Introduction to Econometrics

Keita, Moussa

September 2015

INTRODUCTION A L'ECONOMETRIE

Par
Moussa Keita, PhD*
Septembre 2015
(Version 1)

*Ecole d'Economie, Université d'Auvergne clermont Ferrand 1

Contact info: Email : keitam09@ymail.com

Codes JEL: C1, C2

Mots clés: Econométrie, Modèle linéaire, Variable qualitative, logit, probit, MCO, Maximum de vraisemblance

AVANT-PROPOS

Ce manuscrit propose une introduction à l'analyse économétrique. Il est articulé autour de deux grandes parties structurées sur six chapitres. La première partie, consacrée à l'étude du modèle linéaire, est constituée de quatre chapitres. Le premier présente quelques concepts statistiques utiles en Econométrie alors que le second et le troisième chapitre sont consacrés à l'étude du modèle linéaire simple et du modèle linéaire multiple. Quant au quatrième chapitre, il se focalise sur l'étude du modèle linéaire généralisé (utilisé en cas de violation de certaines hypothèses standards du modèle linéaire). Dans cette première partie, un accent particulier est mis sur les méthodes d'estimation telles que la méthode des moindres carrés ordinaires et la méthode de maximum de vraisemblance. Une large discussion est également menée sur les techniques d'inférence et sur les approches de tests d'hypothèses. La seconde partie du travail est consacrée à l'étude des modèles à variable dépendante qualitative. Dans cette partie, deux classes de modèles sont étudiées : celles des modèles à variable dépendante dichotomique (probit et logit standard) et celle des modèles à variables dépendante polytomique (probit et logit multinomiaux ordonnés et non ordonnés).

Le manuscrit étant toujours en cours de progression, nous restons réceptifs à toutes critiques et suggestions de nature à améliorer le contenu du travail.

CHAPITRE 1. CONCEPTS STATISTIQUES DE BASE EN ECONOMETRIE

1.1. Notions de série statistique

On s'intéresse à deux variables x et y mesurées sur n unités d'observation. Pour chaque unité, on obtient alors donc deux mesures. La série statistique est alors une suite de n couples des valeurs prises par les deux variables sur chaque individu. Cela peut se présenter comme suit :

X	x_1	x_2	x_3	...	x_n
Y	y_1	y_2	y_3	...	y_n

Chacune des deux variables peut être soit quantitative, soit qualitative.

Par exemple lorsqu'on mesure le poids (X) et la taille (Y) de 20 individus, les informations obtenues peuvent être présentées sous forme de séries statistiques comme suit :

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	60	61	64	67	68	69	70	70	72	73	75	76	78	80	85	90	96	96	98	101
Y	155	162	157	170	164	162	169	170	178	173	180	175	173	175	179	175	180	185	189	187

1.2. Tendance centrale et de dispersion d'une série

1.2.1. La moyenne

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

1.2.2. La variance:

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{1.2a}$$

1.2.3. L'écart-type

$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sqrt{S^2} \tag{1.2b}$$

1.2.4. Covariance (de deux séries)

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad (1.2c)$$

Notons qu'en développant cette expression, on retrouve une nouvelle expression de la covariance :

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y} \qquad (1.2d)$$

A travers cette expression, on peut donner l'expression développée de la variance S_x^2. En effet :

$$S_x^2 = S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i x_i) - \bar{x}\bar{x}$$

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 \qquad (1.2e)$$

1.2.5. Le coefficient de corrélation linéaire (entre deux séries)

Le coefficient de corrélation linéaire mesure le degré de dépendance linéaire entre deux variables. Il est égal à la covariance des deux variables divisée par le produit leur écart-type :

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \qquad (1.3)$$

Le coefficient est compris entre -1 et 1. Pour le cas de deux variables indépendantes la corrélation est égale à 0.

1.2.6. Le coefficient de détermination

Le coefficient de détermination est le carré du coefficient de corrélation. C'est le carré de la covariance divisée par le produit des variances

$$R_{xy}^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} \qquad (1.4)$$

Le coefficient de détermination est compris entre 0 et 1.

1.3. Quelques rappels sur l'opérateur d'espérance E(.)

1.3.1 Définition et propriétés de l'opérateur d'espérance

De façon simple, l'espérance d'une variable correspond à la moyenne de cette variable lorsque n est grand. Elle se calcule par la formule suivante :

$$\mathrm{E(X)} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.5}$$

Dans certains cas, au lieu d'utiliser $\bar{x}$, on utilise E(X). Par exemple, pour calculer la variance, on écrit :

$$\mathrm{VAR(X)} = \frac{1}{n}\sum_{i=1}^{n}\big(x_i - \mathrm{E(X)}\big)^2$$

De plus, sachant que $\mathrm{VAR(X)} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$, on peut réécrire cette expression comme suit :

$$\mathrm{VAR(X)} = \mathrm{E(X^2)} - \big(\mathrm{E(X)}\big)^2 \tag{1.6}$$

Avec

$$\mathrm{E(X^2)} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Cette formulation de la variance s'avère d'une importance capitale dans beaucoup de démonstrations. Elle peut se généraliser quelle que soit la variable considérée. Soit une variable Z telle que $Z = XY$, on a :

$$\mathrm{E(Z)} = \mathrm{E}(XY) = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i$$

$$\mathrm{VAR(Z)} = \frac{1}{n}\sum_{i=1}^{n}\big(x_i y_i - \mathrm{E(Z)}\big)^2$$

Or on sait que :

$$\mathrm{VAR(Z)} = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i)^2 - \big(\mathrm{E(Z)}\big)^2$$

Par conséquent :

$$\mathrm{VAR(Z)} = \mathrm{E(Z^2)} - \big(\mathrm{E(Z)}\big)^2$$

Ainsi de façon explicite, on peut écrire :

$$\mathrm{VAR(XY)} = \mathrm{E(X^2Y^2)} - \big(\mathrm{E(XY)}\big)^2$$

Et lorsque les deux variables X et Y sont indépendantes alors $\mathrm{E(XY)} = \mathrm{E(X)} * \mathrm{E(Y)}$. Ainsi, on a :

$$\mathrm{VAR(XY)} = \mathrm{E(X^2Y^2)} - \big(\mathrm{E(X)E(Y)}\big)^2 \tag{1.7}$$

On peut élargir ce type de raisonnement au cas de la covariance entre X et Y. En effet,

$$\mathrm{COV(X,Y)} = \frac{1}{n}\sum_{i=1}^{n}\big(x_i - \mathrm{E(X)}\big)\big(y_i - \mathrm{E(Y)}\big) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \mathrm{E(X)E(Y)}$$

$$\mathrm{COV(X,Y)} = \mathrm{E(XY)} - \mathrm{E(X)E(Y)} \tag{1.8}$$

Si les deux variables sont indépendantes $\mathrm{E(XY)} = \mathrm{E(X)E(Y)}$, Par conséquent

$$\mathrm{COV(X,Y)} = 0$$

Aussi, la formule de la covariance peut être généralisée quelle que soit la puissance de X et de Y. On a :

$$\mathrm{COV}(\mathrm{X}^r, \mathrm{Y}^r) = \mathrm{E}(\mathrm{X}^r\mathrm{Y}^r) - \mathrm{E}(\mathrm{X}^r)\mathrm{E}(\mathrm{Y}^r) \tag{1.9}$$

Exemple :

$$\mathrm{COV}(\mathrm{X}^2, \mathrm{Y}^2) = \mathrm{E}(\mathrm{X}^2\mathrm{Y}^2) - \mathrm{E}(\mathrm{X}^2)\mathrm{E}(\mathrm{Y}^2)$$

1.3.2. Quelques utilisations de l'opérateur d'espérance $E(.)$

1.3.2.1. Calcul de l'espérance dans le cas de la somme ou du produit de deux variables aléatoires

- **Espérance d'une somme** $\mathrm{E(X+Y)}$**:**

$$E(X+Y) = E(X) + E(Y) \tag{1.10}$$

- **Espérance d'un produit** $\mathrm{E(XY)}$**:**

On sait que :

$$\mathrm{COV}(X, \mathrm{Y}) = \mathrm{E(XY)} - \mathrm{E(Y)E(Y)} \Rightarrow$$

$$\mathrm{E(XY)} + \mathrm{COV}(X, \mathrm{Y}) + \mathrm{E(Y)E(Y)} \tag{1.11a}$$

Si les deux variables sont indépendantes $\mathrm{COV}(X, \mathrm{Y}) = 0$. Ainsi, on a :

$$\mathrm{E(XY)} = \mathrm{E(X)E(Y)} \tag{1.11b}$$

L'espérance du produit de 2 variables aléatoires indépendantes est le produit des espérances.

1.3.2.2. Calcul de la variance dans le cas de la somme ou du produit de deux variables aléatoires

- **Variance d'une somme** $\mathrm{VAR}(\mathrm{X}+\mathrm{Y})$**:**

$$VAR(X+Y) = E((X+Y)^2) - [E(X+Y)]^2$$

$$= E(\mathrm{X}^2 + \mathrm{Y}^2 + 2XY) - [E(X) + E(Y)]^2$$

$$= E(\mathrm{X}^2) + E(\mathrm{Y}^2) + 2E(XY) - [E(X)]^2 - [E(Y)]^2 - 2E(X)E(Y)$$

$$= E(\mathrm{X}^2) - [E(X)]^2 + E(\mathrm{Y}^2) - [E(Y)]^2 + 2[E(XY) - E(X)E(Y)]$$

$$VAR(X) + VAR(Y) + 2\mathrm{COV}(\mathrm{X},\mathrm{Y})$$

$$VAR(X+Y) = VAR(X) + VAR(Y) + 2\mathrm{COV}(\mathrm{X},\mathrm{Y}) \qquad (1.12)$$

Si les deux variables sont indépendantes, on a $2\mathrm{COV}(\mathrm{X},\mathrm{Y}) = 0$ ainsi on a:

$$VAR(X+Y) = VAR(X) + VAR(Y)$$

- **Variance d'un produit** $\mathrm{VAR}(\mathrm{XY})$**:**

$$\mathrm{VAR}(\mathrm{XY}) = \mathrm{E}(\mathrm{X}^2\mathrm{Y}^2) - [\mathrm{E}(\mathrm{XY})]^2$$

Or $\quad \mathrm{COV}(X,\mathrm{Y}) = \mathrm{E}(\mathrm{XY}) - \mathrm{E}(\mathrm{X})\mathrm{E}(\mathrm{Y}) \Rightarrow$

$$\mathrm{E}(\mathrm{XY}) = \mathrm{COV}(X,\mathrm{Y}) + \mathrm{E}(\mathrm{X})\mathrm{E}(\mathrm{Y})$$

$$\mathrm{COV}(\mathrm{X}^2,\mathrm{Y}^2) = \mathrm{E}(\mathrm{X}^2\mathrm{Y}^2) - \mathrm{E}(\mathrm{X}^2)\mathrm{E}(\mathrm{Y}^2) \Rightarrow$$

$$\mathrm{E}(\mathrm{X}^2\mathrm{Y}^2) = \mathrm{COV}(\mathrm{X}^2,\mathrm{Y}^2) + \mathrm{E}(\mathrm{X}^2)\mathrm{E}(\mathrm{Y}^2)$$

Ainsi, on a : $\mathrm{VAR}(\mathrm{XY}) = \mathrm{COV}(\mathrm{X}^2,\mathrm{Y}^2) + \mathrm{E}(\mathrm{X}^2)\mathrm{E}(\mathrm{Y}^2) - \left[COV(X,Y) + \mathrm{E}(\mathrm{X})[\mathrm{E}(\mathrm{Y})]\right]^2$

On sait que : $\mathrm{E}(\mathrm{X}^2) = \mathrm{VAR}(\mathrm{X}) + [\mathrm{E}(\mathrm{X})]^2$ et $\mathrm{E}(\mathrm{Y}^2) = \mathrm{VAR}(\mathrm{Y}) + [\mathrm{E}(\mathrm{Y})]^2$

Ainsi : $\mathrm{VAR}(\mathrm{XY}) = \mathrm{COV}(\mathrm{X}^2,\mathrm{Y}^2) + [\,\mathrm{VAR}(\mathrm{X}) + [\mathrm{E}(\mathrm{X})]^2].[\,\mathrm{VAR}(\mathrm{Y}) + [\mathrm{E}(\mathrm{Y})]^2] - \left[COV(X,Y) + \mathrm{E}(\mathrm{X})[\mathrm{E}(\mathrm{Y})]\right]^2 \qquad (1.13a)$

A noter que dans le cas de l'indépendance:

$$\mathrm{COV}(\mathrm{X}^2,\mathrm{Y}^2) = \mathrm{COV}(\mathrm{X},\mathrm{Y}) = 0$$

Ce qui permet donc de réduire la formule à:

$$\mathrm{VAR}(\mathrm{XY}) = [\,\mathrm{VAR}(\mathrm{X}) + [\mathrm{E}(\mathrm{X})]^2].[\,\mathrm{VAR}(\mathrm{Y}) + [\mathrm{E}(\mathrm{Y})]^2] - \left[\mathrm{E}(\mathrm{X})[\mathrm{E}(\mathrm{Y})]\right]^2$$

Ainsi, en développant cette expression, on retrouve la formule initiale

$$\mathrm{VAR(XY)} = \mathrm{VAR(X)VAR(Y)} + \mathrm{VAR(X)[E(Y)]^2} + \mathrm{VAR(Y)[E(X)]^2} \qquad (1.13b)$$

1.3.2.3. Autres formules particulières

$$\mathrm{VAR(aX + b)} = \mathrm{a^2VAR(X)}$$

$$\mathrm{COV(aX; Y)} = \mathrm{aCOV(X; Y)}$$

1.4. Rappel sur les lois statistiques usuelles

1.4.1. La loi normale et le théorème central limite

Soit X une variable aléatoire suivant une loi normale de moyenne μ et de variance σ^2 . En considérant la variable aléatoire $\bar{X}$ telle que $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, cette la variable suit également une loi normale de moyenne μ mais de variance $\frac{\sigma^2}{n}$. Ainsi en centrant cette variable par μ et réduisant par la racine carrée de $\frac{\sigma^2}{n}$, on obtient une variable aléatoire Z suivant une loi normale de moyenne zéro et de variance égale à 1. Ce qui se présente comme suit :

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \rightsquigarrow N(0,1) \qquad (1.14)$$

Cette propriété dénommée théorème centrale limite se résume alors comme suit :

$$X \rightsquigarrow N(\mu, \sigma^2) \Rightarrow \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

1.4.2. La loi de khi-deux

Soit $X_1, \ldots, X_p$ une suite de variables aléatoires indépendantes normales, centrées réduites (c'est à dire de moyenne nulle et de variance égale à 1), alors la variable aléatoire :χ_p^2 est définie comme la somme des carrées de la variable X. Ainsi, on a :

$$\chi_p^2 = \sum_{i=1}^{p} X_i^2 \qquad (1.15)$$

suit une loi de khi-2 à p degrés de liberté.

Ainsi avec cette variable, on peut montrer que :

$$E(\chi_p^2) = p$$

$$VAR(\chi_p^2) = 2p$$

1.4.3. La loi de Student

La loi de Student est définie à partir du rapport entre une variable normale centrée réduite et une loi de khi-deux à p degrés de liberté. Elle est traduite comme suit :

$$t_p = \frac{X}{\sqrt{\chi_p^2/p}} \tag{1.16}$$

1.4.4. La loi de Fisher

La loi de Fisher est définie à partir du rapport entre deux variables indépendantes suivant chacune une loi de khi-deux respectivement à p et q degrés de liberté. Elle est traduite comme suit :

$$F_{p,q} = \frac{\chi_p^2/p}{\chi_q^2/q} \tag{1.17}$$

NB : Il est facile de montrer que le carré d'une variable de Student à q degrés de liberté est une variable de Fisher à 1 et q degrés de liberté.

1.5. Rappel sur les tests d'hypothèses

1.5.1. Forme générale d'un test d'hypothèse

1.5.1.1. Tests d'hypothèses simples

Le test d'hypothèse consiste à énoncer deux hypothèses sur un paramètre θ, dont une seule est vraie. Par exemple, on peut tester

-l'hypothèse nulle H0 que $\theta = \theta_0$,

-l'hypothèse alternative H1 que $\theta = \theta_1$.

L'objectif est de prendre une décision sur H0 qui consistera à rejeter H0 ou à ne pas rejeter H0. La décision est prise sur base des données observées, et peut donc conduire à deux types d'erreurs :

• Rejeter H0 alors que H0 est vraie, cette erreur est appelée erreur de première espèce. Elle est notée α. A noter que 1- α représente le seuil de confiance.

• Ne pas rejeter H0 alors que H0 est fausse, cette erreur est appelée erreur de deuxième espèce. Elle est notée β avec 1- β qui représente alors la puissance du test.

	H0 est vraie	H0 est fausse
Rejeter H0	α	1- β
Ne pas rejeter H0	$1-\alpha$	β

1.5.1.2. Tests d'hypothèses composites

Dans la pratique, les tests sont généralement des tests composites. En effet, les hypothèses sont généralement du type "Le paramètre θ est-il strictement plus grand qu'une certaine valeur θ_0 ?" Ce type d'hypothèse composite amène à la construction de test du type :

$$1\begin{cases}H_0\ \theta=\theta_0\\H_1\ \theta\neq\theta_0\end{cases}\quad 2\begin{cases}H_0\ \theta=\theta_0\\H_1\ \theta<\theta_0\end{cases}\quad 3\begin{cases}H_0\ \theta=\theta_0\\H_1\ \theta>\theta_0\end{cases}\quad 4\begin{cases}H_0\ \theta\geq\theta_0\\H_1\ \theta<\theta_0\end{cases}\qquad 5\begin{cases}H_0\ \theta\leq\theta_0\\H_1\ \theta>\theta_0\end{cases}$$

L'égalité doit toujours apparaître dans l'hypothèse nulle. Si la question est : " θ est-il strictement plus grand que θ_0?" On posera l'hypothèse alternative :

$$\text{H1} : \theta \geq \theta_0 \text{ et donc H0} : \theta \leq \theta_0.$$

1.5.1.3. Test de comparaison à une valeur théorique

Soit X une variable aléatoire suivant une loi normale de moyenne m et de variance σ^2. On montre d'abord que si $X \rightsquigarrow N(m,\sigma^2)$ alors :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow N\left(m,\frac{\sigma^2}{n}\right)$$

Ainsi avec le théorème central limite, on a

$$Z = \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1).$$

Cependant lorsque la variance σ^2 n'est pas connue, on utilise l'expression de la variance estimée telle que : $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$. Mais sachant que la variance estimée est la somme des carrés d'une loi normale, alors elle est distribuée selon une loi de χ^2 à n-1 degrés de libertés. Ainsi en appliquant le théorème central limite, on trouve :

$$T = \frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \rightsquigarrow t(n-1) \quad (1.18)$$

Cette propriété montre bien que la connaissance de la variance a donc une implication importante dans la conduite des tests d'hypothèse.

1.5.1.4. Détermination de la P-value d'un test

La P-value est la probabilité correspondant à la statistique Z, T, χ_p^2 ou F obtenue dans la construction du test. En prenant le cas d'une statistique qui suit une loi normale, la pvalue correspond à la probabilité à Z lue dans la table de la loi normale. Elle se définit comme la probabilité d'obtenir une statistique z qui soit supérieure à la statistique calculée Z :

$$Pvalue = P(z \geq Z)$$

$$Pvalue = 1 - P(z < Z)$$

Ainsi pour obtenir la pvalue, on lit d'abord dans la table de la loi normale la probabilité $P(z < Z)$. Ensuite, on calcule la pvalue.

La pvalue fournit aussi une règle décision dans le test. En effet, lorsque la pvalue est inférieure au seuil α, on rejette H0. Mais lorsque la pvalue est supérieure au seuil α on ne peut pas rejeter H0.

1.5.2. Test bilatéral

Soit X une variable aléatoire suivant une loi normale de moyenne m et de variance σ^2. On souhaite tester l'hypothèse suivante :

$$\begin{cases} H_0 \ \mu = m \\ H_1 \ \mu \neq m \end{cases}$$

1.5.2.1. Cas où la variance σ^2 est connue :

Lorsque σ^2 est connue, avec le théorème central limite, on peut poser :

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

En fixant le seuil de première espèce α, la statistique de ce test se présente comme suit (compte tenu du caractère bilatéral) :

$$P\left(\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} < -Z^*_{1-\frac{\alpha}{2}}\right) + P\left(\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} > Z^*_{1-\frac{\alpha}{2}}\right) = \alpha$$

$$P\left(Z < -Z^*_{1-\frac{\alpha}{2}}\right) + P\left(\mathrm{Z} > Z^*_{1-\frac{\alpha}{2}}\right) = \alpha$$

Mais sachant que $P\left(Z < -Z^*_{1-\frac{\alpha}{2}}\right) = P\left(Z > Z^*_{1-\frac{\alpha}{2}}\right)$ (propriété d'une loi symétrique), on a :

$$2\,P\left(\mathrm{Z} > Z^*_{1-\frac{\alpha}{2}}\right) = \alpha$$

Où Z est la statistique du test calculée et $Z^*_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite (lue dans la table de loi normale centrée réduite).

Par ailleurs sachant que $P\left(Z < -Z^*_{1-\frac{\alpha}{2}}\right) + P\left(\mathrm{Z} > Z^*_{1-\frac{\alpha}{2}}\right) = \alpha$, cela signifie que :

$$P\left(-Z^*_{1-\frac{\alpha}{2}} < Z < Z^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(|Z| < Z^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Dès lors on peut utiliser l'une des deux expressions pour prendre la décision du test : soit $2\,P\left(\mathrm{Z} > Z^*_{1-\frac{\alpha}{2}}\right) = \alpha$ qui exprime le seuil d'erreur ou $P\left(|Z| < Z^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$ qui exprime le seuil de confiance. Dans l'un ou l'autre des cas, on compare la valeur Z calculée à la valeur de $Z^*_{1-\frac{\alpha}{2}}$ lue dans la table de la loi normale. Ainsi lorsque $Z > Z^*_{1-\frac{\alpha}{2}}$, on rejette l'hypothèse H0. En revanche lorsque $Z < Z^*_{1-\alpha}$, on ne peut pas rejeter H0.

La région critique de ce test (encore appelée région de rejet de H0) se définit telle que :

$$RC = \left\{\left|\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}\right| > Z^*_{1-\frac{\alpha}{2}}\right\} \; soit \quad \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} \notin \left[-Z^*_{1-\frac{\alpha}{2}}\,;\, Z^*_{1-\frac{\alpha}{2}}\right]$$

Ou

$$RC = \left\{|\bar{X}-m| > Z^*_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} \; soit \; \bar{X} \notin \left]m - Z^*_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,;\, \mathrm{m} + Z^*_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

Connaissant donc la région critique, on peut définir la région d'acceptation de H0 sachant que $\left(\left|\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}\right| < Z^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$. Dès lors, la région d'acceptation se définit comme suit.

$$RA = \left\{ -Z^*_{1-\alpha} < \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} < Z^*_{1-\frac{\alpha}{2}} \right\}$$

Ou

$$RA = \left\{ \mathrm{m} - Z^*_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \bar{X} < \mathrm{m} + Z^*_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right\}$$

Ainsi connaissant la région de rejet ou la région d'acceptation, on peut donner une autre règle de décision par rapport au test. En effet après avoir calculée la moyenne $\bar{X}$ sur l'échantillon, on regarde si sa valeur appartient ou pas à la région d'acception. Ainsi si $\bar{X}$ appartient à l'intervalle RA, on ne peut pas rejeter H0. Par contre si $\bar{X}$ appartient de RC, on rejette H0.

1.5.2.2. Cas où la variance σ^2 n'est pas connue

Lorsque la variance σ^2 n'est pas connue, on utilise la variance estimée telle que : $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Dès lors, en appliquant le théorème central limite on trouve une loi de Student qui se présente comme suit :

$$\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \rightsquigarrow t(n-1)$$

Dans cette configuration, en fixant le seuil de première espèce α, la statistique du test se présente comme suit (compte tenu au du caractère bilatéral) :

$$P\left(\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < -T^*_{1-\frac{\alpha}{2}}\right) + P\left(\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} > T^*_{1-\frac{\alpha}{2}}\right) = \alpha$$

$$P\left(T < -T^*_{1-\frac{\alpha}{2}}\right) + P\left(T > T^*_{1-\frac{\alpha}{2}}\right) = \alpha$$

Mais sachant que $P\left(T < -T^*_{1-\frac{\alpha}{2}}\right) = P\left(T > T^*_{1-\frac{\alpha}{2}}\right)$ (propriété d'une loi symétrique), on a :

$$2\,P\left(T > T^*_{1-\frac{\alpha}{2}}\right) = \alpha$$

Où T est la statistique du test calculée et $T^*_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite (lue dans la table de loi normale centrée réduite).

Par ailleurs sachant que $P\left(T < -T^*_{1-\frac{\alpha}{2}}\right) + P\left(T > T^*_{1-\frac{\alpha}{2}}\right) = \alpha$, cela signifie que :

$$P\left(-T^*_{1-\frac{\alpha}{2}} < T < T^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(|T| < T^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Dès lors on peut utiliser l'une des deux expressions pour prendre la décision du test : soit $2\,P\left(T > T^*_{1-\frac{\alpha}{2}}\right) = \alpha$ qui exprime le seuil d'erreur ou $P\left(|T| < T^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$ qui exprime le seuil de confiance. Dans l'un ou l'autre des cas, on compare la valeur T calculée à la valeur de $T^*_{1-\frac{\alpha}{2}}$ lue dans la table de la loi normale. Ainsi lorsque $T > T^*_{1-\frac{\alpha}{2}}$, on rejette l'hypothèse H0. En revanche lorsque $T < T^*_{1-\alpha}$, on ne peut pas rejeter H0.

La région critique de ce test (encore appelée région de rejet de H0) se définit telle que :

$$RC = \left\{ \left| \frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \right| > T^*_{1-\frac{\alpha}{2}} \right\} \; soit \quad \frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \notin \left[-T^*_{1-\frac{\alpha}{2}} \,;\, T^*_{1-\frac{\alpha}{2}} \right]$$

Ou

$$RC = \left\{ |\bar{X}-m| > T^*_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n-1}} \right\} \; soit \; \bar{X} \notin \left] m - T^*_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n-1}} \,;\, \mathrm{m} + T^*_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n-1}} \right[$$

Connaissant donc la région critique, on peut définir la région d'acceptation de H0 sachant que $\left(\left| \frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \right| < T^*_{1-\frac{\alpha}{2}} \right) = 1-\alpha$. Dès lors, la région d'acceptation se définit comme suit.

$$RA = \left\{ -T^*_{1-\alpha} < \frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < T^*_{1-\frac{\alpha}{2}} \right\}$$

Ou

$$RA = \left\{ \mathrm{m} - T^*_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n-1}} < \bar{X} < \mathrm{m} + T^*_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n-1}} \right\}$$

1.5.3. Test unilatéral (à droite)

$$\begin{cases} H_0 \ \mu = m \\ H_1 \ \mu > m \end{cases}$$

1.5.3.1. Cas où la variance σ^2 est connue

Lorsque la variance est connue la statistique du test sous H_0 suit une loi normale $N(0,1)$. Ainsi connaissant le seuil d'erreur α on définit la région critique telle que :

$$P\left(\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} > Z^*_{1-\alpha} \right) = \alpha$$

$$P(Z > Z^*_{1-\alpha}) = \alpha$$

Où Z est la statistique du test calculée et $Z^*_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite (lue dans la table de loi normale centrée réduite).

Ainsi lorsque $Z > Z^*_{1-\alpha}$, on rejette l'hypothèse H0. En revanche lorsque $Z < Z^*_{1-\alpha}$, on ne peut pas rejeter H0.

La région critique de ce test (encore appelée région de rejet de H0) se définit alors comme suit :

$$RC = \left\{ \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} > Z^*_{1-\alpha} \right\}$$

Ou

$$RC = \left\{ \bar{X} > \mathrm{m} + Z^*_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\}$$

Connaissant donc la région critique, on peut définir la région d'acceptation de H0 sachant que $\left(\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} < Z^*_{1-\alpha} \right) = 1-\alpha$. Dès lors, la région d'acceptation se définit comme suit.

$$RA = \left\{ \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} < Z^*_{1-\alpha} \right\} \ soit \ \ \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \in]-\infty \, ; \, Z^*_{1-\alpha}]$$

Ou

$$RA = \left\{\bar{X} < \mathrm{m} + Z^*_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right\} \; soit \;\; \bar{X} \in \left]-\infty\,;\; \mathrm{m} + Z^*_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right]$$

Détermination de la P-value du test :

La pvalue de ce test se détermine en calculant la probabilité $1 - P(z < Z)$ où $P(z < Z)$ est la probabilité correspondant à Z dans la table de la loi normale.

1.5.3.2. Cas où la variance σ^2 n'est pas connue :

Lorsque σ^2 n'est pas connue, on a :

$$\frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \rightsquigarrow T(n-1)$$

Ainsi connaissant le seuil d'erreur α on définit la région critique du test telle que :

$$P\left(\frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} > T^*_{1-\alpha}\right) = \alpha$$

$$P(T > T^*_{1-\alpha}) = \alpha$$

Où T est la statistique du test calculée et $T^*_{1-\alpha}$ le quantile d'ordre $1 - \alpha$ de la loi de la loi de Student.

Ainsi lorsque $T > T^*_{1-\alpha}$, on rejette l'hypothèse H0. Et lorsque $T < T^*_{1-\alpha}$, on ne peut pas rejeter H0.

La région critique du test se définit comme suit :

$$RC = \left\{\frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} > T^*_{1-\alpha}\right\}$$

Ou

$$RC = \left\{\bar{X} > \mathrm{m} + T^*_{1-\alpha}\frac{\hat{\sigma}}{\sqrt{n-1}}\right\}$$

Connaissant donc la région critique, on peut définir la région d'acceptation de H0 sachant que $\left(\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < T^*_{1-\alpha}\right) = 1 - \alpha$. Dès lors, la région d'acceptation se définit comme suit.

$$RA = \left\{ \frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < T_{1-\alpha}^{*} \right\} \; soit \quad \frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \in]-\infty \; ; \; T_{1-\alpha}^{*}]$$

Ou

$$RA = \left\{ \bar{X} < \mathrm{m} + T_{1-\alpha}^{*} \frac{\hat{\sigma}}{\sqrt{n-1}} \right\} \; soit \; \bar{X} \in \left] -\infty \; ; \; \mathrm{m} + T_{1-\alpha}^{*} \frac{\hat{\sigma}}{\sqrt{n-1}} \right]$$

Détermination de la P-value du test :

La pvalue de ce test se détermine en calculant la probabilité $1 - P(t < T)$ où $P(t < T)$ est la probabilité correspondant à t dans la table de la loi normale.

1.5.4. Test unilatéral (à gauche)

$$\begin{cases} H_0 \;\; \mu = m \\ H_1 \;\; \mu < m \end{cases}$$

1.5.4.1. Cas où la variance σ^2 est connue

Lorsque la variance est connue la statistique du test sous H_0 suit une loi normale $N(0,1)$. Ainsi connaissant le seuil d'erreur α on définit la région critique telle que :

$$P\left(\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} < -Z_{1-\alpha}^{*} \right) = \alpha$$

$$P(Z < -Z_{1-\alpha}^{*}) = \alpha$$

Où Z est la statistique du test calculée et $Z_{1-\alpha}^{*}$ le quantile d'ordre $1 - \alpha$ de la loi normale centrée réduite. Ainsi lorsque $Z < Z_{1-\alpha}^{*}$, on rejette l'hypothèse H0. Et lorsque $Z > Z_{1-\alpha}^{*}$, on ne peut pas rejeter H0.

La région critique du test se définit alors comme suit :

$$RC = \left\{ \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} < -Z_{1-\alpha}^{*} \right\}$$

Ou

$$RC = \left\{ \bar{X} < \mathrm{m} - Z_{1-\alpha}^{*} \frac{\sigma}{\sqrt{n}} \right\}$$

Sachant que $\left(\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} > -Z_{1-\alpha}^{*}\right) = 1-\alpha$, on peut définir la région d'acceptation de H0 comme suit.

$$RA = \left\{ \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} > -Z_{1-\alpha}^{*} \right\} \; soit \quad \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} \in]-Z_{1-\alpha}^{*} ; +\infty]$$

Ou

$$RA = \left\{ \bar{X} > \mathrm{m} - Z_{1-\alpha}^{*} \frac{\sigma}{\sqrt{n}} \right\} \; soit \; \bar{X} \in \left] \mathrm{m} - Z_{1-\alpha}^{*} \frac{\sigma}{\sqrt{n}} ; +\infty \right]$$

Détermination de la P-value du test :

La pvalue de ce test se détermine en calculant la probabilité $1 - P(z < Z)$ où $P(z < Z)$ est la probabilité correspondant à Z dans la table de la loi normale.

1.5.4.2. Cas où la variance σ^2 n'est pas connue

Lorsque σ^2 n'est pas connue, sachant que $\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \rightsquigarrow T(n-1)$ et connaissant le seuil d'erreur α on définit la région critique du test telle que :

$$P\left(\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < -T_{1-\alpha}^{*}\right) = \alpha$$

$$P(T < -T_{1-\alpha}^{*}) = \alpha$$

Où T est la statistique du test calculée et $T_{1-\alpha}^{*}$ le quantile d'ordre $1-\alpha$ de la loi de la loi de Student.

Ainsi lorsque $T < -T_{1-\alpha}^{*}$, on rejette l'hypothèse H0. Et lorsque $T > -T_{1-\alpha}^{*}$, on ne peut pas rejeter H0.

La région critique du test se définit comme suit :

$$RC = \left\{ \frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < -T_{1-\alpha}^{*} \right\}$$

Ou

$$RC = \left\{ \bar{X} < \mathrm{m} - T_{1-\alpha}^{*} \frac{\hat{\sigma}}{\sqrt{n-1}} \right\}$$

Connaissant donc la région critique, on peut définir la région d'acceptation de H0 sachant que $\left(\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} > -T^*_{1-\alpha}\right) = 1-\alpha$. Dès lors, la région d'acceptation se définit comme suit.

$$RA = \left\{ \frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} > -T^*_{1-\alpha} \right\} \; soit \quad \frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \in \left]-T^*_{1-\alpha};\ +\infty\ \right[$$

Ou

$$RA = \left\{ \bar{X} > \mathrm{m} - T^*_{1-\alpha} \frac{\hat{\sigma}}{\sqrt{n-1}} \right\} \; soit \; \bar{X} \in \left] \mathrm{m} - T^*_{1-\alpha} \frac{\hat{\sigma}}{\sqrt{n-1}}; +\infty \right]$$

Détermination de la P-value du test :

La pvalue de ce test se détermine en calculant la probabilité $1 - P(t < T)$ où $P(t < T)$ est la probabilité correspondant à t dans la table de la loi normale.

1.6. Les règles d'utilisation des tables statistiques usuelles

1.6.1. Utilisation de la table de la loi normale centrée réduite

La table de la loi normale centrée réduite présente sur la première ligne et la première colonne les valeurs des fractiles Z (encore appelés quantiles). Dans la première colonne, on lit la valeur du quantile Z à un décimal près. Et sur la première ligne, on lit le nombre de décimaux restants à 10-2 près. Ainsi c'est en faisant la somme d'un élément en colonne et d'un élément en ligne qu'on obtient la valeur de Z. Par exemple $Z = 1.96$ s'obtient en faisant 1,9+0,06 où 1,9 provient de la première colonne alors que 0,06 provient de la première ligne. Ainsi pour trouver la valeur de n'importe quel fractile, on procède à cette décomposition. Par exemple $Z = 3.37$ se lit en décomposant 3.3 (en colonne) et 0,07 (en ligne).

En plus de la première colonne extérieure et la première ligne extérieure (qui permettent de connaitre la valeur de Z), on se réfère aux cellules intérieures pour lire les probabilités associées aux fractiles. La probabilité associée à une fractile correspond à la valeur contenue dans la cellule qui se trouve au croisement des deux membres qui forment la valeur de la fractile. Par exemple, sachant que 2,47 est formée par 2,4(en ligne) et 0,07 (en colonne), alors, la probabilité correspondant à 2,47 se trouve au croisement entre 2,4 et 0,07. Cette valeur est égale à 0,993.

