

Né Attribué par la bibliothèque



--	--	--	--	--	--	--	--	--	--

Année univ.: 2019/2020



République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

## La régression linéaire simple et multiple

Mémoire présenté en vue de l'obtention du diplôme de

Master Académique

Université de Saida - Dr Moulay Tahar

Discipline : MATHÉMATIQUES

Spécialité : Analyse stochastique, statistique des processus et applications  
(ASSPA)

par

**Fatima Becharef<sup>1</sup>**

Sous la direction de

Soutenue le 16/09/2020 devant le jury composé de

**Dr S.Soltani**

Mr. Lahcene Yahyaoui	Université Dr. Moulay Tahar. Saïda	Président
Mme. Sara Soltani	Université Dr. Moulay Tahar. Saïda	Encadreur
Melle. Saâdia Rahmani	Université Dr. Moulay Tahar. Saïda	Examinatrice
Melle. Lamia Bousmaha	Université Dr. Moulay Tahar. Saïda	Examinatrice

---

1. e-mail : bechareffatima96@gmail.com

# *REMERCIEMENTS*

Je souhaite avant tout remercier Dieu le tout puissant de m'avoir donné la volonté, le courage et la patience de réaliser mon objectif.

Mes remerciements les plus sincères vont tout d'abord à mon encadreur de mémoire *Mme S.Soltani* pour son suivi, son aide et ses conseils. je tiens à lui exprimer ma gratitude .

Je tiens également à remercier les membres du jury, Docteur *L. Yahyaoui*, Docteur *S.Rahmani* et Docteur *L.Bousmaha* qui ont bien voulu accepter de porter leur jugement sur ce modeste travail.

Je remercie aussi tous mes professeurs pour son aide durant toutes ces années.

Et un grand remerceiments à mes très chers parents qui ont toujours veillé à mon bien-être et qui ont fait de moi ce qu'on est aujourd'hui. Et à ma chère soeur Malak, et mes frères Hicham et Ali.

Mes derniers remerciements, mais non les moindres s'adresse à tous mes amies, et à tous ceux qui ont contribué de loin ou de près à la réalisation de ce travail.

à tous, merci !

# Table des matières

<b>1</b>	<b>La régression linéaire simple et multiple</b>	<b>8</b>
1.1	Le modèle de régression linéaire simple . . . . .	8
1.1.1	Définition . . . . .	8
1.1.2	Hypothèses du modèle . . . . .	8
1.1.3	Estimation des paramètres par la méthode des Moindres Carrés Ordinaires (MCO) . . . . .	9
1.1.4	Calcul des espérances mathématiques des estimateurs . . . . .	12
1.1.5	Estimation de la variance des erreurs . . . . .	29
1.1.6	Analyse de la variance et le coefficient de détermination . . . . .	31
1.2	Les modèles de régression linéaire multiple . . . . .	34
1.2.1	Présentation . . . . .	34
1.2.2	Hypothèses du modèle . . . . .	35
1.2.3	Estimation et propriétés des estimateurs . . . . .	35
1.2.4	Estimateur de la variance de l'erreur et la matrice de variance covariance des coefficients de régression . . . . .	37
1.2.5	Equation d'analyse de la variance et qualité d'un ajustement . . . . .	38
<b>2</b>	<b>Tests et intervalles de confiance</b>	<b>40</b>
2.1	Cas de modèle linéaire simple . . . . .	40
2.1.1	Test de nullité de $\hat{a}$ . . . . .	40

2.1.2	Test de nullité de $\hat{b}$ . . . . .	42
2.1.3	Analyse de la variance et test de Fisher . . . . .	45
2.1.4	Prévision à l'aide d'un modèle de régression simple . . . . .	47
2.2	Cas de modèle linéaire multiple . . . . .	48
2.2.1	Le test de student . . . . .	48
2.2.2	Test de Fisher (test de signification globale du modèle de régression ) . . .	49
2.2.3	La prévision dans le modèle de la régression multiple . . . . .	51
<b>3</b>	<b>Application et conclusion</b>	<b>52</b>
3.1	Exemple d'application . . . . .	52
3.2	Conclusion . . . . .	64
	<b>Bibliographie</b>	<b>65</b>

# Introduction générale

En statistiques, en économétrie et en apprentissage automatique, un modèle de régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.

On parle aussi de modèle linéaire ou de modèle de régression linéaire.

Parmi les modèles de régression linéaire, le plus simple est l'ajustement affine. Celui-ci consiste à rechercher la droite permettant d'expliquer le comportement d'une variable statistique  $Y$  comme étant une fonction affine d'une autre variable statistique  $X$ .

En général, le modèle de régression linéaire désigne un modèle dans lequel l'espérance conditionnelle de  $Y$  connaissant  $X$  est une fonction affine des paramètres. Cependant, on peut aussi considérer des modèles dans lesquels c'est la médiane conditionnelle de  $Y$  connaissant  $X$  ou n'importe quel quantile de la distribution de  $Y$  connaissant  $X$  qui est une fonction affine des paramètres.

Le modèle de régression linéaire est souvent estimé par la méthode des moindres carrés mais il existe aussi de nombreuses autres méthodes pour estimer ce modèle. On peut par exemple estimer le modèle par maximum de vraisemblance ou encore par inférence bayésienne. Bien qu'ils soient souvent présentés ensemble, le modèle linéaire et la méthode des moindres carrés ne désignent pas la même chose. Le modèle linéaire désigne une classe de modèles qui peuvent être estimés par un grand nombre de méthodes, et la méthode des moindres carrés désigne une méthode d'estimation. Elle peut être utilisée pour estimer différents types de modèles.

# Historique

*Ruder Josip Boskovic* [5] est le premier scientifique à calculer les coefficients de régression linéaire, en 1755-1757, quand il entreprit de mesurer la longueur de cinq méridiens terrestres en minimisant la somme des valeurs absolues.

*Pierre-Simon de Laplace* [5] utilise cette méthode pour mesurer les méridiens dans ( Sur les degrés mesurés des méridiens et sur les longueurs observées sur pendule ) en 1789.

La première utilisation de la méthode des moindres carrés est attribuée à *Adrien-Marie Legendre*[1] en 1805 ou à *Carl Friedrich Gauss* [5] qui dit l'avoir utilisée à partir de 1795.

*Carl Friedrich Gauss* démontre, en 1821, le théorème connu aujourd'hui sous le nom de théorème de *Gauss-Markov* [6] qui exprime sous certaines conditions la qualité des estimateurs, *Andrei Markov* [6] le redécouvre en 1900 .

C'est à *Francis Galton* [5] qu'est accordée la paternité de l'expression ( régression linéaire) en 1886 . Dans son article, *Galton* exprime la taille des fils en fonction de la taille des pères. Il constate un phénomène de ( régression vers la moyenne ) .

Plus tard la colinéarité des variables explicatives est devenue un sujet de recherche important. En 1970, *Arthur E. Hoerl* [10] et *Robert W. Kennard* [10] proposent la régression pseudo-orthogonale (Ridge Regression), une des méthodes d'estimation conçues pour pallier la présence de colinéarité de certaines variables explicatives en imposant des contraintes sur les coefficients.

*La méthode du lasso* (Lasso Regression), ayant le même objectif en utilisant une technique analogue, a été créée en 1996 par *Robert Tibshirani* [12].

# Organisation du mémoire

Dans le premier chapitre, on étudie le modèle de la régression linéaire simple et multiple, puis on estime leurs paramètres par la méthode des moindres carrés et on va donner des estimateurs de quelques paramètres statistiques.

Dans le deuxième chapitre on va faire des tests sur les paramètres de la régression linéaire.

Et on termine par un exemple d'application .

# Chapitre 1

## La régression linéaire simple et multiple

### 1.1 Le modèle de régression linéaire simple

#### 1.1.1 Définition

Le modèle de régression linéaire simple est une variable endogène (dépendante) expliquée par une seule variable exogène (indépendante) mise sous forme mathématique suivante :

$$Y_t = aX_t + b + \varepsilon_t, \quad t = 1 \dots n$$

avec :

$Y_t$  : la variable endogène (dépendante, à expliquer) à la date  $t$  ;

$X_t$  : la variable exogène (indépendante, explicative) à la date  $t$  ;

$a, b$  : sont deux paramètres à estimer ;

$\varepsilon_t$  : l'erreur aléatoire du modèle ;

$n$  : nombre d'observations.

#### 1.1.2 Hypothèses du modèle

Le modèle repose sur les hypothèses suivantes :

(H1)  $\mathbb{E}(\varepsilon_t) = 0$  , l'erreur centrée ;



- (H2)  $\mathbb{E}(\varepsilon_t^2) = \sigma_\varepsilon^2$ , la variance de l'erreur est constante ;
- (H3)  $\text{cov}(\varepsilon_t, \varepsilon_{t'}) = 0$ , si  $\varepsilon_t \neq \varepsilon_{t'}$ , les erreurs sont indépendantes ;
- (H4) La normalité des erreurs,  $\varepsilon_t \rightarrow \mathcal{N}(0, \sigma_\varepsilon^2)$  ;
- (H5)  $\text{cov}(x_t, \varepsilon_t) = 0$ , l'erreur est indépendante de la variable exogène ;
- (H6) La variable exogène  $X_t$  n'est pas aléatoire ;
- (H7) Le modèle est linéaire en  $X$  par rapport aux paramètres.

### 1.1.3 Estimation des paramètres par la méthode des Moindres Carrés Ordinaux (MCO)

Soit le modèle suivant :

$$Y_t = aX_t + b + \varepsilon_t$$

Le principe des moindres carrés consiste à rechercher les valeurs des paramètres qui minimisent la somme des carrés des résidus :

$$\min_{(a,b) \in \mathbb{R}^2} \sum_{t=1}^n \varepsilon_t^2 = \min_{(a,b) \in \mathbb{R}^2} \sum_{t=1}^n \left( Y_t - \hat{Y}_t \right)^2 = \min_{(a,b) \in \mathbb{R}^2} \sum_{t=1}^n \left( Y_t - aX_t - b \right)^2 = \min_{(a,b) \in \mathbb{R}^2} \sum_{t=1}^n F^2$$

Pour que cette fonction ait un minimum, il faut que les dérivées par-rapport à  $a$  et  $b$  soient nuls.

$$\begin{cases} \frac{\partial F}{\partial a} = 0 \Leftrightarrow 2 \sum_{t=1}^n \left( Y_t - aX_t - b \right) (-X_t) = 0; \\ \frac{\partial F}{\partial b} = 0 \Leftrightarrow 2 \sum_{t=1}^n \left( Y_t - aX_t - b \right) (-1) = 0. \end{cases}$$

D'après les deux équations, on obtient :

$$\sum_{t=1}^n Y_t X_t = a \sum_{t=1}^n X_t^2 + b \sum_{t=1}^n X_t \quad (1.1)$$

$$\sum_{t=1}^n Y_t = a \sum_{t=1}^n X_t + nb \quad (1.2)$$

En notans  $\hat{a}$  et  $\hat{b}$  les solutions des équations (1.1) et (1.2), d'après (1.2) on obtient :

$$\hat{b} = \frac{\sum_{t=1}^n Y_t}{n} - \hat{a} \frac{\sum_{t=1}^n X_t}{n}$$

Donc

$$\hat{b} = \bar{Y} - \hat{a}\bar{X}$$

$$\text{puisque } \left( \bar{Y} = \frac{\sum_{i=1}^n Y_t}{n} \right) \text{ et } \left( \bar{X} = \frac{\sum_{i=1}^n X_t}{n} \right).$$

En remplaçant la valeur de  $\hat{b}$  dans l'équation (1.1), on obtient :

$$\begin{aligned} \sum_{t=1}^n Y_t X_t &= \left( \bar{Y} - \hat{a}\bar{X} \right) \sum_{t=1}^n X_t + \hat{a} \sum_{t=1}^n X_t^2 \Leftrightarrow \sum_{t=1}^n Y_t X_t = \bar{Y} \sum_{t=1}^n X_t - \hat{a}\bar{X} \sum_{t=1}^n X_t + \hat{a} \sum_{t=1}^n X_t^2 \\ &\Leftrightarrow \sum_{t=1}^n Y_t X_t - \bar{Y} \sum_{t=1}^n X_t = \hat{a} \left( \sum_{t=1}^n X_t^2 - \bar{X} \sum_{t=1}^n X_t \right) \end{aligned}$$

