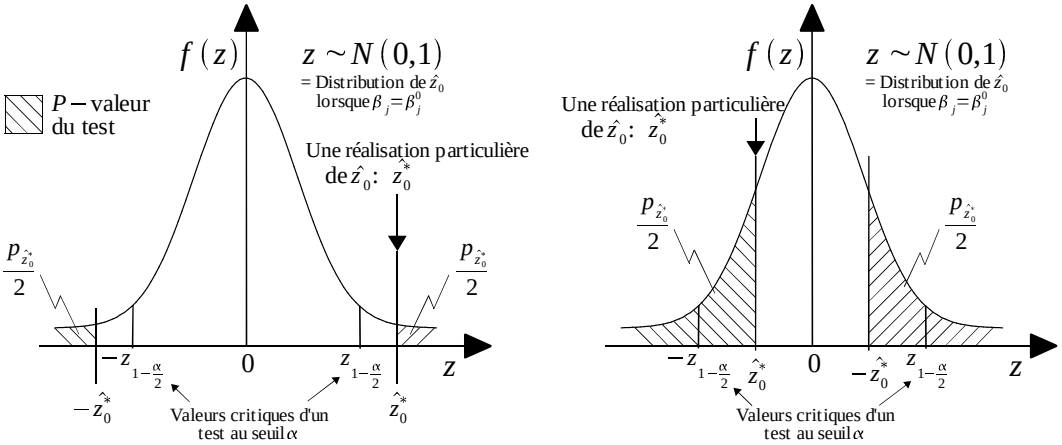
culier, il est naturel de se demander quelle est la *valeur minimale* du seuil α du test (bilatéral ou unilatéral selon le test réalisé) pour laquelle on peut rejeter H_0. Cette valeur minimale de α est appelée la *P-valeur du test*.

Désignons par $\hat{z}_o^* = \frac{\hat{\beta}_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}$ la valeur de la statistique de test $\hat{z}_o$ obtenue pour un échantillon particulier.

Dans le cas d'un *test bilatéral* ($H_0 : \beta_j = \beta_j^o$ contre $H_1 : \beta_j \neq \beta_j^o$), la P-valeur $p_{\hat{z}_o^*}$ du test pour cet échantillon particulier est donnée par :

$$p_{\hat{z}_o^*} = I\!P(|z| > |\hat{z}_o^*|)\,, \quad \text{où } z \sim N(0,1)$$

Graphiquement :



Graphique 24 : P-valeur d'un test bilatéral

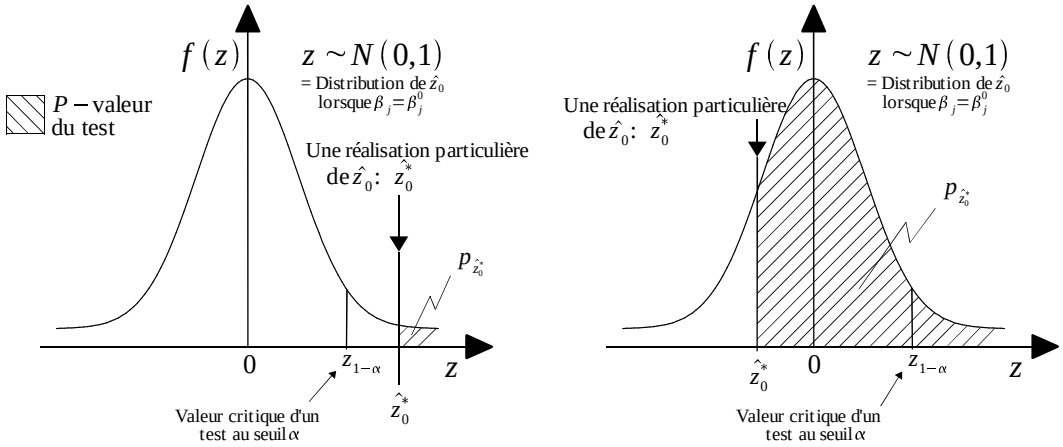
Dans le cas d'un *test unilatéral à droite* ($H_0 : \beta_j \leq \beta_j^o$ contre $H_1 : \beta_j > \beta_j^o$), la P-valeur $p_{\hat{z}_o^*}$ du test pour cet échantillon particulier est donnée par :

$$p_{\hat{z}_o^*} = I\!P(z > \hat{z}_o^*)\,, \quad \text{où } z \sim N(0,1)\,,$$

et de façon symétrique, dans le cas d'un *test unilatéral à gauche* ($H_0 : \beta_j \geq \beta_j^o$ contre $H_1 : \beta_j < \beta_j^o$), la P-valeur $p_{\hat{z}_o^*}$ du test pour cet échantillon particulier est donnée par :

$$p_{\hat{z}_o^*} = I\!P(z < \hat{z}_o^*)\,, \quad \text{où } z \sim N(0,1)$$

Graphiquement (cas du test unilatéral à droite) :



Graphique 25 : P-valeur d'un test unilatéral à droite

Comme le suggère les graphiques ci-dessus, un test (bilatéral ou unilatéral) au seuil α rejettera H_0 pour tout α supérieure à la P-valeur $p_{\hat{z}_o^*}$ du test, et ne rejettera pas H_0 pour tout α inférieur (ou égal) à la P-valeur $p_{\hat{z}_o^*}$ du test. La P-valeur $p_{\hat{z}_o^*}$ est donc bien la valeur minimale du seuil α du test pour laquelle on peut rejeter H_0.

Plus la P-valeur $p_{\hat{z}_o^*}$ du test (bilatéral ou unilatéral) est petite, plus on peut rejeter H_0 à un seuil α petit, càd. avec un risque de première espèce (dans le cas d'un test unilatéral, un risque de première espèce maximum) petit, et donc plus il est crédible que H_0 est fausse.

On notera que la P-valeur d'un test n'est pas la probabilité que H_0 soit vraie[43]. C'est la probabilité, *sous l'hypothèse que H_0 est vraie*, d'obtenir pour la statistique de test $\hat{z}_o$ une valeur 'aussi extrême' que la valeur observée $\hat{z}_o^*$. D'où le fait que plus cette probabilité est petite, plus on peut être confiant dans le fait que H_0 est effectivement fausse.

La P-valeur d'un test est toujours reportée par les logiciels économétriques. Une bonne pratique empirique est de toujours reporter, outre la valeur de la statistique de test obtenue, la P-valeur du test que l'on a effectué. De cette façon, le lecteur peut facilement se faire sa propre opinion quant à la plausibilité de l'hypothèse nulle H_0 testée.

4.2.2. Cas où σ^2 est inconnu

En pratique, la statistique de test $\hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)}$ ne peut pas être calculée car $s.e.(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)} = \sqrt{\sigma^2 q_{jj}}$ (où $q_{jj} = [(X'X)^{-1}]_{jj}$) dépend de la variance du terme d'erreur σ^2 qui est inconnue.

[43] H_0 est soit vraie, soit fausse, pas vraie ou fausse avec une certaine probabilité.

Comme on l'a fait pour le calcul des intervalles de confiance, on peut contourner ce problème en remplaçant σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$.

On a vu à la Section 4.1.2 que, sous les hypothèses A1 à A6, lorsqu'on remplace σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$, on a pour $\hat{\beta}_j$ ($j = 1, 2$; cf. équation (4.7)) :

$$\hat{t} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{s}^2 q_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)} \sim t(n-2)\,,$$

de sorte que si la vraie valeur de β_j *est égale à* β_j^o, on a :

$$\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \sim t(n-2)\,, \tag{4.12}$$

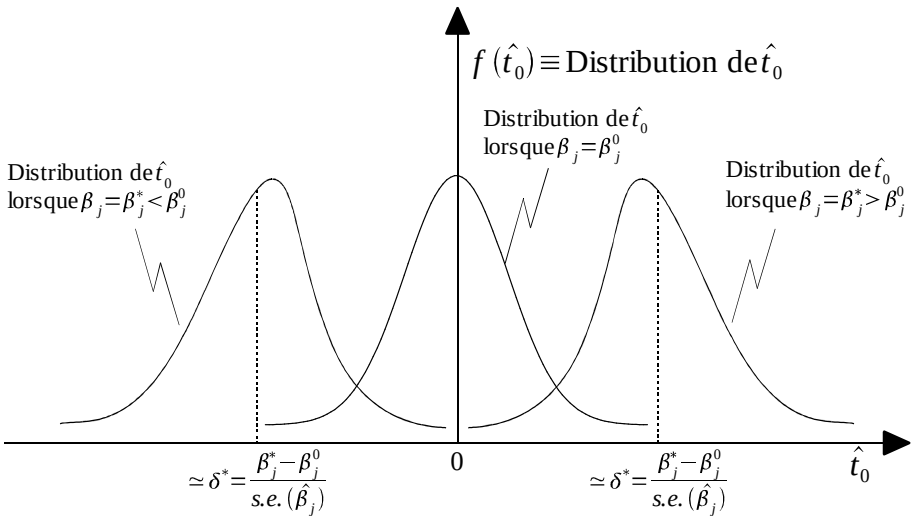
et on peut montrer que si la vraie valeur de β_j *est différente de* β_j^o *et par exemple égale à* β_j^* ($\beta_j^* \neq \beta_j^o$), on a :

$$\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \sim t(\delta^*, n-2)\,, \tag{4.13}$$

où $t(\delta^*, n-2)$ désigne la *loi de Student non-centrale*[44] à $(n-2)$ degrés de liberté et le paramètre de non-centralité δ^* est égal à :

$$\delta^* = \frac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}$$

Graphiquement :



Graphique 26 : Distribution de $\hat{t}_o$

En d'autres termes, si $\beta_j = \beta_j^o$, $\hat{t}_o$ suit une loi de Student à $(n-2)$ degrés de

[44] Par définition, si $z \sim N(\delta, 1)$, $v \sim \chi^2(m)$, et z et v sont indépendamment distribués, alors : $t = \frac{z}{\sqrt{\frac{v}{m}}} \sim t(\delta, m)$.

liberté, tandis que si $\beta_j = \beta_j^* \ (\neq \beta_j^o)$, le même $\hat{t}_o$ suit une loi de Student toujours à $(n-2)$ degrés de liberté, mais décentré (par rapport à zéro), avec un paramètre de non-centralité égal à $\delta^* = \frac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}$.

Notons que, pour n grand (disons $n > 30$), la loi de Student $t(n-2)$ est approximativement la même que la loi normale standardisée $N(0,1)$, et la loi de Student non-centrale $t(\delta^*, n-2)$ est pareillement approximativement la même que la loi normale décentrée $N(\delta^*, 1)$.

On constate qu'à la transposition loi normale / loi de Student près, le comportement de la statistique $\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)}$ est identique celui de la statistique $\hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)}$.

Ainsi, par analogie à ce que nous avons établi à la Section 4.2.1 pour le cas où σ^2 est connu[45] :

1- Un test *bilatéral* au seuil α de $\mathrm{H}_0 : \ \beta_j = \beta_j^o$ contre $\mathrm{H}_1 : \ \beta_j \neq \beta_j^o$ est donné par la règle de décision :

$$\begin{cases} \text{- Rejet de } \mathrm{H}_0 \text{ si } |\hat{t}_o| = \left| \dfrac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \right| > t_{n-2;1-\frac{\alpha}{2}} \\ \text{- Non-rejet de } \mathrm{H}_0 \text{ sinon} \end{cases}$$

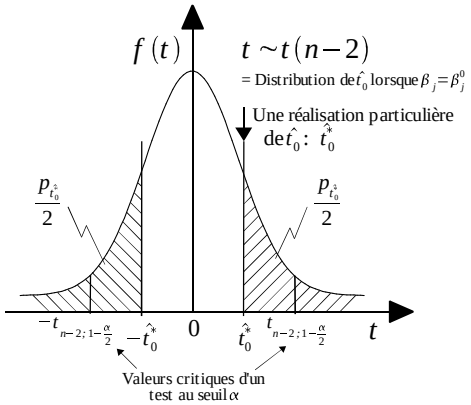
où la valeur critique $t_{n-2;1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $t(n-2)$, ou de façon totalement équivalente par la règle de décision :

$$\begin{cases} \text{- Rejet de } \mathrm{H}_0 \text{ si } \ \beta_j^o \text{ n'appartient pas à l'intervalle de} \\ \qquad\qquad\qquad\quad \text{confiance (4.9) à } (1-\alpha) \times 100\% \text{ pour } \beta_j \\ \text{- Non-rejet de } \mathrm{H}_0 \text{ sinon} \end{cases}$$

et la P-valeur de ce test, pour un échantillon particulier, est donnée par :

$$p_{\hat{t}_o^*} = I\!P(|t| > |\hat{t}_o^*|), \quad \text{où } t \sim t(n-2)$$

Graphiquement :



Graphique 27 : P-valeur d'un t-test bilatéral

[45] Le lecteur est invité à vérifier par lui-même que les mêmes résultats s'appliquent bien.

2- Un test *unilatéral à droite* au seuil α de H_0 : $\beta_j \leq \beta_j^o$ contre H_1 : $\beta_j > \beta_j^o$ est donné par la règle de décision :

$$\begin{cases} \text{- Rejet de } H_0 \text{ si } \hat{t}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} > t_{n-2;1-\alpha} \\ \text{- Non-rejet de } H_0 \text{ sinon} \end{cases}$$

où la valeur critique $t_{n-2;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $t(n-2)$, et la P-valeur de ce test, pour un échantillon particulier, est donnée par :

$$p_{\hat{t}_o^*} = I\!P(t > \hat{t}_o^*)\,, \quad \text{où } t \sim t(n-2)$$

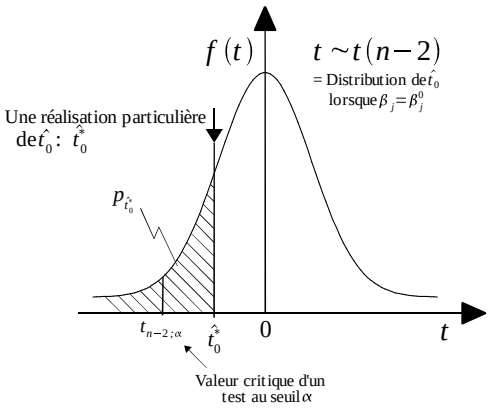
De façon symétrique, un test *unilatéral à gauche* au seuil α de H_0 : $\beta_j \geq \beta_j^o$ contre H_1 : $\beta_j < \beta_j^o$ est donné par la règle de décision :

$$\begin{cases} \text{- Rejet de } H_0 \text{ si } \hat{t}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} < t_{n-2;\alpha} (= -t_{n-2;1-\alpha}) \\ \text{- Non-rejet de } H_0 \text{ sinon} \end{cases}$$

où la valeur critique $t_{n-2;\alpha}$ $(= -t_{n-2;1-\alpha})$ est le quantile d'ordre α de la loi $t(n-2)$, et la P-valeur de ce test, pour un échantillon particulier, est donnée par :

$$p_{\hat{t}_o^*} = I\!P(t < \hat{t}_o^*)\,, \quad \text{où } t \sim t(n-2)$$

Graphiquement (cas du test unilatéral à gauche) :



Graphique 28 : P-valeur d'un t-test unilatéral à gauche

Les interprétations en termes de *risque de première espèce* et de *puissance*, ainsi que l'interprétation de la P-valeur de ces tests, sont identiques (à la transposition loi normale / loi de Student près) à celles développées pour le cas où σ^2 est connu : de ce point de vue, rien de nouveau.

4.2.3. Terminologie et précisions d'interprétation

Lorsqu'on est amené à *rejeter* au *seuil* α, pour un échantillon particulier, l'hypothèse nulle du test :

1- bilatéral de $H_0 : \beta_j = \beta_j^o$ contre $H_1 : \beta_j \neq \beta_j^o$, on dit que le paramètre estimé $\hat{\beta}_j$ *est (statistiquement) significativement différent de* β_j^o *au seuil* α.

2- unilatéral à droite de $H_0 : \beta_j \leq \beta_j^o$ contre $H_1 : \beta_j > \beta_j^o$, on dit que le paramètre estimé $\hat{\beta}_j$ *est (statistiquement) significativement supérieur à* β_j^o *au seuil* α.

3- unilatéral à gauche de $H_0 : \beta_j \geq \beta_j^o$ contre $H_1 : \beta_j < \beta_j^o$, on dit que le paramètre estimé $\hat{\beta}_j$ *est (statistiquement) significativement inférieur à* β_j^o *au seuil* α.

Lorsqu'on est amené à *ne pas rejeter* H_0 au *seuil* α, on dit $\hat{\beta}_j$ *n'est pas (statistiquement) significativement*, selon les cas, *différent de*, *supérieur à*, ou *inférieur à* β_j^o *au seuil* α.

Un cas particulier important de t-test bilatéral de β_j^o est celui où $\beta_j^o = 0$, càd. celui où on teste $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$. Dans ce cas, la statistique de test $\hat{t}_o$ se réduit à $\hat{t}_o = \frac{\hat{\beta}_j}{s.\hat{e}.(\hat{\beta}_j)}$, statistique qu'on appelle couramment *t-statistique* (de $\hat{\beta}_j$). La statistique $\hat{t}_o = \frac{\hat{\beta}_j}{s.\hat{e}.(\hat{\beta}_j)}$ et sa P-valeur sont calculées en standard pour $\hat{\beta}_1$ et $\hat{\beta}_2$ par tous les logiciels économétriques.

Lorsque l'hypothèse nulle $H_0 : \beta_j = 0$ (contre $H_1 : \beta_j \neq 0$) est *rejetée* au *seuil* α, on dit que $\hat{\beta}_j$ *est (statistiquement) significatif au seuil* α, et lorsqu'elle n'est *pas rejetée* au *seuil* α, on dit que $\hat{\beta}_j$ *n'est (statistiquement) significatif au seuil* α.

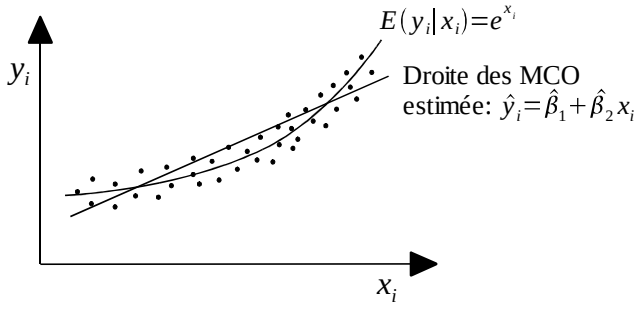
Ce test a généralement peu de sens pour l'intercept β_1. Par contre, pour la pente β_2, il est très important car il revient à tester, dans le cadre de la spécification du modèle de régression linéaire :

$$\begin{aligned} H_0' : E(y_i|x_i) &= \beta_1 , && \text{i.e., } E(y_i|x_i) \text{ est une constante,} \\ & && \text{elle ne dépend pas de } x_i \\ \text{contre } H_1' : E(y_i|x_i) &= \beta_1 + \beta_2 x_i , && \text{i.e., } E(y_i|x_i) \text{ est une fonction} \\ & && \text{linéaire de } x_i \end{aligned}$$

La mise en oeuvre de ce test appelle plusieurs remarques :

1- Le fait de trouver $\hat{\beta}_2$ significatif ne garantit pas que $E(y_i|x_i)$ est bien une fonction

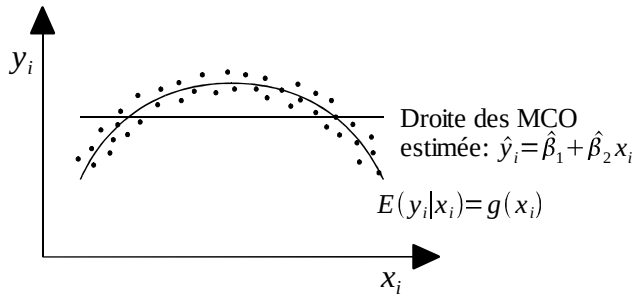
linéaire de x_i. Graphiquement :



Graphique 29 : $\hat{\beta}_2$ significatif avec $E(y_i|x_i)$ non linéaire

Dans l'exemple graphique ci-dessus, la vraie relation est non linéaire ($E(y_i|x_i) = e^x$) et, si la taille d'échantillon n'est pas trop petite, $\hat{\beta}_2$ apparaîtra certainement comme significatif.

2- A contrario, le fait de ne pas trouver $\hat{\beta}_2$ significatif ne signifie pas nécessairement que $E(y_i|x_i)$ ne dépend pas de x_i. C'est seulement une absence de preuve que $E(y_i|x_i)$ dépend de x_i. Cette absence de preuve peut très bien être due à une précision d'estimation trop faible (= puissance de test réduite), ou encore au fait que la vraie relation, qui est non linéaire, reste 'cachée' lorsqu'on considère un modèle linéaire. Graphiquement :



Graphique 30 : $\hat{\beta}_2$ non significatif avec $E(y_i|x_i)$ non linéaire

Dans l'exemple graphique ci-dessus, $\hat{\beta}_2$ apparaîtra certainement comme non significatif.

3- Il ne faut pas confondre '$\hat{\beta}_2$ est (très) significatif' — càd. $\hat{t}_o = \frac{\hat{\beta}_2}{s.\hat{e}.(\hat{\beta}_2)}$ a une (très) grande valeur, ou ce qui revient au même, la P-valeur du test est (très) petite —, et 'x_i a un effet (très) important sur $E(y_i|x_i)$' : lorsque la précision d'estimation est (très) grande (i.e., $s.\hat{e}.(\hat{\beta}_2)$ est (très) petit), on peut très bien avoir que $\hat{\beta}_2$ est (très) significatif et en même temps que l'effet de x_i sur $E(y_i|x_i)$, qui est reflété par la valeur de $\hat{\beta}_2$, est dérisoire (i.e., $\hat{\beta}_2$ est (très) petit). Cette remarque est liée à une caractéristique générale des tests d'hypothèse (quels qu'ils soient) qu'il convient de toujours garder à l'esprit[46] : lorsque la précision d'estimation est grande, un rejet de H_0 (dans le cas qui nous occupe, H_0 : $\beta_2 = 0$), même

[46] Au même titre qu'il convient de toujours garder à l'esprit qu'un rejet de H_0 est une preuve d'autant plus convaincante que H_0 est fausse que le seuil α auquel on rejette H_0 est petit, et qu'un non-rejet de H_0, au moins lorsque la précision d'estimation est limitée, ne constitue pas une preuve convaincante que H_0 est vraie.

très marqué, ne signifie pas nécessairement qu'on en est très loin (dans le cas qui nous occupe, β_2 fortement éloigné de 0).

4.2.4. Exemple : la fonction de consommation de HGL (2008)

Pour les données de Hill, Griffiths et Lim (2008) considérée à la Section 2.2.3, qui pour rappel considère le modèle de fonction de consommation :

$$y_i = \beta_1 + \beta_2 x_i + e_i \,,$$

où x_i désigne le revenu d'un ménage (en centaines de \$) et y_i les dépenses alimentaires de ce ménage (en \$), en utilisant le logiciel GRETL, on obtient comme tableau de résultats d'estimation :

Model 1:
OLS, using observations 1-40
Dependent variable: y

	coefficient	std. error	t-ratio	p-value
const	83.4160	43.4102	1.922	0.0622 *
x	10.2096	2.09326	4.877	1.95e-05 ***

Mean dependent var	283.5735	S.D. dependent var	112.6752
Sum squared resid	304505.2	S.E. of regression	89.51700
R-squared	0.385002	Adjusted R-squared	0.368818
F(1, 38)	23.78884	P-value(F)	0.000019
Log-likelihood	-235.5088	Akaike criterion	475.0176
Schwarz criterion	478.3954	Hannan-Quinn	476.2389

La signification des rubriques reportées par GRETL est :

- coefficient : paramètre estimé $\hat{\beta}_j$
- std. error : écart-type estimé $s.\hat{e}.(\hat{\beta}_j)$
- t-ratio : t-statistique $\hat{t}_o = \frac{\hat{\beta}_j}{s.\hat{e}.(\hat{\beta}_j)}$, i.e. la statistique de test de $\mathrm{H}_0 : \beta_j = 0$ contre $\mathrm{H}_1 : \beta_j \neq 0$
- p-value : la P-valeur du test de $\mathrm{H}_0 : \beta_j = 0$ contre $\mathrm{H}_1 : \beta_j \neq 0$
- Mean dependent var : la valeur moyenne des y_i, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$
- S.D. dependent var : l'écart-type des y_i, $\sigma_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})}$
- Sum squared resid : la somme des carrés des résidus $= \hat{e}'\hat{e}$
- S.E. of regression : l'écart-type estimé de l'erreur $= \sqrt{\hat{s}^2}$
- Log-likelihood : la log-vraisemblance de l'estimateur MV (cf. Section 2.2.2)

Les autres rubriques seront explicitées dans la suite.

Sur base de ce tableau de résultats, si on note que pour $(n-2) = 38$ et $\alpha = 0,05$, on a $t_{n-2;1-\frac{\alpha}{2}} = t_{38;0,975} = 2,024$ et $t_{n-2;1-\alpha} = t_{38;0,95} = 1,686$, on peut :

1- calculer un intervalle de confiance à 95% pour β_1 :

$$\begin{aligned} \hat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_1) &= 83,42 \pm 2,024 \times 43,41 \\ &= [-4,44\,;\, 171,28]\ , \end{aligned}$$

2- calculer un intervalle de confiance à 95% pour β_2 :

$$\begin{aligned} \hat{\beta}_2 \pm t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_2) &= 10,21 \pm 2,024 \times 2,09 \\ &= [5,98\,;\, 14,44] \end{aligned}$$

3- voir que la statistique de test $\hat{t}_o$ du t-test de $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ est égale à $1,922$, et que H_0 peut être rejetée au *seuil minimum* de $0,0622$ (= P-valeur du test).

4- voir que la statistique de test $\hat{t}_o$ du t-test de $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$ est égale à $4,877$, et que H_0 peut être rejetée au *seuil minimum* de $0,0000195$ (= P-valeur du test).

5- effectuer un test de $H_0 : \beta_1 \leq 0$ contre $H_1 : \beta_1 > 0$. On a $\hat{t}_o = 1,922$, et la P-valeur du test est égale à $\frac{0,0622}{2} = 0,0311$, de sorte que H_0 peut être rejetée au *seuil minimum* de $0,0311$.

6- effectuer un test de $H_0 : \beta_2 \geq 20$ contre $H_1 : \beta_2 < 20$. On obtient :

$$\hat{t}_o = \frac{10,21 - 20}{2,09} = -4,68$$

On peut rejeter H_0 au seuil de 5% car $\hat{t}_o = -4,68 < t_{38;0,05} = -t_{38;0,95} = -1,686$. La P-valeur du test[47] est en fait égale à 1,79e-05, de sorte que H_0 peut être rejetée au *seuil minimum* de $0,0000179$.

4.3. Intervalle de confiance, test d'hypothèse et non-normalité

Nous avons obtenu les intervalles de confiance et tests d'hypothèse de β_1 et β_2 en supposant que, outre les hypothèses A1 à A5, l'hypothèse optionnelle de normalité A6 du modèle était satisfaite. Qu'en est-il si, comme on peut couramment s'y attendre en pratique, cette dernière hypothèse n'est pas remplie ?

Comme nous allons le voir, lorsqu'on renonce à l'hypothèse A6 de normalité, les procédures que nous avons établies restent valables, mais seulement asymptotiquement, en grand échantillon.

[47] La P-valeur peut être calculée en utilisant le 'p-value finder' de GRETL.

On a vu à la Section 3.4.2 que, sous les hypothèses A1 à A5, sans faire appel à l'hypothèse de normalité A6, on a *asymptotiquement* (lorsque $n \to \infty$) :

$$\left[V(\hat{\beta})\right]^{-\frac{1}{2}} \left(\hat{\beta} - \beta\right) \xrightarrow{d} N(0, I)\,, \quad \text{où } V(\hat{\beta}) = \sigma^2 (X'X)^{-1}\,,$$

ce qui implique (pour $j = 1, 2$) :

$$\hat{z} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 q_{jj}}} \xrightarrow{d} N(0,1)\,, \quad \text{où } q_{jj} = \left[(X'X)^{-1}\right]_{jj}\,,$$

soit, sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand :

$$\hat{z} = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \approx N(0,1)\,, \quad \text{où } s.e.(\hat{\beta}_j) = \sqrt{\sigma^2 q_{jj}}\,, \tag{4.14}$$

et donc :

$$\begin{cases} \hat{z}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \approx N(0,1)\,, & \text{si } \beta_j = \beta_j^o \\ \hat{z}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \approx N\left(\dfrac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}, 1\right)\,, & \text{si } \beta_j = \beta_j^* (\neq \beta_j^o) \end{cases} \tag{4.15}$$

Les résultats (4.14) et (4.15) sont des *versions asymptotiques* (valables uniquement pour n grand) des *résultats exacts* de distribution d'échantillonnage (4.2) et (4.10)-(4.11) sur lesquels nous nous sommes appuyés pour obtenir, respectivement, des intervalles de confiance et des tests d'hypothèse de β_j, ceci sous l'hypothèse de normalité A6 et dans le cas où σ^2 est connu.

Sous l'hypothèse de normalité A6 et pour le cas où σ^2 n'est pas connu, nous avons vu que, pour l'essentiel, le remplacement de σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$ faisait simplement passer de lois normales à des lois de Student.

Asymptotiquement, lorsque n est grand, on peut montrer que le remplacement de σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$ ne modifie pas les distributions d'échantillonnage en jeu, de sorte qu'on a aussi, sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand :

$$\hat{t} = \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)} \approx N(0,1)\,, \quad \text{où } s.\hat{e}.(\hat{\beta}_j) = \sqrt{\hat{s}^2 q_{jj}}\,, \tag{4.16}$$

et donc :

$$\begin{cases} \hat{t}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \approx N(0,1)\,, & \text{si } \beta_j = \beta_j^o \\ \hat{t}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \approx N\left(\dfrac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}, 1\right)\,, & \text{si } \beta_j = \beta_j^* (\neq \beta_j^o) \end{cases} \tag{4.17}$$

Les résultats (4.16) et (4.17) sont des *versions asymptotiques* (valables uniquement pour n grand) des *résultats exacts* de distribution d'échantillonnage (4.7) et (4.12)-(4.13) sur lesquels nous nous sommes appuyés pour obtenir, respectivement, des intervalles de confiance et des tests d'hypothèse de β_j, ceci sous l'hypothèse de normalité A6 et dans le cas où σ^2 est inconnu.

On sait par ailleurs que lorsque $n \to \infty$, la loi de Student $t(n-2)$ tend vers la loi normale $N(0,1)$, de sorte que les quantiles de la loi de Student $t(n-2)$ et de la loi normale $N(0,1)$ s'égalisent.

De ces éléments, on peut conclure[48] que les procédures d'intervalles de confiance et de tests d'hypothèse pour β_j obtenues aux Sections 4.1.2 et 4.2.2, qui sont *exactes en échantillon fini* sous l'hypothèse de normalité A6, restent valables *asymptotiquement*, à titre approximatif, pour n grand, sous les seules hypothèses A1 à A5.

En pratique, on considère généralement qu'une taille d'échantillon $n \geq 30$ est un minimum pour que l'approximation asymptotique soit d'une précision raisonnable.

[48] Le lecteur est invité à refaire le raisonnement en détail. On notera que, dans les calculs des intervalles de confiance et des tests d'hypothèse pour n grand, plutôt que d'utiliser les valeurs critiques (quantiles) de la loi de Student, on pourrait très bien utiliser celles de la loi normale. L'usage veut cependant qu'on utilise en pratique toujours celles de la loi de Student.

Chapitre 5

Prévision, R^2, unités de mesure et forme fonctionnelle

5.1. Prévision

Un des objectifs du modèle de régression est de faire des prévisions. A cet égard, on peut distinguer deux types de prévision :

1- une prévision de l'*espérance* de y sachant x_0 :

$$E(y_0) = \beta_1 + \beta_2 x_0 \,,$$

càd. de la valeur moyenne de y parmi la sous-population pour laquelle $x = x_0$.

2- une prévision de la *valeur* de y sachant x_0 :

$$y_0 = \beta_1 + \beta_2 x_0 + e_0 \,,$$

càd. de la valeur de y pour un individu pris au hasard parmi la sous-population pour laquelle $x = x_0$.

Dans les deux cas, il s'agit d'une prévision *conditionnelle* à la valeur de x_0 que l'on se donne, qui est donc fixe et connue.

On notera au passage que le type de prévision (1) est en fait davantage une *estimation* qu'une *prévision* : contrairement à (2) qui cherche à prédire une variable aléatoire y_0, (1) s'efforce de prédire une quantité non-stochastique $E(y_0)$.

Dans la suite, on suppose que y_0, tout comme les observations $y_1, y_2, ..., y_n$, satisfait les hypothèses A1 à A5 du modèle, plus éventuellement l'hypothèse de normalité A6.

5.1.1. Prévision de l'espérance de y sachant x_0

Sachant x_0, un estimateur / prédicteur naturel de :

$$E(y_0) = \beta_1 + \beta_2 x_0$$

est tout simplement :

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0 \,,$$

soit, sous forme matricielle :

$$\hat{y}_0 = X_0 \hat{\beta} \,, \quad \text{où } X_0 = \begin{bmatrix} 1 & x_0 \end{bmatrix}$$

5.1.1.1. Propriétés d'échantillonnage

β étant estimé, $\hat{y}_0$ et l'*erreur de prévision* $\hat{p}_0 = \hat{y}_0 - E(y_0)$ sont des variables aléatoires, qui ont une certaine distribution d'échantillonnage.

L'espérance de $\hat{y}_0$ et $\hat{p}_0$ sont données par :

$$\begin{aligned} E(\hat{y}_0) &= E(X_0\hat{\beta}) = X_0 E(\hat{\beta}) && (\text{car } X_0 \text{ fixe}) \\ &= X_0\beta = E(y_0) && (\text{car } E(\hat{\beta}) = \beta) \end{aligned}$$

et

$$\begin{aligned} E(\hat{p}_0) &= E(\hat{y}_0 - E(y_0)) = E(\hat{y}_0) - E(y_0) \\ &= E(y_0) - E(y_0) = 0 \end{aligned}$$

Comme $E(\hat{y}_0) = E(y_0)$ et $E(\hat{p}_0) = 0$, on dit que $\hat{y}_0$ est un estimateur / prédicteur *non biaisé* de $E(y_0) = \beta_1 + \beta_2 x_0$.

La variance de la prévision $\hat{y}_0$, qui est égale à la variance de l'erreur de prévision $\hat{p}_0$, est donnée par :

$$\begin{aligned} Var(\hat{y}_0) &= E\left[(\hat{y}_0 - E(\hat{y}_0))^2\right] = E\left[(\hat{y}_0 - E(y_0))^2\right] = Var(\hat{p}_0) \\ &= E\left[\Big(X_0(\hat{\beta} - \beta)\Big)^2\right] && (\text{car } \hat{y}_0 = X_0\hat{\beta} \text{ et} \\ & && \quad E(\hat{y}_0) = E(y_0) = X_0\beta) \\ &= E\left[X_0(\hat{\beta} - \beta)(\hat{\beta} - \beta)' X_0'\right] \\ &= X_0 E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] X_0' && (\text{car } X_0 \text{ fixe}) \,, \end{aligned}$$

soit, puisque $V(\hat{\beta}) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right]$:

$$Var(\hat{y}_0) = X_0 V(\hat{\beta}) X_0' = Var(\hat{p}_0) \tag{5.1}$$

En utilisant les expressions (3.3) - (3.5), on peut vérifier que, sous forme détaillée,

cela donne :

$$Var(\hat{y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] = Var(\hat{p}_0) \tag{5.2}$$

Comme on peut le constater, la variance $Var(\hat{y}_0) = Var(\hat{p}_0)$ de (l'erreur de) la prévision dépend, d'une part, de x_0 (plus particulièrement de son écart $|x_0 - \bar{x}|$ au point moyen de l'échantillon), et d'autre part, de $V(\hat{\beta})$. Ainsi, on peut s'attendre à une prévision de $E(y_0)$ d'autant plus précise que β est estimé de façon précise et que l'on cherche à prédire $E(y_0)$ pour une valeur x_0 proche du point moyen de l'échantillon $\bar{x}$. On notera que lorsque la taille d'échantillon $n \to \infty$, $V(\hat{\beta}) \to 0$ et donc $Var(\hat{y}_0) = Var(\hat{p}_0)$ tend aussi vers 0 : la prévision tend à être 'parfaite', exacte.

Sous l'hypothèse A6 de normalité des y_i, on sait que $\hat{\beta}$ est distribué *de façon exacte* selon une loi normale. Comme $\hat{y}_0 = X_0\hat{\beta}$ et $\hat{y}_0 - E(y_0) = X_0(\hat{\beta} - \beta)$ sont des combinaisons linéaires de $\hat{\beta}$, et qu'une combinaison linéaire d'un vecteur distribué selon une loi normale suit également une loi normale (cf. Section 2.3.1), sous les hypothèses A1 à A6, on a :

$$\hat{y}_0 \sim N(X_0\beta, X_0V(\hat{\beta})X_0') \tag{5.3}$$

et

$$\hat{p}_0 \sim N(0, X_0V(\hat{\beta})X_0') \tag{5.4}$$

Si l'hypothèse A6 de normalité des y_i n'est pas remplie, on sait que $\hat{\beta}$ est seulement *asymptotiquement* distribué selon une loi normale, et les résultats de distribution (5.3) et (5.4) tiennent seulement asymptotiquement (pour $n \to \infty$). Formellement, sous les seules hypothèses A1 à A5, on a ainsi :

$$\frac{\hat{y}_0 - X_0\beta}{\sqrt{X_0V(\hat{\beta})X_0'}} \xrightarrow{d} N(0,1)$$

et

$$\frac{\hat{p}_0}{\sqrt{X_0V(\hat{\beta})X_0'}} \xrightarrow{d} N(0,1)\,,$$

soit, sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand :

$$\hat{y}_0 \approx N(X_0\beta, X_0V(\hat{\beta})X_0')$$

et

$$\hat{p}_0 \approx N(0, X_0V(\hat{\beta})X_0')$$

Avant de voir comment on peut utiliser ces résultats pour construire un intervalle de prévision pour $E(y_0)$, on notera qu'un estimateur *convergent* et *non biaisé* (sous

les hypothèses A1 à A5) de la variance de (l'erreur de) la prévision :

$$Var(\hat{y}_0) = Var(\hat{p}_0) = X_0 V(\hat{\beta}) X_0' = \sigma^2 X_0 (X'X)^{-1} X_0'$$

est simplement obtenu en remplaçant la variance inconnue du terme d'erreur σ^2 par son estimateur (convergent et non biaisé) $\hat{s}^2$:

$$V\hat{a}r(\hat{y}_0) = V\hat{a}r(\hat{p}_0) = X_0 \hat{V}(\hat{\beta}) X_0' = \hat{s}^2 X_0 (X'X)^{-1} X_0' \,,$$

et qu'à partir de $V\hat{a}r(\hat{y}_0) = V\hat{a}r(\hat{p}_0)$, un estimateur *convergent*, mais *pas non biaisé*, de l'écart-type $s.e.(\hat{y}_0) = s.e.(\hat{p}_0)$ de (l'erreur de) la prévision est donné par :

$$s.\hat{e}.(\hat{y}_0) = s.\hat{e}.(\hat{p}_0) = \sqrt{V\hat{a}r(\hat{y}_0)} = \sqrt{V\hat{a}r(\hat{p}_0)}$$

5.1.1.2. Intervalle de prévision

De façon semblable à ce que nous avons fait pour construire des intervalles de confiance pour β_1 et β_2, on peut s'appuyer sur la distribution d'échantillonnage de l'erreur de prévision $\hat{p}_0$ pour construire un *intervalle de prévision* pour $E(y_0)$, càd. un intervalle de valeurs plausibles pour l'espérance de y sachant x_0.

Notons que comme l'estimateur MCO ou les intervalles de confiance, un intervalle de prévision est aussi une *règle de décision*, càd. une recette ou une formule qui décrit comment utiliser les observations d'un échantillon pour établir un intervalle de valeurs plausibles pour la valeur que l'on cherche à prédire, en l'occurrence ici l'espérance de y sachant x_0.

Lorsque l'hypothèse A6 de normalité des y_i est satisfaite, on vient de voir que, sous les hypothèses A1 à A6, l'erreur de prévision :

$$\hat{p}_0 = \hat{y}_0 - E(y_0)$$

est telle que :

$$\hat{p}_0 \sim N(0, X_0 V(\hat{\beta}) X_0') \,, \quad \text{où } V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \,,$$

de sorte que :

$$\hat{z}_{p_0} = \frac{\hat{p}_0}{\sqrt{\sigma^2 X_0 (X'X)^{-1} X_0'}} = \frac{\hat{y}_0 - E(y_0)}{s.e.(\hat{p}_0)} \sim N(0,1) \tag{5.5}$$

On pourrait, en supposant que σ^2 est connu, construire un intervalle de prévision pour $E(y_0)$ en s'appuyant sur le seul résultat d'échantillonnage (5.5). Cela donnerait cependant un intervalle de prévision qui ne peut pas être appliqué en pratique, puisqu'en pratique σ^2 n'est pas connu.

De façon semblable à ce que nous avons déjà fait à plusieurs reprises, on peut contourner ce problème en remplaçant la valeur inconnue de σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$. Ce remplacement de σ^2 par $\hat{s}^2$ a simplement pour effet,

comme on l'a déjà vu dans d'autres circonstances, de faire passer la distribution de $\hat{z}_{p_0}$ d'une loi normale à une loi de Student. En effet, on sait que, sous les hypothèses A1 à A6, on a (cf. Section 4.1.2) :

$$\hat{v} = \frac{(n-2)\hat{s}^2}{\sigma^2} \sim \chi^2(n-2)\,,$$

et on peut par ailleurs montrer que $\hat{z}_{p_0}$ et $\hat{v}$ sont indépendamment distribués, de sorte que de la définition de la loi de Student[49], on a :

$$\hat{t}_{p_0} = \frac{\hat{z}_{p_0}}{\sqrt{\dfrac{\hat{v}}{n-2}}} = \frac{\dfrac{\hat{y}_0 - E(y_0)}{\sqrt{\sigma^2 X_0 (X'X)^{-1} X_0'}}}{\sqrt{\dfrac{\hat{s}^2}{\sigma^2}}} \sim t(n-2)\,,$$

soit, en simplifiant :

$$\hat{t}_{p_0} = \frac{\hat{y}_0 - E(y_0)}{\sqrt{\hat{s}^2 X_0 (X'X)^{-1} X_0'}} = \frac{\hat{y}_0 - E(y_0)}{s.\hat{e}.(\hat{p}_0)} \sim t(n-2) \tag{5.6}$$

$\hat{t}_{p_0}$ suivant une loi de Student $t(n-2)$, on a :

$$I\!P\left(-t_{n-2;1-\frac{\alpha}{2}} \leq \frac{\hat{y}_0 - E(y_0)}{s.\hat{e}.(\hat{p}_0)} \leq t_{n-2;1-\frac{\alpha}{2}}\right) = 1-\alpha\,,$$

où la valeur critique $t_{n-2;1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $t(n-2)$. On en déduit :

$$I\!P\left(\hat{y}_0 - t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{p}_0) \leq E(y_0) \leq \hat{y}_0 + t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{p}_0)\right) = 1-\alpha\,, \tag{5.7}$$

soit un *intervalle de prévision à* $(1-\alpha)\times 100\%$ pour $E(y_0)$:

$$\left[\hat{y}_0 - t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{p}_0)\,;\, \hat{y}_0 + t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{p}_0)\right]\,, \tag{5.8}$$

où $s.\hat{e}.(\hat{p}_0) = \sqrt{\hat{s}^2 X_0 (X'X)^{-1} X_0'} = \sqrt{X_0 \hat{V}(\hat{\beta}) X_0'}$.

Etant donné (5.7), sous les hypothèses A1 à A6, il y a une probabilité $1-\alpha$ que l'intervalle *stochastique* (5.8) recouvre la vraie valeur (inconnue) de $E(y_0) = X_0\beta$.

Appliqué à un échantillon particulier, l'intervalle de prévision (5.8) à $(1-\alpha)\times 100\%$ pour $E(y_0)$[50] synthétise de façon très parlante l'information disponible tant sur le niveau (prévision ponctuelle) que sur la variabilité d'échantillonnage, et donc la précision, de la prévision réalisée.

Pour α fixé, la largeur de l'intervalle de prévision (5.8), qui synthétise la précision de la prévision réalisée, dépend de l'écart-type *estimé* $s.\hat{e}.(\hat{p}_0)$ de l'erreur de prévision

[49] Pour rappel, si $z \sim N(0,1)$, $v \sim \chi^2(m)$ et que z et v sont indépendamment distribués, alors : $t = \frac{z}{\sqrt{\frac{v}{m}}} \sim t(m)$. Cf. l'annexe B de Hill, Griffiths et Lim (2008).

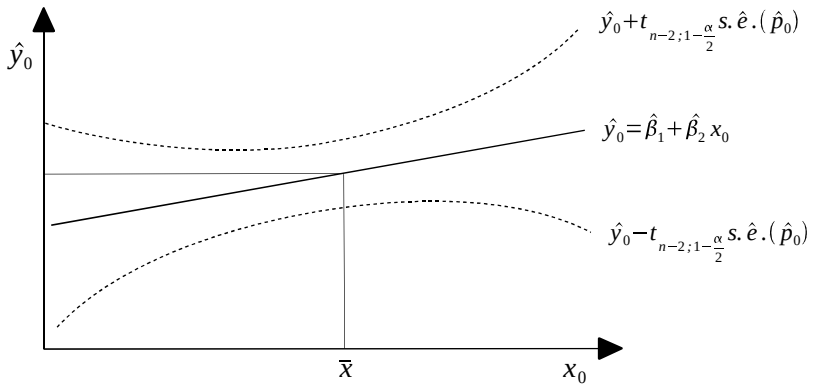
[50] Notons au passage que l'intervalle de prévision (5.8) n'est en fait rien d'autre qu'un intervalle de confiance pour une combinaison linéaire de β_1 et β_2 : $X_0\beta = \beta_1 + \beta_2 x_0$.

$\hat{p}_0$, qui lui-même dépend :

1- de la valeur de x_0 (au travers de $X_0 = \begin{bmatrix} 1 & x_0 \end{bmatrix}$),

2- de la précision d'estimation estimée de β (au travers de $\hat{V}(\hat{\beta})$).

Clairement, plus $\hat{V}(\hat{\beta})$ est petit (au sens matriciel), plus $s.\hat{e}.(\hat{p}_0)$ sera petit. La dépendance de $s.\hat{e}.(\hat{p}_0)$ par rapport à x_0 est illustrée par le graphique ci-dessous :



Graphique 31 : Intervalle de prévision pour $E(y_0)$

On voit que, toutes autres choses étant égales, l'intervalle de prévision est d'autant plus large que x_0 est éloigné de la moyenne empirique $\bar{x}$ des x_i de l'échantillon, ce qui est aisément vérifié si on examine la forme détaillée de $s.\hat{e}.(\hat{p}_0)$ qui est donnée par (elle découle de (5.2)) :

$$s.\hat{e}.(\hat{p}_0) = \sqrt{\hat{s}^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})}$$

On notera encore que comme $\hat{V}(\hat{\beta}) \to 0$ lorsque $n \to \infty$, on a aussi que $s.\hat{e}.(\hat{p}_0) \to 0$ lorsque $n \to \infty$.

L'intervalle de prévision (5.8) suppose que, outre les hypothèses A1 à A5, l'hypothèse optionnelle de normalité A6 du modèle était satisfaite. Comme les procédures d'intervalles de confiance et de tests d'hypothèse pour β_j, cet intervalle de prévision reste toutefois valable lorsqu'on renonce à l'hypothèse de normalité, mais à nouveau seulement asymptotiquement, en grand échantillon. En effet, nous avons vu à la section précédente que, sous les seules hypothèses A1 à A5, on a toujours :

$$\hat{p}_0 \approx N(0, X_0 V(\hat{\beta}) X_0') \,, \quad \text{où } V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \,,$$

de sorte que :

$$\hat{z}_{p_0} = \frac{\hat{p}_0}{\sqrt{\sigma^2 X_0 (X'X)^{-1} X_0'}} = \frac{\hat{y}_0 - E(y_0)}{s.e.(\hat{p}_0)} \approx N(0,1)$$

Asymptotiquement, lorsque n est grand, on peut montrer que le remplacement

de σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$ ne modifie pas la distribution d'échantillonnage en jeu, de sorte qu'on a aussi, sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand :

$$\hat{t}_{p_0} = \frac{\hat{y}_0 - E(y_0)}{\sqrt{\hat{s}^2 X_0 (X'X)^{-1} X_0'}} = \frac{\hat{y}_0 - E(y_0)}{s.\hat{e}.(\hat{p}_0)} \approx N(0,1) \tag{5.9}$$

Le résultat (5.9) est la *version asymptotique* (valable uniquement pour n grand) du *résultat exact* de distribution d'échantillonnage (5.6) sur lequel nous nous sommes appuyé pour obtenir l'intervalle de prévision (5.8) sous l'hypothèse de normalité A6.

Si on se rappelle que lorsque $n \to \infty$, la loi de Student $t(n-2)$ tend vers la loi normale $N(0,1)$, de sorte que les quantiles de la loi de Student $t(n-2)$ et de la loi normale $N(0,1)$ s'égalisent, on peut voir[51] que l'intervalle de prévision (5.8), qui est *exact en échantillon fini* sous l'hypothèse A6 de normalité, reste bien valable *asymptotiquement*, à titre approximatif, pour n grand, sous les seules hypothèses A1 à A5.

5.1.2. Prévision de la valeur de y sachant x_0

Sachant x_0, on peut encore utiliser :

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0 = X_0 \hat{\beta}$$

comme prédicteur de :

$$y_0 = \beta_1 + \beta_2 x_0 + e_0 = X_0 \beta + e_0$$

5.1.2.1. Propriétés d'échantillonnage

L'espérance du prédicteur $\hat{y}_0$ n'est pas égale à la valeur y_0 que l'on cherche à prédire :

$$E(\hat{y}_0) = X_0\beta \neq X_0\beta + e_0 = y_0$$

Cependant, l'*erreur de prévision* $\hat{f}_0 = \hat{y}_0 - y_0$ est elle bien d'espérance nulle. En effet :

$$\begin{aligned} E(\hat{f}_0) &= E(\hat{y}_0 - y_0) = E\left[X_0\hat{\beta} - X_0\beta - e_0\right] \\ &= X_0 E(\hat{\beta}) - X_0\beta - E(e_0) \qquad (\text{car } X_0 \text{ fixe}) \\ &= 0 \qquad (\text{car } E(\hat{\beta}) = \beta \text{ et } E(e_0) = 0) \end{aligned}$$

[51] On notera que, dans le calcul de l'intervalle de prévision pour n grand, plutôt que d'utiliser les valeurs critiques (quantiles) de la loi de Student, on pourrait très bien utiliser celles de la loi normale. L'usage veut cependant qu'on utilise en pratique toujours celles de la loi de Student.

Comme $E(\hat{f}_0) = 0$, on dit encore que $\hat{y}_0$ est un *prédicteur non biaisé* de y_0.

La variance de l'erreur de prévision $\hat{f}_0$ est donnée par :

$$\begin{aligned} Var(\hat{f}_0) &= E\left[(\hat{f}_0 - E(\hat{f}_0))^2\right] = E\left(\hat{f}_0^2\right) && (\text{car } E(\hat{f}_0) = 0) \\ &= E\left[\left(X_0(\hat{\beta} - \beta) - e_0\right)^2\right] && (\text{car } \hat{f}_0 = X_0(\hat{\beta} - \beta) - e_0) \\ &= E\left[X_0(\hat{\beta} - \beta)(\hat{\beta} - \beta)' X_0' - 2X_0(\hat{\beta} - \beta)e_0 + e_0^2\right] \\ &= X_0 E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] X_0 - 2X_0 E\left[(\hat{\beta} - \beta)e_0\right] + E(e_0^2)\,, \end{aligned}$$

où la dernière égalité découle du fait que X_0 est fixe. Comme $E(e_0^2) = \sigma^2$, $E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] = V(\hat{\beta})$ et :

$$\begin{aligned} E\left[(\hat{\beta} - \beta)e_0\right] &= E\left[(X'X)^{-1} X' e e_0\right] && (\text{car } \hat{\beta} - \beta = (X'X)^{-1}X'e) \\ &= (X'X)^{-1} X' E(e e_0) && (\text{car } X \text{ fixe}) \\ &= 0 && (\text{car } E(e e_0) = \begin{bmatrix} Cov(e_1, e_0) \\ \vdots \\ Cov(e_n, e_0) \end{bmatrix} = 0)\,, \end{aligned}$$

on trouve finalement :

$$Var(\hat{f}_0) = \sigma^2 + X_0 V(\hat{\beta}) X_0'\,, \tag{5.10}$$

soit, sous forme détaillée :

$$Var(\hat{f}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \tag{5.11}$$

L'expression (5.10) de la variance $Var(\hat{f}_0)$ de l'erreur de prévision $\hat{f}_0$ est très semblable à l'expression (5.1) de la variance $Var(\hat{p}_0)$ de l'erreur de prévision $\hat{p}_0$. On a :

$$Var(\hat{f}_0) = \sigma^2 + Var(\hat{p}_0)$$

Ainsi, lorsque $\hat{y}_0$ est utilisé en tant que prédicteur de y_0, la variabilité de l'erreur de prévision est supérieure à ce qu'elle est lorsque le même $\hat{y}_0$ est utilisé en tant que prédicteur de $E(y_0)$. Lorsque $\hat{y}_0$ est utilisé en tant que prédicteur de $E(y_0)$, la seule source de variabilité provient du fait que $\hat{\beta}$ est estimé. Lorsque $\hat{y}_0$ est utilisé en tant que prédicteur de y_0, la variabilité de l'erreur de prévision provient du fait que $\hat{\beta}$ est estimé et du terme d'erreur e_0, càd. de la variabilité de y_0 autour de son espérance $E(y_0) = \beta_1 + \beta_2 x_0$, d'où le terme σ^2 supplémentaire qui apparaît dans l'expression de $Var(\hat{f}_0)$.