Cette compréhension de la structure de la table de loi normale est extrêmement importante car elle servira à déterminer les fractiles (lorsque l'on connait les

probabilités) ou à l'inverse déterminer les probabilités (lorsque l'on connait les fractiles).

1.6.1.1 Lecture des fractiles connaissant les probabilités α

Dans une optique de détermination de la statistique d'un test suivant une loi normale et dont le seuil d'erreur est α, on lit le fractile correspondant à α. Pour cela, on calcule d'abord$(1-\alpha)+\frac{\alpha}{2}$ c'est-à-dire $1-\frac{\alpha}{2}$.. Ensuite, on recherche cette valeur dans les cellules intérieures de la table. Une fois cette valeur identifiée, on fait la somme des deux cellules extérieures (en ligne et en colonne) dont le croisement correspond à cette valeur $1-\frac{\alpha}{2}$ lue dans la table. Par exemple, pour le trouver la statistique (le fractile) correspondant à α =5 %, on calcule d'abord $1-\frac{\alpha}{2}$ (soit 0,975). Ensuite, en recherchant 0,975 dans les cellules intérieures de la table, on constate que cette valeur se trouve au croisement entre 1,9 et 0.06. Par conséquent le fractile correspondant à 5% est 1,96.

Notons aussi que cette valeur peut être obtenue avec la plus part des logiciels statistiques et économétriques plus ou moins spécialisés. Par exemple, certaines fonctions de Microsoft™ Excel® fournissent les valeurs contenues dans les tables statistiques usuelles. Pour obtenir le fractile correspondant à au seuil α, on utilise la formule suivante :

$$= loi.normale.standard.inverse(1-\frac{\alpha}{2})$$

Où α représente la probabilité (et correspond généralement au seuil d'erreur).

Remarque :

Puisque la loi normale est une loi symétrique, si l'on veut déterminer la valeur opposée du fractile (en vue par exemple de la détermination d'un intervalle de confiance (ou autre), on considère juste l'opposé de ce fractile pour trouver la borne inférieure de l'encadrement.

1.6.1.2. Lecture des probabilités α connaissant les fractiles

Lorsque l'on connait le fractile, pour déterminer la probabilité correspondante par lecture d'une table de la loi normale centrée et réduite, on décompose d'abord ce fractile en deux éléments (selon la méthode précédemment discutée). Ensuite, on recherche la cellule intérieure de la table se trouvant au croisement (de la ligne et de la colonne extérieure) des deux valeurs. Ainsi, après avoir déterminé cette valeur notée P, on calcule α telle que $1-\frac{\alpha}{2}=P$ soit $\alpha=2(1-P)$. Cette valeur α correspond donc à la probabilité recherchée. Par exemple, pour trouver la probabilité correspondant à 1,53, on décompose d'abord cette valeur entre 1,5 et 0,03. En recherchant dans les cellules intérieures de la table la probabilité se

trouvant au croisement de ces deux valeurs, on trouve 0,937. Ainsi puisque cette valeur équivaut à $1-\frac{\alpha}{2}$, on peut alors en déduire α comme $\alpha = 2(1-0{,}937)$. Soit $\alpha = 0{,}126 = 12{,}6\%$

Notons aussi que valeur de la probabilité peut s'obtenir en utilisant également les fonctions de Microsoft Excel. Pour cela, on peut utiliser la formule suivante :

$$= loi.normale.standard.n(q; VRAI)$$

Où q représente la valeur du fractile dont on cherche à déterminer la probabilité.

Remarque :

Dans une démarche de test, la valeur de la probabilité ainsi calculée correspond généralement à la p.value lorsque la détermination de la probabilité porte sur une statistique de test. En effet dans un test, on connaît à priori le seuil théorique α. Ce seuil est utilisé pour lire la statistique théorique (ou seuil critique) du test. On a alors le couple (α ; S^*). Ensuite en construisant le test et en calculant la statistique du test sous l'hypothèse nulle, on obtient S. Ce qui manque alors c'est la probabilité associée à cette statistique calculée. Elle est dénommée la p.value p_0. Dès lors, pour déterminer la p.value afin de former le couple (p_0 ; S), on lit la probabilité correspondant à S dans la table. C'est donc après avoir définie cette probabilité qu'on forme la règle de décision du test :

- Si $S > S^* \Rightarrow p_0 < \alpha$ alors on rejette H0.
- Si $S < S^* \Rightarrow p_0 > \alpha$ alors on ne peut pas rejeter H0.

1.6.1.2. Lecture de la table de la loi normale dans le cas d'un encadrement de la fractile

- **Cas d'un encadrement de type** : $P\,(-b < Z < b)$

Lorsque le fractile Z est une valeur encadrée par une borne supérieure b et une borne inférieure -b, pour lire la probabilité pour que Z soit compris entre -b et b, on procède d'abord à un développement comme suit :

$$P\,(-b < Z < b) = P(Z < b) - P(Z < -b)$$

Or, sachant les propriété d'une loi symétrique, on a: $P(Z < -b) = P(Z > b)$. Mais on sait aussi que $P(Z > b) = 1 - P(Z < b)$. Dès lors, on a : $P\,(-b < Z < b) = P(Z < b) - [\,1 - P(Z < b)]$. Au final, après développement, on trouve : $P\,(-b < Z < b) = 2P(Z < b) - 1$.

Cela montre donc que dans une loi symétrique, pour trouver la probabilité d'un encadrement symétrique (qui correspond en général au seuil de confiance), il faut simplement multiplier par 2 la probabilité obtenue en considérant uniquement la borne supérieure (en suivant les méthodes de lectures précédemment présentées). Ensuite retrancher 1 pour trouver la probabilité de l'encadrement (au seuil de confiance). Par exemple, quand on demande de calculer la probabilité pour que Z soit comprise entre -2,72 et 2,72. On lit d'abord la probabilité associée à 2,72 (soit 0,9967). Ensuite, on multiplie cette valeur par 2 et on retranche 1. On trouve alors 0,9934. Ainsi le seuil d'erreur α s'obtient simplement comme est 1-0,9934 soit 0,66%. Il faut noter que dans un encadrement α n'est pas calculée telle que $1-\frac{\alpha}{2}=P$ mais comme $1-\alpha=P$.

- **Cas d'un encadrement de type** $P\,(Z<b)$ ou de type $P\,(Z>b)$

Lorsqu'il s'agit d'un encadrement de type $P\,(Z<b)$, on garde sans aucune transformation la valeur lue dans la table (ou obtenue par la fonction : $=loi.normale.standard.n(q;VRAI)$. Ainsi, le seuil d'erreur α s'obtient en utilisant la relation $1-\alpha=P$.

Mais quand il s'agit d'un encadrement de type $P(U>b)$, on lit d'abord $P(U<b)$, ensuite on calcule $P(U>b)=1-P(U<b)$. Ainsi, le seuil d'erreur α s'obtient en utilisant la relation $1-\alpha=P$.

1.6.2. Utilisation de la table de Student

1.6.2.1. Lecture des fractiles connaissant les probabilités α

En général la table de Student se présente de telle sorte que les lignes correspondent aux degrés de liberté et les colonnes correspondent aux valeurs des probabilités. Pour utiliser une table se présentant sous ce format, on retrouve d'abord le degré de liberté, puis on lit sur la ligne correspondante (de gauche à droite) jusqu'à trouver la première valeur de t^* supérieure au t calculé. Et on retient en haut de la colonne la valeur P correspondante à cette valeur.

NB : Néanmoins, il faut noter que dans la table de Student, le quantile $1-\frac{\alpha}{2}$ se lit dans la colonne $P=\alpha$ alors que le quantile $1-\alpha$ se lit dans la colonne $P=2\alpha$. Cette distinction est importante car, elle permet de différencier la lecture de la table selon qu'il s'agit d'un test bilatéral $(1-\frac{\alpha}{2})$ ou d'un test unilatéral $(1-\alpha)$.

Notons aussi que pour déterminer le fractile d'ordre $1-\frac{\alpha}{2}$ ou $1-\alpha$ de la loi Student, on peut aussi utiliser la fonction Excel :

$= loi.student.inverse(\alpha\,; ddl)$ pour le cas d'un test bilatéral et
$= loi.student.inverse(2\alpha\,; ddl)$ pour le cas d'un test unilatéral

Aussi lorsque l'on veut déterminer la valeur symétrique (opposée) d'un fractile en vue, par exemple, de la détermination d'un intervalle de confiance, etc, on prend juste l'opposé du fractile calculée puisque la loi de Student est une loi symétrique.

Par ailleurs, il faut aussi noter que lorsque n est grand (n>30), on peut approximer la loi de Student par la loi normale. Dès lors, on peut utiliser la table de la loi normale comme décrite précédemment.

1.6.2.2. Lecture des probabilités connaissant les fractiles

Pour déterminer la probabilité α dans une table de la loi de Student, on se sert uniquement du fractile et du nombre de degré de liberté en prenant le chemin inverse qui conduit à la détermination du fractile. Dans la table de Student, on se place sur la ligne correspondant au nombre de degrés de liberté et on se déplace de gauche vers la droite en essayant d'identifier la valeur la plus proche possible du fractile recherché. Une fois la valeur du fractile identifiée, on retrouve la valeur de la probabilité en lisant dans le libellé de la colonne correspondant en haut de la table.
Cette procédure peut aussi être mise en œuvre en utilisant les fonctions d'Excel spécifiée comme suit :

$$= loi.student(q; ddl; 2)$$

Où q représente le fractile dont on cherche la probabilité ; ddl représente le nombre de degrés de liberté. L'option 2 signifie que le logiciel doit fournir directement la valeur α. En effet, en mettant l'option 1, on obtient $\frac{\alpha}{2}$ qu'il va falloir ensuite multiplier par 2 pour retrouver α.

1.6.3. Utilisation de la table de khi-deux

1.6.3.1. Lecture des fractiles connaissant les probabilités α

La lecture d'une table de khi-deux se fait de la même manière que la lecture de la table de Student discutée précédemment notamment pour ce qui concerne la recherche la recherche du fractile correspondant à un seuil donné.

Cependant la procédure diffère significativement lorsqu'il s'agit des encadrements car la loi de khi-deux n'est pas une loi symétrique. En effet, à la différence des précédentes lois, la loi de khi-deux n'est pas symétriquement distribuée autour de 0. Par conséquent lorsqu'on veut procéder, par exemple, à

un encadrement, on doit d'abord définir la probabilité associée chaque fractile constituant les bornes. Par exemple pour encadrer une valeur U dans la perspective de la détermination d'un intervalle de confiance etc.., on calcule d'abord deux probabilités :

$$p1 = \frac{\alpha}{2} \; et \;\; p2 = \frac{\alpha}{2} + (1 - \alpha).$$

Ensuite, on lit les fractiles correspondant à chaque probabilité (en utilisant les degrés de liberté). Ensuite, on encadre U telle que $q1 < U < q2$ où $q1$ et $q2$ représentent respectivement les fractiles correspondants à $p1$ et $p2$. Cet encadrement se fait donc de telle sorte que $P(q1 < U < q2) = 1 - \alpha$. Où $1 - \alpha$ est le seuil de confiance. Pour executer cette prodcéure sous excel, on procède comme suit : $= loi.khideux.inverse(p1; ddl)$ et $= loi.khideux.inverse(p2; ddl)$.

Les valeurs obtenues servent donc à construire l'intervalle de confiance.

1.6.3.2. Lecture des probabilités connaissant les fractiles

Là également, il n'y pas de différence entre la procédure de lecture d'une loi de khi-deux et une loi de Student pour ce qui concerne la recherche d'une probabilité simple. Par conséquent, on peut se référer à la méthode discutée pour la loi de Student.

En revanche lorsqu'il s'agit de déterminer la probabilité lorsque le fractile est fourni sous forme d'encadrement, la procédure est un peu particulière.

En effet, Puisque, nous avons deux bornes, pour obtenir la probabilité correspondant à la fractile inférieure (c'est-à-dire pour obtenir $\frac{\alpha}{2}$), on recherche juste cette valeur dans la colonne où l'on identifié le fractile. Ensuite, on multiplie par 2 pour obtenir α. De la même manière, on peut se servir de la fractile supérieure pour déterminer la probabilité correspondant à $\frac{\alpha}{2} + (1 - \alpha)$ qui permet ensuite d'obtenir α. Toutes ces procédure peuvent être mis en œuvre sous excel, en utilisant l'une des formules suivantes :

$$= loi.khideux(q1; ddl)$$
$$= loi.khideux(q2; ddl$$

L'une ou l'autre de ces deux valeurs obtenues permettra alors de calculer α et par conséquent 1-α .

CHAPITRE 2 : LE MODELE LINEAIRE SIMPLE

Nous cherchons à étudier la relation entre deux variables Y et X. Y est la variable que on cherche à expliquer (ou à prédire), on parle de variable endogène (dépendante) ; X est la variable explicative (prédictive), on parle de variable exogène (indépendante). L'équation mathématique qui permet de relier Y à X est appelée modèle. Ce modèle est dit simple lorsqu'il existe qu'une seule variable explicative. La forme générale du modèle linéaire simple est la suivante :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (2.1)$$

y_i représente la variable expliquée (encore appelée variable dépendante). x_i représentent la variable explicative (encore appelée variable indépendante). ε_i représente les perturbations aléatoires ou les résidus. Les coefficients β_0 et β_1 sont les paramètres à estimer. Ils représentent les effets des variables explicatives sur la variable expliquée. Par exemple, β_1 mesure l'impact de la variable x_i sur la variable y_i.

Il faut simplement remarquer que dans l'équation (2.1), les variables y et x sont observées alors que les paramètres β_0, β_1 et les perturbations aléatoires sont inobservés. Le terme aléatoire ε_i, que l'on appelle l'erreur du modèle permet de résumer toute l'information qui n'est pas prise en compte dans la relation linéaire que l'on cherche à établir entre Y et X c.-à-d. les problèmes de spécifications, l'approximation par la linéarité, etc.

Les paramètres β_0, β_1 sont estimés dans une procédure appelée régression. La régression linéaire consiste à trouver une droite qui ajuste au mieux un nuage de points formé par les couples (X, Y). Pour cela plusieurs techniques peuvent être utilisées dont notamment la méthode des moindres carrés ordinaires et la méthode de maximum de vraisemblance.

2.1. Estimation par les moindres carrés ordinaires

L'estimation du modèle linéaire simple par les Moindres Carrés Ordinaires MCO consiste à chercher la droite qui minimise la somme des carrés des résidus :

$$l(\beta_0, \beta_1) = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} \varepsilon^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (2.2)$$

Le minimum de cette fonction s'obtient en annulant les dérivées partielles par rapport à β_0 et β_1.

$$\begin{cases} \dfrac{\partial l(\beta_0,\beta_1)}{\beta_0} = -\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i) = 0 & (2.3a) \\ \dfrac{\partial l(\beta_0,\beta_1)}{\beta_1} = -\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)x_i = 0 & (2.3b) \end{cases}$$

Par simplification, on a :

$$\begin{cases} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)x_i = 0 \end{cases}$$

Ces deux équations sont appelées « ***équation normales*** ». On peut simplement noter que comme $y_i - \beta_0 - \beta_1 x_i = \varepsilon_i$, alors les équation normales peuvent aussi s'écrire sous la nouvelle forme suivante :

$$\begin{cases} \sum_{i=1}^{n} \varepsilon_i = 0 \\ \sum_{i=1}^{n} x_i \varepsilon_i = 0 \end{cases}$$

Ces deux propriétés sont extrêmement importantes. Elles montrent d'une part que la somme des résidus est nulle et d'autre part que le produit croisé entre les résidus et la variable explicative est aussi nulle. Mais comme on le verra un peu plus loin, la première propriété n'est plus vérifiée lorsqu'il n'y pas de constante dans le modèle. Ce qui a aussi quelques implications.

En revenant à la forme explicite des équation normales, on a un système de deux équations à deux inconnues, qui peuvent également s'écrire :

$$\begin{cases} \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0 \end{cases}$$

En divisant la première équation par n, on retrouve :

$$\begin{cases} \bar{y} - \beta_0 - \beta_1 \bar{x} = 0 \\ \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0 \end{cases}$$

La première équation montre que la droite passe par le point moyen ($\bar{x}$, $\bar{y}$). Ce qui permet donc de poser que :

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \qquad (2.4a)$$

En remplaçant β_0 par sa valeur dans la seconde équation divisée par n, on a :

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta_1 \bar{x})\bar{x} - \beta_1 \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \beta_1 \left(\frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2 \right) = 0$$

$$S_{xy} - \beta_1 S_x^2 = 0 \qquad (2.4b)$$

Ainsi la solution au problème de minimisation de la somme des carrés de résidus se présente comme suit :

$$\begin{cases} \hat{\beta}_1 = \dfrac{S_{xy}}{S_x^2} & (2.5a) \\ \hat{\beta}_0 = \bar{y} - \dfrac{S_{xy}}{S_x^2}\bar{x} & (2.5b) \end{cases}$$

$$\begin{cases} \hat{\beta}_1 = \dfrac{COV(x,y)}{VAR(x)} & (2.6a) \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} & (2.6b) \end{cases}$$

La droite de régression s'écrit alors comme :

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i \qquad (2.7)$$

Avec $\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$ et $\hat{\beta}_0 = \overline{y} - \frac{S_{xy}}{S_x^2}\overline{x}$

2.1.1. Les valeurs ajustées (ou valeurs prédites) du modèle

Les valeurs ajustées notées $\widehat{y_i}$ sont obtenues au moyen de la droite de régression :

$$\widehat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2.8}$$

2.1.2. Les hypothèses de base sur les résidus de régression

Le résidu est la partie inexpliquée de y_i par la droite de régression. C'est la différence entre la valeur observée y_i de la variable dépendante et sa valeur ajustée $\widehat{y_i}$. Ils sont calculés comme suit :

$$\varepsilon_i = y_i - \widehat{y_i} \tag{2.9}$$

Propriétés des résidus

La première hypothèse qui conditionne la validité de l'estimateur des moindres carrés ordinaires est que l'espérance des résidus soit nulle. Cette propriété se traduit par l'expression suivante :

$$E(\varepsilon_i) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i = \frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y_i}) = 0$$

$$E(\varepsilon_i) = 0 \tag{2.10}$$

Cette propriété est l'une des hypothèses fondamentales dans l'estimation par les moindres carrés ordinaires.

La seconde hypothèse concernant les résidus est celle d'homocédasticité qui signifie que la variance des résidus est constante. Elle est notée comme suit :

$$VAR(\varepsilon_i) = E({\varepsilon_i}^2) - [E(\varepsilon_i)]^2 = E({\varepsilon_i}^2) = \sigma_\varepsilon^2 \tag{2.11}$$

Une troisième hypothèse est la non-corrélation entre les résidus et les variables explicatives du modèle. On dit alors qu'il y a orthogonalité (ou indépendance) entre les résidus et les variables explicatives. Cette indépendance se traduit par une covariance nulle entre la série des résidus ε_i et la série des x_i. Ce qui se traduit comme suit :

$$COV(x_i, \varepsilon_i) = E(x_i\varepsilon_i) - E(x_i)E(\varepsilon_i) = E(x_i\varepsilon_i) = \frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i = 0$$

$$COV(x_i, \varepsilon_i) = \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i = 0 \qquad (2.12a)$$

La quatrième hypothèse stipule que les résidus sont non corrélés entre eux, en d'autres termes la covariance entre deux résidus i et j est toujours égale à 0.

$$COV(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) - E(\varepsilon_i)E(\varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \quad \forall\, i \neq j \qquad (2.12b)$$

Le non-respect de l'une des quatre hypothèses entraîne l'invalidité de l'estimation par les moindres carrés ordinaires. Dès lors, il apparaît important de s'assurer que ces hypothèses soient vérifiées avant d'utiliser cette méthode d'estimation.

Par ailleurs, il existe une hypothèse supplémentaire sur la distribution des résidus. C'est l'hypothèse de normalité des résidus. On suppose que les résidus suivent une loi normale de moyenne nulle et de variance σ_ε^2.

$$\varepsilon_i \rightsquigarrow N(0, \sigma_\varepsilon^2)$$

Il faut noter que cette hypothèse n'est pas une condition nécessaire de la validité de l'estimateur des MCO. Celui-ci cherche simplement à minimiser la somme des carrés des résidus. Peu importe donc la loi suivie la série des résidus dans cette méthode d'estimation. Il faut juste que les résidus soient indépendants et identiquement distribués.

En plus de ces cinq hypothèses sur la série des résidus, il existe aussi des hypothèses sur la série des variables explicatives. En effet, on suppose que la série de X n'est pas stochastique c'est-à-dire que X est non aléatoire. Cette hypothèse signifie que son espérance et sa variance sont constantes. En revanche la variable Y est un variable stochastique car sa valeur est influencée par les perturbations provenant des ε_i

2.1.3. Décomposition de la somme des carrés

2.1.3.1. Somme des carrés expliquée (SCE)

La SCE est la somme des écarts à la moyenne des valeurs ajustées (prédites)

$$SCE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad (2.13)$$

La SCE indique la variabilité expliquée par le modèle c.-à-d. la variation de Y expliquée par X.

2.1.3.2. Somme des carrés résiduelle (SCR)

La SCR est la somme des carrés des écarts aléatoires:

$$SCR = \sum_{i=1}^{n} {\varepsilon_i}^2 \qquad (2.14)$$

La SCR indique la variabilité non-expliquée (résiduelle) par le modèle c.-à-d. l'écart entre les valeurs observées de Y et celles prédites par le modèle.

2.1.3.3. Somme des carrés totale (SCT)

La SCT indique la variabilité totale de Y c.-à-d. l'information disponible dans les données. Elle est la somme de la SCE et de la SCR :

Démonstration

On sait que :

$$y_i = \hat{y_i} + \varepsilon_i$$

Ajoutons de part et d'autre de l'équation la moyenne de $\bar{y}$. On a :

$$y_i - \bar{y} = \hat{y_i} - \bar{y} + \varepsilon_i$$

Ainsi, on a :

$$(y_i - \bar{y}) = (\hat{y_i} - \bar{y}) + \varepsilon_i$$

Elevons les deux membres d'équation au carré on a :

$$(y_i - \bar{y})^2 = [(\hat{y_i} - \bar{y}) + \varepsilon_i]^2$$

En développant le membre de droite, on a :

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

En passant l'opérateur somme, on a :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2 + \sum_{i=1}^{n}{\varepsilon_i}^2 + 2\sum_{i=1}^{n}(\hat{y_i} - \bar{y})\varepsilon_i$$

Or d'après les équations normales, $\sum_{i=1}^{n}(\hat{y_i} - \bar{y})\varepsilon_i$ est nulle car, on a :

$$\sum_{i=1}^{n}(\hat{y_i} - \bar{y})\varepsilon_i = \sum_{i=1}^{n}\hat{y_i}\varepsilon_i - \sum_{i=1}^{n}\bar{y}\varepsilon_i = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i)\varepsilon_i - \bar{y}\sum_{i=1}^{n}\varepsilon_i$$

$$\sum_{i=1}^{n}(\widehat{y_i}-\bar{y})\varepsilon_i=\hat{\beta}_0\sum_{i=1}^{n}\varepsilon_i+\hat{\beta}_1\sum_{i=1}^{n}x_i\varepsilon_i-\bar{y}\sum_{i=1}^{n}\varepsilon_i$$

On sait d'après les équations normales que :

$$\sum_{i=1}^{n}\varepsilon_i=0$$

Par conséquent :

$$2\sum_{i=1}^{n}(\widehat{y_i}-\bar{y})\varepsilon_i=0$$

Ce qui, au final permet d'écrire que :

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\widehat{y_i}-\bar{y})^2+\sum_{i=1}^{n}{\varepsilon_i}^2$$

D'où : $$SCT=SCE+SCR \tag{2.15}$$

Ainsi, on a bien la somme des carrés totale égale à la somme des carrés expliquées et la somme des carrés résiduelle.

Nb : Cette propriété tient uniquement lorsqu'il existe une constante dans la régression. Mais lorsqu'il n'y a pas de constante dans l'équation estimée, la propriété sur la somme des carrés totale n'est plus vérifiée car $\sum_{i=1}^{n}\varepsilon_i\neq 0$.

Il faut aussi savoir que lorsque cette propriété n'est pas vérifiée, le test de significativité globale ou le calcul du R^2 ne sont plus valables.

2.1.4. Equation de décomposition de la variance

L'équation d'analyse de la variance se déduit de l'équation de la somme des carrés. Ainsi, en fonction de la SCT, la SCE et la SCR, on obtient la variance totale (VT), variance expliquée (VE) et variance résiduelle(VR) en divisant les sommes des carrés correspondantes par le nombre d'observations n.

$$VE=\frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i}-\bar{y})^2 \tag{2.16}$$

$$VR=\frac{1}{n}\sum_{i=1}^{n}{\varepsilon_i}^2 \tag{2.17}$$

$$VT = VE + VR = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}{\varepsilon_i}^2 \quad (2.18)$$

Sachant par ailleurs que la variance totale de y est $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ alors :

$$VT = S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}{\varepsilon_i}^2$$

La VE indique la variabilité expliquée par le modèle c.-à-d. la variation de Y expliquée par X. La VR indique la variabilité non-expliquée (résiduelle) par le modèle c.-à-d. l'écart entre les valeurs observées de Y et celles prédites par le modèle. La VT indique la variabilité totale de Y c.-à-d. l'information disponible dans les données.

2.1.5. Le coefficient de détermination : R^2 et R^2 ajusté

A partir de l'équation de décomposition de la variance, on peut déduire un indicateur synthétique capable d'indiquer la proportion de variance de la variable Y expliquée par le modèle. C'est le coefficient de détermination R^2.

En effet, en reprenant l'expression de la variance totale (ou de la somme des carrés totale), on pose :

$$VT = VE + VR$$

En divisant chaque membre de l'équation par VT, on trouve :

$$1 = \frac{VE}{VT} + \frac{VR}{VT}$$

Le coefficient de détermination de y par x noté R_{xy}^2 se définit comme la part de la variance de y expliquée par x. Cette définition correspond à l'expression $\frac{VE}{VT}$ qui indique le rapport entre la variance totale et la variance expliquée. Dans ce cas, on a :

$$R_{xy}^2 = \frac{VR}{VT} = 1 - \frac{VR}{VT} \quad (2.19a)$$

En remplaçant VR et VT par leur valeur et en simplifiant par n, on obtient :

$$R_{xy}^2 = 1 - \left(\frac{\sum_{i=1}^{n}{\varepsilon_i}^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}{\varepsilon_i}^2}\right)$$

$$R_{xy}^2 = 1 - \left(\frac{\sum_{i=1}^{n}{\varepsilon_i}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right)$$

$$R_{xy}^2 = 1 - \left(\frac{SCR}{SCT}\right) \tag{2.19b}$$

Le R_{xy}^2 est un indicateur de la qualité de l'ajustement. Il est compris entre 0 et 1. Quand il est proche de 1, le modèle sera considéré de bonne qualité.

Même si le R^2 reste un bon indicateur de la qualité de l'ajustement, son principal inconvénient provient du fait qu'il augmente mécaniquement lorsque l'on ajoute des variables explicatives supplémentaires. Ce qui signifie qu'il suffit d'ajouter arbitrairement les variables explicatives pour que le R^2 augmente. Ce qui affaiblit la parcimonie du modèle, c'est-à-dire sa capacité à décrire la réalité avec un nombre restreint de variables. C'est pour cette raison qu'on introduit la notion de R^2 ajusté calculé comme suit :

$$R_{ajusté}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SCR}{SCT} \tag{2.20}$$

Où $n-k-1$ est le nombre de degrés de liberté avec k le nombre de variables du modèle (constante exclue).

Propriétés : Lien entre le coefficient de détermination R^2 et le coefficient de corrélation r_{xy}.

Le coefficient de détermination est le carré du coefficient de corrélation.

Démonstration :

$$R^2 = \frac{SCE}{SCT} = \frac{VE}{VT} = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Dans un premier temps, remplaçons $\hat{y}_i$ par son expression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, on a :

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Ensuite $\hat{\beta}_0$ par son expression $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, on obtient

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_1(x_i - \bar{x})\right)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}_1{}^2\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_1(x_i - \bar{x})\right)^2\right)}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}_1{}^2 V(X)}{V(Y)}$$

$$= \frac{\left(\frac{COV(X,Y)}{V(X)}\right)^2 V(X)}{V(Y)} = \frac{(COV(X,Y))^2 V(X)}{(V(X))^2 V(Y)} = \frac{(COV(X,Y))^2}{V(X)V(Y)}$$

$$R^2 = \left(\frac{COV(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}\right)^2 = \left(r_{xy}\right)^2$$

$$R^2 = \left(r_{xy}\right)^2 \tag{2.21}$$

Par ailleurs, connaissant cette propriété, on peut le coefficient de corrélation linéaire multiple tel que $R = \sqrt{R^2}$ mais aussi le coefficient de corrélation simple. Mais dans ce cas, on utilise le signe du coefficient de régression simple comme suit :

$$r_{xy} = signe\left(\hat{\beta}_0\right) * R \tag{2.22}$$

2.1.6. Calcul de la variance estimée des résidus

La variance des résidus notée σ_ε^2 est en pratique inobservée. On doit alors proposer une valeur estimée de ce paramètre. Pour cela, on part de la formule de la variance résiduelle précédemment présentée en y apportant quelques ajustements. En effet :

$$\hat{\sigma}_\varepsilon^2 = \frac{n}{n-(k+1)}VR \tag{2.23}$$

$$Avec\ VR = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Où VR est la variance résiduelle calculée sur la série des résidus estimés, n est le nombre d'observations, k est le nombre de variables explicatives (ici égal à 1). Ainsi $(k+1)$ est le nombre total de paramètres à estimer, également appelé nombre de degré de liberté.

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k+1)} \tag{2.24}$$

Cette normalisation par $(k+1)$ a pour but d'obtenir un estimateur sans biais de de la variance des résidus car la $VR = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ est biaisé.

2.2. Propriétés des estimateurs : biais et convergence

Deux propriétés sont importantes dans l'évaluation d'un estimateur. (1) Est-ce qu'il est sans biais c.-à-d. est-ce qu'en moyenne nous obtenons la vraie valeur du paramètre ? (2) Est-ce qu'il est convergent c.-à-d. est-ce que l'estimation devient de plus en plus précise lorsque que la taille de l'échantillon tend vers l'infini ?

Le biais d'estimation est évalué en fonction de l'espérance tandis que la précision est évaluée en fonction de la variance de l'estimateur.

2.2.1. Le biais d'estimation

On dit que qu'un estimateur est sans biais si son espérance est égale à la vraie valeur du paramètre estimé. Par exemple, on dit qu'un estimateur $\hat{\theta}$ est sans biais si :

$$E(\hat{\theta}) = \theta$$

Où θ est la vraie valeur du paramètre.

Pour vérifier cette propriété sur les estimateurs des MCO de $\hat{\beta}_0$ $\hat{\beta}_1$ et $\hat{\sigma}_\varepsilon^2$, on calcule leur espérance respective. Ainsi, on a :

$$E(\hat{\beta}_k) = \beta_k \qquad k = 0;1 \tag{2.25}$$

$$E(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^2 \tag{2.26}$$

Pour arriver à ces résultats, il y a deux étapes principales. Dans un premier temps, on exprime le paramètre estimé en fonction de sa vraie valeur; Et dans un deuxième temps, on passe cette expression à l'opérateur d'espérance mathématique. D'une manière générale, en se basant sur les hypothèses de base énoncées sur le modèle linéaire, l'expression de simplifie généralement et se réduit à la vraie valeur si l'estimateur est non biaisé.

Démonstrations :

Démontrons que l'estimateur $\hat{\beta}_1$ de β_1 est sans biais.

On sait que : $\hat{\beta}_1 = \frac{COV(X,Y)}{VAR(X)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Faisons apparaître β_1 dans cette expression en remplaçant $(y_i - \bar{y})$. Pour cela, on part du modèle initial :

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

Passons d'abord cette équation à l'opérateur somme et divisons par n pour calculer $\bar{y}$. On a :

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \beta_0 + \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i$$

On obtient ainsi :

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$$

Formons la différence de cette équation avec l'équation initiale pour trouver $(y_i - \bar{y})$. On a :

$$-\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ \bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} \end{cases}$$
$$(y_i - \bar{y}) = \beta_1 (x - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})$$

On remplace $(y_i - \bar{y})$ par son expression dans l'expression de $\hat{\beta}_1$. On a :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})[\beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

En développant cette expression, on trouve :

$$\hat{\beta}_1 = \frac{\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Par ailleurs sachant que $\bar{\varepsilon} = 0$, on a :

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

On peut maintenant passer cette expression à l'opérateur d'espérance pour voir si $\frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (qui représente potentiellement le biais d'estimation) est nul. On a :

$$E(\hat{\beta}_1) = E\left(\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$E(\hat{\beta}_1) = E(\beta_1) + E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Mais puisque la variable x_i n'est pas stochastique (non aléatoire), alors $\frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ est aussi non stochastique. Or on sait que l'espérance d'une variable non stochastique est égale à une constante indépendante des erreurs. Par conséquent $E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(\varepsilon_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Notons aussi que β_1 est non stochastique, par conséquent $(\beta_1) = \beta_1$. Dès lors :

$$E(\hat{\beta}_1) = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(\varepsilon_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

D'après la première hypothèse sur les résidus, on a : $E(\varepsilon_i)$=0. Ainsi

$\sum_{i=1}^{n}(x_i - \bar{x})E(\varepsilon_i) = 0$. On obtient donc finalement que :

$$E(\hat{\beta}_1) = \beta_1$$

Ce qui permet de montrer que $\hat{\beta}_1$ est sans biais.

Cette démonstration permet de tirer une conclusion essentielle : l'estimateur des moindres carrés ordinaires (MCO) est sans biais, si et seulement si les deux hypothèses suivantes sont respectées : La variable X n'est pas stochastique (X est non aléatoire) et l'espérance de l'erreur est nulle.

Pour ce qui concerne $\hat{\beta}_0$ on peut démontrer qu'il est aussi sans biais en suivant la même procédure que pour $\hat{\beta}_1$. En effet, on sait que :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

En remplaçant $\bar{y}$ par son expression $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$ tirée de l'équation de l'équation initial du modèle, on obtient :

$$\hat{\beta}_0 = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{\varepsilon}$$

En passant cette équation à l'opérateur d'espérance sachant que $\bar{\varepsilon} = 0$, on a :

$$E(\hat{\beta}_0) = \beta_0 + \bar{x}\, E(\beta_1 - \hat{\beta}_1)$$

Sachant que $\hat{\beta}_1$ est un estimateur sans biais de β_1, alors $E(\beta_1 - \hat{\beta}_1) = 0$. Par conséquent :

$$E(\hat{\beta}_0) = \beta_0$$

Ce qui montre que l'estimateur $\hat{\beta}_0$ de β_0 est sans biais.

2.2.2. Convergence d'un estimateur

Un estimateur $\hat{\theta}$ sans biais de θ est convergent si et seulement si sa variance tend vers 0 lorsque le nombre d'observations (taille de l'échantillon) tend vers l'infini.

$$VAR(\hat{\theta}) \to 0 \; quand \; n \to \infty$$

Où n est le nombre d'observations.

Un estimateur consistant est un estimateur qui converge en probabilité vers le paramètre qu'il tente d'estimer et qui se concentre de plus en plus autour de ce paramètre à mesure que le nombre d'observations augmente. Ainsi, la variance d'un estimateur consistant diminue à mesure qu'on augmente la taille de l'échantillon.

Démonstrations :

Ainsi, pour démontrer qu'un estimateur est convergent, il faut d'abord expliciter l'expression de la variance de cet estimateur et montrer, par la suite que cette variance tend vers 0 lorsque n tend vers ∞.