D'où

$$\begin{aligned} \hat{a} &= \frac{\sum_{t=1}^n X_t Y_t - \bar{Y} \sum_{t=1}^n X_t}{\sum_{t=1}^n X_t^2 - \bar{X} \sum_{t=1}^n X_t} \\ &= \frac{\sum_{t=1}^n X_t Y_t - n\bar{Y} \left( \frac{\sum_{t=1}^n X_t}{n} \right)}{\sum_{t=1}^n X_t^2 - n\bar{X} \left( \frac{\sum_{t=1}^n X_t}{n} \right)} \end{aligned}$$

$$\begin{aligned}
&= \frac{\sum_{t=1}^n X_t Y_t - n \bar{X} \bar{Y}}{\sum_{t=1}^n X_t^2 - n \bar{X}^2} \\
&= \frac{\frac{1}{n} \sum_{t=1}^n (Y_t - \bar{Y})(X_t - \bar{X})}{\frac{1}{n} \sum_{t=1}^n (X_t - \bar{X})^2} \\
&= \frac{cov(X, Y)}{var(X)}
\end{aligned}$$

**Conclusion :**

les estimateurs des *MCO* du modèle de régression linéaire simple

$Y_t = b + aX_t + \varepsilon_t$  sont :

$$\hat{a} = \frac{\sum_{t=1}^n (Y_t - \bar{Y})(X_t - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2} = \frac{cov(X, Y)}{var(X)}$$

et  $\hat{b} = \bar{Y} - \hat{a}\bar{X}$

**Différentes écritures du modèle de régression linéaire simple :**

Le modèle théorique ( modèle non ajusté) :

$$Y_t = aX_t + b + \varepsilon_t$$

Le modèle estimé ( modèle ajusté) :

$$Y_t = \hat{a}X_t + \hat{b} + e_t$$

Avec :

$$\hat{Y}_t = \hat{a}X_t + \hat{b}$$

$$\text{et } e_t = Y_t - \hat{Y}_t = Y_t - \hat{a}X_t - \hat{b}$$

$e_t$  : est le résidu du modèle.

#### 1.1.4 Calcul des espérances mathématiques des estimateurs

##### ► Calcul de l'espérance de $\hat{a}$

Soit le modèle suivant :  $Y_t = \hat{a}X_t + \hat{b} + e_t$

D'après la méthode des *MCO*, on a :

$$\hat{a} = \frac{\sum_{t=1}^n (Y_t - \bar{Y})(X_t - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2}$$

En posant

$$x_t = X_t - \bar{X} \text{ et } y_t = Y_t - \bar{Y}$$

Nous obtenons

$$\hat{a} = \frac{\sum_{t=1}^n x_t y_t}{\sum_{t=1}^n x_t^2} \quad (1.3)$$

On remplace la valeur  $y_t$  dans (1.3), on obtient :

$$\hat{a} = \frac{\sum_{t=1}^n x_t (Y_t - \bar{Y})}{\sum_{t=1}^n x_t^2} = \frac{\sum_{t=1}^n x_t Y_t}{\sum_{t=1}^n x_t^2} - \bar{Y} \frac{\sum_{t=1}^n x_t}{\sum_{t=1}^n x_t^2}$$

Comme  $\sum_{t=1}^n x_t = 0$

alors

$$\hat{a} = \frac{\sum_{t=1}^n x_t Y_t}{\sum_{t=1}^n x_t^2} \quad (1.4)$$

car

$$\begin{aligned} \sum_{t=1}^n x_t &= \sum_{t=1}^n (X_t - \bar{X}) \\ &= \sum_{t=1}^n X_t - \sum_{t=1}^n \bar{X} \\ &= n \left( \frac{\sum_{t=1}^n X_t}{n} \right) - n\bar{X} \\ &= n\bar{X} - n\bar{X} = 0 \end{aligned}$$

On remplace maintenant  $Y_t = aX_t + b + \varepsilon_t$  dans l'équation (1.4), on aura :

$$\begin{aligned} \hat{a} &= \frac{\sum_{t=1}^n x_t (aX_t + b + \varepsilon_t)}{\sum_{t=1}^n x_t^2} \\ &= \frac{b \sum_{t=1}^n x_t + a \sum_{t=1}^n x_t X_t + \sum_{t=1}^n x_t \varepsilon_t}{\sum_{t=1}^n x_t^2} \\ &= \frac{a \sum_{t=1}^n x_t X_t + \sum_{t=1}^n x_t \varepsilon_t}{\sum_{t=1}^n x_t^2} \end{aligned}$$

$$\begin{aligned}
&= \frac{a \sum_{t=1}^n x_t (x_t + \bar{X})}{\sum_{t=1}^n x_t^2} + \frac{\sum_{t=1}^n x_t \varepsilon_t}{\sum_{t=1}^n x_t^2} \\
&= \frac{a \sum_{t=1}^n x_t^2}{\sum_{t=1}^n x_t^2} + \frac{\bar{X} a \sum_{t=1}^n x_t}{\sum_{t=1}^n x_t^2} + \frac{\sum_{t=1}^n x_t \varepsilon_t}{\sum_{t=1}^n x_t^2}
\end{aligned}$$

car  $X_t = x_t + \bar{X}$ .

Comme  $\sum_{t=1}^n x_t = 0$  (on l'a déjà démontré), il résulte alors :

$$\hat{a} = a + \frac{\sum_{t=1}^n x_t \varepsilon_t}{\sum_{t=1}^n x_t^2}$$

En passant à l'espérance mathématique, on trouve :

$$\begin{aligned}
\mathbb{E}(\hat{a}) &= \mathbb{E}(a) + \mathbb{E} \left( \frac{\sum_{t=1}^n x_t \varepsilon_t}{\sum_{t=1}^n x_t^2} \right) \\
&= \mathbb{E}(a) + \left( \frac{\sum_{t=1}^n x_t \mathbb{E}(\varepsilon_t)}{\sum_{t=1}^n x_t^2} \right)
\end{aligned}$$

Or, d'après l'hypothèse **(H1)**,

$$\mathbb{E}(\varepsilon_t) = 0$$

Finalement :

$$\mathbb{E}(\hat{a}) = a$$

$a$  est un estimateur sans biais.

► Calcul de l'espérance de  $\hat{b}$

On a :

$$\begin{aligned}\hat{b} &= \bar{Y} - \hat{a}\bar{X} \\ \hat{b} &= \bar{Y} - \bar{X} \left( \frac{\sum_{t=1}^n x_t y_t}{\sum_{t=1}^n x_t^2} \right) \\ \hat{b} &= \bar{Y} - \bar{X} \left( \frac{\sum_{t=1}^n x_t}{\sum_{t=1}^n x_t^2} \right) (Y_t - \bar{Y}) \\ \hat{b} &= \bar{Y} - \bar{X} \left( \frac{\sum_{t=1}^n x_t Y_t}{\sum_{t=1}^n x_t^2} \right)\end{aligned}$$

D'où

$$\hat{b} = \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X} x_t}{\sum_{t=1}^n x_t^2} \right) Y_t \quad (1.5)$$

$$\text{Car} \left( \bar{Y} = \frac{\sum_{t=1}^n Y_t}{n} \right).$$

Or que :

$$Y_t = aX_t + b + \varepsilon_t$$

donc :

$$\hat{b} = \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) (aX_t + b + \varepsilon_t)$$

$$\hat{b} = \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) aX_t + \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) b + \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t$$

$$\hat{b} = a \left( \frac{\sum_{t=1}^n X_t}{n} \right) - a\bar{X} \left( \frac{\sum_{t=1}^n x_t X_t}{\sum_{t=1}^n x_t^2} \right) + \frac{\sum_{t=1}^n b}{n} - b\bar{X} \left( \frac{\sum_{t=1}^n x_t}{\sum_{t=1}^n x_t^2} \right) + \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t$$

$$\hat{b} = a\bar{X} - a\bar{X} \left( \frac{\sum_{t=1}^n x_t X_t}{\sum_{t=1}^n x_t^2} \right) + b + \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t$$

Comme  $X_t = x_t + \bar{X}$ , on déduit :

$$\hat{b} = a\bar{X} - a\bar{X} \left( \frac{\sum_{t=1}^n x_t(x_t + \bar{X})}{\sum_{t=1}^n x_t^2} \right) + b + \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t$$

$$\hat{b} = a\bar{X} - a\bar{X} \left( \frac{\sum_{t=1}^n x_t^2}{\sum_{t=1}^n x_t^2} \right) - a\bar{X}^2 \left( \frac{\sum_{t=1}^n x_t}{\sum_{t=1}^n x_t^2} \right) + b + \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t$$



On obtient alors :

$$\hat{b} = b + \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t$$

En passant à l'espérance mathématique, on trouve :

$$\begin{aligned} \mathbb{E}(\hat{b}) &= \mathbb{E}(b) + \mathbb{E} \left[ \left( \sum_{t=1}^n \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t \right] \\ \mathbb{E}(\hat{b}) &= \mathbb{E}(b) \left[ \left( \sum_{t=1}^n \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \mathbb{E}(\varepsilon_t) \right] \end{aligned}$$

$$\mathbb{E}(\hat{b}) = \mathbb{E}(b) \text{ car } \mathbb{E}(\varepsilon_t) = 0$$

$$\text{Finalement } \mathbb{E}(\hat{b}) = b$$

$\beta_0$  est un estimateur sans biais.

► **Calcul de la variance de  $\hat{a}$**

Par définition, la variance de  $(\hat{a})$  est donnée par :

$$var(\hat{a}) = \mathbb{E} \left[ \hat{a} - \mathbb{E}(\hat{a}) \right]^2$$

$$\text{Et } \mathbb{E}(\hat{a}) = \hat{a}.$$

d'un autre coté, on sait que :

$$\hat{a} = a + \frac{\sum_{t=1}^n x_t \varepsilon_t}{\sum_{t=1}^n x_t^2}$$

Ce qui implique :

$$\hat{a} - a = \frac{\sum_{t=1}^n x_t \varepsilon_t}{\sum_{t=1}^n x_t^2}$$

Alors on déduit que :

$$\begin{aligned} \text{var}(\hat{a}) &= \mathbb{E}(\hat{a} - a)^2 = \mathbb{E} \left( \frac{\sum_{t=1}^n x_t \varepsilon_t}{\sum_{t=1}^n x_t^2} \right)^2 = \frac{1}{\left( \sum_{t=1}^n x_t^2 \right)^2} \mathbb{E} \left( \sum_{t=1}^n x_t \varepsilon_t \right)^2 \\ &= \frac{1}{\left( \sum_{t=1}^n x_t^2 \right)^2} \mathbb{E} \left[ x_1^2 \varepsilon_1^2 + x_2^2 \varepsilon_2^2 + \dots + x_n^2 \varepsilon_n^2 + 2x_1 x_2 \varepsilon_1 \varepsilon_2 + \dots + 2x_{n-1} x_n \varepsilon_{n-1} \varepsilon_n \right] \end{aligned}$$

D'après les hypothèses **(H1)**, **(H2)** et **(H3)** du modèle de régression simple, on obtient :

$$\text{var}(\hat{a}) = \frac{1}{\left( \sum_{t=1}^n x_t^2 \right)^2} \left( \sigma_\varepsilon^2 x_1^2 + \sigma_\varepsilon^2 x_2^2 + \sigma_\varepsilon^2 x_3^2 + \dots + \sigma_\varepsilon^2 x_n^2 \right)$$

D'où :

$$\begin{aligned} \text{var}(\hat{a}) &= \frac{\sigma_\varepsilon^2 \sum_{t=1}^n x_t^2}{\left( \sum_{t=1}^n x_t^2 \right)^2} \\ \text{var}(\hat{a}) &= \frac{\sigma_\varepsilon^2}{\sum_{t=1}^n x_t^2} \\ \text{var}(\hat{a}) &= \frac{\sigma_\varepsilon^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \end{aligned}$$

► **Calcul de la variance de  $\hat{b}$**

D'après les propriétés de l'estimateur  $\hat{b}$  on a :

$$\hat{b} = b + \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t \Rightarrow \hat{b} - b = \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t$$