Rappelons que lorsque $\hat{y}_0$ est utilisé en tant que prédicteur de $E(y_0)$ et que la

taille d'échantillon $n \to \infty$, $V(\hat{\beta}) \to 0$, de sorte que $Var(\hat{p}_0)$ tend aussi vers 0 : la prévision tend à être 'parfaite', exacte. Ce n'est plus le cas lorsque $\hat{y}_0$ est utilisé en tant que prédicteur de y_0 : lorsque la taille d'échantillon $n \to \infty$, on a $Var(\hat{f}_0) \to \sigma^2$, autrement dit, il reste toujours la variabilité associée au terme d'erreur e_0.

Sous l'hypothèse A6 de normalité des y_i, y_0 est par hypothèse distribué de façon normale, et on sait que $\hat{\beta}$ est aussi distribué *de façon exacte* selon une loi normale. Comme $\hat{f}_0 = \hat{y}_0 - y_0 = X_0\hat{\beta} - y_0$ est une combinaison linéaire de $\hat{\beta}$ et y_0, et qu'une combinaison linéaire d'un vecteur distribué selon une loi normale suit également une loi normale (cf. Section 2.3.1), sous les hypothèses A1 à A6, on a :

$$\hat{f}_0 \sim N(0, \sigma^2 + X_0 V(\hat{\beta}) X_0') \tag{5.12}$$

Si l'hypothèse A6 de normalité des y_i n'est pas remplie, on sait que $\hat{\beta}$ est toujours, *asymptotiquement*, distribué selon une loi normale. Mais ce n'est *plus* le cas de y_0. Contrairement à ce que nous avons vu pour le cas de $\hat{p}_0$, le résultat de distribution (5.12) *ne tient donc pas*, même asymptotiquement, pour n grand, sous les seules hypothèses A1 à A5.

Avant de voir comment on peut utiliser le résultat de distribution (5.12) pour construire un intervalle de prévision pour y_0, on notera qu'un estimateur *convergent* et *non biaisé* (sous les hypothèses A1 à A5) de la variance de l'erreur de prévision :

$$Var(\hat{f}_0) = \sigma^2 + X_0 V(\hat{\beta}) X_0' = \sigma^2(1 + X_0(X'X)^{-1}X_0')$$

est à nouveau simplement obtenu en remplaçant la variance inconnue du terme d'erreur σ^2 par son estimateur (convergent et non biaisé) $\hat{s}^2$:

$$\hat{Var}(\hat{f}_0) = \hat{s}^2 + X_0 \hat{V}(\hat{\beta}) X_0' = \hat{s}^2(1 + X_0(X'X)^{-1}X_0')\,,$$

et qu'à partir de $\hat{Var}(\hat{f}_0)$, un estimateur *convergent*, mais *pas non biaisé*, de l'écart-type $s.e.(\hat{f}_0)$ de l'erreur de prévision est donné par :

$$\hat{s.e.}(\hat{f}_0) = \sqrt{\hat{Var}(\hat{f}_0)}$$

5.1.2.2. Intervalle de prévision

De façon semblable à ce que nous avons fait pour l'intervalle de prévision pour $E(y_0)$, en s'appuyant sur la distribution d'échantillonnage de l'erreur de prévision $\hat{f}_0$, on peut construire un *intervalle de prévision* pour y_0, càd. un intervalle de valeurs plausibles pour la valeur de y sachant x_0.

On vient de voir que, lorsque l'hypothèse A6 de normalité des y_i est satisfaite, sous les hypothèses A1 à A6, l'erreur de prévision :

$$\hat{f}_0 = \hat{y}_0 - y_0$$

est telle que :

$$\hat{f}_0 \sim N(0, \sigma^2 + X_0 V(\hat{\beta}) X_0') \,, \quad \text{où } V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \,,$$

de sorte que :

$$\hat{z}_{f_0} = \frac{\hat{f}_0}{\sqrt{\sigma^2(1 + X_0(X'X)^{-1}X_0')}} = \frac{\hat{y}_0 - y_0}{s.e.(\hat{f}_0)} \sim N(0,1) \tag{5.13}$$

Comme précédemment, on pourrait, en supposant que σ^2 est connu, construire un intervalle de prévision pour y_0 en s'appuyant sur le seul résultat d'échantillonnage (5.13). Cela donnerait cependant un intervalle de prévision qui ne peut pas être appliqué en pratique, puisqu'en pratique σ^2 n'est pas connu.

On peut à nouveau contourner ce problème en remplaçant la valeur inconnue de σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$. Ce remplacement de σ^2 par $\hat{s}^2$ a simplement pour effet de faire passer la distribution de $\hat{z}_{f_0}$ d'une loi normale à une loi de Student. En effet, on sait que, sous les hypothèses A1 à A6, on a (cf. Section 4.1.2) :

$$\hat{v} = \frac{(n-2)\hat{s}^2}{\sigma^2} \sim \chi^2(n-2) \,,$$

et on peut encore montrer que $\hat{z}_{f_0}$ et $\hat{v}$ sont indépendamment distribués, de sorte que de la définition de la loi de Student[52], on a :

$$\hat{t}_{f_0} = \frac{\hat{z}_{f_0}}{\sqrt{\frac{\hat{v}}{n-2}}} = \frac{\frac{\hat{y}_0 - y_0}{\sqrt{\sigma^2(1 + X_0(X'X)^{-1}X_0')}}}{\sqrt{\frac{\hat{s}^2}{\sigma^2}}} \sim t(n-2) \,,$$

soit, en simplifiant :

$$\hat{t}_{f_0} = \frac{\hat{y}_0 - y_0}{\sqrt{\hat{s}^2(1 + X_0(X'X)^{-1}X_0')}} = \frac{\hat{y}_0 - y_0}{s.\hat{e}.(\hat{f}_0)} \sim t(n-2)$$

$\hat{t}_{f_0}$ suivant une loi de Student $t(n-2)$, on a :

$$I\!P\left(-t_{n-2;1-\frac{\alpha}{2}} \leq \frac{\hat{y}_0 - y_0}{s.\hat{e}.(\hat{f}_0)} \leq t_{n-2;1-\frac{\alpha}{2}} \right) = 1 - \alpha \,,$$

où la valeur critique $t_{n-2;1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $t(n-2)$. On en déduit :

$$I\!P\left(\hat{y}_0 - t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{f}_0) \leq y_0 \leq \hat{y}_0 + t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{f}_0) \right) = 1 - \alpha \,, \tag{5.14}$$

[52] Pour rappel, si $z \sim N(0,1)$, $v \sim \chi^2(m)$ et que z et v sont indépendamment distribués, alors : $t = \frac{z}{\sqrt{\frac{v}{m}}} \sim t(m)$.

soit un *intervalle de prévision à* $(1-\alpha)\times 100\%$ pour y_0 :

$$\left[\hat{y}_0 - t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{f}_0)\,;\, \hat{y}_0 + t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{f}_0)\right], \tag{5.15}$$

où $s.\hat{e}.(\hat{f}_0) = \sqrt{\hat{s}^2(1 + X_0(X'X)^{-1}X_0')} = \sqrt{\hat{s}^2 + X_0\hat{V}(\hat{\beta})X_0'}$.

Etant donné (5.14), sous les hypothèses A1 à A6, il y a une probabilité $1-\alpha$ que l'intervalle *stochastique* (5.15) recouvre la vraie valeur (inconnue) — et elle-même *stochastique* — de $y_0 = X_0\beta + e_0$.

Appliqué à un échantillon particulier, l'intervalle de prévision (5.15) à $(1-\alpha)\times 100\%$ pour y_0 synthétise de façon très parlante l'information disponible tant sur le niveau (prévision ponctuelle) que sur la variabilité d'échantillonnage, et donc la précision, de la prévision réalisée.

Pour α fixé, la largeur de l'intervalle de prévision (5.15), qui synthétise la précision de la prévision réalisée, dépend de l'écart-type *estimé* $s.\hat{e}.(\hat{f}_0)$ de l'erreur de prévision $\hat{f}_0$, qui lui-même dépend :

1- de la valeur estimée $\hat{s}^2$ de σ^2.

2- de la valeur de x_0 (au travers de $X_0 = \left[\begin{array}{cc} 1 & x_0 \end{array}\right]$),

3- de la précision d'estimation estimée de β (au travers de $\hat{V}(\hat{\beta})$).

L'écart-type estimé $s.\hat{e}.(\hat{f}_0)$ sera d'autant plus petit que $\hat{s}^2$ est petit, que $\hat{V}(\hat{\beta})$ est petit (au sens matriciel), et finalement que x_0 est proche de la moyenne empirique $\bar{x}$ des x_i de l'échantillon, ce qui se vérifie si on examine la forme détaillée de $s.\hat{e}.(\hat{f}_0)$ qui est donnée par (elle découle de (5.11)) :

$$s.\hat{e}.(\hat{f}_0) = \sqrt{\hat{s}^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2})}$$

La dépendance de $s.\hat{e}.(\hat{f}_0)$ par rapport à x_0 peut être représentée de façon semblable au Graphique 31 (cf. p. 75). Simplement, l'intervalle de prévision est ici plus large, puisque $s.\hat{e}.(\hat{f}_0) > s.\hat{e}.(\hat{p}_0)$.

On notera encore que lorsque $n \to \infty$, comme $\hat{V}(\hat{\beta}) \to 0$, on a $s.\hat{e}.(\hat{f}_0) \to \sigma$ et non vers 0 comme c'était le cas pour $s.\hat{e}.(\hat{p}_0)$.

Pour conclure, on notera finalement que si l'hypothèse A6 de normalité des y_i n'est pas remplie, contrairement au cas de l'intervalle de prévision pour $E(y_0)$, l'intervalle de prévision (5.15) pour y_0 *ne tient pas*, même asymptotiquement, pour n grand, sous les seules hypothèses A1 à A5. Il *ne peut* donc *pas* être utilisé, à titre approximatif pour n grand, lorsque l'hypothèse A6 de normalité n'est pas satisfaite. Cela découle du fait que résultat de distribution (5.12) sur lequel on s'est appuyé pour construire (5.15) ne tient pas, même asymptotiquement, pour n grand, sous les seules hypothèses A1 à A5.

5.1.3. Exemple : la fonction de consommation de HGL (2008)

Pour les données de Hill, Griffiths et Lim (2008) considérée à la Section 2.2.3, qui pour rappel considère le modèle de fonction de consommation :

$$y_i = \beta_1 + \beta_2 x_i + e_i \,,$$

où x_i désigne le revenu d'un ménage (en centaines de \$) et y_i les dépenses alimentaires de ce ménage (en \$), on a déjà vu (cf. Section 4.2.4) qu'en utilisant le logiciel GRETL, on obtient le tableau de résultats d'estimation suivant :

Model 1:
OLS, using observations 1-40
Dependent variable: y

	coefficient	std. error	t-ratio	p-value
const	83.4160	43.4102	1.922	0.0622 *
x	10.2096	2.09326	4.877	1.95e-05 ***

Mean dependent var	283.5735	S.D. dependent var	112.6752
Sum squared resid	304505.2	S.E. of regression	89.51700
R-squared	0.385002	Adjusted R-squared	0.368818
F(1, 38)	23.78884	P-value(F)	0.000019
Log-likelihood	-235.5088	Akaike criterion	475.0176
Schwarz criterion	478.3954	Hannan-Quinn	476.2389

De ce tableau de résultats, on peut calculer $\hat{s}^2$ et le prédicteur des dépenses alimentaires $\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0$ pour un revenu de 2000 \$, soit $x_0 = 20$:

$$\begin{aligned} \hat{s}^2 &= \frac{304505,2}{(40-2)} = 8013,29 \\ \hat{y}_i &= 83,42 + 10,21 \times 20 = 287,62\,\$ \end{aligned}$$

Toujours en utilisant GRETL, on obtient pour $\hat{V}(\hat{\beta})$:

Covariance matrix of regression coefficients:

const	x	
1884.44	-85.9032	const
	4.38175	x

Sur base de ce résultat complémentaire, si on note que pour $(n-2) = 38$ et $\alpha = 0,05$, on a $t_{n-2;1-\frac{\alpha}{2}} = t_{38;0,975} = 2,024$, on peut calculer[53], toujours pour $x_0 = 20$:

[53] Notons que $s.\hat{e}.(\hat{p}_0)$ et $s.\hat{e}.(\hat{f}_0)$ peuvent aisément être calculés en utilisant les capacités de calcul matriciel de GRETL. Les quantiles de la loi de Student peuvent de même être obtenus en utilisant les 'Statistical tables' de GRETL.

1- un intervalle de prévision à 95% pour $E(y_0)$:

$$\begin{aligned}\hat{y}_0 \pm t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{p}_0) &= \hat{y}_0 \pm t_{n-2;1-\frac{\alpha}{2}} \sqrt{X_0 \hat{V}(\hat{\beta}) X_0'} \\ &= 287,62 \pm 2,024 \times 14,178 \\ &= [258,92\,;\,316,32]\end{aligned}$$

2- un intervalle de prévision à 95% pour y_0 :

$$\begin{aligned}\hat{y}_0 \pm t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{f}_0) &= \hat{y}_0 \pm t_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{s}^2 + X_0 \hat{V}(\hat{\beta}) X_0'} \\ &= 287,62 \pm 2,024 \times 90,633 \\ &= [104,18\,;\,471,06]\end{aligned}$$

On constate que l'intervalle de prévision pour y_0 est bien plus large que l'intervalle de prévision pour $E(y_0)$. Cela découle simplement du fait que, comme suggéré par le graphique des données reproduit à la Section 2.2.3, les dépenses alimentaires pour un revenu donné varient fortement d'un ménage à l'autre ($\hat{s}^2$ élevé). Ainsi, si on peut prédire avec une bonne précision la valeur moyenne des dépenses alimentaires des ménages ayant un revenu de 2000 \$, on ne peut par contre pas prédire avec précision la valeur des dépenses alimentaires d'un ménage pris au hasard parmi les ménages ayant un revenu de 2000 \$.

5.2. Le coefficient de détermination : R^2

Le coefficient de détermination, communément appelé et noté R^2, fournit une mesure du degré d'ajustement du modèle aux données. Il est reporté par tous les logiciels économétriques[54]. Il est défini comme décrit ci-après.

Une fois le modèle estimé, on peut décomposer chaque observation y_i en une partie *expliquée* par le modèle $\hat{y}_i$, et une partie *non expliquée* ou *résiduelle* $\hat{e}_i$:

$$y_i = \hat{y}_i + \hat{e}_i\,,$$

où $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ et $\hat{e}_i = y_i - \hat{y}_i$. Sous forme matricielle :

$$Y = \hat{Y} + \hat{e} = X\hat{\beta} + \hat{e} \tag{5.16}$$

De (5.16), on peut tirer :

$$\begin{aligned}Y'Y &= (X\hat{\beta} + \hat{e})'(X\hat{\beta} + \hat{e}) \\ &= \hat{\beta}' X'X\hat{\beta} + \hat{\beta}' X'\hat{e} + \hat{e}'X\hat{\beta} + \hat{e}'\hat{e} \\ &= \hat{Y}'\hat{Y} + \hat{e}'\hat{e}\,,\end{aligned}$$

puisque d'après la condition de premier ordre (2.22) définissant $\hat{\beta}$, $X'\hat{e} = 0$. On a

[54] Dans GRETL, il est reporté sous la rubrique 'R-squared'.

donc :

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} \hat{e}_i^2 \tag{5.17}$$

De la relation (2.8) établie à la Section 2.2.1, on sait que :

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Par ailleurs, puisque $\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0$ (cf. l'équation (2.1) à la Section 2.2.1), on a :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \overline{\hat{y}}$$

En soustrayant $n\bar{y}^2 = n\overline{\hat{y}}^2$ des deux membres de (5.17), on obtient dès lors la décomposition :

$$\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} \hat{y}_i^2 - n\overline{\hat{y}}^2 + \sum_{i=1}^{n} \hat{e}_i^2 \,,$$

soit :

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}})^2}_{\text{SCE}} + \underbrace{\sum_{i=1}^{n} \hat{e}_i^2}_{\text{SCR}} \,,$$

où SCT désigne la somme des carrés totaux (centrés), SCE la somme des carrés expliqués (centrés), et SCR la somme des carrés des résidus.

Cette décomposition est connue sous le nom d'*équation d'analyse de la variance* car en divisant ses deux membres par n, on a :

$$\underbrace{Var_e(y_i)}_{\text{Variance totale}} = \underbrace{Var_e(\hat{y}_i)}_{\text{Variance expliquée}} + \underbrace{Var_e(\hat{e}_i)}_{\text{Variance résiduelle}}$$

où $Var_e(.)$ désigne la variance empirique. Notons au passage qu'au contraire de (5.17), cette décomposition *n'est pas valable* si le modèle n'inclut pas une constante (un intercept), car dans ce cas on n'a pas $\bar{y} = \overline{\hat{y}}$.

Le coefficient de détermination, noté R^2, est basé sur cette décomposition. Il est défini par :

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}} = \frac{Var_e(\hat{y}_i)}{Var_e(y_i)}$$

Le R^2 mesure la *part de la variance* des y_i *expliquée* par la régression, ou plus précisément, la part de la variance des y_i qui peut être *linéairement associée* à la

variation des x_i. Par construction, le R^2 est toujours compris entre 0 et 1 :

$$0 \leq R^2 \leq 1$$

avec : - $R^2 = 1$ si et seulement si SCR $= 0$.
- $R^2 = 0$ si et seulement si SCE $= 0$, soit si et seulement si SCT $=$ SCR.

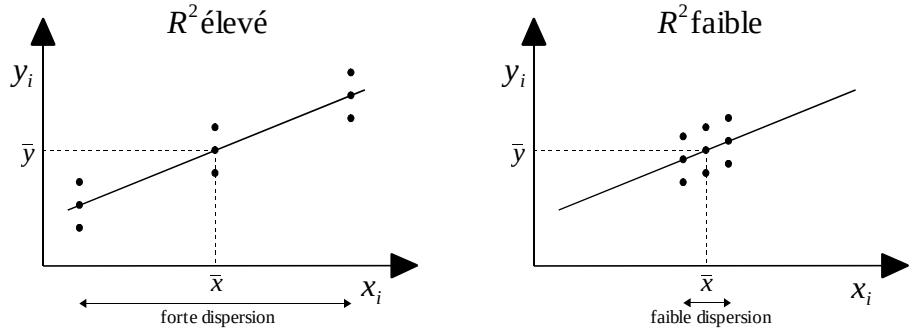
Plusieurs points méritent encore d'être épinglés :

1- La propriété $0 \leq R^2 \leq 1$ ne tient pas nécessairement si le modèle n'inclut pas une constante ou est estimé par une autre méthode que les MCO (ou le MV sous l'hypothèse de normalité).

2- La somme des carrés totaux (SCT) d'une régression est égale à la somme des carrés des résidus (SCR) de la régression des y_i sur une constante (sans autre variable explicative).

3- Le R^2 est une *mesure descriptive*. Un modèle ayant un R^2 élevé n'est pas un 'bon' modèle, ou un modèle 'correct'. Un 'bon' modèle est un modèle qui satisfait les hypothèses sur lesquelles il est fondé : linéarité de l'espérance conditionnelle, homoscédasticité, non-corrélation. Typiquement, le R^2 est plutôt faible (de l'ordre de $0,3$ - $0,5$) lorsqu'on analyse des données en coupe, et (très) élevé ($0,9$ et plus) lorsqu'on analyse des données chronologiques.

4- Le R^2 est souvent interprété comme une mesure globale de la 'capacité prédictive' du modèle. C'est cependant loin d'en être une mesure parfaite. En effet, comme illustré par le graphique ci-dessous, pour les mêmes paramètres estimés $\hat{\beta}_1$ et $\hat{\beta}_2$, et une même somme de carrés des résidus SCR, et donc un même $\hat{s}^2$, le R^2 augmente mécaniquement avec la dispersion des x_i :



Graphique 32 : R^2 et dispersion des x_i

De ce point de vue, $\sqrt{\hat{s}^2}$ semble être une mesure mieux adaptée (même si, contrairement au R^2, elle dépend des unités de mesure, cf. infra). D'autres mesures sont possibles, par exemple, l'erreur absolue moyenne en pourcentage :

$$EAMP = \frac{1}{n} \sum_{i=1}^{n} |\frac{\hat{e}_i}{y_i}|$$

5.2.1. R^2 et corrélation

Dans le cadre du modèle de régression linéaire simple, on peut établir des liens intéressants entre R^2 et *coefficient de corrélation*. On peut ainsi montrer que :

1- le R^2 est égal au carré du coefficient de corrélation empirique $\rho_e(x_i, y_i)$ entre x_i et y_i :

$$R^2 = \left(\rho_e(x_i, y_i)\right)^2 , \tag{5.18}$$

le coefficient de corrélation empirique entre x_i et y_i étant défini par :

$$\rho_e(x_i, y_i) = \frac{Cov_e(x_i, y_i)}{\sqrt{Var_e(x_i)}\sqrt{Var_e(y_i)}} ,$$

où $Cov_e(.,.)$ désigne la covariance empirique[55] et $Var_e(.)$ la variance empirique[56]. L'égalité (5.18) implique que la régression linéaire simple de y_i sur x_i et la régression linéaire simple inverse de x_i sur y_i ont un même R^2 (car $\rho_e(x_i, y_i) = \rho_e(y_i, x_i)$). Cela montre qu'un R^2 élevé ne constitue en aucun cas, comme on pourrait à première vue le croire, une preuve de causalité de x_i vers y_i (ou à l'inverse de y_i vers x_i), de même que, dans la même veine, un $\hat{\beta}_2$ (fortement) significatif ne constitue en aucun cas une telle preuve.

2- le R^2 est égal au carré du coefficient de corrélation empirique $\rho_e(y_i, \hat{y}_i)$ entre y_i et $\hat{y}_i$:

$$R^2 = \left(\rho_e(y_i, \hat{y}_i)\right)^2 , \tag{5.19}$$

où :

$$\rho_e(y_i, \hat{y}_i) = \frac{Cov_e(y_i, \hat{y}_i)}{\sqrt{Var_e(y_i)}\sqrt{Var_e(\hat{y}_i)}}$$

En d'autres termes, le R^2 reflète le degré de corrélation entre y_i et son prédicteur $\hat{y}_i$.

5.3. Unités de mesure

Les paramètres et les statistiques calculés dans le cadre du modèle de régression linéaire simple ne sont pas sans unités de mesure : ils dépendent des unités de mesure des observations (x_i, y_i). Ainsi, dans le modèle :

$$y_i = \beta_1 + \beta_2 x_i + e_i , \tag{5.20}$$

avec :

$$E(e_i) = 0, \quad V(e_i) = \sigma^2 \quad \text{et} \quad Cov(e_i, e_j) = 0 ,$$

où y_i est le poids en kg d'un individu et x_i est sa taille en cm, β_1 et l'erreur e_i se

---

[55] Pour rappel, $Cov_e(x_i, y_i) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

[56] Pour rappel, $Var_e(.) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

mesurent en kg, tandis que $\beta_2 = \frac{dE(y_i)}{dx_i}$ est mesuré en kg/cm et que σ^2 est mesuré en kg^2.

Que se passe-t-il si on change les unités de mesure de x_i et/ou y_i ? Pour le voir, supposons que x_i et y_i soient maintenant mesurés de telle sorte que :

$$\begin{aligned} y_i^* &= ay_i \quad \Leftrightarrow \quad y_i = \frac{y_i^*}{a}\,, \qquad (a>0) \\ x_i^* &= cx_i \quad \Leftrightarrow \quad x_i = \frac{x_i^*}{c}\,, \qquad (c>0)\,, \end{aligned}$$

où x_i^* et y_i^* désignent les variables dans les nouvelles unités de mesure. Par exemple, si y_i est maintenant mesuré en gr plutôt qu'en kg, on a $a = 1000$, et si x_i est maintenant mesuré en m plutôt qu'en cm, on a $c = \frac{1}{100}$.

Exprimé dans les nouvelles unités de mesure, le modèle (5.20) devient :

$$\begin{aligned} & \frac{y_i^*}{a} = \beta_1 + \beta_2 \frac{x_i^*}{c} + e_i \\ \Leftrightarrow \quad & y_i^* = a\beta_1 + \frac{a}{c}\beta_2 x_i^* + ae_i \\ \Leftrightarrow \quad & y_i^* = \beta_1^* + \beta_2^* x_i^* + e_i^*\,, \end{aligned} \tag{5.21}$$

où :

$$\beta_1^* = a\beta_1\,, \quad \beta_2^* = \frac{a}{c}\beta_2\,, \quad e_i^* = ae_i\,,$$

et :

$$E(e_i^*) = 0\,, \quad Var(e_i^*) = a^2 Var(e_i) = a^2\sigma^2 = \sigma^{*2}\,,$$

$$Cov(e_i^*, e_j^*) = E(e_i^* e_j^*) = a^2 E(e_i e_j) = a^2 Cov(e_i, e_j) = 0$$

On constate que si le changement d'unités de mesure affecte les unités de mesure (et donc les valeurs et l'interprétation) des paramètres et de l'erreur, en revanche, la structure des hypothèses du modèle (linéarité, homoscédasticité, non-corrélation) reste elle inchangée : on peut passer sans difficulté du modèle (5.20) au modèle (5.21), et vice-versa. Dans le modèle (5.21), exprimés dans les nouvelles unités de mesure, on a simplement que β_1^* et l'erreur e_i^* se mesurent maintenant en gr (plutôt qu'en kg) , tandis que $\beta_2^* = \frac{dE(y_i^*)}{dx_i^*}$ est maintenant mesuré en gr/m (plutôt qu'en kg/cm) et que σ^{*2} est maintenant mesuré en gr^2 (plutôt qu'en kg^2).

Nous venons de voir l'impact d'un changement d'unités de mesure sur la structure des hypothèses (pas de changement) et les vraies valeurs des paramètres (changements correspondant aux modifications des unités de mesure) du modèle. Qu'en est-il de l'impact de ce changement d'unités de mesure sur les valeurs *estimées* des paramètres (et autres statistiques) du modèle ? Pour le voir, on peut comparer les valeurs estimées sur base du modèle initial aux valeurs estimées sur base du modèle exprimé dans les nouvelles unités de mesure.

L'estimation du modèle initial (5.20) donne :

$$\hat{\beta}_2 = \frac{Cov_e(x_i, y_i)}{Var_e(x_i)}\,, \qquad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$$V\hat{a}r(\hat{\beta}_2) = \frac{\hat{s}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,, \qquad V\hat{a}r(\hat{\beta}_1) = \hat{s}^2 \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2\,, \qquad R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{e}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$V\hat{a}r(\hat{p}_0) = \hat{s}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right), \qquad V\hat{a}r(\hat{f}_0) = \hat{s}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0$$

D'autre part, l'estimation du modèle (5.21) exprimé dans les nouvelles unités de mesure donne :

$$\begin{aligned}
\hat{\beta}_2^* &= \frac{Cov_e(x_i^*, y_i^*)}{Var_e(x_i^*)} = \frac{acCov_e(x_i, y_i)}{c^2 Var_e(x_i)} = \frac{a}{c}\hat{\beta}_2 \\
\hat{\beta}_1^* &= \bar{y}^* - \hat{\beta}_2^* \bar{x}^* = a\bar{y} - \frac{a}{c}\hat{\beta}_2 c\bar{x} = a\hat{\beta}_1 \\
\hat{e}_i^* &= y_i^* - \hat{\beta}_1^* - \hat{\beta}_2^* x_i^* = ay_i - a\hat{\beta}_1 - \frac{a}{c}\hat{\beta}_2 c x_i = a\hat{e}_i \\
\hat{s}^{*2} &= \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^{*2} = \frac{1}{n-2}\sum_{i=1}^{n}(a\hat{e}_i)^2 = a^2\hat{s}^2 \\
V\hat{a}r(\hat{\beta}_2^*) &= \frac{\hat{s}^{*2}}{\sum_{i=1}^{n}(x_i^* - \bar{x}^*)^2} = \frac{a^2\hat{s}^2}{c^2\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{a^2}{c^2}V\hat{a}r(\hat{\beta}_2) \\
V\hat{a}r(\hat{\beta}_1^*) &= \hat{s}^{*2}\frac{\sum_{i=1}^{n} x_i^{*2}}{n\sum_{i=1}^{n}(x_i^* - \bar{x}^*)^2} = a^2\hat{s}^2\frac{c^2\sum_{i=1}^{n} x_i^2}{c^2 n\sum_{i=1}^{n}(x_i - \bar{x})^2} = a^2 V\hat{a}r(\hat{\beta}_1) \\
R^{*2} &= 1 - \frac{\sum_{i=1}^{n} \hat{e}_i^{*2}}{\sum_{i=1}^{n}(y_i^* - \bar{y}^*)^2} = 1 - \frac{a^2\sum_{i=1}^{n} \hat{e}_i^2}{a^2\sum_{i=1}^{n}(y_i - \bar{y})^2} = R^2 \\
V\hat{a}r(\hat{p}_0^*) &= \hat{s}^{*2}\left(\frac{1}{n} + \frac{(x_0^* - \bar{x}^*)^2}{\sum_{i=1}^{n}(x_i^* - \bar{x}^*)^2}\right) = a^2\hat{s}^2\left(\frac{1}{n} + \frac{c^2(x_0 - \bar{x})^2}{c^2\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = a^2 V\hat{a}r(\hat{p}_0)
\end{aligned}$$

$$\begin{aligned} \hat{Var}(\hat{f}_0^*) &= \hat{s}^{*2}(1+\frac{1}{n}+\frac{(x_0^*-\bar{x}^*)^2}{\sum_{i=1}^{n}(x_i^*-\bar{x}^*)^2}) = a^2\hat{s}^2(1+\frac{1}{n}+\frac{c^2(x_0-\bar{x})^2}{c^2\sum_{i=1}^{n}(x_i-\bar{x})^2}) = a^2\hat{Var}(\hat{f}_0) \\ \hat{y}_0^* &= \hat{\beta}_1^*+\hat{\beta}_2^* x_0^* = a\hat{\beta}_1+\frac{a}{c}\hat{\beta}_2 c x_0 = a\hat{y}_0 \end{aligned}$$

Des expressions ci-dessus, on peut voir que :

1- Les paramètres estimés $\hat{\beta}_1^*$, $\hat{\beta}_2^*$ et $\hat{s}^{*2}$ sont reliés à $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{s}^2$ de la même façon que leurs vraies valeurs β_1^*, β_2^* et σ^{*2} sont reliés à β_1, β_2 et σ^2 : ils sont modifiés de la même façon que le changement d'unités de mesure,

2- les t-statistiques $\hat{t}_o^* = \frac{\hat{\beta}_j^*}{s.\hat{e}.(\hat{\beta}_j^*)}$ de test de H_0 : $\beta_j^* = 0$ contre H_1 : $\beta_j^* \neq 0$ sont inchangées. Il en est de même pour tous les t-tests (bilatéraux ou unilatéraux) si on ajuste la valeur testée sous H_0 de la même façon que le changement d'unités de mesure,

3- Les intervalles de confiance pour β_j^* sont modifiés de la même façon que le changement d'unités de mesure,

4- le R^2 est inchangé : il ne dépend pas des unités de mesure,

5- les intervalles de prévision pour $\hat{y}_0^*$ et $E(\hat{y}_0^*)$ sont modifiés de la même façon que le changement d'unités de mesure de y_i. Ils sont inchangés pour un changement d'unités de mesure de x_i si on ajuste bien la valeur de x_0 de la même façon.

Plutôt que de changer les unités de mesure de x_i et/ou y_i en les *multipliant* par une constante, il arrive que l'on *ajoute* une constante à x_i et/ou y_i. Dans ce cas, seul l'intercept du modèle est affecté. En effet, si on modifie x_i et y_i de telle sorte que :

$$\begin{aligned} y_i^* &= y_i + a \quad \Leftrightarrow \quad y_i = y_i^* - a \\ x_i^* &= x_i + c \quad \Leftrightarrow \quad x_i = x_i^* - c\,, \end{aligned}$$

le modèle initial (5.20) devient :

$$\begin{aligned} y_i^* - a &= \beta_1 + \beta_2\,(x_i^* - c) + e_i \\ \Leftrightarrow \qquad y_i^* &= (\beta_1 + a - c\beta_2) + \beta_2 x_i^* + e_i \\ \Leftrightarrow \qquad y_i^* &= \beta_1^* + \beta_2 x_i^* + e_i\,, \end{aligned}$$

où :

$$\beta_1^* = \beta_1 + a - c\beta_2$$

Dans ce cas, on peut montrer que :

1- concernant β_2, estimation, t-tests et intervalle de confiance restent inchangés,

2- concernant β_1, estimation (on aura : $\hat{\beta}_1^* = \hat{\beta}_1 + a - c\hat{\beta}_2$), t-tests et intervalle de confiance sont modifiés,

3- le R^2 et $\hat{s}^2$ restent inchangés,

4- les intervalles de prévision pour $\hat{y}_0^*$ et $E(\hat{y}_0^*)$ sont translatés de la même façon

que y_i. Ils restent inchangés pour l'ajout d'une constante à x_i si on ajuste bien la valeur de x_0 de la même façon.

5.4. Forme fonctionnelle

Comme on l'a déjà évoqué, l'hypothèse de linéarité du modèle standard requiert seulement que le modèle soit linéaire dans les *paramètres*, pas nécessairement dans les *variables*. Ainsi, l'ensemble des propriétés et procédures d'inférence décrites jusqu'ici sont valables pour la classe de modèles :

$$y_i^* = \beta_1 + \beta_2 x_i^* + e_i \,,$$

où :

$$\begin{aligned} y_i^* &= f_1(y_i), \qquad \text{(i.e., une fonction } \textit{connue} \text{ de } y_i) \\ x_i^* &= f_2(x_i), \qquad \text{(i.e., une fonction } \textit{connue} \text{ de } x_i) \end{aligned}$$

et :

$$E(e_i) = 0, \quad V(e_i) = \sigma^2 \quad \text{et} \quad Cov(e_i, e_j) = 0$$

De cette façon, le modèle de régression linéaire simple permet de modéliser des relations *non-linéaires* entre *variables*. Les formes non-linéaires les plus couramment utilisées sont décrites ci-dessous.

5.4.1. Le modèle lin-log

Le modèle lin-log s'écrit :

$$y_i = \beta_1 + \beta_2 \ln x_i + e_i \,, \qquad (x_i > 0)$$

On notera qu'il ne peut être utilisé que si tous les x_i sont strictement positifs.

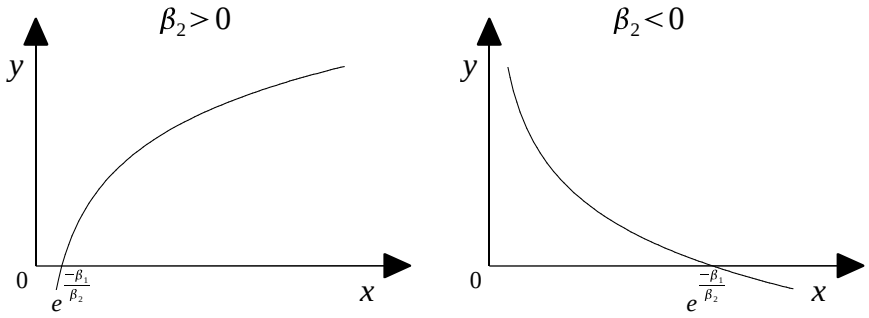
Pour la partie systématique $y = \beta_1 + \beta_2 \ln x$ du modèle, on a :

$$\frac{dy}{dx} = \beta_2 \frac{1}{x}$$

et

$$E_{y,x} = \frac{dy}{dx}\frac{x}{y} = \beta_2 \frac{1}{x}\frac{x}{y} = \frac{\beta_2}{y} = \frac{\beta_2}{\beta_1 + \beta_2 \ln x}$$

Graphiquement :



Graphique 33 : $y = \beta_1 + \beta_2 \ln x$

Dans ce modèle, le paramètre β_2 s'interprète comme une semi-élasticité :

$$\beta_2 = \frac{dy}{d \ln x} = \frac{dy}{\frac{dx}{x}}$$

β_2 mesure la variation *absolue* de y pour une variation *relative* (unitaire) de x.

Le paramètre β_2 étant une semi-élasticité, il est insensible à une modification des unités de mesure de x_i. Une telle modification d'unités de mesure n'a d'influence que sur l'intercept du modèle. En effet, si les unités de mesure de x_i sont modifiées de telle sorte que :

$$x_i^* = cx_i \quad \Leftrightarrow \quad \ln x_i^* = \ln c + \ln x_i \quad \Leftrightarrow \quad \ln x_i = \ln x_i^* - \ln c\,, \qquad (c > 0)\,,$$

le modèle initial devient :

$$\begin{aligned} & y_i = \beta_1 + \beta_2(\ln x_i^* - \ln c) + e_i \\ \Leftrightarrow \qquad & y_i = \beta_1^* + \beta_2 \ln x_i^* + e_i\,, \end{aligned}$$

où :

$$\beta_1^* = \beta_1 - \beta_2 \ln c$$

On constate qu'une modification des unités de mesure de x_i a, dans ce modèle, les mêmes effets que l'ajout d'une constante à x_i analysé à la Section 5.3.

Ce modèle pourrait par exemple être utilisé pour modéliser une fonction de consommation (avec $\beta_2 > 0$), dont la propension marginale à consommer est décroissante.

5.4.2. Le modèle log-lin

Le modèle log-lin, encore appelé modèle exponentiel, s'écrit :

$$\ln y_i = \beta_1 + \beta_2 x_i + e_i\,, \qquad (y_i > 0)$$

On notera qu'il ne peut être utilisé que si tous les y_i sont strictement positifs.

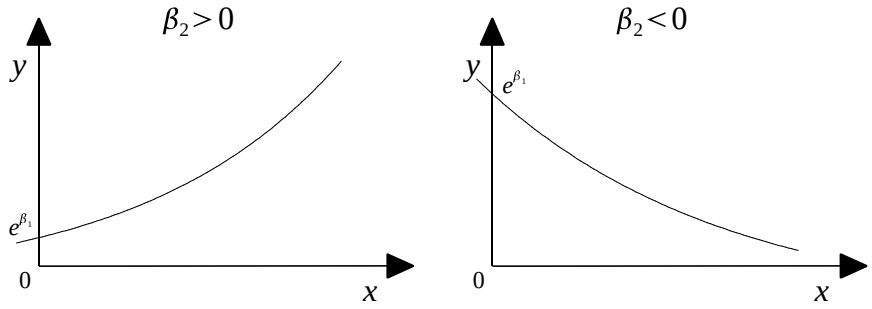
Pour la partie systématique $\ln y = \beta_1 + \beta_2 x \Leftrightarrow y = e^{\beta_1+\beta_2 x}$ du modèle, on a :

$$\frac{dy}{dx} = \beta_2 e^{\beta_1+\beta_2 x} = \beta_2 y$$

et

$$E_{y,x} = \frac{dy}{dx}\frac{x}{y} = \beta_2 x$$

Graphiquement :



Graphique 34 : $\ln y = \beta_1 + \beta_2 x \Leftrightarrow y = e^{\beta_1+\beta_2 x}$

Dans ce modèle, le paramètre β_2 s'interprète aussi comme une semi-élasticité :

$$\beta_2 = \frac{d \ln y}{dx} = \frac{\frac{dy}{y}}{dx}$$

β_2 mesure la variation *relative* de y pour une variation *absolue* (unitaire) de x.

Le paramètre β_2 est ici insensible à une modification des unités de mesure de y_i. Une telle modification d'unités de mesure n'a d'influence que sur l'intercept du modèle. En effet, si les unités de mesure de y_i sont modifiées de telle sorte que :

$$y_i^* = a y_i \quad \Leftrightarrow \quad \ln y_i^* = \ln a + \ln y_i \quad \Leftrightarrow \quad \ln y_i = \ln y_i^* - \ln a\,, \qquad (a > 0)\,,$$

le modèle initial devient :

$$\begin{aligned} & \ln y_i^* - \ln a = \beta_1 + \beta_2 x_i + e_i \\ \Leftrightarrow \qquad & \ln y_i^* = \beta_1^* + \beta_2 x_i + e_i\,, \end{aligned}$$

où :

$$\beta_1^* = \beta_1 + \ln a$$

On constate qu'une modification des unités de mesure de y_i a, dans ce modèle, les mêmes effets que l'ajout d'une constante à y_i analysé à la Section 5.3.

Un usage classique de ce modèle est son utilisation pour modéliser une fonction de salaire (salaire en fonction du niveau d'éducation, avec $\beta_2 > 0$), pour laquelle on peut s'attendre à ce que le rendement marginal, en termes de salaire, d'une année

d'étude supplémentaire soit croissant.

5.4.3. Le modèle log-log

Le modèle log-log s'écrit :

$$\ln y_i = \beta_1 + \beta_2 \ln x_i + e_i\,, \qquad (x_i > 0,\ y_i > 0)$$

On notera qu'il ne peut être utilisé que si tous les x_i et tous les y_i sont strictement positifs.

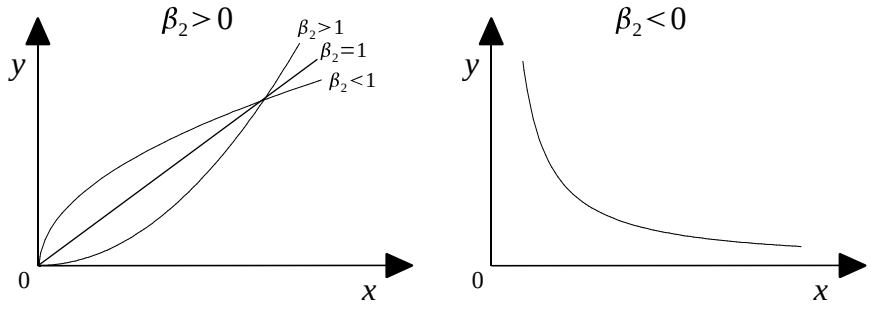
Pour la partie systématique $\ln y = \beta_1 + \beta_2 \ln x \Leftrightarrow y = e^{\beta_1} x^{\beta_2}$ du modèle, on a :

$$\frac{dy}{dx} = \beta_2 e^{\beta_1} x^{\beta_2 - 1} = \beta_2 \frac{e^{\beta_1} x^{\beta_2}}{x} = \beta_2 \frac{y}{x}$$

et

$$E_{y,x} = \frac{dy}{dx}\frac{x}{y} = \beta_2$$

Graphiquement :



Graphique 35 : $\ln y = \beta_1 + \beta_2 \ln x \Leftrightarrow y = e^{\beta_1} x^{\beta_2}$

Dans ce modèle, le paramètre β_2 s'interprète comme une élasticité :

$$\beta_2 = \frac{d \ln y}{d \ln x} = \frac{\frac{dy}{y}}{\frac{dx}{x}}$$

β_2 mesure la variation *relative* de y pour une variation *relative* (unitaire) de x.

Le paramètre β_2 étant une élasticité, il est insensible à une modification des unités de mesure de x_i et de y_i. Une telle modification d'unités de mesure n'a d'influence que sur l'intercept du modèle. En effet, si les unités de mesure de x_i et de y_i sont modifiées de telle sorte que :

$$\begin{array}{rcl} y_i^* = a y_i & \Leftrightarrow & \ln y_i^* = \ln a + \ln y_i \Leftrightarrow \ln y_i = \ln y_i^* - \ln a\,, \quad (a > 0) \\ x_i^* = c x_i & \Leftrightarrow & \ln x_i^* = \ln c + \ln x_i \Leftrightarrow \ln x_i = \ln x_i^* - \ln c\,, \quad (c > 0)\,, \end{array}$$

le modèle initial devient :

$$\begin{aligned} & \ln y_i^* - \ln a = \beta_1 + \beta_2(\ln x_i^* - \ln c) + e_i \\ \Leftrightarrow \qquad & \ln y_i^* = \beta_1^* + \beta_2 \ln x_i^* + e_i \,, \end{aligned}$$

où :

$$\beta_1^* = \beta_1 + \ln a - \beta_2 \ln c$$

On constate encore qu'une modification des unités de mesure de x_i et/ou de y_i a, dans ce modèle, les mêmes effets que l'ajout d'une constante à x_i et/ou y_i analysé à la Section 5.3.

Des usages classiques de ce modèle sont son utilisation pour modéliser une fonction de demande (quantité en fonction du prix, avec $\beta_2 < 0$), une fonction d'offre (quantité en fonction du prix, avec $\beta_2 > 0$), ou encore une fonction de production (output en fonction d'un input, avec $\beta_2 > 0$).

5.4.4. Remarques

1- Si ils constituent bien les modèles non-linéaires (dans les variables) les plus utilisés en pratique, les modèles lin-log, log-lin et log-log ne sont pas les seules formes fonctionnelles non-linéaires possibles. A titre d'exemple, on pourrait considérer :

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i^2 + e_i \\ \ln y_i &= \beta_1 + \beta_2 \frac{1}{x_i} + e_i \,, \end{aligned}$$

etc... De façon générale, le choix d'une forme fonctionnelle spécifique peut être guidé par les caractéristiques attendues (en particulier en termes de dérivée et d'élasticité) à priori de la relation théorique d'intérêt.

2- Au modèle log-log :

$$\ln y_i = \beta_1 + \beta_2 \ln x_i + e_i \,,$$

correspond, pour y_i lui-même, le modèle *non-linéaire* à *erreur multiplicative* :

$$y_i = e^{\beta_1} x^{\beta_2} v_i \,,$$

où[57] :

$$v_i = e^{e_i} \simeq 1 + e_i$$

s'interprète comme une erreur *relative* (i.e., proportionnelle), et dont l'espérance et la variance (conditionnelle à x_i) sont approximativement (si la variance $\sigma^2 =$

[57] L'approximation $e^x \simeq 1 + x$ tient pour x au voisinnage de zéro.

$Var(e_i)$ est petite, et donc e_i pas trop éloigné de zéro) égales à[58] :

$$\begin{aligned} E(y_i) &\simeq e^{\beta_1} x^{\beta_2} \\ Var(y_i) &\simeq \sigma^2 \left[e^{\beta_1} x^{\beta_2}\right]^2 = \sigma^2 \left[E(y_i)\right]^2 \end{aligned}$$

De même, au modèle log-lin :

$$\ln y_i = \beta_1 + \beta_2 x_i + e_i \,,$$

correspond, pour y_i lui-même, le modèle *non-linéaire* à *erreur multiplicative* :

$$y_i = e^{\beta_1 + \beta_2 x_i} v_i \,,$$

où, comme ci-dessus :

$$v_i = e^{e_i} \simeq 1 + e_i$$

s'interprète comme une erreur *relative* (i.e., proportionnelle), et dont l'espérance et la variance (conditionnelle à x_i) sont approximativement (si la variance $\sigma^2 = Var(e_i)$ est petite, et donc e_i pas trop éloigné de zéro) égales à[59] :

$$\begin{aligned} E(y_i) &\simeq e^{\beta_1 + \beta_2 x_i} \\ Var(y_i) &\simeq \sigma^2 \left[e^{\beta_1 + \beta_2 x_i}\right]^2 = \sigma^2 \left[E(y_i)\right]^2 \end{aligned}$$

On voit ainsi qu'au modèle log-log et log-lin correspondent des modèles non seulement *non-linéaire* pour l'espérance (conditionnelle) de y_i, mais aussi *hétéroscédastiques*, dont la variance (conditionnelle) de y_i est proportionnelle au carré de son espérance (et donc l'écart-type de y_i proportionnel à l'espérance). Cette faculté qu'ont ces modèles de rendre compte, au travers d'un modèle de régression linéaire simple, de relations non seulement non-linéaires mais aussi hétéroscédastiques (en conséquence de l'erreur proportionnelle) est une des principales raisons de leur très fréquente utilisation en pratique.

3- Comme pour tout modèle de régression standard, dans un modèle log-log ou log-lin, un *prédicteur ponctuel* non biaisé de :

$$\ln y_0 = X_0 \beta + e_0 \,,$$

où $X_0 = \left[\begin{array}{cc} 1 & \ln x_0 \end{array}\right]$ dans le cas du modèle log-log et $X_0 = \left[\begin{array}{cc} 1 & x_0 \end{array}\right]$ dans le cas du modèle log-lin, est donné par :

$$\widehat{\ln y_0} = X_0 \hat{\beta}$$

et un *intervalle de prévision à* $(1-\alpha)\times 100\%$ pour $\ln y_0$ est donné par :

$$\left[\widehat{\ln y_0} - t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{f}_0) \,;\, \widehat{\ln y_0} + t_{n-2;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{f}_0)\right] \,,$$

où $s.\hat{e}.(\hat{f}_0) = \sqrt{\hat{s}^2 + X_0 \hat{V}(\hat{\beta}) X_0'}$.

---

[58] Ces expressions sont obtenues en utilisant l'approximation $v_i = e^{e_i} \simeq 1 + e_i$ et les propriétés de e_i : $E(e_i) = 0$ et $Var(e_i) = \sigma^2$.

[59] A nouveau, ces expressions sont obtenues en utilisant l'approximation $v_i = e^{e_i} \simeq 1 + e_i$ et les propriétés de e_i : $E(e_i) = 0$ et $Var(e_i) = \sigma^2$.

De ce prédicteur ponctuel et de cet intervalle de prévision pour $\ln y_0$, on peut déduire un *prédicteur ponctuel* et un *intervalle de prévision à* $(1-\alpha)\times 100\%$ pour y_0 donnés respectivement par :

$$\hat{y}_0 = e^{\widehat{\ln y_0}} = e^{X_0\hat{\beta}}$$

et

$$\left[e^{\widehat{\ln y_0} - t_{n-2;1-\frac{\alpha}{2}}\, s.\hat{e}.(\hat{f}_0)}\,;\, e^{\widehat{\ln y_0} + t_{n-2;1-\frac{\alpha}{2}}\, s.\hat{e}.(\hat{f}_0)}\right]$$

càd. obtenu en prenant simplement l'exponentielle du prédicteur ponctuel et des bornes de l'intervalle de prévision pour $\ln y_0$. Notons qu'au contraire de $\widehat{\ln y_0}$ qui est un prédicteur *non biaisé* de $\ln y_0$, $\hat{y}_0 = e^{\widehat{\ln y_0}}$ n'est *pas* un prédicteur *non biaisé*[60] de y_0. Notons également que la validité de l'intervalle de prévision ci-dessus requiert l'hypothèse A6 de normalité.

[60] En pratique son biais est cependant faible si la variance $\sigma^2 = Var(e_i)$ est petite. Si la variance $\sigma^2 = Var(e_i)$ n'est pas petite, un meilleur prédicteur ponctuel de y_0 est donné par $\hat{y}_0 = \hat{\alpha}e^{\widehat{\ln y_0}} = \hat{\alpha}e^{X_0\hat{\beta}}$, où $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{e}_i}$. Nous ignorerons cette complication dans le cadre de ce cours.

Chapitre 6

Le modèle de régression linéaire multiple

6.1. Du modèle économique au modèle économétrique

6.1.1. Un modèle économique

La théorie du capital humain de G. Becker (1964)[61] suggère que le salaire obtenu par un individu est fonction de sa productivité, qui elle-même dépend de son niveau d'éducation et de son expérience professionnelle. De façon formelle, cette assertion peut être décrite par la relation théorique :

$$y = f(x_2, x_3)\,, \text{ avec } \frac{\partial y}{\partial x_2} > 0 \text{ et } \frac{\partial y}{\partial x_3} > 0,$$

où y = salaire, x_2 = nombre d'années d'étude et x_3 = nombre d'années d'expérience.