On a démontré précédemment que :

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Ainsi sachant que la variance d'une somme est égale à la somme des variances dans le cas des variables indépendantes, l'expression ci-dessus permet donc d'écrire que :

$$VAR(\hat{\beta}_1) = VAR(\beta_1) + VAR\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

β_1 étant une valeur fixe (non aléatoire), $VAR(\beta_1) = 0$. Ainsi on a :

$$VAR(\hat{\beta}_1) = VAR\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

En changeant l'indice au dénominateur on a :

$$VAR(\hat{\beta}_1) = VAR\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right)$$

Par ailleurs, x_i étant une variable non stochastique, alors $\frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$ est non stochastique et on peut noter cette expression comme $\frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \omega_i$. Ainsi on a :

$$VAR(\hat{\beta}_1) = VAR\left(\sum_{i=1}^{n}\omega_i\varepsilon_i\right) = E\left[\left(\sum_{i=1}^{n}\omega_i\varepsilon_i\right)^2\right]$$

On peut, par ailleurs démontrer que :

$$\left(\sum_{i=1}^{n}\omega_i\varepsilon_i\right)^2 = \sum_{i=1}^{n}(\omega_i\varepsilon_i)^2 + 2\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\omega_i\omega_j\varepsilon_i\,\varepsilon_j\right)$$

Ainsi, on a :

$$VAR(\hat{\beta}_1) = E\left[\sum_{i=1}^{n}(\omega_i\varepsilon_i)^2 + 2\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\omega_i\omega_j\varepsilon_i\,\varepsilon_j\right)\right]$$

$$VAR(\hat{\beta}_1) = \sum_{i=1}^{n} \omega_i{}^2 E(\varepsilon_i{}^2) + 2\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \omega_i \omega_j E(\varepsilon_i\, \varepsilon_j)\right)$$

Or on sait que : $E(\varepsilon_i{}^2) = VAR(\varepsilon_i) = \sigma_\varepsilon^2$ et que $E(\varepsilon_i \varepsilon_j) = COV(\varepsilon_i, \varepsilon_j) = 0$. Par conséquent :

$$VAR(\hat{\beta}_1) = \sigma_\varepsilon^2 \sum_{i=1}^{n} \omega_i{}^2$$

Mais sachant que $\omega_i = \frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$ alors on a : :

$$VAR(\hat{\beta}_1) = \sigma_\varepsilon^2 \sum_{i=1}^{n}\left(\frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right)^2 = \frac{\sigma_\varepsilon^2}{\left(\sum_{j=1}^{n}(x_j - \bar{x})^2\right)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{\sigma_\varepsilon^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$VAR(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Ainsi, connaissant cette expression de la variance, on peut facilement montrer que celle-ci tend vers 0 lorsque n tend vers l'infini. En effet :

$$n \to \infty \Rightarrow \sum_{i=1}^{n}(x_i - \bar{x})^2 \to \infty \Rightarrow$$

$$VAR(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\infty} \to 0$$

Ce qui permet donc de démontrer l'estimateur $\hat{\beta}_1$ est convergent.

En suivant la même démarche, on peut montrer que $\hat{\beta}_0$ est aussi un estimateur convergent car :

$$VAR(\hat{\beta}_0) = \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

Ainsi, on peut montrer que :

$$n \to \infty \Rightarrow \sum_{i=1}^{n}(x_i - \bar{x})^2 \to \infty \Rightarrow$$

$$VAR(\hat{\beta}_0) \to 0$$

Tableau 2.1 : Récapitulatif sur les paramètres

Estimateur	Expression simple	Expression développée	Espérance	Variance
s	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$	$\beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{\varepsilon}$	β_0	$\sigma_\varepsilon^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]$
$\hat{\beta}_1$	$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$	β_1	$\frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$
$\hat{\sigma}_\varepsilon^2$	$\frac{\sum_{i=1}^n \hat{\varepsilon}_i{}^2}{n-(k+1)}$	$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-(k+1)}$	σ_ε^2	

$$COV(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma_\varepsilon^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{2.27}$$

2.3. Inférence statistique

Jusque maintenant, nous avons étudié les estimateurs et leur variance. Mais pour pouvoir faire des tests ou établir des intervalles de confiance, il faut disposer de toute la distribution des estimateurs. C'est pourquoi cette section se focalise sur l'étude des distributions des paramètres estimés.

2.3.1. Les lois de distributions des paramètres estimés

2.3.1.1. Distribution de la variance estimée de l'erreur

Il apparaît important de connaître la distribution de la variance estimée de l'erreur pour pouvoir déterminer la distribution des coefficients estimés lorsque nous introduirons σ_ε^2 dans les expressions de leur variance. En effet, on sait par hypothèse que : $\varepsilon_i \rightsquigarrow N(0, \sigma_\varepsilon^2)$. Avec le théorème centrale limite sur ε_i, on a :

$$\frac{\varepsilon_i}{\sigma_\varepsilon} \rightsquigarrow N(0,1)$$

Et comme $\hat{\varepsilon}_i$ est une réalisation de ε_i, il en vient que :

$$\frac{\hat{\varepsilon}_i}{\sigma_{\hat{\varepsilon}}} \rightsquigarrow N(0,1)$$

En passant l'opérateur de somme au carré de cette expression, on obtient une distribution à $\chi^2(n-k-1)$ où k est le nombre de variables explicatives et $k+1$ le nombre de paramètres à estimer. Ainsi, on a :

$$\sum_{i=1}^n \left(\frac{\hat{\varepsilon}_i}{\sigma_{\hat{\varepsilon}}}\right)^2 \rightsquigarrow \chi^2(n-k-1)$$

$$\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}}{\sigma_{\hat{\varepsilon}}^2} \rightsquigarrow \chi^2(n-k-1$$

$$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sigma_{\hat{\varepsilon}}^2} \rightsquigarrow \chi^2(n-k-1)$$

Ou, de manière équivalente, en se référant à l'estimateur de la variance de l'erreur, on a :

$$\frac{\hat{\sigma}_{\hat{\varepsilon}}^2}{\sigma_{\hat{\varepsilon}}^2} \rightsquigarrow \frac{\chi^2(n-k-1)}{(n-k-1)} \tag{2.28}$$

$$\text{Avec } \hat{\sigma}_{\hat{\varepsilon}}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-(k+1)}$$

2.3.1.2. Distribution des coefficients estimés

Lorsque la variance des erreurs est connue, les coefficients estimés suivent une loi normale.

Pour le coefficient $\hat{\beta}_0$, on a :

$$\hat{\beta}_0 \rightsquigarrow N(\beta_0, \sigma_{\hat{\beta}_0}^2)$$

$$\text{où } \sigma_{\hat{\beta}_0}^2 = \sigma_{\varepsilon}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

Avec le théorème central limite, on peut réécrire :

$$\left(\frac{\hat{\beta}_0-\beta_0}{\sigma_{\hat{\beta}_0}}\right) \rightsquigarrow N(0,1)$$

Aussi, pour le coefficient $\hat{\beta}_1$ on peut écrire :

$$\hat{\beta}_1 \rightsquigarrow N(\beta_1, \sigma_{\hat{\beta}_1}^2)$$

$$\text{Où } \sigma_{\hat{\beta}_1}^2 = \frac{\sigma_{\varepsilon}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Avec le théorème central limite, on peut réécrire :

$$\left(\frac{\hat{\beta}_1-\beta_1}{\sigma_{\hat{\beta}_1}}\right) \rightsquigarrow N(0,1)$$

D'une manière générale quelle que soit la variable j lorsque σ_{ε}^2 est connue, on peut écrire :

$$\left(\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}}\right) \rightsquigarrow N(0,1) \tag{2.29}$$

2.3.1.3. Distribution des variances des coefficients estimés

Soit le modèle estimé $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. En considérant d'abord la constante, on sait que :

$$\sigma^2_{\hat{\beta}_0} = \sigma^2_{\hat{\varepsilon}} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

Ainsi, on peut en déduire que :

$$\hat{\sigma}^2_{\hat{\beta}_0} = \hat{\sigma}^2_{\hat{\varepsilon}} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

En faisant le rapport entre ces deux expressions, on trouve :

$$\frac{\hat{\sigma}^2_{\hat{\beta}_0}}{\sigma^2_{\hat{\beta}_0}} = \frac{\hat{\sigma}^2_{\hat{\varepsilon}} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}{\sigma^2_{\hat{\varepsilon}} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} = \frac{\hat{\sigma}^2_{\hat{\varepsilon}}}{\sigma^2_{\hat{\varepsilon}}}$$

Ainsi, on peut poser l'égalité suivante :

$$\frac{\hat{\sigma}^2_{\hat{\beta}_0}}{\sigma^2_{\hat{\beta}_0}} = \frac{\hat{\sigma}^2_{\hat{\varepsilon}}}{\sigma^2_{\hat{\varepsilon}}}$$

Cette équation signifie que le rapport entre la variance estimée et la vraie variance de coefficient est égal au rapport entre la variance estimée et la vraie variance de l'erreur. Et puisque $\frac{\hat{\sigma}^2_{\hat{\varepsilon}}}{\sigma^2_{\hat{\varepsilon}}} \rightsquigarrow \frac{\chi^2(n-k-1)}{(n-k-1)}$, alors, on peut en déduire que :

$$\frac{\hat{\sigma}^2_{\hat{\beta}_0}}{\sigma^2_{\hat{\beta}_0}} = \frac{\hat{\sigma}^2_{\hat{\varepsilon}}}{\sigma^2_{\hat{\varepsilon}}} \rightsquigarrow \frac{\chi^2(n-k-1)}{(n-k-1)}$$

Avec $\sigma^2_{\hat{\beta}_0} = \sigma^2_{\hat{\varepsilon}} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$ et $\hat{\sigma}^2_{\hat{\beta}_0} = \hat{\sigma}^2_{\hat{\varepsilon}} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$

Par ailleurs, on sait que $\left(\frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}}\right) \rightsquigarrow N(0,1)$. Et comme la loi de Student est définie comme le rapport entre une loi normale centrée réduite et la racine carrée d'une loi du χ^2 normalisée par ses degrés de liberté, on peut écrire que :

$$\frac{\left(\frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}}\right)}{\left(\frac{\hat{\sigma}_{\hat{\beta}_0}}{\sigma_{\hat{\beta}_0}}\right)} = \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-k-1)}{(n-k-1)}}}$$

Cette expression permet donc de montrer que :

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \rightsquigarrow T(n-k-1)$$

Et de manière analogue, on peut montrer que :

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \rightsquigarrow T(n-k-1)$$

Ainsi, d'une manière générale, lorsque la variance des erreurs $\sigma_{\hat{\varepsilon}}^2$ n'est pas connue et qu'il est estimé par $\hat{\sigma}_{\hat{\varepsilon}}^2$, pour chaque variable j, on peut écrire :

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \rightsquigarrow T(n-k-1) \qquad (2.30)$$

En utilisant cette expression sert à la fois dans les tests sur les coefficients mais aussi dans de nombreux calculs tels que celui des intervalles de confiance de ces paramètres.

2.3.2. Test de significativité des coefficients estimés

2.3.3.1. Test de significativité de $\boldsymbol{\beta_0}$

Test bilatéral

Ce test se formule avec l'hypothèse suivante :

$$\begin{cases} H_0 \;\; \beta_0 = 0 \\ H_1 \;\; \beta_0 \neq 0 \end{cases}$$

Connaissant la distribution de probabilité de $\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}}$, on peut calculer la statistique de Student sous H0 comme suit :

$$t_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - 0}{\hat{\sigma}_{\hat{\beta}_0}} = \frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}} \qquad (2.31)$$

On peut simplement constater que la statistique de Student est égale au rapport entre la valeur estimée du coefficient et la valeur estimée de l'écart-type (racine carrée de la variance estimée). Cela correspond au rapport entre une variable

suivant une loi normale et une variable suivant une loi de khi-deux. Cette statistique suit dont une loi de Student à $(n-k-1)$ degré de libertés (égale à $n-2$ dans le cas du modèle linéaire simple où $k=1$).

Puisqu'il s'agit ici d'un test bilatéral, la région critique sera déterminée à gauche et à droite de 0. Par ailleurs, la loi de Student est une loi symétrique, c'est-à-dire que la probabilité d'une fractile positive est égale à la probabilité de la fractile opposée (négative). Par conséquent on considère la valeur absolue de la statistique de Student. Ainsi, le seuil critique du test sera divisé en deux valeurs (les deux parties symétriques de la distribution). Dès la statistique de Student sera défini par rapport à $t^*_{1-\frac{\alpha}{2}}$ qui représente la statistique lue dans la table de Student en considérant le nombre de degré de liberté $(n-k-1)$.

Ainsi, si $\left|t_{\hat{\beta}_0}\right| > t^*_{1-\frac{\alpha}{2}}$, ce qui implique la p.value p_0 du test sera inférieure à α. Dans ce cas, on rejette l'hypothèse de nullité du coefficient. En revanche, si $\left|t_{\hat{\beta}_0}\right| < t^*_{1-\frac{\alpha}{2}}$, cela implique que la p.value p_0 est supérieure à α. Dans ce cas, on ne peut pas rejeter l'hypothèse de nullité du coefficient.

De manière analogue, on peut mettre en œuvre le test significativité sur le coefficient β_1.

Test unilatéral

Le test bilatéral se contente de tester si le paramètre estimé est égal ou pas à une valeur donnée. Mais au cas où l'hypothèse nulle est rejetée, il ne permet pas de dire si le paramètre estimé est inférieure ou supérieure à la valeur spécifiée. Le test unilatéral vise à examiner une telle éventualité. Par exemple, on veut tester si un paramètre β_k est égal à une valeur fixée $\beta_{0,k}$, contre l'hypothèse alternative qu'il est strictement supérieur à cette valeur. L'hypothèse nulle s'écrit donc H0 : $\beta_k = \beta_{0,k}$, contre l'hypothèse alternative Ha : $\beta_k > \beta_{0,k}$. Dans cette configuration, on calcule la statistique de test de Student comme suit :

$$t_{\hat{\beta}_k} = \frac{\hat{\beta}_k - \beta_{0,k}}{\hat{\sigma}_{\hat{\beta}_k}} \tag{2.32}$$

Et on compare cette statistique à la valeur lue dans la table de Student au seuil de $1-\alpha$ et à $(n-k-1)$degrés de libertés $t^*_{1-\alpha}$.

Si $t_{\hat{\beta}_k} > t^*_{1-\alpha}$, cela signifie que la p.value p_0 est inférieure à α. Dans ce cas, on rejette l'hypothèse nulle. En revanche, lorsque $t_{\hat{\beta}_k} < t^*_{1-\alpha}$, cela signifie que la p.value p_0 est supérieure à α. On ne peut donc pas rejeter l'hypothèse nulle.

En revanche si on veut tester que le paramètre β_k est égal à une valeur fixée $\beta_{0,k}$, contre l'hypothèse alternative qu'il est strictement inférieur à cette valeur.

L'hypothèse alternative devient Ha : $\beta_k < \beta_{0,k}$. Dans cette configuration, on calcule toujours la statistique de test de Student comme suit :

$$t_{\widehat{\beta}_k} = \frac{\hat{\beta}_k - \beta_{0,k}}{\hat{\sigma}_{\widehat{\beta}_k}}$$

Et on compare cette statistique à l'opposé de la valeur lue dans la table de Student au seuil de $1-\alpha$ et à $(n-k-1)$degrés de libertés $-t^*_{1-\alpha}$
Dès lors, si $t_{\widehat{\beta}_k} < -t^*_{1-\alpha}$, ce qui signifie aussi que la p.value p_0 est inférieure à α. Dans ce cas, on rejette l'hypothèse nulle. En revanche, lorsque $t_{\widehat{\beta}_k} > -t^*_{1-\alpha}$, cela signifie que la p.value p_0 est supérieure à α. Dans ce cas, on ne peut pas rejeter l'hypothèse nulle.

Remarque sur la lecture de la p.value dans une table de Student

La p.value p_0 se détermine à partir de la table de Student en lisant la probabilité correspondant à la fractile $\left|t_{\widehat{\beta}_0}\right|$ compte tenu du nombre de degré de liberté. Pour déterminer la p.value, on fait le chemin inverse qui a permis de déterminer $t^*_{1-\frac{\alpha}{2}}$. En effet, pour déterminer $t^*_{1-\frac{\alpha}{2}}$, on s'est servit de α et du nombre de degré de liberté $(n-k-1)$ pour lire $t^*_{1-\frac{\alpha}{2}}$. Mais pour déterminer la p.value p_0, on se sert de $\left|t_{\widehat{\beta}_0}\right|$ et du nombre de degré de liberté $(n-k-1)$ pour lire la valeur de la probabilité correspondante. Se référer au chapitre 1 à la section 1.6 qui donne les indications de lecture d'une table statistique pour déterminer soit le fractile (valeur de la statistique) soit la probabilité (p.value).

2.3.3.2. Test de significativité globale du modèle

Pour tester la significativité globale du modèle, on se base sur la statistique F :

$$F_{p,q} = \frac{\dfrac{SCE}{(n-k)-(n-k-1)}}{\dfrac{SCR}{(n-k)}} \qquad (2.33a)$$

Où k est le nombre de variables explicatives (sans la constante) ; $(n-k)-(n-k-1)=1$. Avec p et q le nombre de degrés de libertés du dénominateur et numérateur. Ici $p = n-k$ et $q = 1$.

$$F_{p,q} = \frac{\dfrac{SCE}{1}}{\dfrac{SCR}{(n-k)}}$$

Cette statistique indique la mesure dans laquelle la variance expliquée est significativement supérieure à la variance résiduelle. Dans ce cas, on peut

considérer que l'explication emmenée par la régression traduit une relation qui existe réellement dans la population

Test à partir du coefficient de détermination

Le test de significativité peut également se faire à partir du coefficient de détermination. La statistique du Fisher se présente alors comme suit :

$$F = \frac{\frac{R^2}{1}}{\frac{(1-R^2)}{(n-k)}} \qquad (2.33b)$$

Calcul en utilisant la distribution sous H0.

Il existe une troisième approche de test de significativité globale qui consiste à utiliser la distribution des sommes des carrés des résidus sous H0 (Tous les coefficients sont nuls sauf la constante). Sous H0, la SCE est distribué selon un $\chi^2(1)$ et SCR selon un $\chi^2(n-k)$ de fait pour F nous avons :

$$F = \frac{\frac{\chi^2((n-k)-(n-k-1))}{1}}{\frac{\chi^2(n-k)}{(n-k)}} \qquad (2.33c)$$

Règles de décision

La région critique du test, correspondant au rejet de H0, au risque α est définie pour les valeurs anormalement élevées de F c.-à-d :

$$F > F_{1-\alpha}(1, n-k)$$

Décision à partir de la p-value p_0. La p-value p_0 correspond à la probabilité que la loi de Fisher dépasse la statistique calculée F. Ainsi, la règle de décision au risque α devient : $p_0 < \alpha$

Si $F > F_{1-\alpha}(1, n-k)$, ce qui implique que $p_0 < \alpha$. Alors, on rejette H0.

2.3.3. Intervalle de confiance des paramètres estimés

Pour ne pas se limiter à une estimation ponctuelle $\hat{\theta}$ d'un paramètre θ, on préfère généralement définir un intervalle [L ,U] dans lequel pourrait se trouver le paramètre avec une certaine probabilité notée $1-\alpha$ où α est le seuil d'erreur.

Soit $\hat{\beta}$ l'estimateur d'un paramètre β. On souhaite déterminer un intervalle de confiance du paramètre β. Pour cela, on suit la démarche suivante :

On sait que : $\frac{\hat{\beta}-\beta}{\hat{\sigma}_{\hat{\beta}}} \rightsquigarrow T(n-k-1)$. Ainsi connaissant la valeur de α on peut poser que :

$$P\left(\left|\frac{\hat{\beta}-\beta}{\hat{\sigma}_{\hat{\beta}}}\right| > T^*_{1-\frac{\alpha}{2}}(n-k-1)\right) = \alpha$$

Où $T^*_{1-\frac{\alpha}{2}}(n-k-1)$ représente la statistique lue dans la table de Student en considérant le seuil d'erreur α et le nombre de degré de liberté $(n-k-1)$.

Dès lors l'intervalle de confiance se déduit de cette probabilité telle que :

$$P\left(\left|\frac{\hat{\beta}-\beta}{\hat{\sigma}_{\hat{\beta}}}\right| > T^*_{1-\frac{\alpha}{2}}(n-k-1)\right) = 1-\alpha$$

Soit :

$$P\left(-T^*_{1-\frac{\alpha}{2}}(n-k-1) < \frac{\hat{\beta}-\beta}{\hat{\sigma}_{\hat{\beta}}} < T^*_{1-\frac{\alpha}{2}}(n-k-1)\right) = 1-\alpha$$

D'où :

$$P\left(\hat{\beta} - T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\beta}} < \beta < \hat{\beta} + T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\beta}}\right) = 1-\alpha$$

Ainsi, $1-\alpha$ chance que β soit compris entre $\hat{\beta} - T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\beta}}$ et $\hat{\beta} + T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\beta}}$. Ainsi l'intervalle de confiance de β se définit comme suit :

$$IC_{\beta} = \left[\hat{\beta} - T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\beta}} \ ; \ \hat{\beta} + T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\beta}}\right] \qquad (2.34)$$

Dès lors, pour tout modèle estimé de type $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, on peut fournir les intervalles de confiance suivants :

$$IC_{\beta_0} = \left[\hat{\beta}_0 - T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\beta}_0} \ ; \ \hat{\beta}_0 + T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\beta}_0}\right]$$

$$IC_{\beta_1} = \left[\hat{\beta}_1 - T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\beta}_1} \ ; \ \hat{\beta}_1 + T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\beta}_1}\right]$$

Où $T^*_{1-\frac{\alpha}{2}}(n-k-1)$ représente la statistique lue dans la table de Student en considérant le seuil d'erreur α et le nombre de degré de liberté $(n-k-1)$.

2.3.4. Prédiction à l'intérieur de l'échantillon et intervalle de confiance de la droite de régression

Déterminer l'intervalle de confiance de la droite de régression c'est fournir un intervalle de confiance pour la valeur prédite $\widehat{y_i}$ à l'intérieur de l'échantillon qui a servi à estimer les paramètres. Ce calcul se base essentiellement sur la valeur estimée de la constante (qui représente la moyenne de y lorsque $x = 0$).

Soit $y_i = \beta_0 + \beta_1 x_i$. L'estimation de ce modèle permet d'obtenir les valeur prédites telles que $\widehat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. La formule générale de l'intervalle de confiance de cette droite se calcule comme suit :

$$IC_{\widehat{y_i}} = \left(\hat{\beta}_1 x_i + \hat{\beta}_0\right) \pm T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\varepsilon}} \sqrt{\left[\frac{1}{n} + \frac{(x_i - \bar{x})}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}\right]}$$

La quantité $h_i = \left[\frac{1}{n} + \frac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right]$ est appelée le levier (leverage). Elle permet d'identifier les observations aberrantes. Ainsi avec h_i, on peut réécrire :

$$IC_{\widehat{y_i}} = \left(\hat{\beta}_1 x_i + \hat{\beta}_0\right) \pm T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\varepsilon}}\sqrt{h_i} \qquad (2.35)$$

NB : L'intervalle de confiance se calcule pour chaque individu i. Ainsi, en représentant la série des bornes inférieures et la série des bornes supérieures, on détermine l'intervalle de confiance de la droite estimée.

2.3.5. Prédiction hors-échantillon et erreur de prédiction

Le but ultime de l'estimation d'un modèle est de pouvoir réaliser des prédictions de la variable y pour un individu i qui ne figure pas dans l'échantillon d'origine (échantillon ayant servi à estimer les paramètres). Il s'agit alors d'une prédiction hors-échantillon (out-of-sample). On distingue généralement deux types de prédiction : une prédiction ponctuelle et ou une prédiction par intervalle de confiance.

2.3.5.1. Prédiction ponctuelle de y hors-échantillon

Soit x_i^* la valeur de x observée pour un individu i n'appartenant pas à l'échantillon initial. En partant du modèle $y_i^* = \beta_0 + \beta_1 x_i^* + \varepsilon_i^*$, où y_i^* n'est pas observée, la prédiction ponctuelle hors-échantillon se fait en remplaçant x_i^* dans l'équation estimée telle que :

$$\hat{y}_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i^*$$

On peut facilement montrer que cette prédiction est sans biais. Pour cela on montre soit que $E(\hat{y}_i^*) = y_i^*$. En effet

$$E(\hat{y}_i^*) = E(\hat{\beta}_0) + E\big(\hat{\beta}_1\big)E(x_i^*)$$

$$= \beta_0 + \beta_1 E(x_i^*)$$

Avec $E(x_i^*) = x_i^*$, d'où :

$$E(\hat{y}_i^*) = \beta_0 + \beta_1 x_i^* = y_i^*$$

L'estimation ponctuelle $\hat{y}_i^*$ de y_i^* est donc sans biais.

Ainsi, en utilisant l'équation $\hat{y}_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i^*$, on peut calculer l'erreur de prédiction du modèle $\hat{\varepsilon}_i^*$ qui est la valeur estimée de ε_i^* (non observable). Ainsi, on a :

$$\hat{\varepsilon}_i^* = \hat{y}_i^* - y_i^*$$

Là aussi, on peut montrer que l'erreur de prévision est en moyenne nulle c'est-à-dire que $E(\hat{\varepsilon}_i^*) = 0$. En effet, on sait que :

$$\hat{\varepsilon}_i^* = \hat{y}_i^* - y_i^*$$

$$= \big(\hat{\beta}_0 + \hat{\beta}_1 x_i^*\big) - (\beta_0 + \beta_1 x_i^* + \varepsilon_i^*)$$

$$\hat{\varepsilon}_i^* = \big(\hat{\beta}_0 - \beta_0\big) + \big(\beta_1 - \hat{\beta}_1\big)x_i^* - \varepsilon_i^*$$

On passant cette égalité à l'opérateur d'espérance, on a :

$$E(\hat{\varepsilon}_i^*) = E\big(\hat{\beta}_0 - \beta_0\big) + E\big(\beta_1 - \hat{\beta}_1\big)E(x_i^*) - E(\varepsilon_i^*)$$

$$E(\hat{\varepsilon}_i^*) = 0 + 0 \times E(x_i^*) + 0$$

$$E(\hat{\varepsilon}_i^*) = 0$$

En conclusion, pour obtenir une prévision ponctuelle et une estimation ponctuelle de l'erreur de prédiction, on utilise respectivement les deux formules ci-dessous :

$$\hat{y}_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i^* \tag{2.36}$$

$$\hat{\varepsilon}_i^* = \hat{y}_i^* - y_i^* \tag{2.37}$$

2.3.5.2. Prédiction par intervalle de confiance de *y* hors-échantillon

Pour construire l'intervalle de confiance de la prédiction de y hors échantillon, nous devons d'abord calculer l'intervalle de confiance de l'erreur de prédiction. Pour cela, nous nous servons de la variance de l'erreur de prédiction et la loi de distribution de l'erreur de prédiction.

Variance de l'erreur de prédiction

L'erreur de prévision théorique notée ε_i^* a une espérance nulle, $E(\varepsilon_i^*) = 0$ et une variance $V(\varepsilon_i^*)$ égale $\sigma_{\varepsilon_i^*}^2 = E[(\varepsilon_i^*)^2]$.

Mais puisque l'erreur de prévision estimée $\hat{\varepsilon}_i^*$ est une réalisation de ε_i^*, il en vient que $V(\hat{\varepsilon}_i^*) = \sigma_{\hat{\varepsilon}_i^*}^2 = E[(\hat{\varepsilon}_i^*)^2]$. Le développement de cette expression permet de montrer que :

$$\sigma_{\hat{\varepsilon}_i^*}^2 = \sigma_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})}{\sum_{j=1}^n \left(x_j - \bar{x}\right)^2}\right] \tag{2.38}$$

Ce qui peut se réécrire telle que :

$$\sigma_{\hat{\varepsilon}_i^*}^2 = \sigma_\varepsilon^2 [1 + h_i]$$

Avec $h_i = \left[\frac{1}{n} + \frac{(x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2}\right]$ appelée levier.

De là, on peut obtenir la variance estimée de l'erreur de prévision estimée comme suit :

$$\hat{\sigma}_{\hat{\varepsilon}_i^*}^2 = \hat{\sigma}_\varepsilon^2 [1 + h_i]$$

Distribution de l'erreur de prédiction

On sait par hypothèse que $\frac{\varepsilon_i^*}{\sigma_{\varepsilon_i^*}} \rightsquigarrow N(0,1)$ (application du théorême central limite).

Et comme $\hat{\varepsilon}_i^*$ est une réalisation de ε_i^*, il en vient également que :

$$\frac{\hat{\varepsilon}_i^*}{\sigma_{\hat{\varepsilon}_i^*}} \rightsquigarrow N(0,1)$$

Ce qui permet d'écrire que :

$$\frac{\hat{\varepsilon}_i^*}{\sigma_{\hat{\varepsilon}_i^*}} = \frac{\hat{y}_i^* - y_i^*}{\sigma_{\hat{\varepsilon}_i^*}} \rightsquigarrow N(0,1)$$

Tout comme pour les autres paramètres estimés du modèle, on sait aussi que le rapport entre la variance estimée et la vraie variance suit une loi de khi-deux à $n - k - 1$. Dès lors, on peut écrire que :

$$\frac{\hat{\sigma}_{\hat{\varepsilon}_i^*}^2}{\sigma_{\hat{\varepsilon}_i^*}^2} \rightsquigarrow \frac{\chi^2(n - k - 1)}{(n - k - 1)}$$

On obtient alors deux distributions que sont : $\frac{\hat{\varepsilon}_i^*}{\sigma_{\hat{\varepsilon}_i^*}} \rightsquigarrow N(0,1)$ et $\frac{\hat{\sigma}_{\hat{\varepsilon}_i^*}^2}{\sigma_{\hat{\varepsilon}_i^*}^2} \rightsquigarrow \frac{\chi^2(n-k-1)}{(n-k-1)}$.

Ainsi, en faisant le rapport entre une loi normale et la racine carrée d'une loi de de khi-deux, on obtient une loi de student. On a :

$$\frac{\frac{\hat{\varepsilon}_i^*}{\sigma_{\hat{\varepsilon}_i^*}}}{\sqrt{\frac{\hat{\sigma}_{\hat{\varepsilon}_i^*}^2}{\sigma_{\hat{\varepsilon}_i^*}^2}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-k-1)}{(n-k-1)}}} \rightsquigarrow T(n-k-1)$$

$$\frac{\frac{\hat{\varepsilon}_i^*}{\sigma_{\hat{\varepsilon}_i^*}}}{\frac{\hat{\sigma}_{\hat{\varepsilon}_i^*}}{\sigma_{\hat{\varepsilon}_i^*}}} \rightsquigarrow T(n-k-1)$$

Ce qui finalement donne :

$$\frac{\hat{\varepsilon}_i^*}{\hat{\sigma}_{\hat{\varepsilon}_i^*}} \rightsquigarrow T(n-k-1) \qquad (2.39)$$

On en conclut alors que l'erreur de prévision suit une loi de student à $n-k-1$ degrés de liberté. Dès lors, pour toute prédiction de type $\hat{y}_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i^*$, on peut fournir un intervalle de confiance pour l'erreur de prévision $\hat{\varepsilon}_i^*$ comme suit :

$$IC_{\hat{\varepsilon}_i^*} = \left[-T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\varepsilon}_i^*} \; ; \; T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\varepsilon}_i^*}\right] \qquad (2.40)$$

Où $T^*_{1-\frac{\alpha}{2}}(n-k-1)$ représente la statistique lue dans la table de Student en considérant le seuil d'erreur α et le nombre de degrés de liberté $(n-k-1)$. Et où

$$\hat{\sigma}_{\hat{\varepsilon}_i^*}^2 = \hat{\sigma}_{\varepsilon}^2[1+h_i]$$

$$\text{Avec } h_i = \left[\frac{1}{n} + \frac{(x_i-\bar{x})}{\sum_{j=1}^n (x_j-\bar{x})^2}\right] \text{ et } \hat{\sigma}_{\varepsilon}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i{}^2}{n-k-1}$$

Calcul de l'intervalle de confiance de la prédiction de y

Avec l'expression de l'intervalle de confiance de l'erreur de prédiction, on calcule l'intervalle de confiance de la prédiction de y comme suit :

$$IC_{\hat{y}_i^*} = \hat{y}_i^* \pm T^*_{1-\frac{\alpha}{2}}(n-k-1)\hat{\sigma}_{\hat{\varepsilon}_i^*} \qquad (2.41)$$

$$\text{Avec } \hat{y}_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i^*$$

2.3.6. Linéarisation des modèles non-linéaires

Dans la plupart des cas, les modèles à estimer ne se présentent pas initialement sous la forme linéaire. Il faut donc procéder à une linéarisation avant de mettre en œuvre les techniques d'estimation. Les principaux types de modèles sont :

- *Le modèle log-linéaire*

$$y_i = a_0 x_i^{\beta_1}$$

Pour linéariser ce modèle, on procède à une transformation logarithmique. Ainsi, on a :

$$ln(y_i) = \beta_0 + \beta_1 ln(x_i)$$

$$Avec\ \beta_0 = ln(a_0)$$

- *Le modèle exponentiel*

$$y_i = e^{(\beta_0 + \beta_1 x_i)}$$

Là aussi, on procède à une transformation logarithmique pour linéariser ce modèle :

$$ln(y_i) = \beta_0 + \beta_1 x_i$$

- *Le modèle logarithmique*

$$y_i = \beta_0 + \beta_1 ln(x_i)$$

Ce modèle peut être directement estimé car il se présente déjà sous forme linéaire.

- *Le modèle hyperbolique*

$$y_i = \frac{\beta}{x_i - x_{0i}} + y_{0i}$$

Le modèle hyperbolique peut aussi être estimé car il est déjà linéaire[1] en β. L'équation se présente alors simplement comme suit :

$$\widetilde{y_i} = \beta \widetilde{x_i}$$

Avec $\widetilde{y_i} = y_i - y_{0i}$ et $\widetilde{x_i} = \frac{1}{x_i - x_{0i}}$

[1] Il faut noter qu'on peut distinguer deux types de linéarité : la linéarité en fonction des paramètres et la linéarité en fonction des variables. La linéarité dont il est question ici est déterminée en fonction des paramètres et non en fonction des variables. C'est seulement la non-linéarité des paramètre qui peut nécessiter une linéarisation du modèle.

- *Le modèle logistique*

$$y_i = y_{min,i} + \frac{y_{max,i} - y_{min,i}}{1 + e^{(\beta_0 + \beta_1 x_i)}}$$

Le modèle logistique est linéarisé en prenant le logarithme tel que :

$$ln\left(\frac{y_{max,i} - y_{min,i}}{y_i - y_{min,i}}\right) = \beta_0 + \beta_1 x_i$$

La forme linéarisée du modèle logistique est :

$$ln(\tilde{y_i}) = \beta_0 + \beta_1 x_i$$

Avec $\tilde{y_i} = \frac{y_{max,i} - y_{min,i}}{y_i - y_{min,i}}$

- *Le modèle parabolique*

Le modèle parabolique se présente sous la forme suivante :

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

Ce modèle est déjà linéaire par rapport aux paramètres. Par conséquent, il n'y a pas besoin de procéder à quelle que transformation que ce soit. Il faut simplement noter, en revanche, que la relation entre x et y n'est pas linéaire. Ici, il s'agit d'une relation parabolique encore d'une relation quadratique.