Par définition, la variance de  $\hat{b}$  est donnée par :  $var(\hat{b}) = \mathbb{E} \left[ \hat{b} - \mathbb{E}(\hat{b}) \right]^2 = \mathbb{E}(\hat{b} - b)^2$

Puisque  $\mathbb{E}(\hat{b}) = b$  nous obtenons :

$$var(\hat{b}) = \mathbb{E} \left( \hat{b} - b \right)^2$$

Alors on déduit que :

$$\begin{aligned} var(\hat{b}) &= \mathbb{E} \left[ \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t \right]^2 \\ var(\hat{b}) &= \mathbb{E} \left[ \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t \right]^2 + 2\mathbb{E} \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_{n-1}}{\sum_{t=1}^n x_{n-1}^2} \right) \left( \frac{1}{n} - \frac{\bar{X}x_n}{\sum_{t=1}^n x_n^2} \right) \varepsilon_{n-1} \varepsilon_n \\ var(\hat{b}) &= \mathbb{E} \left[ \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t \right]^2 \end{aligned}$$

D'après les hypothèses (1),(2),et(3) du modèle de régression simple, on obtient :

$$var(\hat{b}) = \sigma_\varepsilon^2 \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^n x_t^2} \right)^2$$

Il résulte alors :

$$var(\hat{b}) = \sigma_\varepsilon^2 \sum_{t=1}^n \left( \frac{1}{n^2} - \frac{2\bar{X}x_t}{n \sum_{t=1}^n x_t^2} + \frac{\bar{X}^2 x_t^2}{\left( \sum_{t=1}^n x_t^2 \right)^2} \right)$$

Puisque  $\sum_{t=1}^n x_t = 0$  alors :

$$var(\hat{b}) = \sigma_\varepsilon^2 \left( \frac{1}{n} + \frac{\bar{X}^2 \sum_{t=1}^n x_t^2}{\left( \sum_{t=1}^n x_t^2 \right)^2} \right) \quad (1.6)$$

$$var(\hat{b}) = \sigma_\varepsilon^2 \left( \frac{\sum_{t=1}^n x_t^2 + n\bar{X}^2}{n \sum_{t=1}^n x_t^2} \right) \quad (1.7)$$

Et comme  $X_t = x_t + \bar{X}$  donc on déduit que :

$$\begin{aligned} var(\hat{b}) &= \sigma_\varepsilon^2 \left( \frac{\sum_{t=1}^n x_t^2 + \sum_{t=1}^n \bar{X}^2}{n \sum_{t=1}^n x_t^2} \right) \\ &= \sigma_\varepsilon^2 \left( \frac{\sum_{t=1}^n (x_t^2 + \bar{X}^2)}{n \sum_{t=1}^n x_t^2} \right) \\ &= \sigma_\varepsilon^2 \left( \frac{\sum_{t=1}^n X_t^2}{n \sum_{t=1}^n (X_t - \bar{X})^2} \right) \end{aligned}$$

### Conclusion

Les variances des paramètres  $\hat{a}$  et  $\hat{b}$  du modèle  $Y_t = aX_t + b + \varepsilon_t$  sont :

$$var(\hat{a}) = \frac{\sigma_\varepsilon^2}{\sum_{t=1}^n (X_t - \bar{X})^2}$$

$$\text{et } var(\hat{b}) = \sigma_\varepsilon^2 \left( \frac{\sum_{t=1}^n X_t^2}{n \sum_{t=1}^n (X_t - \bar{X})^2} \right)$$

### ► Calcul de la covariance de $(\hat{a}, \hat{b})$ :

Par définition, la covariance entre  $\hat{a}$  et  $\hat{b}$  se calcule comme suit :

$$cov(\hat{a}, \hat{b}) = \mathbb{E} \left[ \left( \hat{a} - \mathbb{E}(\hat{a}) \right) \left( \hat{b} - \mathbb{E}(\hat{b}) \right) \right] = \mathbb{E} \left[ (\hat{a} - a)(\hat{b} - b) \right]$$

Comme  $\hat{a}$  et  $\hat{b}$  sont sans biais,

Alors :

$$cov(\hat{a}, \hat{b}) = \mathbb{E} \left[ \frac{\sum_{t=1}^n x_t \varepsilon_t}{\sum_{t=1}^n x_t^2} * \sum_{t=1}^n \left( \frac{1}{n} - \frac{\bar{X} x_t}{\sum_{t=1}^n x_t^2} \right) \varepsilon_t \right]$$

$$cov(\hat{a}, \hat{b}) = \mathbb{E} \left[ \frac{\sum_{t=1}^n \varepsilon_t \sum_{t=1}^n x_t \varepsilon_t}{n \sum_{t=1}^n x_t^2} - \frac{\bar{X} \sum_{t=1}^n x_t \varepsilon_t \sum_{t=1}^n x_t \varepsilon_t}{\left( \sum_{t=1}^n x_t^2 \right)^2} \right]$$

$$cov(\hat{a}, \hat{b}) = \mathbb{E} \left[ \frac{\sum_{t=1}^n \varepsilon_t \sum_{t=1}^n x_t \varepsilon_t}{n \sum_{t=1}^n x_t^2} \right] - \bar{X} \mathbb{E} \left[ \frac{\left( \sum_{t=1}^n x_t \varepsilon_t \right)^2}{\left( \sum_{t=1}^n x_t^2 \right)^2} \right]$$

D'après les hypothèses **(H1)** ,**(H2)** et **(H3)** du modèle de régression simple, il résulte que :

$$cov(\hat{a}, \hat{b}) = -\frac{\bar{X}\sigma_\varepsilon^2}{\sum_{t=1}^n x_t^2}$$

► **La convergence du  $\hat{a}$  et  $\hat{b}$**

$$\text{On a : } var(\hat{a}) = \frac{\sigma_\varepsilon^2}{\sum_{t=1}^n (X_t - \bar{X})^2}$$

lorsque  $n \rightarrow +\infty$

$$\text{alors } \sum_{t=1}^n (X_t - \bar{X})^2 \rightarrow +\infty$$

et  $var(\hat{a}) \rightarrow 0$

On déduit que  $\hat{a}$  est convergent.

De même, pour  $\hat{b}$

D'après l'équation (1.6), on a :

$$var(\hat{b}) = \sigma_\varepsilon^2 \left( \frac{1}{n} + \frac{\bar{X}^2 \sum_{t=1}^n x_t^2}{\left( \sum_{t=1}^n x_t^2 \right)^2} \right)$$

Et comme  $x_t = X_t - \bar{X}$  alors :

$$var(\hat{b}) = \sigma_\varepsilon^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \right)$$

lorsque  $n \rightarrow +\infty$

$$\text{alors } \sum_{t=1}^n (X_t - \bar{X})^2 \rightarrow +\infty$$

et  $var(\hat{b}) \rightarrow 0$

On déduit que  $\hat{b}$  est convergent.

**Théorème 1.1.1 (Gauss Markov)**

Soit le modèle suivant :  $Y_t = aX_t + b + \varepsilon_t$

un estimateur de moindres carrés est un estimateur de Gauss Markov (blue) s'il est sans biais, linéaire et possède une variance minimale.

**Preuve du Théorème (1.1.1)**

Pour démontrer ce théorème, on définit un autre estimateur linéaire sans biais sous la forme suivante :

$$\alpha = \sum_{t=1}^n A_t Y_t$$

Par la suite on compare la variance de  $\alpha$  avec la variance de  $\hat{a}$  et celui qui a une variance minimale on dira qu'il est le meilleur estimateur.

Démontrons d'abord est ce que  $\hat{a}$  est linéaire et sans biais.

**►  $\hat{a}$  est-il linéaire ?**

D'après les propriétés du paramètre  $\hat{a}$  on a :

$$\begin{aligned} \hat{a} &= \frac{\sum_{t=1}^n x_t y_t}{\sum_{t=1}^n x_t^2} \\ \hat{a} &= \frac{\sum_{t=1}^n x_t (Y_t - \bar{Y})}{\sum_{t=1}^n x_t^2} \\ \hat{a} &= \frac{\sum_{t=1}^n x_t Y_t}{\sum_{t=1}^n x_t^2} - \bar{Y} \frac{\sum_{t=1}^n x_t}{\sum_{t=1}^n x_t^2} \end{aligned}$$

Comme  $\sum_{t=1}^n x_t = 0$ , on obtient :

$$\hat{a} = \frac{\sum_{t=1}^n x_t Y_t}{\sum_{t=1}^n x_t^2}$$

En posant

$$V_t = \frac{x_t}{\sum_{t=1}^n x_t^2}$$

Alors  $\hat{a}$  s'écrit sous la forme suivante :

$$\sum_{t=1}^n A_t Y_t = \sum_{t=1}^n V_t Y_t$$

Ce qui fait que  $\hat{a}$  est linéaire.

►  **$\hat{a}$  est-il sans biais ?**

On a :

$$\mathbb{E}(\hat{a}) = \mathbb{E}\left[\sum_{t=1}^n V_t Y_t\right]$$

$$\mathbb{E}(\hat{a}) = \mathbb{E}\left[\sum_{t=1}^n V_t (aX_t + b + \varepsilon_t)\right]$$

$$\mathbb{E}(\hat{a}) = a\mathbb{E}\left[\sum_{t=1}^n V_t X_t\right] + b\mathbb{E}\left[\sum_{t=1}^n V_t\right] + \mathbb{E}\left[\sum_{t=1}^n V_t \varepsilon_t\right]$$

Connaissant que :

$$\sum_{t=1}^n V_t = 0, \sum_{t=1}^n x_t = 0 \text{ et } \mathbb{E}(\varepsilon_t) = 0$$

Alors :

$$\mathbb{E}(\hat{a}) = a\mathbb{E}\left[\sum_{t=1}^n V_t X_t\right]$$

Sachant que :



$$V_t = \frac{x_t}{\sum_{t=1}^n x_t^2}$$

Ce qui fait :

$$\mathbb{E}(\hat{a}) = a \mathbb{E} \left( \frac{\sum_{t=1}^n x_t X_t}{\sum_{t=1}^n x_t^2} \right) \quad (1.8)$$

Et comme  $x_t = X_t - \bar{X} \Rightarrow X_t = x_t + \bar{X}$

Nous remplaçons la valeur de  $X_t$  dans (1.8), nous obtenons :

$$\begin{aligned} \mathbb{E}(\hat{a}) &= a \mathbb{E} \left( \frac{\sum_{t=1}^n x_t X_t}{\sum_{t=1}^n x_t^2} \right) \\ \mathbb{E}(\hat{a}) &= a \mathbb{E} \left[ \frac{\sum_{t=1}^n x_t (x_t + \bar{X})}{\sum_{t=1}^n x_t^2} \right] \\ \mathbb{E}(\hat{a}) &= a \mathbb{E} \left( \frac{\sum_{t=1}^n x_t^2}{\sum_{t=1}^n x_t^2} \right) + a \mathbb{E} \left( \frac{\bar{X} \sum_{t=1}^n x_t}{\sum_{t=1}^n x_t^2} \right) \end{aligned}$$

Et puisque :  $\sum_{t=1}^n x_t = 0$

Finalement  $\mathbb{E}(\hat{a}) = a$  est un paramètre sans biais.

De ces deux démonstrations (linéaire et sans biais) on retient que :

$$\sum_{t=1}^n V_t = 0 \text{ et } \sum_{t=1}^n V_t X_t = 1$$

►  $\hat{a}$  possède-t-il une variance minimale ?