6.1.2. Le modèle économétrique

Comme dans le cas du modèle de régression simple, on cherche une *contrepartie empirique* de la relation théorique $y = f(x_2, x_3)$, une contrepartie empirique prenant la forme d'un *modèle probabiliste paramétré*, et on regarde les données dont on dispose comme des *réalisations particulières* des variables aléatoires de ce modèle, pour une valeur particulière des paramètres du modèle.

Le plus simple est de raisonner en supposant que les observations dont on dispose sont des données en coupe obtenues par tirages aléatoires d'individus dans une population. Comme dans le cas du modèle de régression simple, le modèle obtenu

[61] Becker, G.S. (1964), *Human Capital : A Theoretical and Empirical Analysis, with Special Reference to Education*, University of Chicago Press.

pourra s'appliquer à des données chronologiques en se plaçant dans une optique de modélisation (cf. Section 2.1.2).

Au travers de l'épreuve aléatoire 'tirer un individu au hasard dans la population et noter la valeur de son salaire y, de son niveau d'étude x_2 et de son niveau d'expérience x_3', on peut représenter la population par une *distribution de probabilité jointe* $f(y, x_2, x_3)$, inconnue et à priori complexe, qui correspond à la distribution de fréquence des triplets de variables (y, x_2, x_3) dans la population.

Lorsqu'on cherche à expliquer y en fonction de x_2 et de x_3, l'information pertinente est concentrée dans la *distribution conditionnelle* $f(y|x_2, x_3)$ qui, pour chaque valeur du couple (x_2, x_3), correspond à la distribution de fréquence des différentes valeurs de y dans la population.

Si on cherchait à expliquer y en fonction de x_2 seulement, la distribution conditionnelle pertinente serait $f(y|x_2)$ qui, pour chaque valeur de x_2 (quel que soit la valeur de x_3), correspond à la distribution de fréquence des différentes valeurs de y dans la population. La distribution conditionnelle $f(y|x_2)$ peut être obtenue de la distribution jointe $f(y, x_2)$, qui elle-même peut être obtenue (par marginalisation) de $f(y, x_2, x_3)$.

La distribution conditionnelle $f(y|x_2, x_3)$ peut être résumée par l'*espérance conditionnelle* de y sachant (x_2, x_3) — aussi appelée *courbe de régression* de y en (x_2, x_3) — qui, pour chaque valeur du couple (x_2, x_3), correspond à la valeur moyenne de y dans la population. Il en est de même pour la distribution conditionnelle $f(y|x_2)$, qui peut être résumée par l'*espérance conditionnelle* de y sachant x_2 — aussi appelée *courbe de régression* de y en x_2 — qui, pour chaque valeur de x_2 (quel que soit la valeur de x_3), correspond à la valeur moyenne de y dans la population. De manière générale, on a :

$$E(y|x_2, x_3) = g(x_2, x_3) \qquad \text{(i.e., une fonction de } x_2 \text{ et } x_3\text{)}$$

et

$$E(y|x_2) = g^*(x_2) \qquad \text{(i.e., une fonction de } x_2\text{)}$$

Comme dans le cas du modèle de régression simple, l'espérance conditionnelle de y sachant (x_2, x_3) constitue, dans le modèle de régression multiple, la contrepartie empirique de la relation théorique $y = f(x_2, x_3)$ d'intérêt.

Plusieurs points méritent d'être épinglés :

1- Les espérances conditionnelles $E(y|x_2, x_3) = g(x_2, x_3)$ et $E(y|x_2) = g^*(x_2)$ ne sont pas sans liens. On peut en effet montrer que[62] :

$$E(y|x_2) = \sum_{x_3} E(y|x_2, x_3) f(x_3|x_2)\,,$$

autrement dit que $E(y|x_2) = g^*(x_2)$ est, pour chaque valeur de x_2, une moyenne

[62] Dans le cas continu, $E(y|x_2) = \int_{-\infty}^{\infty} E(y|x_2, x_3) f(x_3|x_2) dx_3$

pondérée par les probabilités conditionnelles $f(x_3|x_2)$ des $E(y|x_2, x_3) = g(x_2, x_3)$ évaluées aux différentes valeurs possibles de x_3.

2- Si les espérances conditionnelles $E(y|x_2, x_3) = g(x_2, x_3)$ et $E(y|x_2) = g^*(x_2)$ ne sont pas sans liens, elles peuvent néanmoins être très différentes. Ainsi, on peut très bien avoir :

$$E(y|x_2) = \alpha \qquad \text{(i.e., une constante, ne dépend pas de } x_2)$$

et

$$E(y|x_2, x_3) = g(x_2, x_3) \qquad \text{(i.e., une fonction de } x_2 \text{ et } x_3)$$

De même, on peut très bien avoir :

$$E(y|x_2) = g^*(x_2) \qquad \text{(i.e., une fonction de } x_2)$$

et

$$E(y|x_2, x_3) = g(x_3) \qquad \text{(i.e., une fonction de } x_3 \text{ seulement)}$$

Par contre, si $E(y|x_2, x_3) = \alpha$ (i.e., une constante), on a nécessairement $E(y|x_2) = \alpha$ (i.e., une constante).
De manière générale, une variable peut ainsi apparaître ou non pertinente selon l'*ensemble* des variables conditionnantes pris en compte. De même, l'effet marginal $\frac{\partial E(y|.)}{\partial x_j}$ d'une variable x_j sera généralement différent selon l'ensemble des variables conditionnantes considéré.

3- Les espérances conditionnelles $E(y|x_2, x_3) = g(x_2, x_3)$ et $E(y|x_2) = g^*(x_2)$ répondent à des questions différentes concernant la relation entre y et x_2 dans la population : $E(y|x_2)$ représente la façon dont y dépend de x_2, aucune autre variable n'étant maintenue constante, tandis de $E(y|x_2, x_3)$ représente la façon dont y dépend de x_2, la variable x_3 étant maintenue constante.
Dans cette optique, pour obtenir la relation entre y et x_2 *toutes autres choses étant égales*, autrement dit ce à quoi la théorie économique fait généralement référence lorsqu'elle parle de l'existence d'une relation entre deux variables, il faut considérer $E(y|x_2, x_3, ..., x_k)$, où $x_3, ..., x_k$ est l'ensemble des variables (autres que x_2) qui influencent systématiquement y.

4- A priori, tout choix d'ensemble de variables conditionnantes est légitime, en particulier pour la prévision. Tout dépend de la question à laquelle on cherche à répondre et, de façon plus pragmatique, de l'information (des variables) disponibles.

5- Finalement, on notera que si le choix d'un ensemble de variables conditionnantes plus large met par définition en lumière une information plus précise, il y a un revers à la médaille : une relation est a priori d'autant plus difficile à modéliser par une forme paramétrique simple et à estimer avec précision que l'ensemble des variables conditionnantes est large.

Avant de poursuivre, illustrons les concepts évoqués ci-avant pour une population hypothétique dont la distribution (discrète) jointe du salaire ($= y$), du nombre

d'années d'étude (= x_2) et du nombre d'années d'expérience (= x_3) est donnée par :

$f(y, x_2, x_3)$		1000	1500	2000	2500
$x_2 = 12$	$x_3 = 5$	0,05	0,03	0,02	0
	$x_3 = 10$	0,08	0,2	0,08	0,04
$x_2 = 16$	$x_3 = 5$	0,08	0,2	0,12	0
	$x_3 = 10$	0	0,02	0,05	0,03

De la distribution jointe $f(y, x_2, x_3)$, on peut déduire les distributions (marginales) jointes $f(y, x_2)$ et $f(x_2, x_3)$, la distribution marginale $f(x_2)$, et la distribution conditionnelle $f(x_3|x_2)$. Elles sont données[63] par :

$$f(y, x_2) = \sum_{x_3} f(y, x_2, x_3)\,, \quad f(x_2, x_3) = \sum_{y} f(y, x_2, x_3)\,,$$
$$f(x_2) = \sum_{x_3} f(x_2, x_3) \quad \text{et} \quad f(x_3|x_2) = \frac{f(x_2, x_3)}{f(x_2)}$$

On obtient :

$f(y, x_2)$	1000	1500	2000	2500
$x_2 = 12$	0,13	0,23	0,1	0,04
$x_2 = 16$	0,08	0,22	0,17	0,03

$f(x_2, x_3)$	$x_3 = 5$	$x_3 = 10$	$f(x_2)$
$x_2 = 12$	0,1	0,4	0,5
$x_2 = 16$	0,4	0,1	0,5

et

$f(x_3|x_2)$	$x_3 = 5$	$x_3 = 10$
$x_2 = 12$	0,2	0,8
$x_2 = 16$	0,8	0,2

Des distributions jointes $f(y, x_2, x_3)$ et $f(x_2, x_3)$, on peut déduire la distribution conditionnelle et l'espérance conditionnelle de y sachant (x_2, x_3). De même, de la distribution jointe $f(y, x_2)$ et de la distribution marginale $f(x_2)$, on peut déduire la distribution conditionnelle et l'espérance conditionnelle de y sachant x_2 seulement. Elles sont données[64] par :

$$f(y|x_2, x_3) = \frac{f(y, x_2, x_3)}{f(x_2, x_3)} \quad \text{et} \quad E(y|x_2, x_3) = \sum_{y} y f(y|x_2, x_3)\,,$$
$$f(y|x_2) = \frac{f(y, x_2)}{f(x_2)} \quad \text{et} \quad E(y|x_2) = \sum_{y} y f(y|x_2)$$

[63] Dans le cas continu, $f(y, x_2) = \int_{-\infty}^{\infty} f(y, x_2, x_3) dx_3$, $f(x_2, x_3) = \int_{-\infty}^{\infty} f(y, x_2, x_3) dy$ et $f(x_2) = \int_{-\infty}^{\infty} f(x_2, x_3) dx_3$, $f(x_3|x_2)$ ayant la même définition.

[64] Dans le cas continu, $E(y|x_2, x_3) = \int_{-\infty}^{\infty} y f(y|x_2, x_3) dy$, $f(y|x_2, x_3)$ ayant la même définition, et $E(y|x_2) = \int_{-\infty}^{\infty} y f(y|x_2) dy$, $f(y|x_2)$ ayant également la même définition.

On obtient :

$f(y\|x_2,x_3)$		1000	1500	2000	2500	$E(y\|x_2,x_3)$
$x_2=12$	$x_3=5$	0,5	0,3	0,2	0	1350
	$x_3=10$	0,2	0,5	0,2	0,1	1600
$x_2=16$	$x_3=5$	0,2	0,5	0,3	0	1550
	$x_3=10$	0	0,2	0,5	0,3	2050

et

$f(y\|x_2)$	1000	1500	2000	2500	$E(y\|x_2)$
$x_2=12$	0,26	0,46	0,2	0,08	1550
$x_2=16$	0,16	0,44	0,34	0,06	1650

On peut par ailleurs vérifier qu'on a bien :

$$E(y|x_2) = \sum_{x_3} E(y|x_2,x_3) f(x_3|x_2)$$

L'espérance conditionnelle $E(y|x_2,x_3) = g(x_2,x_3)$ définit un modèle probabiliste de la relation théorique $y = f(x_2,x_3)$ d'intérêt, dont les variables aléatoires (y,x_2,x_3) ont des probabilités de réalisation décrites par la distribution jointe inconnue $f(y,x_2,x_3)$. On obtient un *modèle probabiliste paramétré* de la relation théorique d'intérêt si on suppose une forme fonctionnelle, dépendant de paramètres, pour $g(x_2,x_3)$. De façon semblable au modèle de régression linéaire simple, le modèle de régression linéaire multiple standard suppose :

$$E(y|x_2,x_3) = \beta_1 + \beta_2 x_2 + \beta_2 x_3 \qquad \text{(i.e., une fonction linéaire de } x_2 \text{ et } x_3\text{)}$$

Pour cette forme fonctionnelle, on a :

$$\begin{aligned}
\frac{\partial E(y|x_2,x_3)}{\partial x_2} &= \frac{\partial g(x_2,x_3)}{\partial x_2} = \beta_2 && \text{(i.e., une constante)} \\
\frac{\partial E(y|x_2,x_3)}{\partial x_3} &= \frac{\partial g(x_2,x_3)}{\partial x_3} = \beta_3 && \text{(i.e., une constante)} \\
\frac{\partial^2 E(y|x_2,x_3)}{\partial x_2^2} &= \frac{\partial^2 E(y|x_2,x_3)}{\partial x_3^2} = 0 && \\
\frac{\partial^2 E(y|x_2,x_3)}{\partial x_2 \partial x_3} &= \frac{\partial^2 E(y|x_2,x_3)}{\partial x_3 \partial x_2} = 0 && \text{(i.e., pas d'interaction)}
\end{aligned}$$

Autrement dit, les effets marginaux de x_2 et de x_3 sont constants, et en particulier ne présentent pas d'interactions (l'effet marginal de x_2 ne dépend pas de x_3, et vice-versa).

Si le modèle de régression linéaire multiple est correct, chaque observation (y_i, x_{i2}, x_{i3}) satisfait le modèle probabiliste :

$$E(y_i|x_{i2},x_{i3}) = \beta_1 + \beta_2 x_{i2} + \beta_2 x_{i3}\,, \qquad i = 1,...,n,$$

où les β_j sont des paramètres inconnus à estimer et, avant observation, (y_i, x_{i2}, x_{i3})

sont des variables aléatoires.

Comme dans le cas du modèle de régression simple, on s'appuie sur un ensemble d'hypothèses statistiques complémentaires qui, pour l'essentiel, peuvent être relâchées si nécessaire.

Ces hypothèses sont les suivantes :

1- $Var(y_i|x_{i2}, x_{i3}) = \sigma^2$ (i.e., homoscédasticité)

2- Les x_{ij} sont fixes, non-stochastiques (+ une hypothèse d'indépendance linéaire, cf. infra). Cette hypothèse équivaut à raisonner conditionnellement aux valeurs des x_{ij} observées dans l'échantillon[65] et, comme pour le modèle de régression simple, permet de recourir à l'écriture simplifiée :

$$\begin{array}{l} E(y_i) = \beta_1 + \beta_2 x_{i2} + \beta_2 x_{i3} \\ Var(y_i) = \sigma^2 \end{array} \qquad i = 1, ..., n$$

3- $Cov(y_i, y_j) = 0\,, \quad \forall i \neq j$ (i.e., non-corrélation)

4- De façon optionnelle, $y_i \sim N(\beta_1 + \beta_2 x_{i2} + \beta_2 x_{i3}, \sigma^2)$ (i.e., normalité)

L'introduction d'un terme d'erreur $e_i = y_i - \beta_1 - \beta_2 x_{i2} - \beta_2 x_{i3}$ permet de reformuler le modèle et ses hypothèses de la même façon que dans le cas du modèle de régression simple :

A1 $y_i = \beta_1 + \beta_2 x_{i2} + \beta_2 x_{i3} + e_i\,, \qquad i = 1, ..., n$

A2 $E(e_i) = 0 \;\Leftrightarrow\; E(y_i) = \beta_1 + \beta_2 x_{i2} + \beta_2 x_{i3}\,, \qquad i = 1, ..., n$

A3 $Var(e_i) = \sigma^2 = Var(y_i)\,, \qquad i = 1, ..., n$

A4 $Cov(e_i, e_j) = 0 = Cov(y_i, y_j)\,, \qquad \forall i \neq j$

A5 les x_{ij} sont non-stochastiques (+ une hypothèse d'indépendance linéaire, cf. infra)

A6 (optionnel) $e_i \sim N(0, \sigma^2) \;\Leftrightarrow\; y_i \sim N(\beta_1 + \beta_2 x_{i2} + \beta_2 x_{i3}, \sigma^2)\,, \qquad i = 1, ..., n$

Les remarques faites à la fin de la Section 2.1.4, en particulier concernant la (non-)vie propre de l'erreur e_i, sa non-observabilité et son origine (variables non prises en compte et variabilité naturelle), sont ici toujours d'application.

6.1.3. Formulation générale du modèle et de ses hypothèses sous forme matricielle

De façon générale, le modèle de régression linéaire multiple permet de prendre en compte un nombre k quelconque de variables explicatives.

---

[65] Au sens strict, elle correspond au cas d'un *échantillonnage stratifié*.

On note :

$$X_i = \begin{bmatrix} 1 & x_{i2} & \cdots & x_{ik} \end{bmatrix} \quad \text{et} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Par définition, on a :

$$\begin{aligned} & y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i \\ \Leftrightarrow \quad & y_i = X_i \beta + e_i \,, \qquad\qquad i = 1, ..., n \end{aligned}$$

En empilant les n observations de l'échantillon, on peut écrire :

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \,, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \text{et} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

de sorte que :

$$\begin{cases} y_1 = \beta_1 + \beta_2 x_{12} + ... + \beta_k x_{1k} + e_1 \\ y_2 = \beta_1 + \beta_2 x_{22} + ... + \beta_k x_{2k} + e_1 \\ \qquad\qquad \vdots \\ y_n = \beta_1 + \beta_2 x_{n2} + ... + \beta_k x_{nk} + e_1 \end{cases}$$

$$\Leftrightarrow \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

soit, de façon compacte :

$$Y = X\beta + e$$

Sur base de cette notation matricielle, les hypothèses A1 - A6 du modèle de régression linéaire multiple s'écrivent :

A1	$Y = X\beta + e$
A2	$E(e) = 0 \;\Leftrightarrow\; E(Y) = X\beta$
A3 - A4	$V(e) = \sigma^2 I = V(Y)$
A5	X est non-stochastique et $\text{rg}(X) = k$
A6	(optionnel) $e \sim N(0, \sigma^2 I) \;\Leftrightarrow\; Y \sim N(X\beta, \sigma^2 I)$

On voit que, hormis le changement de dimension de X et de β qui reflète le fait que l'on permet maintenant d'avoir un nombre k quelconque de variables explicatives (intercept compris), rien n'a changé par rapport à la formulation sous forme matricielle du modèle de régression simple et de ses hypothèses. Celui-ci apparaît maintenant simplement comme le cas particulier où $k = 2$.

Notons encore, comme dans le cas du modèle de régression simple, que l'hypothèse $\text{rg}(X) = k$ requiert que les k colonnes de X soient linéairement indépendantes, ce qui est le cas s'il n'existe pas de constantes non nulles $c_1, c_2, ..., c_k$ telles que :

$$c_1 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} + ... + c_k \begin{bmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{bmatrix} = 0 \,,$$

autrement dit si aucune colonne de X n'est une combinaison linéaire exacte des autres colonnes de X.

6.2. Estimation MCO des paramètres du modèle

L'estimateur MCO est défini par :

$$\begin{aligned} \hat{\beta} &= \text{Argmin}_\beta (Y - X\beta)'(Y - X\beta) \\ &= \text{Argmin}_\beta e'e \end{aligned}$$

Il est obtenu en recherchant le minimum de la fonction :

$$\begin{aligned} SCR(\beta) &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - 2\beta' X'Y + \beta' X'X\beta \end{aligned}$$

En suivant l'approche de la Section 2.3.3, on vérifie aisément (faites-le !) que la *condition de premier ordre* définissant $\hat{\beta}$ s'écrit :

$$X'(Y - X\hat{\beta}) = X'\hat{e} = 0 \quad \Leftrightarrow \quad X'X\hat{\beta} = X'Y \,,$$

soit, sous forme détaillée :

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i3} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}^2 & \sum_{i=1}^n x_{i2}x_{i3} & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \sum_{i=1}^n x_{i3} & \sum_{i=1}^n x_{i3}x_{i2} & \sum_{i=1}^n x_{i3}^2 & \cdots & \sum_{i=1}^n x_{i3}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i2} & \sum_{i=1}^n x_{ik}x_{i3} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \sum_{i=1}^n x_{i3}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{bmatrix} ,$$

et que l'estimateur MCO s'écrit :

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

On voit qu'à nouveau, hormis le changement de dimension de X et de β, rien

n'a changé par rapport à la forme matricielle de l'estimateur MCO du modèle de régression simple.

Plusieurs remarques méritent encore d'être faites :

1- Sous forme détaillée, la première ligne de la condition $X'\hat{e} = 0$ définissant $\hat{\beta}$ donne : $\sum_{i=1}^{n} \hat{e}_i = 0$. La somme des résidus de la régression est donc nulle. Notons que ce n'est pas le cas si le modèle n'inclut pas une constante.

2- Sous forme détaillée, la première ligne de la condition $X'X\hat{\beta} = X'Y$ définissant $\hat{\beta}$ donne, en réarrangant :

$$\bar{y} = \beta_1 + \beta_2 \bar{x}_2 + ... + \beta_k \bar{x}_k$$

L'hyperplan de régression passe donc par le point moyen de l'échantillon. Notons comme ci-dessus que ce n'est pas le cas si le modèle n'inclut pas une constante.

3- Comme dans le cas du modèle de régression simple, l'hypothèse A5 que $\text{rg}(X) = k$ assure que $\text{rg}(X'X) = k$, soit que $X'X$ est non-singulière, et donc inversible (cf. Section 2.3.3).

4- L'estimateur MV de β sous l'hypothèse A6 de normalité est toujours, dans le cadre du modèle de régression multiple, identique à l'estimateur MCO. De même, l'estimateur MV de σ^2 est toujours donné par :

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i^2 = \frac{\hat{e}'\hat{e}}{n}, \quad \hat{e} = Y - X\hat{\beta}$$

5- On peut aisément vérifier (faites-le !) que l'ensemble des résultats complémentaires décrits à la Section 2.3.4 reste valable dans le cadre du modèle de régression multiple, et en particulier que la matrice de projection $M_X = I - X(X'X)^{-1}X'$ a la même interprétation et possède les mêmes propriétés que dans le cas du modèle de régression simple : M_X est symétrique et idempotente, et $\hat{e} = M_X e$.

6.3. Propriétés de l'estimateur MCO

Le passage du modèle de régression de 2 à k variables explicatives ne modifie en rien les propriétés statistiques de l'estimateur MCO. De même, l'interprétation de ces propriétés reste inchangée.

6.3.1. Propriétés d'échantillonnage

En suivant l'approche de la Section 3.1, on vérifie aisément (faites-le !) que, sous

les hypothèses A1, A2 et A5, on a toujours :

$$E(\hat{\beta}) = \begin{bmatrix} E(\hat{\beta}_1) \\ \vdots \\ E(\hat{\beta}_k) \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = \beta \,,$$

autrement dit que $\hat{\beta}$ est un estimateur non biaisé de β, et que, sous les hypothèses A1 à A5, on a encore :

$$V(\hat{\beta}) = \begin{bmatrix} Var(\hat{\beta}_1) & \cdots & Cov(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_k, \hat{\beta}_1) & \cdots & Var(\hat{\beta}_k) \end{bmatrix} = \sigma^2 (X'X)^{-1}$$

On peut montrer que, sous forme détaillée, pour $k = 3$, cela donne :

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{(1 - \rho_e(x_{i2}, x_{i3})^2) \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2} \tag{6.1}$$

$$Var(\hat{\beta}_3) = \frac{\sigma^2}{(1 - \rho_e(x_{i2}, x_{i3})^2) \sum_{i=1}^{n} (x_{i3} - \bar{x}_3)^2} \tag{6.2}$$

$$Cov(\hat{\beta}_2, \hat{\beta}_3) = \frac{-\rho_e(x_{i2}, x_{i3})\sigma^2}{(1 - \rho_e(x_{i2}, x_{i3})^2) \sqrt{\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2} \sqrt{\sum_{i=1}^{n} (x_{i3} - \bar{x}_3)^2}} \tag{6.3}$$

où $\rho_e(x_{i2}, x_{i3})$ désigne la corrélation empirique entre x_{i2} et x_{i3} :

$$\rho_e(x_{i2}, x_{i3}) = \frac{Cov_e(x_{i2}, x_{i3})}{\sqrt{Var_e(x_{i2})} \sqrt{Var_e(x_{i3})}}$$

Sur base des expressions détaillées (6.1), (6.2) et (6.3), on peut voir que les facteurs déterminant $V(\hat{\beta})$ sont[66] :

1- La variance σ^2 du terme d'erreur :

$$\text{si } \sigma^2 \nearrow \text{ , alors } Var(\hat{\beta}_2),\ Var(\hat{\beta}_3) \text{ et } |Cov(\hat{\beta}_2, \hat{\beta}_3)| \nearrow$$

Autrement dit, plus la dispersion des y_i autour du plan de régression $E(y_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$ est grande, moins la précision d'estimation est grande.

2- La dispersion des variables explicatives :

$$\text{si } \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 \text{ et } \sum_{i=1}^{n} (x_{i3} - \bar{x}_3)^2 \nearrow \text{ , alors } \begin{array}{l} Var(\hat{\beta}_2),\ Var(\hat{\beta}_3) \\ \text{et } |Cov(\hat{\beta}_2, \hat{\beta}_3)| \searrow \end{array}$$

Autrement dit, plus la dispersion des x_{ij} est grande, plus la précision d'estimation

[66] L'intercept $\hat{\beta}_1$ est ici non considéré.

est grande.

3- La taille n de l'échantillon :

$$\text{si } n \nearrow \text{ , } \sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2 \text{ et } \sum_{i=1}^{n}(x_{i3}-\bar{x}_3)^2 \nearrow \text{ , alors } \begin{array}{l} Var(\hat{\beta}_2),\, Var(\hat{\beta}_3) \\ \text{et } |Cov(\hat{\beta}_2,\hat{\beta}_3)| \searrow \end{array}$$

Autrement dit, plus la taille d'échantillon est grande, plus la précision d'estimation est grande.

4- La corrélation entre les variables explicatives :

$$\text{si } |\rho_e(x_{i2},x_{i3})| \nearrow \text{ , alors } Var(\hat{\beta}_2),\, Var(\hat{\beta}_3) \text{ et } |Cov(\hat{\beta}_2,\hat{\beta}_3)| \nearrow$$

Autrement dit, plus la corrélation entre les variables explicatives x_{ij} est grande, moins la précision d'estimation est grande.

Nous avions déjà identifié les trois premiers facteurs dans le cadre du modèle de régression simple. La seule nouveauté ici est le rôle de la corrélation entre les variables explicatives (hors intercept). Intuitivement, si les variables explicatives sont fortement corrélées, cela signifie que, dans l'échantillon, leurs valeurs bougent toujours de concert, de sorte qu'il est difficile d'estimer précisément l'effet marginal propre (i.e., le paramètre de β_j) de chacune d'elles. Pour pouvoir estimer précisément l'effet propre des différentes variables, il est nécessaire qu'elles varient, au moins partiellement, indépendamment les unes des autres.

En suivant l'approche de la Section 3.2, on peut encore aisément vérifier (faites-le !) que le théorème Gauss - Markov est toujours d'application, de sorte que $\hat{\beta}$ est le meilleur estimateur sans biais de β. En d'autres termes, sous les hypothèses A1 à A5 :

$$V(\hat{\beta}^*) \geq V(\hat{\beta}),$$

pour tout autre estimateur linéaire sans biais $\hat{\beta}^*$ de β que l'estimateur MCO $\hat{\beta}$. Cela implique que pour tout vecteur a de dimension $k \times 1$, on a :

$$Var(a'\hat{\beta}^*) \geq Var(a'\hat{\beta}),$$

autrement dit que la variance de toute combinaison linéaire de $\hat{\beta}^*$ est toujours supérieure ou égale à la variance de la même combinaison linéaire de $\hat{\beta}$, et donc en particulier que :

$$Var(\hat{\beta}_j^*) \geq Var(\hat{\beta}_j), \quad j = 1, ..., k$$

Finalement, en suivant l'approche des Sections 3.3 et 3.4, on peut à nouveau aisément vérifier (faites-le !) que si, aux hypothèses A1 à A5, on ajoute l'hypothèse A6 de normalité, on a de façon exacte en *échantillon fini* :

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}),$$

tandis que, sous les seules hypothèses A1 à A5 (sans invoquer A6 donc), on a

asymptotiquement :

$$\hat{\beta} \xrightarrow{p} \beta$$

et

$$\left[V(\hat{\beta})\right]^{-\frac{1}{2}} \left(\hat{\beta} - \beta\right) \xrightarrow{d} N(0, I)\,,$$

soit, exprimé sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand (au moins $n > 30$) :

$$\hat{\beta} \approx N(\beta, \sigma^2 (X'X)^{-1})$$

6.3.2. Estimateur de σ^2 et de $V(\hat{\beta})$

On peut estimer la matrice de variance-covariance $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ de l'estimateur MCO $\hat{\beta}$ en remplaçant la variance σ^2 du terme d'erreur du modèle par un estimateur de cette quantité.

Un estimateur naturel de σ^2 est donné par l'estimateur MV :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{\hat{e}'\hat{e}}{n}\,, \quad \hat{e} = Y - X\hat{\beta}$$

En suivant l'approche de la Section 3.5.1, on peut aisément vérifier (faites-le !) que, bien que convergent, cet estimateur est biaisé puisque, sous les hypothèses A1 à A5, on a :

$$\begin{aligned} E(\hat{\sigma}^2) &= E\left(\frac{e'M_X e}{n}\right) \qquad\qquad (\text{car } \hat{e} = M_X e,\ \text{cf. Section 2.3.4}) \\ &= \frac{\sigma^2}{n}\mathrm{tr}\,[M_X] \\ &= \frac{n-k}{n}\sigma^2 < \sigma^2 \end{aligned} \tag{6.4}$$

De (6.4), on peut déduire que, sous les hypothèses A1 à A5, un estimateur *convergent* et *non biaisé* de σ^2 est donné par :

$$\hat{s}^2 = \frac{1}{n-k}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{\hat{e}'\hat{e}}{n-k}$$

Dans le cas du modèle de régression simple, on avait simplement $k = 2$.

Sur base de l'estimateur $\hat{s}^2$, sous les hypothèses A1 à A5, un estimateur *convergent* et *non biaisé* de $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ est donné par :

$$\hat{V}(\hat{\beta}) = \hat{s}^2(X'X)^{-1}$$

Des éléments diagonaux $\hat{Var}(\hat{\beta}_j)$ $(j = 1, ..., k)$ de cet estimateur $\hat{V}(\hat{\beta})$, on obtient des estimateurs *convergents*, mais *pas non biaisés* des écarts-types $s.e.(\hat{\beta}_j)$ des différents $\hat{\beta}_j$ en prenant :

$$s.\hat{e}.(\hat{\beta}_j) = \sqrt{\hat{Var}(\hat{\beta}_j)}\,, \quad j = 1, ..., k$$

6.4. Intervalles de confiance et tests d'hypothèse de β_j

On sait que, sous les hypothèses A1 à A6 (y.c. donc l'hypothèse de normalité), on a de façon exacte :

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})\,,$$

ce qui implique, pour $j = 1, ..., k$, que :

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))\,,$$

où $Var(\hat{\beta}_j) = \sigma^2 q_{jj}$, avec $q_{jj} = [(X'X)^{-1}]_{jj}$, de sorte que :

$$\hat{z} = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim N(0, 1)\,,$$

où $s.e.(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)} = \sqrt{\sigma^2 q_{jj}}$. En particulier, lorsque $\beta_j = \beta_j^o$, on a :

$$\hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \sim N(0, 1)\,,$$

et lorsque $\beta_j = \beta_j^* \neq \beta_j^o$, on a :

$$\hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \sim N\left(\frac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}, 1\right)$$

En suivant l'approche de la Section 4.1.2, on peut aisément vérifier (faites-le !) que, sous les hypothèses A1 à A6, on a :

$$\frac{\hat{e}'\hat{e}}{\sigma^2} = \frac{e' M_X e}{\sigma^2} \sim \chi^2(n-k)$$

$$\Leftrightarrow \quad \hat{v} = \frac{(n-k)\hat{s}^2}{\sigma^2} \sim \chi^2(n-k) \qquad (\text{car } \hat{e}'\hat{e} = (n-k)\hat{s}^2)$$

Dans le cas du modèle de régression simple, on avait simplement $k = 2$.

On peut encore montrer que $\hat{z}$ et $\hat{v}$ sont indépendamment distribués, de sorte

que de la définition de la loi de Student, on a :

$$\hat{t} = \frac{\hat{z}}{\sqrt{\frac{\hat{v}}{n-k}}} = \frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 q_{jj}}}}{\sqrt{\frac{\hat{s}^2}{\sigma^2}}} \sim t(n-k)\,,$$

soit, en simplifiant :

$$\hat{t} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{s}^2 q_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)} \sim t(n-k)$$

En particulier, lorsque $\beta_j = \beta_j^o$, on a :

$$\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \sim t(n-k)\,,$$

tandis que lorsque $\beta_j = \beta_j^* \neq \beta_j^o$, on peut montrer qu'on a :

$$\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \sim t(\delta^*, n-k)\,,$$

où $t(\delta^*, n-k)$ désigne la loi de Student non-centrale à $(n-k)$ degrés de liberté et le paramètre de non-centralité δ^* est égal à $\delta^* = \frac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}$.

On voit que, mis à part la modification du nombre de degrés de liberté sur les lois de Student impliquées ($(n-k)$ au lieu de $(n-2)$), les résultats ci-dessus sont en tout point identiques aux résultats sur lesquels nous nous sommes appuyés au Chapitre 4 pour construire des intervalles de confiance pour β_j $(j = 1, 2)$ et des tests d'hypothèses (bilatéraux et unilatéraux) pour β_j $(j = 1, 2)$ dans le modèle de régression linéaire simple.

On en conclut qu'au nombre de degrés de libertés près ($(n-k)$ au lieu de $(n-2)$), les procédures décrites au Chapitre 4 pour les intervalles de confiance pour β_j (cf. Section 4.1.2) et pour les tests d'hypothèses de β_j (cf. Section 4.2.2) restent d'application (le seul changement notable est le nombre de degrés de liberté de la loi de Student impliqué : $(n-k)$ au lieu de $(n-2)$) dans la cadre du modèle de régression linéaire multiple. De même, l'interprétation de ces procédures reste identique.

Si on renonce à l'hypothèse A6 de normalité , on sait que, sous les seules hypothèses A1 à A5, on a toujours asymptotiquement :

$$\left[V(\hat{\beta})\right]^{-\frac{1}{2}} \left(\hat{\beta} - \beta\right) \xrightarrow{d} N(0, I)\,, \quad \text{où } V(\hat{\beta}) = \sigma^2 (X'X)^{-1}\,,$$

ce qui implique, pour $j = 1, ..., k$, que :

$$\hat{z} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 q_{jj}}} \xrightarrow{d} N(0, 1)\,, \quad \text{où } q_{jj} = \left[(X'X)^{-1}\right]_{jj}\,,$$

soit, sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand :

$$\hat{z} = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \approx N(0,1)\,, \quad \text{où } s.e.(\hat{\beta}_j) = \sqrt{\sigma^2 q_{jj}}\,,$$

et donc :

$$\begin{cases} \hat{z}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \approx N(0,1)\,, & \text{si } \beta_j = \beta_j^o \\ \hat{z}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \approx N\left(\dfrac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}, 1\right), & \text{si } \beta_j = \beta_j^* (\neq \beta_j^o) \end{cases}$$

Asymptotiquement, lorsque n est grand, on peut encore montrer que le remplacement de σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$ ne modifie pas les distributions d'échantillonnage en jeu, de sorte qu'on a aussi, sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand :

$$\hat{t} = \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)} \approx N(0,1)\,, \quad \text{où } s.\hat{e}.(\hat{\beta}_j) = \sqrt{\hat{s}^2 q_{jj}}\,,$$

et donc :

$$\begin{cases} \hat{t}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \approx N(0,1)\,, & \text{si } \beta_j = \beta_j^o \\ \hat{t}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \approx N\left(\dfrac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}, 1\right), & \text{si } \beta_j = \beta_j^* (\neq \beta_j^o) \end{cases}$$

On sait par ailleurs que lorsque $n \to \infty$, la loi de Student $t(n-k)$ tend vers la loi normale $N(0,1)$, de sorte que les quantiles de la loi de Student $t(n-k)$ et de la loi normale $N(0,1)$ s'égalisent.

Comme à la Section 4.3 du Chapitre 4, de ces éléments, on peut conclure que les procédures d'intervalles de confiance pour β_j $(j = 1, ..., k)$ et de tests d'hypothèse de β_j $(j = 1, ..., k)$ suggérées ci-dessus, qui sont *exactes en échantillon fini* sous l'hypothèse de normalité A6, restent valables *asymptotiquement*, à titre approximatif, pour n grand, sous les seules hypothèses A1 à A5. En bref, toujours rien de nouveau, sinon le changement du nombre de degrés de liberté de la loi de Student impliqué : $(n-k)$ au lieu de $(n-2)$.

On notera finalement que les remarques de terminologie et d'interprétation développées à la Section 4.2.3 dans le cadre du modèle de régression simple restent d'application dans le cadre du modèle de régression linéaire multiple. Ainsi, en particulier :

1- Le test bilatéral de $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$ pour $j = 2, ..., k$ (pour β_1, ce test a généralement peu de sens) revient, dans le cadre du modèle de régression

linéaire multiple, à tester, pour par exemple $j = 2$ et $k = 3$:

$$\mathrm{H}_0' : \; E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_3 x_{i3} \,, \qquad \text{i.e., } E(y_i|x_{i2}, x_{i3}) \text{ ne dépend pas de } x_{i2} \text{ et est linéaire en } x_{i3}$$

$$\text{contre } \mathrm{H}_1' : \; E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_2 x_{i2} + \beta_2 x_{i3i} \,, \qquad \text{i.e., } E(y_i|x_{i2}, x_{i3}) \text{ est une fonction linéaire de } x_{i2} \text{ et } x_{i3}$$

2- Le fait de trouver les $\hat{\beta}_j$ significatifs (pour $j = 2, ..., k$) ne garantit en rien que $E(y_i|x_{i2}, ..., x_{ik})$ est bien une fonction linéaire des x_{ij}.

3- A contrario, le fait de ne pas trouver un $\hat{\beta}_j$ significatif ne signifie pas nécessairement que $E(y_i|x_{i2}, ..., x_{ik})$ ne dépend pas de x_{ij}. C'est seulement une absence de preuve que $E(y_i|x_{i2}, ..., x_{ik})$ dépend de x_{ij}.

4- Il ne faut pas confondre '$\hat{\beta}_j$ est (très) significatif' et 'x_{ij} a un effet (très) important sur $E(y_i|x_{i2}, ..., x_{ik})$'. Pas question donc d'évaluer l'importance relative des effets des variables x_{ij} sur $E(y_i|x_{i2}, ..., x_{ik})$ sur base de la plus ou moins grande significativité de leur paramètre estimé $\hat{\beta}_j$. On notera au passage que comparer l'importance relative des effets des variables x_{ij} sur $E(y_i|x_{i2}, ..., x_{ik})$ est une opération délicate lorsque les variables en jeu ne sont pas directement comparables (par exemple, le revenu et le nombre de membres d'un ménage).

5- Pour conclure, on rappellera que des modèles impliquant des variables explicatives (ensembles conditionnants) différentes étant des modèles différents (cf. les commentaires de la Section 6.1.2), le fait que le paramètre estimé d'une variable donnée apparaisse significatif (par abus de langage, on dit souvent : que la variable apparaisse significative) dans un modèle et pas dans un autre n'a a priori rien de contradictoire.

6.5. Prévision et intervalles de prévision

Comme dans le cadre du modèle de régression linéaire simple, le prédicteur :

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_{02} + ... + \hat{\beta}_k x_{0k} = X_0 \hat{\beta} \,, \quad \text{où } X_0 = \begin{bmatrix} 1 & x_{02} & \cdots & x_{0k} \end{bmatrix} \,,$$

peut à la fois être utilisé comme estimateur / prédicteur de :

$$E(y_0) = \beta_1 + \beta_2 x_{02} + ... + \beta_k x_{0k} \,,$$

càd. de l'espérance de y sachant $(x_{02}, ...x_{0k})$, et comme prédicteur de :

$$y_0 = \beta_1 + \beta_2 x_{02} + ... + \beta_k x_{0k} + e_0 \,,$$

càd. de la valeur de y sachant $(x_{02}, ...x_{0k})$.

En suivant l'approche de la Section 5.1, on peut aisément vérifier (faites-le !) que les erreurs de prévision $\hat{p}_0 = \hat{y}_0 - E(y_0)$ et $\hat{f}_0 = \hat{y}_0 - y_0$ sont, dans le modèle de régression multiple, toujours telles que, sous les hypothèses A1 à A6 (y.c. donc

l'hypothèse de normalité), on a :

$$\begin{aligned}
E(\hat{p}_0) &= 0\,, \quad \text{i.e. } \hat{y}_0 \text{ est un estimateur / prédicteur non biaisé de } E(y_0) \\
E(\hat{f}_0) &= 0\,, \quad \text{i.e. } \hat{y}_0 \text{ est un prédicteur non biaisé de } y_0 \\
Var(\hat{p}_0) &= X_0 V(\hat{\beta}) X_0' = \sigma^2 X_0 (X'X)^{-1} X_0' \\
Var(\hat{f}_0) &= \sigma^2 + X_0 V(\hat{\beta}) X_0' = \sigma^2 (1 + X_0 (X'X)^{-1} X_0') \\
\hat{p}_0 &\sim N(0, Var(\hat{p}_0)) \quad \Leftrightarrow \quad \hat{z}_{p_0} = \frac{\hat{p}_0}{\sqrt{\sigma^2 X_0 (X'X)^{-1} X_0'}} \sim N(0,1) \\
\hat{f}_0 &\sim N(0, Var(\hat{f}_0)) \quad \Leftrightarrow \quad \hat{z}_{f_0} = \frac{\hat{f}_0}{\sqrt{\sigma^2 (1 + X_0 (X'X)^{-1} X_0')}} \sim N(0,1)
\end{aligned}$$

Toujours en suivant l'approche de la Section 5.1, et en notant qu'on peut montrer que $\hat{z}_{p_0}$ et $\hat{z}_{f_0}$ sont indépendamment distribués de $\hat{v} = \frac{(n-k)\hat{s}^2}{\sigma^2} \sim \chi^2(n-k)$ (sur le résultat $\hat{v} \sim \chi^2(n-k)$, voir la Section 6.4), on vérifie aisément (faites-le !) que, sous les hypothèses A1 à A6, on a encore :

$$\hat{t}_{p_0} = \frac{\hat{p}_0}{\sqrt{\hat{s}^2 X_0 (X'X)^{-1} X_0'}} = \frac{\hat{y}_0 - E(y_0)}{s.\hat{e}.(\hat{p}_0)} \sim t(n-k)$$

et

$$\hat{t}_{f_0} = \frac{\hat{f}_0}{\sqrt{\hat{s}^2 (1 + X_0 (X'X)^{-1} X_0')}} = \frac{\hat{y}_0 - y_0}{s.\hat{e}.(\hat{f}_0)} \sim t(n-k)$$

On voit à nouveau que, mis à part la modification du nombre de degrés de liberté de la loi de Student impliquée ($(n-k)$ au lieu de $(n-2)$), et évidemment le changement de dimension de X et de X_0, les résultats ci-dessus sont en tout point identiques aux résultats sur lesquels nous nous sommes appuyés à la Section 5.1 pour construire des intervalles de prévision pour $E(y_0)$ et y_0 dans le cadre du modèle de régression linéaire simple.

On en conclut qu'au nombre de degrés de libertés ($(n-k)$ au lieu de $(n-2)$) et à la dimension de X et de X_0 près, les intervalles de prévision pour $E(y_0)$ et y_0 décrits respectivement à la Section 5.1.1.2 et à la Section 5.1.2.2 restent d'application (le changement le plus notable est le nombre de degrés de liberté de la loi de Student impliqué : $(n-k)$ au lieu de $(n-2)$) dans le cadre du modèle de régression linéaire multiple. De même, l'interprétation de ces intervalles de prévision reste inchangée.

Si on renonce à l'hypothèse A6 de normalité, comme $\hat{p}_0 = \hat{y}_0 - E(y_0) = X_0(\hat{\beta} - \beta)$ est une combinaison linéaire de $\hat{\beta}$ et que $\hat{\beta}$ est toujours *asymptotiquement* distribué de façon normale, on a encore que $\hat{p}_0$ est *asymptotiquement* distribué de façon normale. Ainsi, formellement, sous les hypothèses A1 à A5, on a :

$$\frac{\hat{p}_0}{\sqrt{\sigma^2 X_0 (X'X)^{-1} X_0'}} = \frac{\hat{y}_0 - E(y_0)}{s.e.(\hat{p}_0)} \xrightarrow{d} N(0,1)\,,$$

soit, sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand :

$$\frac{\hat{y}_0 - E(y_0)}{s.e.(\hat{p}_0)} \approx N(0,1)\,,$$

où $s.e.(\hat{p}_0) = \sqrt{\sigma^2 X_0(X'X)^{-1}X_0'} = \sqrt{X_0 V(\hat{\beta})X_0'}$, et, comme lorsque n est grand, on peut montrer que le remplacement de σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$ ne modifie pas la distribution d'échantillonnage en jeu, on a aussi, sous forme d'approximation :

$$\frac{\hat{y}_0 - E(y_0)}{s.\hat{e}.(\hat{p}_0)} \approx N(0,1)\,,$$

où $s.\hat{e}.(\hat{p}_0) = \sqrt{\hat{s}^2 X_0(X'X)^{-1}X_0'} = \sqrt{X_0 \hat{V}(\hat{\beta})X_0'}$.

Etant donné la convergence de la loi de Student $t(n-k)$ vers la loi normale $N(0,1)$ lorsque $n \to \infty$, comme à la Section 5.1.1.2, on peut conclure que l'intervalle de prévision pour $E(y_0)$ suggéré ci-dessus, qui est *exact en échantillon fini* sous l'hypothèse A6 de normalité, reste encore valable *asymptotiquement*, à titre approximatif, pour n grand, sous les seules hypothèses A1 à A5.

On notera pour terminer que pour les mêmes raisons que celles invoquées à la Section 5.1.2.2, un même résultat asymptotique *ne tient pas* dans le cas de l'intervalle de prévision pour y_0.

6.6. Exemple : les ventes d'une chaîne de fast-food de HGL (2008)

Hill, Griffiths et Lim (2008) s'intéressent[67] à l'effet de la politique de prix et de publicité sur les ventes d'une chaîne de fast-food. Le modèle considéré est :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

où y_i désigne les recettes mensuelles de vente (en milliers de \$), x_{i2} le prix de vente unitaire (en \$), et x_{i3} le montant des dépenses publicitaires mensuelles (en milliers de \$).

L'échantillon est composé d'observations (en coupe) de (y_i, x_{i2}, x_{i3}) dans 75 succursales de la chaîne de fast-food situées dans des petites villes américaines de taille comparable.

Les questions d'intérêt sont :

1- La demande est-elle rigide ($\beta_2 > 0$) ou au contraire élastique ($\beta_2 < 0$) ?

2- Les dépenses publicitaires sont-elles efficaces ($\beta_3 > 0$) ? rentables ($\beta_3 > 1$) ?

[67] Voir p. 106 et suivantes.

En utilisant le logiciel GRETL, on obtient comme statistique descriptive :

Summary statistics, using the observations 1 - 75

	Mean	Median	Minimum	Maximum
y	77.375	76.500	62.400	91.200
x2	5.6872	5.6900	4.8300	6.4900
x3	1.8440	1.8000	0.5000	3.1000

	Std. Dev.	C.V.	Skewness	Ex. kurtosis
y	6.4885	0.083859	-0.010631	-0.74467
x2	0.51843	0.091158	0.061846	-1.3328
x3	0.83168	0.45102	0.037087	-1.2951

et comme tableau de résultats d'estimation :

Model 1:
OLS, using observations 1-75
Dependent variable: y

	coefficient	std. error	t-ratio	p-value
const	118.914	6.35164	18.72	2.21e-029 ***
x2	-7.90785	1.09599	-7.215	4.42e-010 ***
x3	1.86258	0.683195	2.726	0.0080 ***

Mean dependent var	77.37467	S.D. dependent var	6.488537
Sum squared resid	1718.943	S.E. of regression	4.886124
R-squared	0.448258	Adjusted R-squared	0.432932
F(2, 72)	29.24786	P-value(F)	5.04e-10
Log-likelihood	-223.8695	Akaike criterion	453.7390
Schwarz criterion	460.6915	Hannan-Quinn	456.5151

L'interprétation des coefficients estimés est la suivante :

- $\hat{\beta}_1 =$ l'intercept (ordonnée à l'origine) : il représente ici le niveau moyen *théorique* des ventes pour un prix et un montant de dépenses publicitaires nulles. Ce niveau théorique est estimé à 118914 $ (attention aux unités de mesure !). Ce montant n'a pas d'interprétation économique.
- $\hat{\beta}_2 =$ l'effet marginal de x_{i2} (x_{i3} étant maintenu constant) : il représente ici la recette marginale obtenue d'un accroissement unitaire du prix. Dans cet exemple, il est estimé qu'une augmentation du prix de 1 $ diminue les recettes mensuelles de vente moyenne de $7907,85$ $ (attention aux unités de mesure !).
- $\hat{\beta}_3 =$ l'effet marginal de x_{i3} (x_{i2} étant maintenu constant) : il représente ici la recette marginale obtenue d'un accroissement unitaire des dépenses publicitaires. Dans cet exemple, il est estimé qu'une augmentation des dépenses publicitaires de 1000 $ accroît les recettes mensuelles de vente moyenne de $1862,58$ $ (attention aux unités de mesure !).

Sur base du tableau des résultats d'estimation, si on note[68] que pour $(n-k) = 72$ et $\alpha = 0,05$, on a $t_{n-k;1-\frac{\alpha}{2}} = t_{72;0,975} = 1,993$ et $t_{n-k;1-\alpha} = t_{72;0,95} = 1,666$, on peut :

1- calculer un intervalle de confiance à 95% pour β_2 :

$$\begin{aligned} \hat{\beta}_2 \pm t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_2) &= -7,908 \pm 1,993 \times 1,096 \\ &= [-10,092\,;\, -5,724] \end{aligned}$$

2- calculer un intervalle de confiance à 95% pour β_3 :

$$\begin{aligned} \hat{\beta}_3 \pm t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_3) &= 1,863 \pm 1,993 \times 0,683 \\ &= [0,502\,;\, 3,224] \end{aligned}$$

3- voir que la statistique de test $\hat{t}_o$ du t-test de $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$ est égale à -7, 215, et que H_0 peut être rejetée au *seuil minimum* de 4, 42e-010 (= P-valeur du test). On peut donc rejeter l'hypothèse nulle que $E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_3 x_{i3}$, càd. que $E(y_i|x_{i2}, x_{i3})$ ne dépend pas de x_{i2}.

4- voir que la statistique de test $\hat{t}_o$ du t-test de $H_0 : \beta_3 = 0$ contre $H_1 : \beta_3 \neq 0$ est égale à $2,726$, et que H_0 peut être rejetée au *seuil minimum* de $0,008$ (= P-valeur du test). On peut donc rejeter l'hypothèse nulle que $E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_2 x_{i2}$, càd. que $E(y_i|x_{i2}, x_{i3})$ ne dépend pas de x_{i3}.

5- effectuer un test de $H_0 : \beta_2 \geq 0$ contre $H_1 : \beta_2 < 0$. On a $\hat{t}_o$ = -7, 215, et la P-valeur du test est égale à $\frac{4,42\text{e-}010}{2} = 2,21\text{e-}010$, de sorte H_0 peut être rejetée au *seuil minimum* de 2, 21e-010. On peut donc affirmer que $\hat{\beta}_2$ est statistiquement significativement inférieur à 0, autrement dit que le demande est élastique.

6- effectuer un test de $H_0 : \beta_3 \leq 0$ contre $H_1 : \beta_3 > 0$. On a $\hat{t}_o = 2,726$, et la P-valeur du test est égale à $\frac{0,008}{2} = 0,004$, de sorte H_0 peut être rejetée au *seuil minimum* de $0,004$. On peut donc affirmer que $\hat{\beta}_3$ est statistiquement significativement supérieur à 0, autrement dit que les dépenses publicitaires sont efficaces (elles accroissent les ventes).

7- effectuer un test de $H_0 : \beta_3 \leq 1$ contre $H_1 : \beta_3 > 1$. On obtient :

$$\hat{t}_o = \frac{1,863 - 1}{0,683} = 1,264$$

On ne peut pas rejeter H_0 au seuil de 5% car $\hat{t}_o = 1,264 < t_{n-k;1-\alpha} = t_{72;0,95} = 1,666$. La P-valeur du test[69] est en fait égale à $0,105$, de sorte H_0 ne peut être rejetée qu'au *seuil minimum* de $0,105$. Sauf à prendre un risque de première espèce relativement élevé (supérieur à $0,105$), on ne peut donc pas affirmer que $\hat{\beta}_3$ est statistiquement significativement supérieur à 1, autrement dit que les dépenses publicitaires sont rentables (elles accroissent les recettes d'un montant au moins égal à leur coût). Si le paramètre estimé $\hat{\beta}_3$ (= $1,863$) suggèrent bien qu'elles sont rentables, il apparaît que sa précision d'estimation est trop faible pour pouvoir conclure avec confiance que c'est bien le cas.