2.4. Estimateur du maximum de vraisemblance

L'estimation par les moindres carrés ordinaires ne fait aucune hypothèse sur la loi de distribution des perturbations aléatoires. Il est simplement supposé que celles-ci sont indépendantes et identiquement distribuées. L'estimation par maximum de vraisemblance consiste à faire une hypothèse sur la distribution de probabilité de ε_i . En effet, on suppose, que les ε_i suivent une loi normale de moyennes nulle et de variance σ_ε^2 . En reprenant, l'équation (2.1) du modèle linéaire initial, on a :

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

L'hypothèse de normalité des résidus se présente comme suit :

$$\varepsilon_i \rightsquigarrow N(0, \sigma_\varepsilon^2)$$

Il faut aussi noter que comme $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ et que $\varepsilon_i \rightsquigarrow N(0, \sigma_\varepsilon^2)$ alors on aura $y_i \rightsquigarrow N(\beta_0 + \beta_1 \bar{x}, \sigma_\varepsilon^2)$. Car :

$$E(y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = E(\beta_0 + \beta_1 x_i) + E(\varepsilon_i) = \beta_0 + \beta_1 \bar{x}$$

$$VAR(y_i) = VAR(\beta_0 + \beta_1 x_i + \varepsilon_i) = VAR(\beta_0 + \beta_1 x_i) + VAR(\varepsilon_i) = \sigma_\varepsilon^2$$

Fonction de densité et fonction de vraisemblance

Une variable aléatoire Z est dite normale de moyenne μ et de variance σ^2 si sa densité vaut :

$$f(z_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{z_i - \mu}{\sigma}\right)^2\right) \tag{2.42}$$

Ainsi, en supposant que la variable ε_i suit une loi normale $N(0, \sigma_\varepsilon^2)$ et sachant que : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ on peut spécifier sa fonction de densité telle que : $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$. On aura alors :

$$f(y_i) = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma_\varepsilon}\right)^2\right)$$

Ainsi, pour estimer les paramètres β_0 ,β_1 et σ_ε par maximum de vraisemblance, on pose d'abord l'expression de la fonction de vraisemblance. Pour n observations de la variable y_i, la fonction de vraisemblance s'obtient en faisant le produit des n fonction de densités, puisque les n observations sont supposées indépendantes et identiquement distribuées (hypothèse i.d.d) :

Ainsi, en notant par $L(\beta_0, \beta_1, \sigma_\varepsilon^2)$, la fonction de vraisemblance, on a :

$$L(\beta_0, \beta_1, \sigma_\varepsilon^2) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \left[\frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(y_i - \beta_0 - \beta_1 x_i)^2\right)\right]$$

$$L(\beta_0, \beta_1, \sigma_\varepsilon^2) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right)$$

Il est souvent plus facile de chercher à maximiser le logarithme de la fonction de vraisemblance plutôt que la fonction elle-même, car le logarithme est une transformation monotone croissante. Cette transformation n'a aucune incidence sur les paramètres. Ainsi, en prenant le logarithme de la fonction de vraisemblance, on a :

$$LogL(\beta_0, \beta_1, \sigma_\varepsilon^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_j\right)^2$$

La méthode du maximum de vraisemblance consiste à choisir les paramètres $\beta_0, \beta_1, \sigma_\varepsilon^2$ et σ_ε^2 de sorte à maximiser cette fonction de log-vraisemblance. Pour

cela, il faut dériver la dérivée et retrouver les conditions de premier ordre afin d'en déduire chacun des paramètres.

$$\begin{cases} \dfrac{\partial LogL(\beta_0,\beta_1,\sigma_\varepsilon^2)}{\beta_0} = \dfrac{2}{2\sigma_\varepsilon^2}\displaystyle\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i) = 0 \\ \\ \dfrac{\partial l(\beta_0,\beta_1)}{\beta_1} = \dfrac{2}{2\sigma_\varepsilon^2}\displaystyle\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)x_i = 0 \end{cases}$$

On obtient un système de deux équations à deux inconnues, qui peuvent également s'écrire :

$$\begin{cases} \bar{y} - \beta_0 - \beta_1\bar{x} = 0 \\ \\ \displaystyle\sum_{i=1}^{n} x_i y_i - \beta_0\sum_{i=1}^{n}x_i - \beta_1\sum_{i=1}^{n}{x_i}^2 = 0 \end{cases}$$

Ce qui permet donc de poser que $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \beta_1\bar{x} \qquad (2.43a)$$

En remplaçant $\hat{\beta}_0$ par sa valeur dans la seconde équation divisée par n, on a :

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\bar{y}-\beta_1\bar{x})\bar{x} - \beta_1\frac{1}{n}\sum_{i=1}^{n}{x_i}^2 = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \beta_1\left(\frac{1}{n}\sum_{i=1}^{n}{x_i}^2 - \bar{x}^2\right) = 0$$

$$COV(x,y) - \beta_1 VAR(x) \qquad (2.43b)$$

Ainsi la solution au problème de minimisation de la somme des carrés de résidus se présente comme suit :

$$\begin{cases} \hat{\beta}_1 = \dfrac{COV(x,y)}{VAR(x)} & (2.44a) \\ \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} & (2.44b) \end{cases}$$

Ce qui montre bien que les estimateurs de maximum de vraisemblance sont équivalents aux estimateurs des MCO lorsque les résidus sont normaux.

CHAPITRE 3. LE MODELE LINEAIRE MULTIPLE

L'équation générale du modèle linéaire se présente comme suit :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad (3.1)$$

où y représente la variable expliquée (encore appelée variable dépendante). x_1, $x_2 \ldots$ x_k représentent les variables explicatives (encore appelée variable indépendantes). ε représente une perturbation aléatoire ou les résidus. Les coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ sont les paramètres à estimer. Ils représentent les effets des variables explicatives sur les variables expliquées. Par exemple, β_1 mesure l'impact de la variable x_1 sur la variable y.

Il faut simplement remarquer que dans l'équation (3.1), les variables y et x sont observées alors que les paramètres $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ et les perturbations aléatoires sont inobservés.

3.1. Estimation par Moindre Carrés ordinaires

Comme pour le modèle linéaire simple, l'estimation du modèle linéaire multiple par les Moindres Carrés Ordinaires (MCO) consiste à choisir les paramètres de façons à minimiser la somme des carrés des résidus :

$$l(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = \quad \text{Min} \sum_{i=1}^{n} \varepsilon^2 = \quad \underset{\beta_0, \beta_1}{\text{Min}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \quad (3.2)$$

Le minimum de cette fonction s'obtient en annulant les dérivées partielles par rapport à chacun des paramètres $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Pour cela, il faut dériver la fonction $l(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$ par rapport à chacun de ses arguments pour en dériver les conditions de premiers ordres qui forment les équations normales.

$$\begin{cases} \dfrac{\partial l(.)}{\beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}) = 0 & (i) \\ \dfrac{\partial l(.)}{\beta_1} = -2\sum_{i=1}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}) = 0 & (ii) \\ \dfrac{\partial l(.)}{\beta_2} = -2\sum_{i=1}^{n} x_{i2}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}) = 0 & (ii) \\ \cdots \\ \cdots \\ \dfrac{\partial l(.)}{\beta_k} = -2\sum_{i=1}^{n} x_{ik}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}) = 0 & (\ldots) \end{cases}$$

Après simplification par -2, le développement des facteurs et la distribution de l'opérateur somme, on obtient le suivant ci-dessous :

$$\begin{cases} \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1\sum_{i=1}^{n} x_{i1} - \beta_2\sum_{i=1}^{n} x_{i2} - \cdots - \beta_k\sum_{i=1}^{n} x_{ik} = 0 & (i) \\ \sum_{i=1}^{n} x_{i1}y_i - \beta_0\sum_{i=1}^{n} x_{i1} - \beta_1\sum_{i=1}^{n}(x_{i1})^2 - \beta_2\sum_{i=1}^{n} x_{i1}x_{i2} - \cdots - \beta_k\sum_{i=1}^{n} x_{i1}x_{ik} = 0 & (ii) \\ \sum_{i=1}^{n} x_{i2}y_i - \beta_0\sum_{i=1}^{n} x_{i2} - \beta_1\sum_{i=1}^{n} x_{i2}x_{i1} - \beta_2\sum_{i=1}^{n}(x_{i2})^2 - \cdots - \beta_k\sum_{i=1}^{n} x_{i2}x_{ik} = 0 & (iii) \\ \cdots \\ \cdots \\ \sum_{i=1}^{n} x_{ik}y_i - \beta_0\sum_{i=1}^{n} x_{ik} - \beta_1\sum_{i=1}^{n} x_{ik}x_{i1} - \beta_2\sum_{i=1}^{n} x_{ik}x_{i2} - \cdots - \beta_k\sum_{i=1}^{n}(x_{ik})^2 = 0 & (\ldots) \end{cases}$$

On obtient alors un système k+1 équations (dites équations normales) à k+1 inconnues.

3.1.1. Résolution du système par substitution

Pour résoudre ce système, on procède par la méthode de substitution.

D'abord, on tire β_0 de l'équation (i) en divisant d'abord celle-ci par n. Ce qui donne :

$$\beta_0 = \bar{y} - \beta_1\overline{x_1} - \beta_2\overline{x_2} - \cdots - \beta_k\overline{x_k}$$

Ensuite, on remplace β_0 par son expression dans toutes les autres équations. Ensuite, on tire β_1 de l'équation (ii) et l'on replace par sa valeur dans toutes les autres équations. Ce processus de substitution continue jusqu'à obtenir β_k dont la valeur ne dépend plus que des expressions connues. Dès lors, après avoir calculé β_k, on peut calculer β_{k-1} et ainsi de suite jusqu'à β_0 tel que :

$$\beta_0 = \bar{y} - \hat{\beta}_1\overline{x_1} - \hat{\beta}_2\overline{x_2} - \cdots - \hat{\beta}_k\overline{x_k} \tag{3.3}$$

Cette méthode de résolution s'avère très longue et couteuse en calcul, c'est pourquoi, on lui préfère la méthode matricielle de résolution. La mise en œuvre de la méthode matricielle nécessite d'abord une écriture matricielle des données.

3.1.2. Représentation matricielle des données

La représentation matricielle des données est une étape fondamentale dans l'estimation des paramètres dans le modèle linéaire multiple. En considérant, l'équation (3.1) mettant en relation une variable dépendante y un ensemble de K variables explicatives auquel on adjoint une constante, si l'on dispose de N observations sur chacune de ces variables, on peut proposer une formulation matricielle de l'équation (3.1).

D'abord l'empilement des n observations permet d'obtenir un vecteur-colonne c'est à dire une matrice à n lignes et une colonne. Ce vecteur sera noté Y avec :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{3.4a}$$

Notons que cette notation est équivalente à l'expression :

$$Y = (y_1, \quad y_2, \ldots, \ldots y_n)' \tag{3.4b}$$

Cette deuxième notation de Y se présente sous la forme du transposé d'un vecteur-ligne.

Les k variables de l'équation (1) peuvent également s'écrire sous la forme matricielle. Mais, dans un premier temps, on présente la matrice sans prendre en compte la constante β_0, la matrice se présente comme suit :

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \tag{3.5}$$

Sans prendre en compte le paramètre β_0, la matrice X a pour dimension N × K.

En prenant en compte le paramètre β_0, on insère un vecteur colonne rempli de 1 dans la matrice X. Ainsi, la matrice X devient :

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{2k} \end{pmatrix} \tag{3.6}$$

Avec la prise en compte du paramètre β_0, la matrice X a finalement pour dimension N × (K + 1).

La série des ε_i doit aussi être mise sous la forme matricielle. Celle-ci se présente sous la forme d'un vecteur-colonne ou sous la forme de transposé d'un vecteur-ligne. On a alors:

$$\epsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = (\varepsilon_1, \quad \varepsilon_2, \dots, \dots \varepsilon_n)' \tag{3.7}$$

En fin les coefficients peuvent être regroupés sous une forme matrice, principalement sous la forme d'un vecteur-colonne. Cette représentation des coefficients se fait comme suit :

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \tag{3.8}$$

En rassemblant ces différent éléments avec leur nouvelles formulations, l'équation (1) se présente comme suit :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{2k} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{3.9}$$

Ainsi, de façon synthétique, l'équation se présente comme suit :

$$Y = X\,\beta + \epsilon \tag{3.10}$$

L'estimation de l'équation (3.10) par les moindres carrés ordinaires consiste à minimiser la somme des carrés des résidus qui constitue le vecteur ϵ. L'équation de minimisation se présente comme suit :

$$L(\beta) = \min_{\beta_0,\beta_1,\beta_2 \dots \beta_k} \sum_{i=1}^{n} {\varepsilon_i}^2 = \min_{\beta} \epsilon'\epsilon = \min_{\beta} (Y - X\,\beta)'(Y - X\,\beta) \tag{3.11}$$

Il faut simplement noter que comme $\epsilon = Y - X\,\beta$, alors $\epsilon'\epsilon$ représente la forme matricielle de la somme des carrés des résidus $\sum_{i=1}^{n} {\varepsilon_i}^2$. En effet, une matrice multipliée par sa transposée donne toujours la somme des carrés de ses éléments.

Pour obtenir le minimum de $L(\beta)$, on annule le vecteur des dérivées matricielles :

$$\frac{\partial L(\beta)}{\beta} = -2X'(Y - X\,\beta) = 0 \tag{3.12}$$

Ce qui donne :

$$X'X\,\beta \;=\; X'Y \tag{3.13}$$

Cette expressions traduit la matrice des équations normales.

En multipliant, les deux membres de l'équation par l'inverse de $X'X$, c'est-à-dire $(X'X)^{-1}$, on trouve β telque :

$$\beta = (X'X)^{-1}X'Y \tag{3.14}$$

3.1.3. Correspondance entre la méthode de substitution et la méthode matricielle

Pour trouver la correspondance entre la méthode de substitution et la méthode matricielle, on part des équations normales qui forment le système d'équations

$$\begin{cases} \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_{i1} - \beta_2 \sum_{i=1}^{n} x_{i2} - \cdots - \beta_k \sum_{i=1}^{n} x_{ik} = 0 & (i) \\ \sum_{i=1}^{n} x_{i1}y_i - \beta_0 \sum_{i=1}^{n} x_{i1} - \beta_1 \sum_{i=1}^{n} (x_{i1})^2 - \beta_2 \sum_{i=1}^{n} x_{i1}x_{i2} - \cdots - \beta_k \sum_{i=1}^{n} x_{i1}x_{ik} = 0 & (ii) \\ \sum_{i=1}^{n} x_{i2}y_i - \beta_0 \sum_{i=1}^{n} x_{i2} - \beta_1 \sum_{i=1}^{n} x_{i2}x_{i1} - \beta_2 \sum_{i=1}^{n} (x_{i2})^2 - \cdots - \beta_k \sum_{i=1}^{n} x_{i2}x_{ik} = 0 & (iii) \\ \cdots \\ \sum_{i=1}^{n} x_{ik}y_i - \beta_0 \sum_{i=1}^{n} x_{ik} - \beta_1 \sum_{i=1}^{n} x_{ik}x_{i1} - \beta_2 \sum_{i=1}^{n} x_{ik}x_{i2} - \cdots - \beta_k \sum_{i=1}^{n} (x_{ik})^2 = 0 & (\ldots) \end{cases}$$

Regroupons d'un côté les membres qui contiennent les paramètres, on a :

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{i1} + \beta_2 \sum_{i=1}^{n} x_{i2} + \cdots + \beta_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i & (i) \\ \beta_0 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} (x_{i1})^2 + \beta_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \beta_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i & (ii) \\ \beta_0 \sum_{i=1}^{n} x_{i2} + \beta_1 \sum_{i=1}^{n} x_{i2}x_{i1} + \beta_2 \sum_{i=1}^{n} (x_{i2})^2 + \cdots + \beta_k \sum_{i=1}^{n} x_{i2}x_{ik} = \sum_{i=1}^{n} x_{i2}y_i & (iii) \\ \cdots \\ \cdots \\ \beta_0 \sum_{i=1}^{n} x_{ik} + \beta_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \beta_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \beta_k \sum_{i=1}^{n} (x_{ik})^2 = \sum_{i=1}^{n} x_{ik}y_i & (\ldots) \end{cases}$$

A partir de cette formulation, on construit une forme matricielle qui se présente comme suit :

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} (x_{i1})^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} (x_{i2})^2 & \cdots & \cdots & \sum_{i=1}^{n} x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \cdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \cdots & \sum_{i=1}^{n} (x_{ik})^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \cdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \sum_{i=1}^{n} x_{i2}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{pmatrix}$$

Avec cette forme, on peut en déduire la correspondance suivante :

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} (x_{i1})^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} (x_{i2})^2 & \cdots & \cdots & \sum_{i=1}^{n} x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \cdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \cdots & \sum_{i=1}^{n} (x_{ik})^2 \end{pmatrix} = X'X \quad ;$$

$$\begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \\ \sum_{i=1}^{n} x_{i2} y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik} y_i \end{pmatrix} = X'Y \quad ; \quad \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{pmatrix} = \beta$$

Avec cette décomposition, on constate que la matrice $X'X$ est constituée dans sa première colonne par la somme des variables, dans sa deuxième colonne par la somme du croisement de la variable x_{i1} avec les autres variables, dans la troisième colonne par le croisement de la variable x_{i2} avec les autres variables et dans la dernière colonne par le croisement de la variable x_{ik} avec les autres variables[2].

On peut également constater que la matrice que la matrice $X'Y$ est un vecteur-colonne dont les éléments sont la somme des croisements entre les colonnes de la matrice X et les éléments de la matrice Y.

Par ailleurs, le modèle linéaire simple étant un cas particulier du modèle multiple, on peut montrer que la forme matricielle des équations normales dans le modèle linéaire simple se présente comme suit :

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} (x_i)^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} (x_i)^2 \end{pmatrix} = X'X \; ; \; \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} = X'Y \; ; \; \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \beta$$

[2] Il faut simplement noter le fait que la matrice X étant constituée dans sa première colonne par des 1 alors la matrice X'X est, en réalité constituée dans sa première colonne par le croisement de 1 avec les autres variables. Ce qui équivaut à faire la somme de ces variables.
On peut aussi constater que la matrice X'X est une matrice symétrique dont la diagonale est constituée de la somme des carrées des variables, car les éléments diagonaux correspondent aux croisements des variables avec elles-mêmes.

3.1.4. Calcul des valeurs prédites

$$\hat{Y} = X\,\hat{\beta} \tag{3.15}$$

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

Le vecteur des valeurs ajustées peut être interprété comme la projection de Y sur le sous-espace engendré par les colonnes de la matrice X.

$$\hat{Y} = P_X Y \tag{3.16}$$

Où P_X est l'opérateur de projection.

$$P_X = X(X'X)^{-1}X' \tag{3.17}$$

3.1.5. Calcul des valeurs résiduelles

Le vecteur des résidus s'obtient simplement en faisant la différence entre le vecteur des valeurs observées de Y et le vecteur de valeurs prédites $\hat{Y}$.

$$\epsilon = Y - \hat{Y} \tag{3.18}$$

Propriétés

Le vecteur des résidus ϵ est orthogonal à la fois au vecteur des valeurs prédites et à la matrice X. Ce qui signifie mathématiquement que $\epsilon'\hat{Y} = 0$ et $\epsilon'X = 0$

3.1.6. Calcul de la variance totale, expliquée et résiduelle

Soit le vecteur-colonne $\bar{Y}$ rempli uniquement par la moyenne de la variable y telle que : $\bar{Y} = \begin{pmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix}$ avec $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, la variance totale s'obtient par la formule suivante :

$$S_Y^2 = \frac{1}{n}(Y - \bar{Y})'(Y - \bar{Y}) \tag{3.19}$$

Quant à la variance expliquée, elle s'obtient dans les configuration identiques en remplaçant le vecteur des valeurs observées de Y par le vecteur de valeurs prédites $\hat{Y}$. Ainsi, on a :

$$S_{\hat{Y}}^2 = \frac{1}{n}\left(\hat{Y} - \bar{Y}\right)'\left(\hat{Y} - \bar{Y}\right) \tag{3.20}$$

La variance résiduelle s'obtient de la même manière, à la simple différence que la moyenne des résidus étant nulle, ce facteur n'est pas inclus dans la formule. La formule de la variance résiduelle se présente comme suit :

$$S_\epsilon^2 = \frac{1}{n}\epsilon'\epsilon \ = \frac{1}{n}(Y-\hat{Y})'(Y-\hat{Y}) \tag{3.21}$$

Il faut simplement rappeler que la variance totale est la somme de la variance expliquée et de la variance résiduelle. Par conséquent :

$$S_Y^2 = S_{\hat{Y}}^2 + S_\epsilon^2 \tag{3.22}$$

Calcul du coefficient de détermination :

$$R_{xy}^2 = \frac{S_{\hat{Y}}^2}{S_Y^2} = 1 - \left(\frac{S_\epsilon^2}{S_Y^2}\right) \tag{3.23}$$

La racine carrée du coefficient de détermination est appelée le coefficient de corrélation multiple.

3.1.7. Matrice de variance-covariance

La matrice de variance-covariance est une matrice carrée qui contient les covariances entre l'ensemble des variables explicatives prises deux à deux. En considérant par exemple quatre variables x_1, x_2, x_3 et x_4, la matrice de variance-covariance se présente sous la forme suivante :

	x_1	x_2	x_3	x_4
x_1	$S_{x_1}^2$	$S_{x_1x_2}$	$S_{x_1x_3}$	$S_{x_1x_4}$
x_2	$S_{x_2x_1}$	$S_{x_2}^2$	$S_{x_2x_3}$	$S_{x_2x_4}$
x_3	$S_{x_3x_1}$	$S_{x_3x_2}$	$S_{x_3}^2$	$S_{x_3x_4}$
x_4	$S_{x_4x_1}$	$S_{x_4x_2}$	$S_{x_4x_3}$	$S_{x_4}^2$

Les éléments sur la diagonale représentent les variances alors que les éléments hors diagonale sont les covariances. La matrice de variance-covariance est une matrice symétrique puisque $S_{x_jx_k} = S_{x_kx_j}$ où $S_{x_jx_k}$ représente la covariance entre la variable x_j et la variable x_k. En utilisant une notation matricielle, la matrice de variance-covariance se présente comme suit :

$$S = \begin{pmatrix} S_{x_1}^2 & S_{x_1x_2} & \dots & S_{x_1x_k} \\ S_{x_2x_1} & S_{x_2}^2 & \dots & S_{x_2x_k} \\ \dots & \dots & \dots & \dots \\ S_{x_kx_1} & S_{x_kx_2} & \dots & S_{x_k}^2 \end{pmatrix} \tag{3.24}$$

3.1.8. La matrice de corrélation

La matrice des corrélations se présente sous le même format que la matrice de variance-covariance en remplaçant les covariances $S_{x_2x_1}$par les coefficients de corrélation r_{xy}. Cependant sur les diagonales, les variances $S_{x_1}^2$ seront remplacées par 1 car le coefficient de corrélation d'une variable avec elle-même est égal à 1. Ainsi la matrice de corrélation se présente comme suit :

$$R = \begin{pmatrix} 1 & r_{x_1x_2} & \cdots & r_{x_1x_k} \\ r_{x_2x_1} & 1 & \cdots & r_{x_2x_k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{x_kx_1} & S_{x_kx_2} & \ldots & 1 \end{pmatrix} \tag{3.25}$$

La matrice de corrélation est une matrice symétrique puisque $r_{x_jx_k} = r_{x_kx_j}$ où $r_{x_jx_k}$est le coefficient de corrélation entre la variable x_j et la variable x_k.

Exercices d'application

Exercice 1 : On considère le modèle suivant $y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \varepsilon_i$ avec i=1, 2,…, 10. Les données du problème sont les suivantes :

$$\sum_{i=1}^{n}(y_i)^2 = 177, \sum_{i=1}^{n} y_i = 10\,, \sum_{i=1}^{n} x_{i1}y_i = 20\,, \sum_{i=1}^{n} x_{i2}y_i = 40\,, \sum_{i=1}^{n}(x_{i1})^2 = 5, \sum_{i=1}^{n}(x_{i2})^2 = 20,$$

$$\sum_{i=1}^{n} x_{i1} = \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} x_{i1}x_{i2} = 0.$$

1- Estimer les paramètres de ce modèle
2- Calculer la SCT, calculer la SCE et en déduire la SCR
3- Calculer le R².

Exercice 2 : Dans un modèle où on cherche un ajustement linéaire de Y sur X et la constante, on dispose des résultats suivants portant sur 52 observations :

$$\widehat{y_t} = -0.43x_t + 1.286$$

$$\bar{x} = 1.063\,;\; S_y^2 = 0.00137\,;S_x^2 = 0.00686$$

Déterminez successivement les valeurs du coefficient de corrélation linéaire entre X et Y, le coefficient de détermination, la SCT, SCE et la SCR.

Exercice 3 : A partir des données du tableau ci-dessous et en utilisant le modèle linéaire estimer une fonction de production de type Cobb-Douglas

Entreprise(i)	Travail(xi)	Capital (zi)	Production (yi)
1	73	80	60
2	81	90	120
3	88	95	190
4	86	95	250
5	87	980	300
6	96	110	360
7	10	120	380
8	11	130	430
9	12	150	440

3.2. Propriétés des estimateurs

3.2.1. Esperance et Biais d'estimation

Lorsque les hypothèses de base du modèle linéaire sont satisfaites, l'estimateur des MCO est sans biais. Un estimateur est dit sans biais si son espérance mathématique est égale à la vraie valeur du paramètre à estimer. En effet étant donné que l'estimateur des MCO est $\hat{\beta}_{mco} = (X'X)^{-1}X'Y$, on peut décomposer cette expression comme suit :

$$\hat{\beta}_{mco} = (X'X)^{-1}X'(X\beta + \epsilon\ )$$

$$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon$$

Or $(X'X)^{-1}X'X = I$ où I est la matrice identité. Ainsi :

$$\hat{\beta}_{mco} = \beta + (X'X)^{-1}X'\epsilon$$

Dès lors :

$$E\left(\ \hat{\beta}_{mco}\right) = E(\beta + (X'X)^{-1}X'\epsilon)$$

$$= E(\beta) + E((X'X)^{-1}X'\epsilon)$$

$$= \beta + (X'X)^{-1}X'E(\epsilon)$$

Or $E(\epsilon) = 0$, ainsi,

$$E\left(\hat{\beta}_{mco}\right) = \beta \tag{3.26}$$

Ce qui montre que l'estimateur MCO est sans biais car l'espérance est égale à la vraie valeur du paramètre recherché.

3.2.2. Variance et Convergence

Par ailleurs, l'estimateur étant une variable aléatoire, on peut mesurer sa variance. En effet :

$$VAR\left(\ \hat{\beta}_{mco}\right) = VAR(\beta + (X'X)^{-1}X'\epsilon)$$

$$= VAR(\beta) + VAR((X'X)^{-1}X'\epsilon)$$

$$= 0 + (X'X)^{-1}X'VAR(\epsilon)X(X'X)^{-1}$$

$$= (X'X)^{-1}X'I\sigma_{\varepsilon}^{2}X(X'X)^{-1}$$

$$= \sigma_{\varepsilon}^{2}[(X'X)^{-1}X'X](X'X)^{-1}$$

$$VAR\left(\ \hat{\beta}_{mco}\right) = \sigma_{\varepsilon}^{2}(X'X)^{-1} \qquad (3.27)$$

Théorème de Gauss-Markov

Selon le théorème de Gauss-Markov, l'estimateur $\hat{\beta}_{mco}$ est le meilleur estimateur dans la classe des estimateurs linéaires sans biais de β car sa variance est la plus faible. En effet, selon ce théorème il n'existe pas d'autres estimateurs linéaires (sans biais) présentant une plus petite variance que celle de l'estimateur des MCO. Les estimateurs des MCO sont BLUE (best linear unbiased estimator). On dit qu'ils sont efficaces.

Pour démontrer cette propriété, considérons l'estimateur MCO avec

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

Où A est une matrice

Par la suite considérons un estimateur linéaire quelconque β^* telle que :

$$\beta^* = CY$$

Où C est une matrice.

Considérons une matrice B telle que :

$$B = C - A$$

On peut alors redéfinir β^* tel que :

$$\beta^* = (B + A)Y$$

Ce qui donne : $$\beta^* = (B + (X'X)^{-1}X')Y$$

Calculons l'espérance de β^*, on a :

$$E(\beta^*) = E[(B + (X'X)^{-1}X')Y]$$

$$= E[(B + (X'X)^{-1}X')(X\beta + \epsilon)]$$

$$= E[BX\beta + (X'X)^{-1}X'X\beta + B\epsilon + (X'X)^{-1}X'\epsilon]$$

$$= BX\beta + I\beta$$

$$\beta^* = (BX + I)\beta$$

Ainsi pour que β^* soit sans biais, il faut que $BX = 0$.

Calculons maintenant la variance :

$$VAR(\beta^*) = VAR[(B + (X'X)^{-1}X')Y]$$

$$= (B + (X'X)^{-1}X')I\sigma_\varepsilon^2(B + (X'X)^{-1}X')'$$

$$VAR(\beta^*) = \{BB' + BX(X'X)^{-1} + (X'X)^{-1}X'B' + (X'X)^{-1}\}\sigma_\varepsilon^2$$

Sachant $BX(X'X)^{-1} = 0$; $(X'X)^{-1}X'B' = 0$,

$$VAR(\beta^*) = \{BB' + (X'X)^{-1}\}\sigma_\varepsilon^2$$

La matrice BB' est semi-définie positive. Tous les éléments de sa diagonale sont positifs. Donc, le meilleur estimateur est obtenu quand B = 0. Ainsi, lorsque $BB' = 0$, on obtient l'estimateur $\hat{\beta}_{mco}$. C'est pourquoi on dit que $\hat{\beta}_{mco}$ est un estimateur BLUE (Best Linear Unbiased Estimator).

3.2.3. Distribution de probabilité des estimateurs

Propriété 1 :

Dans le modèle linéaire général avec des résidus normaux, on a :

$$\hat{\beta} \rightsquigarrow N(\beta, \sigma_\varepsilon^2(X'X)^{-1}) \tag{3.28}$$

Propriété 2 : Dans le modèle linéaire général avec les résidus normaux, on a le corolaire suivant :

$$(\hat{\beta} - \beta)'\frac{(X'X)}{\sigma_\varepsilon^2}(\hat{\beta} - \beta) \rightsquigarrow \chi_p^2 \tag{3.29}$$

En effet, en appliquant le théorème central limite sur le vecteur $\hat{\beta}$, on a :

$$\frac{(\hat{\beta} - \beta)}{(\sigma_\varepsilon^2(X'X)^{-1})^{\frac{1}{2}}} \rightsquigarrow N(0, I)$$

$$\Rightarrow \frac{((X'X)^{-1})^{-\frac{1}{2}}}{\sigma_\varepsilon}(\hat{\beta} - \beta) \rightsquigarrow N(0, I)$$

$$\Rightarrow \frac{(X'X)^{\frac{1}{2}}}{\sigma_\varepsilon}\left(\hat{\beta}-\beta\right) \rightsquigarrow N(0,I) \tag{3.30}$$

En élevant cette expression au carré et en faisant la somme de ces carrés, on retrouve une loi de khi-deux:

$$\left(\hat{\beta}-\beta\right)'\frac{\left((X'X)^{\frac{1}{2}}\right)'}{\sigma_\varepsilon}\ \frac{\left((X'X)^{\frac{1}{2}}\right)}{\sigma_\varepsilon}\left(\hat{\beta}-\beta\right) \rightsquigarrow \chi_p^2$$

$(X'X)^{\frac{1}{2}}$ est une matrice idempotente par conséquent $\left((X'X)^{\frac{1}{2}}\right)'\left((X'X)^{\frac{1}{2}}\right) = (X'X)$.
Ainsi, on a :

$$\left(\hat{\beta}-\beta\right)'\frac{(X'X)}{\sigma_\varepsilon^2}\left(\hat{\beta}-\beta\right) \rightsquigarrow \chi_p^2$$

Par ailleurs, on pouvait se servir d'une autre propriété pour démontrer que cette distribution suit une loi de khi-deux. En effet, soit un vecteur aléatoire U de distribution normale, de moyenne nulle et de variance I. Si P est une matrice symétrique, idempotente et de rang p, alors $U'PU$ est une variable aléatoire qui suit une loi de χ_p^2 à p degrés de liberté.

Démonstration :

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

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

$$\left(\hat{\beta}-\beta\right)'\frac{(X'X)}{\sigma_\varepsilon^2}\left(\hat{\beta}-\beta\right) = (\epsilon'X(X'X)^{-1})\frac{(X'X)}{\sigma_\varepsilon^2}((X'X)^{-1}X'\epsilon)$$

$$= \frac{\epsilon'}{\sigma_\varepsilon}X(X'X)^{-1}X'\frac{\epsilon}{\sigma_\varepsilon}$$

Et comme la matrice $X(X'X)^{-1}X'$ est symétrique et idempotente et de rang p et que $\frac{\epsilon\prime}{\sigma_\varepsilon}$ est un vecteur multinormal, par conséquent, on a bien :

$$\left(\hat{\beta}-\beta\right)'\frac{(X'X)}{\sigma_\varepsilon^2}\left(\hat{\beta}-\beta\right) \rightsquigarrow \chi_p^2$$

Propriété 3 : Dans le modèle linéaire général avec les résidus normaux, on a le corolaire suivant (tiré de l'expression de la démonstration précédente):

$$\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma_\varepsilon^2} \rightsquigarrow \chi_{n-p}^2$$

En effet, $\hat{\epsilon} = Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = P_X^{\perp}\epsilon$

Avec $P_X^{\perp} = I - X(X'X)^{-1}X'$.

Or $P_X^{\perp}$ est une matrice symétrique et idempotente et de rang n-p. On obtient alors :

$$\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma_\varepsilon^2} = \frac{\epsilon'}{\sigma_\varepsilon} P_X^{\perp\prime} P_X^{\perp} \frac{\epsilon}{\sigma_\varepsilon} \rightsquigarrow \chi^2_{n-p}$$

$$= \frac{\epsilon'}{\sigma_\varepsilon} P_X^{\perp} \frac{\epsilon}{\sigma_\varepsilon}$$

Ainsi

$$\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma_\varepsilon^2} = \frac{\epsilon'}{\sigma_\varepsilon} P_X^{\perp} \frac{\epsilon}{\sigma_\varepsilon} \rightsquigarrow \chi^2_{n-p}$$

$$\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma_\varepsilon^2} \rightsquigarrow \chi^2_{n-p} \Rightarrow \frac{(n-p)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \rightsquigarrow \chi^2_{n-p}$$

Tableau 2 : Récapitulatif sur les paramètres du modèle linéaire multiple

Estimateur	Expression simple	Expression développée	Esperance	Variance
$\hat{\beta}$	$(X'X)^{-1}X'Y$	$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon$	β	$\sigma_\varepsilon^2(X'X)^{-1}$
$\hat{\sigma}_\varepsilon^2$	$\frac{\hat{\epsilon}'\hat{\epsilon}}{n-p}$	$\frac{(Y-X\beta)'(Y-X\beta)'}{n-p}$	σ_ε^2	

Résumé sur les lois de probabilité des paramètres

$$\hat{\beta} \rightsquigarrow N(\beta, \sigma_\varepsilon^2(X'X)^{-1})$$

$$(\hat{\beta} - \beta)' \frac{(X'X)}{\sigma_\varepsilon^2} (\hat{\beta} - \beta) \rightsquigarrow \chi^2_p$$

$$\frac{(n-p)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \rightsquigarrow \chi^2_{n-p}$$

3.3. Tests d'hypothèses sur les coefficients estimés

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad j = 1, 2, .. p$$

3.3.1. Test sur les coefficients individuels β_j

Ce test se formule avec l'hypothèse suivante :

$$\begin{cases} H_0 \;\; \beta_j = \beta j_0 \\ H_1 \;\; \beta_j \neq \beta_{j0} \end{cases}$$

Où β_j est le j-ième élément de β.

Sous H0, on a

$$\widehat{\beta_J} \rightsquigarrow N\left(\beta_{j0}, \sigma^2_{\widehat{\beta_J}}\right)$$

$$\text{Où} \quad \sigma^2_{\widehat{\beta}_j} = [\sigma^2_\varepsilon (X'X)^{-1}]_{jj}$$

$\sigma^2_{\widehat{\beta_J}}$représente en fait au j-ième élément diagonal de la matrice de variance-covariance $\sigma^2_\varepsilon (X'X)^{-1}$. C'est donc la variance du paramètre $\widehat{\beta_J}$.