On suppose qu'il existe un autre estimateur sans biais linéaire défini comme suit :

$$\alpha = \sum_{t=1}^n A_t Y_t$$

Avec :  $\mathbb{E}(\alpha) = a$

Et  $A_t = V_t + M_t$

On a  $Y_t = aX_t + b + \varepsilon_t$

En passant à l'espérance mathématique :

$$\mathbb{E}(\alpha) = \mathbb{E} \left[ \sum_{t=1}^n A_t Y_t \right]$$

$$\mathbb{E}(\alpha) = \mathbb{E} \left[ \sum_{t=1}^n A_t (aX_t + b + \varepsilon_t) \right]$$

$$\mathbb{E}(\alpha) = a \mathbb{E} \left[ \sum_{t=1}^n A_t X_t \right] + b \mathbb{E} \left[ \sum_{t=1}^n A_t \right]$$

$$\mathbb{E}(\alpha) = a \tag{1.9}$$

$$\text{Car } \mathbb{E} \left[ \sum_{t=1}^n A_t \varepsilon_t \right] = 0$$

pour que l'équation (1.9) soit vérifiée c'est-à-dire  $\mathbb{E}(\alpha) = a$ , il faut que  $\mathbb{E}(\alpha) = a$  il faut que :

$$\sum_{t=1}^n A_t X_t = 1 \tag{1.10}$$

et

$$\sum_{t=1}^n A_t = 0 \tag{1.11}$$

$$\Leftrightarrow \begin{cases} \sum_{t=1}^n A_t X_t = 1 \Rightarrow \sum_{t=1}^n (V_t + M_t) X_t = 1 \Rightarrow \sum_{t=1}^n V_t X_t + \sum_{t=1}^n M_t X_t = 1 \\ \sum_{t=1}^n A_t = 0 \Rightarrow \sum_{t=1}^n (V_t + M_t) = 0 \Rightarrow \sum_{t=1}^n V_t = 0 \text{ et } \sum_{t=1}^n M_t = 0 \end{cases}$$

Maintenant, nous calculons la  $var(\alpha)$  :

Sous les conditions (1.10) et (1.11) et d'après la définition de la variance on a :

$$var(\alpha) = \mathbb{E} \left[ \alpha - \mathbb{E}(\alpha) \right]^2 = \mathbb{E} \left( \alpha - a \right)^2$$

D'un autre coté on a :

$$\alpha = \sum_{t=1}^n A_t Y_t = \sum_{t=1}^n A_t \left( a X_t + b + \varepsilon_t \right) = a \sum_{t=1}^n A_t X_t + b \sum_{t=1}^n A_t + \sum_{t=1}^n A_t \varepsilon_t$$

Sous l'hypothèse, que les conditions (1.10) et (1.11) soient vérifiées :

$$\alpha = a + \sum_{t=1}^n A_t \varepsilon_t \Rightarrow \alpha - a = \sum_{t=1}^n A_t \varepsilon_t$$

Nous obtenons :

$$\begin{aligned} var(\alpha) &= \mathbb{E} \left( \sum_{t=1}^n A_t \varepsilon_t \right)^2 \\ &= \mathbb{E} \left( A_1 \varepsilon_1 + A_2 \varepsilon_2 + \dots A_n \varepsilon_n \right)^2 \\ &= \mathbb{E} \left( A_1^2 \varepsilon_1^2 + A_2^2 \varepsilon_2^2 + \dots A_n^2 \varepsilon_n^2 + 2A_1 A_2 \varepsilon_1 \varepsilon_2 + \dots + 2A_{n-1} A_n \varepsilon_{n-1} \varepsilon_n \right) \\ &= \mathbb{E} \left( \sum_{t=1}^n A_t^2 \varepsilon_t^2 + 2 \sum_{t=1}^n \sum_{t'=1}^n A_t A_{t'} \varepsilon_t \varepsilon_{t'} \right) \end{aligned}$$

D'après les hypothèses du modèle de régression simple :

$$\mathbb{E}(\varepsilon_t^2) = \sigma_\varepsilon^2 \text{ et } \mathbb{E}(\varepsilon_t \varepsilon_{t'}) = 0$$

On déduit alors :

$$\begin{aligned}
 var(\alpha) &= \sigma_\varepsilon^2 \sum_{t=1}^n A_t^2 \\
 var(\alpha) &= \sigma_\varepsilon^2 \sum_{t=1}^n \left( V_t M_t \right)^2 \\
 var(\alpha) &= \sigma_\varepsilon^2 \sum_{t=1}^n \left( V_t^2 + M_t^2 + 2V_t M_t \right) \\
 var(\alpha) &= \sigma_\varepsilon^2 \sum_{t=1}^n V_t^2 + \sigma_\varepsilon^2 \sum_{t=1}^n M_t^2 + 2\sigma_\varepsilon^2 \sum_{t=1}^n V_t M_t
 \end{aligned}$$

Sachant que :

$$\sum_{t=1}^n V_t = 0 \text{ alors } \sum_{t=1}^n M_t + \sum_{t=1}^n V_t = 0 \Rightarrow \sum_{t=1}^n M_t = 0$$

Et

$$\sum_{t=1}^n V_t X_t = 1 \text{ alors } \sum_{t=1}^n M_t X_t + \sum_{t=1}^n V_t X_t = 1 \Rightarrow \sum_{t=1}^n M_t X_t = 0$$

On déduit :

$$\sum_{t=1}^n V_t M_t = \frac{\sum_{t=1}^n x_t M_t}{\sum_{t=1}^n x_t^2} = \frac{\sum_{t=1}^n X_t M_t}{\sum_{t=1}^n x_t^2} - \bar{X} \frac{\sum_{t=1}^n M_t}{\sum_{t=1}^n x_t^2} = 0$$

Ce qui résulte :

$$var(\alpha) = \sigma_\varepsilon^2 \sum_{t=1}^n V_t^2 + \sigma_\varepsilon^2 \sum_{t=1}^n M_t^2 \text{ car } \left( \sum_{t=1}^n V_t M_t = 0 \right)$$

Nous remplaçons  $V_t$  par  $V_t = \frac{x_t}{\sum_{t=1}^n x_t^2}$  on trouve donc :

$$var(\alpha) = \sigma_\varepsilon^2 \sum_{t=1}^n \left[ \frac{x_t^2}{\left( \sum_{t=1}^n x_t^2 \right)^2} \right] + \sigma_\varepsilon^2 \sum_{t=1}^n M_t^2$$

$$var(\alpha) = \sigma_\varepsilon^2 \frac{\sum_{t=1}^n x_t^2}{\sum_{t=1}^n (x_t^2)} + \sigma_\varepsilon^2 \sum_{t=1}^n M_t^2$$

$$var(\alpha) = \frac{\sigma_\varepsilon^2}{\sum_{t=1}^n x_t^2} + \sigma_\varepsilon^2 \sum_{t=1}^n M_t^2$$

Or que :

$$var(\hat{a}) = \frac{\sigma_\varepsilon^2}{\sum_{t=1}^n x_t^2}$$

Finalement :

$$var(\alpha) = var(\hat{a}) + \sigma_\varepsilon^2 \sum_{t=1}^n M_t^2$$

On remarque que  $var(\alpha) > var(\hat{a})$  puisque  $\sigma_\varepsilon^2 \sum_{t=1}^n M_t^2 > 0$ .

On conclut que le paramètre ( $\hat{a}$ ) a une variance minimale, ce qui fait qu'il est le meilleur estimateur (estimateur blue).

### Remarque

Même procédure pour le paramètre ( $\hat{b}$ ) :

On suppose  $\beta = \sum_{t=1}^n B_t Y_t$  avec  $B_t = W_t + N_t$  tel que  $W_t = \frac{1}{n} - \bar{X} V_t$

### 1.1.5 Estimation de la variance des erreurs

Soit le modèle de régression simple :  $Y_t = aX_t + b + \varepsilon_t$

Sachant que le résidu est :

$$e_t = Y_t - \hat{Y}_t$$

On a :  $\hat{Y}_t = \hat{a}X_t + \hat{b}$ ;

et  $\bar{Y} = a\bar{X} + b + \bar{\varepsilon}$

Alors :

$$e_t = aX_t + b + \varepsilon_t - \hat{a}X_t - \hat{b}$$

on remplace  $\hat{b}$  par sa valeur, on obtient :

$$e_t = aX_t + b + \varepsilon_t - \hat{a}X_t - \bar{Y} + \hat{a}\bar{X}$$

On remplace aussi  $\bar{Y}$  par sa valeur, on obtient :

$$e_t = aX_t + b + \varepsilon_t - \hat{a}X_t - a\bar{X} - b - \bar{\varepsilon} + \hat{a}\bar{X}$$

$$e_t = (a - \hat{a})X_t - (a - \hat{a})\bar{X} + \varepsilon_t - \bar{\varepsilon}$$

$$e_t = (a - \hat{a})(X_t - \bar{X}) + \varepsilon_t - \bar{\varepsilon}$$

$$e_t = (a - \hat{a})x_t + \varepsilon_t - \bar{\varepsilon}$$

car  $x_t = (X_t - \bar{X})$

D'où

$$\sum_{t=1}^n e_t^2 = \sum_{t=1}^n \left[ (a - \hat{a})x_t + (\varepsilon_t - \bar{\varepsilon}) \right]^2$$

$$\sum_{t=1}^n e_t^2 = \sum_{t=1}^n (\varepsilon_t - \bar{\varepsilon})^2 + (a - \hat{a})^2 \sum_{t=1}^n x_t^2 + 2(a - \hat{a}) \sum_{t=1}^n x_t (\varepsilon_t - \bar{\varepsilon})$$

On passant à l'espérance  $\mathbb{E}\left(\sum_{t=1}^n e_t^2\right)$  on obtient :

$$\mathbb{E}\left(\sum_{t=1}^n e_t^2\right) = (n-1)\sigma_\varepsilon^2 + \sigma_\varepsilon^2 - 2\sigma_\varepsilon^2 = (n-2)\sigma_\varepsilon^2$$

On déduit :

$$\sigma_\varepsilon^2 = \frac{\mathbb{E}\left(\sum_{t=1}^n e_t^2\right)}{n-2}$$

fnalement :

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{t=1}^n e_t^2}{n-2}$$

$\hat{\sigma}_\varepsilon^2$  est un estimateur sans biais.

### 1.1.6 Analyse de la variance et le coefficient de détermination

Pour calculer le coefficient de détermination, nous démontrons d'abord les deux relations :

1.  $\sum_{t=1}^n e_t = 0$ , La somme des résidus est nulle (la droite de régression passe par le point moyen cela est valable uniquement pour les modèles contenant le terme constant),
2.  $\sum_{t=1}^n Y_t = \sum_{t=1}^n \hat{Y}_t$ , légalité entre la moyenne de la série à expliquer et la moyenne de la série ajustée.

On démontre d'abord que :  $\sum_{t=1}^n e_t = \sum_{t=1}^n (Y_t - \hat{Y}_t) = 0$

On sait que :

$$Y_t = \hat{Y}_t + e_t = \hat{a}X_t + \hat{b} + e_t \Leftrightarrow \sum_{t=1}^n Y_t = \hat{a} \sum_{t=1}^n X_t + n\hat{b} + \sum_{t=1}^n e_t \Rightarrow \sum_{t=1}^n e_t = n\bar{Y} - \hat{a}n\bar{X} - n\hat{b}$$

On remplace  $\hat{b}$  par sa valeur on obtient alors :

$$\sum_{t=1}^n e_t = n\bar{Y} - \hat{a}n\bar{X} - n(\bar{Y} - \hat{a}\bar{X})$$

D'où :

$$\sum_{t=1}^n e_t = n\bar{Y} - \hat{a}n\bar{X} - n\bar{Y} + \hat{a}n\bar{X}$$

Donc :  $\sum_{t=1}^n e_t = 0$

Puisque  $\sum_{t=1}^n e_t = 0$  on déduit alors :

$$\sum_{t=1}^n e_t = \sum_{t=1}^n (Y_t - \hat{Y}_t) = 0 \Rightarrow \sum_{t=1}^n Y_t - \sum_{t=1}^n \hat{Y}_t = 0$$

On conclue :

$$\sum_{t=1}^n Y_t = \sum_{t=1}^n \hat{Y}_t \Rightarrow \bar{Y} = \bar{\hat{Y}}$$

A partir de ces deux équations nous pourrions déduire la fonction fondamentale d'analyse de la variance.

On a :

$$Y_t - \hat{Y}_t = e_t \Rightarrow Y_t = \hat{Y}_t + e_t$$

D'où :

$$Y_t - \bar{Y} = \hat{Y}_t + e_t - \bar{Y} \Rightarrow (Y_t - \bar{Y})^2 = (\hat{Y}_t - \bar{Y})^2 + e_t^2 + 2(\hat{Y}_t - \bar{Y})e_t$$

Passant aux sommes on trouve :

$$\sum_{t=1}^n (Y_t - \bar{Y})^2 = \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 + \sum_{t=1}^n e_t^2 + 2 \sum_{t=1}^n (\hat{Y}_t - \bar{Y})e_t$$

Comme :  $\sum_{t=1}^n e_t = 0$

et  $\sum_{t=1}^n Y_t = \sum_{t=1}^n \hat{Y}_t$

On déduit alors :

$$\sum_{t=1}^n (\hat{Y}_t - \bar{Y})e_t = 0$$

Il résulte :

$$\sum_{t=1}^n (Y_t - \bar{Y})^2 = \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 + \sum_{t=1}^n e_t^2$$



Qu'on peut écrire comme suit

$$\begin{aligned}\sum_{t=1}^n (Y_t - \bar{Y})^2 &= \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 + \sum_{t=1}^n e_t^2 \\ \Rightarrow SCT &= SCE + SCR\end{aligned}\tag{1.12}$$

l'équation (1.12) : appelée l'équation d'analyse de la variance.