---

[68] Pour rappel, les quantiles de la loi de Student peuvent être obtenus en utilisant les 'Statistical tables' de GRETL.

[69] Pour rappel, la P-valeur peut être calculée en utilisant le 'p-value finder' de GRETL.

Toujours sur base du tableau des résultats d'estimation, on peut encore calculer $\hat{s}^2$ et le prédicteur des recettes mensuelles de vente $\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_{02} + \hat{\beta}_3 x_{03}$ pour un prix de $5,5\,\$$, soit $x_{02} = 5,5$, et un montant de dépenses publicitaires de $1800\,\$$, soit $x_{03} = 1,8$ (ces valeurs sont proches du point moyen de l'échantillon, cf. le tableau des statistiques descriptives) :

$$\begin{aligned} \hat{s}^2 &= \frac{1718,94}{(75-3)} = 23,874 \\ \hat{y}_i &= 118,91 - 7,91 \times 5,5 + 1,86 \times 1,8 = 78,753 \end{aligned}$$

Toujours en utilisant GRETL, on obtient pour $\hat{V}(\hat{\beta})$:

Covariance matrix of regression coefficients:

const	x2	x3	
40.3433	-6.79506	-0.748421	const
	1.2012	-0.0197422	x2
		0.466756	x3

Sur base de ce résultat complémentaire, on peut calculer[70], toujours pour $x_{02} = 5,5$ et $x_{03} = 1,8$:

1- un intervalle de prévision à 95% pour $E(y_0)$:

$$\begin{aligned} \hat{y}_0 \pm t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{p}_0) &= \hat{y}_0 \pm t_{n-k;1-\frac{\alpha}{2}} \sqrt{X_0 \hat{V}(\hat{\beta}) X_0'} \\ &= 78,753 \pm 1,993 \times 0,601 \\ &= [77,555\,;\,79,951] \end{aligned}$$

2- un intervalle de prévision à 95% pour y_0 :

$$\begin{aligned} \hat{y}_0 \pm t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{f}_0) &= \hat{y}_0 \pm t_{n-k;1-\frac{\alpha}{2}} \sqrt{\hat{s}^2 + X_0 \hat{V}(\hat{\beta}) X_0'} \\ &= 78,753 \pm 1,993 \times 4,923 \\ &= [68,941\,;\,88,565] \end{aligned}$$

On constate à nouveau, et pour les mêmes raisons que celles déjà évoquées, que l'intervalle de prévision pour y_0 est bien plus large que l'intervalle de prévision pour $E(y_0)$: l'intervalle de prévision pour les recettes de vente moyenne (sachant $x_{02} = 5,5$ et $x_{03} = 1,8$) donne l'intervalle 77555 $ - 79951 $, soit un intervalle assez précis, tandis que l'intervalle de prévision pour les recettes de ventes d'une succursale prise au hasard (parmi celles pour lesquelles $x_{02} = 5,5$ et $x_{03} = 1,8$) donne l'intervalle 69941 $ - 88565 $, soit un intervalle bien plus large.

[70] Pour rappel, notons que $s.\hat{e}.(\hat{p}_0)$ et $s.\hat{e}.(\hat{f}_0)$ peuvent aisément être calculés en utilisant les capacités de calcul matriciel de GRETL.

6.7. Le coefficient de détermination multiple : R^2

En suivant l'approche de la Section 5.2, on peut aisément vérifier (faites-le !) que, comme dans le modèle de régression simple, de la décomposition de y_i en une partie expliquée $\hat{y}_i$ et une partie résiduelle $\hat{e}_i$:

$$\begin{aligned} & y_i = \hat{y}_i + \hat{e}_i \\ \Leftrightarrow \quad & Y = \hat{Y} + \hat{e} = X\hat{\beta} + \hat{e}\,, \end{aligned}$$

on peut obtenir la décomposition :

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}_{\text{SCE}} + \underbrace{\sum_{i=1}^{n}\hat{e}_i^2}_{\text{SCR}}\,,$$

où SCT désigne la somme des carrés totaux (centrés), SCE la somme des carrés expliqués (centrés), et SCR la somme des carrés des résidus, qu'on peut encore écrire sous la forme de l'équation d'analyse de la variance :

$$\underbrace{Var_e(y_i)}_{\text{Variance totale}} = \underbrace{Var_e(\hat{y}_i)}_{\text{Variance expliquée}} + \underbrace{Var_e(\hat{e}_i)}_{\text{Variance résiduelle}}$$

où $Var_e(.)$ désigne la variance empirique. On se rappellera que cette décomposition *n'est pas valable* si le modèle n'inclut pas une constante (un intercept).

On peut dès lors pareillement définir un *coefficient de détermination multiple*, toujours noté R^2 :

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}} = \frac{Var_e(\hat{y}_i)}{Var_e(y_i)}$$

qui mesure la *part de la variance* des y_i *expliquée* par la régression, ou plus précisément, la part de la variance des y_i qui peut être *linéairement associée* à la variation des variables $(x_{i2}, ..., x_{ik})$. Par construction, sauf si le modèle n'inclut pas une constante, on a toujours :

$$0 \leq R^2 \leq 1$$

avec : - $R^2 = 1$ si et seulement si SCR $= 0$.
- $R^2 = 0$ si et seulement si SCE $= 0$, soit si et seulement si SCT $=$ SCR.

Plusieurs points méritent encore d'être épinglés :

1- On rappellera que le R^2 est une *mesure descriptive* (rien de plus), et est souvent interprété comme une mesure globale (mais imparfaite, cf. Section 5.2) de la 'capacité prédictive' du modèle.

2- On peut encore montrer que le R^2 est toujours égal au carré du coefficient de corrélation empirique $\rho_e(y_i, \hat{y}_i)$ entre y_i et $\hat{y}_i$:

$$R^2 = \left(\rho_e(y_i, \hat{y}_i)\right)^2$$

En d'autres termes, le R^2 reflète toujours le degré de corrélation entre y_i et son

prédicteur $\hat{y}_i$.

3- Le R^2 augmente de façon automatique lorsqu'on augmente le nombre de variables explicatives dans une régression multiple. Pour contourner ce problème, on défini parfois un R^2 *ajusté*, noté $\bar{R}^2$, comme[71] :

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

Le $\bar{R}^2$ est parfois utilisé pour sélectionner, sur base de sa 'capacité prédictive', parmi des régressions incluant différents ensembles de variables explicatives, le modèle qui présente le $\bar{R}^2$ le plus élevé. Il s'agit d'une pratique assez peu recommandable.

4- Il existe des décompositions du R^2 visant à attribuer aux différentes variables explicatives la part de la variance expliquée par la régression. On parle alors de R^2 *partiels*. Comme le R^2, ces mesures sont purement descriptives et imparfaites.

6.8. Unités de mesure

Comme dans le cas du modèle de régression simple, les paramètres et les statistiques calculées dans le cadre du modèle de régression linéaire multiple ne sont pas sans unités de mesure : ils dépendent des unités de mesure des observations.

Il est intuitif et pas très difficile de vérifier que les modifications d'unités de mesure des variables ont le même effet dans le cadre du modèle de régression linéaire multiple que dans le cadre du modèle de régression linéaire simple. Ainsi :

1- Une modification des unités de mesure de y_i ou l'ajout d'une constante à y_i a les mêmes effets que dans le modèle simple.

2- Une modification des unités de mesure ou l'ajout d'une constante à une des variables explicatives x_{ij} a les mêmes effets que dans le modèle simple. Il en est de même d'une modification simultanée des unités de mesures de plusieurs variables explicatives (simple addition des effets, pas d'interactions). Dans le cas de l'ajout simultané de constantes à plusieurs variables explicatives, les effets 's'accumulent' sur l'intercept du modèle.

6.9. Forme fonctionnelle

On a vu à la Section 5.4, dans le cadre du modèle de régression linéaire simple, que l'on pouvait, en transformant de façon adéquate y et/ou x, modéliser des *relations non-linéaires* entre les variables x et y, tout en gardant un modèle *linéaire dans les paramètres*.

[71] Il est reporté par GRETL sous la rubrique 'Adjusted R-squared'.

Il va de soi que cette possibilité est également d'application dans le cadre du modèle de régression multiple.

En pratique, les modèles de ce type les plus couramment utilisés sont :

1- Le modèle log-log :

$$\ln y_i = \beta_1 + \beta_2 \ln x_{i2} + ... + \beta_k \ln x_{ik} + e_i \,, \qquad (x_{ij} > 0,\ y_i > 0)$$

On notera qu'il ne peut être utilisé que si tous les x_{ij} et tous les y_i sont strictement positifs. Un exemple classique de son utilisation est donné par l'estimation d'une fonction de production Cobb-Douglas :

$$\begin{aligned} & y = Ak^{\alpha} l^{\beta} \\ \Leftrightarrow \quad & \ln y = \ln A + \alpha \ln k + \beta \ln l \,, \end{aligned}$$

où y désigne la valeur ajouté de la firme, k son stock de capital et l son nombre de travailleurs.

2- Le modèle log-lin :

$$\ln y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i \,, \qquad (y_i > 0)$$

On notera qu'il ne peut être utilisé que si tous les y_i sont strictement positifs. Un exemple classique de son utilisation est donné par l'estimation d'une fonction de salaire (log du salaire en fonction du niveau d'étude et d'expérience professionnelle).

3- Le modèle lin-log :

$$y_i = \beta_1 + \beta_2 \ln x_{i2} + ... + \beta_k \ln x_{ik} + e_i \,, \qquad (x_{ij} > 0)$$

On notera qu'il ne peut être utilisé que si tous les x_{ij} sont strictement positifs.

Les variantes de ces modèles, où les variables explicatives sont pour partie sous forme logarithmique et pour partie non-transformées, sont également très courantes.

Bien entendu, l'interprétation des paramètres de ces différents modèles (en particulier en termes d'élasticité, de semi-élasticité, ou encore de dérivée) est semblable à celle développée à la Section 5.4.

On notera encore que, comme développé à la Section 5.4.4 pour le cas du modèle régression simple, les modèles log-log et log-lin correspondent, pour y_i lui-même, à des modèles non seulement *non-linéaires*, mais aussi *hétéroscédastiques*. De même, on notera que pour ces modèles, du prédicteur ponctuel $\widehat{\ln y_0}$ et de l'intervalle de prévision pour $\ln y_0$ obtenus de la façon habituelle, on peut déduire, en prenant simplement l'exponentielle du prédicteur ponctuel et des bornes de l'intervalle de prévision pour $\ln y_0$, un prédicteur ponctuel et un intervalle de prévision pour y_0 lui-même.

6.9.1. Régression polynomiale

Transformer la variable dépendante et/ou les variables explicatives de modèle de régression standard n'est pas le seul moyen de modifier la forme fonctionnelle du modèle, tout en gardant un modèle linéaire dans les paramètres. On peut encore le faire en ajoutant des puissances et/ou des produits croisés des variables explicatives.

Revenons un instant au cas du modèle de régression linéaire simple standard :

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

On sait qu'on peut modéliser une relation non-linéaire entre x et y en transformant x_i et/ou y_i. Alternativement, on peut utiliser un modèle de *régression polynomiale* du type :

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3 + ... + e_i$$

où x_i^2, x_i^3,... sont traités comme des variables explicatives additionnelles. Pour cette forme fonctionnelle, on a :

$$\frac{dE(y_i|x_i)}{dx_i} = \beta_2 + 2\beta_3 x_i + 3\beta_4 x_i^2 + ...,$$

autrement dit, l'effet de x_i sur y_i n'est pas constant, mais lui-même une fonction de x_i.

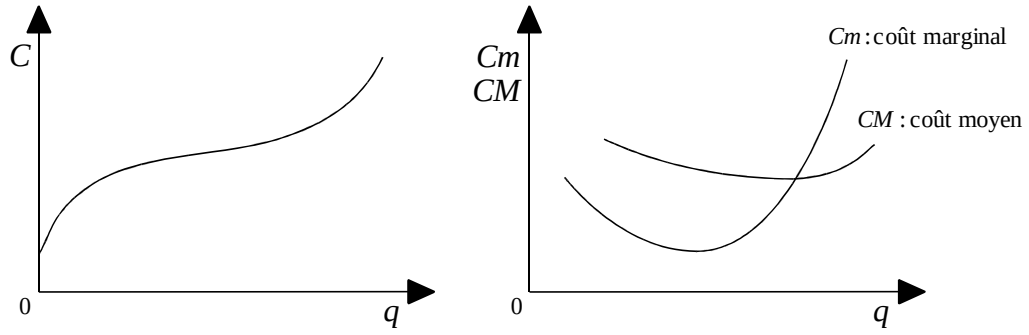
Pour modéliser une fonction de coût (minimum) correspondant à des rendements non-proportionnels, on peut par exemple utiliser le modèle de régression polynomiale :

$$C_i = \beta_1 + \beta_2 q_i + \beta_3 q_i^2 + \beta_4 q_i^3 + e_i\,,$$

où C_i désigne le coût total de production, et q_i le volume de production. Sur base de ce modèle, pour le coût marginal, on a :

$$\frac{dE(C_i|q_i)}{dq_i} = \beta_2 + 2\beta_3 q_i + 3\beta_4 q_i^2$$

A priori, on s'attend à obtenir $\beta_2 > 0$, $\beta_3 < 0$ et $\beta_4 > 0$, autrement dit, un coût marginal d'abord décroissant, puis croissant. Graphiquement :



Graphique 36 : Fonction de coût avec rendements non-proportionnels

De la même façon, outre la possibilité de transformer la variable dépendante et/ou les variables explicatives, la forme fonctionnelle d'un modèle de régression linéaire multiple standard tel que :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

peut être modifiée en considérant le modèle *régression multiple polynomiale* (ici quadratique) :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i3}^2 + \beta_6 (x_{i2} x_{i3}) + e_i$$

où, de même, x_{i2}^2, x_{i3}^2 et $(x_{i2}x_{i3})$ sont traités comme des variables explicatives additionnelles. Pour cette forme fonctionnelle, on a :

$\frac{\partial E(y_i \mid x_{i2}, x_{i3})}{\partial x_{i2}}$	$=$	$\beta_2 + 2\beta_4 x_{i2} + \beta_6 x_{i3}$,	i.e., une fonction (ici linéaire) de x_{i2} et x_{i3}
$\frac{\partial E(y_i \mid x_{i2}, x_{i3})}{\partial x_{i3}}$	$=$	$\beta_3 + 2\beta_5 x_{i2} + \beta_6 x_{i2}$,	i.e., une fonction (ici linéaire) de x_{i2} et x_{i3}
$\frac{\partial^2 E(y_i \mid x_{i2}, x_{i3})}{\partial x_{i2}^2}$	$=$	$2\beta_4$,	i.e., une constante
$\frac{\partial^2 E(y_i \mid x_{i2}, x_{i3})}{\partial x_{i3}^2}$	$=$	$2\beta_5$,	i.e., une constante
$\frac{\partial^2 E(y_i \mid x_{i2}, x_{i3})}{\partial x_{i2} \partial x_{i3}}$	$=$	$\frac{\partial^2 E(y_i \mid x_{i2}, x_{i3})}{\partial x_{i3} \partial x_{i2}} = \beta_6$,	i.e., existence d'une interaction entre x_{i2} et x_{i3}

Par exemple, dans une fonction de salaire, on peut s'attendre à ce que non seulement le niveau d'éducation ($= Educ_i$) et d'expérience professionnelle ($= Expe_i$) influence le salaire ($= Sal_i$) moyen de façon non-linéaire, mais aussi à ce que ces deux facteurs interagissent (l'effet de l'éducation sur le salaire dépend de l'expérience professionnelle, et de même, l'effet de l'expérience professionnelle sur le salaire dépend de l'éducation). Pour capturer ces caractéristiques, on peut considérer le modèle :

$$Sal_i = \beta_1 + \beta_2 Educ_i + \beta_3 Expe_i + \beta_4 Educ_i^2 + \beta_5 Expe_i^2 + \beta_6 (Educ_i Expe_i) + e_i$$

On peut évidemment aussi combiner régression polynomiale et transformations de variables. Ainsi, un modèle couramment utilisé, par exemple pour modéliser une fonction de production plus générale que la fonction Cobb-Douglas, est le modèle log-log polynomial (ici quadratique), aussi appelé modèle 'translog' :

$$\ln y_i = \beta_1 + \beta_2 \ln x_{i2} + \beta_3 \ln x_{i3} + \beta_4 (\ln x_{i2})^2 + \beta_5 (\ln x_{i3})^2 + \beta_6 (\ln x_{i2} \ln x_{i3}) + e_i$$

Un autre modèle couramment utilisé, par exemple pour modéliser une fonction de salaire, est le modèle log-lin polynomial (ici quadratique) :

$$\ln y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i3}^2 + \beta_6 (x_{i2} x_{i3}) + e_i$$

Les propriétés de ces formes fonctionnelles peuvent être analysées de la même façon que ci-dessus. A ce propos, on notera que, dans une régression polynomiale, pour facilité l'interprétation des paramètres, il est utile de centrer (par exemple autour de leur moyenne) les variables explicatives du modèle. Ainsi, dans le cas du modèle translog, si les variables explicatives sont centrées en x_{i2}^* et x_{i3}^*, la partie systématique du modèle s'écrit :

$$\begin{aligned} \ln y_i \;=\; & \beta_1 + \beta_2 (\ln x_{i2} - \ln x_{i2}^*) + \beta_3 (\ln x_{i3} - \ln x_{i3}^*) + \beta_4 (\ln x_{i2} - \ln x_{i2}^*)^2 \\ & + \beta_5 (\ln x_{i3} - \ln x_{i3}^*)^2 + \beta_6 [(\ln x_{i2} - \ln x_{i2}^*)(\ln x_{i3} - \ln x_{i3}^*)] \end{aligned}$$

de sorte qu'on a :

$$\frac{\partial \ln y_i}{\partial \ln x_{i2}} = \beta_2 + 2\beta_4 (\ln x_{i2} - \ln x_{i2}^*) + \beta_6 (\ln x_{i3} - \ln x_{i3}^*) \;, \qquad \text{i.e., une fonction de } \ln x_{i2} \text{ et } \ln x_{i3}$$

$$\frac{\partial \ln y_i}{\partial \ln x_{i3}} = \beta_3 + 2\beta_5 (\ln x_{i3} - \ln x_{i3}^*) + \beta_6 (\ln x_{i2} - \ln x_{i2}^*) \;, \qquad \text{i.e., une fonction de } \ln x_{i2} \text{ et } \ln x_{i3}$$

et donc que β_2 et β_3 s'interprètent directement comme l'élasticité[72] de y_i par rapport à respectivement x_{i2} et x_{i3}, pour $x_{i2} = x_{i2}^*$ et $x_{i3} = x_{i3}^*$. Si les variables explicatives n'étaient pas centrées, β_2 et β_3 s'interpréteraient de la même façon, mais pour $x_{i2} = 1$ et $x_{i3} = 1$ ($\Leftrightarrow \ln x_{i2} = \ln x_{i3} = 0$), ce qui sauf exception correspond à un point (x_{i2}, x_{i3}) sans intérêt.

[72] qui n'est pas constante, mais varient en fonction de x_{i2} et x_{i3}.

Chapitre 7

Test de Fisher, colinéarité et problèmes de spécification

7.1. Le test de Fisher (F-test)

Dans le chapitre précédent, on a vu comment tester séparément les paramètres β_j $(j = 1, ..., k)$, par exemple comment tester $\mathrm{H}_0 : \beta_2 = 0$ contre $\mathrm{H}_1 : \beta_2 \neq 0$ dans le modèle de régression :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \,,$$

ce qui revient à tester :

$$\mathrm{H}_0' : E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_3 x_{i3} \,, \qquad \text{i.e., } E(y_i|x_{i2}, x_{i3}) \text{ ne dépend pas de } x_{i2} \text{ et est linéaire en } x_{i3}$$

$$\text{contre } \mathrm{H}_1' : E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} \,, \quad \text{i.e., } E(y_i|x_{i2}, x_{i3}) \text{ est une fonction linéaire de } x_{i2} \text{ et } x_{i3}$$

On peut souhaiter tester des hypothèses plus élaborées. Par exemple, dans le modèle de fonction de production Cobb-Douglas :

$$\ln y_i = \beta_1 + \beta_2 \ln x_{i2} + \beta_3 \ln x_{i3} + e_i \,,$$

où y_i désigne la valeur ajouté de la firme, x_{i2} son stock de capital et x_{i3} son nombre de travailleurs, on peut souhaiter tester non seulement :

$$(1) \qquad \mathrm{H}_0 : \beta_2 = 0 \text{ contre } \mathrm{H}_1 : \beta_2 \neq 0 \quad \text{et} \quad \mathrm{H}_0 : \beta_3 = 0 \text{ contre } \mathrm{H}_1 : \beta_3 \neq 0 \,,$$

autrement dit la significativité de $\hat{\beta}_2$ et de $\hat{\beta}_3$, mais encore :

(2) $\mathrm{H}_0 : \beta_2 = \beta_3 = 0$ contre $\mathrm{H}_1 : \beta_2 \neq 0$ et/ou $\beta_3 \neq 0$, ce qui revient à tester :

$$\begin{aligned} \mathrm{H}'_0 &: E(\ln y_i | x_{i2}, x_{i3}) = \beta_1 , && \text{i.e., } E(\ln y_i | x_{i2}, x_{i3}) \text{ ne dépend} \\ & && \text{ni de } x_{i2}, \text{ ni de } x_{i3} \\ \text{contre } \mathrm{H}'_1 &: E(\ln y_i | x_{i2}, x_{i3}) = \beta_1 + \beta_2 \ln x_{i2} && \text{i.e., } E(\ln y_i | x_{i2}, x_{i3}) \text{ est une fonct.} \\ & \quad + \beta_3 \ln x_{i3} , && \text{linéaire de } \ln x_{i2} \text{ et } \ln x_{i3} \end{aligned}$$

càd. la significativité de la régression dans son ensemble.

(3) $\mathrm{H}_0 : \beta_2 + \beta_3 = 1$ contre $\mathrm{H}_1 : \beta_2 + \beta_3 \neq 1$, ce qui revient à tester :

$$\begin{aligned} \mathrm{H}'_0 &: E(\ln y_i | x_{i2}, x_{i3}) = \beta_1 + \beta_2 \ln x_{i2} + (1 - \beta_2) \ln x_{i3} \\ \text{contre } \mathrm{H}'_1 &: E(\ln y_i | x_{i2}, x_{i3}) = \beta_1 + \beta_2 \ln x_{i2} + \beta_3 \ln x_{i3} \end{aligned}$$

càd. que les rendements d'échelle sont égaux à 1.

(4) $\mathrm{H}_0 : \beta_2 - \beta_3 = 0$ contre $\mathrm{H}_1 : \beta_2 - \beta_3 \neq 0$, ce qui revient à tester :

$$\begin{aligned} \mathrm{H}'_0 &: E(\ln y_i | x_{i2}, x_{i3}) = \beta_1 + \beta_2 (\ln x_{i2} + \ln x_{i3}) \\ \text{contre } \mathrm{H}'_1 &: E(\ln y_i | x_{i2}, x_{i3}) = \beta_1 + \beta_2 \ln x_{i2} + \beta_3 \ln x_{i3} \end{aligned}$$

càd. que les élasticités partielles de y_i par rapport à x_{i2} et x_{i3} sont égales.

Ces différents tests sont tous des cas particuliers du *test général* :

$$\begin{aligned} \mathrm{H}_0 &: R_0 \beta = r_0 \\ \text{contre } \mathrm{H}_1 &: R_0 \beta \neq r_0 , && \text{i.e., au moins 1 des } q \\ & && \text{restrictions est fausse} \end{aligned}$$

dans le modèle de régression :

$$Y = X\beta + e ,$$

où R_0 est une matrice $q \times k$ de constantes ($q \leq k$; q = le nbr. de restrictions et k = le nbr. de paramètres) et r_0 un vecteur $q \times 1$ de constantes.

Dans l'exemple du modèle de fonction de production Cobb-Douglas décrit ci-dessus, $k = 3$ et on obtient les tests (1), (2), (3) et (4) en prenant :

- pour les tests (1) : respectivement $R_0 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$ et $r_0 = 0$, et $R_0 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$ et $r_0 = 0$.
- pour le test (2) : $R_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ et $r_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
- pour le test (3) : $R_0 = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$ et $r_0 = 1$.
- pour le test (4) : $R_0 = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}$ et $r_0 = 0$.

7.1.1. La procédure de test

7.1.1.1. Cas où σ^2 est connu

Pour simplifier, on commence par considérer le cas où σ^2 est connu.

On sait que, sous les hypothèses A1 à A6, on a :

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

où $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$, de sorte que :

$$(R_0\hat{\beta} - r_0) \sim N(R_0\beta - r_0, \sigma^2 R_0(X'X)^{-1}R_0')$$

En effet, $(R_0\hat{\beta} - r_0)$ est une combinaison linéaire de $\hat{\beta}$. Il est donc lui-même distribué selon une loi normale et :

$$\begin{aligned} E(R_0\hat{\beta} - r_0) &= R_0E(\hat{\beta}) - r_0 \\ &= R_0\beta - r_0 \end{aligned}$$

et

$$\begin{aligned} V(R_0\hat{\beta} - r_0) &= V(R_0\hat{\beta}) = R_0V(\hat{\beta})R_0' \\ &= \sigma^2 R_0(X'X)^{-1}R_0' \end{aligned} \tag{7.1}$$

Ainsi, lorsque la vraie valeur de β est telle que $R_0\beta = r_0$, càd. que H_0 *est vraie*, de la propriété (2.18) de la loi normale multivariée[73], on a :

$$\begin{aligned} \hat{\chi}_0^2 &= (R_0\hat{\beta} - r_0)' \left[R_0V(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= \frac{1}{\sigma^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \sim \chi^2(q)\,, \end{aligned} \tag{7.2}$$

où $\chi^2(q)$ désigne la *loi du khi-carré*[74] à q degrés de liberté, tandis que si la vraie valeur de β est telle que $R_0\beta \neq r_0$, càd. que H_0 *est fausse*, on peut montrer qu'on a :

$$\begin{aligned} \hat{\chi}_0^2 &= (R_0\hat{\beta} - r_0)' \left[R_0V(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= \frac{1}{\sigma^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \sim \chi^2(\delta^*, q)\,, \end{aligned} \tag{7.3}$$

où $\chi^2(\delta^*, q)$ désigne la *loi du khi-carré non-centrale*[75] à q degrés de liberté et le paramètre de non-centralité δ^* est égal à $\delta^* = (R_0\beta - r_0)' \left[R_0V(\hat{\beta})R_0'\right]^{-1} (R_0\beta - r_0)$.

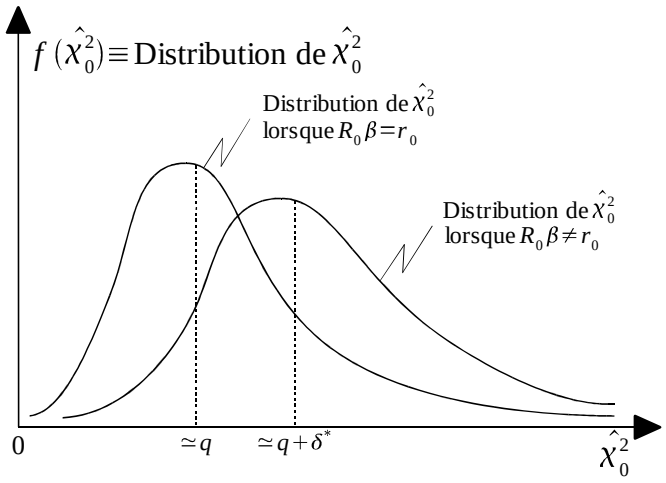
---

[73] Cf. Section 2.3.1.

[74] Cf. l'annexe B de Hill, Griffiths et Lim (2008).

[75] Par définition, si $X \sim N(m, \Sigma)$, où X est un vecteur de dimension $q \times 1$, alors : $X'\Sigma^{-1}X \sim \chi^2(\delta, q)$, où $\delta = m'\Sigma^{-1}m$.

Autrement dit, si H_0 est vraie (i.e., $R_0\beta = r_0$), $\hat{\chi}_0^2$ suit une loi du khi-carré standard, tandis que si H_0 est fausse (i.e., $R_0\beta \neq r_0$), le même $\hat{\chi}_0^2$ suit une loi du khi-carré non-centrale, dont le paramètre de non-centralité δ^* est d'autant plus grand que H_0 est fausse. En effet, $\delta^* = (R_0\beta - r_0)' \left[R_0 V(\hat{\beta}) R_0'\right]^{-1} (R_0\beta - r_0)$ est une forme quadratique dont la matrice $\left[R_0 V(\hat{\beta}) R_0'\right]^{-1}$ est définie positive[76]. Cela implique que plus $R_0\beta - r_0$ est différent de zéro (i.e., $R_0\beta \neq r_0$), plus δ^* est grand. Graphiquement :



Graphique 37 : Distribution de $\hat{\chi}_0^2$

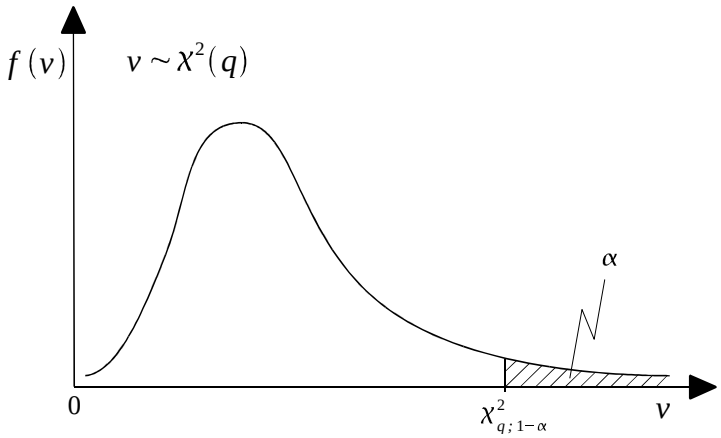
Etant donné ses propriétés, on peut utiliser $\hat{\chi}_0^2$ comme *statistique de test* pour tester H_0 : $R_0\beta = r_0$ contre H_1 : $R_0\beta \neq r_0$ (i.e., au moins 1 des q contraintes est fausse).

Un test au seuil α de H_0 : $R_0\beta = r_0$ contre H_1 : $R_0\beta \neq r_0$ est donné par la règle de décision :

$$\begin{cases} \text{- Rejet de } H_0 \text{ si } \hat{\chi}_0^2 > \chi^2_{q;1-\alpha} \\ \text{- Non-rejet de } H_0 \text{ sinon} \end{cases}$$

où la *valeur critique* $\chi^2_{q;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(q)$, càd. la valeur $\chi^2_{q;1-\alpha}$ telle que $I\!P(v \leq \chi^2_{q;1-\alpha}) = 1 - \alpha$, où $v \sim \chi^2(q)$. Graphiquement :

[76] De (7.1), on a $V(R_0\hat{\beta}) = R_0 V(\hat{\beta}) R_0'$. Autrement dit, $R_0 V(\hat{\beta}) R_0'$ est la matrice de variance-covariance de $R_0\hat{\beta}$, et est donc une matrice définie positive (sauf dans le cas pathologique où R_0 ou X n'est pas de rang plein). Comme l'inverse d'une matrice définie positive est elle-même toujours définie positive, $\left[R_0 V(\hat{\beta}) R_0'\right]^{-1}$ est donc aussi une matrice définie positive.

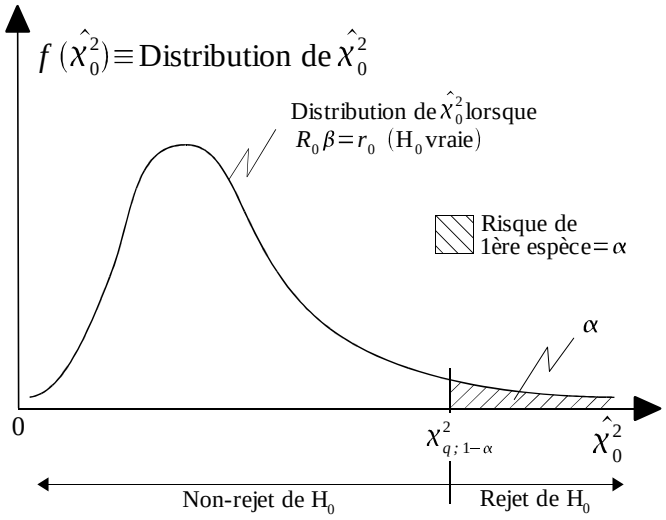


Graphique 38 : Quantile d'ordre $1-\alpha$ de la loi $\chi^2(q)$

Par construction, le seuil α du test est égal au *risque de première espèce* du test :

$$I\!P(\mathrm{RH_0} \,|\, \mathrm{H_0 \text{ est vraie}}) = \alpha$$

Graphiquement :

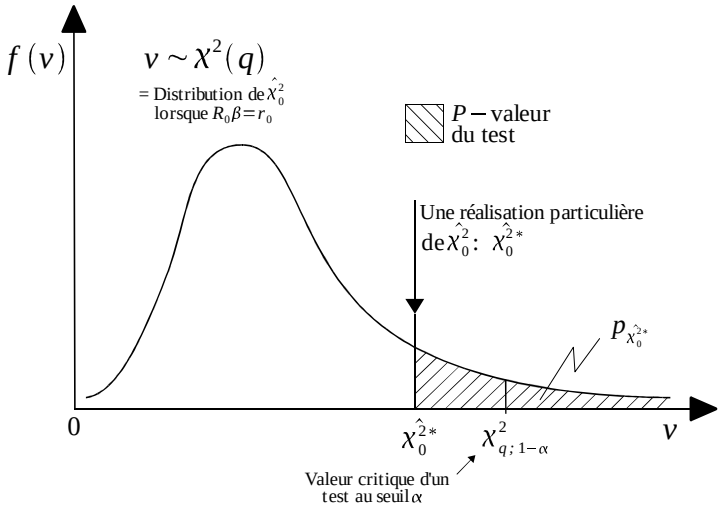


Graphique 39 : Risque de première espèce du χ^2-test

Soit $\hat{\chi}_0^{2*}$ la valeur de la statistique de test $\hat{\chi}_0^2$ obtenue pour un échantillon particulier. Pour cet *échantillon particulier*, la *P-valeur* du test est donnée par :

$$p_{\hat{\chi}_0^{2*}} = I\!P(v > \hat{\chi}_0^{2*}), \quad \text{où } v \sim \chi^2(q)$$

Graphiquement :



Graphique 40 : P-valeur du χ^2-test

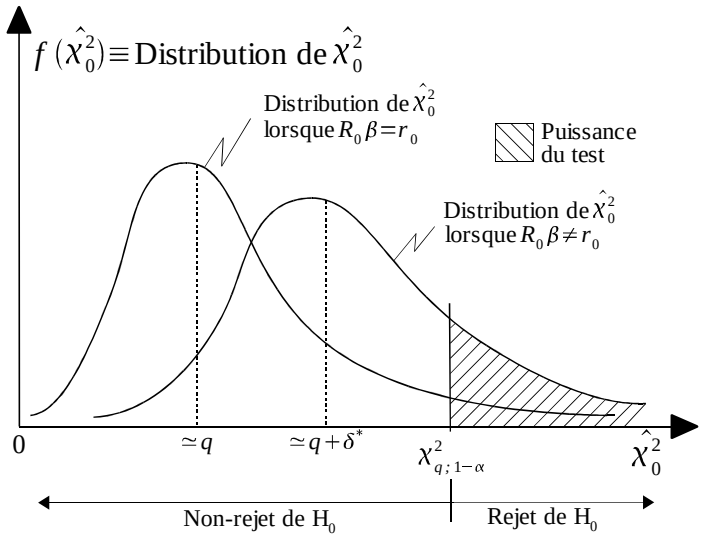
Comme toujours, la P-valeur $p_{\hat{\chi}_0^{2*}}$ du test est la *valeur minimale* du seuil α du test pour laquelle on peut, pour un échantillon particulier, rejeter H_0.

Pour α fixé, la *puissance du test*, càd. $I\!P(\mathrm{RH}_0 \,|\, \mathrm{H}_0 \text{ est fausse})$, sera d'autant plus grande que le paramètre de non-centralité $\delta^* = (R_0\beta - r_0)' \left[R_0 V(\hat{\beta}) R_0'\right]^{-1} (R_0\beta - r_0)$ est grand, càd. que :

1- H_0 est fausse (i.e., $|R_0\beta - r_0|$ est grand).

2- la précision d'estimation de β est grande (i.e., $V(\hat{\beta})$ est petite, au sens matriciel).

Graphiquement :



Graphique 41 : Puissance du χ^2-test

Comme pour tout test, plus on peut rejeter H_0 pour α petit, plus on peut être

confiant dans le fait que H_0 est effectivement fausse. Cependant, il faut se garder, en particulier lorsque la précision d'estimation n'est pas très grande, d'interpréter un non-rejet de H_0 comme une preuve convaincante que H_0 est vraie (car au contraire du risque de première espèce, la puissance du test n'est pas sous contrôle). On gardera aussi à l'esprit que lorsque la précision d'estimation est grande, un rejet de H_0, même très marqué, ne signifie pas nécessairement qu'on en est loin.

7.1.1.2. Cas où σ^2 est inconnu

En pratique, la variance du terme d'erreur σ^2 est inconnue. Comme d'habitude, on peut cependant la remplacer par son estimateur convergent et non biaisé $\hat{s}^2$.

On a vu à la section précédente que, sous les hypothèses A1 à A6, lorsque la vraie valeur de β est telle que $R_0\beta = r_0$, càd. que H_0 est vraie, on a :

$$\hat{\chi}_0^2 = \frac{1}{\sigma^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \sim \chi^2(q)$$

et on sait par ailleurs que, toujours sous les hypothèses A1 à A6, on a aussi (cf. Section 6.4):

$$\hat{v} = \frac{(n-k)\hat{s}^2}{\sigma^2} \sim \chi^2(n-k)$$

On peut encore montrer que $\hat{\chi}_0^2$ et $\hat{v}$ sont indépendamment distribuées, de sorte que, de la définition de la loi de Fisher[77], sous les hypothèses A1 à A6, lorsque la vraie valeur de β est telle que $R_0\beta = r_0$, càd. que H_0 est vraie, on a :

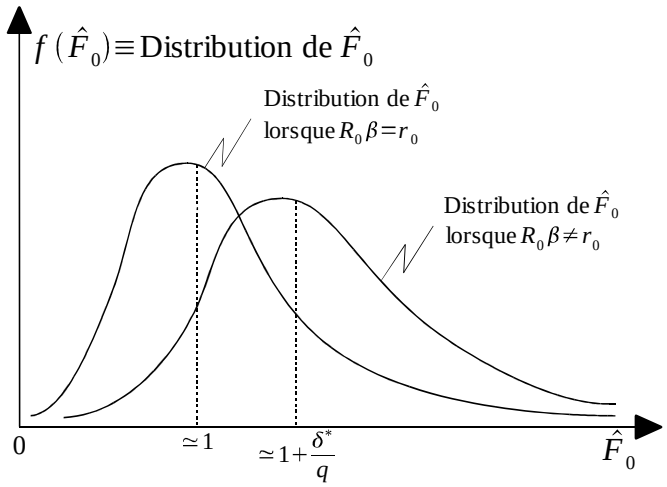
$$\begin{aligned}
\hat{F}_0 &= \frac{\frac{\hat{\chi}_0^2}{q}}{\frac{\hat{v}}{n-k}} = \frac{\frac{1}{q\sigma^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0)}{\frac{\hat{s}^2}{\sigma^2}} \\
&= \frac{1}{q\hat{s}^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\
&= \frac{1}{q}(R_0\hat{\beta} - r_0)' \left[R_0\hat{V}(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \sim F(q, n-k)\,,
\end{aligned}$$

tandis que lorsque la vraie valeur de β est telle que $R_0\beta \neq r_0$, càd. que H_0 est fausse, on peut montrer qu'on a :

[77] Si $v_1 \sim \chi^2(m_1)$, $v_2 \sim \chi^2(m_2)$ et que v_1 et v_2 sont indépendamment distribués, alors : $F = \frac{\frac{v_1}{m_1}}{\frac{v_2}{m_2}} \sim F(m_1, m_2)$. Cf. l'annexe B de Hill, Griffiths et Lim (2008).

$$\begin{aligned}\hat{F}_0 &= \frac{1}{q\hat{s}^2}(R_0\hat{\beta}-r_0)'\left[R_0(X'X)^{-1}R_0'\right]^{-1}(R_0\hat{\beta}-r_0)\\ &= \frac{1}{q}(R_0\hat{\beta}-r_0)'\left[R_0\hat{V}(\hat{\beta})R_0'\right]^{-1}(R_0\hat{\beta}-r_0)\sim F(\delta^*,q,n-k)\,,\end{aligned}$$

où $F(\delta^*, q, n-k)$ désigne la *loi de Fisher non-centrale*[78] à q et $(n-k)$ degrés de liberté, et le *paramètre de non-centralité* δ^* est égal à $\delta^* = (R_0\beta - r_0)'\left[R_0V(\hat{\beta})R_0'\right]^{-1}(R_0\beta - r_0)$. Graphiquement :



Graphique 42 : Distribution de $\hat{F}_0$

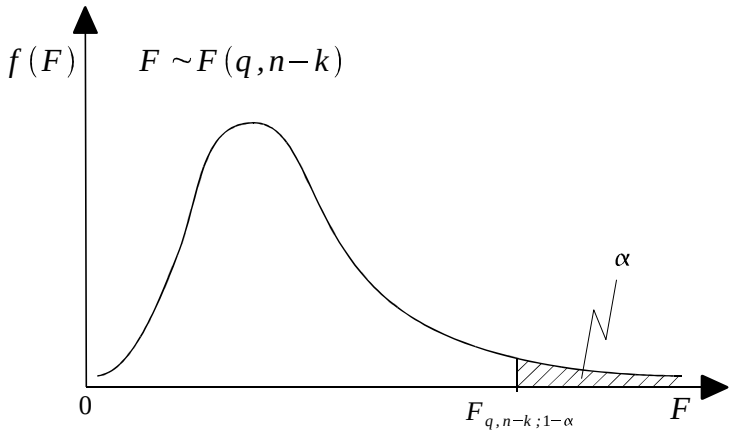
On constate qu'à la transposition loi du khi-carré / loi de Fisher près, le comportement de la statistique $\hat{F}_0$ — qui, notez-le, est égale à la statistique $\hat{\chi}_0^2$ divisée par q et où σ^2 est remplacé par $\hat{s}^2$ — est en tout point semblable à celui de le statistique $\hat{\chi}_0^2$.

Ainsi, de façon semblable au cas où σ^2 est connu, un test au seuil α de H_0 : $R_0\beta = r_0$ contre H_1 : $R_0\beta \neq r_0$ est donné par la règle de décision :

$$\begin{cases} \text{- Rejet de } \mathrm{H}_0 \text{ si } \hat{F}_0 > F_{q,n-k;1-\alpha} \\ \text{- Non-rejet de } \mathrm{H}_0 \text{ sinon} \end{cases}$$

où la *valeur critique* $F_{q,n-k;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher $F(q, n-k)$, càd. la valeur $F_{q,n-k;1-\alpha}$ telle que $\mathbb{P}(F \leq F_{q,n-k;1-\alpha}) = 1-\alpha$, où $F \sim F(q, n-k)$. Graphiquement :

[78] Par définition, Si $v_1 \sim \chi^2(\delta, m_1)$, $v_2 \sim \chi^2(m_2)$ et que v_1 et v_2 sont indépendamment distribués, alors : $F = \frac{\frac{v_1}{m_1}}{\frac{v_2}{m_2}} \sim F(\delta, m_1, m_2)$.

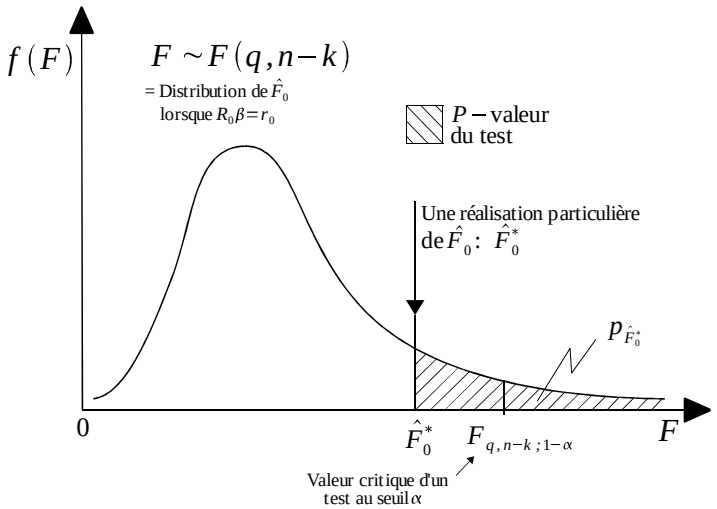


Graphique 43 : Quantile d'ordre $1-\alpha$ de la loi $F(q, n-k)$

La *P-valeur* de ce test, pour un *échantillon particulier*, est donnée par :

$$p_{\hat{F}_0^*} = I\!P(F > \hat{F}_0^*)\,, \quad \text{où } F \sim F(q, n-k)$$

Graphiquement :



Graphique 44 : P-valeur du F-test

Les interprétations en termes de risque de première espèce, de puissance, ainsi que l'interprétation de la P-valeur de ce test sont identiques à celles développées pour le cas où σ^2 est connu.

Pour conclure, on notera que la statistique de test :

$$\hat{F}_0 = \frac{1}{q}(R_0\hat{\beta} - r_0)' \left[R_0\hat{V}(\hat{\beta})R_0'\right]^{-1}(R_0\hat{\beta} - r_0) \tag{7.4}$$

peut être réécrite sous une forme différente. On peut en effet montrer que :

$$\hat{F}_0 = \frac{(\hat{e}_c'\hat{e}_c - \hat{e}'\hat{e})/q}{\hat{e}'\hat{e}/(n-k)} = \frac{(\mathrm{SCR}_c - \mathrm{SCR})/q}{\mathrm{SCR}/(n-k)} \tag{7.5}$$

où : SCR = la somme des carrés des résidus de la régression
$Y = X\beta + e$.
SCR_c = la somme des carrés des résidus de la régression
$Y = X\beta + e$, où β est estimé sous la contrainte $R_0\beta = r_0$
(moindres carrés contraints).

La somme des carrés des résidus contraints (= SCR_c) est très facile à obtenir lorsque la contrainte $R_0\beta = r_0$ est de forme simple. Par exemple, pour le test de $\text{H}_0 : \beta_2 = \beta_3 = 0$ contre $\text{H}_1 : \beta_2 \neq 0$ et/ou $\beta_3 \neq 0$ dans la régression :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i \,,$$

SCR_c est simplement donné par la somme des carrés des résidus de la régression contrainte :

$$y_i = \beta_1 + \beta_4 x_{i4} + e_i \,,$$

càd. de la régression initiale d'où on a retiré les variables x_{i2} et x_{i3} (puisque sous la contrainte, $\beta_2 = \beta_3 = 0$).

La forme (7.5) du F-test montre que ce test peut être regardé comme un test qui examine si l'imposition des contraintes impliquées par $\text{H}_0 : R_0\beta = r_0$ accroît significativement ou non (si elle l'accroît trop fortement, on rejette H_0) la somme des carrés des résidus d'une régression.

La plupart des logiciels économétriques (GRETL en particulier) permettent de calculer de façon très simple le F-test sous sa forme générale (7.4). Il suffit de spécifier les contraintes $R_0\beta = r_0$, et le logiciel reporte alors la valeur de la statistique $\hat{F}_0$ et la P-valeur du test. La forme (7.5) du F-test n'est en pratique vraiment utile que lorqu'on ne dispose pas d'un logiciel qui permet de le calculer de façon simple sous sa forme générale (7.4).

7.1.2. F-test et non-normalité

Nous avons obtenu le F-test en supposant que, outre les hypothèses A1 à A5, l'hypothèse optionnelle de normalité A6 du modèle était satisfaite. Qu'en est-il si, comme on peut couramment s'y attendre en pratique, cette dernière hypothèse n'est pas remplie ?

Comme nous allons le voir, lorsqu'on renonce à l'hypothèse A6 de normalité, le F-test reste valable, mais seulement asymptotiquement, en grand échantillon.

On sait que, sous les hypothèses A1 à A5, sans faire appel à l'hypothèse A6 de normalité, on a *asymptotiquement* (lorsque $n \to \infty$) :

$$\left[V(\hat{\beta})\right]^{-\frac{1}{2}} \left(\hat{\beta} - \beta\right) \xrightarrow{d} N(0, I) \,, \quad \text{où } V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \,,$$

soit, sous forme d'approximation utilisable en échantillon fini pour n suffisamment

grand :

$$\hat{\beta} \approx N(\beta, \sigma^2(X'X)^{-1}) \,,$$

de sorte qu'on a encore, sous forme d'approximation :

$$(R_0\hat{\beta} - r_0) \approx N(R_0\beta - r_0, \sigma^2 R_0(X'X)^{-1}R_0')$$

Ainsi, lorsque la vraie valeur de β est telle que $R_0\beta = r_0$, càd. que H_0 *est vraie*, on a toujours, sous forme d'approximation :

$$\begin{aligned} \hat{\chi}_0^2 &= (R_0\hat{\beta} - r_0)' \left[R_0 V(\hat{\beta}) R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= \frac{1}{\sigma^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(q) \,, \end{aligned} \tag{7.6}$$

tandis que si la vraie valeur de β est telle que $R_0\beta \neq r_0$, càd. que H_0 *est fausse*, on peut montrer qu'on a encore, sous forme d'approximation :

$$\begin{aligned} \hat{\chi}_0^2 &= (R_0\hat{\beta} - r_0)' \left[R_0 V(\hat{\beta}) R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= \frac{1}{\sigma^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(\delta^*, q) \,, \end{aligned} \tag{7.7}$$

où $\delta^* = (R_0\beta - r_0)' \left[R_0 V(\hat{\beta}) R_0'\right]^{-1} (R_0\beta - r_0)$.

Les résultats (7.6) et (7.7) sont des *versions asymptotiques* (valables uniquement pour n grand) des *résultats exacts* de distribution d'échantillonnage (7.2) et (7.3) sur lesquels nous nous sommes appuyés pour obtenir un χ^2-test sous l'hypothèse A6 de normalité et dans le cas où σ^2 est connu.

Sous l'hypothèse A6 de normalité et pour le cas où σ^2 n'est pas connu, nous avons vu que, pour l'essentiel, le remplacement de σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$ faisait passer de la statistique $\hat{\chi}_0^2$ à la statistique $\hat{F}_0$ et de lois khi-carrés à des lois de Fisher[79].

Asymptotiquement, lorsque n est grand, on peut montrer que le remplacement de σ^2 par son estimateur convergent et non biaisé $\hat{s}^2$ ne modifie pas les distributions d'échantillonnage en jeu, de sorte qu'on a aussi, sous forme d'approximation, lorsque H_0 est vraie (i.e., $R_0\beta = r_0$) :

$$\begin{aligned} \hat{\chi}_0^{2\prime} &= (R_0\hat{\beta} - r_0)' \left[R_0 \hat{V}(\hat{\beta}) R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= \frac{1}{\hat{s}^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(q) \,, \end{aligned}$$

[79] Pour rappel, la statistique $\hat{F}_0$ est égale à la statistique $\hat{\chi}_0^2$ divisée par q, et où σ^2 est remplacé par $\hat{s}^2$.

et lorsque H_0 est fausse (i.e., $R_0\beta \neq r_0$) :

$$\begin{aligned}\hat{\chi}_0^{2\prime} &= (R_0\hat{\beta} - r_0)' \left[R_0 \hat{V}(\hat{\beta}) R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= \frac{1}{\hat{s}^2}(R_0\hat{\beta} - r_0)' \left[R_0 (X'X)^{-1} R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(\delta^*, q)\,,\end{aligned}$$

où $\delta^* = (R_0\beta - r_0)' \left[R_0 V(\hat{\beta}) R_0'\right]^{-1} (R_0\beta - r_0)$.

Ainsi, le χ^2-test obtenu sous l'hypothèse A6 de normalité et dans le cas où σ^2 est connu reste valable *asymptotiquement*, à titre approximatif pour n grand, sous les seules hypothèses A1 à A5 et avec σ^2 remplacé par $\hat{s}^2$. Ce χ^2-test est couramment appelé *test du khi-carré* ou encore *test de Wald*.