La valeur estimée de cette variance s'écrit telle que :

$$\hat{\sigma}^2_{\widehat{\beta}_j} = [\hat{\sigma}^2_\varepsilon (X'X)^{-1}]_{jj}$$

On sait par ailleurs que :

$$\frac{(n-p)\hat{\sigma}^2_{\widehat{\beta}}}{\sigma^2_{\widehat{\beta}}} = \frac{(n-p)\hat{\sigma}^2_\varepsilon}{\sigma^2_\varepsilon} \rightsquigarrow \chi^2_{n-p}$$

Ce qui implique donc que :

$$\frac{(n-p)\hat{\sigma}^2_{\widehat{\beta}_j}}{\sigma^2_{\widehat{\beta}_j}} \rightsquigarrow \chi^2_{n-p}$$

De plus on peut montrer que :

$$\frac{\hat{\beta}_j - \beta j_0}{\sigma_{\widehat{\beta}_j}} \rightsquigarrow N(0,1)$$

Le rapport entre ces deux dernières expression telle que :

$$\frac{\frac{\hat{\beta}_j - \beta j_0}{\sigma_{\hat{\beta}_j}}}{\frac{\hat{\sigma}_{\hat{\beta}_j}}{\sigma_{\hat{\beta}_j}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-p}}{(n-p)}}} \rightsquigarrow T(n-p)$$

$$\frac{\hat{\beta}_j - \beta j_0}{\hat{\sigma}_{\hat{\beta}_j}} \rightsquigarrow T(n-p) \qquad (3.31)$$

On retombe alors dans le cadre d'un test de Student classique.

3.3.2. Test sur une combinaison linéaire de coefficients (Test de Wald)

Le test de Wald se présente sous la forme d'une combinaison linéaire des coefficients formulé comme suit :

$$\begin{cases} H_0 \;\; R\beta = r \\ H_1 \;\; R\beta \neq r \end{cases}$$

R est une matrice q× q avec q≤ p et r un vecteur colonne de dimension q. En outre R est supposée de rang q.

La forme générale de la matrice R et du vecteur r est la suivante :

$$R = \begin{pmatrix} R_{10} & R_{11} & \dots & R_{1p} \\ R_{20} & R_{21} & \dots & R_{2p} \\ \dots & \dots & \dots & \dots \\ R_{q0} & R_{q1} & \dots & R_{qp} \end{pmatrix}$$

$$r = \begin{pmatrix} r_1 \\ r_2 \\ \dots \\ r_q \end{pmatrix}$$

Lorsque q=1 alors la matrice R se réduit à un vecteur-ligne et le vecteur r se réduit à un scalaire. $R = \begin{pmatrix} R_{10} & R_{11} \dots & R_{1p} \end{pmatrix}$ et r = r_1. Ce cas se présente lorsqu'il n'existe qu'une seule contrainte linéaire qui peut se présenter comme suit :

$$H_0 : \begin{pmatrix} R_{10} & R_{11} \dots & R_{1p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = r_1$$

Soit

$$H_0 : R_{10}\beta_0 + R_{11}\beta_1 + R_{12}\beta_2 + \ldots R_{1p}\beta_p = r_1$$

Ce test inclut le test sur les coefficients individuels comme un cas particulier. En effet, dans le cas d'un test sur coefficient individuel où l'hypothèse nulle est $H_0 \ \beta_j = \beta_{j0}$, la contrainte se présentera alors comme suit :

$$H_0 : 0\beta_0 + 0\beta_1 + \cdots + 1\beta_j + \ldots + 0\beta_p = \beta j_0$$

Dans ce cas, $R = (0 \quad 0 \ldots 1 \ldots 0)$ et $r_1 = \beta j_0$

Le test inclut également le test de significativité globale comme un cas particulier en attribuant 1 à l'ensemble des éléments diagonaux de R (matrice identité) et 0 en dehors de ces éléments. En plus, on attribue 0 à r. En effet lorsqu'on a plusieurs contraintes, l'hypothèse nulle du test de Wald se présente comme suit :

$$H0\begin{cases} R_{10}\beta_0 + R_{11}\beta_1 + R_{12}\beta_2 + \ldots R_{1p}\beta_p = r_1 \\ R_{20}\beta_0 + R_{21}\beta_1 + R_{22}\beta_2 + \ldots R_{2p}\beta_p = r_2 \\ \ldots \quad \ldots \ldots \ldots \quad \ldots \ldots \quad \ldots \\ R_{q0}\beta_0 + R_{q1}\beta_1 + R_{q2}\beta_2 + \ldots R_{qp}\beta_p = r_q \end{cases}$$

Pour effectuer le test de significativité globale (nullité de tous les paramètres exception faite de la constante), on a la formulation suivante :

$$H0\begin{cases} 0\beta_0 + 1\beta_1 + 0\beta_2 + \ldots 0\beta_p = 0 \\ 0\beta_0 + 0\beta_1 + 1\beta_2 + \ldots 0\beta_p = 0 \\ \ldots \quad \ldots \ldots \ldots \quad \ldots \ldots \quad \ldots \\ 0\beta_0 + 0\beta_1 + 0\beta_2 + \ldots 1\beta_p = 0 \end{cases}$$

Ce qui permet donc d'écrire :

$$\begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \ldots \\ 0 \end{pmatrix}$$

Ainsi, la matrice R se présente comme suit :

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

En somme pour déterminer la matrice R du test de significativité sous forme de test de contrainte, on attribue 1 à tous les éléments diagonaux en dehors de la colonne destinée à la constante. Et 0 à l'ensemble des r_q (Vecteur nul pour r).

$$H_0 : \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & .. & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \dots \\ r_q \end{pmatrix}$$

L'hypothèse nulle se présente alors comme suit :

$$H0 \begin{cases} \beta_1 = 0 \\ \beta_2 = 0 \\ \dots \\ \beta_p = 0 \end{cases}$$

Dans tous les autres cas, quel que soit la nature de la combinaison, l'hypothèse nulle du test se présente comme suit :

$$H_0 \;\; R\beta = r \tag{3.32}$$

Calculons d'abord $R\hat{\beta} - r$ sous H0.
Sous H0, on a :

$$R\hat{\beta} - r = R(X'X)^{-1}X'Y - r$$
$$= R(X'X)^{-1}X'(X\beta + \epsilon) - r$$

$$= R(X'X)^{-1}X'\epsilon + (R\beta - r)$$

Sous H0,$(R\beta - r) = 0$, ainsi

$$R\hat{\beta} - r = R(X'X)^{-1}X'\epsilon$$

Calculons ensuite la variance de $R\hat{\beta} - r$ (en vue de formuler le théorême central-limite) pour déterminer sa distribution de loi. On a :

$$VAR\left(R\hat{\beta} - r\right) = VAR\left(R\hat{\beta}\right)$$
$$= R\left[VAR\left(\hat{\beta}\right)\right]R'$$

$$= R(\sigma_\varepsilon^2(X'X)^{-1})R'$$

$$VAR\left(R\hat{\beta} - r\right) = VAR\left(R\hat{\beta}\right) = \sigma_\varepsilon^2 R(X'X)^{-1}R'$$

Appliquons maintenant le théorème central-limite sur $R\hat{\beta} - r$ (encore appelée forme quadratique de $VAR\left(R\hat{\beta}\right)$. On a :

$$\left(R\hat{\beta}-r\right)'\left(VAR\left(R\hat{\beta}\right)\right)^{-1}\left(R\hat{\beta}-r\right)=\frac{1}{\sigma_\varepsilon^2}\epsilon W\epsilon'$$

Où

$$W=X(X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}R(X'X)^{-1}X'$$

Comme W est une matrice idempotente, alors, on peut montrer que :

$$\frac{1}{\sigma_\varepsilon^2}\epsilon W\epsilon' \rightsquigarrow \chi_q^2$$

Ainsi

$$\left(R\hat{\beta}-r\right)'\left(VAR\left(R\hat{\beta}\right)\right)^{-1}\left(R\hat{\beta}-r\right)\rightsquigarrow\chi_q^2$$

Avec $VAR\left(R\hat{\beta}\right)=\sigma_\varepsilon^2R(X'X)^{-1}R.$

Toutefois on ne peut réaliser directement le test khi-deux à partir de cette formule du fait que l'expression de la forme quadratique dépend de σ_ε^2 (qui est inconnue). Il faut alors partir de la loi distribution du rapport des variances (estimée et vraie valeur). Ainsi, on a :

$$\frac{(n-p)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}\rightsquigarrow\chi_{n-p}^2\Leftrightarrow$$

$$\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma_\varepsilon^2}\rightsquigarrow\chi_{n-p}^2\Leftrightarrow$$

On sait, en plus qu'en faisant le rapport entre deux lois de khi-deux divisées par leur degré de liberté respective, on obtient une loi de Fisher. Ainsi, on a :

$$\frac{\dfrac{\left(R\hat{\beta}-r\right)'\{R(X'X)^{-1}R'\}^{-1}\left(R\hat{\beta}-r\right)}{\sigma_\varepsilon^2}}{\dfrac{\dfrac{\hat{\epsilon}'\hat{\epsilon}}{\sigma_\varepsilon^2}}{n-p}}=\rightsquigarrow F(q,n-p)$$

Au final, en utilisant l'expression de la forme quadratique du vecteur $\hat{\beta}$ et du rapport des variances des erreurs, on obtient une loi de Fisher telle que :

$$F=\frac{\dfrac{\left(R\hat{\beta}-r\right)'\{R(X'X)^{-1}R'\}^{-1}\left(R\hat{\beta}-r\right)}{q}}{\dfrac{\hat{\epsilon}'\hat{\epsilon}}{n-p}}\rightsquigarrow F(q,n-p)\qquad(3.33)$$

Dès lors si le Fisher calculé est supérieur au Fisher lu dans la table, on rejette l'hypothèse nulle $H_0\ R\beta=r$. Dans ce cas, la contrainte spécifiée n'est pas valide.

3.4. Estimateur des moindres carrés contraints

Pour mettre en perspective la notion d'estimations sous contraintes linéaires, partons du cas d'une fonction de production macroéconomique de type Cobb-Douglas à rendement d'échelle constant. Soit :

$$logY_t = \alpha + \mu logN_t + \gamma logK_t + \delta t + u_t \qquad (3.34)$$

Où Y_t est le niveau de production, N_t la quantité de main d'œuvre, K_t le stock de capital et t une tendance et u_t le terme d'erreur.
Pour commencer on suppose que $\mu + \gamma = 1$, ce qui traduit une situation de rendement d'échelle constant. Définissons une contrainte supplémentaire selon laquelle $\alpha = 2\delta$.
Pour estimer ce modèle, il faut d'abord spécifier la matrice des coefficients de contrainte R. En effet on a :

$$\begin{cases} \mu + \gamma = 1 \\ \alpha - 2\delta = 0 \end{cases}$$

Sachant que le vecteur des coefficients se présente tel que $\beta = \begin{pmatrix} \alpha \\ \mu \\ \gamma \\ \delta \end{pmatrix}$ alors, les deux contraintes peuvent s'écrire sous la forme matricielle comme suit :

$$\begin{pmatrix} 0\alpha + & 1\mu + & 1\gamma + & 0\delta \\ 1\alpha & 0\mu & 0\gamma & -2\delta \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Ce qui donne :

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} \alpha \\ \mu \\ \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

La contrainte s'écrit alors comme $R\beta = r$ avec $R = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & -2 \end{pmatrix}$ et $r = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

Ainsi, connaissant R et r, l'estimation du modèle linéaire peut directement se déduire de l'estimateur des moindres carrés ordinaires non contraintes par la formule suivante :

$$\hat{\beta}_{CC} = \hat{\beta}_{mco} + (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}\left(r - R\hat{\beta}_{mco}\right) \qquad (3.35)$$

On peut simplement faire remarquer que lorsque $\left(r - R\hat{\beta}_{mco}\right) = 0$, c'est-à-dire lorsque la contrainte est valide alors l'estimateur des moindres carrés contraints équivaut à l'estimateur des moindres carrés ordinaires

$$\hat{\beta}_{CC} = \hat{\beta}_{mco}$$

3.4.1. Propriété de l'estimateur des moindres carrés contraints

Espérance de $\hat{\beta}_{CC}$:

L'espérance de $\hat{\beta}_{CC}$ s'écrit comme suit :

$$E\left(\hat{\beta}_{CC}\right) = \beta - \ (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}(R\beta - r)$$

Il apparait de cette expression que lorsque les contraintes sont valides $(R\beta - r)$ alors l'estimateur du moindre carrés contraint est sans biais car :

$$E\left(\hat{\beta}_{CC}\right) = \beta \qquad (3.36a)$$

Par ailleurs, on peut montrer que si les contraintes sont valides, alors l'estimateur des moindres carrés contraints est optimal parmi les estimateurs linéaires sans biais de β vérifiant la contrainte.

Variance estimée des résidus

L'estimateur de la variance est fondé sur la somme des carrés des résidus contraints. Il est définit de la façon suivante :

$$\hat{\sigma}^2_{\varepsilon_{cc}} = \frac{\hat{\epsilon}'_{cc}\hat{\epsilon}_{cc}}{n - p + q} \qquad (3.36b)$$

Avec $\hat{\epsilon}_{cc} = Y - X\hat{\beta}_{CC}$, p le nombre de paramètres du modèle y compris la constante et q le nombre de contraintes.

3.4.2. Le test de Fisher (sur la validité des contraintes)

Pour tester la validité des contraintes, on tombe dans le cadre classique d'un test de Wald contraintes linéaires. L'hypothèse nulle de ce test est la suivante :

$$\begin{cases} H_0 \ R\beta = r \\ H_1 \ R\beta \neq r \end{cases}$$

Nous disposons de deux manières pour calculer la statistique de Fisher.

La première méthode est d'utiliser la forme quadratique de la matrice des contraintes et le rapport des variances des résidus. La formule adoptée dans cette méthode est la suivante :

$$F = \frac{\frac{\left(R\hat{\beta}_{mco} - r\right)'\{R(X'X)^{-1}R'\}^{-1}\left(R\hat{\beta}_{mco} - r\right)}{q}}{\frac{\hat{\epsilon}'\hat{\epsilon}}{n-p}} \rightsquigarrow F(q, n-p)$$

Dans ce test, on a pas besoin de calculer l'estimateur du moindre carrés contraints ($\hat{\beta}_{cc}$).

Dans la seconde méthode, on estime à la fois l'estimateur $\hat{\beta}_{mco}$ et l'estimateur $\hat{\beta}_{cc}$. Ensuite, on calcule la statistique de Fisher à partir de la somme des carrés des résidus de chaque modèle. La formule se présente comme suit :

$$F = \frac{\left(\frac{SCR_{CC} - SCR_{mco}}{ddl_{CC} - ddl_{mco}}\right)}{\left(\frac{SCR_{mco}}{ddl_{mco}}\right)} \rightsquigarrow F(ddl_{CC} - ddl_{mco}, ddl_{mco})$$

Où SCR_{CC} et SCR_{mco} représentent respectivement la somme des carrés des résidus issus du modèle de moindre carrés contraints et du modèle de moindre carrés ordinaires. ddl_{CC} et ddl_{mco} sont les degrés de liberté respectifs.

$$ddl_{CC} = n - p + q$$

$$ddl_{mco} = n - p$$

n est le nombre d'observations, p est le nombre de paramètres du modèle et q le nombre de contraintes.

Dans les deux formules, la statistique de Fisher est à comparer à la valeur tabulée en vue du rejet ou du non-rejet de H0.

3.4.3. La statistique de Fisher dans le cadre du test de Chow (ou test de changement de régime)

Dans des cas où l'on souhaite savoir si un modèle de comportement reste stable entre deux sous-périodes, on peut utiliser le Test de Fisher dans une configuration connue sous le test de Chow.

En effet, si l'on dispose d'observations sur deux sous-périodes de $t = 1$ à T_1 et de $t = T1 + 1$ à $t = T1 + T2$, on tente alors de modéliser ces données par deux modèles :

$$\begin{cases} Y_1 = X_1\beta_1 + u_1 \; si \; t = 1, \dots, T_1 \\ Y_2 = X_2\beta_2 + u_2 \quad si \;\; t = T_1 + 1, \dots, T \end{cases}$$

L'hypothèse nulle de ce test est la suivante :

$$\begin{cases} H_0 \ \beta_1 = \beta_2 \\ H_1 \ \beta_1 \neq \beta_2 \end{cases}$$

Ce test est en fait un cas particulier du test de Fisher dans la mesure dans le sens où la contrainte est définie par H0 avec H1 qui représente l'estimations non contrainte. Dans cette configuration, la statistique de Fisher se calcule comme suit :

$$F = \frac{\left(\frac{SCR_{H_0} - SCR_{H_1}}{ddl_{H_0} - ddl_{H_1}}\right)}{\left(\frac{SCR_{H_1}}{ddl_{H_1}}\right)} \rightsquigarrow F\left(ddl_{H_0} - ddl_{H_1}, ddl_{H_1}\right)$$

Où SCR_{H_0} et SCR_{H_1} représentent respectivement la somme des carrés des résidus issus du modèle du modèle estimé sous H_0 (modèle contraint) et la somme des carrés des résidus issus du modèle du modèle estimé sous H_1.

$SCR_{H_0} = SCR_0$ obtenu sur l'échantillon total.

$$SCR_{H_1} = SCR_1 + SCR_2$$

SCR_{H_1} est la somme de deux sommes carrés des résidus (obtenue en estimant le modèle sur chacun des deux sous-échantillons).

ddl_{H_0} et ddl_{H_1} représentent les degrés de liberté respectifs de SCR_{H_0} et de SCR_{H_1}.

$$ddl_{H_0} = p$$

$$ddl_{H_1} = T_1 + T_2 - 2p$$

où p est le nombre de paramètres du modèle y compris la constante. T_1 est le nombre d'observations de la première période et T_2 le nombre d'observation de la seconde période.

Lorsque la statistique calculée est supérieure à la valeur tabulée aux degrés de libertés définis alors on rejette l'hypothèse nulle. Ce qui signifie que le coefficient n'est pas stable entre les deux périodes.

3.5. Estimation par maximum de vraisemblance

Le principe de l'estimation par maximum de vraisemblance consiste à faire une hypothèse sur la distribution de probabilité de ε_i . En effet, on suppose, que les ε_i suivent une loi normale de moyennes nulle et de variance σ_ε^2 .

$$\varepsilon_i \rightsquigarrow N(0, \sigma_\varepsilon^2)$$

En reprenant, l'équation du modèle linéaire initiale, on a :

$$Y = X\beta + \epsilon \tag{3.36}$$

L'hypothèse de normalité du vecteur des résidus se présente comme suit :

$$\epsilon \rightsquigarrow N(0, I\sigma_\varepsilon^2)$$

Il faut aussi noter que comme $Y = X\beta + \epsilon$ et que $\epsilon \rightsquigarrow N(0, I\sigma_\varepsilon^2)$ alors on aura :

$$Y \rightsquigarrow N(X\beta, I\sigma_\varepsilon^2)$$

En effet

$$E(Y) = E(X\beta + \epsilon) = E(X\beta) + E(\epsilon) = X\beta$$

$$VAR(Y) = VAR(X\beta + \epsilon) = VAR(X\beta) + VAR(\epsilon) = 0 + I\sigma_\varepsilon^2 = I\sigma_\varepsilon^2$$

Fonction de densité et fonction de vraisemblance

Une variable aléatoire X est dite normale de moyenne μ et de variance σ^2 si sa densité vaut :

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right) \tag{3.37}$$

Lorsque X est une matrice telle que $= (x_1, x_2, \dots, x_k)'$, on parle de distribution multinormale où la moyenne est $\mu = (\mu_1, \mu_2, \dots, \mu_k)'$ et de variance-covariance noté $\sum$ Dans ce cas, la fonction de densité se présente comme suit :

$$f(X) = \frac{1}{(2\pi)^{\frac{1}{2}}|\sum|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left((X-\mu)\sum^{-1}\right)^2\right) \tag{3.38}$$

Connaissant la fonction de densité d'une loi multinormale, on peut spécifier la fonction de densité de la variable ϵ sachant que $Y = X\beta + \epsilon$. On obtient alors :

$$f(Y) = \frac{1}{(2\pi)^{\frac{1}{2}}(I\sigma_\varepsilon^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left((Y-X\beta)(I\sigma_\varepsilon^2)^{-1}\right)^2\right)$$

$$f(Y) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(Y - X\beta)^2\right) \tag{3.39}$$

Ainsi, pour estimer, l'équation (3.36) par maximum de vraisemblance, on pose d'abord l'expression de la fonction de vraisemblance. Ainsi, en notant par $L(\beta_0, \beta_1, \beta_2, \dots \beta_k, \sigma_\varepsilon^2)$, la fonction de vraisemblance est :

$$L(\beta_0, \beta_1, \beta_2, \dots \beta_k, \sigma_\varepsilon^2) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \left[\frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \right) \right]$$

$$L(\beta_0, \beta_1, \beta_2, \dots \beta_k, \sigma_\varepsilon^2) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_j \right)^2 \right)$$

Cependant, étant donné le nombre de paramètre à estimer, pour alléger l'étape des dérivations, il est judicieux de présenter la fonction de vraisemblance sous la forme matricielle afin de faciliter la dérivation. En effet, la forme matricielle de la fonction de vraisemblance est Y fonction de va

$$L(\beta, \sigma_\varepsilon^2) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(Y - X\beta)'(Y - X\beta)\right)$$

Par ailleurs, comme il est plus plus facile de chercher à maximiser le logarithme de la fonction de vraisemblance plutôt que la fonction elle-même, on prend le logarithme de la fonction de vraisemblance :

$$LogL(\beta, \sigma_\varepsilon^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}(Y - X\beta)'(Y - X\beta)$$

La méthode du maximum de vraisemblance consiste à choisir les paramètres β et σ_ε^2 de sorte à maximiser cette fonction de vraisemblance. Pour cela, il faut dériver la dérivée et retrouver les conditions de premier ordre afin d'en déduire chacun des paramètres. En dérivant cette fonction par rapport à β et à σ_ε^2, on trouve :

$$\begin{cases} \dfrac{\partial LogL(\beta, \sigma_\varepsilon^2)}{\beta} = \dfrac{1}{\sigma_\varepsilon^2}(X'Y - X'X\beta) = 0 \\ \dfrac{\partial LogL(\beta, \sigma_\varepsilon^2)}{\sigma_\varepsilon^2} = -\dfrac{n}{2}\sigma_\varepsilon^2 + \dfrac{1}{2(\sigma_\varepsilon^2)^2}(Y - X\beta)'(Y - X\beta) = 0 \end{cases}$$

En résolvant ce système on trouve :

$$\hat{\beta}_{mv} = (X'X)^{-1}X'Y \tag{3.40}$$

On constate alors que sous l'hypothèse de normalité, l'estimateur de maximum de vraisemblance (MV) est égal à l'estimateur des moindres carrés ordinaires.

Si $\epsilon \rightsquigarrow N(0, I\sigma_\varepsilon^2) \Leftrightarrow Y \rightsquigarrow N(X\beta, I\sigma_\varepsilon^2) \Rightarrow \hat{\beta}_{mv} = \hat{\beta}_{mco} = (X'X)^{-1}X'$

On montre par ailleurs que la variance de l'estimateur MV est :

$$VAR\left(\hat{\beta}_{MV}\right) = \sigma_\varepsilon^2 (X'X)^{-1} \tag{3.41}$$

Cependant σ_ε^2 étant, en général inconnu, il faut considérée sa valeur estimée $\hat{\sigma}_\varepsilon^2$:

$$\hat{\sigma}_\varepsilon^2 = \frac{\epsilon'\epsilon}{n-k} \tag{3.42}$$

Où n est le nombre d'observations et k est le nombre de paramètres à estimer (y compris la constante). k correspond au rang de la matrice. Ainsi $n-k$ est le nombre de degrés de liberté. Avec cette variance estimée, la variance de l'estimateur MV est :

$$VAR\left(\hat{\beta}_{MV}\right) = \hat{\sigma}_\varepsilon^2 (X'X)^{-1} \tag{3.43}$$

Cette expression représente la variance estimée de l'estimateur de MV. Il faut aussi noter son espérance est égale β car

$$E\left(\hat{\beta}_{MV}\right) = \beta \tag{3.44}$$

Propriétés

Un estimateur est efficace ou de variance minimum si sa variance est plus petite ou égale que tous les estimateurs du paramètre. Et un estimateur $\hat{\theta}$ est dit convergent, s'il converge en probabilité vers le paramètre à estimer, c'est-à-dire :

$$\lim_{n\to\infty} Pr\left(\left|\hat{\theta} - \theta\right| > \tau\right) = 0 \tag{3.45}$$

Où τ est une quantité arbitrairement petite.

Si l'estimateur du maximum de vraisemblance admet une solution unique, alors cet estimateur est convergent et asymptotiquement efficace du paramètre. De plus, cet estimateur converge en loi vers une normale.

$$\hat{\beta}_{MV} \rightsquigarrow N(\beta, \hat{\sigma}_\varepsilon^2 (X'X)^{-1}) \tag{3.46}$$

Cependant, l'estimateur du maximum de vraisemblance n'est pas nécessairement sans biais. L'estimateur du maximum de vraisemblance de σ_ε^2 est en effet biaisé.

Exercices d'application

Exercice 1 : Soit y une suite de variables aléatoires (v.a.) indépendantes et identiquement distribuées (iid) suivant une loi normale de moyenne μ et de variance σ^2 si sa densité vaut :

1. Estimez μ par la méthode du maximum de vraisemblance si l'on suppose que σ^2 est connue.

2. Estimez σ^2 par la méthode du maximum de vraisemblance si l'on suppose que μ est connue.

3. Estimez μ et σ^2 par la méthode du maximum de vraisemblance si l'on suppose que μ et σ^2 sont inconnues.

Exercice 2 On considère un modèle linéaire multiple où une variable dépendante Y est expliquée par un ensemble de variable X tel que :

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix}$$

En supposant n observations indépendantes et identiquement distribuées (iid) suivant une loi normale. Ecrivez la fonction de vraisemblance lorsque le vecteur de paramètre est $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

2. Ecrivez la fonction de vraisemblance de manière scalaire (et non sous la forme matricielle).

3. Annulez les dérivées partielles par rapport à β_0, β_1, σ^2.

CHAPITRE 4. LE MODELE LINEAIRE GENERALISE

Les modèles que nous avons étudiés jusque-là ont été élaborés sur les hypothèses fondamentales suivantes:

1.) $E(\varepsilon_i) = 0$
2.) $V(\varepsilon_i) = \sigma^2$
3.) $COV(\varepsilon_i, \varepsilon_j) = 0$
4.) $COV(x_i, \varepsilon_i) = 0$

Sous ces hypothèses, l'estimation des paramètres du modèle par la méthode des moindres carrés demeurent les meilleurs estimateurs linéaires sans biais. Mais la violation de l'une de ces hypothèses, entraîne divers problèmes économétriques notamment des biais d'estimation, l'inefficience des estimateurs, etc. Le but de cette section est de présenter les différentes méthodes d'estimation compte tenu du problème économétrique soulevé par la violation de l'hypothèse.

D'une manière générale, trois principaux problèmes sont rencontrés : l'hétéroscédasticité, l'autocorrélation des erreurs ou l'endogéneité (ou problème de corrélation entre les variables explicatives et les résidus).

4.1. Test de normalité des résidus

D'une manière générale, la normalité des résidus est construite à partir des hypothèses 1) et 2) que sont respectivement : $E(\varepsilon_i) = 0$; $V(\varepsilon_i) = \sigma^2$.

La procédure couramment utilisée pour tester la normalité des résidus est le test de Jarque-Bera. La statistique du test de Jarque-Bera est calculée à partir des deux caractéristiques principales d'une distribution normale : le coefficient d'asymétrie (skewness) et le coefficient d'aplatissement (kurtosis). Ces deux coefficients se calculent comme suit :

$$S = \frac{E[(\varepsilon_i)^3]}{(\sigma_i)^3} \qquad (4.1a)$$

$$K = \frac{E[(\varepsilon_i)^4]}{(\sigma_i)^4} \qquad (4.1b)$$

Le coefficient d'asymétrie correspond au rapport entre le moment d'ordre 3 et le cube de l'écart-type. Le coefficient d'aplatissement, quant à lui, correspond au rapport entre le moment d'ordre 4 et l'écart-type élevé à la puissance 4.

Pour une distribution normale, le coefficient d'asymétrie est nécessairement nul (puisque le moment d'ordre 3 existe). De même pour une loi normale, le coefficient d'aplatissement est égal à 3. Le test de normalité consiste donc à tester conjointement :

$$H_0 : S = 0 \; et \; K = 3$$

La statistique de test proposée alors par Jarque-Bera est la suivante :

$$JB = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \rightsquigarrow \chi^2(2) \qquad (4.2)$$

Où n est la taille de l'échantillon (nombre d'observations).

Lorsque la statistique JB est supérieure au khi-deux lu dans la table, on rejette H_0. Dans ce cas, les résidus ne suivent pas une loi normale.

4.2. Test d'hétéroscédasticité

Jusque-là, nous avons fait l'hypothèse que la variance du terme d'erreur était constante conditionnellement aux variables explicatives. On dit alors qu'il y a homoscédasticité. Intuitivement, cela veut dire que la variance du terme d'erreur est constante peu importe le niveau des variables explicatives. Dans cette section, nous étudions ce qui arrive lorsque cette hypothèse est relâchée.

L'hétéroscédasticité traduit une situation où la variance des erreurs n'est plus identique (constante) d'un individu à un autre. En présence d'hétéroscédasticité, l'estimateur des moindres carrés ordinaires reste toujours sans biais. Mais, il est inefficace et sa variance est biaisée. L'estimateur n'est plus BLUE. Néanmoins des corrections peuvent être apportées à l'estimateur pour le rendre efficace avec une variance plus faible et non biaisée.

On est en présence d'hétéroscédasticité lorsque :

$$V(\varepsilon_i) = \sigma_i^2 \; \neq \sigma^2 \qquad (4.3)$$

De façon matricielle, en partant du modèle tel que:

$$Y = X\beta + \epsilon$$

La matrice de variance-covariance qui s'exprime telle que $E(\epsilon'\epsilon)$ est une matrice diagonale dont la forme générale est la suivante :

$$\Omega = \begin{pmatrix} \sigma_{\varepsilon_1}^2 & 0 & \dots & \dots & 0 \\ 0 & \sigma_{\varepsilon_2}^2 & \dots & \dots & \dots \\ \dots & \dots & \sigma_{\varepsilon_i}^2 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \sigma_{\varepsilon_n}^2 \end{pmatrix}$$

$$E(\epsilon'\epsilon) = \Omega = Diag(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$$

A cause de l'hétéroscédasticité, on a $\Omega \neq \sigma^2 I_N$.

$$\Omega = \begin{pmatrix} \sigma_{\varepsilon_1}^2 & 0 & \dots & \dots & 0 \\ 0 & \sigma_{\varepsilon_2}^2 & \dots & \dots & \dots \\ \dots & \dots & \sigma_{\varepsilon_i}^2 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \sigma_{\varepsilon_n}^2 \end{pmatrix} \neq \sigma_\varepsilon^2 I_N$$

Puisque l'hétéroscédasticité invalide les résultats des tests (particulièrement les tests t et F) et fait en sorte que les MCO ne sont plus BLUE, il apparaît utile de pouvoir tester sa présence. Nous pouvons toujours tracer un graphique de "points" et observer s'il y a de l'hétéroscédasticité, cependant il est préférable de procéder à un test formel. Les trois prochaines sous-sections portent sur de tels tests. Dans cette section, nous allons d'abord discuter des méthodes de détection de l'hétéroscédasticité avant de présenter la méthode de correction.

Détection graphique

En général, l'hétéroscédasticité peut être visible à partir d'un graphique lorsqu'on représente le nuage de points des résidus en fonction des différentes variables explicatives. Par exemple, en présence d'hétéroscédasticité, le nuage de points tend à « s'élargir », tel un entonnoir c'est-à-dire où la variance des résidus n'est pas constante et tend à être plus ou moins forte selon l'intervalle de Y considéré. Même s'il y a bien une relation linéaire entre les variables X et Y, mais le nuage des résidus rend une forme en "entonnoir" ce qui signifie que les estimations de Y en fonction de X sont très bonnes pour des valeurs petites de Y mais beaucoup plus médiocres pour des valeurs élevées de Y. On peut aussi avoir 1 « entonnoir » dans le sens inverse (ie qui se réduit avec les valeurs de X). Dans ce cas, l'estimation est très bonne pour les valeurs élevées de Y mais mauvaise pour des faibles valeurs.

Cependant pour détecter rigoureusement l'hétéroscédasticité, il est nécessaire d'avoir recours à des tests explicites. Plusieurs tests sont, à cet effet, utilisés : le test de Goldfeld-Quandt, le test de Breush-Pagan et le test de White, etc.

4.2.1. Le test Goldfeld-Quandt

Le test Goldfeld-Quandt repose sur l'hypothèse que la variance des perturbations est une fonction monotone d'une ou des variables explicatives X. L'idée du test est de comparer les variances des perturbations sur deux sous-échantillons de tailles n_1 et n_2 telles que n_1 correspond aux premières observations et n_2 les dernières. On choisit alors les n_1 et n_2 observations de manière à séparer suffisamment les deux sous-échantillons dans le but de s'assurer que la variance puisse significativement être différente entre les deux sous-échantillons.

Pour présenter ce test, on suppose d'abord que X est constitué d'une seule variable suivant le modèle :

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

Pour mettre en œuvre le test, on ordonne d'abord, de façon croissante, les observations en fonction de x_i de telle sorte que x_i soit inférieur x_{i+1}. Ensuite, on exclue m observations centrales telles que :

$$n_1 = n_2 = \frac{n-m}{2}$$

Où n est le nombre d'observation total, m est le nombre centrales à exclure afin d'obtenir n_1 et n_2.

Ainsi, si σ_1^2 est la variance des perturbations sur le premier sous échantillons et que σ_2^2 est la variance sur le second sous échantillon, on peut alors formuler l'hypothèse de test de l'hétéroscédasticité comme suit :

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

L'hypothèse nulle H_0 traduit l'homoscédastcité alors que l'hypothèse H_1 traduit l'hétérocédastcité.

Le test de Goldfeld-Quandt est fondé sur la statistique suivante :

$$GQ = \frac{\sigma_2^2}{\sigma_1^2} = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{\frac{\sum_{i=1}^{n_2}(y_i - \hat{y}_i)^2}{n_2 - 2}}{\frac{\sum_{j=1}^{n_1}(y_j - \hat{y}_j)^2}{n_1 - 2}} = \frac{SCR_2}{SCR_1}$$

$$GQ = \frac{SCR_2}{SCR_1} \tag{4.4}$$

Où SCR_2 est la somme des carrés des résidus issue de l'estimation par MCO du modèle sur le second sous-échantillon et SCR_1 la somme des carrés des résidus obtenue sur le premier sous-échantillon

La statistique GQ est le rapport entre deux variables aléatoires suivant une loi de khi-deux divisées par leur degrés de libertés respectifs ($n_2 - 2$ et $n_1 - 2$), alors la statistique GQ suit une loi de Fisher. On a alors :

$$GQ = \frac{SCR_2}{SCR_1} \rightsquigarrow F(n_2 - 2\,, n_1 - 2\,)$$

Lorsque GQ est supérieure à $F(n_2 - 2\,, n_1 - 2\,)$, on rejette H_0, signifiant alors l'existence de l'hétéroscédasticité dans les données.

L'un des principaux avantages du test de Goldfeld-Quandt est la facilité de sa mise en œuvre. Cependant, son principal inconvénient est la difficulté du choix de variables pour trier les observations lorsque l'on dispose de plusieurs variables explicatives. Par ailleurs, ce test n'est pas très adapté aux petits échantillons du simple fait qu'il nécessite toujours un découpage de l'échantillon initial.