Avec :

$$SCT = \sum_{t=1}^n (Y_t - \bar{Y})^2 : \text{désigne la variabilité totale ;}$$

$$SCE = \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 : \text{désigne la variabilité expliquée ;}$$

$$SCR = \sum_{t=1}^n e_t^2 = \sum_{t=1}^n (Y_t - \hat{Y}_t)^2 : \text{désigne la variabilité des résidus.}$$

### Coefficient de détermination

De l'équation (1.12) on peut déduire le coefficient de détermination

$$\begin{aligned}(1.12) \Rightarrow \frac{SCT}{SCT} &= \frac{SCE}{SCT} + \frac{SCR}{SCT} \\ \Rightarrow 1 &= \frac{SCE}{SCT} + \frac{SCR}{SCT}\end{aligned}$$

D'où

$$R^2 = 1 - \frac{SCR}{SCT} = \frac{SCE}{SCT}$$

$0 \leq R^2 \leq 1$ , plus la valeur de  $R^2$  est proche de 1, plus le modèle est plus significatif.

## 1.2 Les modèles de régression linéaire multiple

### 1.2.1 Présentation

Le modèle multiple est une généralisation du modèle simple dans lequel figurent plusieurs variables explicatives :

$$Y_t = a_0 + a_1X_{1t} + a_2X_{2t} + \dots + a_kX_{kt} + \varepsilon_t, \quad t = 1 \dots n$$

Avec :

$Y_t$  : variable à expliquer à la date  $t$  ;

$X_{1t}$  : variable explicative 1 à la date  $t$  ;

$\vdots$

$X_{kt}$  : variable explicative  $k$  à la date  $t$  ;

$a_0, a_1, \dots, a_k$  : les paramètres du modèle ;

$\varepsilon_t$  : l'erreur aléatoire du modèle ;

$n$  : nombre d'observations.

### La forme matricielle

Pour faciliter l'écriture de certains résultats, on a habituellement recours aux notations matricielles en écrivant le modèle observation par observation, nous obtenons :

$$Y_1 = a_0 + a_1x_{11} + a_2x_{21} + \dots + a_kx_{k1} + \varepsilon_1$$

$$Y_2 = a_0 + a_1x_{12} + a_2x_{22} + \dots + a_kx_{k2} + \varepsilon_2$$

$\vdots$

$$Y_t = a_0 + a_1x_{1t} + a_2x_{2t} + \dots + a_kx_{kt} + \varepsilon_t$$

$$Y_n = a_0 + a_1x_{1n} + a_2x_{2n} + \dots + a_kx_{kn} + \varepsilon_n$$

Soit sous la forme matricielle :

$$Y_{(n,1)} = X_{(n,k+1)}A_{(k+1,1)} + \varepsilon_{(n,1)}$$

Avec :

$$Y_{(n,1)} = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_t \\ Y_n \end{pmatrix}; A = \begin{pmatrix} a_0 \\ a_1 \\ \cdot \\ \cdot \\ a_t \\ a_n \end{pmatrix}; X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdot & \cdot & x_{k1} \\ 1 & x_{12} & x_{22} & \cdot & \cdot & x_{k2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{1t} & x_{2t} & \cdot & \cdot & x_{kt} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{1n} & x_{2n} & \cdot & \cdot & x_{kn} \end{pmatrix}; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_t \\ \varepsilon_n \end{pmatrix}$$

Avant d'estimer le modèle, on cite d'abord les hypothèses sur lesquelles il se repose.

### 1.2.2 Hypothèses du modèle

Le modèle repose sur les hypothèses suivantes :

- (H0) les valeurs  $x_{it}$  sont observées sans erreurs ;
- (H1)  $\mathbb{E}(\varepsilon_t) = 0$  , espérance nulle ;
- (H2)  $\mathbb{E}(\varepsilon_t^2) = \sigma_\varepsilon^2$ , la variance de l'erreur est constante  $\forall t$  ;
- (H3)  $\mathbb{E}(\varepsilon_t, \varepsilon_{t'}) = 0$  si  $t \neq t'$  ,indépendance des erreurs ;
- (H4)  $cov(x_{it}, \varepsilon_t) = 0$ , l'erreur indépendant des  $x_{it}$  ;
- (H5) absence de colinéarité entre les variables explicatives  $\Rightarrow (X'X)$  régulière et  $(X'X)^{-1}$  existe ;
- (H6)  $\left(\frac{X'X}{n}\right)$  tend vers une matrice finie non singulière ;
- (H7)  $n > k + 1$  nombre d'observations est supérieur aux nombre des séries explicatives.

### 1.2.3 Estimation et propriétés des estimateurs

#### ► Estimation des coefficients de régression :

Soit le modèle :

$$Y_t = X_t A + \varepsilon_t \tag{1.13}$$

Afin d'estimer le vecteur  $(A)$  composé des coefficients  $a_0, a_1, \dots, a_k$  nous appliquons la méthode des moindres carrés ordinaire (MCO) qui consiste à minimiser la somme des carrés des erreurs,

soit :

$$\min_{(a_0, \dots, a_k) \in \mathbb{R}^{k+1}} \sum_{t=1}^n \varepsilon_t^2 = \min_{(a_0, \dots, a_k) \in \mathbb{R}^{k+1}} \varepsilon' \varepsilon = \min \left( Y - X \hat{A} \right)' \left( Y - X \hat{A} \right) = \min_{(a_0, \dots, a_k) \in \mathbb{R}^{k+1}} F \quad (1.14)$$

Avec  $\varepsilon'$  : est le transposé du vecteur  $\varepsilon$ .

Pour minimiser cette fonction par rapport au vecteur  $(A)$  nous différencions  $F$  par rapport au même vecteur et on obtient :

$$\frac{\partial F}{\partial A} = -2X'Y + 2X'X\hat{A} = 0 \Rightarrow \hat{A} = (X'X)^{-1}X'Y$$

Avec  $(X'X)$  matrice de dimension  $(k+1, k+1)$  est inversible.

Le modèle estimé s'écrit :

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 x_{1t} + \hat{a}_2 x_{2t} + \dots + \hat{a}_k x_{kt} + e_t$$

Avec :  $e_t = Y_t - \hat{Y}_t$

où  $e_t$  : résidu, est l'écart entre la valeur observée de la variable à expliquée et sa valeur estimée, elle est connue.

## Conclusion

l'estimateur de *MCO* du modèle de régression linéaire multiple est :

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

### ► Propriétés des estimateurs :

#### • Estimateur sans biais :

Soit le modèle  $Y = XA + \varepsilon$

On peut s'écrire :

$$\begin{cases} Y = X\hat{A} + e \\ \hat{Y} = X\hat{A} \end{cases} \Rightarrow e = Y - \hat{Y}$$

Nous obtenons :

$$\begin{aligned} \hat{A} &= (X'X)^{-1}X'Y \\ \hat{A} &= (X'X)^{-1}X'(XA + \varepsilon) \end{aligned}$$

$$\begin{aligned}
\hat{A} &= (X'X)^{-1}X'(XA) + (X'X)^{-1}X'\varepsilon \\
\hat{A} &= A + (X'X)^{-1}X'\varepsilon \\
\hat{A} - A &= (X'X)^{-1}X'\varepsilon
\end{aligned} \tag{1.15}$$

D'où

$$\mathbb{E}(\hat{A}) = A + (X'X)^{-1}X'\mathbb{E}(\varepsilon)$$

avec :  $\mathbb{E}(\varepsilon) = 0$

Finalement

$\mathbb{E}(\hat{A}) = A \Rightarrow$  l'estimateur est sans biais.

#### 1.2.4 Estimateur de la variance de l'erreur et la matrice de variance covariance des coefficients de régression

On a :

$$var(\hat{A}) = \mathbb{E}[(\hat{A} - A)(\hat{A} - A)'] \tag{1.16}$$

En remplaçant (1.15) dans (1.16) on obtient :

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

Avec  $\sigma_\varepsilon^2$  matrice diagonale =

$$\begin{pmatrix}
\sigma_\varepsilon^2 & 0 & . & . & 0 \\
0 & \sigma_\varepsilon^2 & 0 & . & . \\
. & . & . & . & . \\
. & . & . & . & 0 \\
0 & . & . & 0 & \sigma_\varepsilon^2
\end{pmatrix}$$

Et puisque  $\sigma_\varepsilon^2$  est inconnu donc on l'estime.

##### ► Estimateur de la variance de l'erreur :

Soit le modèle :  $Y_t = XA + \varepsilon_t$

On a :  $e = Y - \hat{Y}$

$\hat{Y} = X\hat{A}$  d'où

$$e = Y - X\hat{A} \tag{1.17}$$

En remplaçant  $\hat{A} = (X'X)^{-1}X'Y$  dans l'équation (1.17) on aura :

$$e = Y - X(X'X)^{-1}X'Y = [I - X(X'X)^{-1}X']Y$$

On pose  $M$  matrice idempotente,  $M = I - X(X'X)^{-1}X'$   
alors

$$\sum_{t=1}^n e_t^2 = e'e = \varepsilon'M\varepsilon = \varepsilon'M\varepsilon$$

$$\mathbb{E}(e'e) = \sigma_\varepsilon^2 I_n [I - X(X'X)^{-1}X']^t = \sigma_\varepsilon^2 I_n (n - k - 1)$$

On obtient alors :

$$\hat{\sigma}_\varepsilon^2 = \frac{e'e}{n - k - 1}$$

Avec :  $\text{var}(\hat{A}) = \hat{\sigma}_\varepsilon^2 (X'X)^{-1}$  matrice variance-covariance.

### 1.2.5 Equation d'analyse de la variance et qualité d'un ajustement

Comme le modèle simple on a :

1.  $\sum_{i=1}^n Y_t = \sum_{i=1}^n \hat{Y}_t \Rightarrow \bar{Y} = \bar{\hat{Y}}$
2.  $\sum_{i=1}^n e_t = 0$

De ces deux relations nous déduisons l'équation fondamentale de l'analyse de la variance :

$$\sum_{t=1}^n \left( Y_t - \bar{Y} \right)^2 = \sum_{t=1}^n \left( \hat{Y}_t - \bar{\hat{Y}} \right)^2 + \sum_{t=1}^n e_t^2$$

$$SCT = SCE + SCR$$

Avec :

$SCT$  : variabilité totale.

$SCE$  : variabilité expliquée.

$SCR$  : variabilité résiduelle.

Cette équation permet de juger la qualité d'ajustement d'un modèle, en effet plus  $SCE$  est proche du  $SCT$  meilleur est l'ajustement globale du modèle. Cependant ces valeurs dépendant des unités de mesure, c'est pourquoi on préfère utiliser le nombre sans dimensions :

$$R^2 = \frac{\sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^n (Y_t - \bar{Y})^2} = 1 - \frac{\sum_{t=1}^n e_t^2}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

$R^2$  = Coefficient de détermination : mesure la proportion de la variance de  $Y$  expliquée par la régression de  $Y$  sur  $X$ .

- Si  $n < k$  alors on calcule le coefficient de détermination corrigé.