Le F-test décrit à la section précédente, qui est *exact en échantillon fini* sous l'hypothèse de normalité A6, est, pour n grand, *asymptotiquement équivalent* à ce χ^2-test. En effet, on a d'une part :

$$\hat{F}_0 = \frac{\hat{\chi}_0^{2\prime}}{q}\,,$$

et d'autre part, lorsque $n \to \infty$:

$$F_{q,n-k;1-\alpha} \simeq \frac{\chi^2_{q;1-\alpha}}{q} \quad \Leftrightarrow \quad qF_{q,n-k;1-\alpha} \simeq \chi^2_{q;1-\alpha}\,,$$

où $F_{q,n-k;1-\alpha}$ et $\chi^2_{q;1-\alpha}$ désignent les quantiles d'ordre $1-\alpha$ de respectivement la loi $F(q, n-k)$ et la loi $\chi^2(q)$. Ce dernier résultat vient du fait que si $F \sim F(m_1, m_2)$, lorsque $m_2 \to \infty$, on a $m_1 F \xrightarrow{d} \chi^2(m_1)$, autrement dit, que si une variable aléatoire F est distribuée selon une loi de Fisher $F(m_1, m_2)$, lorsque $m_2 \to \infty$, la variable aléatoire $m_1 F$ tend en distribution vers la loi $\chi^2(m_1)$, soit sous forme d'approximation pour n grand : $m_1 F \approx \chi^2(m_1)$.

On en déduit que le F-test, qui est *exact en échantillon fini* sous l'hypothèse A6 de normalité, reste également valable *asymptotiquement*, à titre approximatif, pour n grand, sous les seules hypothèses A1 à A5.

En pratique, pour tester en grand échantillon $H_0 : R_0\beta = r_0$ contre $H_0 : R_0\beta = r_0$ sans faire appel à l'hypothèse A6 de normalité, on peut ainsi indifféremment utiliser le χ^2-test (basé sur la statistique $\hat{\chi}_0^{2\prime}$) ou le F-test. L'usage veut qu'on utilise généralement le F-test, car il est non seulement valable, à titre approximatif, pour n grand et sans faire appel à l'hypothèse de normalité, mais aussi exact (quel que soit n, en particulier n petit) si l'hypothèse de normalité tient (ce qui n'est pas le cas du χ^2-test).

7.1.3. Cas particuliers du F-test

7.1.3.1. Le F-test de la significativité d'un paramètre

On considère le test de $\mathrm{H}_0 : \beta_j = 0$ contre $\mathrm{H}_1 : \beta_j \neq 0$ dans la régression :

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i \,,$$

càd. le test de la significativité de $\hat{\beta}_j$.

On a vu que ce test pouvait être réalisé à l'aide d'un t-test (test de Student). Il peut également être réalisé à l'aide d'un F-test, qui est totalement équivalent au t-test.

Sous sa forme générale (7.4), la statistique de test $\hat{F}_0$ de $\mathrm{H}_0 : \beta_j = 0$ contre $\mathrm{H}_1 : \beta_j \neq 0$ est obtenue en prenant $q = 1$, $r_0 = 0$ et $R_0 = \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}$, i.e., pour R_0, un vecteur $1 \times k$ avec un 1 en $j^{\text{ième}}$ position et des 0 partout ailleurs. Pour ces valeurs, on a :

$$R_0\hat{\beta} = \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_j \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \hat{\beta}_j$$

et

$$\begin{aligned} R_0\hat{V}(\hat{\beta})R_0' &= \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \hat{Var}(\hat{\beta}_1) & \cdots & \hat{Cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \ddots & \vdots \\ \hat{Cov}(\hat{\beta}_k, \hat{\beta}_1) & \cdots & \hat{Var}(\hat{\beta}_k) \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \\ &= \hat{Var}(\hat{\beta}_j) = \hat{s}^2 q_{jj} \,, \quad \text{où } q_{jj} = \left[(X'X)^{-1}\right]_{jj} \,, \end{aligned}$$

de sorte que :

$$\hat{F}_0 = \frac{\hat{\beta}_j^2}{\hat{Var}(\hat{\beta}_j)} = \left(\frac{\hat{\beta}_j}{s.\hat{e}.(\hat{\beta}_j)}\right)^2 = \hat{t}_o^2 \,, \qquad \text{où } s.\hat{e}.(\hat{\beta}_j) = \sqrt{\hat{s}^2 q_{jj}}$$

On constate que la statistique $\hat{F}_0$ est tout simplement égale au carré de la statistique $\hat{t}_o$ sur laquelle est fondé le t-test. Sachant que si $t \sim t(\nu)$, alors $t^2 \sim F(1, \nu)$, ce qui implique que $F_{1,n-k;1-\alpha} = \left(t_{n-k;1-\frac{\alpha}{2}}\right)^2$, où $F_{1,n-k;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $F(1, n-k)$ et $t_{n-k;1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $t(n-k)$, on voit que le t-test et le F-test de $\mathrm{H}_0 : \beta_j = 0$ contre $\mathrm{H}_1 : \beta_j \neq 0$ sont totalement équivalents.

On peut vérifier (faites-le !) que l'équivalence t-test/F-test tient également pour le test de $\mathrm{H}_0 : \ \beta_j = \beta_j^o$ contre $\mathrm{H}_1 : \ \beta_j \neq \beta_j^o$, quelle que soit la valeur de β_j^o. On notera encore que le F-test n'est équivalent qu'au t-test *bilatéral*, pas à un t-test *unilatéral*.

Le F-test de $\mathrm{H}_0 : \ \beta_j = 0$ contre $\mathrm{H}_1 : \ \beta_j \neq 0$ peut également être facilement obtenu de la forme (7.5) de la statistique $\hat{F}_0$. Dans ce cas, outre la somme des carrés des résidus de la régression non-contrainte (= SCR), on a besoin de la somme des carrés des résidus de la régression contrainte (= SCR_c), qui est ici simplement donnée par la somme des carrés des résidus de la régression :

$$y_i = \beta_1 + ... + \beta_{j-1} x_{i(j-1)} + \beta_{j+1} x_{i(j+1)} + ... + \beta_k x_{ik} + e_i \, ,$$

càd. de la régression initiale d'où on a retiré la variable x_{ij} (puisque sous la contrainte, $\beta_j = 0$). Dans le cas d'un test de $\mathrm{H}_0 : \ \beta_j = \beta_j^o$ contre $\mathrm{H}_1 : \ \beta_j \neq \beta_j^o$, SCR_c serait donné par la somme des carrés des résidus de la régression :

$$(y_i - \beta_j^o x_{ij}) = \beta_1 + ... + \beta_{j-1} x_{i(j-1)} + \beta_{j+1} x_{i(j+1)} + ... + \beta_k x_{ik} + e_i \, ,$$

càd. la régression initiale d'où on a retiré la variable x_{ij} et dont la variable dépendante y_i est remplacée par $(y_i - \beta_j^o x_{ij})$.

7.1.3.2. Le F-test d'une combinaison linéaire scalaire de paramètres

On considère à titre d'exemple le test de $\mathrm{H}_0 : \ \beta_2 + \beta_3 = 1$ contre $\mathrm{H}_1 : \ \beta_2 + \beta_3 \neq 1$ dans la régression :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + e_i$$

Sous sa forme générale (7.4), la statistique de test $\hat{F}_0$ de $\mathrm{H}_0 : \ \beta_2 + \beta_3 = 1$ contre $\mathrm{H}_1 : \ \beta_2 + \beta_3 \neq 1$ est obtenue en prenant $q = 1$, $r_0 = 1$ et $R_0 = \begin{bmatrix} 0 & 1 & 1 & 0 & \cdots & 0 \end{bmatrix}$, où R_0 est un vecteur $1 \times k$. Pour ces valeurs, on a :

$$R_0 \hat{\beta} = \begin{bmatrix} 0 & 1 & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \hat{\beta}_2 + \hat{\beta}_3$$

et

$$R_0 \hat{V}(\hat{\beta}) R_0' = \begin{bmatrix} 0 & 1 & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \hat{Var}(\hat{\beta}_1) & \cdots & \hat{Cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \ddots & \vdots \\ \hat{Cov}(\hat{\beta}_k, \hat{\beta}_1) & \cdots & \hat{Var}(\hat{\beta}_k) \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{aligned} &= \hat{Var}(\hat{\beta}_2) + 2\hat{Cov}(\hat{\beta}_2, \hat{\beta}_3) + \hat{Var}(\hat{\beta}_3) \\ &= \hat{Var}(\hat{\beta}_2 + \hat{\beta}_3) = \hat{Var}(R_0\hat{\beta}) \end{aligned}$$

de sorte que :

$$\hat{F}_0 = \frac{\Big((\hat{\beta}_2 + \hat{\beta}_3) - 1\Big)^2}{\hat{Var}(\hat{\beta}_2 + \hat{\beta}_3)} = \left(\frac{(\hat{\beta}_2 + \hat{\beta}_3) - 1}{s.\hat{e}.(\hat{\beta}_2 + \hat{\beta}_3)}\right)^2 = \hat{t}_o^2 \,,$$

où $s.\hat{e}.(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{\hat{Var}(\hat{\beta}_2 + \hat{\beta}_3)}$.

On constate que la statistique $\hat{F}_0$ apparaît à nouveau comme le carré de ce qui ressemble fort à la statistique d'un t-test qui testerait l'égalité à 1 de la combinaison linéaire scalaire $R_0\beta = \beta_2 + \beta_3$: $\hat{t}_0 = \frac{R_0\hat{\beta}-1}{s.\hat{e}.(R_0\hat{\beta})}$.

Ce résultat suggère qu'un test de l'égalité à une constante r_0 d'une combinaison linéaire *scalaire* $R_0\beta$ de β, càd. un test de $\mathrm{H}_0 : R_0\beta = r_0$ contre $\mathrm{H}_1 : R_0\beta \neq r_0$, où R_0 est un vecteur $1 \times k$ (i.e., un vecteur ligne ; donc une seule restriction), peut être réalisé au travers d'un t-test, et que ce t-test (bilatéral) est totalement équivalent au F-test de cette même restriction sur les paramètres.

C'est bien le cas. En effet, sous les hypothèses A1 à A6, on sait qu'on a (cf. Section 6.4) :

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}) \,,$$

de sorte que, si R_0 est un vecteur $1 \times k$, on a :

$$\begin{aligned} & R_0\hat{\beta} \sim N(R_0\beta, \sigma^2 R_0(X'X)^{-1}R_0') \\ \Leftrightarrow \quad & \hat{z} = \frac{R_0\hat{\beta} - R_0\beta}{s.e.(R_0\hat{\beta})} \sim N(0,1) \,, \end{aligned}$$

où $s.e.(R_0\hat{\beta}) = \sqrt{Var(R_0\hat{\beta})} = \sqrt{\sigma^2 R_0(X'X)^{-1}R_0'} = \sqrt{R_0 V(\hat{\beta}) R_0'}$.

En particulier, lorsque $R_0\beta = r_0$, on a :

$$\hat{z}_o = \frac{R_0\hat{\beta} - r_0}{s.e.(R_0\hat{\beta})} \sim N(0,1) \,,$$

tandis que lorsque $R_0\beta \neq r_0$, on a :

$$\hat{z}_o = \frac{R_0\hat{\beta} - r_0}{s.e.(R_0\hat{\beta})} \sim N\left(\frac{R_0\beta - r_0}{s.e.(R_0\hat{\beta})}, 1\right)$$

Par ailleurs, on sait que, sous les hypothèses A1 à A6, on a aussi (cf. Section

6.4) :

$$\hat{v} = \frac{(n-k)\hat{s}^2}{\sigma^2} \sim \chi^2(n-k)\,,$$

et on peut encore montrer que $\hat{z}$ et $\hat{v}$ sont indépendamment distribués, de sorte que, de la définition de la loi de Student, on a :

$$\hat{t} = \frac{\hat{z}}{\sqrt{\dfrac{\hat{v}}{n-k}}} = \frac{R_0\hat{\beta} - R_0\beta}{s.\hat{e}.(R_0\hat{\beta})} \sim t(n-k)\,,$$

où $s.\hat{e}.(R_0\hat{\beta}) = \sqrt{V\hat{a}r(R_0\hat{\beta})} = \sqrt{\hat{s}^2 R_0(X'X)^{-1}R_0'} = \sqrt{R_0\hat{V}(\hat{\beta})R_0'}$.

En particulier, lorsque $R_0\beta = r_0$, on a :

$$\hat{t}_o = \frac{R_0\hat{\beta} - r_0}{s.\hat{e}.(R_0\hat{\beta})} \sim t(n-k)\,,$$

tandis que lorsque $R_0\beta \neq r_0$, on peut montrer qu'on a :

$$\hat{t}_o = \frac{R_0\hat{\beta} - r_0}{s.\hat{e}.(R_0\hat{\beta})} \sim t(\delta^*, n-k)\,, \quad \text{où } \delta^* = \frac{R_0\beta - r_0}{s.e.(R_0\hat{\beta})}$$

On voit qu'au remplacement de β_j par $R_0\beta$, de $\hat{\beta}_j$ par $R_0\hat{\beta}$ et de β_j^o par r_0 près, les résultats ci-dessus sont identiques à ceux sur lesquels nous nous sommes appuyés à la Section 6.4 pour construire des t-tests (bilatéral et unilatéral) de β_j et un intervalle de confiance pour β_j.

On en conclut que, lorsque R_0 est un vecteur $1 \times k$ (i.e., un vecteur ligne ; donc une seule restriction) :

1- Un test de $\mathrm{H}_0 : R_0\beta = r_0$ contre $\mathrm{H}_1 : R_0\beta \neq r_0$ peut être effectué au travers d'un t-test *bilatéral* standard basé sur la statistique :

$$\hat{t}_o = \frac{R_0\hat{\beta} - r_0}{s.\hat{e}.(R_0\hat{\beta})}\,, \quad \text{où } s.\hat{e}.(R_0\hat{\beta}) = \sqrt{\hat{s}^2 R_0(X'X)^{-1}R_0'} = \sqrt{R_0\hat{V}(\hat{\beta})R_0'}\,,$$

et que pour la même raison que celle évoquée à la section précédente (i.e., si $t \sim t(\nu)$, alors $t^2 \sim F(1,\nu)$, ce qui implique que $F_{1,n-k;1-\alpha} = \left(t_{n-k;1-\frac{\alpha}{2}}\right)^2$), ce test est totalement équivalent au F-test de la même restriction.

2- Des tests de $\mathrm{H}_0 : R_0\beta \geq r_0$ contre $\mathrm{H}_1 : R_0\beta < r_0$ et de $\mathrm{H}_0 : R_0\beta \leq r_0$ contre $\mathrm{H}_1 : R_0\beta > r_0$ peuvent pareillement être effectués au travers de t-tests *unilatéraux* standards, toujours basé sur la même statistique $\hat{t}_o = \frac{R_0\hat{\beta} - r_0}{s.\hat{e}.(R_0\hat{\beta})}$.

3- Un intervalle de confiance à $(1-\alpha)\times 100\%$ pour $R_0\beta$ est de façon semblable donné par (vérifiez-le !) :

$$\left[R_0\hat{\beta} - t_{n-k;1-\frac{\alpha}{2}}\, s.\hat{e}.(R_0\hat{\beta})\,;\; R_0\hat{\beta} + t_{n-k;1-\frac{\alpha}{2}}\, s.\hat{e}.(R_0\hat{\beta})\right]$$

On peut aisément vérifier, en reproduisant un raisonnement semblable à ceux réalisés à plusieurs reprises (faites-le !), que les t-tests et l'intervalle de confiance ci-dessus, qui sont *exacts en échantillon fini* sous l'hypothèse A6 de normalité, restent valables *asymptotiquement*, à titre approximatif pour n grand, sous les seules hypothèses A1 à A5.

En revenant à notre exemple de départ, on notera que le F-test de $\mathrm{H}_0 : \beta_2+\beta_3 = 1$ contre $\mathrm{H}_1 : \beta_2 + \beta_3 \neq 1$ dans la régression :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + e_i$$

peut également être facilement obtenu de la forme (7.5) de la statistique $\hat{F}_0$. Dans ce cas, outre la somme des carrés des résidus de la régression non-contrainte (= SCR), on a à nouveau besoin de la somme des carrés des résidus de la régression contrainte (= SCR_c), qui est ici simplement donnée par la somme des carrés des résidus de la régression (en utilisant $\beta_2 = 1 - \beta_3$) :

$$(y_i - x_{i2}) = \beta_1 + \beta_3(x_{i3} - x_{i2}) + ... + \beta_k x_{ik} + e_i \,,$$

ou, de façon équivalente, de la régression (en utilisant $\beta_3 = 1 - \beta_2$) :

$$(y_i - x_{i3}) = \beta_1 + \beta_2(x_{i2} - x_{i3}) + ... + \beta_k x_{ik} + e_i$$

7.1.3.3. Le F-test de la significativité de la régression dans son ensemble

On considère le test de $\mathrm{H}_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$ contre $\mathrm{H}_1 : \beta_2 \neq 0$ et/ou $\beta_3 \neq 0$ et/ou ... et/ou $\beta_k \neq 0$ dans la régression :

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i$$

càd. un test de la significativité de la régression dans son ensemble.

Sous sa forme générale (7.4), la statistique de test $\hat{F}_0$ de ce test est obtenue en prenant $q = k - 1$,

$$R_0 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \quad \text{et} \quad r_0 = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

où R_0 est une matrice $(k-1) \times k$ et r_0 est un vecteur $(k-1) \times 1$. Pour ces valeurs, on a :

$$R_0\hat{\beta} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

et

$$
\begin{aligned}
& R_0 \hat{V}(\hat{\beta}) R_0' \\
= & \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \hat{Var}(\hat{\beta}_1) & \cdots & \hat{Cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \ddots & \vdots \\ \hat{Cov}(\hat{\beta}_k, \hat{\beta}_1) & \cdots & \hat{Var}(\hat{\beta}_k) \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \\
= & \begin{bmatrix} \hat{Var}(\hat{\beta}_2) & \cdots & \hat{Cov}(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \ddots & \vdots \\ \hat{Cov}(\hat{\beta}_k, \hat{\beta}_2) & \cdots & \hat{Var}(\hat{\beta}_k) \end{bmatrix}
\end{aligned}
$$

de sorte que :

$$
\hat{F}_0 = \frac{1}{k-1} \begin{bmatrix} \hat{\beta}_2 & \cdots & \hat{\beta}_k \end{bmatrix} \begin{bmatrix} \hat{Var}(\hat{\beta}_2) & \cdots & \hat{Cov}(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \ddots & \vdots \\ \hat{Cov}(\hat{\beta}_k, \hat{\beta}_2) & \cdots & \hat{Var}(\hat{\beta}_k) \end{bmatrix}^{-1} \begin{bmatrix} \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}
$$

On peut aisément vérifier (faites-le !) que tout F-test de la nullité jointe (ou de l'égalité à des constantes) d'un sous-ensemble des paramètres du modèle a la même structure — une forme quadratique avec au centre l'inverse de la matrice de variance - covariance des paramètres considérés — que ci-dessus.

Sous la forme (7.5) de la statistique $\hat{F}_0$, ce test est particulièrement simple puisque la somme des carrés des résidus de la régression contrainte ($= \text{SCR}_c$) est ici tout simplement la somme des carrés des résidus de y_i sur une constante, qui est égale à la somme des carrés totaux SCT de la régression initiale, de sorte que :

$$
(\text{SCR}_c - \text{SCR}) = \text{SCT} - \text{SCR} = \text{SCE}\,,
$$

et donc que :

$$
\hat{F}_0 = \frac{(\hat{e}_c'\hat{e}_c - \hat{e}'\hat{e})/(k-1)}{\hat{e}'\hat{e}/(n-k)} = \frac{\text{SCE}/(k-1)}{\text{SCR}/(n-k)}
$$

Le F-test de la significativité de la régression dans son ensemble est reporté en standard par virtuellement tous les logiciels économétriques[80].

7.1.4. Test joint versus tests individuels

Il est important de bien voir les différences entre un test joint de paramètres et des tests individuels de ces mêmes paramètres.

On considère à titre d'exemple le F-test joint de H_0 : $\beta_2 = \beta_3 = 0$ contre H_1 :

[80] Il est reporté par GRETL sous les rubriques 'F(.,.)' (= statistique de test $\hat{F}_0$) et 'P-value(F)' (= P-valeur du test).

$\beta_2 \neq 0$ et/ou $\beta_3 \neq 0$ et les tests individuels (t-test ou F-test : ils sont équivalents) de $\mathrm{H}_0^1 : \beta_2 = 0$ contre $\mathrm{H}_1^1 : \beta_2 \neq 0$ et de $\mathrm{H}_0^2 : \beta_3 = 0$ contre $\mathrm{H}_1^2 : \beta_3 \neq 0$ dans le modèle de régression :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$$

On notera les éléments suivants :

1- Le F-test joint de H_0 et les tests individuels de H_0^1 et H_0^2 ne répondent pas à la même question. Bien qu'ils partagent la même hypothèse alternative :

$$\mathrm{H}_1' = \mathrm{H}_1^{1\prime} = \mathrm{H}_1^{2\prime} : E(y_i|x_{i2}, x_{i3}, x_{i4}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} \,,$$

autrement dit que $E(y_i|.)$ est une fonction linéaire de x_{i2}, x_{i3} et x_{i4}, sans restrictions sur β_1, β_2, β_3 et β_4, ils reviennent à tester des hypothèses nulles différentes :

a- dans le cas du test joint :

$$\mathrm{H}_0' : E(y_i|x_{i2}, x_{i3}, x_{i4}) = \beta_1 + \beta_4 x_{i4} \,,$$

autrement dit que $E(y_i|.)$ est une fonction linéaire de x_{i4}, sans restrictions sur β_1 et β_4, et qui ne dépend ni de x_{i2}, ni de x_{i3}.

b- dans le cas des tests individuels, d'une part :

$$\mathrm{H}_0^{1\prime} : E(y_i|x_{i2}, x_{i3}, x_{i4}) = \beta_1 + \beta_3 x_{i3} + \beta_4 x_{i4} \,,$$

autrement dit que $E(y_i|.)$ est une fonction linéaire de x_{i3} et x_{i4}, sans restrictions sur β_1, β_3 et β_4, et qui ne dépend pas de x_{i2}, et d'autre part :

$$\mathrm{H}_0^{2\prime} : E(y_i|x_{i2}, x_{i3}, x_{i4}) = \beta_1 + \beta_2 x_{i2} + \beta_4 x_{i4} \,,$$

autrement dit que $E(y_i|.)$ est une fonction linéaire de x_{i2} et x_{i4}, sans restrictions sur β_1, β_2 et β_4, et qui ne dépend pas de x_{i3}.

2- Le F-test joint de H_0 ne se réduit pas à une simple addition ou combinaison des F-tests (ou t-tests) individuels de H_0^1 et de H_0^2. En effet, les statistiques $\hat{F}_{01}$ et $\hat{F}_{02}$ des F-tests de H_0^1 et de H_0^2 sont données par (cf. Section 7.1.3.1) :

$$\hat{F}_{01} = \frac{\hat{\beta}_2^2}{\hat{Var}(\hat{\beta}_2)} \quad \text{et} \quad \hat{F}_{02} = \frac{\hat{\beta}_3^2}{\hat{Var}(\hat{\beta}_3)} \,,$$

tandis que la statistiques $\hat{F}_0$ du F-test de H_0 est donnée par (cf. Section 7.1.3.3) :

$$\hat{F}_0 = \frac{1}{2} \begin{bmatrix} \hat{\beta}_2 & \hat{\beta}_3 \end{bmatrix} \begin{bmatrix} \hat{Var}(\hat{\beta}_2) & \hat{Cov}(\hat{\beta}_2, \hat{\beta}_3) \\ \hat{Cov}(\hat{\beta}_3, \hat{\beta}_2) & \hat{Var}(\hat{\beta}_3) \end{bmatrix}^{-1} \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

On remarquera en particulier le fait que $\hat{F}_0$ fait intervenir la covariance entre $\hat{\beta}_2$ et $\hat{\beta}_3$, ce qui n'est le cas ni de $\hat{F}_{01}$, ni de $\hat{F}_{02}$.

3- On peut se faire un idée plus précise des différences entre le F-test joint de H_0 et les (F- ou t-) tests individuels de H_0^1 et de H_0^2 en examinant graphiquement leur

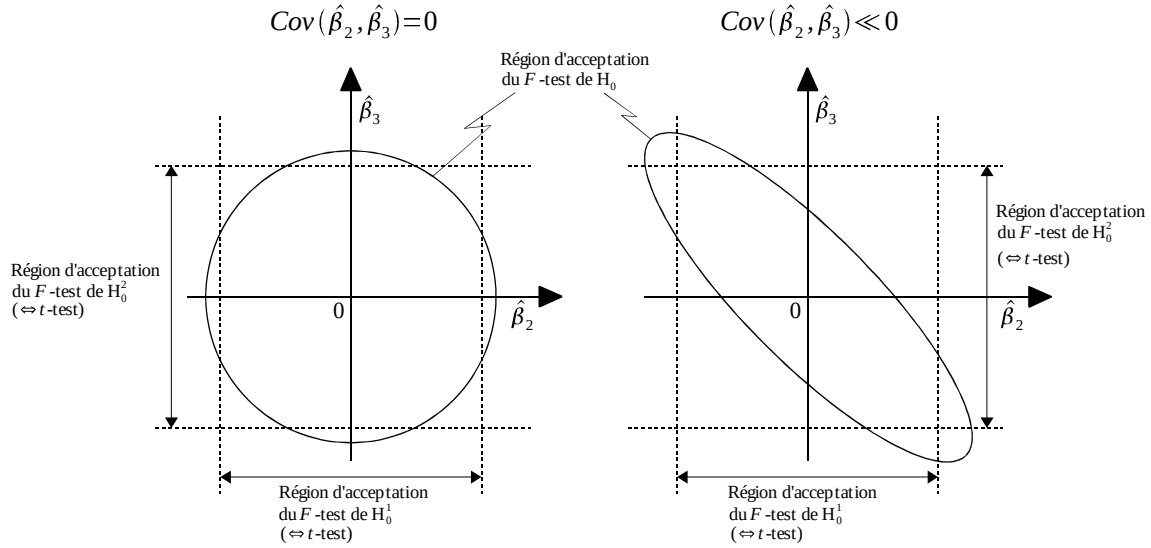
région respective d'acceptation (et donc de rejet). En grand échantillon (pour n grand), les régions d'acceptation (i.e., de non-rejet) des F-tests individuels au seuil α sont données par les ensembles de valeurs de $\hat{\beta}_j$ $(j = 2, 3)$ qui sont telles que[81] :

$$\hat{F}'_{0j} = \frac{\hat{\beta}_j^2}{Var(\hat{\beta}_j)} \simeq \hat{F}_{0j} \leq F_{1,n-k;1-\alpha} \,,$$

où $F_{1,n-k;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher $F(1, n-k)$, et la région d'acceptation du F-test joint au seuil α est donnée par l'ensemble des couples de valeurs $(\hat{\beta}_2, \hat{\beta}_3)$ qui sont telles que :

$$\hat{F}'_0 = \frac{1}{2} \left[\begin{array}{cc} \hat{\beta}_2 & \hat{\beta}_3 \end{array} \right] \left[\begin{array}{cc} Var(\hat{\beta}_2) & Cov(\hat{\beta}_2, \hat{\beta}_3) \\ Cov(\hat{\beta}_3, \hat{\beta}_2) & Var(\hat{\beta}_3) \end{array} \right]^{-1} \left[\begin{array}{c} \hat{\beta}_2 \\ \hat{\beta}_3 \end{array} \right] \simeq \hat{F}_0 \leq F_{2,n-k;1-\alpha} \,,$$

où $F_{2,n-k;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher $F(2, n-k)$. Graphiquement :



Graphique 45 : Région d'acceptation du test joint et des tests individuels

On voit en particulier que la région d'acceptation du F-test joint dépend fortement de la covariance entre $\hat{\beta}_2$ et $\hat{\beta}_3$. Lorsque cette covariance est égale à zéro (et que les variances de $\hat{\beta}_2$ et de $\hat{\beta}_3$ sont égales), la région d'acceptation du F-test joint est un cercle, qui est proche du carré formé par l'intersection des régions d'acceptation des tests individuels. Lorsque cette covariance est différente de zéro (positive ou négative), la région d'acceptation du F-test joint est une ellipse, inclinée vers le bas (si $Cov(\hat{\beta}_2, \hat{\beta}_3) < 0$) ou vers le haut (si $Cov(\hat{\beta}_2, \hat{\beta}_3) > 0$), et dont la surface est (fortement si $|Cov(\hat{\beta}_2, \hat{\beta}_3)|$ est très différente de zéro) plus petite que la surface du carré formé par l'intersection des régions d'acceptation des tests individuels.

[81] Ci-dessous, pour simplifier, on fait comme si les variances et covariances des paramètres étaient connues. Pour n grand, cela n'a en fait aucune importance (le comportement asymptotique des statistiques de tests est le même que les variances et covariances soient connues ou doivent être estimées).

4- On pourrait être tenté de tester l'hypothèse nulle jointe H_0 en s'appuyant sur les tests individuels de H_0^1 et de H_0^2. Plutôt que d'utiliser le F-test joint, on pourrait ainsi décider d'accepter H_0 si H_0^1 et H_0^2 sont *toutes les deux* acceptées, et de rejeter H_0 si *une au moins* des hypothèses nulles H_0^1 et H_0^2 est rejetée. Graphiquement, cela reviendrait à accepter H_0 lorsque $(\hat{\beta}_2, \hat{\beta}_3)$ appartient au carré formé par l'intersection des régions d'acceptation des tests individuels, et à rejeter H_0 sinon. Tester l'hypothèse nulle jointe H_0 sur base d'un tel *test induit* (plutôt que sur base du F-test joint) n'est pas une très bonne idée. Pour deux raisons :

a- Contrairement à ce qu'on pourrait croire, le *risque de première espèce* α_I d'un tel test induit n'est pas égal au seuil α des tests individuels qui le composent, mais compris entre α et 2α. On a en effet[82] :

$$\begin{aligned}
\alpha_I &= I\!P(\mathrm{RH}_0 \,|\, \mathrm{H}_0 \text{ est vraie}) = I\!P(\mathrm{RH}_0^1 \text{ ou } \mathrm{RH}_0^2 \,|\, \mathrm{H}_0 \text{ est vraie}) \\
&= I\!P(\mathrm{RH}_0^1 \,|\, \mathrm{H}_0 \text{ est vraie}) + I\!P(\mathrm{RH}_0^2 \,|\, \mathrm{H}_0 \text{ est vraie}) \\
&\quad - I\!P(\mathrm{RH}_0^1 \text{ et } \mathrm{RH}_0^2 \,|\, \mathrm{H}_0 \text{ est vraie}) \\
&= \alpha + \alpha - I\!P(\mathrm{RH}_0^1 \text{ et } \mathrm{RH}_0^2 \,|\, \mathrm{H}_0 \text{ est vraie}) \\
&= 2\alpha - I\!P(\mathrm{RH}_0^1 \text{ et } \mathrm{RH}_0^2 \,|\, \mathrm{H}_0 \text{ est vraie})
\end{aligned}$$

où $I\!P(\mathrm{RH}_0^1 \text{ et } \mathrm{RH}_0^2 \,|\, \mathrm{H}_0 \text{ est vraie})$ est toujours compris entre 0 et α, et est égal à α^2 si RH_0^1 et RH_0^2 sont indépendants (i.e., pour n grand, si $Cov(\hat{\beta}_2, \hat{\beta}_3) = 0$).
Ainsi, si chacun des tests individuels est effectué au seuil de 5%, le test induit aura un risque de première espèce α_I compris entre 5% et 10%. Pour avoir un test induit dont le risque de première espèce α_I est au maximum de 5% (et donc comparable au F-test joint au seuil de 5%), chacun des tests individuels devrait être effectué au seuil de $2,5\%$.

b- A risque de première espèce comparable, càd. si les seuils de tests individuels sont ajustés comme suggéré ci-dessus de façon à ce que α_I soit au maximum égal au seuil α du F-test joint, la *puissance* d'un tel test induit est quasi-toujours *inférieure* à celle du F-test joint, et cela est d'autant plus vrai que la valeur absolue de la covariance entre $\hat{\beta}_2$ et $\hat{\beta}_3$ est élevée. Graphiquement, la plus grande puissance du F-test est liée au fait que la surface de la région d'acceptation du F-test joint est (fortement si $|Cov(\hat{\beta}_2, \hat{\beta}_3)|$ est très différente de zéro) plus petite que la surface du carré formé par l'intersection des régions d'acceptation des tests individuels. Le F-test étant généralement plus puissant, il est évidemment préférable au test induit.

5- Etant donné les éléments développés ci-dessus, on devine que le F-test joint et les tests individuels peuvent aboutir à des conclusions apparemment contradictoires. En pratique, il est très rare que le F-test joint ne rejette pas l'hypothèse nulle jointe H_0 alors que l'une au moins des hypothèses nulles H_0^1 et de H_0^2 est rejetée

[82] Pour rappel, de la théorie du calcul des probabilités, pour A et B désignant deux événements quelconques, on a : (1) $I\!P(A \text{ ou } B) = I\!P(A) + I\!P(B) - I\!P(A \text{ et } B)$, (2) $I\!P(A \text{ et } B) \leq I\!P(A)$ et $I\!P(A \text{ et } B) \leq I\!P(B)$, et (3) si A, B sont deux événements indépendants, alors $I\!P(A \text{ et } B) = I\!P(A)I\!P(B)$. Les mêmes règles de calcul tiennent pour les probabilités conditionnelles.

sur base des tests individuels[83]. Par contre, il est relativement fréquent (chaque fois que $Var(\hat{\beta}_2)$, $Var(\hat{\beta}_3)$ *et* $|Cov(\hat{\beta}_2, \hat{\beta}_3)|$ sont élevées, voir la Section 7.3 ci-dessous pour un exemple typique) que le F-test joint rejette l'hypothèse nulle jointe H_0 alors que ni l'hypothèse nulle H_0^1 ni l'hypothèse nulle H_0^2 n'est rejetée sur base des tests individuels[84].

Pour conclure, on notera que l'ensemble des considérations développées ci-dessus vaut pour tout F-test joint d'une hypothèse nulle multiple (i.e., testant conjointement plusieurs restrictions) comparé aux tests individuels des hypothèses nulles qui le composent (i.e., les différentes restrictions prises séparément).

7.2. Exemple : les ventes d'une chaîne de fast-food de HGL (2008)

Hill, Griffiths et Lim (2008) considèrent[85] une extension de leur modèle visant à évaluer l'effet de la politique de prix et de publicité sur les ventes d'une chaîne de fast-food. Le modèle considéré est :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i3}^2 + e_i \tag{7.8}$$

où y_i désigne les recettes mensuelles de vente (en milliers de \$), x_{i2} le prix de vente unitaire (en \$), et x_{i3} le montant des dépenses publicitaires mensuelles (en milliers de \$).

La conjecture sous-jacente à cet extension est que l'effet des dépenses publicitaires sur les recettes n'est probablement pas linéaire (i.e., l'effet marginal de x_{i3} n'est probablement pas constant), comme le suppose le modèle initial (cf. Section 6.6).

Dans le modèle étendu (7.8), on s'attend à avoir $\beta_3 > 0$ et $\beta_4 < 0$, de sorte que l'effet marginal des dépenses publicitaires sur les recettes, mesuré par $\frac{\partial E(y_i)}{\partial x_{i3}} = \beta_3 + 2\beta_4 x_{i3}$, soit une fonction décroissante des dépenses publicitaires.

En utilisant le logiciel GRETL, on obtient :

---

[83] Cela se produit lorsque le couple $(\hat{\beta}_2, \hat{\beta}_3)$ obtenu pour un échantillon particulier appartient à la région d'acceptation de test joint (cercle ou ellipse, cf. le Graphique 45), mais n'appartient pas au carré formé par l'intersection des régions d'acceptation des tests individuels.

[84] Cela se produit lorsque le couple $(\hat{\beta}_2, \hat{\beta}_3)$ obtenu pour un échantillon particulier n'appartient pas à la région d'acceptation de test joint (cercle ou ellipse, cf. le Graphique 45), mais appartient au carré formé par l'intersection des régions d'acceptation des tests individuels.

[85] Voir p. 140 et suivantes.

Model 2:
OLS, using observations 1-75
Dependent variable: y

	coefficient	std. error	t-ratio	p-value
const	109.719	6.79905	16.14	1.87e-025 ***
x2	-7.64000	1.04594	-7.304	3.24e-010 ***
x3	12.1512	3.55616	3.417	0.0011 ***
x3_2	-2.76796	0.940624	-2.943	0.0044 ***

Mean dependent var	77.37467	S.D. dependent var	6.488537
Sum squared resid	1532.084	S.E. of regression	4.645283
R-squared	0.508235	Adjusted R-squared	0.487456
F(3, 71)	24.45932	P-value(F)	5.60e-11
Log-likelihood	-219.5540	Akaike criterion	447.1080
Schwarz criterion	456.3780	Hannan-Quinn	450.8094

et

Null hypothesis: the regression parameters are zero for the variables
x3, x3_2

Asymptotic test statistic:
Wald chi-square(2) = 16.8827, with p-value = 0.000215757
F-form: F(2,71) = 8.44136, with p-value = 0.000514159

Sur base des résultats reportés ci-dessus, on peut :

1- voir que la statistique de test $\hat{F}_0$ du F-test de $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ contre $H_1 : \beta_2 \neq 0$ et/ou $\beta_3 \neq 0$ et/ou $\beta_4 \neq 0$ est égale à $24,45932$, et que H_0 peut être rejetée au *seuil minimum* de 5,60e-11 (= P-valeur du test). On peut donc rejeter l'hypothèse nulle que $E(y_i|x_{i2}, x_{i3}) = \beta_1$, càd. que $E(y_i|x_{i2}, x_{i3})$ ne dépend pas ni de x_{i2}, ni de x_{i3}. Autrement dit, il apparaît que la régression est fortement significative dans son ensemble.

2- voir que la statistique de test $\hat{F}_0$ du F-test de $H_0 : \beta_3 = \beta_4 = 0$ contre $H_1 : \beta_3 \neq 0$ et/ou $\beta_4 \neq 0$ est égale à $8,44136$, et que H_0 peut être rejetée au *seuil minimum* de 0.000514159 (= P-valeur du test). On peut donc rejeter l'hypothèse nulle que $E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_2 x_{i2}$, càd. que $E(y_i|x_{i2}, x_{i3})$ ne dépend pas de x_{i3}. Autrement dit, il apparaît que les dépenses publicitaires ont un effet fortement significatif sur les recettes (à prix de vente constant).

3- voir que la statistique de test $\hat{t}_o$ du t-test de $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$ est égale à $-7,304$, et que H_0 peut être rejetée au *seuil minimum* de 3,24e-010 (= P-valeur du test). On peut donc rejeter l'hypothèse nulle que $E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_3 x_{i3} + \beta_4 x_{i3}^2$, càd. que $E(y_i|x_{i2}, x_{i3})$ ne dépend pas de x_{i2}. Autrement dit, il apparaît que le prix de vente a un effet fortement significatif sur les recettes (à dépenses publicitaires constantes).

4- effectuer un test de $H_0 : \beta_3 \leq 0$ contre $H_1 : \beta_3 > 0$. On a $\hat{t}_o = 3,417$, et la P-valeur du test est égale à $\frac{0,0011}{2} = 0,00055$, de sorte H_0 peut être rejetée au *seuil minimum* de $0,00055$. Il apparaît donc que $\hat{\beta}_3$ est, comme attendu, statistiquement significativement supérieur à 0.

5- effectuer un test de $H_0 : \beta_4 \geq 0$ contre $H_1 : \beta_4 < 0$. On a $\hat{t}_o$ = -2,943, et la P-valeur du test est égale à $\frac{0,0044}{2} = 0,0022$, de sorte H_0 peut être rejetée au *seuil minimum* de $0,0022$. Il apparaît donc que $\hat{\beta}_4$ est, comme attendu, statistiquement significativement inférieur à 0.

Le bénéfice, en termes de recettes, d'une unité supplémentaire de dépenses publicitaires est donné par :

$$\frac{\partial E(y_i)}{\partial x_{i3}} = \beta_3 + 2\beta_4 x_{i3}$$

Le coût additionnel de cette unité supplémentaire de dépenses publicitaires est le coût de la publicité elle-même, plus le coût de production des unités supplémentaires qui seront vendues grâce à la publicité. Si on néglige ce second aspect, le montant optimal x^*_{i3} de dépenses publicitaires doit satisfaire (recette marginale = coût marginal) :

$$\beta_3 + 2\beta_4 x^*_{i3} = 1$$

Une estimation du montant optimal de dépenses publicitaires est peut être obtenue en remplaçant β_3 et β_4 par leur estimation :

$$\hat{x}^*_{i3} = \frac{1 - \hat{\beta}_3}{2\hat{\beta}_4} = \frac{1 - 12,1512}{2(-2,76796)} = 2,014$$

Le montant optimal de dépenses publicitaires est donc estimé à 2014 $ (attention aux unités de mesure !).

Sur base de son expérience, un manager de la chaîne de fast-food pense que le niveau optimal des dépenses publicitaires est de 1900 $, et que pour un prix de vente conjointement fixé à 6 $, on devrait obtenir en moyenne une recette de 80000 $. On peut tester la compatibilité de ces conjectures avec les données en testant :

$$H_0 : \beta_3 + 2\beta_4(1,9) = 1 \quad \text{et} \quad \beta_1 + \beta_2(6) + \beta_3(1,9) + \beta_4(1,9)^2 = 80$$
contre H_1 : au moins 1 des deux restrictions est fausse

En utilisant le logiciel GRETL, on obtient :

```
Restriction set
  1: b[x3] + 3.8*b[x3_2] = 1
  2: b[const] + 6*b[x2] + 1.9*b[x3] + 3.61*b[x3_2] = 80

Test statistic: F(2,71) = 5.74123, with p-value = 0.00488466
```

On voit que la statistique de test $\hat{F}_0$ du F-test joint est égale à $5,74123$, et que

H_0 peut être rejetée au *seuil minimum* de 0,00488466 (= P-valeur du test). On peut donc rejeter l'affirmation du manager, càd. l'hypothèse nulle jointe que le niveau optimal des dépenses publicitaires est de 1900 \$, et que pour un prix de vente conjointement fixé à 6 \$, on devrait obtenir en moyenne une recette de 80000 \$.

Le rejet de l'hypothèse nulle jointe est-elle due à la conjecture relative au niveau optimal des dépenses publicitaires, à la conjecture relative à la recette moyenne obtenue pour un prix de vente de 6 \$ et des dépenses publicitaires de 1900 \$, ou aux deux conjectures ? On peut tenter de répondre à cette question en testant séparément, d'une part :

$$H_0 : \beta_3 + 2\beta_4(1,9) = 1 \text{ contre } H_1 : \beta_3 + 2\beta_4(1,9) \neq 1 ,$$

et d'autre part :

$$\begin{aligned} H_0 \quad &: \quad \beta_1 + \beta_2(6) + \beta_3(1,9) + \beta_4(1,9)^2 = 80 \\ \text{contre } H_1 \quad &: \quad \beta_1 + \beta_2(6) + \beta_3(1,9) + \beta_4(1,9)^2 \neq 80 \end{aligned}$$

En utilisant encore le logiciel GRETL, on obtient :

```
Restriction
  b[x3] + 3.8*b[x3_2] = 1

Test statistic: F(1,71) = 0.936195, with p-value = 0.336543
```

et

```
Restriction
  b[const] + 6*b[x2] + 1.9*b[x3] + 3.61*b[x3_2] = 80

Test statistic: F(1,71) = 10,8721, with p-value = 0,00152693
```

On constate que si la conjecture relative au niveau optimal des dépenses publicitaires ne peut pas (à moins de prendre un risque de première espèce de 33% ou plus) être rejetée, la conjecture relative à la recette moyenne obtenue pour un prix de vente de 6 \$ et des dépenses publicitaires de 1900 \$ apparaît elle très fortement rejetée par les données.

7.3. Colinéarité

On dit qu'il y a *colinéarité parfaite* entre les régresseurs d'une régression multiple lorsque une (ou plusieurs) des variables explicatives de la régression est une *combinaison linéaire exacte* des (ou d'un sous-ensemble des) autres variables explicatives du modèle. Dans cette situation, l'hypothèse A5 $\text{rg}(X) = k$ n'est pas satisfaite (i.e., $\text{rg}(X) \neq k$), de sorte que la matrice $(X'X)$ est singulière et donc non-inversible, et l'estimateur MCO $\hat{\beta} = (X'X)^{-1} X'Y$ n'est tout simplement pas défini (cf. Section 6.2).

On parle (d'un problème) de colinéarité lorsqu'on est proche d'une situation de

colinéarité parfaite.

Par exemple, si on cherche à expliquer la consommation d'un ménage ($= y_i$) en fonction de son revenu ($= x_{i2}$) et de sa fortune mobilière ($= x_{i3}$) au travers du modèle :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \,,$$

il y a de fortes chance que, dans un échantillon de données en coupe, x_{i2} et x_{i3} varient systématiquement ensemble, autrement dit que la corrélation empirique $\rho_e(x_{i2}, x_{i3})$ entre x_{i2} et x_{i3} soit très élevée (de l'ordre 0,95 ou plus), ce qui traduit le fait que x_{i2} et x_{i3} sont proches de satisfaire une relation linéaire exacte.

Lorsque deux ou plusieurs variables explicatives d'une régression multiple sont (fortement) colinéaires, on observera typiquement en pratique que :

1- les variances et covariances (en valeurs absolues) de leurs paramètres sont (très) élevées. Ainsi, malgré un R^2 éventuellement élevé, les coefficients estimés de ces variables apparaîtront *individuellement* peu ou pas significatifs. Par contre, ils pourraient très bien apparaître *conjointement* (au travers d'un F-test joint de leur nullité) significatifs. De même, certaines fonctions de ces paramètres (en particulier, $X_0\beta = E(y_0)$ pour des valeurs de X_0 proches des valeurs observées dans l'échantillon) pourraient très bien être estimées de façon (très) précise.
Ainsi, pour l'exemple donné ci-dessus, conformément aux expressions (6.1) - (6.3) de la Section 6.3.1, si $\rho_e(x_{i2}, x_{i3})$ est proche de 1, les variances $Var(\hat{\beta}_2)$ et $Var(\hat{\beta}_3)$ seront (très) élevées, et la covariance $Cov(\hat{\beta}_2, \hat{\beta}_3)$ sera négative et également (très) élevée en valeur absolue. Les paramètres estimés $\hat{\beta}_2$ et $\hat{\beta}_3$ ont donc toutes les chances d'apparaître individuellement peu ou pas significatifs, ce qui traduit simplement le fait qu'étant donné que x_{i2} et x_{i3} varient systématiquement ensemble, il est difficile de séparer leur effet marginal propre — d'où des estimations peu précises — sur y_i. Par contre, il est fort probable qu'on pourra estimer de façon précise leur effet marginal conjoint $\beta_2 + \beta_3 = \frac{\partial E(y_i)}{\partial x_{i2}} + \frac{\partial E(y_i)}{\partial x_{i2}}$. En effet, on a :

$$Var(\hat{\beta}_2 + \hat{\beta}_3) = Var(\hat{\beta}_2) + Var(\hat{\beta}_2) + 2Cov(\hat{\beta}_2, \hat{\beta}_3) \,,$$

de sorte que la covariance, fortement négative, peut très bien compenser les variances élevées. Pour le même type de raison, $\hat{\beta}_2$ et $\hat{\beta}_3$ apparaîtront très certainement conjointement (au travers du F-test de la significativité de la régression dans son ensemble) significatifs, et pour des valeurs de X_0 proches des valeurs observées dans l'échantillon, les prévisions $\hat{y}_0 = X_0\hat{\beta}$ seront probablement (en tout cas en tant qu'estimateur/prédicteur de $E(y_0)$) assez précises.

2- les résultats d'estimation sont (très) sensibles à la suppression de quelques observations et/ou d'une variable apparemment non pertinente (car non significative). Cela découle simplement du fait que les paramètres sont estimés de façon peu précise, et sont donc très variables d'un échantillon à l'autre.

On notera encore les points suivants :

1- Ce qui caractérise fondamentalement les effets de la colinéarité, ce n'est pas que les variances des paramètres sont élevées, ou que les résultats d'estimation sont peu robustes à la suppression de quelques observations : un échantillon de (très) petite taille ou une (très) faible dispersion des variables explicatives produit les mêmes effets. Ce qui caractérise fondamentalement les effets de la colinéarité, c'est que les variances des paramètres sont élevées, et que simultanément les valeurs absolues des covariances des paramètres sont également élevées. C'est cela qui rend possible le fait que des paramètres peuvent apparaître individuellement peu ou pas significatifs, tout en étant conjointement très significatifs.

2- Une colinéarité importante au sein des variables explicatives n'implique pas nécessairement des variances élevées pour les paramètres, et donc des estimations individuelles peu précises de ces paramètres. De ce point de vue, une forte colinéarité (dans l'exemple ci-dessus, $\rho_e(x_{i2}, x_{i3})$ est proche de 1) peut très bien être compensée par la grande taille de l'échantillon et/ou une forte dispersion des variables explicatives.

3- Les données dont on dispose étant presque toujours de nature non-expérimentale, il existe quasi toujours un certain degré de colinéarité au sein des variables explicatives d'une régression. Cette colinéarité n'est pas en soi un problème. Elle ne devient un problème que si elle est (le principal) responsable d'une forte imprécision des paramètres estimés, qui rendent les résultats obtenus peu exploitables.

4- Pour identifier les relations éventuelles de colinéarité au sein d'un ensemble de variables explicatives, on peut :

   a- examiner les corrélations entre les différentes variables explicatives prises deux à deux.

   b- si les relations à la base de la colinéarité semblent plus complexes (i.e., impliquent simultanément plusieurs variables), examiner les régressions — en particulier leur R^2 — de chacune des variables explicatives sur les autres variables explicatives.

5- Lorsqu'on est confronté à un problème de colinéarité, on peut :

   a- chercher à obtenir de 'meilleures' données : moins colinéaires, plus nombreuses, plus dispersées. En pratique, ce n'est le plus souvent pas possible.

   b- atténuer le problème en imposant des contraintes sur les paramètres. On peut en effet montrer qu'une estimation contrainte (moindres carrés sous contrainte) des paramètres d'un modèle de régression améliore toujours la précision d'estimation des paramètres. Malheureusement, imposer des contraintes sur les paramètres crée des biais si les contraintes sont incorrectes. Cette solution n'est donc envisageable que dans les situations où la théorie suggère des restrictions a priori pertinentes sur les paramètres[86], ce qui est loin d'être en pratique souvent le cas.

[86] Pour un exemple de ce type, voir la Section 6.5 de Hill, Griffiths et Lim (2008). Nous n'en dirons pas plus dans le cadre de ce cours.

c- simplement reconnaître, comme l'indique la forte imprécision d'estimation des paramètres, que l'information contenue dans l'échantillon dont on dispose est trop faible pour pouvoir obtenir des estimations précises, ou autrement dit, que le modèle est trop complexe en regard de l'information contenue dans l'échantillon.

7.4. Problèmes de spécification

7.4.1. Forme fonctionnelle

Supposons que l'on cherche à expliquer le salaire y_i d'un individu en fonction de son niveau d'éducation x_{i2} et de son niveau d'expérience professionnelle x_{i3}. On le sait, le modèle de régression définit comme contrepartie empirique de la relation théorique d'intérêt l'espérance conditionnelle de y_i sachant (x_{i2}, x_{i3}). De façon générale, on a :

$$E(y_i|x_{i2}, x_{i3}) = g(x_{i2}, x_{i3}) \qquad \text{(i.e., une fonction de } x_{i2} \text{ et } x_{i3})$$

Le modèle de régression linéaire standard suppose, au travers des hypothèses[87] A1 et A2, que la fonction $g(x_{i2}, x_{i3})$ est linéaire :

$$E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\,,$$

de sorte que la relation théorique d'intérêt peut être estimée sur base du modèle de régression standard :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \tag{7.9}$$

En pratique, rien n'assure que cette hypothèse de forme fonctionnelle est bien correcte. Elle peut cependant aisément être testée.

Pour tester que la forme fonctionnelle du modèle de régression standard (7.9) est bien correcte, il suffit de considérer le modèle étendu :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i3}^2 + \beta_6 (x_{i2} x_{i3}) + e_i\,,$$

càd. le modèle original (7.9) auquel sont ajoutés les carrés et le produit croisé des variables (x_{i2}, x_{i3}), et de tester à l'aide d'un F-test H_0 : $\beta_4 = \beta_5 = \beta_6 = 0$ contre H_1 : $\beta_4 \neq 0$ et/ou $\beta_5 \neq 0$ et/ou $\beta_6 \neq 0$.

De façon générale, pour tester la forme fonctionnelle du modèle de régression linéaire multiple :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + e_i\,, \tag{7.10}$$

[87] auxquelles il convient d'ajouter l'hypothèse A5 que X est non-stochastique pour pouvoir écrire de façon simplifiée $E(y_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$.

il suffit de considérer un modèle étendu comprenant, outre le modèle original (7.10), tous les carrés et produits croisés des variables $(x_{i2}, x_{i3}, ..., x_{ik})$, et de tester à l'aide d'un F-test l'hypothèse nulle que les paramètres associés aux variables ajoutées — les carrés et produits croisés des variables $(x_{i2}, x_{i3}, ..., x_{ik})$ — sont égaux à zéro.