4.2.2. Le test Breush-Pagan

Le test Breush-Pagan est un prolongement (ou une généralisation) du test Goldfeld-Quandt qui propose une forme fonctionnelle de hétéroscédasticité. En d'autres termes, le test Breush-Pagan consiste à expliquer l'hétéroscédasticité en fonctions des valeurs des variables explicatives. Cette équation se présente comme suit :

$$\sigma_i^2 = \sigma_0^2 + \alpha_1 x_i \qquad (4.5)$$

L'hypothèse nulle du test Breush-Pagan se formule alors comme suit :

$$\begin{cases} H_0 : \alpha_1 = 0 \;\; \Rightarrow \sigma_i^2 = \sigma_0^2 \\ H_1 : \alpha_1 \neq 0 \Rightarrow \sigma_i^2 = \; \sigma_0^2 + \alpha_1 x_i \end{cases}$$

Le test Breush-Pagan se résume alors à un test sur la nullité du coefficient associé à la variable x dans l'explication de la variance des résidus. Et pour un ensemble de variable X, le principe est de tester la nullité jointe des coefficients.

La mise en œuvre du test Breush-Pagan comporte trois étapes.

1- Dans un premier temps, on approxime les σ_i^2 par les $\hat{\varepsilon}_i^2$ obtenus de l'estimation par MCO du modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
2- Dans un second temps, on estime par MCO le modèle suivant :
$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_i + u_i$$
3- Dans un troisième temps, on calcule la statistique de test BP telle que :

$$BP = nR^2 \quad (4.6)$$

Où R^2 est le R^2 de l'estimation de l'étape 2. Sous H_0 R^2 doit être nul. Mais d'une manière générale, la statistique BP suit asymptotiquement une loi de Khi-deux à k-1 degrés de liberté où k est le nombre de variables explicatives (y compris la constante). n représente la taille de l'échantillon.

Lorsque la statistique BP est supérieure au khi-deux tabulé, on rejette H_0 retenant ainsi la présence d'hétéroscédasticité.

4.2.3. Le test de White

Le test d'hétéroscédasticité de White est une généralisation de la forme fonctionnelle de l'hétéroscédasticité de Breusch-Pagan en y introduisant des formes quadratiques. En supposant par exemple deux variables x_{1i} et x_{2i}, l'équation d'hétéroscédasticité de White peut être présentée comme suit :

$$\sigma_i^2 = \sigma_0^2 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{1i}^2 + \gamma_4 x_{2i}^2 + \gamma_5 (x_{2i} * x_{2i}) + u_i \quad (4.7)$$

L'hypothèse nulle du test peut alors se formuler comme suit :

$$\begin{cases} H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = 0 \\ H_1 : \gamma_1 \neq 0 \; ou \; \gamma_2 \neq 0 \; ou \; \gamma_3 \neq 0 \; ou \; \gamma_4 \neq 0 \; ou \; \gamma_5 \neq 0 \end{cases}$$

Ainsi, tout comme le test Breush-Pagan, le test de White se résume à un test sur la nullité des coefficients du modèle d'explication de la variance des résidus.

Le test se met en œuvre en trois étapes :

1- Dans un premier temps, on approxime les σ_i^2 par les $\hat{\varepsilon}_i^2$ obtenus de l'estimation par MCO du modèle $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$.
2- Dans un second temps, on estime par MCO le modèle suivant :
$$\hat{\varepsilon}_i^2 = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{1i}^2 + \gamma_4 x_{2i}^2 + \gamma_5 (x_{2i} * x_{2i}) + u_i$$
3- Dans un troisième temps, on calcule la statistique de test W telle que :
$$W = nR^2 \quad (4.8)$$

Où R^2 est le R^2 de l'estimation de l'étape 2. Sous H_0 R^2 doit être nul. Et d'une manière générale, la statistique W suit une loi de Khi-deux à $\frac{k(k+1)}{2} - 1$ degrés de liberté où k est le nombre de variables explicatives (y compris la constante). n représente la taille de l'échantillon.

Lorsque la statistique W est supérieure au khi-deux tabulé, on rejette H_0 retenant ainsi la présence d'hétéroscédasticité.

NB : lorsque la taille de l'échantillon est faible, on peut utiliser la seconde version du test de White qui se présente sous la forme d'un test de Fisher. La statistique de ce test est la suivante :

$$W = \frac{\left(SCR_{H_0} - SCR_{H_1}\right)}{SCR_{H_1}} \rightsquigarrow F(N-1\,,N-k\,) \tag{4.9}$$

Où SCR_{H_0} est la somme des carrés des résidus sous H_0 c'est-à-dire la somme des carrés des résidus obtenu par régression de $\hat{\varepsilon}_i^2$ uniquement sur la constante γ_0. SCR_{H_1} est la SCR obtenue en estimant par MCO l'équation complète. k est le nombre de variables explicatives (y compris la constante).

4.2.4. Correction de l'hétéroscédasticité

4.2.4.1. Cas où la matrice de variance-covariance est connue

Dans le cas où la variance est connue, si les unités statistiques sont par exemple, des entreprises, la variance peut être liée à un effet de taille notée z (par ex. le nombre de travailleurs). Dans ce cas, la matrice de variance-covariance s'écrit comme suit:

$$\Omega = \begin{pmatrix} \sigma_{\varepsilon_1}^2 & 0 & \dots & \dots & 0 \\ 0 & \sigma_{\varepsilon_2}^2 & \dots & \dots & \dots \\ \dots & \dots & \sigma_{\varepsilon_i}^2 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \sigma_{\varepsilon_n}^2 \end{pmatrix} = \sigma_0^2 \begin{pmatrix} z_1 & 0 & \dots & \dots & 0 \\ 0 & z_2 & \dots & \dots & \dots \\ \dots & \dots & z_i & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & z_n \end{pmatrix} = \sigma_0^2 Z$$

$$\text{Où } \; Z = Diag(z_1, z_2, \dots, z_n)$$

Les valeurs z_i sont strictement positives.

Ainsi connaissant la matrice de variance-covariance, on peut corriger l'hétéroscédasticité en utilisant les moindres carrées généralisé (MCG) comme suit.

Soit le modèle suivant :

$$Y = X\beta + \epsilon$$

Si nous multiplions cette équation par l'inverse de la racine carrée de la matrice de variance-covariance, on obtient :

$$\Omega^{-\frac{1}{2}} Y = \Omega^{-\frac{1}{2}} X\beta + \Omega^{-\frac{1}{2}} \epsilon$$

Ce qui peut se réécrire comme suit :

$$Y^* = X^*\beta + \epsilon^*$$

Avec $Y^* = \Omega^{-\frac{1}{2}} Y$; $X^* = \Omega^{-\frac{1}{2}} X$; $\epsilon^* = \Omega^{-\frac{1}{2}} \epsilon$

En appliquant les MCO sur cette équation transformée, on obtient l'estimateur des Moindres Carrées Généralisés qui se présente comme suit.

$$\hat{\beta}_{MCG} = \left(X'\Omega^{-1}X\right)^{-1}\left(X'\Omega^{-1}Y\right) \tag{4.10}$$

Ce qui permet ainsi de corriger l'hétéroscédasticité.

Par ailleurs, on peut aussi montrer que cet estimateur est aussi équivalent à $\hat{\beta}_{MCG} = (X'Z^{-1}X)^{-1}(X'Z^{-1}Y)$ car en remplaçant Ω par $\sigma_0^2 Z$, on a :

$$\hat{\beta}_{MCG} = (X'(\sigma_0^2 Z)^{-1}X)^{-1}(X'(\sigma_0^2 Z)^{-1}Y) = \frac{\sigma_0^2}{\sigma_0^2}(X'(Z)^{-1}X)^{-1}(X'(Z)^{-1}Y)$$

$$\hat{\beta}_{MCG} = (X'Z^{-1}X)^{-1}(X'Z^{-1}Y) \tag{4.11}$$

Cette expression montre alors que l'estimateur des moindres carrés généralisés (MCG) peut être construit sans qu'il soit nécessaire de se poser la question sur la valeur de σ_0^2.

Par ailleurs, il est possible de corriger le problème de l'hétéroscédasticité en suivant les étapes décrites ci-dessous.

D'abord, construire la matrice H telle que :

$$H = \begin{pmatrix} 1/\sqrt{z_1} & 0 & \dots & \dots & 0 \\ 0 & 1/\sqrt{z_2} & \dots & \dots & \dots \\ & & 1/\sqrt{z_i} & & \\ \dots & \dots & & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 1/\sqrt{z_n} \end{pmatrix}$$

A noter simplement que :

$$M'M = Z^{-1} = \sigma_0^2 \Omega^{-1}$$

Ensuite, on multiplie tous les membres de l'équation $Y = X\beta + \epsilon$ par la matrice H, on trouve :

$$MY = MX\beta + M\epsilon$$

On peut simplement faire remarquer que $E(M\epsilon) = 0$ et que

$$VAR(M\epsilon) = M\Omega M = M\sigma_0^2 ZM = \sigma_0^2 I = \sigma_\varepsilon^2 = cste$$

Cette relation montre donc qu'en multipliant l'équation initiale par la matrice M, on retrouve l'homoscédasticité. L'application des MCO sur le modèle transformé se présente comme suit :

$$\hat{\beta}_{MCP} = (X'MMX)^{-1}(X'MMY)$$

En résumé, pour corriger l'hétéroscédasticité, il suffit de multiplier toute les variables par l'inverse de la racine carrée de la variance des erreurs ($\frac{1}{\sqrt{\sigma_i^2}}$). On dit alors qu'on a "sphéricisé " le modèle. On est donc ramené au modèle linéaire classique sur lequel on peut appliquer les moindres carrés ordinaires. Dans le cas d'un modèle linéaire multiple, on multiplie tous les vecteurs par l'inverse de la racine carrée de la matrice de variance-covariance des perturbations.

L'estimateur obtenu en appliquant les moindres carrés ordinaires sur les variables transformées est appelé estimateur *Moindres Carrés Généralisés* (MCG). Cet estimateur est sans biais et optimal parmi les estimateurs sans bais linéaires.

4.2.4.2. Cas où la matrice de variance-covariance est inconnue

Dans la plupart des cas, la matrice de variance-covariance Ω n'est pas connue. De plus, il n'existe souvent aucune hypothèse sur la forme de l'hétéroscédasticité permettant d'estimer la matrice de variance-covariance. Cependant, même en l'absence de telle hypothèse des estimateurs ont été proposés dans le but de pouvoir estimer la matrice de variance-covariance. L'estimateur le plus courant est celui de White. Dans cet estimateur, la matrice de variance-covariance se présente comme suit :

$$\widehat{\Omega} = \begin{pmatrix} \hat{\varepsilon}_1^2 & 0 & \dots & \dots & 0 \\ 0 & \hat{\varepsilon}_2^2 & \dots & \dots & \dots \\ \dots & \dots & \hat{\varepsilon}_i^2 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \hat{\varepsilon}_n^2 \end{pmatrix}$$

Où les $\hat{\varepsilon}_i$ représentent les résidus de l'estimation du modèle initiale.

Bien que $\hat{\varepsilon}_i^2$ soient des estimateurs biaisés des σ_i^2, ils restent des estimateurs convergents. Pour corriger l'hétéroscédasticité dans cette configuration, White propose l'estimateur suivant :

$$\hat{\beta}_{MCQG} = \left(X'\widehat{\Omega}^{-1}X\right)^{-1}\left(X'\widehat{\Omega}^{-1}Y\right) \qquad (4.12)$$

En résumé, lorsque les variances σ_i^2 (ou la matrice de variance-covariance) ne sont pas connue, il faut proposer une approximation de ces variances en se servant des résidus au carré. Ces résidus seront ensuite utilisés pour transformer les variables auquel on applique les MCO. L'estimateur des moindres carrés est alors appelée *Moindres Carrés Quasi-Généralisés*(MCQG). Il faut noter que cet estimateur n'est pas nécessairement optimal.

En résumé, lorsque la matrice de variance-covariance n'est pas connue, la démarche de correction de l'hétéroscédasticité se présente comme suit :

- Estimer le modèle initial par MCO et récupérer les résidus $\hat{\varepsilon}_i$
- Approximer les variances individuelles σ_i^2 par les carrés $\hat{\varepsilon}_i^2$
- Estimer l'équation traduisant la forme fonctionnelle de l'hétéroscédasticité (formule de White)
- Utiliser les paramètres de cette estimation pour calculer les valeurs prédites de σ_i^2 pour obtenir $\hat{\sigma}_i^2$
- Multiplier les variables par l'inverse $(\frac{1}{\sqrt{\hat{\sigma}_i^2}})$ afin d'obtenir les variables transformées
- Appliquer les MCO sur ce modèle transformé pour obtenir les paramètres.

4.3. Test d'autocorrélation des erreurs

La présence d'autocorrélation des erreurs correspond à la violation de l'hypothèse $COV(\varepsilon_i, \varepsilon_j) = 0$. En effet on a $COV(\varepsilon_i, \varepsilon_j) = \rho\, \sigma_{\varepsilon_i}^2 \neq 0$ où $\sigma_{\varepsilon_i}^2$ est la variance des erreurs et ρ un paramètre compris entre -1 et 1. On est donc en présence d'autocorrélation.

A noter, toutefois, qu'en présence d'autocorrélation, l'estimateur MCO reste toujours sans biais mais sa variance n'est plus minimale. L'autocorrélation des erreurs est, généralement, fréquente dans les séries temporelles ou dans les données spatiales. Pour illustrer le problème d'autocorrélation, partons par exemple d'un modèle de série temporelle dont l'équation est la suivante :

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \,; t = 1, 2, \dots, T$$

En supposant qu'il y a une autocorrélation d'ordre 1, la série des ε_t peut s'écrire comme un processus autorégressif d'ordre 1 dit AR(1). Dans ce cas, on a l'équation suivante :

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \qquad (4.13)$$

Où u_t est un bruit blanc, indépendant et identiquement distribué, de moyenne 0 et de variance $\sigma_{u_t}^2$

Pour mettre en évidence l'autocorrélation des erreurs, il faut alors tester la nullité du coefficient ρ dans ce processus[3]. Plusieurs tests sont alors disponibles notamment : le test de Durbin-Watson ou le test de Box-Pierce, etc.

[3] Il faut noter que l'autocorrélation peut aussi être d'ordre supérieur.

4.3.1. Le test d'autocorrélation de Durbin-Watson

L'hypothèse nulle du test de Durbin-Watson est la suivante :

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

La statistique de test de Durbin-Watson s'écrit comme suit :

$$DW = 2(1 - \hat{\rho}) = \frac{\sum_{t=2}^{T}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{T}(\hat{\varepsilon}_t)^2} \tag{4.14}$$

Il faut noter que la statistique DW est comprise entre 0 et 4. Si $\hat{\rho} = -1$ (autocorrélation négative), alors $DW = 4$ et si $\hat{\rho} = 1$ (autocorrélation positive), alors $DW = 0$. Sous H_0, $\hat{\rho} = 0$, donc $DW = 2$. Ainsi, lorsque la statistique DW est proche de 2, cela signifie une absence d'autocorrélation des erreurs. Dans ce cas, on ne peut pas rejeter l'hypothèse nulle.

Cependant, pour une interprétation plus précise du test de Durbin-Watson, on se réfère à une table qui donne les valeurs critiques (d1 et d2). Les valeurs d1 et d2 sont fournies dans la table de DW et présentées en fonction à la fois du nombre d'observations et du nombre de variables explicatives. En fonction du seuil d'erreur retenu (alpha), on lit dans la table les deux valeurs d1 et d2 avec lesquelles on construit la table de décision suivante :

0	d1	d2	4-d2	4-d1	4
Autocorrélation positive	?	Pas d'autocorrélation	?	Autocorrélation négative	

Les interprétation sont les suivantes :

$DW \in [0\ ;\ d1]$: Autocorrélation positive
$DW \in]d1\ ;\ d2[$: On ne peut pas conclure
Si $DW \in [d2\ ;\ 4 - d2]$: Pas d'autocorrélation
Si $DW \in]4 - d2\ ;\ 4 - d1[$: On ne peut pas conclure
Si $DW \in [4 - d1\ ;\ 4]$: Autocorrélation positive.

Par exemple, supposons que nous ayons estimé un modèle linéaire simple (une seule variable explicative) sur 17 observations. Supposons qu'à la suite de cette estimation nous avons calculé le DW qui est égal à 1,05. Pour tester la présence d'autocorrélation au seuil de 5%, nous suivons les étapes suivantes :

D'abord, dans la table de Durbin-Watson, les valeurs d1 et d2 qui correspondent à un modèle avec une variable explicative k=1 (constante incluse dans la régression) et le nombre d'observation n=17 sont respectivement 1,13 et 1,38.

Dans cette configuration, on remarque que DW est comprise entre 0 et 1,13. On rejette alors H0 et conclut à une autocorrélation positive.

NB : Pour pouvoir utiliser la table de DW, plusieurs conditions doivent être satisfaites. D'abord, le nombre d'observations doit être supérieur à 15 car il n'y pas de valeurs tabulées pour les tailles d'échantillon en dessous de 15. Ensuite, le modèle estimé doit inclure une constante car la table originale de DW est construite à partir des modèles incluant une constante. Toutefois, il existe une autre version de la table conçue pour les modèles sans constante. Enfin, la variable dépendante retardée ne doit pas figurer parmi les régresseurs. Et si c'est le cas il faudrait utiliser la statistique h de Durbin. En effet, lorsque le modèle est de la forme $y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \varepsilon_t$ avec $\varepsilon_t = \rho\varepsilon_{t-1} + u_t \; ; t = 1, 2, \dots, T$. Dans cette configuration, la statistique de Durbin-Watson n'est pas directement utilisable. Il faut alors apporter des corrections en utilisant la statistique de Durbin qui se présente comme suit :

$$h = \left(1 - \frac{DW}{2}\right)\sqrt{\frac{T}{\left(1 - T\,\hat{\sigma}^2_{\hat{\beta}_2}\right)}} \tag{4.15}$$

Où DW est la statistique ordinaire de Durbin-Watson, $\hat{\sigma}^2_{\hat{\beta}_2}$ est la variance estimée du coefficient $\hat{\beta}_2$, T est le nombre d'observations.

Sachant que $DW = 2(1 - \hat{\rho})$ où $\hat{\rho}$ est la valeur estimée du coefficient de corrélation, alors, on peut montrer que :

$$h = \hat{\rho}\sqrt{\frac{T}{\left(1 - T\,\hat{\sigma}^2_{\hat{\beta}_2}\right)}} \tag{4.16}$$

Les hypothèses du test sont équivalentes à celles de Durbin-Watson. En revanche la statistique h suit une loi normale (0, 1). Ainsi, l'hypothèse du test se présente comme suit :

$$\begin{cases} H_0 : h = 0 \Leftrightarrow \rho = 0 \\ H_1 : h \neq 0 \Leftrightarrow \rho \neq 0 \end{cases}$$

En revanche, pour la règle de décision, on compare la valeur h calculée à la valeur $Z_{1-\frac{\alpha}{2}}$ de la table de loi normale centrée réduite au seuil d'erreur α. Ainsi, lorsque $|h| > Z_{1-\frac{\alpha}{2}}$, on rejette l'hypothèse nulle d'absence d'autocorrélation.

4.3.2. Le test d'autocorrélation de Box-Pierce

La statistique du test proposé par Box-Pierce est la suivante :

$$Q_{BP} = T\sum_{j=1}^{p} {r_j}^2 \quad avec$$

$$r_j = \frac{\sum_{t=j+1}^{T}\left(\hat{\varepsilon}_t\hat{\varepsilon}_{t-j}\right)}{\sum_{t=1}^{T}(\hat{\varepsilon}_t)^2}$$

Cette statistique suit une loi de khi-deux à p degrés de liberté où p est le nombre de retards considéré.

$$Q_{BP} = T\sum_{j=1}^{p} {r_j}^2 \rightsquigarrow \chi^2(p) \tag{4.17}$$

Lorsque la statistique calculée est supérieure à la statistique de khi-deux lue dans la table, on rejette l'hypothèse nulle d'absence d'autocorrélation.

4.3.3. Le test d'autocorrélation de Ljung-Box

Le test de Ljung-Box est une version améliorée du test de Box-Pierce. La statistique de test se présente comme suit :

$$Q_{LB} = T(T+2)\sum_{j=1}^{p}\frac{{r_j}^2}{T-j} \rightsquigarrow \chi^2(p) \tag{4.18}$$

Lorsque la statistique calculée est supérieure à la statistique de khi-deux lue dans la table, on rejette l'hypothèse nulle d'absence d'autocorrélation.

Remarque : Les tests de Box-Pierce et de Ljung-Box sont plus généraux que celui de Durbin-Watson, car ils permettent de détecter des autocorrélations d'ordre supérieur. Ils sont qualifiés de test portemanteau car on peut augmenter le nombre de retard, au fur et à mesure, lorsqu'on cherche des autocorrélations d'ordre supérieur.

4.3.4. Correction de l'autocorrélation

L'estimateur MCO en présence d'autocorrélation est sans biais mais sa variance n'est pas minimale. Il faut donc pouvoir apporter des corrections. Dans le cas d'un modèle autorégressif d'ordre 1 tel que :

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \,; t = 1, 2, \dots, T$$

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

$E(u_t) = 0$; $E({u_t}^2) = V(u_t) = \sigma_u^2$; $E(u_t\varepsilon_{t-1}) = 0$; $E(u_t u_{t'}) = 0 \ \forall\, t \neq t'$

Pour un nombre de retard p, on peut montrer que ε_t s'écrit comme suit :

$$\varepsilon_t = \rho^p \varepsilon_{t-p} + \sum_{j=1}^{p} \rho^j u_{t-j}$$

Mais, on peut négliger le premier membre puisque :

$$\rho^p \to 0 \; si \; p \to \infty$$

Ainsi, on a :

$$\varepsilon_t = \sum_{j=1}^{p} \rho^j u_{t-j}$$

On en déduit alors que ε_t est une fonction linéaire des u_t.

En utilisant cette expression, calculons l'espérance, la variance et la covariance des erreurs. On a :

Espérance :

$$E(\varepsilon_t) = E\left(\sum_{j=1}^{p} \rho^j u_{t-j}\right) = \sum_{j=1}^{p} \rho^j E(u_{t-j}) = 0$$

$$E(\varepsilon_t) = 0$$

Variance

$$V(\varepsilon_t) = \sigma_\varepsilon^2 = V\left(\sum_{j=1}^{p} \rho^j u_{t-j}\right) = V\left(\rho u_{t-1} + \rho^2 u_{t-2} + \cdots + \rho^p u_{t-p}\right)$$

$= V(\rho u_{t-1}) + V(\rho^2 u_{t-2}) + \cdots + V(\rho^p u_{t-p})$ car $Cov\,(u_t; u_{t'}) = E(u_t u_{t'}) = 0 \ \ \forall\, t \neq t'$

Ainsi :

$$V(\varepsilon_t) = \rho E[(u_{t-1})^2] + \rho^2 E[(u_{t-2})^2] + \cdots + \rho^p E\left[\left(u_{t-p}\right)^2\right]$$

$$\text{Or} \ \ E[(u_{t-1})^2] = E[(u_{t-2})^2] = \cdots = E\left[\left(u_{t-p}\right)^2\right] = \sigma_u^2$$

Donc $$V(\varepsilon_t) = (\rho + \rho^2 + \cdots + \rho^p)\, \sigma_u^2$$

Or $\rho + \rho^2 + \cdots + \rho^p = \frac{1-\rho^p}{1-\rho}$ (somme p premier termes d'une suite géométrique de raison ρ de premier terme 1).

$$V(\varepsilon_t) = \left(\frac{1-\rho^p}{1-\rho}\right) \sigma_u^2$$

Mais comme $\rho^p \to 0 \; si \; p \to \infty$ alors : $V(\varepsilon_t) = \sigma_\varepsilon^2 = \frac{\sigma_u^2}{(1-\rho)}$

Covariance

$$Cov\left(\varepsilon_t; \varepsilon_{t-p}\right) = E\left(\varepsilon_t \varepsilon_{t-p}\right) = \rho^p V\left(\varepsilon_{t-p}\right) = \rho^p \, \sigma_\varepsilon^2$$

On peut donc en déduire le coefficient de corrélation ρ comme suit :

$$\rho = \frac{Cov(\varepsilon_t; \varepsilon_{t-p})}{\sqrt{V(\varepsilon_t)}\sqrt{V(\varepsilon_{t-p})}}$$

Connaissant alors ρ, on peut alors proposer une correction de l'autocorrélation en transformant le modèle autorégressif initial comme suit :

$$\begin{cases} y_t^* = y_t - \rho y_{t-1} \\ x_t^* = x_t - \rho x_{t-1} \\ \beta_0^* = \beta_0(1-\rho) \end{cases}$$

Avec cette transformation, l'équation à estimer se présente alors comme suit :

$$y_t^* = \beta_0^* + \beta_1^* x_t^* + \varepsilon_t^*$$

Ce modèle peut donc être estimé par MCO afin d'obtenir les paramètres β_0^* et β_1^*.

Cependant, ρ n'étant pas connu, il faut alors utiliser une approximation de cette valeur afin de poursuivre la correction. Plusieurs méthodes ont été proposées dont les principales sont : la méthode Cochranne-Orcutt, la méthode de Durbin mais aussi d'autres méthodes comme celle de Hildreth-Lu.

4.3.4.1. La méthode de correction de Cochranne-Orcutt

La méthode de correction de l'autocorrélation proposée par Cochranne-Orcutt est une méthode qui se réalise en quatre étapes de manière itérative.

En partant de l'équation $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$; où $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, les étapes sont les suivantes :

1- Dans la première étape, on estime l'équation par MCO pour obtenir les paramètres $\hat{\beta}_0$ et $\hat{\beta}_1$.
2- Dans la seconde étape, on utilise les résidus de cette première estimation pour calculer le coefficient de corrélation $\hat{\rho}_1$ en utilisant la formule de Box-Pierce :

$$\hat{\rho}_1 = \frac{\sum_{t=2}^{T}(\hat{\varepsilon}_t\hat{\varepsilon}_{t-1})}{\sum_{t=1}^{T}(\hat{\varepsilon}_t)^2}$$

3- Dans la troisième étape, on utilise ce coefficient pour calculer les variables transformées telles que :

$$\begin{cases} y_t^* = y_t - \hat{\rho}_1 y_{t-1} \\ x_t^* = x_t - \hat{\rho}_1 x_{t-1} \end{cases}$$

4- Dans la quatrième étape, on estime par MCO le modèle transformé:

$$y_{t,1}^* = \beta_{0,1}^* + \beta_{1,1}^* x_{t,1}^* + \varepsilon_{t,1}^*$$

En utilisant les informations de cette quatrième étape, on revient à la première étape. En effet, après avoir estimé le modèle transformé, on récupère les résidus et on calcule de nouveau le coefficient de corrélation $\hat{\rho}_2$ (reprise étape 1 et 2). On transforme de nouveau les variables (déjà transformées dans la première itération), on forme la nouvelle équation transformée, qui est ensuite estimée par MCO. Cette équation se présente alors comme suit :

$$y_{t,2}^* = \beta_{0,2}^* + \beta_{1,2}^* x_{t,2}^* + \varepsilon_{t,2}^*$$

Et les paramètres estimés sont : $\hat{\beta}_{0,2}^*$ et $\hat{\beta}_{1,2}^*$. A la i-ième itération, l'équation transformée se présente comme suit :

$$y_{t,i}^* = \beta_{0,i}^* + \beta_{1,i}^* x_{t,i}^* + \varepsilon_{t,i}^*$$

Et les paramètres estimés sont : $\hat{\beta}_{0,i}^*$ et $\hat{\beta}_{1,i}^*$

Ce processus itératif continue jusqu'à ce que l'écart entre deux valeurs consécutives $\hat{\rho}_{i+1}$ et $\hat{\rho}_i$ soit significativement négligeable. Cet écart est bien sûr évalué de manière arbitraire. Mais il doit être apprécié par rapport à l'écart initial entre les deux premières valeurs de $\hat{\rho}$ ($\hat{\rho}_1$ et $\hat{\rho}_2$).

4.3.4.2. La méthode de correction de Durbin

La méthode de Durbin est une méthode correction qui se réalise en deux étapes. Dans la première étape, on estime l'équation suivante :

$$y_t = \beta_0 + \rho y_{t-1} + \beta_1 x_t - \beta_2 x_{t-1} + v_t$$

Avec la contrainte que $\beta_2 = \rho\beta_1$. L'estimation de cette équation permet donc d'obtenir $\hat{\rho}$.

Dans la seconde étape, on utilise la valeur estimée de ρ pour calculer les variables transformées :

$$\begin{cases} y_t^* = y_t - \hat{\rho}_1 y_{t-1} \\ x_t^* = x_t - \hat{\rho}_1 x_{t-1} \end{cases}$$

Ces variables transformées permettent ensuite de formuler l'équation :

$$y_t^* = \beta_0^* + \beta_1^* x_t^* + \varepsilon_t^*$$

Cette équation est alors estimée par MCO pour déduire les paramètres.

4.3.4.3. Autres méthodes de correction de l'autocorrélation

Il existe diverses autres méthodes de correction. Nous évoquons par exemple la méthode de Hildreth-Lu. La méthode de Hildreth-Lu consiste à choisir plusieurs de ρ entre -1 et 1, pour pouvoir ensuite estimer un grand nombre de fois l'équation transformée $y_t^* = \beta_0^* + \beta_1^* x_t^* + \varepsilon_t^*$. La méthode consiste alors à retenir l'estimation qui fournit la plus faible somme des carrés des résidus. Par ailleurs, il faut aussi signaler que l'autocorrélation ne se présente pas sous la forme d'un processus AR, les méthodes précédemment discutées ne sont plus valables. Il faut alors utiliser d'autres méthodes. Par exemple, lorsque l'autocorrélation se présente sous la forme d'un moyenne mobile telle que :

$$y_t = \beta_0 + \rho y_{t-1} + \beta_1 x_t - \beta_2 x_{t-1} + \varepsilon_t$$

$$\varepsilon_t = u_t + \rho u_{t-1}$$

Pour éliminer cette forme d'autocorrélation, il faut adopter la méthode d'estimation par maximum de vraisemblance.

4.4. Autres cas de violation des hypothèses de base du modèle linéaire

En plus de la violation des hypothèses de normalité, d'homocédasticité, d'absence d'autocorrélation, il existe d'autres cas de violation qui rendent invalide le modèle linéaire de base.

L'un de ces cas de violation est la corrélation entre les variables explicatives et les erreurs ($COV(x_i, \varepsilon_i) \neq 0$). Ce problème est généralement connu sous le terme d'endogéneité. La variable x_{ik} est dite endogène lorsqu'elle est liée aux erreurs du modèle. Les variables explicatives non corrélées avec les erreurs sont dites variables exogènes. La correction de l'endogénéité nécessite généralement l'utilisation de la méthode de variable instrumentale (MVI). La MVI consiste à choisir des variables supplémentaires (appelées instruments) non corrélées avec les erreurs mais fortement corrélés avec la variable. Ces instruments doivent avoir un pouvoir explicatif suffisamment fort sur la variable endogène. Mais elles ne doivent avoir aucune corrélation avec la variable dépendante du modèle. En présence d'endogénéité, la méthode communément utilisée pour estimer le modèle est celle des Doubles Moindres Carrés (DMC) ou *Two-stage least squares (2SLS)*. La DMC consiste dans un premier temps à régresser par MCO la variable endogène sur ses instruments et sur les autres variables exogènes du modèle. Et dans un second temps, à calculer la valeur prédite de la variable endogène et introduire cette valeur prédite dans le modèle initiale pour ensuite estimer celui-ci par MCO. Il existe d'autres méthodes d'estimation telles que la méthode des moments généralisées, la méthode de maximum de vraisemblance à

information limitée, etc…Ces méthodes d'estimations n'ont pas été développées dans ce document.

En dehors du problème d'endogéneité, on peut aussi signaler d'autres problèmes tels que la multicolinéarité, la mauvaise spécification du modèle ou l'omission des variables explicatives pouvant chacun aboutir à biaiser les estimations du modèle linéaire. Il convient donc tester l'existence de ces problèmes et le cas échéant les corriger avant toute estimation.

CHAPITRE 5. MODELES A VARIABLE DEPENDANTE DICHOTOMIQUE

5.1. Présentation

Dans les précédents chapitres, les variables dépendantes étaient supposées être de nature continue. Cependant, dans de nombreuses estimations économétriques on s'intéresse à des variables de nature qualitative (exemples : choix entre deux moyens de transport, défaut de paiement d'un client, échec à un examen d'évaluation, etc…). Dans ce chapitre, nous nous intéressons essentiellement aux des modèles dont la variable dépendante est qualitative binaire (encore appelée variable dichotomique).

Une variable dichotomique est une variable qualitative traduisant la présence ou l'absence d'un évènement probabiliste. La présence de l'évènement est généralement codée 1 et l'absence de l'évènement codée par 0. Une variable dichotomique est donc une variable qui prend deux valeurs 0 et 1.

Dans un modèle à variable dépendante binaire, il s'agit d'étudier la probabilité d'observer 1. Par exemple, on souhaite analyser la probabilité de réussite à un examen d'un groupe d'étudiant. On considère un échantillon d'étudiants qui ont déjà passé l'examen. Et après la proclamation des résultats, on relève sur chacun des individus s'il a réussi ou pas. Pour cela, on construit une variable binaire qui prend 1 si l'individu a réussi à l'examen et 0 sinon. Soit y_i cette variable dépendante dichotomique. Ensuite, on réunit un ensemble de variables explicatives (telles que le nombre le nombre d'heures de révision, etc..) afin d'élaborer le modèle.

Dans tout ce qui suit, y_i sera codée de la manière suivante :

$$y_i = \begin{cases} 1 \; si \; l'individu \; présente \; le \; phénomène \; étudié \\ 0 \; sinon \end{cases} \qquad (5.1)$$

Pour étudier la probabilité pour que $y_i = 1$, on dispose d'un ensemble de k variables explicatives $x_{i1},\ x_{i2}, \dots, x_{ik}$ que l'on peut rassembler dans un vecteur X_i. Le modèle de probabilité se présente alors comme suit :

$$P(y_i = 1/X_i) = F(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) = F(X_i\beta) \qquad (5.2)$$

Où $P(y_i = 1/X_i)$ représente la probabilité que y_i soit égal à 1 conditionnellement aux caractéristiques $x_{i1},\ x_{i2}, \dots, x_{ik}$. β est un vecteur constitués de $k+1$ paramètres : β_0, β_1, β_2, …β_k. $F(.)$ est la fonction de répartition de la quantité $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ ($X_i\beta$ sous la forme matricielle). Les propriétés de la

fonction $F(.)$ sont telles que pour tout variable z, $\lim_{Z\to-\infty} F(z) = 0$ et $\lim_{Z\to+\infty} F(z) = 1$. $F(.)$ est donc une fonction continue positive et comprise entre 0 et 1.

En se basant sur ces précédentes propriétés, le modèle à choix discret s'écrit comme suit :

$$\begin{cases} P(y_i = 1) = F(X_i\beta) & (5.3a) \\ P(y_i = 0) = 1 - F(X_i\beta) & (5.3b) \end{cases}$$

D'une manière générale, le modèle s'écrit comme suit :

$$y_i = F(X_i\beta) + \varepsilon_i \quad (5.4)$$

Où ε_i est le terme d'erreur.

A travers cette expression on peut faire une analogie avec le modèle linéaire. On constatera, en effet, que dans le modèle de probabilité y_i est expliquée non pas avec la valeur directe de $X_i\beta$ mais avec une transformation monotone croissante $F(.)$ qui représente la fonction de répartition.

S'agissant du terme d'erreur ε_i, il présente les propriétés suivantes :

$$E(\varepsilon_i) = 0$$

$$E(\varepsilon_i\varepsilon_j) = 0$$

Par ailleurs, comme y_i ne prend que deux valeurs 0 et 1, on peut étudier directement la distribution de ε_i. En effet :

Pour $y_i = 1$, $\varepsilon_i = 1 - F(X_i\beta)$ avec une probabilité égale à $F(X_i\beta)$

Pour $y_i = 0$, $\varepsilon_i = -F(X_i\beta)$ avec une probabilité égale à $1 - F(X_i\beta)$

Connaissant ces deux valeurs de ε_i, on peut calculer son esperance. En effet :

$$E(\varepsilon_i) = P(y_i = 0)(\varepsilon_i/y_i = 0) + P(y_i = 1)(\varepsilon_i/y_i = 1)$$

$$E(\varepsilon_i) = [1 - F(X_i\beta)][-F(X_i\beta)] + [F(X_i\beta)][1 - F(X_i\beta)]$$

$$E(\varepsilon_i) = 0$$

Ce qui montre alors que les erreurs sont, en moyenne, nulles dans le modèle dichotomique.