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

On a  $\bar{R}^2 \leq R^2$  et si  $n$  est grand  $\bar{R}^2 \cong R^2$

## Chapitre 2

# Tests et intervalles de confiance

### 2.1 Cas de modèle linéaire simple

#### 2.1.1 Test de nullité de $\hat{a}$

Soit le modèle suivant :

$$Y_t = aX_t + b + \varepsilon_t$$

On sait que

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{t=1}^n e_t^2}{n-2} \quad (2.1)$$

Et  $\varepsilon_t \rightarrow \mathcal{N}(0, \sigma_\varepsilon^2)$

De (2.1) on a :

$$\begin{aligned} (n-2)\hat{\sigma}_\varepsilon^2 &= \sum_{t=1}^n e_t^2 \\ \Rightarrow \frac{(n-2)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} &= \frac{\sum_{t=1}^n e_t^2}{\sigma_\varepsilon^2} \rightarrow \chi_{n-2}^2 \end{aligned}$$



D'un autre coté on a :

$$\hat{a} \rightarrow \mathcal{N} \left( a, \frac{\sigma_\varepsilon^2}{n \sum_{t=1} x_t^2} \right)$$

D'où on obtient la variable centrée réduite  $Z_1$  :

$$Z_1 = \frac{\hat{a} - a}{\sqrt{\frac{\sigma_\varepsilon^2}{n \sum_{t=1} x_t^2}}} \rightarrow \mathcal{N}(0, 1)$$

D'après la définition de la loi de Student qui est : le rapport d'une loi centrée réduite et la racine carrée d'une loi de khideux divisée par le nombre de ses degrés de liberté.

On applique cette définition on obtient alors :

$$T_c = \frac{Z_1}{\sqrt{\frac{(n-2) \hat{\sigma}_\varepsilon^2}{n-2}}}$$

$$T_c = \frac{\sigma_\varepsilon Z_1}{\hat{\sigma}_\varepsilon}$$

$$T_c = \frac{\sigma_\varepsilon}{\hat{\sigma}_\varepsilon} \frac{\hat{a} - a}{\sqrt{\frac{\sigma_\varepsilon^2}{n \sum_{t=1} x_t^2}}}$$

$$T_c = \frac{\hat{a} - a}{\hat{\sigma}_{\hat{a}}} \rightarrow t_{(n-2, \frac{\theta}{2})} \quad (2.2)$$

### Hypothèses du test

A partir du résultats (2.2) on peut effectuer un test qui consiste à tester l'hypothèse suivante :

$$\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0 \end{cases}$$

### La statistique du test

Sous l'hypothèse  $H_0$ , on obtient la valeur critique ( $T_c$ ) tel que :

$$T_c = \left| \frac{\hat{a} - a}{\hat{\sigma}_{\hat{a}}} \right| \rightarrow T_t \left( n - 2, \frac{\theta}{2} \right)$$

Avec :

$T_c$  : la valeur critique de la statistique ( $T$ ) (dite calculée) ;

$\hat{a}$  : l'estimateur du paramètre  $a$  ;

$\hat{\sigma}_{\hat{a}}$  : l'écart-type du paramètre  $a$  ;

$\theta$  : Le seuil donné, en général  $\theta = 5\%$  ;

$n - 2$  : degré de liberté ;

$T_t$  : la valeur de la statistique Student ( $T$ ) lue à partir de la table statistique.

### Règle de décision

- Si  $|T_c| < T_t$ , on accepte l'hypothèse  $H_0$  ;
- Si  $|T_c| \geq T_t$ , on rejette l'hypothèse  $H_0$ .

#### 2.1.2 Test de nullité de $\hat{b}$

$$\text{On a : } \hat{b} \rightarrow \mathcal{N} \left( b, \sigma_{\varepsilon}^2 \left( \frac{\sum_{t=1}^n x_t^2}{n \sum_{t=1}^n (X_t - \bar{X})^2} \right) \right)$$

D'où on obtient la variable centrée réduite  $Z_2$  :

$$Z_2 = \frac{\hat{b} - b}{\sqrt{\sigma_{\varepsilon}^2 \left( \frac{\sum_{t=1}^n X_t^2}{n \sum_{t=1}^n (X_t - \bar{X})^2} \right)}} \rightarrow \mathcal{N}(0, 1)$$

On appliquant la définition de la loi de Student on obtient :

$$\begin{aligned}
 T_c &= \frac{Z_2}{\sqrt{\frac{(n-2)\hat{\sigma}_\varepsilon^2}{n-2}}} \\
 T_c &= \frac{\sigma_\varepsilon Z_2}{\hat{\sigma}_\varepsilon} \\
 T_c &= \frac{\sigma_\varepsilon}{\hat{\sigma}_\varepsilon} \frac{\hat{b} - b}{\sqrt{\sigma_\varepsilon^2 \left( \frac{\sum_{t=1}^n X_t^2}{n \sum_{t=1}^n (X_t - \bar{X})^2} \right)}} \\
 T_c &= \frac{\hat{b} - b}{\hat{\sigma}_{\hat{b}}} \rightarrow t_{(n-2, \frac{\theta}{2})} \tag{2.3}
 \end{aligned}$$

### Hypothèses du test

A partir du résultats (2.3) on peut effectuer un test qui consiste à tester l'hypothèse suivante :

$$\begin{cases} H_0 : b = 0 \\ H_1 : b \neq 0 \end{cases}$$

### La statistique du test

Sous l'hypothèse  $H_0$ , on obtient la valeur critique ( $T_c$ ) tel que :

$$T_c = \left| \frac{\hat{b} - b}{\hat{\sigma}_{\hat{b}}} \right| \rightarrow T_t \left( n - 2, \frac{\theta}{2} \right)$$

Avec :

$T_c$  : la valeur critique de la statistique ( $T$ ) (dite calculée) ;

$\hat{b}$  : l'estimateur du paramètre  $b$  ;

$\hat{\sigma}_{\hat{b}}$  : l'écart-type du paramètre  $b$  ;

$\theta$  : Le seuil donné, en général  $\theta = 5\%$ ;

$n - 2$  : degré de liberté;

$T_t$  : la valeur de la statistique Student ( $T$ ) lue à partir de la table statistique.

### Règle de décision

- Si  $|T_c| < T_t$ , on accepte l'hypothèse  $H_0$ ;
- Si  $|T_c| \geq T_t$ , on rejette l'hypothèse  $H_0$ .

### Remarques

1. Lorsque on effectue les tests d'hypothèses bilatéraux des deux paramètres  $a$  et  $b$  suivants :

$$\left\{ \begin{array}{l} H_0 : a = \alpha \\ H_1 : a \neq \alpha \end{array} \right. \quad \text{et} \quad \left\{ \begin{array}{l} H_0 : b = \beta \\ H_1 : b \neq \beta \end{array} \right.$$

On prend le seuil  $\frac{\theta}{2}$ .

2. Lorsque on effectue les tests unilatéraux des deux paramètres c'est-à-dire les tests d'hypothèses suivants :

$$\left\{ \begin{array}{l} H_0 : a = \alpha \\ H_1 : a > \alpha, a < \alpha \end{array} \right. \quad \text{et} \quad \left\{ \begin{array}{l} H_0 : b = \beta \\ H_1 : b > \beta, b < \beta \end{array} \right.$$

On prend le seuil  $\theta$ .

### Intervalles de confiances des paramètres $a$ et $b$

Les intervalles de confiances des paramètres  $a$  et  $b$  et au seuil donné  $\theta$  (au niveau de confiance  $(1 - \theta)$ ) sont donnés par :

**Intervalle de confiance de  $a$  :**

$$\hat{a} \pm t_{\frac{\theta}{2}} * \hat{\sigma}_{\hat{a}}$$

**Intervalle de confiance de  $b$  :**

$$\hat{b} \pm T_{\frac{\theta}{2}} * \hat{\sigma}_{\hat{b}}$$

**Intervalles de confiance de  $\sigma^2$** 

On a :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = S^2 \Rightarrow \frac{(n-2)S^2}{\sigma^2} \Rightarrow \chi_{n-2}^2$$

D'où

$$P\left[\chi_1^2 \leq \frac{(n-2)S^2}{\sigma^2} \leq \chi_2^2\right] = 1 - \theta$$

$$P\left[\frac{\chi_1^2}{(n-2)S^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi_2^2}{(n-2)S^2}\right] = 1 - \theta$$

On obtient :

$$P\left[\frac{(n-2)S^2}{\chi_1^2} \leq \sigma^2 \leq \frac{(n-2)S^2}{\chi_2^2}\right] = 1 - \theta$$

**2.1.3 Analyse de la variance et test de Fisher**

On a déjà défini  $SCR$ ,  $SCE$  et  $SCT$ , ses sommes peuvent être utilisées pour tester l'hypothèse suivante : pour  $a$  :

$$\begin{cases} H_0 : a = 0 \wedge b = 0 \\ H_1 : a \neq 0 \wedge b \neq 0 \end{cases}$$

Sous l'hypothèse :  $H_0 : a = 0 \wedge b = 0$ ,

On a :

$$\mathbb{E}(SCT) = (n-1)\sigma^2$$

$$\mathbb{E}(SCE) = (1)\sigma^2$$

$$\mathbb{E}(SCR) = (n-2)\sigma^2$$

Avec  $(n-1)$ ,  $(1)$  et  $(n-2)$  des degrés de libertés de  $SCT$ ,  $SCE$  et  $SCR$  respectivement.

D'autre part, lorsque  $H_0$  est vérifiée on a :

$$\frac{SCT}{n-1} \rightarrow \chi_{n-1}^2,$$

$$\frac{SCR}{n-2} \rightarrow \chi_{n-2}^2,$$

et  $\frac{SCE}{1} \rightarrow \chi_1^2$

Du moment que  $\frac{SCE}{\sigma^2}$  et  $\frac{SCR}{\sigma^2}$  sont indépendants, on peut déduire donc valeur critique  $F$ (Fisher) qui se définit comme suit : C'est le rapport entre deux khi deux ( $\chi$ ) indépendants et leurs degrés de libertés ;

Alors on obtient :

$$F_c = \frac{\frac{\frac{SCE}{\sigma^2}}{1}}{\frac{\frac{SCR}{\sigma^2}}{n-2}} = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} = \frac{(n-2)SCE}{SCR} \rightarrow F_{(1,n-2,1-\theta)}$$

Avec :

$F_c$  : désigne la valeur critique de Fisher calculée.

$F$  : désigne la valeur de Fisher lue à partir de la table statistique de Fisher aux degrés de libertés.

$(1, n-2)$  : se sont des degrés de libertés.

$\theta\%$  : Le seuil donné.

### Règle de décision

- Si  $|F_c| \geq F_{(1,n-2,1-\theta)}$  on rejette l'hypothèse  $H_0$  : c'est-à-dire le modèle est globalement significatif.
- Si  $|F_c| < F_{(1,n-2,1-\theta)}$  on accepte l'hypothèse  $H_0$  : c'est-à-dire le modèle n'est pas globalement significatif.

**Tableau d'analyse de la variance**

Source de variation	Sommes des carrées	Degré de liberté	Moyenne des carrées	F calculé
Variabilité à expliquer X	$SCE = \sum_{t=1}^n (\hat{Y} - \bar{\hat{Y}})^2$	$k = 1$	$\frac{SCE}{1} = MC_{reg}$	$F_c = \frac{MC_{reg}}{MC_{res}}$
Variabilité résiduelle	$SCR = \sum_{t=1}^n e_t^2 = \sum_{t=1}^n (Y - \hat{Y})^2$	$n - 2$	$\frac{SCR}{n - 2} = MC_{res}$	
Variabilité totale	$SCE = \sum_{t=1}^n (Y_t - \bar{Y})^2$	$n - 1$		

#### 2.1.4 Prévision à l'aide d'un modèle de régression simple

Une fois les paramètres du modèle  $Y_t = aX_t + b + \varepsilon_t$  ; sont estimés, le modèle est validé, il est possible d'effectuer les prévisions à l'horizon  $h$ .

Soit le modèle estimé sur la période  $t = 1, \dots, n$  alors :

$$\hat{Y}_n = \hat{a}X_n + \hat{b}$$

Si la valeur  $X_t$  est connue en  $(n + h)$  alors la prévision de la valeur estimée  $\hat{Y}_t$  à l'horizon  $(n + h)$  se calculera par l'équation suivante :

$$\hat{Y}_{n+h} = \hat{a}X_{n+h} + \hat{b}$$

Comme on peut calculer l'intervalle de confiance de  $Y_{n+h}$ .