Notons que dans le modèle original (7.10), rien n'empêche les variables $(x_{i2}, x_{i3}, ..., x_{ik})$ d'être des transformations d'autres variables (comme dans un modèle lin-log ou log-log), ou encore de correspondre aux différentes variables d'une régression elle-même polynomiale. On peut donc de la sorte tester la forme fonctionnelle de n'importe quelle forme de régression linéaire multiple (standard, avec des variables totalement ou partiellement transformées, polynomiale, etc...).

Cette façon de tester la forme fonctionnelle du modèle devient vite peu pratique lorsque le modèle original contient beaucoup de variables explicatives : le modèle étendu contient alors un très grand nombre de variables explicatives.

Une approche alternative, plus parcimonieuse, est donnée par le test RESET de Ramsey (1969)[88]. Plutôt que d'ajouter les carrés et les produits croisés des variables du modèle original, le test RESET suggère de considérer un modèle étendu obtenu en ajoutant au modèle original les puissances $\hat{y}_i^2$, $\hat{y}_i^3$, ... de la valeur prédite $\hat{y}_i = X_i\hat{\beta}$ par le modèle original, et comme précédemment, de tester à l'aide d'un F-test l'hypothèse nulle que les paramètres associés aux variables ajoutées — ici, les puissances $\hat{y}_i^2$, $\hat{y}_i^3$, ... des valeurs prédites $\hat{y}_i = X_i\hat{\beta}$ — sont égaux à zéro. Pour tester la forme fonctionnelle du modèle de régression standard (7.9), en utilisant les puissances de $\hat{y}_i$ jusqu'à l'ordre 3, cela signifie considérer le modèle étendu :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 \hat{y}_i^2 + \beta_5 \hat{y}_i^3 + e_i \,,$$

où $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$, et tester à l'aide d'un F-test H_0 : $\beta_4 = \beta_5 = 0$ contre H_1 : $\beta_4 \neq 0$ et/ou $\beta_5 \neq 0$. En pratique, on utilise rarement les puissances de $\hat{y}_i$ au delà de l'ordre 4. On notera que la valeur prédite $\hat{y}_i = X_i\hat{\beta}$ elle-même (i.e., la puissance d'ordre 1 de $\hat{y}_i$) ne peut pas être inclue dans le modèle étendu car elle est par définition parfaitement colinéaire avec les variables du modèle original ($\hat{y}_i$ est une combinaison linéaire exacte de ces variables). L'idée à la base de cette façon de procéder est que si la forme fonctionnelle du modèle original est incorrecte, les variables $\hat{y}_i^2$, $\hat{y}_i^3$, ... — qui, si on les développe, apparaissent comme des fonctions polynomiales des variables du modèle original[89] — devraient généralement améliorer l'ajustement du modèle, et donc apparaître statistiquement significatives.

Deux points méritent encore d'être soulignés :

1- Toute fonction des variables du modèle original peut être utilisée dans le modèle étendu qui sert à tester la forme fonctionnelle du modèle original. Ainsi, on pourrait très bien tester la forme fonctionnelle du modèle original (7.9) en prenant comme variables additionnelles dans le modèle étendu $ln(x_{i2})$, $ln(x_{i3})$,

[88] Ramsey J.B. (1969), "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis", *Journal of the Royal Statistical Society*, Serie B, 31, p. 350-371. Notons que 'RESET' est l'acronyme de 'Regression Specification Error Test'.

[89] Pour l'exemple du test RESET du modèle de régression standard (7.9), on a ainsi : $\hat{y}_i^2 = (\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3})^2 = \hat{\beta}_1^2 + 2\hat{\beta}_1\hat{\beta}_2 x_{i2} + 2\hat{\beta}_1\hat{\beta}_3 x_{i3} + \hat{\beta}_2^2 x_{i2}^2 + 2\hat{\beta}_2\hat{\beta}_3 x_{i2} x_{i3} + \hat{\beta}_3^2 x_{i3}^2$.

ainsi qu'éventuellement les carrés et le produit croisé de ces variables. On ne peut par contre pas inclure dans le modèle étendu des variables (ou des fonctions de ces variables) qui n'étaient pas présentes dans le modèle original. Ainsi par exemple, pour tester la forme fonctionnelle de modèle original (7.9), on ne peut pas considérer comme modèle étendu :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i \,,$$

où x_{i4} est une variable qui n'était pas présente dans le modèle original, et tester à l'aide d'un F-test (ou d'un t-test) $\mathrm{H}_0 : \ \beta_4 = 0$ contre $\mathrm{H}_1 : \ \beta_4 \neq 0$. En effet, dans ce cas, on ne teste plus :

$$\mathrm{H}_0' : \ E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} \,,$$

autrement dit que $E(y_i|x_{i2}, x_{i3})$ est une fonction linéaire de x_{i2} et x_{i3}, mais plutôt :

$$\mathrm{H}_0'' : \ E(y_i|x_{i2}, x_{i3}, x_{i4}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} \,,$$

autrement dit que $E(y_i|x_{i2}, x_{i3}, x_{i4})$ est une fonction linéaire de x_{i2} et x_{i3}, et qu'elle ne dépend pas de x_{i4}. Or, on peut très bien simultanément avoir que H_0' est vraie et que H_0'' est fausse (à cause du changement de l'ensemble des variables conditionnantes).

2- Lorsqu'on rejette l'hypothèse nulle que la forme fonctionnelle du modèle original est correcte, pour identifier plus précisément la forme de la mauvaise spécification du modèle, il peut être utile d'examiner des graphiques des résidus $\hat{e}_i$ du modèle original en fonction de ses différentes variables explicatives, ou encore en fonction de la valeur prédite $\hat{y}_i = X_i\hat{\beta}$ (du modèle original).

7.4.2. Variables omises

Supposons à nouveau que l'on cherche à expliquer le salaire y_i d'un individu en fonction de son niveau d'éducation x_{i2} et de son niveau d'expérience professionnelle x_{i3}. Supposons par ailleurs que toutes les hypothèses (en particulier les hypothèses A1 et A2 concernant la forme fonctionnelle) du modèle de régression linéaire standard soient correctes, de sorte que la relation théorique d'intérêt peut être estimée sur base du modèle de régression standard :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \tag{7.11}$$

Supposons encore que l'on soit en particulier intéressé par le paramètre β_2 qui indique l'effet d'une année d'étude supplémentaire sur le salaire moyen d'un individu, à niveau d'expérience professionnelle constant (= l'effet marginal de x_{i2}, x_{i3} étant maintenu constant). Supposons finalement que l'on estime le paramètre β_2, non pas

à l'aide de l'estimateur MCO $\hat{\beta}$ du modèle (7.11) :

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = (X'X)^{-1} X'Y\,, \quad \text{où } Y = \begin{bmatrix} \vdots \\ y_i \\ \vdots \end{bmatrix} \text{ et } X = \begin{bmatrix} \vdots & \vdots & \vdots \\ 1 & x_{i2} & x_{i3} \\ \vdots & \vdots & \vdots \end{bmatrix}, \quad (7.12)$$

mais à l'aide de l'estimateur MCO $\hat{\beta}^*_{12}$ du modèle de régression :

$$y_i = \beta_1 + \beta_2 x_{i2} + e_i\,, \quad (7.13)$$

où, pour une raison quelconque[90], la variable x_{i3} est omise. L'estimateur MCO du modèle (7.13), où la variable x_{i3} est omise, est donné par :

$$\hat{\beta}^*_{12} = \begin{bmatrix} \hat{\beta}^*_1 \\ \hat{\beta}^*_2 \end{bmatrix} = (X'_{12}X_{12})^{-1} X'_{12}Y\,, \quad \text{où } Y = \begin{bmatrix} \vdots \\ y_i \\ \vdots \end{bmatrix} \text{ et } X_{12} = \begin{bmatrix} \vdots & \vdots \\ 1 & x_{i2} \\ \vdots & \vdots \end{bmatrix}$$

On sait que, le modèle (7.11) étant correctement spécifié (en particulier les hypothèses A1, A2 et A5 étant correctes), l'estimateur MCO $\hat{\beta}_2$ est un estimateur non biaisé du paramètre d'intérêt β_2. Peut-on en dire autant de l'estimateur MCO $\hat{\beta}^*_2$ du modèle (7.13), où la variable x_{i3} a été omise ? La réponse à cette question est non, sauf dans deux cas particuliers.

Pour le voir, il suffit de calculer l'espérance de $\hat{\beta}^*_{12}$. Notons tout d'abord que, sous forme matricielle, on peut écrire le modèle (7.11) :

$$Y = X\beta + e = X_{12}\beta_{12} + X_3\beta_3 + e\,, \quad \text{où } \beta_{12} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \text{ et } X_3 = \begin{bmatrix} \vdots \\ x_{i3} \\ \vdots \end{bmatrix},$$

de sorte qu'on a :

$$\begin{aligned} \hat{\beta}^*_{12} &= (X'_{12}X_{12})^{-1} X'_{12}Y \\ &= (X'_{12}X_{12})^{-1} X'_{12}(X_{12}\beta_{12} + X_3\beta_3 + e) \qquad (\text{car } Y = X_{12}\beta_{12} + X_3\beta_3 + e) \\ &= (X'_{12}X_{12})^{-1} X'_{12}X_{12}\beta_{12} + (X'_{12}X_{12})^{-1} X'_{12}X_3\beta_3 + (X'_{12}X_{12})^{-1} X'_{12}e \\ &= \beta_{12} + (X'_{12}X_{12})^{-1} X'_{12}X_3\beta_3 + (X'_{12}X_{12})^{-1} X'_{12}e\,, \end{aligned}$$

et donc :

$$\begin{aligned} E(\hat{\beta}^*_{12}) &= E\left[\beta_{12} + (X'_{12}X_{12})^{-1} X'_{12}X_3\beta_3 + (X'_{12}X_{12})^{-1} X'_{12}e\right] \\ &= \beta_{12} + (X'_{12}X_{12})^{-1} X'_{12}X_3\beta_3 + (X'_{12}X_{12})^{-1} X'_{12}E(e) \quad (\text{car } X_{12} \text{ et } X_3 \text{ fixes}) \\ &= \beta_{12} + (X'_{12}X_{12})^{-1} X'_{12}X_3\beta_3 \qquad\qquad\qquad\qquad (\text{car } E(e) = 0)\,, \end{aligned}$$

[90] Par exemple, parce que la variable x_{i3} n'est en pratique pas disponible.

soit :

$$\begin{aligned} E(\hat{\beta}_{12}^*) &= \beta_{12} + \beta_3\hat{\delta} \\ \Leftrightarrow \quad \begin{bmatrix} E(\hat{\beta}_1^*) \\ E(\hat{\beta}_2^*) \end{bmatrix} &= \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \beta_3 \begin{bmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{bmatrix} \end{aligned}$$

où $\hat{\delta} = (X_{12}'X_{12})^{-1}X_{12}'X_3$ n'est autre que l'estimateur MCO $\hat{\delta}$ de la régression de la variable omise x_{i3} sur x_{i2} (et une constante) :

$$x_{i3} = \delta_1 + \delta_2 x_{i2} + e_i$$

De l'équation (2.9) de la Section 2.2.1, on a $\hat{\delta}_2 = \frac{Cov_e(x_{i2},x_{i3})}{Var_e(x_{i2})}$, de sorte qu'en ce qui concerne l'estimateur $\hat{\beta}_2^*$ du paramètre d'intérêt β_2, on obtient finalement :

$$E(\hat{\beta}_2^*) = \beta_2 + \beta_3\frac{Cov_e(x_{i2}, x_{i3})}{Var_e(x_{i2})}$$

On constate que l'estimateur MCO $\hat{\beta}_2^*$ du modèle de régression (7.13) — où la variable x_{i3} a été omise — est un estimateur généralement biaisé du paramètre d'intérêt β_2 du modèle (7.11), sauf si l'une au moins des deux conditions suivantes est remplie :

1- $\beta_3 = 0$ dans le modèle de régression (7.11). Dans ce cas, on a $E(y_i|x_{i2}, x_{i3}) = \beta_1 + \beta_2 x_{i2}$, autrement dit, $E(y_i|x_{i2}, x_{i3})$ est une fonction linéaire de x_{i2}, et ne dépend pas de x_{i3}. Il n'est donc pas étonnant qu'omettre la variable x_{i3} du modèle ne crée pas de biais.

2- La variable omise x_{i3} est non corrélée avec x_{i2} (i.e., $Cov_e(x_{i2}, x_{i3}) = 0$)[91]. Il y a en pratique peu de chance que cette condition soit remplie.

Si aucune de ces deux conditions n'est remplie, l'estimateur MCO $\hat{\beta}_2^*$ est biaisé, et son biais est donné par :

$$Biais(\hat{\beta}_2^*) = E(\hat{\beta}_2^*) - \beta_2 = \beta_3\frac{Cov_e(x_{i2}, x_{i3})}{Var_e(x_{i2})}$$

Le tableau ci-dessous résume le signe du biais de $\hat{\beta}_2^*$ en fonction des signes de β_3 et de la corrélation empirique $\rho_e(x_{i2}, x_{i3})$ entre x_{i2} et x_{i3}.

$Biais(\hat{\beta}_2^*)$	$Cov_e(x_{i2}, x_{i3}) > 0$ $\Leftrightarrow \rho_e(x_{i2}, x_{i3}) > 0$	$Cov_e(x_{i2}, x_{i3}) < 0$ $\Leftrightarrow \rho_e(x_{i2}, x_{i3}) < 0$
$\beta_3 > 0$	positif	négatif
$\beta_3 < 0$	négatif	positif

Ainsi, dans notre exemple de départ où y_i désigne le salaire d'un individu, x_{i2} son

[91] Pour rappel, deux variables sont non corrélées si et seulement si leur covariance est nulle.

niveau d'éducation et x_{i3} son niveau d'expérience professionnelle, on peut supposer que $\beta_3 > 0$, et on peut également s'attendre à avoir en pratique[92] $\rho_e(x_{i2}, x_{i3}) < 0$. Par conséquent, l'effet d'une année d'étude supplémentaire sur le salaire moyen d'un individu, à niveau d'expérience professionnelle constant ($= \beta_2$ dans le modèle (7.11)), s'il est estimé sur base du modèle (7.13) où la variable de niveau d'expérience professionnelle ($= x_{i3}$) a été omise, a toutes les chances d'être sous-évalué ($= Biais(\hat{\beta}_2^*) < 0$) .

L'analyse ci-dessus peut aisément être généralisée au cas de l'omission d'une ou plusieurs variables explicatives dans une régression contenant un nombre quelconque k de variables explicatives. Dans ce cas général, on peut pareillement montrer que l'estimateur MCO $\hat{\beta}_{(.)}^*$ des paramètres du modèle où une ou plusieurs variables explicatives ont été omises est un estimateur biaisé des paramètres du modèle d'intérêt original, sauf si l'une au moins des deux conditions suivantes est remplie :

1- Dans le modèle d'intérêt original, le paramètre de *chacune* des variables omises est égal à zéro (i.e., les variables omises sont en fait non pertinentes).

2- *Chacune* des variables omises est (deux à deux) non corrélée avec *chacune* des autres variables du modèle d'intérêt original.

Dans ce cas général, le signe des biais est cependant nettement plus compliqué à déterminer.

En résumé, lorsqu'on désire estimer l'effet marginal sur y_i d'une variable x_{ij}, d'autres variables (x_{i2},...,x_{ij-1}, x_{ij+1},...,x_{ik}) étant maintenues constantes, sauf cas particulier, il est indispensable de bien inclure ces autres variables dans la régression. Si on ne le fait pas, ou que l'on ne le fait que partiellement (omission de certaines variables seulement), on obtiendra généralement des estimations biaisées[93] des effets que l'on cherche à estimer. Ce résultat général avait déjà été suggéré à la Section 6.1.2, où on a souligné que des modèles ayant des ensembles de variables conditionnantes différents, sont des modèles différents, et répondent à des questions différentes.

Nous venons de voir qu'omettre des variables pertinentes (i.e., dont le paramètre est différent de zéro) d'un modèle de régression crée des biais à l'estimation. Pour conclure, on considère brièvement le cas de l'inclusion, à tort, de variables non pertinentes dans une régression.

Supposons à nouveau qu'on s'intéresse aux paramètres du modèle de régression standard (7.11), que l'on suppose toujours correctement spécifié, mais que plutôt que d'estimer ces paramètres à l'aide de l'estimateur MCO (7.12), on estime ces paramètres à l'aide de l'estimateur MCO standard $\hat{\beta}^*$ du modèle :

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i \,, \tag{7.14}$$

[92] Les individus les plus éduqués sont les individus les plus jeunes (effet de génération). Etant les plus jeunes, ils ont forcément un niveau d'expérience plus faible que les individus les plus âgés, qui sont en moyenne moins éduqués.

[93] Notons que l'expression *estimations biaisées* est un abus de language. Au sens strict, c'est l'estimateur utilisé qui est biaisé.

où a été inclu une variable non pertinente x_{i4}, càd. une variable telle que $E(y_i|x_{i2}, x_{i3}, x_{i4}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$, ce qui implique que dans (7.14), le paramètre $\beta_4 = 0$.

On sait que, le modèle (7.11) étant correctement spécifié, l'estimateur MCO (7.12) est un estimateur non biaisé de $\beta = (\beta_1, \beta_2, \beta_3)'$. Le modèle (7.11) étant correctement spécifié, et la variable x_{i4} étant non pertinente, le modèle de régression (7.14) est aussi correctement spécifié, de sorte l'estimateur MCO de ce modèle incluant la variable non pertinente x_{i4} est aussi non biaisé :

$$E(\hat{\beta}^*) = \begin{bmatrix} E(\hat{\beta}_1^*) \\ E(\hat{\beta}_2^*) \\ E(\hat{\beta}_3^*) \\ E(\hat{\beta}_4^*) \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ 0 \end{bmatrix}$$

Simplement, comme $\beta_4 = 0$, on a $E(\hat{\beta}_4^*) = 0$. En d'autres termes, l'inclusion, à tort, d'une variable non pertinente dans le modèle ne crée pas de biais d'estimation.

L'inclusion, à tort, de la variable non pertinente x_{i4} n'est cependant pas sans conséquence. En effet, du Théorème Gauss-Markov, on sait que le meilleur estimateur linéaire sans biais de β dans le modèle correctement spécifié (7.11) est l'estimateur MCO (7.12). On a donc nécessairement :

$$V(\hat{\beta}_{123}^*) \geq V(\hat{\beta}), \quad \text{où } \hat{\beta}_{123}^* = \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \hat{\beta}_3^* \end{bmatrix}$$

Autrement dit, l'estimateur MCO $\hat{\beta}_{123}^*$ du modèle (7.14) incluant la variable non pertinente x_{i4} a nécessairement, pour le vecteur de paramètre $\beta = (\beta_1, \beta_2, \beta_3)'$, une matrice de variance - covariance supérieure ou égale (au sens matriciel) à celle de l'estimateur MCO $\hat{\beta}$ du modèle (7.11). Ce résultat est également valable dans le cas de l'inclusion, à tort, de plusieurs (plutôt qu'une seule) variables non pertinentes dans une régression.

En résumé, contrairement au cas de l'omission de variables (pertinentes)[94], l'inclusion de variables non pertinentes ne crée pas de biais d'estimation, mais réduit généralement la précision d'estimation, ce qui n'est évidemment pas souhaitable, et est donc à éviter.

7.4.3. Hétéroscédasticité et auto-corrélation

Comme on peut le voir des développements des Sections 3.1 et 6.3.1, les seules hypothèses nécessaires pour obtenir un estimateur non biaisé (et convergent) du

[94] Si elles sont non pertinentes, on est dans un des deux cas particuliers où cela ne pose pas de problème.

vecteur de paramètres β du modèle de régression :

$$Y = X\beta + e \tag{7.15}$$

sont les hypothèses A1 et A2 (+ par commodité l'hypothèse A5 que X est non-stochastique), qui assurent que la forme fonctionnelle du modèle est correctement spécifiée. Ni l'hypothèse A3 d'homoscédasticité, ni l'hypothèse A4 de non-corrélation ne sont nécessaires pour que l'estimateur MCO standard $\hat{\beta} = (X'X)^{-1} X'Y$ soit non biaisé.

Ces hypothèses d'homoscédasticité et de non-corrélation sont cependant cruciales pour la validité de toutes les procédures d'inférence (intervalles de confiance, tests d'hypothèse et intervalles de prévision) que nous avons établies. Sous ces hypothèses additionnelles, nous avons montré que la matrice de variance - covariance de l'estimateur MCO $\hat{\beta}$ est donnée par :

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} , \tag{7.16}$$

et qu'un estimateur non biaisé et convergent de cette matrice de variance - covariance est donné par :

$$\hat{V}(\hat{\beta}) = \hat{s}^2 (X'X)^{-1} \tag{7.17}$$

Cet estimateur de la matrice de variance - covariance de $\hat{\beta}$ (ou des éléments de celle-ci, comme les variances ou les écarts-types estimés des différents paramètres) intervient dans toutes les procédures d'inférence que nous avons étudiés.

Ainsi, si les hypothèses additionnelles A3 et A4 d'homoscédasticité et de non-corrélation ne sont pas satisfaites, les procédures d'inférence que nous avons établies ne sont plus valables. Par ailleurs, lorsque ces hypothèses additionnelles A3 et A4 ne sont pas satisfaites, on a aussi que les conditions du Théorème Gauss-Markov ne sont plus remplies, de sorte que l'estimateur MCO $\hat{\beta}$ n'est plus le meilleur estimateur linéaire sans biais de β. Comme on vient de le voir, l'estimateur MCO $\hat{\beta}$ est toujours non biaisé, et peut donc toujours être utilisé, mais il n'est plus celui qui a la plus petite (au sens matriciel) matrice de variance - covariance parmi les estimateurs linéaires sans biais de β. On peut montrer que dans ce cas le meilleur estimateur linéaire sans biais est un estimateur appelé l'*estimateur des Moindres Carrés Généralisés* (MCG). Nous ne développerons pas ici cet estimateur. Nous allons par contre voir comment on peut modifier les procédures d'inférence standards associées à l'estimateur MCO de façon à ce qu'elles restent valables lorsque les hypothèses additionnelles A3 et A4 ne sont pas satisfaites.

Lorsque l'hypothèse A3 - A4 $V(e) = V(Y) = \sigma^2 I$ d'homoscédasticité et de non-corrélation du modèle standard n'est pas satisfaite, on peut de façon générale avoir :

$$V(e) = V(Y) = \Omega = \begin{bmatrix} \sigma_1^2 & \cdots & \gamma_{1i} & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ \gamma_{i1} & \cdots & \sigma_i^2 & \cdots & \gamma_{in} \\ \vdots & & \vdots & \ddots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{ni} & \cdots & \sigma_n^2 \end{bmatrix} , \tag{7.18}$$

où σ_i^2 et γ_{ij}, qui désignent respectivement les variances et covariances (conditionnelles) des observations, peuvent être des fonctions de X. La forme générale (7.18) permet d'avoir n'importe quelle forme d'hétéroscédasticité et de corrélation entre les observations[95]. Elle est donc par définition toujours correcte. Dans le cas particulier où l'hypothèse A3 - A4 d'homoscédasticité et de non-corrélation est remplie, on a simplement $\Omega = \sigma^2 I$.

On peut aisément obtenir la matrice de variance - covariance de l'estimateur MCO $\hat{\beta}$ pour ce cas général. Sous les hypothèses A1, A2 et A5, on a :

$$\hat{\beta} = \beta + (X'X)^{-1} X'e \quad \text{et} \quad E(\hat{\beta}) = \beta \,,$$

de sorte qu'on obtient :

$$\begin{aligned}
V(\hat{\beta}) &= E\left[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'\right] \\
&= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] && (\text{car } E(\hat{\beta}) = \beta) \\
&= E\left[(X'X)^{-1} X'ee'X\,(X'X)^{-1}\right] && (\text{car } \hat{\beta} - \beta = (X'X)^{-1} X'e)) \\
&= (X'X)^{-1} X'E(ee')X\,(X'X)^{-1} && (\text{car } X \text{ fixe}) \\
&= (X'X)^{-1} X'\Omega X\,(X'X)^{-1} && (\text{car } E(ee') = V(e) = \Omega)
\end{aligned}$$

Sous les seules hypothèses A1, A2 et A5, sans faire aucune hypothèse spécifique sur les variances et covariances (conditionnelles) des observations, la matrice de variance - covariance de l'estimateur MCO $\hat{\beta}$ est donc donnée par :

$$V(\hat{\beta}) = (X'X)^{-1} X'\Omega X\,(X'X)^{-1} \,, \tag{7.19}$$

tandis qu'elle se réduit à la formule standard $V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ si l'hypothèse A3 - A4 d'homoscédasticité et de non-corrélation est remplie, càd. si $\Omega = \sigma^2 I$.

Des procédures d'inférence (intervalles de confiance, tests d'hypothèse et intervalles de prévision[96]) valables sous les seules hypothèses A1, A2 et A5 — donc sans faire appel aux hypothèses A3 et A4 — peuvent être obtenues en utilisant, dans les calculs de toutes ces procédures, un estimateur convergent de la matrice de variance - covariance générale (7.19), en lieu et place de l'estimateur standard $\hat{V}(\hat{\beta}) = \hat{s}^2 (X'X)^{-1}$, qui n'est valable que sous les hypothèses additionnelles A3 et A4.

Dans la quête d'un estimateur convergent de la matrice de variance - covariance générale (7.19), on distingue deux cas : le cas où on peut considérer que les observations sont non corrélées, mais peuvent être hétérocédastiques, et le cas général où les observations peuvent à la fois être corrélées et hétéroscédastiques.

---

[95] Notons que si Ω peut être généralement quelconque, comme c'est une matrice de variance - covariance, elle doit tout de même nécessairement être symétrique et (semi-) définie positive.

[96] Pour l'intervalle de prévision de $E(y_0)$ sachant $(x_{02},...,x_{0k})$, mais *pas* pour l'intervalle de prévision de y_0 sachant $(x_{02},...,x_{0k})$.

7.4.3.1. Hétéroscédasticité

Lorsqu'on analyse des *données en coupe*, on peut généralement considérer, pour des raisons d'échantillonnage[97] ou de modélisation, que les observations sont indépendantes d'un individu à l'autre, et donc non corrélées. L'hypothèse A4 de non-corrélation est donc automatiquement satisfaite, et seule une possible violation de l'hypothèse A3 d'homoscédasticité est à considérer. Dans ce cas, la matrice de variance - covariance $V(e) = V(Y) = \Omega$ des observations est une matrice diagonale, et la matrice de variance - covariance générale (7.19) de l'estimateur MCO $\hat{\beta}$ se réduit à :

$$\begin{aligned} V(\hat{\beta}) &= (X'X)^{-1} X'\Omega X (X'X)^{-1} \\ &= (X'X)^{-1} \left(\sum_{i=1}^{n} X_i' X_i \sigma_i^2 \right) (X'X)^{-1} , \end{aligned} \tag{7.20}$$

où σ_i^2 est la variance (conditionnelle) de l'observation i et $X_i = \begin{bmatrix} 1 & x_{i2} & \cdots & x_{ik} \end{bmatrix}$ désigne la i-ième ligne de la matrice des observations X.

On peut montrer, sous des conditions de régularité générales, qu'un estimateur convergent (mais pas non biaisé) de cette matrice (7.20) de variance - covariance de l'estimateur MCO $\hat{\beta}$ est donné par :

$$\hat{V}_{HC}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{i=1}^{n} X_i' X_i \hat{e}_i^2 \right) (X'X)^{-1} , \tag{7.21}$$

où $\hat{e}_i = y_i - X_i\hat{\beta}$. Cet estimateur, qui est dû à White[98], est généralement appelé[99] *estimateur robuste à l'hétéroscédasticité* de la matrice de variance - covariance de l'estimateur MCO $\hat{\beta}$.

Si, dans les calculs des différentes procédures d'inférence que nous avons étudiées, on remplace l'estimateur standard $\hat{V}(\hat{\beta}) = \hat{s}^2 (X'X)^{-1}$ de la matrice de variance - covariance de $\hat{\beta}$ par cet estimateur robuste $\hat{V}_{HC}(\hat{\beta})$[100], on obtient des procédures d'inférence[101] qui sont valables sous les seules hypothèses A1, A2 et A5, donc sans faire appel à l'hypothèse A3 d'homoscédasticité, étant entendu qu'avec des données en coupe l'hypothèse A4 de non-corrélation est sensée être automatiquement

---

[97] Si les observations sont obtenues par tirage aléatoire avec remise — ou sans remise, si l'échantillon est petit par rapport à la population — d'individus dans une population, elles sont par construction indépendantes.

[98] White H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct test for Heterosckedasticity", *Econometrica*, 48, p. 817-838.

[99] En anglais, on dit *heteroskedasticity robust covariance matrix estimator* ou encore *heteroskedasticity consistent covariance matrix estimator*, la deuxième appellation étant à la source de l'abréviation 'HC' (= Heteroskedasticity Consistent).

[100] Par exemple, pour le calcul de l'intervalle de confiance d'un paramètre β_j, cela signifie remplacer l'estimateur standard $s.\hat{e}.(\hat{\beta}_j)$ de l'écart-type du paramètre par l'estimateur robuste $s.\hat{e}._{HC}(\hat{\beta}_j)$, qui est donné par la racine carrée de l'élément (j, j) de $\hat{V}_{HC}(\hat{\beta})$.

[101] Pour toutes les procédures d'inférence que nous avons étudiées, excepté l'intervalle de prévision de y_0 sachant $(x_{02},..., x_{0k})$.

satisfaite. Notons cependant que ces procédures ainsi modifiées ne sont valables qu'*asymptotiquement*, à titre approximatif pour n grand, et ce même si les y_i sont distribués de façon normale. La plupart des logiciels économétriques (GRETL en particulier) permettent de calculer, de façon optionnelle, la matrice de variance-covariance — et les écart-types — robustes à l'hétéroscédasticité des paramètres estimés[102].

L'estimateur robuste (7.21) est un estimateur convergent de $V(\hat{\beta})$ quelque soit la forme d'hétéroscédasticité présente dans les données. C'est également un estimateur convergent de $V(\hat{\beta})$ si l'hypothèse A3 d'homoscédasticité est en réalité satisfaite. Dans ce dernier cas, il vaut cependant mieux utiliser l'estimateur standard $\hat{V}(\hat{\beta}) = \hat{s}^2 (X'X)^{-1}$, car il est plus précis. Pour savoir en pratique quel estimateur de $V(\hat{\beta})$ utiliser, on peut tester si l'hypothèse A3 d'homoscédasticité est ou non remplie.

Un test de l'hypothèse A3 d'homoscédasticité peut être effectué sur base de la régression auxiliaire :

$$\hat{e}_i^2 = \delta_1 + \delta_2 x_{i2} + \delta_3 x_{i3} + ... + \delta_k x_{ik} + v_i \tag{7.22}$$

càd. de la régression du carré des résidus $\hat{e}_i^2$ sur une constante et les différentes variables explicatives du modèle d'intérêt (7.15).

Si l'hypothèse A3 d'homoscédasticité est vraie, on a $Var(e_i) = E(e_i^2) = \sigma^2$ (i.e., une constante), pour tout $i = 1,..., n$. Comme $\hat{e}_i^2$ est un estimateur convergent de e_i^2, dans la régression auxiliaire (7.22), on s'attend, si l'hypothèse A3 d'homoscédasticité est vraie, à ce que tous les paramètres sauf l'intercept soient non significativement différents de zéro.

Cela peut être formellement testé au travers d'un simple F-test de H_0 : $\delta_2 = ... = \delta_k = 0$ contre H_1 : $\delta_2 \neq 0$ et/ou ... et/ou $\delta_k \neq 0$. Une autre statistique de test, asymptotiquement équivalente au F-test, est cependant plus souvent utilisée pour formellement tester la significativité jointe des paramètres δ_2, ..., δ_k dans la régression auxiliaire (7.22). Il s'agit de la statistique[103] :

$$LM_H = n \times R^2$$

où n est la taille d'échantillon et R^2 est le coefficient de détermination multiple de la régression auxiliaire (7.22). On peut montrer que, sous l'hypothèse nulle H_0 d'homoscédasticité[104], $LM_H \sim \chi^2 (k-1)$, où $(k-1)$ est égal au nombre de variables (hors intercept) incluses dans la régression auxiliaire (7.22), de sorte que la règle de décision du test au seuil α est donnée par :

$$\begin{cases} \text{- Rejet de } H_0 \text{ si } LM_H > \chi^2_{k-1;1-\alpha} \\ \text{- Non-rejet de } H_0 \text{ sinon} \end{cases}$$

[102] L'estimateur robuste $\hat{V}_{HC}(\hat{\beta})$ utilisé par les logiciels économétriques peut en pratique être une variante — asymptotiquement équivalente — de l'estimateur donné par (7.21).

[103] L'abbréviation 'LM' de cette statistique vient du fait qu'il s'agit d'un test dit *du Multiplicateur de Lagrange* (*Lagrange Multiplier test* en anglais).

[104] Si l'hypothèse nulle H_0 d'homoscédasticité est fausse, LM_H suit une loi du khi-carré non-centrale.

où la valeur critique $\chi^2_{k-1;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(k-1)$, et la P-valeur de ce test, pour un *échantillon particulier* où la statistique de test prend la valeur particulière LM^*_H, est donnée par :

$$p_{LM_H} = I\!P(v > LM^*_H)\,, \quad \text{où } v \sim \chi^2(k-1)$$

Ce test est connu sous le nom de *test d'hétéroscédasticité de Breusch-Pagan*, bien qu'il s'agisse en réalité d'une version modifiée par Koenker (1981)[105] du test original proposé par Breusch et Pagan (1979)[106]. On notera que ce test, mais aussi sa version F-test, n'est valable qu'*asymptotiquement*, à titre approximatif pour n grand, et ce même si les y_i sont distribués de façon normale.

Pour conclure, on notera encore que, dans la régression auxiliaire (7.22), à côté des différentes variables explicatives $(x_{i2}, x_{i3}, ..., x_{ik})$ du modèle d'intérêt (7.15), on peut encore ajouter les carrés et les produits croisés de ces variables. Sous cette forme étendue, le test est appelé *test d'hétéroscédasticité de White*[107]. Dans le même esprit que le test RESET, on peut également, dans la régression auxiliaire (7.22), à côté de l'intercept δ_1, plutôt que les variables $(x_{i2}, x_{i3}, ..., x_{ik})$, considérer comme variables explicatives à tester les puissances $\hat{y}_i$, $\hat{y}_i^2$, $\hat{y}_i^3$, ... des valeurs prédites $\hat{y}_i = X_i\hat{\beta}$ du modèle d'intérêt (7.15)[108]. On notera que la valeur prédite $\hat{y}_i = X_i\hat{\beta}$ elle-même (i.e., la puissance d'ordre 1 de $\hat{y}_i$) peut ici être inclue dans le modèle auxiliaire (pas de problème de colinéarité parfaite). En pratique, on utilise rarement les puissances de $\hat{y}_i$ au-delà de l'ordre 4. Cette dernière forme du test est particulièrement indiquée lorsqu'on a des raisons de penser que la variance des observations est liée à leur moyenne.

7.4.3.2. Auto-corrélation

Lorsqu'on analyse des *données chronologiques* ou séries temporelles au travers du modèle :

$$\begin{aligned} & Y = X\beta + e \\ \Leftrightarrow \qquad & y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + ... + \beta_k x_{tk} + e_t\,, \quad t = 1, ..., T, \end{aligned} \tag{7.23}$$

outre une possible hétéroscédasticité, les observations sont couramment sériellement corrélées, càd. corrélées d'une période à l'autre. On parle d'*auto-corrélation*. Cela se produit typiquement lorsqu'on considère une modèle *statique*, càd. un modèle où les variables explicatives n'incluent aucune variable — dépendante (telle que y_{t-1}, y_{t-2},...) ou indépendante (telle que x_{t-1j}, x_{t-2j},...) — retardée[109].

---

[105] Koenker R. (1981), "A Note on Studentizing a Test for Heteroskedasticity", *Journal of Econometrics*, 17, p. 107-112.

[106] Breusch T.S. et Pagan A.R. (1979), "A Simple Test for Heteskedasticity and Random Coefficient Variation", *Econometrica*, 47, p. 987-1007.

[107] Voir note 98 p. 159.

[108] Notons que dans toutes ces variantes, le nombre de degrés de liberté de la loi du χ^2 intervenant dans le test basé sur la statistique LM_H est toujours égal au nombre total de variables explicatives (hors intercept) incluses dans la régression auxiliaire (7.22).

[109] Dans le cas contraire, on parle d'un modèle *dynamique*.

Ainsi, lorsqu'on analyse des données chronologiques, une possible violation à la fois de l'hypothèse A3 d'homoscédasticité et (surtout) de l'hypothèse A4 de non-corrélation est à considérer. Dans ce cas, la matrice de variance - covariance $V(e) = V(Y) = \Omega$ des observations peut être quelconque (pas de restrictions sur les variances et les covariances[110]), et la matrice de variance - covariance générale de l'estimateur MCO $\hat{\beta}$ est donnée par :

$$\begin{aligned} V(\hat{\beta}) &= (X'X)^{-1} X'\Omega X (X'X)^{-1} \\ &= (X'X)^{-1} \left(\sum_{t=1}^{T} X_t' X_t \sigma_t^2 + \sum_{\tau=1}^{T-1} \sum_{t=\tau+1}^{T} (X_t' X_{t-\tau} \gamma_{t(t-\tau)} + X_{t-\tau}' X_t \gamma_{(t-\tau)t}) \right) \\ &\quad \times (X'X)^{-1} , \end{aligned} \tag{7.24}$$

où σ_t^2 est la variance (conditionnelle) de l'observation t, $\gamma_{t(t-\tau)}$ $(= \gamma_{(t-\tau)t})$ la covariance (conditionnelle) entre les observations t et $(t-\tau)$, et $X_t = \begin{bmatrix} 1 & x_{t2} & \cdots & x_{tk} \end{bmatrix}$ désigne la t-ième ligne de la matrice des observations X.

On peut montrer, sous des conditions de régularité générales, qu'un estimateur convergent (mais pas non biaisé) de cette matrice (7.24) de variance - covariance de l'estimateur MCO $\hat{\beta}$ est donné par :

$$\begin{aligned} \hat{V}_{HAC}(\hat{\beta}) &= (X'X)^{-1} \\ &\quad \times \left(\sum_{t=1}^{T} X_t' X_t \hat{e}_t^2 + \sum_{\tau=1}^{q} (1 - \tfrac{\tau}{q+1}) \sum_{t=\tau+1}^{T} (X_t' X_{t-\tau} \hat{e}_t \hat{e}_{t-\tau} + X_{t-\tau}' X_t \hat{e}_{t-\tau} \hat{e}_t) \right) \\ &\quad \times (X'X)^{-1} , \end{aligned} \tag{7.25}$$

où $\hat{e}_t = y_t - X_t \hat{\beta}$. Cet estimateur, qui est dû à Newey et West[111], est généralement appelé[112] *estimateur robuste à l'hétéroscédasticité et l'auto-corrélation* de la matrice de variance - covariance de l'estimateur MCO $\hat{\beta}$. Sa mise en oeuvre pratique requiert de choisir une valeur (entière) pour le paramètre q. Le choix optimal de q dépend notamment de l'importance de l'auto-corrélation présente dans les données. Pour des données annuelles, on peut généralement prendre une valeur faible pour q (disons $q \leq 3$). Une valeur plus élevée de q devrait être choisie pour des données trimestrielles, et plus encore pour des données mensuelles.

Comme à la section précédente, si, dans les calculs des différentes procédures d'inférence que nous avons étudiées, on remplace l'estimateur standard $\hat{V}(\hat{\beta}) = \hat{s}^2 (X'X)^{-1}$ de la matrice de variance - covariance de $\hat{\beta}$ par cet estimateur robuste

---

110 Hormis le fait déjà mentionné que Ω doit nécessairement être symétrique et (semi-) définie positive.

111 Newey W.K. et West K.D. (1987), "A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, 55, p. 703-708.

112 En anglais, on dit *heteroskedasticity and autocorrelation robust covariance matrix estimator* ou encore *heteroskedasticity and autocorrelation consistent covariance matrix estimator*, la deuxième appellation étant à la source de l'abréviation 'HAC' (= Heteroskedasticity and Autocorrelation Consistent).

$\hat{V}_{HAC}(\hat{\beta})$[113], on obtient des procédures d'inférence[114] qui sont valables sous les seules hypothèses A1, A2 et A5, donc sans faire appel ni à l'hypothèse A3 d'homoscédasticité, ni à l'hypothèse A4 de non-corrélation. Notons encore que ces procédures ainsi modifiées ne sont valables qu'*asymptotiquement*, à titre approximatif pour n grand, et ce même si les y_i sont distribués de façon normale. La plupart des logiciels économétriques (GRETL en particulier) permettent à nouveau de calculer, de façon optionnelle, la matrice de variance - covariance — et les écart-types — robustes à l'hétéroscédasticité et l'auto-corrélation des paramètres estimés[115].

L'estimateur robuste (7.25) est un estimateur convergent de $V(\hat{\beta})$ quelque soit la forme d'hétéroscédasticité et d'auto-corrélation présente dans les données. Comme l'estimateur $\hat{V}_{HC}(\hat{\beta})$ (robuste à l'hétéroscédasticité uniquement), c'est également un estimateur convergent de $V(\hat{\beta})$ si les hypothèses A3 et A4 d'homoscédasticité et de non-corrélation sont en réalité satisfaites. Dans ce dernier cas, il vaut cependant à nouveau mieux utiliser l'estimateur standard $\hat{V}(\hat{\beta}) = \hat{s}^2 (X'X)^{-1}$, car il est plus précis. Pour savoir en pratique quel estimateur de $V(\hat{\beta})$ utiliser, on peut tester si les hypothèses A3 et A4 d'homoscédasticité et de non-corrélation sont ou non remplies.

On a vu à la section précédente comment tester l'hypothèse A3 d'homoscédasticité[116]. Un test de l'hypothèse A4 de non-corrélation peut être effectué sur base de la régression auxiliaire :

$$\hat{e}_t = X_t b + \delta_1 \hat{e}_{t-1} + \delta_2 \hat{e}_{t-2} + ... + \delta_p \hat{e}_{t-p} + v_t \tag{7.26}$$

càd. de la régression des résidus $\hat{e}_t$ sur les différentes variables explicatives (y compris l'intercept) du modèle d'intérêt (7.23) et sur les résidus retardés $\hat{e}_{t-1}$, $\hat{e}_{t-2}$,..., $\hat{e}_{t-p}$. En pratique, p peut être choisi de façon semblable au paramètre q de $\hat{V}_{HAC}(\hat{\beta})$. On notera que, pour la validité de ce test, il est essentiel de bien inclure dans cette régression auxiliaire les différentes variables explicatives (y compris l'intercept) du modèle d'intérêt (7.23).

Si l'hypothèse A4 de non-corrélation est vraie, on a $Cov(e_t, e_{t-\tau}) = E(e_t e_{t-\tau}) = 0$, pour tout $t = 1,...,T$ et $\tau = 1,...,p$. Comme $\hat{e}_t$ est un estimateur convergent de e_t, dans la régression auxiliaire (7.26), on s'attend, si l'hypothèse A4 de non-corrélation est vraie, à ce que les paramètres δ_1, δ_2,..., δ_p des résidus retardés soient non significativement différents de zéro.

Cela peut être formellement testé au travers d'un simple F-test de H_0 : $\delta_1 = ... = \delta_p = 0$ contre H_1 : $\delta_1 \neq 0$ et/ou ... et/ou $\delta_p \neq 0$. Comme pour le test d'hétéroscédasticité, une autre statistique de test, asymptotiquement équivalente au F-test, est cependant plus souvent utilisée pour formellement tester la significativité

[113] Par exemple, pour le calcul de l'intervalle de confiance d'un paramètre β_j, cela signifie remplacer l'estimateur standard $s.\hat{e}.(\hat{\beta}_j)$ de l'écart-type du paramètre par l'estimateur robuste $s.\hat{e}._{HAC}(\hat{\beta}_j)$, qui est donné par la racine carré de l'élément (j,j) de $\hat{V}_{HAC}(\hat{\beta})$.

[114] A nouveau, pour toutes les procédures d'inférence que nous avons étudiées, excepté l'intervalle de prévision de y_0 sachant $(x_{02},..., x_{0k})$.

[115] L'estimateur robuste $\hat{V}_{HAC}(\hat{\beta})$ utilisé par les logiciels économétriques peut en pratique être une variante — asymptotiquement équivalente — de l'estimateur donné par (7.25).

[116] Dans le présent contexte de données chronologiques, il suffit de remplacer l'indice i des observations par un indice t des périodes de temps, et le nombre d'observations n par T.

jointe des paramètres δ_1, ..., δ_p dans la régression auxiliaire (7.26). Il s'agit de la statistique[117] :

$$LM_A = T \times R^2$$

où T est la taille d'échantillon et R^2 est le coefficient de détermination multiple de la régression auxiliaire (7.26). On peut montrer que, sous l'hypothèse nulle H_0 de non-corrélation[118], $LM_A \sim \chi^2(p)$, où p est égal au nombre des résidus retardés $\hat{e}_{t-1}$, $\hat{e}_{t-2}$,..., $\hat{e}_{t-p}$ inclus dans la régression auxiliaire (7.26), de sorte que la règle de décision du test au seuil α est donnée par :

$$\begin{cases} \text{- Rejet de } H_0 \text{ si } LM_A > \chi^2_{p;1-\alpha} \\ \text{- Non-rejet de } H_0 \text{ sinon} \end{cases}$$

où la valeur critique $\chi^2_{p;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(p)$, et la P-valeur de ce test, pour un *échantillon particulier* où la statistique de test prend la valeur particulière LM_A^*, est donnée par :

$$p_{LM_A} = I\!P(v > LM_A^*), \quad \text{où } v \sim \chi^2(p)$$

Ce test est connu sous le nom de *test d'auto-corrélation de Breusch*[119]*-Godfrey*[120]. On notera que ce test, mais aussi sa version F-test, n'est à nouveau valable qu'*asymptotiquement*, à titre approximatif pour n grand, et ce même si les y_t sont distribués de façon normale.

7.4.4. Non-normalité

L'hypothèse optionnelle A6 de normalité nous a permis d'obtenir des procédures d'inférence (intervalles de confiance, tests d'hypothèse et intervalles de prévision) *exactes* en *échantillon fini*. Toutefois, nous avons vu qu'elle n'était pas essentielle car, sans cette hypothèse, les mêmes procédures d'inférence restent valables *asymptotiquement*, à titre approximatif pour n grand. Cela est vrai pour toutes les procédures d'inférence que nous avons étudiées, sauf une : l'intervalle de prévision pour la valeur de y sachant $(x_{02}, ...x_{0k})$. La validité de cet intervalle de prévision requiert en effet que l'hypothèse optionnelle A6 de normalité soit satisfaite (cf. Section 6.5). Pour savoir en pratique si cette hypothèse tient ou non, on peut la tester.

Un test de l'hypothèse optionnelle A6 de normalité peut être effectué sur base

[117] L'abréviation 'LM' de cette statistique vient à nouveau du fait qu'il s'agit d'un test dit *du Multiplicateur de Lagrange* (*Lagrange Multiplier test* en anglais).

[118] Si l'hypothèse nulle H_0 de non-corrélation est fausse, LM_A suit une loi du khi-carré non-centrale.

[119] Breusch T.S. (1978), "Testing for Autocorrelation in Dynamic Linear Models", *Australian Economic Papers*, 17, p. 334-355.

[120] Godfrey L.G. (1978), "Testing for Higher Order Serial Correlation in Regression Equations when the Regressors Include Lagged Dependent Variables", *Econometrica*, 46, p. 1303-1310.

de la statistique de test[121] :

$$LM_N = \frac{n}{6}\left(\hat{\alpha}_1^2 + \frac{(\hat{\alpha}_2 - 3)^2}{4}\right),$$

où :

$$\hat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{e}_i^3}{\hat{\sigma}^3}, \quad \hat{\alpha}_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{e}_i^4}{\hat{\sigma}^4}, \quad \hat{e}_i = y_i - X_i\hat{\beta} \text{ et } \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2}$$

Si l'hypothèse A6 de normalité est vraie, on a $e_i \sim N(0, \sigma^2)$, de sorte que $\alpha_1 = E(\frac{e_i^3}{\sigma^3}) = 0$ et $\alpha_2 = E(\frac{e_i^4}{\sigma^4}) = 3$, pour tout $i = 1,...,n$, où α_1 et α_2 désignent respectivement les coefficients d'asymétrie et de kurtosis de la loi normale $N(0, \sigma^2)$[122]. Comme $\hat{\alpha}_1$ et $\hat{\alpha}_2$ sont des estimateurs convergents de α_1 et α_2, on s'attend, si l'hypothèse A6 de normalité est vraie, à ce que la statistique LM_N prenne des valeurs proches de zéro. Formellement, on peut montrer que, sous l'hypothèse nulle H_0 de normalité[123], $LM_N \sim \chi^2(2)$, de sorte que la règle de décision du test au seuil α est donnée par :

$$\begin{cases} \text{- Rejet de } H_0 \text{ si } LM_N > \chi^2_{2;1-\alpha} \\ \text{- Non-rejet de } H_0 \text{ sinon} \end{cases}$$

où la valeur critique $\chi^2_{2;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(2)$, et la P-valeur de ce test, pour un *échantillon particulier* où la statistique de test prend la valeur particulière LM_H^*, est donnée par :

$$p_{LM_N} = I\!P(v > LM_H^*), \quad \text{où } v \sim \chi^2(2)$$

Ce test est connu sous le nom de *test de normalité de Jarque-Bera*[124]. On notera que ce test n'est à nouveau valable qu'*asymptotiquement*, à titre approximatif pour n grand.

7.4.5. Régresseurs stochastiques

Lorsque nous l'avons introduit, nous avons dit que l'hypothèse (peu réaliste) A5 selon laquelle X est non-stochastique était faite pour des raisons de commodité technique, et qu'elle équivalait à raisonner, pour X stochastique, conditionnellement aux valeurs de X observées dans l'échantillon. De fait, l'ensemble des résultats que nous avons établis peuvent de façon équivalente être obtenus sur base des hypothèses plus réalistes :

A1′ $\qquad Y = X\beta + e$

[121] L'abréviation 'LM' de cette statistique vient à nouveau du fait qu'il s'agit d'un test dit *du Multiplicateur de Lagrange* (*Lagrange Multiplier test* en anglais).

[122] Cf. l'annexe B de Hill, Griffiths et Lim (2008).

[123] Si l'hypothèse nulle H_0 de normalité est fausse, LM_N suit une loi du khi-carré non-centrale.

[124] Jarque C.M. et Bera A.K. (1980), "Efficient Tests for Normality, Homosckedasticity and Serial Independence of Regression Residuals", *Economics Letters*, 6, p. 255-259.

A2′	$E(e\|X) = 0 \Leftrightarrow E(Y\|X) = X\beta$
A3′ - A4′	$V(e\|X) = \sigma^2 I = V(Y\|X)$
A5′	$\text{rg}(X) = k$
A6′	(optionnel) $e\|X \sim N(0, \sigma^2 I) \Leftrightarrow Y\|X \sim N(X\beta, \sigma^2 I)$

Les hypothèses A1′ à A6′ sont identiques aux hypothèses A1 à A6, à l'exception du fait que l'hypothèse que X est non-stochastique est remplacée, pour X stochastique, par un conditionnement par rapport aux valeurs de X observées dans l'échantillon. Voici, à titre d'exemple, comment on obtient l'espérance et la matrice de variance - covariance de l'estimateur MCO $\hat{\beta}$ sur base de ces hypothèses. De l'hypothèse A1′ $Y = X\beta + e$, on a :

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} X'Y = (X'X)^{-1} X'(X\beta + e) \\ &= (X'X)^{-1} X'X\beta + (X'X)^{-1} X'e \\ &= \beta + (X'X)^{-1} X'e \end{aligned}$$

de sorte que, de l'hypothèse A2′ $E(e|X) = 0$, on obtient[125] :

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

Comme $E(\hat{\beta}|X)$ ne dépend pas de X, on a encore[126] :

$$E(\hat{\beta}) = E\left[E(\hat{\beta}|X)\right] = E(\beta) = \beta$$

Si on ajoute l'hypothèse A3′ - A4′ $V(e|X) = \sigma^2 I$, on obtient par ailleurs :

$$\begin{aligned} V(\hat{\beta}|X) &= E\left[(\hat{\beta} - E(\hat{\beta}|X))(\hat{\beta} - E(\hat{\beta}|X))'|X\right] \\ &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X\right] && (\text{car } E(\hat{\beta}|X) = \beta) \\ &= E\left[(X'X)^{-1} X'ee'X (X'X)^{-1} |X\right] && (\text{car } \hat{\beta} - \beta = (X'X)^{-1} X'e)) \\ &= (X'X)^{-1} X'E(ee'|X)X (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} X'X (X'X)^{-1} && (\text{car } E(ee'|X) = V(e|X) = \sigma^2 I) \\ &= \sigma^2 (X'X)^{-1} \end{aligned}$$

Dans les calculs ci-dessus, l'hypothèse A5′ n'intervient que pour assurer que $(X'X)$ est inversible, et donc que l'estimateur MCO est bien défini. On peut, de façon semblable, ré-obtenir tous les résultats établis précédemment. Ainsi, toutes les propriétés et procédures d'inférence que nous avons vues sont de façon équivalente

[125] Dans le calcul d'une espérance conditionnelle, tout ce qui est fonction des variables de l'ensemble conditionnant peut être traité comme une constante.