En revanche, l'une des hypothèses systématiquement violée dans le modèle dichotomique est celle de l'homocédasticité. En effet, l'homocédasticité suppose que la variance de ε_i est constante (soit $V(\varepsilon_i) = \sigma^2$). Cependant cette hypothèse ne peut pas être respectée dans le modèle dichotomique car les valeurs des ε_i

dépendent des variables explicatives X_i. Ce qui crée une hétéroécédasticité telle que $V(\varepsilon_i) = \sigma_i^2 \neq \sigma^2$. En effet :

$$V(\varepsilon_i) = E(\varepsilon_i^2) = P(y_i = 0)(\varepsilon_i^2/y_i = 0) + P(y_i = 1)(\varepsilon_i^2/y_i = 1)$$

$$V(\varepsilon_i) = [1 - F(X_i\beta)][-F(X_i\beta)]^2 + [F(X_i\beta)][1 - F(X_i\beta)]^2$$

$$V(\varepsilon_i) = [1 - F(X_i\beta)][F(X_i\beta)]$$

$$V(\varepsilon_i) = F(X_i\beta) - [F(X_i\beta)]^2 \tag{5.5}$$

Cette variance dépendant de la valeur de X_i, par conséquent, elle est hétéroscédastique.

Au final, les caractéristiques fondamentales d'un modèle à variable dépendante dichotomique sont les suivantes :

$$\begin{cases} y_i \in \{0\,,1\} \\ E(y_i) = F(X_i\beta) \\ E(\varepsilon_i) = 0 \\ V(\varepsilon_i) = [1 - F(X_i\beta)][F(X_i\beta)] \\ COV(\varepsilon_i, \varepsilon_i) = 0 \ \ \forall\, i \neq j \end{cases} \tag{5.6}$$

5.2. Choix de la fonction $F(.)$ et nature du modèle

La détermination du type de modèle d'estimation dépend du choix de la fonction de répartition F(.). Ce choix n'est a priori non contraint. Cependant, trois principaux modèles sont généralement utilisés : le modèle probit, le modèle logit et le modèle de probabilité linéaire.

Lorsque F(.) est la fonction de répartition de la loi normale, le modèle est dit probit. En revanche lorsque la fonction de répartition choisie est celle de la loi logistique, le modèle est dit logit. Enfin, lorsque la fonction choisie est une fonction identité de $X_i\beta$, le modèle est dit de probabilité linéaire (MPL). Dans ce dernier cas, les paramètres du modèle sont estimés comme dans le modèle linéaire par les moindres carrés ordinaires.

5.2.1. Le modèle probit

Le modèle probit correspond au choix de la fonction de répartition de la loi normale pour traduire la fonction F(.). On sait aussi que par définition la fonction de répartition est la sommation (ou l'intégrale) de la fonction de densité sur un certain espace. Pour la loi normale, cette fonction de répartition se présente comme suit :

$$F(z) = \int_{-\infty}^{z} \phi(t)\, dt = \Phi(z)$$

Où $\phi(t)$ est la fonction de densité de la loi normale centrée réduite et $\Phi(z)$ sa fonction de répartition. Sachant que $(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$, on peut alors écrire :

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \right) dt \qquad (5.7)$$

Les fonctions F(.) et $\Phi(z)$ sont définies de telle sorte que :

$$\lim_{Z \to -\infty} F(z) = \lim_{Z \to -\infty} \Phi(z) = 0 \text{ et } \lim_{Z \to +\infty} F(z) = \lim_{Z \to +\infty} \Phi(z) = 1.$$

Par ailleurs, $\phi(t) = F'(t) = f(t) = \frac{dF(z)}{dz}$.

Etant donné ces propriétés, le modèle se présente comme suit :

$$P(y_i = 1/X_i) = F(X_i\beta) = \Phi(X_i\beta) = \int_{-\infty}^{X_i\beta} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \right) dt \qquad (5.7)$$

5.2.2. Le modèle logit

Le modèle logit correspond au choix de la fonction logistique définie telle que pour une variable z, on a :

$$F(z) = \frac{e^z}{1 + e^z}$$

En divisant le numérateur et le dénominateur par e^z , on retrouve :

$$F(z) = \frac{1}{1 + e^{-z}}$$

La fonction de répartition de la loi logistique peut également se déduire de la fonction de densité telle que :

$$F(z) = \int_{-\infty}^{z} \phi(t)\, dt = \Phi(z)$$

Où $\phi(t)$ est la fonction de densité de la loi logistique définie comme suit :

$$\phi(t) = \frac{e^z}{1 + e^z} - \frac{e^{2z}}{(1 + e^z)^2}$$

Les fonctions F(.) et $\Phi(z)$ sont définies de telle sorte que $\lim_{Z \to -\infty} F(z) = \lim_{Z \to -\infty} \Phi(z) =$ 0 et $\lim_{Z \to +\infty} F(z) = \lim_{Z \to +\infty} \Phi(z) = 1$. Par ailleurs, $\phi(t) = F'(t) = f(t) = \frac{dF(z)}{dz}$

Etant donné ces propriétés, le modèle se présente comme suit :

$$P(y_i = 1/X_i) = F(X_i\beta) = \frac{1}{1 + e^{-X_i\beta}} \quad (5.8)$$

5.3. Définition du modèle dichotomique à partir d'une variable latente

Les modèles à variables dépendantes dichotomiques sont souvent définies à partir d'une variable latente, c'est-à-dire une variable reflétant un processus inobservable qui gouverne la réalisation de la variable dichotomique observée. Par exemple, considérons un étudiant très autonome et qui, pour préparer son examen de fin d'année, a l'habitude de réviser seul. Supposons maintenant qu'on vienne lui proposer d'adhérer à un club de révision composés d'étudiants de niveaux plus ou moins hétérogènes. Le choix de cet étudiant d'adhérer ou pas à ce groupe de révision est soumis à un processus latent. En effet, lorsque l'utilité procurée par l'adhésion au groupe de révision dépasse d'un certain seuil l'utilité d'une révision personnelle, alors l'étudiant choisira d'adhérer au groupe de révision. Dans ce cas, la variable dichotomique observée prend 1. En revanche lorsque l'utilité procurée par l'adhésion au groupe de révision est en dessous de l'utilité d'une révision personnelle, l'étudiant n'adhèrera pas, il choisira une révision personnelle. Dans ce cas, la variable dichotomique d'adhésion prend 0.

Soit y_i^*, la différence entre l'utilité procurée par l'adhésion à une groupe de révision et celle procurée par une révision personne, le choix de l'étudiant peut être décrit par le processus latent suivant :

$$\begin{cases} y_i = 1 \; si \; y_i^* > 0 \\ y_i = 0 \; si \; y_i^* \leq 0 \end{cases} \quad (5.9)$$

Où z est un seuil inobservable.

La définition d'un modèle dichotomique à partir de ce processus consiste à supposer que la variable latente est régie par un modèle linéaire y_i^* telle que :

$$y_i^* = X_i\beta + u_i \quad (5.10)$$

Où y_i^* est la variable latente inobservée, X_i le vecteur des caractéristiques observables qui gouvernent ce processus latent et β le vecteur des paramètres. u_i représente ici le terme d'erreur du modèle latent. Ce terme d'erreur respecte les propriétés de base comme celles énoncées dans le modèle linéaire. En particulier, on a :

$$E(u_i) = 0$$

$$V(u_i) = \sigma_u^2$$

En se basant sur ces propriétés et en supposant que la série des résidus est distribuée selon une fonction de réparation F (.), on peut alors poser que :

$$P(y_i = 1) = P(y_i^* > 0) = P(X_i\beta + u_i > 0) = P(u_i > -X_i\beta)$$

$$= 1 - P(u_i \leq -X_i\beta) = 1 - F(-X_i\beta)$$

Si la loi de la distribution de u_i est symétrique, on a $F(-X_i\beta) = 1 - F(X_i\beta)$. Par conséquent

$$P(y_i = 1) = 1 - [1 - F(X_i\beta)] = F(X_i\beta)$$

Ce qui permet alors de retrouver :

$$P(y_i = 1) = F(X_i\beta)$$

De la même manière, on peut montrer que :

$$P(y_i = 0) = 1 - F(X_i\beta)$$

On retrouve donc exactement les mêmes formulations données précédemment. De plus, en supposant que la série u_i est distribuée selon une loi normale (resp. logistique) et qu'il est indépendant des variables explicatives, on obtient le modèle Probit (resp. Logit).

5.4. Estimation du modèle dichotomique

5.4.1. Méthode de maximum de vraisemblance (MV)

Soit le modèle dichotomique suivant :

$$y_i = F(X_i\beta) + \varepsilon_i$$

L'estimation du modèle dichotomique par la méthode de maximum de vraisemblance consiste à choisir le vecteur de paramètres β de façon à maximiser la vraisemblance de y_i.

La probabilité d'observer y_i pour un individu peut d'abord s'écrire comme suit :

$$P(y_i/X_i) = [P(y_i = 1/X_i)]^{y_i}[1 - P(y_i = 1/X_i)]^{1-y_i}$$

$$P(y_i/X_i) = [F(X_i\beta)]^{y_i}[1 - F(X_i\beta)]^{1-y_i}$$

La fonction de vraisemblance de y_i peut donc se présenter comme suit :

$$L(\beta) = \prod_{i=1}^{n}([F(X_i\beta)]^{y_i}[1 - F(X_i\beta)]^{1-y_i})$$

Maximiser la fonction de vraisemblance équivaut aussi à maximiser le logarithme de la fonction de la fonction de vraisemblance. C'est pourquoi, on préfère d'abord calculer la fonction $logL(\beta)$.

$$logL(\beta) = \sum_{i=1}^{n} \{y_i log[F(X_i\beta)] + (1 - y_i) log[1 - F(X_i\beta)]\}$$

Le maximum de cette fonction s'obtient en dérivant par rapport au vecteur de paramètres β. Ainsi, on obtient ainsi un vecteur de dérivés encore appelé vecteur gradient $G(\beta)$:

$$\frac{\partial logL(\beta)}{\partial \beta} = G(\beta) = \sum_{i=1}^{n} \left\{ \left(\frac{y_i}{F(X_i\beta)} \right) F'^{(X_i\beta)X_i} - \left(\frac{1 - y_i}{1 - F(X_i\beta)} \right) F'(X_i\beta)X_i \right\}$$

Or $F'(X_i\beta) = f(X_i\beta)$, ce qui implique que

$$\frac{\partial logL(\beta)}{\partial \beta} = G(\beta) = \sum_{i=1}^{n} \left\{ \left(\frac{y_i}{F(X_i\beta)} \right) f(X_i\beta)X_i - \left(\frac{1 - y_i}{1 - F(X_i\beta)} \right) f(X_i\beta)X_i \right\}$$

$$G(\beta) = \sum_{i=1}^{n} \left\{ f(X_i\beta)X_i \left(\frac{y_i - F(X_i\beta)}{F(X_i\beta)\big(1 - F(X_i\beta)\big)} \right) \right\}$$

$$G(\beta) = 0 \Leftrightarrow \sum_{i=1}^{n} \left\{ f(X_i\beta)X_i \left(\frac{y_i - F(X_i\beta)}{F(X_i\beta)\big(1 - F(X_i\beta)\big)} \right) \right\} = 0 \qquad (5.11)$$

Dans le cas du modèle logit, l'expression du gradient se simplifie encore davantage. Cette forme simplifiée se présente alors comme suit :

$$G(\beta) = \sum_{i=1}^{n} \big(y_i - F(X_i\beta)\big)X_i$$

Cette propriété provient du fait que pour la loi logistique, on a :

$$f(X_i\beta) = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} - \frac{e^{2X_i\beta}}{(1 + e^{X_i\beta})^2} = F(X_i\beta)[1 - F(X_i\beta)]$$

β étant de dimension k+1, on obtient alors un système de k+1 équations à k+1 inconnus correspondants aux paramètres. Il ne reste plus alors qu'à spécifier la fonction de distribution F (.) pour obtenir la forme fonctionnelle des équations à résoudre. Cependant la log vraisemblance étant non linéaire (à cause notamment des expressions de $f(X_i\beta)$ et $F(X_i\beta)$), il n'est pas possible de donner une expression analytique simple de ces estimateurs, et leur calcul se fait généralement par la mise en œuvre d'un algorithme d'optimisation. Plusieurs algorithmes ont, à cet effet, été élaborés. On dénombre principalement les

algorithmes de Newton Raphson, de Berndt-Hall-Hall-Hausman, de Davidon-Fletcher-Powell, de Broyden-Fletcher-Goldfarb-Shanno. Mais quel que soit l'algorithme d'optimisation retenue, la démarche de résolution consiste à maximiser la fonction de logvraisemblance en partant des conditions du premier ordre. Cependant, compte tenu de l'absence de solution analytique, le choix optimal du paramètre β doit être déduit par un processus itératif. Cette démarche itérative se décrit comme suit.

Dans un premier temps, des valeurs initiales sont attribuées aux k+1 éléments du vecteur β. C'est la définitions des conditions initiales $\hat{\beta}_{i_0}$.

Dans un second temps, on choisit un autre vecteur $\hat{\beta}_{i_1}$ telle que $logL(\hat{\beta}_{i_1}) \geq logL(\hat{\beta}_{i_0})$. Ce choix continue séquentiellement jusqu'à ce que la condition d'arrêt soit satisfaite (i.e convergence).

La convergence est atteinte lorsque la variation de $\hat{\beta}$ ou de $logL(\hat{\beta})$ entre deux itérations consécutives devient inférieur à un certain seuil (appelé seuil de tolérance, défini arbitrairement).

5.4.2. Propriétés des estimateurs MV

Sous les hypothèses de base du modèle dichotomique, l'estimateur du maximum de vraisemblance $\hat{\beta}$ de β est un estimateur convergent et suit asymptotiquement une loi normale de moyenne β et de matrice de variance covariance égale à l'inverse de la matrice d'information de Fischer I (β) évaluée au point β.

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow N[\beta, I(\beta)^{-1}] \tag{5.12}$$

Où $I(\beta)$ est la matrice d'information de Fisher du vecteur de paramètre β.

Matrice Hessienne et matrice d'information de Fisher d'un modèle dichotomique

La matrice Hessienne associée à la logvraisemblance d'un modèle dichotomique se définit comme la dérivée du vecteur gradient. Elle correspond donc à la dérivée seconde de la fonction de logvraisemblance du modèle. Elle se définit comme suit :

$$H(\beta) = \frac{\partial^2 logL(\beta)}{\partial \beta \partial \beta'} \tag{5.13}$$

$$H(\beta) = \left(-\sum_{i=1}^{n}\left[\frac{y_i}{[F(X_i\beta)]^2} + \frac{1-y_i}{[1-F(X_i\beta)]^2}\right][f(X_i\beta)]^2 X'_i X_i\right) + \left(\sum_{i=1}^{n}\left[\frac{y_i - F(X_i\beta)}{F(X_i\beta)\big(1-F(X_i\beta)\big)}\right] f'(X_i\beta) X'_i X_i\right)$$

Où F(.) est la fonction de répartition, f(.) la fonction de densité et f'(.) la dérivée première de la fonction de densité.

La matrice d'information de Fisher se définit comme l'opposé de l'espérance de la matrice Hessienne.

$$I(\beta) = -E\big(H(\beta)\big) = -E\left(\frac{\partial^2 logL(\beta)}{\partial\beta\partial\beta'}\right)$$

$$I(\beta) = -E\left[\left(-\sum_{i=1}^{n}\left[\frac{y_i}{[F(X_i\beta)]^2} + \frac{1-y_i}{[1-F(X_i\beta)]^2}\right][f(X_i\beta)]^2 X'_i X_i\right) + \left(\sum_{i=1}^{n}\left[\frac{y_i - F(X_i\beta)}{F(X_i\beta)\big(1-F(X_i\beta)\big)}\right] f'(X_i\beta) X'_i X_i\right)\right]$$

Mais d'après le modèle initiale $y_i = F(X_i\beta) + \varepsilon_i$, on a :

$$E(y_i) = F(X_i\beta)$$

Par conséquent l'expression de la matrice d'information s'écrit telle que

$$I(\beta) = -\sum_{i=1}^{n}\left[\frac{[f(X_i\beta)]^2}{F(X_i\beta)\big(1-F(X_i\beta)\big)}\right] X'_i X_i \qquad (5.14)$$

Ainsi l'inverse de cette matrice permet d'obtenir la matrice de variance-covariance du paramètre $\hat{\beta}$.

$$V(\hat{\beta}) = \left(-\sum_{i=1}^{n}\left[\frac{[f(X_i\beta)]^2}{F(X_i\beta)\big(1-F(X_i\beta)\big)}\right] X'_i X_i\right)^{-1} \qquad (5.15)$$

5.5. Le modèle de probabilité linéaire

Le modèle de probabilité linéaire correspond à l'utilisation de la fonction linéaire pour estimer le modèle de probabilité. Ce modèle peut être estimé par les MCO. Cependant il faut noter qu'en dépit de sa simplicité attractive, le modèle de probabilité linéaire présente néanmoins des inconvénients importants. D'abord, il présente un problème de cohérence car il ne tient pas compte de la contrainte que $P(y_i = 1/X_i) = X_i\beta$ à appartenir à l'intervalle [0, 1]. Par ailleurs, comme le modèle dichotomique est toujours hétéroscédastique, l'estimateur de la variance des

moindres carrés ordinaires est biaisé, et il n'est pas possible d'effectuer des tests sans corriger cette hétérocédasticité.

5.6. Les effets marginaux dans le modèle dichotomique

Dans le modèle à variable dépendante dichotomique, les coefficients estimés ne représentent pas, comme dans le modèle linéaire, l'effet partiel des variables explicatives sur la variable explicative. Les coefficients ne peuvent donc pas s'interpréter directement, seuls les signes des coefficients sont interprétables. En effet, lorsque le coefficient associé à une variable explicative x_{ik} est positive, on dira que l'accroissement de cette variable favorise la probabilité de survenue de l'évènement $y_i = 1$. En revanche, lorsque le coefficient de la variable x_{ik} est négatif, cela signifie que l'accroissement de x_{ik} défavorise la survenue de l'évènement. Toutefois l'ampleur de cette influence qu'exerce x_{ik} sur la probabilité $P(y_i = 1)$ ne peut pas être mesurée par le coefficient β_k comme dans le modèle linéaire. Celui-ci doit être mesuré en utilisant les effets marginaux.

L'effet marginal d'une variable explicative sur la probabilité de l'évènement $y_i = 1$ est la variation de la probabilité suite à l'accroissement de la variable explicative d'une unité. Il correspond donc à la dérivée de l'expression de la fonction de probabilité $P(y_i = 1/X_i)$ par rapport à x_{ik}. En effet considérons le modèle suivant :

$$P(y_i = 1/X_i) = F(X_i\beta) = \Phi(X_i\beta) = \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

Où $P(y_i = 1/X_i)$ représente la probabilité que y_i soit égal à 1 conditionnellement aux caractéristiques $x_{i1}, x_{i2}, \ldots, x_{ik}$. β est un vecteur constitués de $k+1$ paramètres : β_0, β_1, β_2, $\ldots\beta_k$. $F(.)$ représente la fonction de répartition de la quantité $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ également notée Φ. L'effet marginal de toute variable x_{ik} se mesure comme suit :

$$EM_{x_{ik}} = \frac{\partial P(y_i = 1/X_i)}{\partial x_{ik}} = \frac{\partial \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}{\partial x_{ik}}$$

$$EM_{x_{ik}} = \beta_k \phi(x_{ik}) \qquad (5.16)$$

Où $EM_{x_{ik}}$ représente l'effet marginal, β_k, le coefficient issu de l'estimation du modèle initial et $\phi(x_{ik})$ la fonction de densité de probabilité. La fonction de densité étant, par définition, positive, l'effet marginal a donc le même signe que le coefficient β_k mais reste plus faible en valeur absolue. Par ailleurs, la valeur de l'effet marginal n'est pas constante car elle dépend des variables explicatives. Mais généralement on détermine la valeur des effets marginaux aux moyennes ou aux médianes des variables explicatives. Dans ces deux cas, les effets marginaux sont calculés comme suit :

Pour la moyenne, on a :

$$EM_{x_{ik}} = \beta_k \phi(\bar{x}_{ik})$$

Pour la médiane, on a :

$$EM_{x_{ik}} = \beta_k \phi(med(x_{ik}))$$

Par ailleurs, lorsque le modèle estimé est celui du probit où la fonction de densité est celui d'une loi normale centrée et réduite (loi symétrique), on peut aussi calculer l'effet marginal au point où celle-ci est maximale. En effet la densité de probabilité de la loi normale centrée réduite est maximale au point 0 car $\phi(0) \approx 0{,}40$. Dans ce cas, l'effet marginal se présente comme suit :

$$EM_{x_{ik}=0/probit} = \beta_k \phi(0)$$

$$EM_{x_{ik}=0/probit} = 0{,}40\beta_k$$

Aussi, lorsque le modèle estimé est le logit où la fonction de densité est celui d'une loi logistique (loi symétrique), on peut calculer l'effet marginal au point où celle-ci est maximale. En effet la densité de probabilité de la loi logistique est maximale au point 0 car $\phi(0) \approx 0{,}25$. Dans ce cas, l'effet marginal se présente comme suit :

$$EM_{x_{ik}=0/logit} = \beta_k \phi(0)$$

$$EM_{x_{ik}=0/logit} = 0{,}25\beta_k$$

Il faut aussi noter que l'une des désavantages de l'effet marginal est qu'il dépendra de l'unité de mesure de la variable explicative. C'est d'ailleurs pourquoi, on préfère utiliser l'élasticité qui mesure dans quelle proportion varie la probabilité $P(y_i = 1/X_i)$ suite à une variation de la variable explicative de 1%. La formule est la suivante :

$$EL_{x_{ik}} = \frac{\partial log P(y_i = 1/X_i)}{\partial log x_{ik}} = \frac{\partial P(y_i = 1/X_i)}{\partial x_{ik}} \frac{x_{ik}}{P(y_i = 1/X_i)} = \beta_k x_{ik} \frac{\phi(x_{ik})}{\Phi(x_{ik})}$$

$$EL_{x_{ik}} = \beta_k x_{ik} \frac{\phi(x_{ik})}{\Phi(x_{ik})} \qquad (5.17)$$

5.7. Les Odds ratio dans le modèle logit

Les coefficients obtenus à partir de l'estimation d'un modèle logit ont aussi une interprétation très intéressante. C'est celle des odds ratio. En effet, on sait que dans le modèle logit :

$$\frac{P(y_i = 1)}{1 - P(y_i = 1)} = e^{X_i\beta}$$

Cette égalité représente l'Odd de l'évènement 1. En prenant cette égalité en logarithme, on retrouve la quantité dénommée log-Odds:

$$ln\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = X_i\beta$$

Ainsi, d'une manière générale, pour calculer les Odds ratios, on calcule l'exponentiel du coefficient β. Ainsi, on a :

$$OR = e^{\beta} \tag{5.18}$$

5.8. Passage du modèle probit au modèle logit

Soit un modèle dichotomique définit à partir d'une variable latente telle que

$$\begin{cases} y_i = 1 \; si \; y_i^* > 0 \\ y_i = 0 \; si \; y_i^* \leq 0 \end{cases}$$

Avec y_i^* défini telle que $y_i^* = X_i\beta + u_i$ avec $E(u_i) = 0$ et $V(u_i) = \sigma_u^2$

Dans cette configuration la série des erreurs u_i est distribuée telle que

$$u_i \rightsquigarrow iid(0, \sigma_u^2)$$

Cependant, lorsque la fonction de répartition donnée est celle issue d'une loi $iid(0,1)$, il faudrait alors normaliser les erreurs de sorte à pouvoir déterminer les paramètres du modèle. En effet, lorsque $u_i \rightsquigarrow iid(0, \sigma_u^2)$ et que $y_i^* = X_i\beta + u_i$, alors l'erreur normalisée se présente comme suit :

$$\frac{u_i}{\sigma_u} \rightsquigarrow iid(0,1)$$

Dès lors, le modèle normalisé se présente tel que :

$$\frac{y_i^*}{\sigma_u} = \frac{X_i\beta}{\sigma_u} + \frac{u_i}{\sigma_u}$$

Ainsi, on peut alors poser que :

$$P(y_i = 1) = P\left(\frac{y_i^*}{\sigma_u} > 0\right) = P\left(\frac{X_i\beta}{\sigma_u} + \frac{u_i}{\sigma_u} > 0\right) = P\left(\frac{u_i}{\sigma_u} > -\frac{X_i\beta}{\sigma_u}\right)$$

$$= 1 - P\left(\frac{u_i}{\sigma_u} \leq -\frac{X_i\beta}{\sigma_u}\right) = 1 - \left[1 - P\left(\frac{u_i}{\sigma_u} < \frac{X_i\beta}{\sigma_u}\right)\right] = P\left(\frac{u_i}{\sigma_u} < \frac{X_i\beta}{\sigma_u}\right) = F\left(\frac{X_i\beta}{\sigma_u}\right)$$

Ainsi, on a :

$$P(y_i = 1) = F\left(\frac{X_i\beta}{\sigma_u}\right) \tag{5.19}$$

Cette propriété montre par exemple que lorsque $u_i \rightsquigarrow N(0, \sigma_u^2)$ alors, pour estimer le modèle probit (dans lequel la fonction de répartition considérée est celle d'une loi normale centrée et réduite), il faut donc procéder à une pré-normalisation de l'équation de la variable latente (en utilisation σ_u^2). Cette propriété permet donc de passer d'une loi normale $N(0, \sigma_u^2)$ à une loi normale $N(0,1)$.

Par ailleurs, on peut aussi montrer qu'il est possible de passer d'une loi iid(0,1) à une loi logistique en normalisation par la variance de la loi logistique.

En effet, sachant que la variance d'une loi logistique est égale à $\frac{\pi^2}{3}$, pour passer d'une loi iid(0,1) à une loi logistique l(0, $\frac{\pi^2}{3}$), on multiplie celle-ci par $\frac{\pi}{\sqrt{3}}$ (écart-type). En effet si on a :

$$u_i \rightsquigarrow N(0, \sigma_u^2) \Rightarrow \frac{u_i}{\sigma_u} \rightsquigarrow N(0,1) \Rightarrow \frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u} \rightsquigarrow l\left(0, \frac{\pi^2}{3}\right)$$

Démonstration :

$$E\left(\frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u}\right) = \frac{\pi}{\sigma_u\sqrt{3}}E(u_i) = \frac{\pi}{\sigma_u\sqrt{3}} \times 0 = 0$$

$$Var\left(\frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u}\right) = E\left[\left(\frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u}\right)^2\right] = \frac{\pi^2}{3\sigma_u^2}E({u_i}^2) = \frac{\pi^2}{3\sigma_u^2}Var(u_i) = \frac{\pi^2}{3\sigma_u^2} \times \sigma_u^2 = \frac{\pi^2}{3}$$

En utilisant cette propriété, on peut passer de l'estimation d'un modèle probit à celle d'un modèle probit par une transformation des fonctions de répartition.

En effet si le modèle initial est tel que :

$$\begin{cases} y_i = 1 \; si \; y_i^* > 0 \\ y_i = 0 \; si \; y_i^* \leq 0 \end{cases}$$

Avec y_i^* défini telle que : $y_i^* = X_i\beta + u_i$ où $u_i \rightsquigarrow N(0, \sigma_u^2)$

Alors, pour estimer un modèle probit, on utilise la fonction de répartition définie telle que :

$$P(y_i = 1) = F\left(\frac{X_i\beta}{\sigma_u}\right) \tag{5.20a}$$

Où F(.) est la fonction de répartition d'une loi normale centrée réduite.

En revanche, pour estimer un modèle logit, on utilise la fonction de répartition définie telle que :

$$P(y_i = 1) = F\left(\frac{\pi}{\sqrt{3}}\frac{X_i\beta}{\sigma_u}\right) \qquad (5.20b)$$

Où F(.) est la fonction de répartition d'une loi logistique.

Les étapes de la normalisation se présentent alors comme suit :

$$y_i^* = X_i\beta + u_i$$

$$\frac{y_i^*}{\sigma_u} = \frac{X_i\beta}{\sigma_u} + \frac{u_i}{\sigma_u}$$

$$\frac{\pi}{\sqrt{3}}\frac{y_i^*}{\sigma_u} = \frac{\pi}{\sqrt{3}}\frac{X_i\beta}{\sigma_u} + \frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u}$$

Ainsi, on peut alors poser que :

$$P(y_i = 1) = P\left(\frac{\pi}{\sqrt{3}}\frac{y_i^*}{\sigma_u} > 0\right) = P\left(\frac{\pi}{\sqrt{3}}\frac{X_i\beta}{\sigma_u} + \frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u} > 0\right)$$

$$= P\left(\frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u} > -\frac{\pi}{\sqrt{3}}\frac{X_i\beta}{\sigma_u}\right) = 1 - P\left(\frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u} \leq -\frac{\pi}{\sqrt{3}}\frac{X_i\beta}{\sigma_u}\right)$$

$$= 1 - \left[1 - P\left(\frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u} < \frac{\pi}{\sqrt{3}}\frac{X_i\beta}{\sigma_u}\right)\right] = P\left(\frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u} < \frac{\pi}{\sqrt{3}}\frac{X_i\beta}{\sigma_u}\right) = F\left(\frac{\pi}{\sqrt{3}}\frac{X_i\beta}{\sigma_u}\right)$$

Ainsi, on a :

$$P(y_i = 1) = F\left(\frac{\pi}{\sqrt{3}}\frac{X_i\beta}{\sigma_u}\right)$$

Compte tenu de ces précédentes propriétés, on peut donc passer d'un modèle à un autre par l'intermédiaire d'un coefficient de proportionnalité. La conséquence directe de cette propriété est qu'on peut aussi déduire les estimateurs d'un modèle à partir d'un autre sans avoir besoin d'estimer celui. En effet, il est établit la propriété fondamentale suivante :

Soit $\hat{\beta}_{probit}$, l'estimateur de β obtenu à partir du modèle probit et $\hat{\beta}_{logit}$, celui obtenu à partir d'un modèle logit, on peut montrer que :

$$\hat{\beta}_{logit} = 1.6\hat{\beta}_{probit} \qquad (5.21)$$

Cette même propriété permet d'établir une relation entre l'estimation du modèle probit et celui du modèle de probabilité linéaire (obtenu par MCO) :

$$\hat{\beta}_{mco} = 0.4\hat{\beta}_{probit} \quad (5.22)$$

5.9. Diagnostics sur la qualité de l'estimation des modèles logit et probit

Dans le but de juger de la précision d'un modèle estimé, un certain nombre de mesures ont été proposées. Il s'agit du Pseudo R² (défini suivant le principe du coefficient de détermination R² étudié dans le modèle linéaire) et la proportion de prédictions correctes (mesure du pouvoir de prédiction).

5.9.1. Le R2 de McFadden

Le R2 McFadden est un indice de qualité d'ajustement construit en comparant la valeur de la logvraisemblance du modèle estimé avec toutes les variables explicatives avec la valeur de la logvraisemblance du modèle estimé avec uniquement la constante. La première désigne le logvraisemblance du modèle non contraint alors que la seconde représente le logvraisemblance du modèle contraint. La formule se présente comme suit :

$$R2 = 1 - \left(\frac{LogL_{nc}}{LogL_c}\right) \quad (5.23)$$

Où $LogL_{nc}$ représente le logvraisemblance du modèle non contraint et $LogL_c$ le logvraisemblance du modèle contraint.

Malgré la proximité apparente de cette formule avec celle du R2 du modèle linéaire, sa justification mathématique de cette mesure n'est pas du tout identique à celle du R2. En effet, la principale idée qui sous-tend cette mesure est la suivante. Pour un modèle bien ajusté, la vraisemblance non restreinte L_{nc} doit être proche de 1 ($LogL_{nc}$ proche de 0). Dans ce cas le pseudo-R2 est proche de 1. Au contraire, pour un modèle mal ajusté, $LogL_{nc}$ sera proche de $LogL_c$ et la valeur du pseudo-R2 sera proche de zéro.

5.9.2. Le pouvoir de prédiction du modèle et le pseudo R2

Une autre façon de mesurer la qualité de l'ajustement est d'examiner la capacité prédictive du modèle estimé. L'idée ici est de calculer la proportion de prédictions correctes du modèle. Le calcul du nombre de prédictions correctes nécessite d'abord de définir une règle à partir de laquelle la prédiction de la probabilité $P(y_i = 1/X_i)$ peut conduire à un prédicteur discret de l'état $y_i = 1$ ou $y_i = 0$. En d'autres termes, il faut définir une valeur seuil à partir de la probabilité prédite ($\hat{P}_i = F(X_i\hat{\beta})$ à partir duquel, on considère que l'évènement $y_i = 1$ est réalisé ou que l'évènement $y_i = 0$ est réalisé. Cette règle est définie comme suit :

$$\hat{y}_i = \begin{cases} 1 \; si \; \hat{P}_i > 0{,}5 \\ 0 \; si \; \hat{P}_i \leq 0{,}5 \end{cases}$$

A partir des valeurs prédites des évènements construites à partir de ce critère, on procède à une confrontation avec les évènements réellement observées y_i afin d'évaluer le degré de concordance du modèle avec la réalité. Cette confrontation se réalise généralement au moyen d'un tableau de contingence qui se présente comme suit :

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

n_{00} représente le nombre de cas où le modèle prédit une absence de l'évènement et cette absence se vérifie bien dans la réalité. n_{10} représente le nombre de cas où le modèle prédit l'absence d'évènement alors que dans la réalité, il y a eu l'évènement. n_{01} représente le nombre de cas où le modèle prédit la survenue de l'évènement alors que dans la réalité l'évènement n'est pas survenue. Et enfin, n_{11} représente le nombre de cas où le modèle prédit la survenue de l'évènement et que cette prédiction se confirme bien dans l réalité. A partir de ces quatre situations, le nombre de prédictions correctes est obtenu en additionnant n_{00} et n_{11}. Ainsi, le pourcentage de prédictions correctes se calcule en rapport simplement ce nombre au nombre d'observation total n ($n = n_{00} + n_{10} + n_{01} + n_{11}$). Cette valeur traduit également le pseudo R^2.

$$PseudoR2 = \left(\frac{n_{00} + n_{11}}{n}\right) \tag{5.24}$$

Le modèle sera alors considéré de bonne qualité lorsque cette proche est proche de 1.

5.10. Test d'hypothèses dans le cadre du modèle dichotomique

5.10.1. Test sur un coefficient

Supposons que l'on veuille tester l'hypothèse $H_0: \beta_j = \beta_{0,j}$ contre l'alternative $H_1: \beta_j \neq \beta_{0,j}$, on peut dans ce cas utiliser un test de Student classique ou bien un test de Wald. Cependant, le choix entre le test de Student et le test de Wald dépend de la connaissance ou non de la variance des résidus $\sigma_{\hat{\varepsilon}}^2$. En effet la

statistique du test du test de Student ou de Wald se présente indifféremment comme suit :

$$\frac{\hat{\beta}_j - \beta_{0,j}}{\hat{\sigma}_{\hat{\beta}_j}}$$

Etant donné que la valeur $\hat{\sigma}^2_{\hat{\beta}_0}$ depend de la valeur $\sigma^2_{\hat{\varepsilon}}$ (Voir par exemple dans le cadre du modèle linéaire), alors la distribution de la statistique de $\frac{\hat{\beta}_j - j}{\hat{\sigma}_{\hat{\beta}_j}}$ depend du fait que $\sigma^2_{\hat{\varepsilon}}$ soit connue ou pas.