Sachant bien sûr la valeur de la variance de l'erreur de prévision qui est :

$$var(e_{n+h}) = var(Y_{n+h} - \hat{Y}_{n+h}) = \hat{\sigma}_\varepsilon^2 \left[ \frac{1}{n} + \frac{(X_{n+h} - \bar{X})^2}{\sum_{i=1}^n x_t^2} + 1 \right]$$

D'où l'intervalle de confiance de la variable  $Y_{n+h}$  au seuil  $(1 - \theta)$  est :

$$Y_{n+h} = \hat{Y}_{n+h} \pm t_{n-2}^{\frac{\theta}{2}} * \hat{\sigma}_\varepsilon \sqrt{\left[ \frac{1}{n} + \frac{(X_{n+h} - \bar{X})^2}{\sum_{i=1}^n x_t^2} + 1 \right]}$$

Avec :

$t_{n-2}^{\frac{\theta}{2}}$  : désigne la valeur de  $t$  Student lue à partir de la table statistique au seuil  $\left(\frac{\theta}{2}\right)$  et au degré de liberté  $(n - 2)$ .

$\hat{\sigma}_\varepsilon$  : désigne l'écart-type de l'erreur en valeur connue.

$X_{n+h}$  : désigne la valeur de la variable exogène à l'horizon  $(n + h)$ .

$\hat{Y}_{n+h}$  : désigne la valeur de la variable endogène estimée à l'horizon  $(n + h)$ .

## 2.2 Cas de modèle linéaire multiple

### 2.2.1 Le test de student

Le test de student tester l'influence directe de la variable explicative sur la variable endogène, revient à tester son coefficient de régression s'il est égale ou différent de 0, pour un seuil choisi, en général  $\theta = 5\%$ .

Le test d'hypothèse est le suivant : est appelé (test bilatéral)

$$\begin{cases} H_0 : a_i = 0; i = 0, \dots, k \\ H_1 : a_i \neq 0; i = 0, \dots, k \end{cases}$$



La statistique de student est la suivante :

$$t_c^\theta = \left| \frac{\hat{a}_i - a_i}{\sigma_{\hat{a}_i}} \right| \rightarrow t_{(n-k-1, \frac{\theta}{2})}$$

### Règle de décision

- Si  $|T_c| \leq T$  on accepte l'hypothèse  $H_0$  : la variable  $x_i$  n'est pas contributive à l'explication de  $Y$ .

Test unilatéral : ce test est utilisé lorsque  $H_1 : a_i > 0$  , ou  $H_1 : a_i < 0$  , et d'après la Remarque ( 2.1.2)

$$t_c^\theta = \left| \frac{\hat{a}_i - a_i}{\sigma_{\hat{a}_i}} \right| \rightarrow t_{(n-k-1, \theta)}$$

### 2.2.2 Test de Fisher (test de signification globale du modèle de régression )

Pour tester si l'ensemble des variables explicatives ont une influence sur la variable à expliquée, on fait le test d'hypothèse suivant :

$$\begin{cases} H_0 : a_1 = a_2 = \dots = a_k = 0 \\ H_1 : \text{il existe au moins, } a_i \neq 0; i = 1, \dots, k \end{cases}$$

A partir de l'équation de l'analyse de la variance on a :

$$\begin{aligned} \sum_{t=1}^n (Y_t - \bar{Y})^2 &= \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 + \sum_{t=1}^n e_t^2 \\ \sum_{t=1}^n (Y_t - \bar{Y})^2 &= \mathbb{E} \left( \sum_{t=1}^n e_t^2 \right) = (n - k - 1) \sigma^2 \end{aligned}$$

d'où

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^n e_t^2}{n - k - 1}$$

$$\hat{\sigma}^2 = \frac{SCR}{n - k - 1} \quad (2.4)$$

Et d'après la définition d'un khideux on a :

$$\frac{(n - k - 1)\hat{\sigma}^2}{\sigma^2} \rightarrow \chi_{n-k-1}^2 \quad (2.5)$$

De (2.4) et (2.5) on obtient

$$\frac{(n - k - 1)\hat{\sigma}^2}{\sigma^2} = \frac{(n - k - 1)SCR}{(n - k - 1)\sigma^2} = \frac{SCR}{\sigma^2} \rightarrow \chi_{n-k-1}^2$$

De même pour  $SCE$  :

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=1}^n \left( \hat{Y}_t - \bar{\hat{Y}} \right)^2 \right] &= (k)\sigma^2 \Rightarrow \quad \hat{\sigma}^2 = \frac{SCE}{k} \\ &\Rightarrow \quad \frac{k\hat{\sigma}^2}{\sigma^2} = \frac{k(SCE)}{k(\sigma^2)} \rightarrow \chi_k^2 \end{aligned}$$

Puisque  $\chi_{n-k-1}^2$  et  $\chi_k^2$  sont indépendants alors :

Sous  $H_0$  :

$$\begin{aligned} F_c &= \frac{\frac{\sum_{t=1}^n \left( \hat{Y}_t - \bar{\hat{Y}} \right)^2}{k}}{\frac{\sum_{t=1}^n e_t^2}{n-k-1}} = \frac{\frac{\chi_k^2}{k}}{\frac{\chi_{n-k-1}^2}{n-k-1}} \\ &= \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} \rightarrow F_{k,n-k-1,\theta\%} \end{aligned}$$

### Règle de décision

Si  $|F_c| \geq F_\theta$  on rejette  $H_0$  , et on accepte  $H_1$ , le modèle est globalement significative.

La régression est jugée significative si la variabilité expliquée est significativement différent de 0.

**Tableau d'analyse de la variance**

Source de variation	Sommes des carrées	Degré de liberté	Moyenne des carrées	F calculé
$x_{1t}, \dots, x_{kt}$	$SCE = \sum_{t=1}^n (\hat{Y} - \bar{\hat{Y}})^2 = \hat{A}X'Y - n\bar{\hat{Y}}^2$	$k$	$\frac{SCE}{k} = MC_{reg}$	$F_c = \frac{MC_{reg}}{MC_{res}}$
Variabilité résiduelle	$SCR = \sum_{t=1}^n e_t^2 = Y'Y - \hat{A}'X'Y$	$n - k - 1$	$\frac{SCR}{n - k - 1} = MC_{res}$	
Variabilité totale	$SCE = \sum_{t=1}^n (Y_t - \bar{Y})^2 = Y'Y - n\bar{Y}^2$	$n - 1$		

### 2.2.3 La prévision dans le modèle de la régression multiple

Le problème consiste à déterminer quelle valeur doit être attribuée à la variable endogène lorsque nous connaissons les valeurs des variables exogènes.

Le modèle général estimé est le suivant :

$$\hat{Y}_t = \hat{a}_0 + \hat{a}_1 x_{1t} + \dots + \hat{a}_k x_{kt} + e_t$$

La prévision pour la date  $t + h$  est la suivante :

$$\hat{Y}_{t+h} = \hat{a}_0 + \hat{a}_1 x_{1t+h} + \dots + \hat{a}_k x_{kt+h} + e_{t+h}$$

L'erreur de prévision est donnée par :

$$e_{t+h} = Y_{t+h} - \hat{Y}_{t+h} \rightarrow \mathcal{N}(0, \sigma_{t+h}^2)$$

L'intervalle au seuil de probabilité  $(1 - \theta)$  est donné par la formule suivante :

$$Y_{t+h} = \hat{Y}_{t+h} \pm t_{(\frac{\theta}{2}, n-k-1)} * \hat{\sigma}_\varepsilon \sqrt{X'_{t+h} (X'X)^{-1} X_{t+h} + 1}$$

Considérant que les hypothèses du modèle linéaire général sont vérifiées, la prévision est sans biais.

## Chapitre 3

# Application et conclusion

Dans ce chapitre, nous allons analyser la régression linéaire simple sur un exemple. Cette présentation va nous permettre d'exposer la régression linéaire dans un cas simple afin de bien comprendre les enjeux de cette méthode, les problèmes posés et les réponses apportées.

### 3.1 Exemple d'application

Etude de la relation entre la tension artérielle et l'âge d'un individu. Les données sont extraites de *Bouyer et al.* (1995) *Epidémiologie. Principes et méthodes quantitatives*, Les éditions INSERM.

#### 1. Objectif

On souhaite savoir si, de façon générale, l'âge a une influence sur la tension artérielle et sous quelle forme cette influence peut être exprimée.

Le but est d'expliquer au mieux comment la tension artérielle varie en fonction de l'âge et éventuellement de prédire la tension à partir de l'âge.

#### 2. Population et variables étudiées

Population générale d'individus.

Sur cette population, on définit deux variables.

- La variable  $Y$  : variable tension ; c'est la variable à expliquer, appelée encore variable à régresser, variable réponse, variable dépendante (VD).
- La variable  $X$  : variable âge ; c'est la variable explicative, appelée également régresseur, variable indépendante (VI).

### 3. Echantillon aléatoire d'individus

Pour l'étude, on doit faire des mesures sur  $n$  individus tirés au sort dans la population.

Données numériques : on observe deux échantillons appariés de  $X$  et  $Y$  de taille :

$$(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)$$

où  $x_i$  et  $y_i$  sont les valeurs de  $X$  et  $Y$  observées sur le  $i^{me}$  individu tiré au sort.

### 4. Modèle exprimant la relation entre $Y$ et $X$

On cherche à exprimer la relation entre la variable tension et la variable âge à l'aide d'une fonction mathématique du type  $y = f(x)$ . Graphiquement cela revient à représenter cette relation à l'aide d'une courbe (graphe de la fonction).

### 5. Choix du modèle

Pour choisir le modèle de relation, on doit faire des observations sur un échantillon d'individus. Les données recueillies sur ces individus sont représentées graphiquement à l'aide d'un nuage de points. Si le nuage a une forme particulière s'apparentant à une courbe mathématique, on choisira la fonction mathématique correspondant à cette courbe.

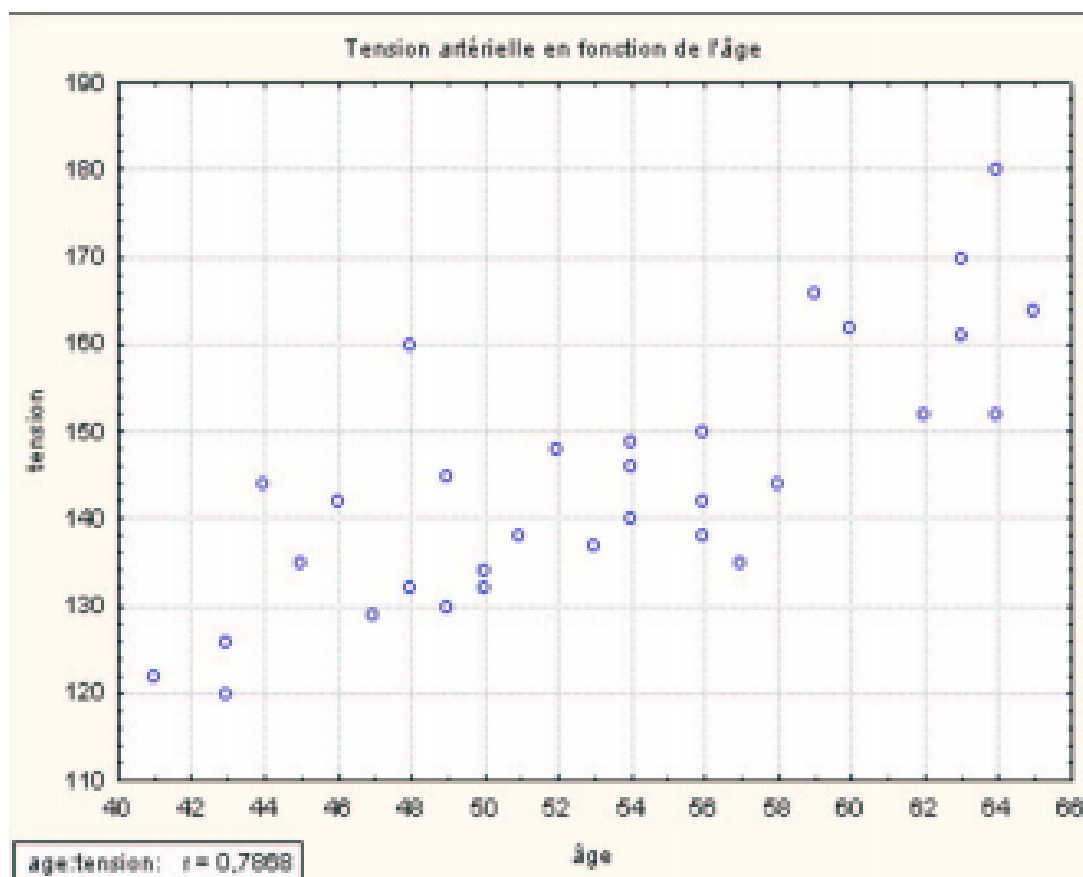
La forme étirée et croissante du nuage suggère une relation positive de type linéaire entre la tension et l'âge.