[126] La *loi des espérances itérées* relie l'espérance conditionnelle et non conditionnelle. De façon générale, on a : $E(Y) = E[E(Y|X)]$.

valables sous les hypothèses plus réalistes A1′ à A5′ (+ optionnellement A6′ pour des résultats exacts en échantillon fini).

On notera néanmoins que le conditionnement par rapport à l'ensemble des valeurs de X — ou de façon équivalente l'hypothèse que X est non-stochastique — n'est pas aussi anodin qu'il n'y paraît à première vue. De façon détaillée, pour chaque observation i, cela signifie par exemple pour l'hypothèse A2′ :

$$E(y_i|X) = E(y_i|X_1, ..., X_i, ..., X_n) = E(y_i|X_i) = X_i\beta\,, \quad i = 1, ..., n,$$

où $X_i = \begin{bmatrix} 1 & x_{i2} & \cdots & x_{ik} \end{bmatrix}$ désigne la i-ième ligne de la matrice des observations X. Autrement dit, cela suppose que $E(y_i|X)$ ne dépend en fait que des variables explicatives X_i de l'observation i, et pas des variables explicatives des observations autres que i. Lorsque cette hypothèse implicite à A2′ est satisfaite, on dit que les variables explicatives sont *strictement exogènes*.

Si cette hypothèse d'*exogénéité stricte* est naturelle lorsqu'on analyse des *données en coupe*, où — pour des raisons d'échantillonnage[127] ou de modélisation — les observations peuvent être considérées comme indépendantes d'un individu i à l'autre, il n'en va pas de même lorsqu'on analyse des *données chronologiques* ou séries temporelles au travers du modèle :

$$\begin{aligned} & Y = X\beta + e \\ \Leftrightarrow \quad & y_t = X_t\beta + e_t\,, \quad t = 1, ..., T, \end{aligned}$$

où $X_t = \begin{bmatrix} 1 & x_{t2} & \cdots & x_{tk} \end{bmatrix}$ désigne la t-ième ligne de la matrice des observations X.

Dans ce cas, l'hypothèse d'*exogénéité stricte* inclue dans A2′ requiert que :

$$E(y_t|X) = E(y_t|X_1, ..., X_t, ..., X_T) = E(y_t|X_t) = X_t\beta\,, \quad t = 1, ..., T,$$

autrement dit que $E(y_t|X)$ ne dépende en fait que des variables explicatives X_t de la période t, et pas des variables explicatives des périodes autres (passées ou futures) que t. Cette hypothèse est assez restrictive car elle exclut par exemple tout phénomène de 'feedback' (i.e., la variable y_t influence les valeurs futures X_{t+1}, X_{t+2},... des variables explicatives), ou encore la présence de *variables dépendantes retardées* parmi les variables explicatives, comme dans le *modèle dynamique autorégressif* (à l'ordre 1):

$$y_t = \beta_1 + \beta_2 y_{t-1} + e_t\,, \quad t = 1, ..., T \tag{7.27}$$

En effet, pour ce modèle dynamique, on a $X_t = \begin{bmatrix} 1 & y_{t-1} \end{bmatrix}$, de sorte que :

$$\begin{aligned} E(y_t|X) &= E(y_t|1, y_0, y_1, ..., y_t, ..., y_{T-1}) = y_t \\ &\neq E(y_t|X_t) = \beta_1 + \beta_2 y_{t-1}\,, \quad t = 1, ..., T, \end{aligned}$$

autrement dit, l'hypothèse d'exogénéité stricte n'est pas satisfaite.

[127] Pour rappel, si les observations sont obtenues par tirage aléatoire avec remise — ou sans remise, si l'échantillon est petit par rapport à la population — d'individus dans une population, elles sont par construction indépendantes.

Il apparaît ainsi que les hypothèses A1′ à A6′ — ou de façon équivalente les hypothèses A1 à A6 qui supposent que X est non-stochastique[128] — sont généralement trop restrictives pour l'analyse des données chronologiques ou séries temporelles. Heureusement, pour l'analyse de ces séries, on peut montrer que les propriétés et procédures d'inférence que nous avons établies sont toujours valables sous les hypothèses moins restrictives :

A1″ $\quad y_t = X_t\beta + e_t$

A2″ $\quad E(e_t|X_t) = 0 \;\Leftrightarrow\; E(y_t|X_t) = X_t\beta$

A3″ $\quad Var(e_t|X_t) = \sigma^2 = V(y_t|X_t)$

A4″ $\quad Cov(e_t, e_s|X_t, X_s) = 0 = Cov(y_t, y_s|X_t, X_s),\ \forall\, t \neq s$

A5″ $\quad \text{rg}(X) = k$

A6″ $\quad$ (optionnel) $\; e_t|X_t \sim N\,(0, \sigma^2) \;\Leftrightarrow\; Y_t|X_t \sim N(X_t\beta, \sigma^2)$

Comme le conditionnement n'est pas fait sur l'ensemble des valeurs de X, mais seulement par rapport à X_t, ces hypothèses n'incluent aucune hypothèse d'exogénéité stricte des variables explicatives. On peut montrer que, sous ces hypothèses A1″ à A5″ (+ optionnellement A6″) les propriétés et procédures d'inférence que nous avons établies sont toujours valables, mais seulement *asymptotiquement*, à titre approximatif pour n grand, et ce même si les y_t sont distribués de façon normale[129]. Ainsi par exemple, sous ces hypothèses, l'estimateur MCO $\hat{\beta}$ n'est plus non biaisé, mais seulement convergent. Cela ne change cependant en pratique rien quand à la façon d'estimer les paramètres du modèle, de calculer des intervalles de confiance ou des tests d'hypothèse, etc...

[128] Dans le modèle dynamique (7.27) qui comprend la variable dépendante retardée comme variable explicative, on ne peut évidemment pas supposer que X est non-stochastique.

[129] L'hypothèse optionnelle A6″ n'est ici plus utile que pour l'intervalle de prévision de y_0 sachant $(x_{02},...x_{0k})$.

Chapitre 8

Variables binaires et modèle logit/probit

Une *variable binaire* (on dit aussi *variable dichotomique* ou *variable muette*)[130] est une variable qui peut prendre seulement deux valeurs distinctes, par convention 0 et 1, et qui est utilisée pour indiquer la présence ou l'absence d'une caractéristique donnée, ou encore la survenance d'un événement particulier. Par exemple :

$$D_i = \begin{cases} 1 & \text{si l'individu } i \text{ est un homme} \\ 0 & \text{sinon} \end{cases}$$

$$D_t = \begin{cases} 1 & \text{si } t = 1940, ..., 1945 \text{ (années de guerre)} \\ 0 & \text{sinon} \end{cases}$$

$$y_i = \begin{cases} 1 & \text{si l'individu } i \text{ est à l'emploi} \\ 0 & \text{sinon} \end{cases}$$

Comme le suggèrent les exemples ci-dessus, une variable binaire peut être utilisée tant comme variable explicative que comme variable dépendante (expliquée).

8.1. Variables explicatives binaires

Les principales utilisations des variables binaires en tant que variables explicatives sont décrites ci-dessous au travers d'exemples.

8.1.1. Comparaison de deux moyennes

On suppose qu'on souhaite estimer les salaires moyens des hommes et des femmes dans une population, et tester s'ils sont ou non différents. Pour cela, on peut utiliser le modèle :

$$y_i = \beta_1 + \beta_2 D_i + e_i \,, \tag{8.1}$$

[130] En anglais, *binary variable* ou *dummy variable*.

où : y_i = le salaire de l'individu i
$D_i = 1$ si l'individu i est un homme, 0 sinon

Pour ce modèle, on a :

$$E(y_i|D_i = 0) = \beta_1 \qquad \text{(i.e., le salaire moyen des femmes)}$$
$$E(y_i|D_i = 1) = \beta_1 + \beta_2 \qquad \text{(i.e., le salaire moyen des hommes)}$$

et un test de l'égalité des salaires moyens des hommes et des femmes revient à tester $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$.

On notera les points suivants :

1- En arrangeant les observations de telle sorte que les n_1 premières regroupent les femmes et les n_2 dernières regroupent les hommes (le nombre total d'observations étant $n = n_1 + n_2$), la matrice des observations X du modèle (8.1) ci-dessus s'écrit :

$$X = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1 \text{ observations} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2 \text{ observations} \end{matrix}$$

On vérifie dès lors aisément (faites-le !) que l'estimateur MCO $\hat{\beta} = (X'X)^{-1}X'Y$ est égal à :

$$\begin{aligned} \hat{\beta} &= \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} n_1 + n_2 & n_2 \\ n_2 & n_2 \end{bmatrix}^{-1} \begin{bmatrix} n_1\bar{y}_1 + n_2\bar{y}_2 \\ n_2\bar{y}_2 \end{bmatrix} \\ &= \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \end{bmatrix} \end{aligned}$$

où $\bar{y}_1$ est le salaire moyen des femmes dans l'échantillon, et $\bar{y}_2$ est le salaire moyen des hommes dans l'échantillon.

2- Au lieu du modèle (8.1), de façon totalement équivalente, on pourrait utiliser le modèle :

$$y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + e_i \,, \tag{8.2}$$

où : $D_{1i} = 1$ si l'individu i est une femme, 0 sinon
$D_{2i} = 1$ si l'individu i est un homme, 0 sinon

Notons que $D_{2i} = 1 - D_{1i}$. Pour ce modèle, on a :

$$E(y_i|D_{1i} = 1, D_{2i} = 0) = \beta_1 \qquad \text{(i.e., le salaire moyen des femmes)}$$
$$E(y_i|D_{1i} = 0, D_{2i} = 1) = \beta_2 \qquad \text{(i.e., le salaire moyen des hommes)}$$

et un test de l'égalité des salaires moyens des hommes et des femmes revient à tester $H_0 : \beta_1 - \beta_2 = 0$ contre $H_1 : \beta_1 - \beta_2 \neq 0$. Pour ce modèle (8.2), en arrangeant les observations comme au point (1) ci-dessus, la matrice des

observations X s'écrit :

$$X = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1 \text{ observations} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2 \text{ observations} \end{matrix}$$

et on vérifie aisément (faites-le !) que l'estimateur MCO $\hat{\beta} = (X'X)^{-1}X'Y$ est égal à :

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}^{-1} \begin{bmatrix} n_1\bar{y}_1 \\ n_2\bar{y}_2 \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix}$$

3- Par contre, on *ne peut pas* utiliser le modèle :

$$y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + e_i \,,$$

car pour ce modèle, les variables de la matrice des observations X sont *parfaitement colinéaires* :

$$X = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1 \text{ observations} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2 \text{ observations} \end{matrix}$$

La première colonne de X est égale à la somme des deux dernières ($D_{1i}+D_{2i} = 1$, $\forall\, i$).

8.1.2. Comparaison de plusieurs moyennes

On suppose qu'on souhaite estimer les salaires moyens des trois régions du pays, et tester s'ils sont ou non différents. Pour cela, on peut utiliser le modèle :

$$y_i = \beta_1 + \beta_2 D_{Bi} + \beta_3 D_{Fi} + e_i \,, \tag{8.3}$$

où : y_i = le salaire de l'individu i
$D_{Bi} = 1$ si l'individu i est bruxellois, 0 sinon
$D_{Fi} = 1$ si l'individu i est flamand, 0 sinon

Pour ce modèle, on a :

$E(y_i|D_{Bi} = 0, D_{Fi} = 0) = \beta_1$ (i.e., le salaire moyen des wallons)
$E(y_i|D_{Bi} = 1, D_{Fi} = 0) = \beta_1 + \beta_2$ (i.e., le salaire moyen des bruxellois)
$E(y_i|D_{Bi} = 0, D_{Fi} = 1) = \beta_1 + \beta_3$ (i.e., le salaire moyen des flamands)

Pour tester l'égalité des salaires moyens dans les trois régions, il suffit de tester $H_0 : \beta_2 = \beta_3 = 0$ contre $H_1 : \beta_2 \neq 0$ et/ou $\beta_3 \neq 0$. Pour tester l'égalité des salaires moyens entre Wallonie et Flandre, on testera $H_0 : \beta_3 = 0$ contre $H_1 : \beta_3 \neq 0$, etc...

On notera encore les points suivants :

1- De façon semblable au cas de la comparaison de deux moyennes, on peut facilement montrer que l'estimateur MCO du modèle (8.3) est simplement égal à $\hat{\beta}_1 = \bar{y}_W$, $\hat{\beta}_2 = \bar{y}_B - \bar{y}_W$ et $\hat{\beta}_3 = \bar{y}_F - \bar{y}_W$, où $\bar{y}_W$, $\bar{y}_B$ et $\bar{y}_F$ désignent, respectivement, le salaire moyen dans l'échantillon des wallons, des bruxellois et des flamands.

2- A nouveau, au lieu du modèle (8.3), de façon totalement équivalente, on pourrait utiliser le modèle (attention au changement de signification des paramètres) :

$$y_i = \beta_1 D_{Wi} + \beta_2 D_{Bi} + \beta_3 D_{Fi} + e_i \,,$$

où $D_{Wi} = 1$ si l'individu i est wallon, 0 sinon, mais *pas* le modèle (pour cause de colinéarité parfaite) :

$$y_i = \beta_1 + \beta_2 D_{Wi} + \beta_3 D_{Bi} + \beta_4 D_{Fi} + e_i$$

8.1.3. Plusieurs critères de classification

On suppose qu'on souhaite estimer le salaire d'un individu en fonction de son sexe et de son niveau d'éducation répartit selon trois niveaux : primaire, secondaire et supérieur. Pour cela, on peut utiliser le modèle :

$$y_i = \beta_1 + \beta_2 D_{Fi} + \beta_3 D_{Pi} + \beta_4 D_{Seci} + e_i \,, \tag{8.4}$$

où : y_i = le salaire de l'individu i
$D_{Fi} = 1$ si l'individu i est une femme, 0 sinon
$D_{Pi} = 1$ si l'individu i possède au plus un diplôme de l'enseignement primaire, 0 sinon
$D_{Seci} = 1$ si l'individu i possède au plus un diplôme de l'enseignement secondaire, 0 sinon

Pour ce modèle, on a :

$E(y_i\|.)$	Primaire ($D_{Pi}=1, D_{Seci}=0$)	Secondaire ($D_{Pi}=0, D_{Seci}=1$)	Supérieur ($D_{Pi}=0, D_{Seci}=0$)
Homme ($D_{Fi}=0$)	$\beta_1 + \beta_3$	$\beta_1 + \beta_4$	β_1
Femme ($D_{Fi}=1$)	$\beta_1 + \beta_2 + \beta_3$	$\beta_1 + \beta_2 + \beta_4$	$\beta_1 + \beta_2$

Sur base de ce modèle, pour tester s'il y a une différence de salaire moyen entre les hommes et les femmes, on testera $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$. Pour tester si les diplômés de l'enseignement supérieur ont un salaire moyen plus élévé que les

diplômés de l'enseignement supérieur, on testera $H_0 : \beta_4 \geq 0$ contre $H_1 : \beta_4 < 0$, etc...

Plusieurs points méritent d'être épinglés :

1- Dans le modèle (8.4) ci-dessus, les estimateurs MCO $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ et $\hat{\beta}_4$ ne correspondent plus, comme précédemment, à des moyennes ou différences de moyennes des salaires observés par catégorie dans l'échantillon (par exemple, $\hat{\beta}_1$ n'est pas égal au salaire moyen dans l'échantillon des hommes ayant un diplôme de l'enseignement supérieur) : il est en effet impossible de capturer 6 moyennes différentes (= nbr. de catégories distinguées) avec seulement 4 paramètres.

2- A nouveau, au lieu du modèle (8.4), de façon totalement équivalente, on pourrait utiliser le modèle (attention au changement de signification des paramètres) :

$$y_i = \beta_1 D_{Hi} + \beta_2 D_{Fi} + \beta_3 D_{Pi} + \beta_4 D_{Seci} + e_i \,,$$

où $D_{Hi} = 1$ si l'individu i est un homme, 0 sinon, mais *pas* le modèle (pour cause de colinéarité parfaite) :

$$y_i = \beta_1 D_{Hi} + \beta_2 D_{Fi} + \beta_3 D_{Pi} + \beta_4 D_{Seci} + \beta_4 D_{Supi} + e_i \,,$$

où $D_{Supi} = 1$ si l'individu i possède un diplôme de l'enseignement supérieur, 0 sinon. En effet, on a :

$$(D_{Hi} + D_{Fi}) = (D_{Pi} + D_{Seci} + D_{Supi}) \;, \quad \forall\, i$$

Le modèle (8.4) ci-dessus suppose que la différence de salaire moyen entre les hommes et les femmes est la même quel que soit le niveau d'éducation, ou ce qui revient au même, que les différences de salaires moyens entre les niveaux d'éducation sont les mêmes quel que soit le sexe. On peut relacher cette hypothèse en considérant le modèle :

$$y_i = \beta_1 + \beta_2 D_{Fi} + \beta_3 D_{Pi} + \beta_4 D_{Seci} + \beta_5 (D_{Fi} D_{Pi}) + \beta_6 (D_{Fi} D_{Seci}) + e_i \,, \qquad (8.5)$$

Pour ce modèle, on a :

$E(y_i\|.)$	Primaire (D_{Pi}= 1, $D_{Seci} = 0$)	Secondaire (D_{Pi}= 0, $D_{Seci} = 1$)	Supérieur (D_{Pi}= 0, $D_{Seci} = 0$)
Homme (D_{Fi}= 0)	$\beta_1 + \beta_3$	$\beta_1 + \beta_4$	β_1
Femme (D_{Fi}= 1)	$\beta_1 + \beta_2 + \beta_3 + \beta_5$	$\beta_1 + \beta_2 + \beta_4 + \beta_6$	$\beta_1 + \beta_2$

Sur base de ce modèle, on peut tester la pertinence du modèle plus restrictif (8.4) en testant $H_0 : \beta_5 = \beta_6 = 0$ contre $H_1 : \beta_5 \neq 0$ et/ou $\beta_6 \neq 0$. Pour tester s'il y a une différence de salaire moyen entre les hommes et les femmes, on testera ici $H_0 : \beta_2 = \beta_5 = \beta_6 = 0$ contre $H_1 : \beta_2 \neq 0$ et/ou $\beta_5 \neq 0$ et/ou $\beta_6 \neq 0$, etc...

On notera pour conclure que dans le modèle (8.5), qui est totalement non contraint (et qui pourrait être reparamétrisé en utilisant une variable binaire pour chacune des 6 catégories distinguées), les estimateurs MCO $\hat{\beta}_j$ correspondent à nouveau à des moyennes ou différences de moyennes des salaires observés par catégorie dans l'échantillon (par exemple, $\hat{\beta}_1$ est ici égal au salaire moyen dans l'échantillon des hommes ayant un diplôme de l'enseignement supérieur).

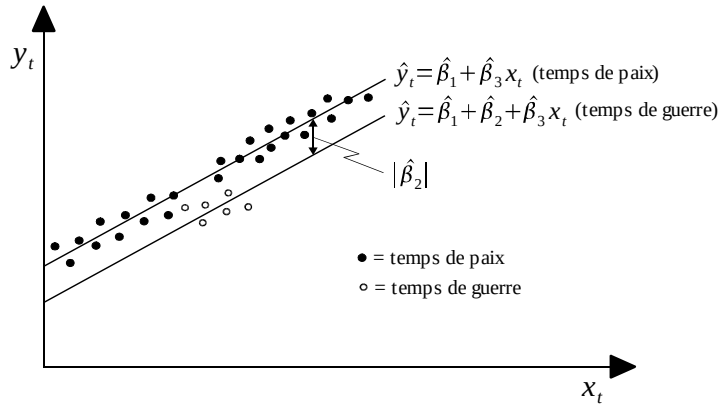
8.1.4. Modifications d'intercept et/ou de pente dans une régression standard

On suppose qu'on souhaite estimer une fonction de consommation au niveau macroéconomique sur la période 1930 - 1950. Durant les années de guerre (1940 - 1945), le niveau de consommation a toutes les chances d'être hors norme. Pour en tenir compte, on peut utiliser le modèle :

$$y_t = \beta_1 + \beta_2 D_t + \beta_3 x_t + e_t \,, \tag{8.6}$$

où : y_t = la consommation de l'année t (par habitant)
x_t = le revenu (PIB) de l'année t (par habitant)
$D_t = 1$ si $t = 1940, 1941, ..., 1945$, 0 sinon

Graphiquement :

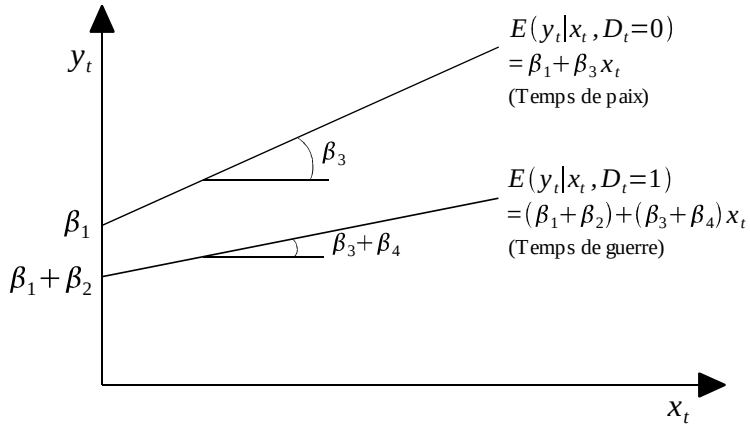


Graphique 46 : La fonction de consommation estimée

Dans le modèle (8.6) ci-dessus, on suppose que les années de guerre ont modifié le niveau de consommation (l'intercept), mais pas la propension marginale à consommer (la pente). Pour tenir compte de ce dernier élément, on peut utiliser le modèle plus général :

$$y_i = \beta_1 + \beta_2 D_t + \beta_3 x_t + \beta_4 D_t x_t + e_t \,, \tag{8.7}$$

Graphiquement :



Graphique 47 : La forme du modèle (8.7)

Sur base du modèle (8.7), on peut tester si la consommation des années de guerre se différencie ou non de la consommation des années de paix en testant H_0 : $\beta_2 = \beta_4 = 0$ contre H_1 : $\beta_2 \neq 0$ et/ou $\beta_4 \neq 0$.

Comme autre exemple de l'utilisation de variabes binaires pour modifier l'intercept et/ou la pente d'une régression standard, supposons qu'on souhaite tester s'il existe une discrimination salariale entre les hommes et le femmes. Pour cela, on peut utiliser le modèle :

$$y_i = \beta_1 + \beta_2 D_{Fi} + \beta_3 Educ_i + \beta_4 Exp_i + e_i\,, \tag{8.8}$$

où : y_i = le salaire de l'individu i
$D_{Fi} = 1$ si l'individu i est une femme, 0 sinon
$Educ_i$ = le nbr. d'années d'étude de l'individu i
Exp_i = le nbr. d'années d'expérience professionnelle de l'individu i

Sur base de ce modèle, pour tester s'il y a discrimination (i.e., si les salaires moyens des hommes et des femmes *pour un même niveau d'éducation et d'expérience professionnelle* sont ou non différents), on testera H_0 : $\beta_2 = 0$ contre H_1 : $\beta_2 \neq 0$.

Le modèle (8.8) ci-dessus suppose que le niveau d'éducation et l'expérience professionnelle ont le même effet sur le salaire moyen des hommes et des femmes. Pour vérifier cette conjecture, et donc la pertinence de ce modèle, on peut utiliser le modèle plus général :

$$y_i = \beta_1 + \beta_2 D_{Fi} + \beta_3 Educ_i + \beta_4 Exp_i + \beta_5 (D_{Fi} Educ_i) + \beta_6 (D_{Fi} Exp_i) + e_i\,, \tag{8.9}$$

Sur base de ce modèle plus général, le test de la conjecture revient à tester H_0 : $\beta_5 = \beta_6 = 0$ contre H_1 : $\beta_5 \neq 0$ et/ou $\beta_6 \neq 0$. Par ailleurs, le test de la présence d'une discrimination salariale — ou à tout le moins de l'existence d'une différence salariale entre les hommes et les femmes de même niveau d'éducation et d'expérience professionnelle — revient à tester H_0 : $\beta_2 = \beta_5 = \beta_6 = 0$ contre H_1 : $\beta_2 \neq 0$ et/ou $\beta_5 \neq 0$ et/ou $\beta_6 \neq 0$.

Les tests de ce dernier type, qui consiste à tester l'égalité de régressions dans deux (ou plusieurs) sous-populations (dans notre exemple, les hommes et les femmes), sont appelés des *tests de Chow*[131]. Dans le cadre de l'analyse de données chronologiques, comme dans le cas du test de $H_0 : \beta_2 = \beta_4 = 0$ contre $H_1 : \beta_2 \neq 0$ et/ou $\beta_4 \neq 0$ dans le modèle (8.7), on parle de *tests de changement structurel*.

8.2. Variables binaires dépendantes

L'utilisation d'une variable binaire comme variable dépendante (expliquée) permet de modéliser la probabilité de posséder une caractéristique donnée, ou la survenance d'un événement particulier, en fonction d'une ou plusieurs variables explicatives.

Supposons qu'on s'intéresse aux chances qu'à un jeune de trouver un emploi dans les six mois suivant sa sortie des études, et ceci en fonction de la longueur de ses études. C'est notre relation d'intérêt.

Comme dans le cas du modèle de régression (simple ou multiple), on cherche une *contrepartie empirique* de la relation d'intérêt, une contrepartie empirique prenant la forme d'un *modèle probabiliste paramétré*, et on regarde les données dont on dispose comme des *réalisations particulières* des variables aléatoires de ce modèle, pour une valeur particulière des paramètres du modèle.

Pour examiner les liens existant entre le fait de trouver un emploi dans les six mois de la sortie des études et le niveau d'éducation, il est naturel de s'appuyer sur des données en coupe obtenues par tirages aléatoires d'individus dans la population des jeunes sortant des études au cours d'une année civile donnée.

Notons y une variable binaire qui prend la valeur 1 si un jeune sortant des études trouve un emploi dans les six mois, et 0 sinon, et x le niveau d'éducation (nombre d'années d'études) du jeune.

Au travers de l'épreuve aléatoire 'tirer un jeune au hasard dans la population et noter la valeur de y (1 s'il trouve du travail dans les six mois de la sortie de ses études, 0 sinon) et de x (son niveau d'éducation)', on peut représenter la population par une *distribution de probabilité jointe* $f(y, x)$, qui correspond à la distribution de fréquence des couples de variables (y, x) dans la population.

Lorsqu'on cherche à expliquer y en fonction de x, l'information pertinente est concentrée dans la *distribution conditionnelle* $f(y|x)$ qui, pour chaque valeur de x, correspond à la distribution de fréquence des différentes valeurs de y dans la population. Comme y est une variable binaire, cette distribution conditionnelle est simplement une loi (conditionnelle) de Bernoulli $\mathcal{B}(p(x))$, dont la fonction de densité est donnée par :

$$f(y|x) = p(x)^y (1 - p(x))^{1-y}, \quad \forall\, y = 0, 1\,,$$

[131] Chow G.C. (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions", *Econometrica*, 52, p. 221-222.

où $p(x)$ est, pour chaque valeur de x, la probabilité que y soit égal à 1 :

$$I\!P(y = 1|x) = f(1|x) = p(x)\,,$$

tandis que la probabilité que y soit égal à 0 est donnée par :

$$I\!P(y = 0|x) = 1 - I\!P(y = 1|x) = f(0|x) = 1 - p(x)$$

La distribution conditionnelle $f(y|x)$ — i.e., la loi (conditionnelle) de Bernoulli $\mathcal{B}(p(x))$ — est entièrement déterminée par la probabilité conditionnelle $I\!P(y = 1|x) = p(x)$ qui, pour chaque valeur de x, correspond à la proportion (ou fréquence) des y qui prennent la valeur 1 dans la population, autrement dit, pour notre exemple, la proportion des jeunes qui trouvent un emploi dans les 6 mois de la sortie de leurs études. Cette probabilité conditionnelle $I\!P(y = 1|x) = p(x)$ constitue la contrepartie empirique de la relation d'intérêt.

La probabilité conditionnelle $I\!P(y = 1|x) = p(x)$ définit un modèle probabiliste de la relation d'intérêt. On obtient un *modèle probabiliste paramétré* de la relation d'intérêt si on suppose une forme fonctionnelle, dépendant de paramètres, pour $p(x)$. De façon générale :

$$I\!P(y = 1|x) = p(x, \beta)\,,$$

où β est un vecteur de paramètres. La seule restriction que doit satisfaire $p(x, \beta)$ est qu'elle doit toujours être comprise entre 0 et 1, quels que soient x et β[132].

Nous avons raisonné ci-dessus en supposant, pour faire simple, qu'il n'y avait qu'une variable explicative. Lorsqu'on considère plusieurs variables explicatives $(x_2, ..., x_k)$, la distribution conditionnelle $f(y|x_2, ..., x_k)$ pertinente est une loi (conditionnelle) de Bernoulli $\mathcal{B}(p(x_2, ..., x_k))$, dont la fonction de densité est donnée par :

$$f(y|x_2, ..., x_k) = p(x_2, ..., x_k)^y (1 - p(x_2, ..., x_k))^{1-y}, \quad \forall\, y = 0, 1\,,$$

où $p(x_2, ..., x_k)$ est, pour chaque valeur de $(x_2, ..., x_k)$, la probabilité que y soit égal à 1 :

$$I\!P(y = 1|x_2, ..., x_k) = f(1|x_2, ..., x_k) = p(x_2, ..., x_k)\,,$$

tandis que la probabilité que y soit égal à 0 est donnée par :

$$I\!P(y = 0|x_2, ..., x_k) = f(0|x_2, ..., x_k) = 1 - p(x_2, ..., x_k)$$

Comme dans le cas simple, la distribution conditionnelle $f(y|x_2, ..., x_k)$ est entièrement déterminée par la probabilité conditionnelle $I\!P(y = 1|x_2, ..., x_k) = p(x_2, ..., x_k)$ qui, pour chaque valeur des variables $(x_2, ..., x_k)$, correspond à la proportion (ou fréquence) des y qui prennent la valeur 1 dans la population. Cette probabilité conditionnelle $I\!P(y = 1|x_2, ..., x_k) = p(x_2, ..., x_k)$ définit un modèle probabiliste de la relation d'intérêt, et un modèle probabiliste paramétré de cette relation d'intérêt est obtenu en choisissant une forme fonctionnelle, dépendant de paramètres, pour

[132] Car une probabilité est toujours comprise entre 0 et 1.

$p(x_2, ..., x_k)$. De façon générale :

$$I\!P(y = 1|x_2, ..., x_k) = p(x_2, ..., x_k; \beta)\,,$$

où β est un vecteur de paramètres. A nouveau, la seule restriction que doit satisfaire $p(x_2, ..., x_k; \beta)$ est qu'elle doit toujours être comprise entre 0 et 1, quels que soient $(x_2, ..., x_k)$ et β.

Si les observations sont obtenues par tirages aléatoires d'individus dans la population et que le modèle est correctement spécifié (i.e., la forme fonctionnelle choisie est correcte), chaque observation (y_i, X_i), où $X_i = (x_2, ..., x_k)$, est telle que :

$$I\!P(y_i = 1|X_i) = p(X_i, \beta) \tag{8.10}$$

et

$$f(y_i|X_i; \beta) = p(X_i, \beta)^{y_i}(1 - p(X_i, \beta))^{1-y_i}, \quad i = 1, ..., n\,, \tag{8.11}$$

où β est un vecteur de paramètres inconnus à estimer et, avant observation, y_i et X_i sont des variables aléatoires.

8.2.1. Le modèle de probabilité linéaire

Une caractéristique remarquable de la distribution conditionnelle (8.11) — i.e., la loi (conditionnelle) de Bernoulli $\mathcal{B}(p(X_i, \beta))$ — est que $p(X_i, \beta)$ est non seulement égal à la probabilité conditionnelle $I\!P(y_i = 1|X_i)$ que y_i soit égal à 1 sachant X_i, mais aussi à l'espérance conditionnelle $E(y_i|X_i)$ de y_i sachant de X_i. En effet, on a :

$$\begin{aligned} E(y_i|X_i) &= \sum_{y_i} y_i f(y_i|X_i; \beta) = 0 \times (1 - p(X_i, \beta)) + 1 \times p(X_i, \beta) \\ &= p(X_i, \beta) = I\!P(y_i = 1|X_i) \end{aligned}$$

Ainsi, si on prend pour $p(X_i, \beta)$ la simple fonction linéaire $p(X_i, \beta) = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}$, on peut réécrire le modèle, qui est alors appelé *modèle de probabilité linéaire*, sous la forme du modèle de régression linéaire :

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i \\ &= X_i\beta + e_i\,, \qquad\qquad i = 1, ..., n\,, \end{aligned} \tag{8.12}$$

où X_i est redéfini de façon à inclure une constante[133] : $X_i = \begin{bmatrix} 1 & x_{i2} & \cdots & x_{ik} \end{bmatrix}$.

Si cette forme fonctionnelle linéaire est correcte, les hypothèses standard A1, A2 et A5 (ou A1′, A2′ et A5′ si X est stochastique) de la régression (8.12) sont satisfaites, de sorte qu'un estimateur non biaisé du vecteur de paramètres β est simplement donné par l'estimateur MCO $\hat{\beta} = (X'X)^{-1}X'Y$ de cette régression.

[133] Dans la suite, X_i sera toujours défini de cette façon, y compris lorsque X_i représente l'ensemble des variables explicatives, car les ensembles conditionnants $(x_2, ..., x_k)$ et $(1, x_2, ..., x_k)$ sont équivalents (l'ajout de la constante n'apporte aucune information complémentaire).

Cette façon de procéder — choix d'une forme fonctionnelle linéaire pour $p(X_i, \beta)$ et estimation des paramètres par l'estimateur MCO standard — pose toutefois deux problèmes :

1- le choix de la forme linéaire $X_i\beta$ pour $p(X_i, \beta)$ ne garantit pas que la probabilité conditionnelle $I\!P(y_i = 1|X_i)$ soit toujours comprise entre 0 et 1. Pour certains X_i, elle peut très bien être inférieure à 0, ou supérieure à 1. Vu autrement, la forme fonctionnelle linéaire $p(X_i, \beta) = X_i\beta$ suppose que l'effet marginal des différentes variables x_{ij} sur la probabilité que y soit égal à 1 est constant, i.e., $\frac{\partial p(X_i,\beta)}{\partial x_{ij}} = \beta_j$ (une constante). C'est impossible, car si cet effet marginal est constant, lorsqu'on augmente x_{ij}, on finira forcément par obtenir une probabilité supérieure à 1, et lorsqu'on diminue x_{ij}, on finira de même forcément par obtenir une probabilité inférieure à 0. Cet effet marginal ne peut donc pas être constant.

2- Si les hypothèses standard A1, A2 et A5 (ou A1′, A2′ et A5′ si X est stochastique) de la régression (8.12) sont satisfaites, il n'en va pas de même de l'hypothèse A3 (ou A3′ si X est stochastique) d'homoscédasticité[134]. En effet, on a :

$$\begin{aligned} Var(y_i|X_i) &= \sum_{y_i}(y_i - E(y_i|X_i))^2 f(y_i|X_i;\beta) \\ &= (0 - p(X_i,\beta))^2 \times (1 - p(X_i,\beta)) + (1 - p(X_i,\beta))^2 \times p(X_i,\beta) \\ &= p(X_i,\beta)\,(1 - p(X_i,\beta)) = X_i\beta(1 - X_i\beta)\,, \end{aligned}$$

autrement dit, la variance conditionnelle $Var(y_i|X_i)$ n'est pas une constante, mais une fonction de X_i.
Comme nous l'avons vu à la Section 7.4.3, la violation de l'hypothèse d'homoscédasticité n'empêche pas l'estimateur MCO standard d'être non biaisé, et il peut donc toujours être utilisé. Par contre, (a) cet estimateur n'est plus celui qui a la plus petite (au sens matriciel) matrice de variance - covariance parmi les estimateurs linéaires sans biais de β et (b) les procédures d'inférence standard qui lui sont associées ne sont plus valables. On sait qu'on peut résoudre le problème de la validité des procédures d'inférence en utilisant l'estimateur robuste à l'hétéroscédasticité $\hat{V}_{HC}(\hat{\beta})$ de la matrice de variance - covariance de l'estimateur MCO en lieu et place de l'estimateur standard $\hat{V}(\hat{\beta}) = \hat{s}^2\,(X'X)^{-1}$. Ce problème n'en est donc pas vraiment un. Le fait que l'estimateur MCO n'est pas optimal en termes de précision d'estimation — on dit qu'il n'est pas *efficace* — est plus ennuyeux car ce type de modèle avec une variable binaire dépendante requiert typiquement un échantillon assez grand pour obtenir des estimations précises. Utiliser un estimateur efficace est donc dans le présent contexte assez important.

Ces limitations n'empêchent pas que le modèle de probabilité linéaire puisse utilement être utilisé dans certaines situations[135]. On lui préfère cependant généralement les modèles logit et probit développés ci-dessous, plus complexe à interpréter

[134] Les observations étant supposées être obtenues par tirages aléatoires — avec remise, ou sans remise, si l'échantillon est petit par rapport à la population — d'individus dans la population, elles sont par construction indépendantes, et donc non-corrélées. L'hypothèse A4 (ou A4′ si X est stochastique) de non-corrélation est donc sensée être automatiquement satisfaite.

[135] Pour une discussion de ce point, voir Wooldridge (2010), Ch 15, dont la référence complète est donnée dans le préambule des notes.

et à estimer, mais qui n'ont pas ces limitations.

8.2.2. Les modèles logit et probit I : spécification

Le modèle logit et le modèle probit sont tous les deux un cas particulier du modèle général :

$$\begin{aligned} I\!P(y_i = 1|X_i) &= p(X_i, \beta) = G\left(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}\right) \\ &= G\left(X_i\beta\right) \qquad\qquad i = 1, ..., n\,, \end{aligned} \tag{8.13}$$

où $G(.)$ est une fonction dont les valeurs sont toujours comprises entre 0 et 1 : $0 < G(z) < 1$, pour tout z.

Dans ce modèle, la probabilité conditionnelle $I\!P(y_i = 1|X_i)$ dépend de X_i uniquement au travers de l'*index* $X_i\beta$, la fonction $G(.)$ étant la *fonction de lien* entre cet index et la probabilité $I\!P(y_i = 1|X_i)$. La fonction de lien $G(.)$ assure que la probabilité $I\!P(y_i = 1|X_i)$ est toujours comprise entre 0 et 1[136]. On notera que l'index $X_i\beta$ est supposé être une fonction linéaire, mais uniquement une fonction linéaire dans les *paramètres*, pas nécessairement dans les *variables* : les variables x_{ij} peuvent être des transformations — par exemple le logarithme — des variables originales, ou inclure des carrés et des produits croisés de variables comme dans une régression polynomiale. L'index $X_i\beta$ peut également inclure des variables explicatives binaires.

Diverses fonctions non-linéaires peuvent être utilisées pour la fonction de lien $G(.)$. Dans le *modèle logit*, la fonction $G(.)$ est la fonction logistique :

$$G(z) = \frac{e^z}{1 + e^z} \tag{8.14}$$

Cette fonction logistique est la fonction de répartition[137] d'une variable aléatoire distribuée selon une loi logistique standard[138]. Dans le *modèle probit*, la fonction $G(.)$ est la fonction de répartition d'une variable aléatoire normale standardisée (qui ne possède pas de forme analytique explicite) :

$$G(z) = \int_{-\infty}^{z} \phi(x)dx\,, \tag{8.15}$$

où $\phi(x)$ est la fonction de densité de la loi normale standardisée : $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.

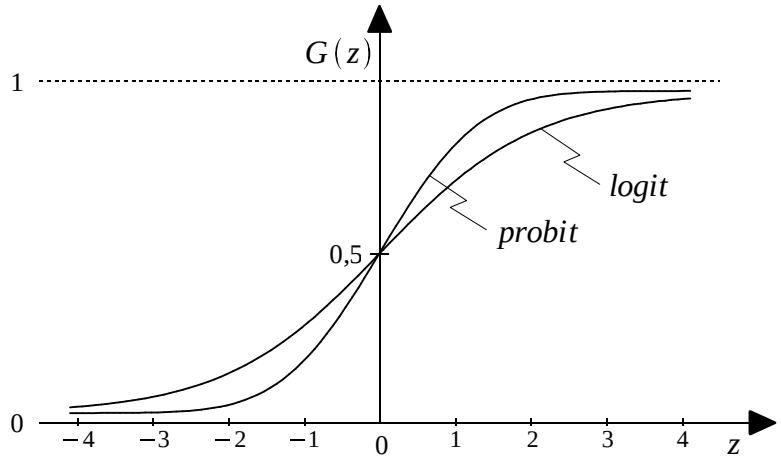
---

[136] Dans le modèle de probabilité linéaire, $G(z) = z$, d'où le fait que la probabilité $I\!P(y_i = 1|X_i)$ n'est pas nécessairement comprise entre 0 et 1.

[137] Pour rappel, la fonction de répartition $F(x)$ d'une variable aléatoire X est définie par $F(x) = I\!P(X \leq x)$.

[138] Une variable aléatoire X est distribuée selon une loi logistique standard si sa fonction de densité est donnée par $f(x) = \frac{e^x}{(1+e^x)^2}$. Cette fonction de densité ressemble à celle d'une loi normale standardisée (forme en cloche, centrée en zéro et symétrique), mais avec une dispersion (variance) plus importante (cf. le Graphique 49 infra).

Graphiquement :



Graphique 48 : La fonction de lien $G(z)$ des modèles logit et probit

Comme on peut le voir, tant pour la fonction de lien logit que probit, $G(z)$ est une fonction strictement croissante de z, $G(0) = 0,5$, $G(z) \to 0$ lorsque $z \to -\infty$, et $G(z) \to 1$ lorsque $z \to +\infty$.

Lorsque, comme dans le cas du modèle logit et du modèle probit, la fonction de lien $G(z)$ correspond à la fonction de répartition d'une variable aléatoire dont la fonction de densité est symétrique par rapport à zéro, le modèle général (8.13) peut être interprété comme dérivé du *modèle à variable latente* :

$$\begin{cases} y_i^* = X_i\beta + e_i \\ y_i = 1 \text{ si } y_i^* > 0, \text{ 0 sinon} \end{cases}, \tag{8.16}$$

où y_i^* est une variable non observable, ou *latente*, qui est supposée être égale à une fonction linéaire $X_i\beta$ des variables explicatives X_i plus un terme d'erreur e_i indépendant de X_i, et y_i une variable binaire observable, qui prend la valeur 1 si $y_i^* > 0$, 0 sinon.

Un exemple classique d'un tel modèle est un modèle de décision basé sur une fonction d'utilité aléatoire. Supposons que y_i^* désigne l'utilité d'un individu i lorsqu'il se rend au travail en transport en commun (plutôt qu'avec un autre moyen de transport). Cette utilité est supposée dépendre, d'une part, au travers de $X_i\beta$, d'un certain nombre de variables telles que le temps de trajet supplémentaire (qui peut être négatif) que représente le fait de prendre les transports en commun (plutôt qu'un autre moyen de transport), du sexe de l'indivdu, de son âge, etc..., et d'autre part, d'un terme aléatoire e_i qui représente les préférences personnelles de l'individu i. La présence de ce terme aléatoire e_i fait que l'utilité y_i^* est aléatoire. L'utilité y_i^* de l'individu i lorsqu'il se rend au travail en transport en commun (plutôt qu'avec un autre moyen de transport) n'est pas observable. Mais son choix, représenté par la variable binaire y_i, de prendre ou non les transports en commun l'est, et on suppose qu'il prend les transports en commun ($y_i = 1$) si son utilité y_i^* est positive, et qu'il prend un autre moyen de transport ($y_i = 0$) sinon.

Si la distribution du terme d'erreur e_i — qui dans notre exemple représente les préférences personnelles de l'individu i — est symétrique par rapport à zéro, et que l'on désigne par $G(z)$ la fonction de répartition de cette distribution, i.e., $G(z) = I\!P(e_i \leq z)$, on a :

$$1 - G(-z) = G(z) \qquad \text{(car la distrib. de } e_i \text{ est symétrique)}\,,$$

de sorte qu'on obtient :

$$\begin{aligned} I\!P(y_i = 1|X_i) &= I\!P(y_i^* > 0|X_i) = I\!P(X_i\beta + e_i > 0|X_i) \\ &= I\!P(e_i > -X_i\beta|X_i) = 1 - I\!P(e_i \leq -X_i\beta|X_i) \\ &= 1 - I\!P(e_i \leq -X_i\beta) \quad \text{(car } e_i \text{ est indépendant de } X_i) \\ &= 1 - G(-X_i\beta) = G(X_i\beta)\,, \end{aligned}$$

ce qui est exactement le modèle (8.13). Dans le cas du modèle logit, la distribution du terme d'erreur e_i correspond à une loi logistique standard, et dans le cas du modèle probit, à une loi normale standardisée.

Dans la plupart des applications empiriques des modèles logit et probit, on est avant tout intéressé par l'effet marginal des différentes variables explicatives x_{ij} — les autres variables étant maintenues constantes — sur la probabilité $I\!P(y_i = 1|X_i)$. La formulation du modèle dans les termes du modèle à variable latente (8.16) pourrait laisser croire qu'on est avant tout intéressé par l'effet marginal des différentes variables explicatives x_{ij} sur la variable latente y_i^*. Ce n'est généralement pas le cas. La variable latente y_i^* est, le plus souvent, une construction de l'esprit (comme dans le modèle de décision basé sur une fonction d'utilité aléatoire) : elle est typiquement non observable et possède rarement une unité de mesure bien définie, de sorte que les valeurs précises des paramètres β_1,..., β_k n'ont en elles-mêmes que peu d'intérêt. Comme nous le verrons ci-dessous, le signe ou l'éventuelle nullité de ces paramètres est par contre important. En bref, sauf cas (très) particulier, il ne faut pas accorder trop d'importance à l'interprétation en termes de variable latente des modèles logit et probit.

Le calcul de l'effet marginal des différentes variables explicatives x_{ij} — les autres variables étant maintenues constantes — sur la probabilité $I\!P(y_i = 1|X_i)$ est rendu compliqué par la forme non-linéaire (8.13) des modèles logit et probit. L'effet marginal de la variable x_{ij} sur la probabilité $I\!P(y_i = 1|X_i)$ est donné par :

$$\frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{ij}} = g(X_i\beta)\beta_j\,, \tag{8.17}$$

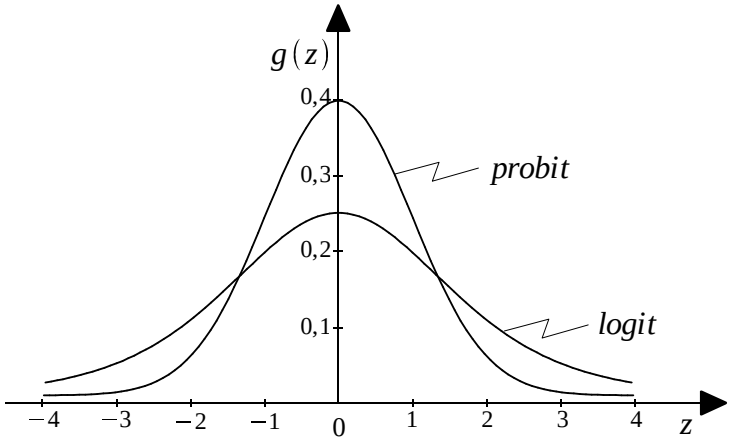
où $g(z) = \frac{dG(z)}{dz}$ est donné, pour le modèle logit, par :

$$g(z) = \frac{e^z}{(1 + e^z)^2}\,, \tag{8.18}$$

et pour le modèle probit, par :

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,, \tag{8.19}$$

càd. par la fonction de densité de, respectivement, la loi logistique standard et la loi normale standardisée. Graphiquement :



Graphique 49 : La fonction $g(z) = \frac{dG(z)}{dz}$ des modèles logit et probit

Comme la fonction $g(.)$ est toujours positive, l'effet marginal $\frac{\partial I\!P(y_i=1|X_i)}{\partial x_{ij}}$ est toujours du même signe que β_j. Son ampleur varie cependant en fonction de *toutes* les variables explicatives X_i au travers de $g(X_i\beta)$. Tant dans le modèle logit que dans le modèle probit, pour β_j fixé, l'effet marginal est maximum lorsque $X_i\beta = 0$, soit lorsque $I\!P(y_i = 1|X_i) = 0,5$, et décroît lorsque $|X_i\beta|$ grandit, soit lorsque $I\!P(y_i = 1|X_i)$ s'écarte de $0,5$ (i.e., tend vers 0 ou 1).

Lorsque l'index $X_i\beta$ contient des variables transformées et/ou des polynômes, la formule (8.17) de l'effet marginal doit être adaptée. Par exemple, pour le modèle :

$$I\!P(y_i = 1|X_i) = G\left(\beta_1 + \beta_2 x_{i2} + \beta_3 \ln x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i4}^2\right) ,$$

on a[139] :

$$\begin{aligned}
\frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{i2}} &= g(X_i\beta)\beta_2 \\
\frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{i3}} &= g(X_i\beta)\frac{\beta_3}{x_{i3}} \\
\frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{i4}} &= g(X_i\beta)(\beta_4 + 2\beta_5 x_{i4})
\end{aligned}$$

On notera finalement que pour des variables explicatives x_{ij} binaires ou discrètes (i.e., prenant un petit nombre de valeurs entières, comme par exemple le nombre d'enfants d'un ménage), la formule (8.17) de l'effet marginal est une approximation qui peut être grossière, en particulier lorsque la valeur de β_j est grande et/ou la valeur de l'index $X_i\beta$ est éloignée de 0 (et donc $I\!P(y_i = 1|X_i)$ éloignée de $0,5$). Il est dans ce cas préférable de calculer l'effet marginal exact. Par exemple, pour le

[139] Notons que si on s'intéresse à la variation (absolue) de $I\!P(y_i = 1|X_i)$ pour une variation *relative* (plutôt qu'absolue) de x_{i3}, on calculera $\frac{\partial I\!P(y_i=1|X_i)}{\partial \ln x_{i3}} = g(X_i\beta)\beta_3$ (plutôt que $\frac{\partial I\!P(y_i=1|X_i)}{\partial x_{i3}} = g(X_i\beta)\frac{\beta_3}{x_{i3}}$).

modèle :

$$I\!P(y_i = 1|X_i) = G\left(\beta_1 + \beta_2 D_i + \beta_3 x_i\right) ,$$

où D_i est une variable binaire, plutôt que l'effet marginal approximatif calculé sur base de la dérivée (8.17) évaluée en $D_i = 0$:

$$\frac{\partial I\!P(y_i = 1|X_i)}{\partial D_i} = g(X_i\beta)\beta_2 = g(\beta_1 + \beta_3 x_i)\beta_2 ,$$

on calculera l'effet marginal exact :

$$\begin{aligned} \frac{\Delta I\!P(y_i = 1|X_i)}{\Delta D_i} &= G(X_i^1\beta) - G(X_i^0\beta) \\ &= G(\beta_1 + \beta_2 + \beta_3 x_i) - G(\beta_1 + \beta_3 x_i) , \end{aligned}$$

où $X_i^1 = [\ 1 \quad 1 \quad x_i\ ]$ et $X_i^0 = [\ 1 \quad 0 \quad x_i\ ]$. De même, pour calculer l'effet marginal exact du passage de $x_{i2} = c$ à $x_{i2} = c + 1$ de la variable discrète x_{i2} dans le modèle :

$$I\!P(y_i = 1|X_i) = G\left(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\right) ,$$

plutôt que l'effet marginal approximatif calculé sur base de la dérivée (8.17) évaluée en $x_{i2} = c$:

$$\frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{i2}} = g(X_i\beta)\beta_2 = g(\beta_1 + \beta_2 c + \beta_3 x_{i3})\beta_2 ,$$

on calculera l'effet marginal exact :

$$\begin{aligned} \frac{\Delta I\!P(y_i = 1|X_i)}{\Delta x_{i2}} &= G(X_i^{c+1}\beta) - G(X_i^c\beta) \\ &= G(\beta_1 + \beta_2(c+1) + \beta_3 x_{i3}) - G(\beta_1 + \beta_2 c + \beta_3 x_{i3}) , \end{aligned}$$

où $X_i^{c+1} = [\ 1 \quad (c+1) \quad x_i\ ]$ et $X_i^c = [\ 1 \quad c \quad x_i\ ]$.