Si $\sigma^2_{\hat{\varepsilon}}$ est connue, alors la statstique $\frac{\hat{\beta}_j - \beta_{0,j}}{\hat{\sigma}_{\hat{\beta}_j}}$ sera distribuée selon une loi normale. La statistique sera alors appellée comme la statistique de Wald.

$$W = \frac{\hat{\beta}_j - \beta_{0,j}}{\hat{\sigma}_{\hat{\beta}_j}} \rightsquigarrow N(0,1) \tag{5.25}$$

On peut aussi montrer que le carré de la statistique de Wald est une loi de Khi-deux, car elle correspond théoriquement à la somme du carré d'une loi normale (0,1) :

$$W^2 = \frac{\left(\hat{\beta}_j - \beta_{0,j}\right)^2}{\hat{\sigma}^2_{\hat{\beta}_j}} \rightsquigarrow \chi^2(1) \tag{5.26}$$

Cette propriété montre donc que le test de Wald peut être réalisé soit à partir d'une distribution normale ou une distribution de khi-deux lorsque la variance des erreurs est connue. En revanche dans le cas où la variance $\sigma^2_{\hat{\varepsilon}}$ n'est pas connue, il faut utiliser la statistique de student. Dans ce cas, la statistique du test se présente comme suit :

$$t^* = \frac{\hat{\beta}_j - \beta_{0,j}}{\hat{\sigma}_{\hat{\beta}_j}} \rightsquigarrow T(n-k-1) \tag{5.27}$$

Nb : Pour obtenir l'écart-type des coefficients, on se sert de la matrice de variance-covariance des paramètres dans laquelle on peut lire $\hat{\sigma}^2_{\hat{\beta}_j}$. La matrice de variance-covariance étant égale à l'inverse de la matrice d'information de Fisher.

5.10.2. Test de Wald sur une contrainte linéaire de coefficients

La démarche de test sur un coefficient individuel peut être généralisée à un ensemble de coefficients sous forme de contraintes linéaires.

D'une manière générale, tous les tests de contraintes linéaires se présentent sous la forme suivante :

$$H_0 \;\; R\beta = r$$

Où R est une matrice q× q avec q≤ k+1 et r un vecteur-colonne de dimension q.
La statistique Wald de ce test se présente alors comme suit :

$$W = \left(R\hat{\beta} - r\right)'(R\hat{V}R')^{-1}\left(R\hat{\beta} - r\right)$$

Où $\hat{V}$ est la matrice de variance-covariance des paramètres.

Cette statistique est distribuée selon un Khi-deux à 1 degré de liberté.

Cette propriété dérive du fait que pour tout coefficient estimé $\hat{\beta}_j$, on a :

$$W = \hat{\beta}_j \hat{\sigma}^2_{\hat{\beta}_j} \rightsquigarrow \chi^2(1) \qquad (5.28)$$

5.10.3. Test du rapport de vraisemblances

De même que le test de Wald peut être utilisé pour tester une contrainte linéaire, on peut aussi utiliser le test de rapport de vraisemblance (likelihood Ratio test) lorsque le modèle estimée par maximum de vraisemblance.

Le test de rapport de vraisemblance consiste à comparer la vraisemblance du modèle estimé sous l'hypothèse H0 (modèle contraint) à la vraisemblance du modèle estimé sous l'hypothèse alternative (modèle non contraint). L'idée du test est que si la contrainte imposée par l'hypothèse H0 est valide, alors la vraisemblance obtenue en estimant le modèle contraint doit être sensiblement la même que celle obtenue en estimant le modèle non contraint. Dans ce cas, d'après le principe de parcimonie (qui stipule que le meilleur d'entre deux modèles ayant les mêmes propriétés par ailleurs est celui pour lequel le nombre de paramètres à estimer est le plus faible), l'hypothèse nulle ne peut pas être rejetée.

La statistique du test de rapport de vraisemblance se construit comme suit :

$$LR = \left(\frac{L_{nc}}{L_c}\right) \qquad (5.29a)$$

Où L_{nc} et L_c représentent respectivement les vraisemblances du modèle non contraint et du modèle contraint.

En élevant cette égalité au carré et en prenant le logarithme, on trouve :

$$-2lnLR = 2(lnL_c - lnL_{nc}) \rightsquigarrow \chi^2(q) \qquad (5.29b)$$

Cette statistique suit un Khi-deux à q degrés de liberté où q représente le nombre de contraintes. Ainsi, en fonction de la valeur de la statistique $2(lnL_c - lnL_{nc})$, l'hypothèse H0 sera rejetée lorsque la statistique est supérieur à la valeur du Khi-deux lue dans la table au seuil $1 - {}^{\alpha}/_{2}$).

5.10.4. Le test du multiplicateur de Lagrange

Le test de multiplicateur de Lagrange (encore appelé test de score) part du principe que si l'hypothèse nulle est satisfaite, les deux estimateurs non contraint $\hat{\beta}_c$ et contraint $\hat{\beta}_{nc}$ ne doivent pas être significativement différents, et que donc la même propriété doit être vérifiée pour le vecteur des conditions du premier ordre de la maximisation de la log vraisemblance. Le test consiste, en pratique, à mesurer la norme du score évalué au point $\hat{\beta}$. Cette statistique se présente alors comme suit :

$$LM = \left(\frac{\partial logL(\beta)}{\partial \beta} / \beta = \hat{\beta}\right)' I(\hat{\beta})^{-1} \left(\frac{\partial logL(\beta)}{\partial \beta} / \beta = \hat{\beta}\right) \rightsquigarrow \chi^2(1) \qquad (5.30)$$

Où $I(\hat{\beta})$ est la matrice d'information de Fisher. Cette statistique suit une loi de khi-deux à q degré de libertés où q est le nombre de contraintes.

CHAPITRE 6. MODELES A VARIABLE DEPENDANTE POLYTOMIQUE

6.1. Présentation

Les modèles à variable dépendante polytomique sont des modèles dans lesquels la variable expliquée prend plus de deux modalités. On dénombre deux grandes catégories de ce genre de modèles qui se distinguent selon qu'on puisse établir ou pas un ordre dans les modalités. Il s'agit des modèles multinomiaux ordonnées et des modèles multinomiaux non ordonnés.

Les modèles multinomiaux ordonnées sont des modèles dans lesquels il existe un certain ordre dans les modalités. Par exemple, lorsqu'on demande à un groupe d'individus d'exprimer leur degré de satisfaction par rapport à la consommation d'un produit, les réponses obtenus peuvent être codées comme suit : 0- *Pas du tout satisfait*, 1-*Plutôt pas satisfait*, 2- *Ni satisfait, ni insatisfait*, 3-*Plutôt satisfait*, 4-*Très satisfait*. On remarque alors un ordre clair dans les modalités, passant de la pire à la meilleure situation (ou réciproquement).

Les modèles multinomiaux non ordonnés sont des modèles dans lesquels il n'existe aucun ordre clair dans les modalités. C'est le cas par exemple lorsqu'on veut étudier le choix des individus entre plusieurs moyens de transport pour faire un trajet. On peut proposer plusieurs modalités : 1-*Vélo*, 2-*Moto*, 3-*Voiture personnelle*, 4-*Bus, 5-Autres*. On ne peut pas mettre un ordre hiérarchique clair entre ces modalités dans la mesure où ce choix appartient uniquement à l'individu qui est en capacité d'estimer le choix le plus favorable pour lui. Ce type de modèle est alors appelé modèle multinomial non ordonnée.

D'une manière générale, un modèle multinomial (qu'il soit ordonné ou non ordonné) est un modèle dans le lequel la variable dépendante qualitative y_i prend un nombre m de modalités supposées mutuellement exclusives et dont la somme des probabilités vaut 1.

Supposons un échantillon constitué de N individus indicés i avec $i = 1, \dots, n$. Ces individus sont supposés choisir entre m modalités indicées k avec $k = 1, \dots, m$. La probabilité pour qu'un individu i choisisse la modalité k est une fonction de ses caractéristiques telles que :

$$Prob(y_i = k) = F(X_i\beta) \quad \forall\, i = 1, \dots, n;\ \forall\, k = 1, \dots, m \qquad (6.1)$$

Où $Prob(y_i = k)$ est la probabilité que l'individu i choisisse la modalité k. $F(.)$ est la fonction de répartition, supposée différer selon les individus (i) mais aussi selon les modalités (k). X_i est la matrice des caractéristiques de l'individu. Et β

représente le vecteur des paramètres du modèle. Etant donné que les modalités sont mutuellement exclusives, la condition suivante est donc vérifiée pour chaque individu:

$$\sum_{k=1}^{m} Prob(y_i = k) = 1 \quad \forall\, i = 1, \dots, n \tag{6.2}$$

La condition (6.2) indique que la somme des probabilités des *m* modalités vaut 1 pour chaque individu.

L'un des corollaires de cette condition est qu'il n'est pas nécessaire de spécifier la fonction de probabilité de toutes les modalités. En effet, étant donné qu'il y a *m* modalités, on peut montrer que la probabilité de la *m-ième* modalité (connaissant celles des *m-1* premières modalités) est la suivante :

$$Prob(y_i = m) = 1 - \left(\sum_{k=1}^{m-1} Prob(y_i = k) \right) \quad \forall\, i = 1, \dots, n \tag{6.3}$$

Ainsi les modèles multinomiaux (qu'ils soient ordonnés ou non ordonnés) présentent ces mêmes propriétés fondamentales. Cependant des différences majeures apparaissent entre ces modèles quant à la spécification des formes fonctionnelles des probabilités mais aussi quant à l'interprétation des résultats.

6.2. Modèles multinomiaux ordonnés : logit et probit ordonnés

Comme signalé ci-dessus, les modèles multinomiaux ordonnés sont des modèles dans lesquels on peut établir un ordre formel entre les modalités. Mais au-delà de cette définition simple, ces modèles peuvent être rigoureusement définis comme des modèles où les modalités prises par la variable dépendante sont des codes attribués aux intervalles de valeurs obtenus par découpage des valeurs continues d'une variable latente. En se basant sur cette définition, le modèle multinomial ordonné se présente comme suit :

$$y_i = \begin{cases} 1 & si\ \ y_i^* < \gamma_1 \\ 2 & si\ \ \gamma_1 \leq y_i^* < \gamma_2 \\ 3 & si\ \ \gamma_2 \leq y_i^* < \gamma_3 \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ m & si\ \ y_i^* > \gamma_m \end{cases} \tag{6.4a}$$

Où y_i représente la variable dépendante pouvant prendre les modalités $1,2,..,m$ qui elles-mêmes correspondent à des intervalles de valeurs d'une variable latentes continue y_i^* définie telle que :

$$y_i^* = X_i\beta + u_i \qquad (4b)$$

$$E(u_i) = 0$$

$$V(u_i) = \sigma_u^2$$

Dans ce modèle $\gamma_1, \gamma_2,\ldots, \gamma_m$ sont des constantes délimitant les intervalles de valeurs de la variable latente. Ce sont des valeurs seuils qui conditionnent les choix des individus. Comme on peut le voir à travers l'équation, l'individu fait le choix de la modalité 1 si la valeur de la variable latente est inférieure à un certain seuil γ. Par contre, il fera le choix m (qui est celui qui domine tous les autres choix) lorsque la valeur de y_i^* dépasse le seuil γ_m. En dehors de ces deux cas extrêmes, il existe aussi des choix intermédiaires correspondant chacun à des intervalles de valeurs spécifiques (voir équation 6.4a).

NB : Dans ce modèle, nous avons décidé de coder la modalité à partir de 1. Le plus souvent, la première modalité est codée par 0. Mais il ne s'agit là que d'un choix de forme qui n'a aucune incidence ni sur les paramètres estimés, ni sur les résultats. Selon le codage adopté ici, la probabilité associée à chaque modalité se présente comme suit :

$$\begin{aligned}
Prob(y_i = 1) &= Prob(y_i^* < \gamma_1) = Prob(X_i\beta + u_i < \gamma_1) \\
&= Prob(u_i < \gamma_1 - X_i\beta) \\
&= F(\gamma_1 - X_i\beta)
\end{aligned}$$

$$\begin{aligned}
Prob(y_i = 2) &= Prob(\gamma_1 \leq y_i^* < \gamma_2) = Prob(\gamma_1 \leq X_i\beta + u_i < \gamma_2) \\
&= Prob(\gamma_1 - X_i\beta \leq u_i < \gamma_2 - X_i\beta) \\
&= F(\gamma_2 - X_i\beta) - F_{ik}(\gamma_1 - X_i\beta)
\end{aligned}$$

$$\begin{aligned}
&\ldots \qquad\qquad \ldots \qquad\qquad\qquad \ldots \qquad\qquad \ldots \\
Prob(y_i = m) &= Prob(y_i^* > \gamma_m) = Prob(X_i\beta + u_i > \gamma_m) \\
&= Prob(u_i > \gamma_m - X_i\beta) \\
&= 1 - Prob(u_i < \gamma_m - X_i\beta) \\
&= 1 - F(\gamma_m - X_i\beta)
\end{aligned}$$

Connaissant les valeurs seuils et la fonction F(.), on peut aussi proposer une formule générale pour obtenir la probabilité de chaque modalité. Cette formule se présente comme suit :

$$Prob(y_i = k) = F(\gamma_k - X_i\beta) - F(\gamma_{k-1} - X_i\beta) \qquad (6.5)$$

Avec $\gamma_0 = -\infty$ et $\gamma_m = +\infty$

Pour estimer le modèle, il faut d'abord calculer la fonction de vraisemblance qui est le double produit des probabilités des modalités à la fois entre les modalités mais aussi entre les individus. La fonction de vraisemblance d'un modèle multinomial ordonné se présente comme suit :

$$L(y,\gamma,\beta) = \prod_{i=1}^{n}\prod_{k=1}^{m}[prob(y_i = k)]$$

$$L(y,\gamma,\beta) = \prod_{i=1}^{n}\prod_{k=1}^{m}[F(\gamma_k - X_i\beta) - F(\gamma_{k-1} - X_i\beta)] \qquad (6.6a)$$

$$\text{Avec } \gamma_0 = -\infty \text{ et } \gamma_m = +\infty$$

On peut alors déterminer les paramètres de ce modèle en maximisant cette fonction de vraisemblance. C'est la procédure de maximum de vraisemblance. Mais maximiser la vraisemblance équivaut aussi à maximiser la log-vraisemblance. C'est pourquoi, la détermination est basée sur la fonction de log-vraisemblance car plus facile à dériver. La fonction de log-vraisemblance se présente alors comme suit :

$$LogL(y,\gamma,\beta) = \sum_{i=1}^{n}\sum_{k=1}^{m} log[F(\gamma_k - X_i\beta) - F(\gamma_{k-1} - X_i\beta)] \qquad (6.6b)$$

$$\text{Avec } \gamma_0 = -\infty \text{ et } \gamma_m = +\infty$$

Connaissant ainsi la fonction F(.), il suffit alors de dériver l'équation (6.6b) afin de rechercher les valeurs des paramètres γ et β qui sont deux vecteurs de paramètres contenant respectivement les valeurs seuils et les coefficients des variables explicatives. Il faut noter que X_i ne contient pas de constante pour des raisons d'identification. En effet, il est impossible d'identifier en même temps le coefficient associé à la constante et les valeurs seuils γ_k.

Dans l'estimation des paramètres du modèle, le choix de la fonction $F(.)$ dépend de la distribution de u_i dans le modèle (6.4b). Lorsque $u_i \rightsquigarrow N(0,1)$ alors la fonction de répartition choisie est celle d'une loi normale centrée réduite. Le modèle ainsi obtenu est appelé *modèle probit multinomial ordonné.* En revanche, lorsque $u_i \rightsquigarrow l(0,\frac{\pi}{3})$ alors la fonction de répartition choisie est celle d'une loi logistique. Le modèle est alors appelé *modèle logit multinomial ordonné.*

Remarque :

Lorsque la série des erreurs u_i est distribuée selon une loi iid telle que la loi normale : $u_i \rightsquigarrow N(0,\sigma_u^2)$ avec $\sigma_u^2 \neq 1$, il faudrait alors normaliser l'ensemble du modèle de sorte que u_i puisse être distribuée $u_i \rightsquigarrow N(0,1)$. Pour cela, on multiplie

tous les membres du modèle par $\frac{1}{\sigma_u}$ tel que : le modèle normalisé se présente tel que : $\frac{y_i^*}{\sigma_u} = \frac{X_i\beta}{\sigma_u} + \frac{u_i}{\sigma_u}$. Notons aussi que pour obtenir un modèle où $u_i \rightsquigarrow l(0, \frac{\pi}{3})$, il faut simplement multiplié le modèle initial par $\frac{\pi}{3} \times \frac{1}{\sigma_u}$ tel que : $\frac{\pi}{\sqrt{3}}\frac{y_i^*}{\sigma_u} = \frac{\pi}{\sqrt{3}}\frac{X_i\beta}{\sigma_u} + \frac{\pi}{\sqrt{3}}\frac{u_i}{\sigma_u}$.

Dans l'un ou l'autre cas, on obtient finalement un modèle transformé ayant les mêmes propriétés qu'un modèle ordinaire sur lequel peut appliquer l'ensemble des démarches précédemment présentées.

Ainsi au final, compte tenu de la nature de la fonction de répartition, on distingue deux grands types de modèles multinomiaux ordonnés : le modèle logit multinomial ordonné et le modèle probit multinomial ordonné. La probabilité de choix de chaque modalité dans chacun des deux modèles se présent comme suit :

Probit ordonné

$$Prob(y_i = k) = \left[\int_{-\infty}^{(\gamma_k - X_i\beta)} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\right) dt\right] - \left[\int_{-\infty}^{(\gamma_{k-1} - X_i\beta)} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\right) dt\right]$$

Logit ordonné

$$Prob(y_i = k) = \left(\frac{e^{(\gamma_k - X_i\beta)}}{1 + e^{(\gamma_k - X_i\beta)}}\right) - \left(\frac{e^{(\gamma_{k-1} - X_i\beta)}}{1 + e^{(\gamma_{k-1} - X_i\beta)}}\right)$$

$$\text{Avec } \gamma_0 = -\infty \text{ et } \gamma_m = +\infty$$

$$\forall\, k = 1, \ldots, m$$

6.3. Les modèles multinomiaux non ordonnés : cas du logit non ordonné

Les modèle multinomiaux non ordonnés sont utilisés lorsqu'on ne peut pas établir d'ordre naturel entre les modalités d'une variable dépendante qualitative. C'est le cas par exemple lorsqu'on propose à un individu de choisir entre trois chemises différent par leur couleur mais identiques sur toutes les autres caractéristique. Dans ce cas l'individu doit choisir entre plusieurs alternatives non nécessairement hiérarchisé. Pour analyser les choix de cet individu, on peut utiliser un modèle multinomial non ordonné. Le modèle le plus fréquent parmi cette classe de modèles est le modèle logit multinomial non ordonné.

Le modèle logit multinomial non ordonné (souvent appelé simplement comme modèle logit multinomial), la probabilité de chaque modalité s'écrit comme suit :

$$Prob(y_i = k) = \frac{e^{X_i\beta_k}}{\sum_{j=1}^{m} e^{X_i\beta_j}} \quad \forall\ k = 1, \dots, m \ ; \quad \forall\ i = 1, \dots, n \qquad (6.7)$$

Tel que présenté, les paramètres du modèle restent indéterminés puisque la probabilité de toute modalité est écrite en fonction de son propre vecteur de paramètre mais aussi en fonction de ceux des autres modalités. Toutefois, cette indétermination peut être levée par la simple normalisation du vecteur de paramètres de l'une des modalités. Ainsi, en normalisant à 0 le coefficient β_k obtenu pour la première modalité tel que $\beta_1 = 0$, on peut alors écrire :

$$Prob(y_i = 1) = \frac{1}{1 + \sum_{j=2}^{m} e^{X_i\beta_j}}$$

$$Prob(y_i = k) = \frac{e^{X_i\beta_k}}{1 + \sum_{j=2}^{m} e^{X_i\beta_j}} \qquad \forall\ k = 2, \dots, m$$

On peut aussi remarquer que lorsque $m = 2$, on tombe dans le cas du modèle logit simple. Il suffirait alors de recoder les modalités 1 et 2 de sorte à retrouver les modèle logit binaire classique.

Tout comme le modèle logit ordonné, l'estimation du modèle logit multinomial (non ordonné) se fait en formulant d'abord la fonction de vraisemblance et de log-vraisemblance telle que :

$$L(y, \gamma, \beta) = \prod_{i=1}^{n} \prod_{k=1}^{m} [prob(y_i = k)]$$

$$L(y, \gamma, \beta) = \prod_{i=1}^{n} \prod_{k=1}^{m} \left(\frac{e^{X_i\beta_k}}{\sum_{j=1}^{m} e^{X_i\beta_j}} \right) \qquad (6.8a)$$

En prenant le logarithme de cette fonction, on retrouve :

$$LogL(y, \beta) = \sum_{i=1}^{n} \sum_{k=1}^{m} (X_i\beta_k) - \sum_{i=1}^{n} log \left(\sum_{k=1}^{m} e^{X_i\beta_k} \right)$$

Et en normalisant le premier le vecteur de paramètre à 0 pour la première modalité, on retrouve finalement :

$$LogL(y, \beta) = \sum_{i=1}^{n} \sum_{k=2}^{m} (X_i\beta_k) - \sum_{i=1}^{n} log \left(1 + \sum_{k=2}^{m} e^{X_i\beta_k} \right) \qquad (6.8b)$$

La dérivation de cette fonction permet en adoptant un algorithme adapté de retrouver les valeurs des paramètres β_k correspondant à chaque modalité. Notons aussi, que contrairement au probit ordonné, le probit "multinomial" permet bien d'estimer les coefficients associés à la constante car la matrice des caractéristiques X_i contient bien un vecteur de constante.

Il existe, par ailleurs, une propriété très importante dans le modèle logit multinomial qui indique que le rapport de probabilité entre deux alternatives est indépendant des vecteurs des paramètres des autres modalités non concernées. Cela se traduit comme suit :

$$\frac{Prob(y_i = j)}{Prob(y_i = l)} = \frac{\frac{e^{X_i\beta_k}}{\sum_{k=1}^{m} e^{X_i\beta_k}}}{\frac{e^{X_i\beta_l}}{\sum_{k=1}^{m} e^{X_i\beta_k}}} = \frac{e^{X_i\beta_k}}{e^{X_i\beta_l}} = e^{X_i(\beta_k - \beta_l)}$$

$$\frac{Prob(y_i = j)}{Prob(y_i = l)} = e^{X_i(\beta_k - \beta_l)} \qquad (6.9)$$

Ainsi les disparités entre deux réponses quelconques ne dépendent que de X_i et des vecteurs paramétriques associés à ces deux réponses β_k et β_l. C'est la condition dite IIA (*Independance of Irrelevant Alternative*).

6.4. Les extension des modèles multinomiaux : logit conditionnel, logit emboité et modèles séquentiels.

6.4.1. Le modèle logit conditionnel

Le modèle logit conditionnel proposé pour la première fois par McFadden est très étroitement lié au modèle multinomial logit non ordonné. Il existe cependant des différences majeures entre ces deux modèles. En effet, dans le modèle logit multinomial, il existe un seul vecteur de variables indépendantes pour chaque observation (caractéristiques des individus) alors qu'il y a m vecteurs différents de paramètres associé à chaque modalité. Mais dans le modèle logit conditionnel, les valeurs des variables indépendantes diffèrent selon les alternatives (caractéristiques relatives aux modalités). C'est le vecteur des paramètres qui est unique. Le modèle logit conditionnel suppose alors que le vecteur de paramètre β est indépendant des modalités.

Pour mieux caractériser le modèle logit conditionnel, prenons le cas d'une localité dont les habitants, pour se rendre au travail, ont le choix entre plusieurs modes de déplacement : 1-*Marche à pied et associés*, 2- *Vélo*, 3- *Voiture personnelle*, 4-*Bus*. On veut alors analyser les déterminants des choix des individus entre ces différents modes de transport. Pour cela, on cherche d'abord à caractériser ces

modalités en déterminant les vecteurs X_{ik}. On peut alors considérer la durée de trajet associée à chaque modalité (temps de trajet) et le coût financier associé à chacune des modalités (dépenses monétaires : ticket de transport, prix du carburant, etc..). On peut alors admettre que ces caractéristiques sont propres aux modalités et non aux individus. D'où la notion de logit conditionnel. Bien entendu, les individus font leur choix en tenant compte de ces facteurs (ici durée et coût) ; et ceux-ci influencent de la même manière les différentes modalités.

Dans le modèle logit multinomial conditionnel, la probabilité que l'individu i choisisse la modalité $k, \forall\, k = 1, .., m$, est définie comme suit :

$$Prob(y_i = k) = \frac{e^{X_{ik}\beta}}{\sum_{j=1}^{m} e^{X_{ij}\beta}} \quad \forall\ k = 1, \dots, m \ ; \ \forall\ i = 1, \dots, n \qquad (6.10)$$

En normalisant les vecteurs de caractéristiques X_{ij} par X_{i1} le vecteur de caractéristique de la première modalité tel que $X_{ij}^{*} = X_{ij} - X_{i1}$, on peut alors écrire :

$$Prob(y_i = 1) = \frac{1}{1 + \sum_{j=2}^{m} e^{X_{ij}^{*}\beta}}$$

$$Prob(y_i = k) = \frac{e^{X_{ik}^{*}\beta}}{1 + \sum_{j=2}^{m} e^{X_{ij}^{*}\beta}} \qquad \forall\ k = 2, \dots, m$$

En prenant l'exemple du mode de transport, on peut considérer que la modalité de référence est la marche à pied. Pour normaliser les vecteurs des caractéristiques, on soustrait respectivement le temps de trajet lié à la marche à pied ainsi que la dépense monétaire associée (en l'occurrence 0) de toutes les autres valeurs sur les autres modalités. On obtient alors les X_{ik}^{*}

L'un des grands avantages du modèle logit conditionnel est sa capacité à prédire la probabilité associée à nouvelle modalité introduite parmi les modalités pré-existantes. En effet, la probabilité associée à une nouvelle modalité s'exprime comme suit :

$$Prob(y_i = m + 1) = \frac{e^{(\hat{X}_{im+1}^{*}\hat{\beta})}}{1 + \sum_{k=2}^{m} e^{X_{ij}^{*}\hat{\beta}} + e^{(\hat{X}_{im+1}^{*}\hat{\beta})}} \qquad \forall\ k = 2, \dots, m \qquad (6.11)$$

Où $m + 1$ représente l'indice de la nouvelle modalité ajoutée aux m modalités existantes. $\hat{\beta}$ est le vecteur de paramètres estimé avec m premières modalités. X_{ij}^{*} est le vecteur de caractéristique associée à la modalité j ; valeur normalisée telle que $X_{ij}^{*} = X_{ij} - X_{i1}$. Quant à $\hat{X}_{im+1}^{*}$, il représente la valeur estimée du vecteur de

caractéristique associée à la modalité $m+1$, normalisée telle que $\hat{X}^*_{im+1} = \hat{X}_{im+1} - X_{i1}$.

La valeur estimée $\hat{X}_{im+1}$ est généralement une valeur hypothétique qu'on attribue à la nouvelle modalité. En effet, reprenons l'exemple du choix entre les modes de transport définis par les modalités suivantes : 1-*Marche à pied et associés*, 2-*Vélo*, 3- *Voiture personnelle*, 4- *Bus*. Supposons maintenant que dans cette localité on veuille introduire un nouveau mode de transport, le tramway (codé par 5). Puisque la modalité était inexistante, on ne peut donc pas déterminer avec précisions les durées de trajet et même le coût financier à cette nouvelle modalité. Il faut alors s'inspirer d'autres expériences à partir d'autres localités afin d'avoir une idée approximative sur les durées de trajet et du coût financier. Ce qui permet alors d'obtenir $\hat{X}_{ij}$ et par ricochet $\hat{X}^*_{im+1}$.

Par ailleurs, tout comme le modèle logit multinomial, le modèle logit conditionnel respecte la condition dite IIA (*Independance of Irrelevant Alternative*). En effet, en calculant le rapport de probabilité entre deux alternatives *j* et *l*, on trouve :

$$\frac{Prob(y_i = j)}{Prob(y_i = l)} = \frac{\frac{e^{X_{ij}\beta}}{\sum_{k=1}^{m} e^{X_{ik}\beta}}}{\frac{e^{X_{il}\beta}}{\sum_{k=1}^{m} e^{X_{ik}\beta}}} = \frac{e^{X_{ij}\beta}}{e^{X_{il}\beta}} = e^{(X_{ij}-X_{il})\beta}$$

$$\frac{Prob(y_i = j)}{Prob(y_i = l)} = e^{(X_{ij}-X_{il})\beta} \qquad (6.12)$$

Cette condition montre que les disparités entre deux réponses quelconques ne dépendent que de X_{ik} et du vecteur de paramètres β. Selon cette condition IAA, l'introduction d'une nouvelle modalité ne modifie pas le rapport de probabilité entre deux modalités quelconques.

Cependant, la condition IAA peut être discutable dans beaucoup de situations. Prenons par exemple le cas d'un pays A dont les habitants pour se rendre dans un pays B ont le choix entre trois types de trajets : 1- le trajet terrestre assuré exclusivement par autocar par une compagnie nommée LandTrans, 2- le trajet maritime assuré par bateau de croisière exclusivement par une compagnie nommée SeaTrans, 3- le trajet aérien assuré par avion de ligne par une compagnie nommée AirTransOne. Supposons par ailleurs qu'un tiers des voyageurs choisisse chaque mode de transport. Dans ce cas, la probabilité associée à chaque modalité est 0,33 et le rapport de probabilité entre les modalités deux à deux vaut 1. Supposons maintenant qu'une nouvelle compagnie de transport aérien soit introduite nommée AirTransTwo. Puisque AirTransOne AirTransTwo sont supposées proposer des services identiques, ces deux compagnies doivent donc avoir les mêmes probabilités d'être choisie. Dans ces

conditions pour que la condition IAA soit toujours vérifiée (rapports de probabilité non influencés par l'ajout d'une modalité supplémentaire), il faudrait que les parts de marché de chaque modalité soit de 0.25 (soit un quart des passagers pour chaque modalité). Ce qui apparait peu réaliste dans la mesure la présence de deux compagnies sur le même segment peut tirer les prix à la baisse dans ce segment et inciter les voyageurs à reporter leur choix sur ce mode de transport. La question peut également se poser lorsqu'on introduit un nouveau mode de transport terrestre en l'occurrence le transport ferroviaire assurée par une compagnie de nommée RailwayTrans. Il devient moins sûr que les probabilités initiales soient maintenues après cette introduction.

D'une manière générale, lorsque l'hypothèse IAA n'est pas vérifiée, il faut alors penser à des modèles alternatifs qui ne sont pas fondés sur cette hypothèse comme par exemple les modèles multinomiaux séquentiels.

6.4.2. Le modèles multinomiaux séquentiels

Les modèles multinomiaux séquentiels sont des modèles dans lesquels il existe d'une part un ordre dans les modalités et où l'atteinte d'une modalité k est conditionnée par l'atteinte des modalités qui précèdent ce niveau. C'est le cas par exemple lorsqu'on veut analyser la probabilité pour qu'un étudiant inscrit dans le système LMD (Licence-Master-Doctorat) atteigne le niveau Doctorat dans son cursus universitaire. On peut alors retenir trois modalités : 1-*Licence* ; 2-*Master* ; 3-*Doctorat*. Pour étudier cette question, on peut se servir d'un modèle multinomial séquentiel car la probabilité d'atteindre le niveau master est conditionnelle à l'obtention du niveau Licence. De même, l'atteinte du niveau Doctorat est conditionnelle à l'obtention du niveau à la fois du niveau Licence et Master. Dans un modèle multinomial séquentiel, la probabilité d'atteinte d'un k est exprimée comme suit :

$$Prob(y_i = k) = \prod_{j=1}^{m-1} \left[\left(1 - F_j(X_i\beta)\right) F_k(X_i\beta)\right] \quad \forall\, k = 1, \dots, m\,; \forall\, i = 1, \dots, n \qquad (6.13)$$

Connaissant l'expression, on peut alors formuler la fonction de vraisemblance en effectuant le triple produit sur les modalités et les individus comme suit :

$$L(y, \beta) = \prod_{i=1}^{n} \prod_{k=1}^{m} [Prob(y_i = k)]$$

$$L(y, \beta) = \prod_{i=1}^{n} \prod_{k=1}^{m} \prod_{j=1}^{m-1} \left[\left(1 - F_j(X_i\beta)\right) F_k(X_i\beta)\right] \qquad (6.14)$$

On met alors en œuvre la procédure de maximisation de vraisemblance en prenant d'abord le logarithme de cette fonction, qui sera ensuite optimisé en utilisant les algorithmes disponibles dans la littérature. Lorsque la fonction F(.) est celle d'une loi logistique, le modèle est alors appelé modèle logit séquentiel. Malgré la complexité des procédures d'estimation de ces, plusieurs logiciels statistiques fournissent aujourd'hui des routines pour estimer aisément les paramètres de ces types de modèle. On peut noter par exemple le module *seqlogit* de stata.

Bibliographie

Alban T. (2000), "Econométrie des Variables Qualitatives", Dunod, Paris

Amemiya T. (1981), "Qualitative Response Models : A Survey", Journal of Economic Litterature,19(4), 481-536

Amemiya T. (1985), "Advanced Econometrics", Cambridge, Harvard University Press.

Araujo C, Brun J.F., Combes J. L, (2006), " Econométrie ", Bréal, Paris.

Behaghel L. (2006), "Lire l'économétrie", collection Repères, La Découverte, Paris

Berkson J. (1951), "Why I prefer Logit to Probit", Biometrics, 7, 327-339.

Bourbonnais, R. (1993), "Econométrie". Dunod, Paris.

Bourbonnais, R.,(1998), "Econométrie". Manuel et exercices corrigés, Dunod, 2è édition

Cameron et Trivedi, (2009), "Microeconometrics: methods and applications", Cambridge University press.

Cohen, M. and Pradel, J. (1993), "Econométrie". Litec, Paris.

Colletaz G. (2001), "Modèles à Variables Expliquées Qualitatives", Miméo Université Orléans

Davidson et MacKinnon, (1993), "Estimation and Inference in Econometric", Oxford University Press

Dodge, Y, Rousson, V.,(2004) "Analyse de régression appliquée", Dunod, 2è édition.

Giraud, R., Chaix, N.(1989), "Econométrie", Presses Universitaires de France (PUF).

Gourieroux C. (1989), “Econométrie des Variables Qualitatives”, Economica, Paris.

Gouriéroux C. et Monfort A. (1996), “Statistique et Modèles Econométriques”, Economica

Greene W, (2008), “Econometric Analysis ”, 6th edition, Upper Saddle River, NJ, Prentice-Hall.

Johnson, J. et DiNardo, J. (1999), “Méthodes Econométriques”. Economica, Paris, 4 edition.

Judge G.G., Miller D.J. et Mittelhammer R.C. (2000), “Econometric Foundations”, CambridgeUniversity Press.

Judge, G., Griffiths, W., Carter Hill, R., Lütkepohl, H., and Lee, T. (1985), “The Theory and Practice of Econometrics”. Wiley, USA, 2 edition.

Labrousse, C. (1983), “Introduction à l'économétrie”, Dunod, Paris.

Maddala. G.S. (1983), “Limited-dependent and Qualitative Variables in Econometrics”, EconometricSociety Monographs, 3, Cambrige University Press.

Morimune K. (1979), “Comparisons of Normal and Logistic Models in the Bivariate Dichitomous Analysis”, Econometrica, 47, 957-975.

Ruud, P. (2000), “An Introduction to classical Econometric Theory”. Oxford University Press, New York,Oxford.

Theil, H. (1979), “Principles of Econometrics”, Wiley Hamilton publication, Canada.

Tobin J. (1958), “Estimation of Relationships for Limited Dependent Variables”, Econometrica,

Wooldridge J. (2003), “Introductory Econometrics, A Modern Approach”, South-Western

Wooldridge J., (2010), “Econometric Analysis of Cross Section and Panel Data ”. Cambridge, MA: MIT Press.