8.2.3. Les modèles logit et probit II : estimateur du maximum de vraisemblance

On suppose que les observations sont constituées de données en coupe obtenues par tirages aléatoires d'individus dans une population, ou à tout le moins qu'elles peuvent, dans une perspective de modélisation, être regardées comme telles. Si on suppose par ailleurs que le modèle logit ou probit est correctement spécifié (i.e., la forme fonctionnelle choisie est correcte), alors les observations (y_i, X_i) sont par hypothèse indépendantes d'un individu à l'autre, et sont telles que :

$$I\!P(y_i = 1|X_i) = G\left(X_i\beta\right) \tag{8.20}$$

et

$$f(y_i|X_i;\beta) = G\left(X_i\beta\right)^{y_i}\left(1 - G\left(X_i\beta\right)\right)^{1-y_i}, \qquad i = 1, ..., n , \tag{8.21}$$

où β est un vecteur de paramètres inconnus à estimer et la fonction de lien $G(.)$ est donnée par (8.14) dans le cas du modèle logit et (8.15) dans le cas du modèle probit.

Les observations étant par hypothèse indépendantes d'un individu à l'autre, la densité jointe des observations $(y_1, ..., y_n)$ sachant $(X_1, ..., X_n)$, appelée *vraisemblance (conditionnelle)*, peut être décomposée comme suit :

$$\begin{aligned} & f(y_1, ..., y_n | X_1, ..., X_n; \beta) \\ = \; & f(y_1|X_1;\beta) \times ... \times f(y_n|X_n;\beta) = \prod_{i=1}^{n} f(y_i|X_i;\beta)\,, \end{aligned}$$

où $f(y_i|X_i;\beta)$ est donné par la fonction de densité (8.21).

En prenant le logarithme de la densité jointe des observations, on obtient la *fonction de log-vraisemblance (conditionnelle)* de l'échantillon :

$$\begin{aligned} L(\beta) &= \ln f(y_1, ..., y_n | X_1, ..., X_n; \beta) \\ &= \sum_{i=1}^{n} \ln f(y_i|X_i;\beta) \\ &= \sum_{i=1}^{n} \left[y_i \ln G\left(X_i\beta\right) + (1-y_i) \ln(1 - G\left(X_i\beta\right)) \right] \end{aligned}$$

L'estimateur du maximum de vraisemblance (MV) $\hat{\beta}$ est défini par la valeur du vecteur de paramètres β qui maximise la vraisemblance[140], ou ce qui revient au même[141], la log-vraisemblance de l'échantillon :

$$\begin{aligned} \hat{\beta} &= \text{Argmax}_\beta L(\beta) = \text{Argmax}_\beta \sum_{i=1}^{n} \ln f(y_i|X_i;\beta) \\ &= \text{Argmax}_\beta \sum_{i=1}^{n} \left[y_i \ln G\left(X_i\beta\right) + (1-y_i) \ln(1 - G\left(X_i\beta\right)) \right] \end{aligned} \tag{8.22}$$

Si la fonction de lien $G(.)$ est donnée par (8.14), $\hat{\beta}$ est appelé l'*estimateur logit*. Si la fonction de lien $G(.)$ est donnée par (8.15), $\hat{\beta}$ est appelé l'*estimateur probit*.

Le problème d'optimisation (8.22) n'a pas de solution analytique. L'estimateur MV $\hat{\beta}$ ne peut être obtenu que numériquement, en utilisant un algorithme d'optimisation approprié. Les logiciels économétriques, en particulier GRETL, s'acquittent très bien et très facilement de cette tâche.

Sous des conditions de régularité générales et si le modèle est bien correctement spécifié, on peut montrer que l'estimateur MV $\hat{\beta}$ est un estimateur *convergent* et *asymptotiquement normal* de β. Formellement :

$$\hat{\beta} \xrightarrow{p} \beta \tag{8.23}$$

[140] Càd. la valeur de β qui rend la plus élevée la probabilité d'observation de l'échantillon dont on dispose. Autrement dit, la valeur de β pour laquelle l'échantillon dont on dispose est le plus probable d'être observé.

[141] Le logarithme étant une fonction strictement croissante, la vraisemblance et la log-vraisemblance ont par construction le même maximum par rapport à β.

et[142]

$$\left[V(\hat{\beta})\right]^{-\frac{1}{2}}\left(\hat{\beta}-\beta\right)\xrightarrow{d}N(0,I)\,, \tag{8.24}$$

où[143] :

$$\begin{aligned} V(\hat{\beta}) &= -\left[\sum_{i=1}^{n}E\left(\frac{\partial^2\ln f(y_i|X_i;\beta)}{\partial\beta\partial\beta'}\right)\right]^{-1} \\ &= \left[\sum_{i=1}^{n}E\left(\frac{g(X_i\beta)^2X_i'X_i}{G\left(X_i\beta\right)\left(1-G\left(X_i\beta\right)\right)}\right)\right]^{-1} \end{aligned} \tag{8.25}$$

soit, exprimé sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand :

$$\hat{\beta}\approx N(\beta,V(\hat{\beta})) \tag{8.26}$$

Dans l'expression (8.25) de la matrice de variance - covariance $V(\hat{\beta})$ de l'estimateur MV $\hat{\beta}$, les fonctions $G(.)$ et $g(.)$ sont données, pour le modèle logit, par respectivement (8.14) et (8.18), et pour le modèle probit, par respectivement (8.15) et (8.19).

De façon semblable au cas du modèle de régression, on peut montrer que la matrice de variance - covariance $V(\hat{\beta})$ de l'estimateur MV $\hat{\beta}$ sera d'autant plus petite (au sens matriciel), et donc la précision d'estimation du vecteur de paramètres β d'autant plus grande, que :

1- les variables explicatives x_{ij} sont dispersées,

2- la taille n de l'échantillon est grande,

3- les variables explicatives x_{ij} sont peu corrélées.

On notera que l'estimateur MV $\hat{\beta}$ n'est pas le seul estimateur possible du vecteur de paramètres β du modèle. Mais c'est le meilleur. En effet, chaque fois que, comme c'est le cas ici , on cherche à estimer les paramètres d'un modèle probabiliste qui spécifie, pour des observations (y_i, X_i) obtenues par échantillonnage aléatoire, la distribution conditionnelle de y_i sachant X_i au travers d'une fonction de densité $f(y_i|X_i;\beta)$ et que ce modèle est correctement spécifié, on peut montrer que l'estimateur MV de β, définit comme $\hat{\beta}=\text{Argmax}_\beta\sum_{i=1}^{n}\ln f(y_i|X_i;\beta)$, fournit toujours un estimateur non seulement convergent et asymptotiquement normal, mais aussi *efficace*, càd. un estimateur ayant la plus petite (au sens matriciel) matrice de variance - covariance — qui est toujours donnée par $V(\hat{\beta})=-\left[\sum_{i=1}^{n}E\left(\frac{\partial^2\ln f(y_i|X_i;\beta)}{\partial\beta\partial\beta'}\right)\right]^{-1}$ — parmi tous les estimateurs convergents et asymptotiquement normaux de β. Autrement dit, on ne peut pas trouver un meilleur estimateur (i.e., un estimateur plus précis) que

[142] Le résultat de normalité asymptotique (8.24) peut de façon équivalente être exprimé comme : $\sqrt{n}(\hat{\beta}-\beta)\xrightarrow{d}N(0,\Sigma)$, où $\Sigma=nV(\hat{\beta})$.

[143] La matrice $\frac{\partial^2\ln f(y_i|X_i;\beta)}{\partial\beta\partial\beta'}$ est la matrice hessienne de la fonction $\ln f(y_i|X_i;\beta)$, càd. une matrice carrée dont les différents éléments (i,j) sont égaux aux dérivées secondes $\frac{\partial^2\ln f(y_i|X_i;\beta)}{\partial\beta_i\partial\beta_j}$.

l'estimateur MV, parmi tous les estimateurs convergents et asymptotiquement normaux de β. Cette propriété générale d'efficacité de l'estimateur MV peut être intuitivement vue comme l'analogue, mais dans un cadre bien plus général, du théorème Gauss-Markov pour le modèle de régression linéaire.

En s'appuyant sur le résultat approximatif[144] (8.26) de distribution d'échantillonnage de $\hat{\beta}$, on peut, de la même façon que dans le cas du modèle de régression linéaire, construire des intervalles de confiance et des tests d'hypothèse relatifs à β, ainsi que des intervalles de prévision (cf. Section 8.2.4 infra). Dans cette perspective, le seul ingrédient encore manquant est un estimateur convergent de la matrice de variance - covariance $V(\hat{\beta})$. On peut montrer qu'un tel estimateur est simplement donné par[145] :

$$\hat{V}(\hat{\beta}) = \left[\sum_{i=1}^{n} \frac{g(X_i\hat{\beta})^2 X_i' X_i}{G(X_i\hat{\beta})(1 - G(X_i\hat{\beta}))}\right]^{-1} \tag{8.27}$$

Des éléments diagonaux $\hat{Var}(\hat{\beta}_j)$ $(j = 1, ..., k)$ de cet estimateur $\hat{V}(\hat{\beta})$, on obtient des estimateurs convergents des écarts-types $s.e.(\hat{\beta}_j)$ des différents $\hat{\beta}_j$ en prenant :

$$s.\hat{e}.(\hat{\beta}_j) = \sqrt{\hat{Var}(\hat{\beta}_j)}\,, \quad j = 1, ..., k$$

Tous les logiciels économétriques, en particulier GRETL, calculent et reportent automatiquement $\hat{V}(\hat{\beta})$ et les écart-types qui en découlent $s.\hat{e}.(\hat{\beta}_j)$.

Une fois le modèle estimé, on peut évaluer la probabilité $I\!P(y_i = 1|X_i)$ pour n'importe quelle valeur de X_i. A cette fin, il suffit de remplacer dans l'expression (8.20) le vecteur de paramètres inconnus β par son estimateur MV $\hat{\beta}$. Un estimateur convergent de la probabilité $I\!P(y_i = 1|X_i)$ est ainsi donné par :

$$\widehat{I\!P}(y_i = 1|X_i) = G(X_i\hat{\beta}) \tag{8.28}$$

De même, on peut estimer l'effet marginal des différentes variables explicatives x_{ij} — les autres variables étant maintenues constantes — sur la probabilité $I\!P(y_i = 1|X_i)$. A nouveau, il suffit de remplacer dans l'expression (8.17) le vecteur de paramètres inconnus β par son estimateur MV $\hat{\beta}$. Si x_{ij} est une variable (au moins approximativement) continue, un estimateur convergent de l'effet marginal de la variable x_{ij} sur la probabilité $I\!P(y_i = 1|X_i)$ est ainsi donné par :

$$\frac{\partial \widehat{I\!P}(y_i = 1|X_i)}{\partial x_{ij}} = g(X_i\hat{\beta})\hat{\beta}_j \tag{8.29}$$

[144] Sauf cas particuliers (comme par exemple l'estimateur MV — qui est égal à l'estimateur MCO — des paramètres du modèle de régression linéaire sous l'hypothèse A6 de normalité), les estimateurs MV ne possèdent pas de propriétés d'échantillonnage exactes en échantillon fini, mais seulement des propriétés asymptotiques, valables pour n grand.

[145] Pour rappel, l'estimateur de la matrice de variance - covariance de l'estimateur MCO est $\hat{V}(\hat{\beta}) = \hat{s}^2(X'X)^{-1}$, qu'on peut encore écrire : $\hat{V}(\hat{\beta}) = \left[\sum_{i=1}^{n} \frac{1}{\hat{s}^2} X_i' X_i\right]^{-1}$. Vu sous cet angle, l'estimateur (8.27) apparaît déjà moins mystérieux (le facteur $\frac{1}{\hat{s}^2}$ est simplement remplacé par $\frac{g(X_i\hat{\beta})^2}{G(X_i\hat{\beta})(1-G(X_i\hat{\beta}))}$).

On peut procéder de la même façon — i.e., remplacer le vecteur de paramètres inconnus β par son estimateur MV $\hat{\beta}$ — lorsque la formule (8.17) de l'effet marginal doit être adaptée pour tenir compte de la présence dans l'index $X_i\beta$ de variables transformées et/ou de polynômes, ou encore lorsqu'il est préférable, en présence de variables explicatives binaires ou discrètes, de calculer l'effet marginal exact plutôt qu'un effet marginal approximatif (sur base de la formule (8.17)).

Les effets marginaux estimés (8.29) ne sont pas constants, mais varient en fonction de la valeur des variables explicatives $X_i = \left[\begin{array}{cccc} 1 & x_{i2} & \cdots & x_{ik} \end{array}\right]$. Pour résumer de façon synthétique les effets marginaux des variables explicatives x_{ij} (au moins approximativement) continues sur la probabilité $I\!P(y_i = 1|X_i)$, il est courant de calculer ces effets marginaux au point moyen de l'échantillon $\bar{X} = \left[\begin{array}{cccc} 1 & \bar{x}_2 & \cdots & \bar{x}_k \end{array}\right]$ au travers de l'expression[146] :

$$\frac{\partial \widehat{I\!P}(y_i = 1|X_i)}{\partial x_{ij}}|_{X_i=\bar{X}} = g(\bar{X}\hat{\beta})\hat{\beta}_j \tag{8.30}$$

Ces effets marginaux calculés au point moyen de l'échantillon peuvent s'interpréter comme les effets marginaux des différentes variables x_{ij} pour un individu moyen, i.e., un individu dont les variables explicatives X_i seraient égales à $\bar{X}$.

On peut encore procéder de la même façon — i.e., résumer les effets marginaux variables en fonction de X_i en calculant ces effets au point moyen de l'échantillon $\bar{X}$ — lorsque la formule (8.17) de l'effet marginal doit être adaptée pour tenir compte de la présence dans l'index $X_i\beta$ de variables transformées et/ou de polynômes, ou encore lorsqu'il est préférable, en présence de variables explicatives binaires ou discrètes, de calculer l'effet marginal exact plutôt qu'un effet marginal approximatif (sur base de la formule (8.17)).

Notons qu'il faut faire attention à la définition du point moyen de l'échantillon lorsque l'index $X_i\beta$ contient des variables transformées, des polynômes et/ou des variables explicatives binaires. Ainsi par exemple, pour le modèle :

$$I\!P(y_i = 1|X_i) = G\left(\beta_1 + \beta_2 x_{i2} + \beta_3 \ln x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i4}^2\right) ,$$

où toutes les variables sont (approximativement) continues, il convient d'utiliser comme point moyen de l'échantillon $\bar{X} = \left[\begin{array}{ccccc} 1 & \bar{x}_2 & \ln \bar{x}_3 & \bar{x}_4 & (\bar{x}_4)^2 \end{array}\right]$, et non $\bar{X} = \left[\begin{array}{ccccc} 1 & \bar{x}_2 & \overline{\ln x_3} & \bar{x}_4 & \overline{x_4^2} \end{array}\right]$, i.e., pour $\ln x_{i3}$, le logarithme de la moyenne de x_{i3} ($= \ln \bar{x}_3$), et non la moyenne du logarithme de x_{i3} ($= \overline{\ln x_3}$), et pour x_{i4}^2, le carré de la moyenne de x_{i4} ($= (\bar{x}_4)^2$), et non la moyenne du carré de x_{i4} ($= \overline{x_4^2}$). Pour ce point moyen $\bar{X}$ correctement défini, les effets marginaux sont alors donnés par :

$$\frac{\partial \widehat{I\!P}(y_i = 1|X_i)}{\partial x_{i2}}|_{X_i=\bar{X}} = g(\bar{X}\hat{\beta})\hat{\beta}_2$$

[146] Une alternative est de calculer les effets marginaux pour tous les points X_i de l'échantillon, puis d'en prendre la moyenne.

$$\frac{\partial \widehat{I\!P}(y_i = 1|X_i)}{\partial x_{i3}}|_{X_i=\bar{X}} = g(\bar{X}\hat{\beta})\frac{\hat{\beta}_3}{\bar{x}_3}$$
$$\frac{\partial \widehat{I\!P}(y_i = 1|X_i)}{\partial x_{i4}}|_{X_i=\bar{X}} = g(\bar{X}\hat{\beta})(\hat{\beta}_4 + 2\hat{\beta}_5\bar{x}_4)$$

De même, pour par exemple le modèle :

$$I\!P(y_i = 1|X_i) = G\left(\beta_1 + \beta_2 D_i + \beta_3 x_i\right) ,$$

où D_i est une variable binaire et x_i une variable (approximativement) continue, on utilisera comme point moyen de l'échantillon $\bar{X} = \left[\begin{array}{ccc} 1 & 0 & \bar{x} \end{array}\right]$ et/ou $\bar{X} = \left[\begin{array}{ccc} 1 & 1 & \bar{x} \end{array}\right]$, plutôt que $\bar{X} = \left[\begin{array}{ccc} 1 & \bar{D} & \bar{x} \end{array}\right]$. Il est en effet difficile de regarder la moyenne $\bar{D}$ d'une variable binaire D_i — qui est égale à la proportion des observations de l'échantillon pour lesquelles $D_i = 1$[147] —, comme représentative d'un quelconque individu moyen. Pour ce (ou ces) point(s) moyen(s) $\bar{X}$ correctement défini(s), l'effet marginal de x_i est alors donné par :

$$\frac{\partial \widehat{I\!P}(y_i = 1|X_i)}{\partial x_i}|_{X_i=\bar{X}} = g(\bar{X}\hat{\beta})\hat{\beta}_3$$

et l'effet marginal exact de D_i par :

$$\begin{aligned} \frac{\Delta I\!P(y_i = 1|X_i)}{\Delta D_i}|_{X_i=\bar{X}} &= G(\bar{X}^1\hat{\beta}) - G(\bar{X}^0\hat{\beta}) \\ &= G(\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3\bar{x}) - G(\hat{\beta}_1 + \hat{\beta}_3\bar{x}) , \end{aligned}$$

où $\bar{X}^1 = \left[\begin{array}{ccc} 1 & 1 & \bar{x} \end{array}\right]$ et $\bar{X}^0 = \left[\begin{array}{ccc} 1 & 0 & \bar{x} \end{array}\right]$

La plupart des logiciels économétriques, en particulier GRETL, calcule et reporte automatiquement des effets marginaux calculés au point moyen de l'échantillon. Ces effets marginaux sont toujours, quelle que soit la forme des variables qui composent l'index $X_i\beta$ du modèle, calculés sur base de la formule de base (8.30), avec $\bar{X} = \left[\begin{array}{cccc} 1 & \bar{x}_2 & \cdots & \bar{x}_k \end{array}\right]$. Ils ne sont donc corrects que si le modèle ne contient ni variables transformées, ni polynômes, ni variables explicatives binaires ou discrètes. Si ce n'est pas le cas, ils doivent être recalculés de la façon indiquée ci-dessus.

Pour se faire une idée du degré d'ajustement ou de la 'capacité prédictive' du modèle que l'on vient d'estimer, il est usuel de calculer le *pourcentage de prévisions correctes* du modèle. Ce pourcentage est obtenu en faisant, pour chaque observation, une prévision $\hat{y}_i$ de la valeur de y_i sachant X_i sur base de la règle :

$$\left\{ \begin{array}{l} \hat{y}_i = 1 \text{ si } \widehat{I\!P}(y_i = 1|X_i) = G(X_i\hat{\beta}) \geq 0,5 \\ \hat{y}_i = 0 \text{ si } \widehat{I\!P}(y_i = 1|X_i) = G(X_i\hat{\beta}) < 0,5 \end{array} \right.$$

Le pourcentage de prévisions correctes du modèle est simplement le pourcentage des observations de l'échantillon pour lesquelles $\hat{y}_i = y_i$, i.e., le pourcentage des observations pour lesquelles la valeur prédite $\hat{y}_i$ est égale à la valeur observée y_i.

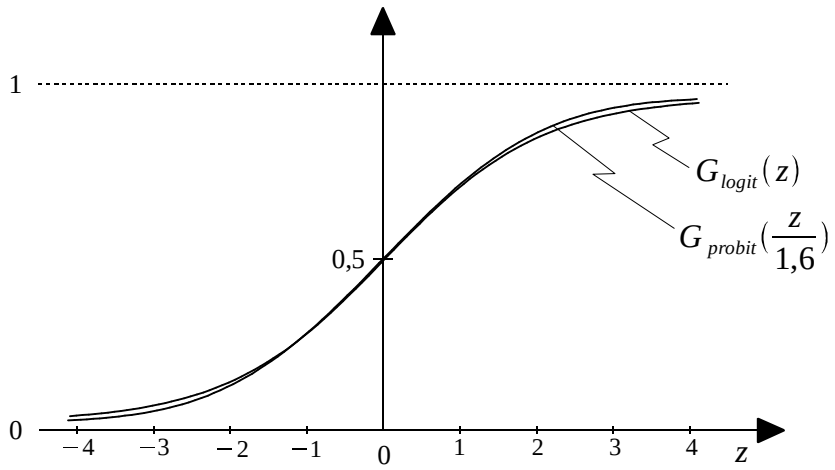
[147] Si par exemple $D_i = 1$ lorsque l'individu i est une femme, et 0 sinon, $\bar{D}$ est la proportion de femmes dans l'échantillon.

Il est reporté par la plupart des logiciels économétriques, en particulier GRETL. Il est parfois ventilé en pourcentages de prévisions correctes parmi, d'une part, les observations telles que $y_i = 1$, et d'autre part, les observations telles que $y_i = 0$. A l'instar du R^2 dans le modèle de régression, il s'agit d'une *mesure descriptive*, intéressante, mais à laquelle il ne faut pas accorder une trop grande importance[148].

Pour conclure cette section, on notera qu'en pratique les modèles logit et probit donnent généralement des résultats très semblables, tant en termes d'estimation des probabilités $I\!P(y_i = 1|X_i)$ qu'en termes d'estimation des effets marginaux $\frac{\partial I\!P(y_i=1|X_i)}{\partial x_{ij}}$. Cela vient du fait que les fonctions de liens (8.14) et (8.15) des modèles logit et probit sont en réalité moins dissemblables qu'il n'y paraît à première vue. En normalisant leur argument de façon adéquate, on a en effet approximativement :

$$G_{logit}(z) \simeq G_{probit}(\frac{z}{1,6}),$$

où $G_{logit}(.)$ et $G_{probit}(.)$ désignent respectivement les fonctions de liens (8.14) et (8.15). Graphiquement :



Graphique 50 : Les fonctions de lien normalisées des modèles logit et probit

A l'estimation, on obtient typiquement $\hat{\beta}_{logit} \simeq 1,6\hat{\beta}_{probit}$, où $\hat{\beta}_{logit}$ et $\hat{\beta}_{probit}$ sont les paramètres estimés des modèles logit et probit, de sorte qu'on a approximativement : $G_{logit}(X_i\hat{\beta}_{logit}) \simeq G_{probit}(X_i\frac{\hat{\beta}_{logit}}{1,6}) \simeq G_{probit}(X_i\hat{\beta}_{probit})$. Le choix d'utiliser en pratique l'un ou l'autre modèle est donc généralement peu crucial[149]. L'interprétation du modèle en termes de variable latente et la popularité de la loi normale explique la popularité du modèle probit. Mais le modèle logit est en pratique plus commode à utiliser, du fait de la disponibilité d'une forme analytique explicite pour la fonction de lien $G(.)$.

[148] Une autre mesure du degré d'ajustement ou de la 'capacité prédictive' du modèle fréquemment reportée par les logiciels économétriques (y compris GRETL) est le pseudo-R^2 de McFadden. Ce pseudo-R^2 n'est cependant pas aussi commode à interpréter que le R^2 standard du modèle de régression.

[149] Les modèles logit et probit ne se différencient (un peu) que pour l'estimation des probalilités $I\!P(y_i = 1|X_i)$ proches de 0 ou de 1.

8.2.4. Les modèles logit et probit III : inférence

On sait que, si le modèle est correctement spécifié, on a asymptotiquement :

$$\left[V(\hat{\beta})\right]^{-\frac{1}{2}}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N(0,I)\,,$$

où :

$$V(\hat{\beta}) = \left[\sum_{i=1}^{n} E\left(\frac{g(X_i\beta)^2 X_i' X_i}{G\left(X_i\beta\right)\left(1-G\left(X_i\beta\right)\right)}\right)\right]^{-1},$$

soit, exprimé sous forme d'approximation utilisable en échantillon fini pour n suffisamment grand :

$$\hat{\beta} \approx N(\beta, V(\hat{\beta})) \tag{8.31}$$

On sait également qu'un estimateur convergent de $V(\hat{\beta})$ est donné par :

$$\hat{V}(\hat{\beta}) = \left[\sum_{i=1}^{n} \frac{g(X_i\hat{\beta})^2 X_i' X_i}{G(X_i\hat{\beta})(1-G(X_i\hat{\beta}))}\right]^{-1} \tag{8.32}$$

En procédant de façon semblable à ce que nous avons fait dans le cadre du modèle de régression linéaire, on peut, sur base de ces résultats, obtenir des intervalles de confiance et des tests d'hypothèse relatifs à β, ainsi que des intervalles de prévision. Ces intervalles et tests ne seront évidemment valables qu'asymptotiquement, à titre approximatif pour n grand.

8.2.4.1. Intervalles de confiance

Le résultat de distribution d'échantillonnage (8.31) implique que, pour $j = 1, ..., k$, on a :

$$\hat{\beta}_j \approx N(\beta_j, Var(\hat{\beta}_j))\,,$$

où $Var(\hat{\beta}_j) = \left[V(\hat{\beta})\right]_{jj}$ désigne l'élément (j,j) de la matrice de variance - covariance $V(\hat{\beta})$, de sorte que :

$$\hat{z} = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \approx N(0,1)\,,$$

où $s.e.(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)}$.

On peut montrer qu'asymptotiquement le remplacement de l'écart-type théorique $s.e.(\hat{\beta}_j)$ par son estimateur convergent $s.\hat{e}.(\hat{\beta}_j)$ ne modifie pas cette distribution d'échantillonnage, de sorte qu'on a aussi :

$$\hat{t} = \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)} \approx N(0,1) \tag{8.33}$$

où $s.\hat{e}.(\hat{\beta}_j) = \sqrt{V\hat{a}r(\hat{\beta}_j)}$ et $V\hat{a}r(\hat{\beta}_j) = \left[\hat{V}(\hat{\beta})\right]_{jj}$ désigne l'élément (j,j) de $\hat{V}(\hat{\beta})$.

Etant donné (8.33), on a :

$$I\!P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)} \leq z_{1-\frac{\alpha}{2}}\right) \simeq 1-\alpha\,,$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $N(0,1)$, dont on peut déduire un intervalle de confiance à $(1-\alpha)\times 100\%$ pour β_j :

$$\left[\hat{\beta}_j - z_{1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j)\,;\, \hat{\beta}_j + z_{1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j)\right] \tag{8.34}$$

Cet intervalle de confiance a la même forme[150] et s'interprète de la même façon que l'intervalle de confiance pour β_j dans le modèle de régression linéaire. Rappelons néanmoins que, dans le présent contexte, les valeurs précises des paramètres β_j — qui représentent[151] les effets marginaux des différentes variables explicatives x_{ij} sur la variable latente y_i^* dans l'interprétation en termes de variable latente du modèle — n'ont généralement en elles-mêmes que peu d'intérêt. Seul leur signe ou leur nullité est généralement interprétable.

Dans la plupart des applications empiriques, on est généralement avant tout intéressé par les valeurs des effets marginaux des différentes variables explicatives x_{ij} sur la probabilité $I\!P(y_i = 1|X_i)$. On a vu que la formule précise de ces effets marginaux varie selon la forme de l'index $X_i\beta$ du modèle, i.e., selon qu'il contient ou non des variables transformées, des polynômes, ou encore des variables explicatives binaires ou discrètes. Dans tous les cas, l'effet marginal de la variable x_{ij} sur la probabilité $I\!P(y_i = 1|X_i)$ est donné par une *fonction non-linéaire* de β :

$$\frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{ij}} = h_j(X_i, \beta)\,, \tag{8.35}$$

et un estimateur ponctuel de cet effet marginal est obtenu en remplaçant β par son estimateur MV $\hat{\beta}$:

$$\frac{\partial \widehat{I\!P}(y_i = 1|X_i)}{\partial x_{ij}} = h_j(X_i, \hat{\beta}) \tag{8.36}$$

Dans le cas de base où x_{ij} est une variable (approximativement) continue et que l'index $X_i\beta$ du modèle ne contient ni variables transformées ni polynômes, on a

---

[150] Simplement, il utilise un quantile de la loi normale plutôt que de la loi de Student — dans le présent contexte, aucun résultat exact en échantillon fini ne justifie l'utilisation, asymptotiquement équivalente pour n grand, de quantiles de la loi de Student au lieu de quantiles de la loi normale —, et bien entendu la définition de $s.\hat{e}.(\hat{\beta}_j)$ est différente.

[151] Si l'index $X_i\beta$ du modèle ne contient ni variables transformées, ni polynômes.

simplement (cf. les équations (8.17) et (8.29)) :

$$h_j(X_i, \beta) = g(X_i\beta)\beta_j \tag{8.37}$$

Comme pour β_j, on peut obtenir un intervalle de confiance pour l'effet marginal $h_j(X_i, \beta)$. Sa dérivation et son calcul pratique sont cependant complexes, du fait que $h_j(X_i, \beta)$ est une fonction non-linéaire de β.

On peut montrer qu'asymptotiquement, pour n grand, on peut approximer la fonction non-linéaire $h_j(X_i, \hat{\beta})$ par une fonction linéaire donnée par son développement de Taylor à l'ordre 1 en $\hat{\beta} = \beta$:

$$\begin{aligned} h_j(X_i, \hat{\beta}) &\simeq h_j(X_i, \beta) + \frac{\partial h_j(X_i, \beta)}{\partial \beta'}(\hat{\beta} - \beta) \\ &\simeq h_j(X_i, \beta) + D(X_i, \beta)(\hat{\beta} - \beta)\,, \end{aligned} \tag{8.38}$$

où :

$$D(X_i, \beta) = \frac{\partial h_j(X_i, \beta)}{\partial \beta'} = \left[\begin{array}{cccc} \frac{\partial h_j(X_i,\beta)}{\partial \beta_1} & \frac{\partial h_j(X_i,\beta)}{\partial \beta_2} & \cdots & \frac{\partial h_j(X_i,\beta)}{\partial \beta_k} \end{array} \right],$$

i.e., un vecteur ligne contenant les dérivées de la fonction $h_j(X_i, \beta)$ par rapport aux différents β_j.

Pour le cas de base où $h_j(X_i, \beta) = g(X_i\beta)\beta_j$, on a[152] :

$$\begin{aligned} & D(X_i, \beta) \\ = & \left[\begin{array}{ccccccc} \beta_j g'(X_i\beta) & \beta_j g'(X_i\beta)x_{i2} & \cdots & \beta_j g'(X_i\beta)x_{ij} + g(X_i\beta) & \cdots & \beta_k g'(X_i\beta)x_{ik} \end{array} \right], \end{aligned}$$

où $g'(z) = \frac{dg(z)}{dz}$ est donné, pour le modèle logit, par :

$$g'(z) = \frac{e^z(1 - e^z)}{(1 + e^z)^3}\,,$$

et pour le modèle probit, par :

$$g'(z) = \frac{-z}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Sur base de l'approximation linéaire (8.38), $h_j(X_i, \hat{\beta})$ est une combinaison linéaire de $\hat{\beta}$. Comme $\hat{\beta} \approx N(\beta, V(\hat{\beta}))$ et qu'une combinaison linéaire d'un vecteur aléatoire normal suit aussi une loi normale, on en déduit qu'on a asymptotiquement[153] :

$$h_j(X_i, \hat{\beta}) \approx N(h_j(X_i, \beta), Var(h_j(X_i, \hat{\beta}))\,, \tag{8.39}$$

---

[152] Pour rappel, $X_i = \left[\begin{array}{cccc} 1 & x_{i2} & \cdots & x_{ik} \end{array} \right]$, i.e., le premier élément de X_i est la constante 1.

[153] Cette façon d'obtenir la distribution asymptotique d'une fonction non-linéaire d'un vecteur de paramètres $\hat{\beta}$ en faisant une approximation linéaire de la fonction en $\hat{\beta} = \beta$ est appelée la *méthode delta* (*delta method* en anglais).

où :

$$Var(h_j(X_i, \hat{\beta})) = D(X_i, \beta) V(\hat{\beta}) D(X_i, \beta)' \tag{8.40}$$

Un estimateur convergent $\hat{Var}(h_j(X_i, \hat{\beta}))$ de la variance $Var(h_j(X_i, \hat{\beta}))$ est simplement obtenu en remplaçant, dans l'expression (8.40), $V(\hat{\beta})$ par son estimateur $\hat{V}(\hat{\beta})$, et évaluant $D(X_i, \beta)$ en $\beta = \hat{\beta}$:

$$\hat{Var}(h_j(X_i, \hat{\beta})) = D(X_i, \hat{\beta}) \hat{V}(\hat{\beta}) D(X_i, \hat{\beta})'$$

Le résultat de distribution d'échantillonnage (8.39) implique que :

$$\hat{z} = \frac{h_j(X_i, \hat{\beta}) - h_j(X_i, \beta)}{s.e.(h_j(X_i, \hat{\beta}))} \approx N(0, 1)\,,$$

où $s.e.(h_j(X_i, \hat{\beta})) = \sqrt{Var(h_j(X_i, \hat{\beta}))}$.

On peut encore montrer qu'asymptotiquement le remplacement de l'écart-type théorique $s.e.(h_j(X_i, \hat{\beta}))$ par son estimateur convergent $s.\hat{e}.(h_j(X_i, \hat{\beta}))$ ne modifie pas cette distribution d'échantillonnage, de sorte qu'on a aussi :

$$\hat{t} = \frac{h_j(X_i, \hat{\beta}) - h_j(X_i, \beta)}{s.\hat{e}.(h_j(X_i, \hat{\beta}))} \approx N(0, 1) \tag{8.41}$$

où $s.\hat{e}.(h_j(X_i, \hat{\beta})) = \sqrt{\hat{Var}(h_j(X_i, \hat{\beta}))}$.

Etant donné (8.41), on a :

$$I\!P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{h_j(X_i, \hat{\beta}) - h_j(X_i, \beta)}{s.\hat{e}.(h_j(X_i, \hat{\beta}))} \leq z_{1-\frac{\alpha}{2}} \right) \simeq 1 - \alpha\,,$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $N(0, 1)$, dont on peut déduire un intervalle de confiance à $(1 - \alpha) \times 100\%$ pour $h_j(X_i, \beta)$:

$$\left[h_j(X_i, \hat{\beta}) - z_{1-\frac{\alpha}{2}} s.\hat{e}.(h_j(X_i, \hat{\beta}))\,;\, h_j(X_i, \hat{\beta}) + z_{1-\frac{\alpha}{2}} s.\hat{e}.(h_j(X_i, \hat{\beta})) \right] \tag{8.42}$$

où $s.\hat{e}.(h_j(X_i, \hat{\beta})) = \sqrt{\hat{Var}(h_j(X_i, \hat{\beta}))} = \sqrt{D(X_i, \hat{\beta}) \hat{V}(\hat{\beta}) D(X_i, \hat{\beta})'}$. Cet intervalle de confiance s'interprète de la façon habituelle, comme l'intervalle de confiance pour β_j.

L'intervalle de confiance (8.42) donne un intervalle de confiance pour l'effet marginal $\frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{ij}} = h_j(X_i, \beta)$ de la variable x_{ij} sur la probabilité $I\!P(y_i = 1|X_i)$ pour une valeur X_i donnée des variables explicatives. Un intervalle de confiance pour cet effet marginal au point moyen de l'échantillon $\bar{X}$ — dont, pour rappel, il faut faire attention à la définition — est simplement obtenu en calculant l'intervalle de confi-

ance pour $X_i = \bar{X}$. Notons que peu de logiciels économétriques — GRETL n'en fait malheureusement pas partie[154] — calcule et reporte automatiquement cet intervalle de confiance.

8.2.4.2. Tests d'hypothèse

On sait que le résultat de distribution d'échantillonnage (8.31) implique que, pour $j = 1, ..., k$, on a :

$$\hat{\beta}_j \approx N(\beta_j, Var(\hat{\beta}_j)) \,,$$

où $Var(\hat{\beta}_j) = \left[V(\hat{\beta})\right]_{jj}$.

Ainsi, si la vraie valeur de β_j est égale à β_j^o, on a :

$$\hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \approx N(0,1) \,,$$

où $s.e.(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)}$, tandis que si la vraie valeur de β_j est différente de β_j^o et par exemple égale à β_j^* ($\beta_j^* \neq \beta_j^o$), on a :

$$\hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j)} \approx N\left(\frac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}, 1\right)$$

On peut à nouveau montrer qu'asymptotiquement le remplacement de l'écart-type théorique $s.e.(\hat{\beta}_j)$ par son estimateur convergent $s.\hat{e}.(\hat{\beta}_j)$ ne modifie pas ces distributions d'échantillonnage, de sorte qu'on a aussi, lorsque $\beta_j = \beta_j^o$:

$$\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \approx N(0,1) \,,$$

et lorsque $\beta_j = \beta_j^* \neq \beta_j^o$:

$$\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \approx N\left(\frac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j)}, 1\right)$$

où $s.\hat{e}.(\hat{\beta}_j) = \sqrt{V\hat{a}r(\hat{\beta}_j)}$ et $V\hat{a}r(\hat{\beta}_j) = \left[\hat{V}(\hat{\beta})\right]_{jj}$.

Comme dans le modèle de régression linéaire, étant donné ses propriétés, on peut utiliser $\hat{t}_o$ comme statistique de test pour tester des hypothèses telles que $\mathrm{H}_0 : \beta_j = \beta_j^o$ contre $\mathrm{H}_1 : \beta_j \neq \beta_j^o$ (test bilatéral) ou $\mathrm{H}_0 : \beta_j \leq \beta_j^o$ (resp. $\beta_j \geq \beta_j^o$) contre $\mathrm{H}_1 : \beta_j > \beta_j^o$ (resp. $\beta_j < \beta_j^o$) (tests unilatéraux). Les règles de décision à appliquer pour des tests au seuil α, ainsi que les P-valeurs de ces tests, sont résumées dans le

[154] Pour calculer l'intervalle de confiance (8.42) avec GRETL, il faut utiliser ses fonctions de calcul matriciel.

tableau suivant :

Test	Règle de décision[155]	P-valeur[156]
$\mathrm{H}_0 : \beta_j = \beta_j^o$	- RH_0 si $\lvert\hat{t}_o\rvert = \left\lvert\frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)}\right\rvert > z_{1-\frac{\alpha}{2}}$	$p_{\hat{t}_o^*} = I\!P(\lvert z\rvert > \lvert\hat{t}_o^*\rvert)$,
contre $\mathrm{H}_1 : \beta_j \neq \beta_j^o$	- NRH_0 sinon	
$\mathrm{H}_0 : \beta_j \leq \beta_j^o$	- RH_0 si $\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} > z_{1-\alpha}$	$p_{\hat{t}_o^*} = I\!P(z > \hat{t}_o^*)$,
contre $\mathrm{H}_1 : \beta_j > \beta_j^o$	- NRH_0 sinon	
$\mathrm{H}_0 : \beta_j \geq \beta_j^o$	- RH_0 si $\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} < z_{\alpha}$	$p_{\hat{t}_o^*} = I\!P(z < \hat{t}_o^*)$
contre $\mathrm{H}_1 : \beta_j < \beta_j^o$	- NRH_0 sinon	

Ces tests, qu'on appelle toujours *t-tests* dans le présent contexte, ont la même forme[157] et s'interprètent — en termes de risque de première espèce, de puissance et de P-valeur — de la même façon que les t-tests de β_j dans le modèle de régression linéaire. Rappelons cependant à nouveau que, dans le présent contexte, seul le signe ou la nullité des β_j est généralement interprétable.

Comme pour le modèle de régression linéaire, tous les logiciels économétriques, en particulier GRETL, calculent et reportent en standard la statistique $\hat{t}_o = \frac{\hat{\beta}_j}{s.\hat{e}.(\hat{\beta}_j)}$ — qu'on appelle toujours *t-statistique* (de $\hat{\beta}_j$) — et la P-valeur du test de $\mathrm{H}_0 : \beta_j = 0$ contre $\mathrm{H}_1 : \beta_j \neq 0$ pour chacun des paramètres du modèle. On peut donc directement voir, sur base des résultats d'estimation, si les différentes variables explicatives x_{ij} ont ou non un effet (statistiquement) significatif sur la probablilité $I\!P(y_i = 1|X_i)$. Pour tester si les différentes variables explicatives x_{ij} ont un effet (statistiquement) significativement positif ou négatif sur la probablilité $I\!P(y_i = 1|X_i)$, comme l'effet marginal $\frac{\partial I\!P(y_i=1|X_i)}{\partial x_{ij}}$ de chacune des variables x_{ij} est toujours — à tout le moins si l'index $X_i\beta$ ne contient ni transformations de variables atypiques[158] ni polynômes[159] — du même signe que β_j, il suffit de tester $\mathrm{H}_0 : \beta_j \leq 0$ contre $\mathrm{H}_1 : \beta_j > 0$ ou $\mathrm{H}_0 : \beta_j \geq 0$ contre $\mathrm{H}_1 : \beta_j < 0$.

Pour pouvoir tester des hypothèses plus élaborées, comme par exemple $\mathrm{H}_0 : \beta_4 = \beta_5 = \beta_6 = 0$ contre $\mathrm{H}_1 : \beta_4 \neq 0$ et/ou $\beta_5 \neq 0$ et/ou $\beta_6 \neq 0$ dans le modèle :

$$I\!P(y_i = 1|X_i) = G\left(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i3}^2 + \beta_6 x_{i2} x_{i3}\right) ,$$

i.e., si une forme linéaire simple (plutôt que polynomiale) en x_{i2} et x_{i3} serait ou non

[155] $z_{1-\frac{\alpha}{2}}$, $z_{1-\alpha}$ et $z_\alpha (= -z_{1-\alpha})$ désignent les quantiles d'ordre $1-\frac{\alpha}{2}$, $1-\alpha$ et α de la loi $N(0,1)$.

[156] $\hat{t}_o^*$ désigne la valeur de la statistique $\hat{t}_o$ obtenue pour un échantillon particulier, et $z \sim N(0,1)$.

[157] Simplement, comme dans le cas de l'intervalle de confiance pour β_j, ils s'appuyent sur la loi normale plutôt que sur la loi de Student, et bien entendu la définition de $s.\hat{e}.(\hat{\beta}_j)$ est différente.

[158] Comme par exemple la fonction inverse $\frac{1}{x}$. Notons que la transformation logarithmique standard $\ln(x)$ ne provoque elle aucune inversion de signe de l'effet marginal.

[159] Lorsque l'index $X_i\beta$ contient des polynômes, les signes des effets marginaux ne sont généralement plus constants, mais variables.

suffisante pour la fonction d'index $X_i\beta$ du modèle, on a besoin d'un test général similaire au F-test du modèle de régression linéaire. Un tel test général de :

$$\mathrm{H}_0 : \; R_0\beta = r_0 \text{ contre } \mathrm{H}_1 : \; R_0\beta \neq r_0$$

où R_0 est une matrice $q \times k$ de constantes ($q \leq k$; q = le nbr. de restrictions et k = le nbr. de paramètres) et r_0 un vecteur $q \times 1$ de constantes, est aisé à obtenir.

Le résultat de distribution d'échantillonnage (8.31) implique qu'on a :

$$(R_0\hat{\beta} - r_0) \approx N(R_0\beta - r_0, R_0V(\hat{\beta})R_0') \,,$$

de sorte que, si la vraie valeur de β est telle que $R_0\beta = r_0$, càd. que H_0 est vraie, on a :

$$\hat{\chi}_0^2 = (R_0\hat{\beta} - r_0)' \left[R_0V(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(q) \,,$$

tandis que si la vraie valeur de β est telle que $R_0\beta \neq r_0$, càd. que H_0 est fausse, on peut montrer qu'on a[160] :

$$\hat{\chi}_0^2 = (R_0\hat{\beta} - r_0)' \left[R_0V(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(\delta^*, q) \,,$$

où $\delta^* = (R_0\beta - r_0)' \left[R_0V(\hat{\beta})R_0'\right]^{-1} (R_0\beta - r_0)$.

On peut encore montrer qu'asymptotiquement le remplacement de $V(\hat{\beta})$ par son estimateur convergent $\hat{V}(\hat{\beta})$ ne modifie pas ces distributions d'échantillonnage, de sorte qu'on a aussi, lorsque H_0 est vraie (i.e., $R_0\beta = r_0$) :

$$\hat{\chi}_0^{2\prime} = (R_0\hat{\beta} - r_0)' \left[R_0\hat{V}(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(q) \,,$$

et lorsque H_0 est fausse (i.e., $R_0\beta \neq r_0$) :

$$\hat{\chi}_0^{2\prime} = (R_0\hat{\beta} - r_0)' \left[R_0\hat{V}(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(\delta^*, q) \,,$$

où $\delta^* = (R_0\beta - r_0)' \left[R_0V(\hat{\beta})R_0'\right]^{-1} (R_0\beta - r_0)$.

Etant donné ses propriétés, on peut utiliser $\hat{\chi}_0^{2\prime}$ comme statistique de test pour tester $\mathrm{H}_0 : \; R_0\beta = r_0$ contre $\mathrm{H}_1 : \; R_0\beta \neq r_0$. La règle de décision à appliquer pour un test au seuil α est donnée par :

$$\begin{cases} \text{- Rejet de } \mathrm{H}_0 \text{ si } \hat{\chi}_0^{2\prime} > \chi^2_{q;1-\alpha} \\ \text{- Non-rejet de } \mathrm{H}_0 \text{ sinon} \end{cases}$$

où la valeur critique $\chi^2_{q;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(q)$, et la P-valeur du test est donnée par :

$$p_{\hat{\chi}_0^{2\prime *}} = I\!P(v > \hat{\chi}_0^{2\prime *})$$

[160] Pour rappel, par définition, si $X \sim N(m, \Sigma)$, où X est un vecteur de dimension $q \times 1$, alors : $X'\Sigma^{-1}X \sim \chi^2(\delta, q)$, où $\delta = m'\Sigma^{-1}m$.

où $\hat{\chi}_0^{2\prime *}$ désigne la valeur de la statistique $\hat{\chi}_0^{2\prime}$ obtenue pour un échantillon particulier, et $v \sim \chi^2(q)$.

Ce χ^2-test, couramment appelé *test du khi-carré* ou encore *test de Wald*, a la même forme[161] et s'interprète — en termes de risque de première espèce, de puissance et de P-valeur — de la même façon que le χ^2-test (basé sur la statistique $\hat{\chi}_0^{2\prime}$, cf. Section 7.1.2) obtenu dans le modèle de régression linéaire. Comme pour le modèle de régression linéaire, la plupart des logiciels économétriques, en particulier GRETL, permettent de le calculer de façon très simple : il suffit de spécifier les contraintes $R_0\beta = r_0$, et le logiciel reporte alors la valeur de la statistique $\hat{\chi}_0^{2\prime}$ et la P-valeur du test.

8.2.4.3. Intervalle de prévision

Comme pour le modèle de régression linéaire, un des objectifs des modèles logit et probit est de faire des prévisions. Un estimateur / prédicteur convergent de la probabilité que y_0 soit égale à 1 sachant que les variables explicatives prennent une valeur $X_0 = \begin{bmatrix} 1 & x_{02} & \cdots & x_{0k} \end{bmatrix}$, càd. de la probabilité :

$$I\!P(y_0 = 1|X_0) = G(X_0\beta)$$

est simplement donné par[162] :

$$\widehat{I\!P}(y_0 = 1|X_0) = G(X_0\hat{\beta}) \tag{8.43}$$

L'estimateur / prédicteur (8.43) fournit une prévision ponctuelle de la probabilité $I\!P(y_0 = 1|X_0) = G(X_0\beta)$. On peut lui associer un intervalle de prévision, càd. un intervalle de valeurs plausibles pour $G(X_0\beta)$. Un tel intervalle de prévision est assez facile à obtenir. Le résultat de distribution d'échantillonnage (8.31) implique qu'on a :

$$X_0\hat{\beta} \approx N(X_0\beta, Var(X_0\hat{\beta}))\,, \tag{8.44}$$

où :

$$Var(X_0\hat{\beta}) = X_0 V(\hat{\beta}) X_0' \tag{8.45}$$

Un estimateur convergent $V\hat{a}r(X_0\hat{\beta})$ de la variance $Var(X_0\hat{\beta})$ est simplement obtenu en remplaçant, dans l'expression (8.45), $V(\hat{\beta})$ par son estimateur $\hat{V}(\hat{\beta})$:

$$V\hat{a}r(X_0\hat{\beta}) = X_0 \hat{V}(\hat{\beta}) X_0'$$

---

[161] Simplement, la définition de $\hat{V}(\hat{\beta})$ est différente. Notons encore que, dans le présent contexte, aucun résultat exact en échantillon fini ne justifie l'utilisation, asymptotiquement équivalente pour n grand, d'une forme du type F-test de ce test.

[162] Cet estimateur correspond, dans le cadre du modèle de régression linéaire, à l'estimateur / prédicteur de l'espérance de y sachant $(x_{02}, ...x_{0k})$.

Le résultat de distribution d'échantillonnage (8.44) implique que :

$$\hat{z} = \frac{X_0\hat{\beta} - X_0\beta}{s.e.(X_0\hat{\beta})} \approx N(0,1)\,,$$

où $s.e.(X_0\hat{\beta}) = \sqrt{Var(X_0\hat{\beta})}$.

On peut encore montrer qu'asymptotiquement le remplacement de l'écart-type théorique $s.e.(X_0\hat{\beta})$ par son estimateur convergent $s.\hat{e}.(X_0\hat{\beta})$ ne modifie pas cette distribution d'échantillonnage, de sorte qu'on a aussi :

$$\hat{t} = \frac{X_0\hat{\beta} - X_0\beta}{s.\hat{e}.(X_0\hat{\beta})} \approx N(0,1)\,, \tag{8.46}$$

où $s.\hat{e}.(X_0\hat{\beta}) = \sqrt{V\hat{a}r(X_0\hat{\beta})}$.

Etant donné (8.46), on a :

$$I\!P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{X_0\hat{\beta} - X_0\beta}{s.\hat{e}.(X_0\hat{\beta})} \leq z_{1-\frac{\alpha}{2}}\right) \simeq 1-\alpha$$

$$\Leftrightarrow\ I\!P\left(X_0\hat{\beta} - z_{1-\frac{\alpha}{2}}s.\hat{e}.(X_0\hat{\beta}) \leq X_0\beta \leq X_0\hat{\beta} + z_{1-\frac{\alpha}{2}}s.\hat{e}.(X_0\hat{\beta})\right) \simeq 1-\alpha\,,$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $N(0,1)$. Finalement, comme la fonction de lien $G(.)$ est — tant dans le modèle logit que probit — strictement croissante, on a encore :

$$I\!P\left(G(X_0\hat{\beta} - z_{1-\frac{\alpha}{2}}s.\hat{e}.(X_0\hat{\beta})) \leq G(X_0\beta) \leq G(X_0\hat{\beta} + z_{1-\frac{\alpha}{2}}s.\hat{e}.(X_0\hat{\beta}))\right) \simeq 1-\alpha\,,$$

dont on peut déduire un intervalle de prévision à $(1-\alpha)\times 100\%$ pour $I\!P(y_0 = 1|X_0) = G(X_0\beta)$:

$$\left[G(X_0\hat{\beta} - z_{1-\frac{\alpha}{2}}s.\hat{e}.(X_0\hat{\beta}))\,;\, G(X_0\hat{\beta} + z_{1-\frac{\alpha}{2}}s.\hat{e}.(X_0\hat{\beta}))\right]\,, \tag{8.47}$$

où $s.\hat{e}.(X_0\hat{\beta}) = \sqrt{V\hat{a}r(X_0\hat{\beta})} = \sqrt{X_0\hat{V}(\hat{\beta})X_0'}$.

Bien que d'une forme un peu différente[163], cet intervalle de prévision s'interprète de la même façon que l'intervalle de prévision pour $E(y_0)$ dans le modèle de régression linéaire.

[163] On notera qu'un intervalle de prévision de forme plus classique pourrait être obtenu en s'appuyant sur la *méthode delta* utilisée à la Section 8.2.4.1.