Le coefficient de corrélation linéaire observé sur l'échantillon est  $r = 0.7868$ .

*Modèle de régression linéaire* : modèle le plus simple qui exprime la relation entre  $Y$  et  $X$  à l'aide d'une fonction linéaire. Graphiquement, la relation est représentée par une droite d'équation  $y = ax + b$ .

#### 1) Ajustement du modèle aux données. Estimation des coefficients de la droite par la méthode des moindres carrés

Le modèle étant posé, il faut estimer numériquement les paramètres du modèle, c'est-à-dire cal-



culer les valeurs numériques des coefficients qui correspondent le mieux aux données. Cela revient à déterminer la droite qui s'ajuste le mieux aux données, c'est-à-dire la droite qui est la plus proche des points.

Selon quel critère et quelles sont les formules permettant d'obtenir des valeurs estimées des coefficients ?

### 1. Formules de calcul des coefficients estimés

$$\hat{a} = \frac{\text{cov}(x, y)}{\text{var}(x)} = r(x, y) \sqrt{\frac{\text{var}(y)}{\text{var}(x)}}$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

Sur l'exemple, on obtient :

$$\bar{x} = 53.85, \bar{y} = 145.32, r = 0.7868, s_x^* = 7.18, s_y^* = 14.39$$

La méthode des moindres carrés fournit les coefficients estimés suivants sur l'exemple :

$$\hat{a} = 1.5771$$

et

$$\hat{b} = 60.3928$$

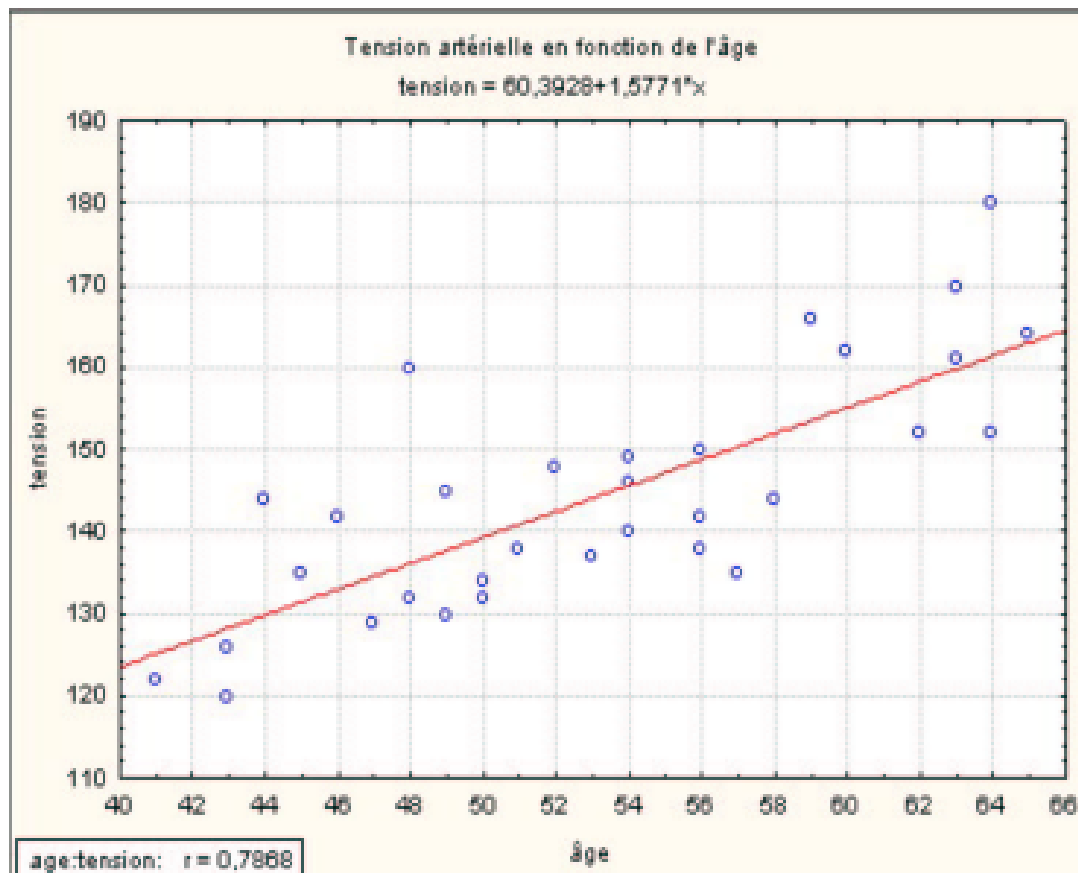
la droite d'équation  $y = 1,5771x + 60,3928$  s'appelle la droite de régression estimée de  $Y$  sur  $X$ .

### 2. droite de régression estimée ou droite des moindres carrés

### 3. Valeurs ajustées, résidus et somme des carrés des résidus

Une fois les coefficients de la droite estimés, on calcule pour chaque individu :

- $\hat{y}_i = \hat{a}x_i + \hat{b}$  s'appelle la valeur ajustée ou prédite de  $Y$  par le modèle.
- $e_i = y_i - \hat{y}_i$  s'appelle le résidu de l'observation  $i$ . C'est l'écart entre la valeur de  $Y$  observée sur l'individu  $n^{me}$  et la valeur prédite. Le résidu  $e_i$  est une approximation du terme d'erreur  $\varepsilon_i$
- la somme des carrés des résidus est  $SCR = \sum_{i=1}^n e_i^2$ . Elle mesure la distance de la droite de régression aux points du nuage de points qui est minimale au sens des moindres carrés.



- La statistique  $\hat{\sigma}^2 = \frac{SCR}{n-2}$  est un estimateur sans biais de  $\sigma^2$

**Exemple numérique :**

l'individu  $n^{\circ}4$  a pour âge  $x_4 = 44$ .

La tension observée est  $y_4 = 144$

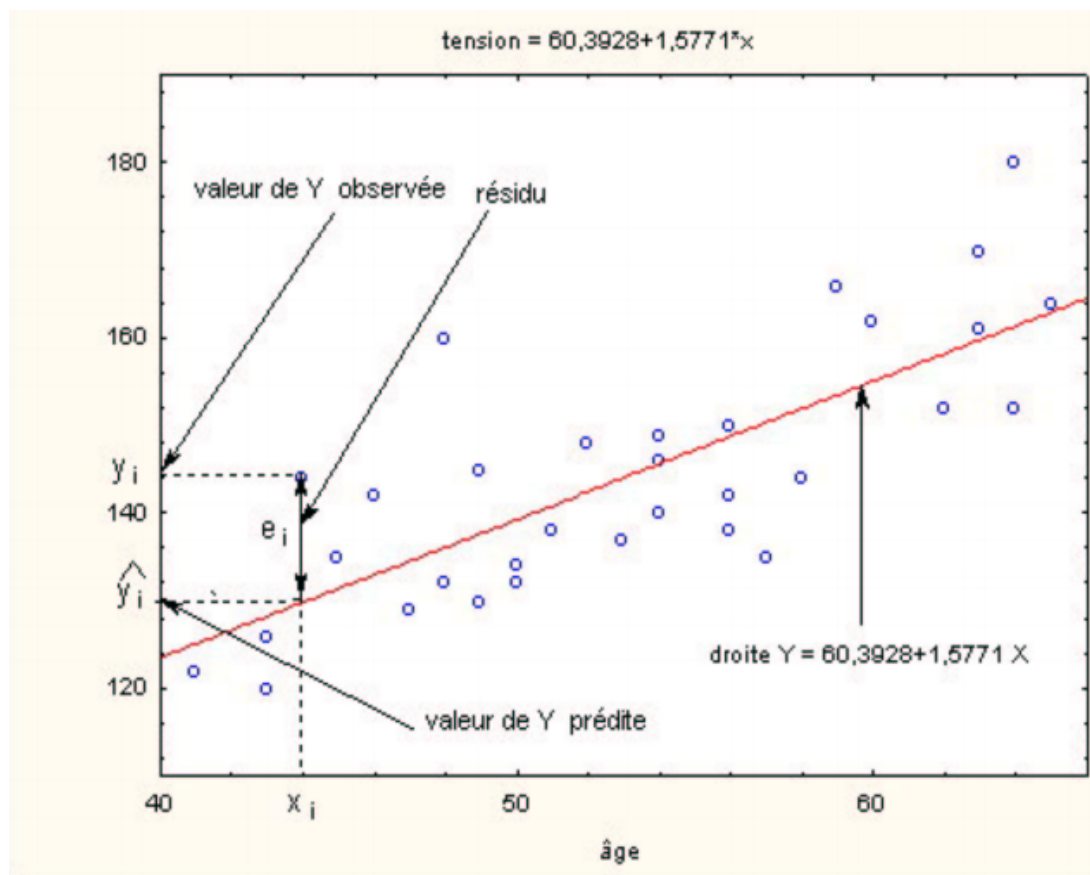
La tension prédite (ou estimée) par le modèle est

$$\hat{y}_4 = 60,3928 + 1,577144 = 129,7852$$

.

Le résidu pour l'observation  $n^{\circ}4$  est  $e_4 = 14.2148$ . Le résidu est positif (point au





dessus de la droite).  $SCR = 2605,569$

$$\hat{\sigma}^2 = \frac{2605.569}{32} = 81.424$$

#### 4. Comment mesurer la qualité de l'ajustement

Pour le modèle choisi,  $Y$  peut varier en fonction :

- de  $X$ , selon la relation linéaire postulée.
- d'autres variables non prises en compte et synthétisées dans le terme d'erreur.

On va mesurer la part de chacune de ces deux sources de variation pour évaluer la qualité de l'ajustement du modèle aux données.

#### Décomposition de la variation totale des observations

On peut tout d'abord écrire la décomposition suivante :

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

On peut montrer la propriété suivante :

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

peut être décomposée en deux parties :

$$SCT = SCR + SCE$$

où  $SCE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  représente la variance expliquée par la régression (mesure la variation des valeurs ajustées autour de la moyenne  $\bar{y}$ ) et  $SCR = \sum_{i=1}^n e_i^2$  représente la variance résiduelle ou non expliquée (partie de la variation totale qui n'est pas expliquée par le modèle de régression).

Dans l'exemple, on obtient :

$$SCT = 6839.442, SCE = 4233.873, SCR = 2605.569$$

### Le coefficient de détermination $R^2$

Afin d'avoir une idée globale de la qualité de l'ajustement linéaire, on définit  $R^2$  le coefficient de détermination qui est le carré du coefficient de corrélation  $R$  :

$$R^2 = \frac{SCE}{SCT}$$

Il mesure la part de la variation totale de  $Y$  expliquée par le modèle de régression sur  $X$ .

Pour l'exemple, on a  $r^2 = 0.619 = \frac{4233.873}{6839.442}$ . Le modèle de régression explique 61,91% de la variation totale.

## 2) Tests

### 1. Test global de significativité de la régression

Il paraît raisonnable de tester la significativité globale du modèle, c'est à dire tester si tous les coefficients sont supposés nuls, excepté la constante. Cela correspond dans le cas de la régression linéaire simple à  $H_0 : a = 0$  contre  $H_1 : a \neq 0$

**La statistique du test** : statistique  $F$  de Fisher

On utilise la statistique, notée  $F$  définie par la formule :

$$F = (n - 2) \frac{R^2}{1 - R^2} = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}}$$

**Loi de  $F$  sous  $H_0$**

La statistique  $F$  suit la loi de Fisher à  $(1, n - 2)$  ddl.

**Région de rejet de  $H_0$**  Sous  $H_0$  , on s'attend à observer une valeur de  $F$  proche de 0.

Plus la valeur de  $F$  est grande et plus elle est en faveur de  $H_1$ .

La région de rejet est située à l'extrémité droite du domaine .

**Décision** Règle basée sur la p-valeur : si  $\theta_{obs} \leq \theta$ , on rejette  $H_0$  au risque d'erreur  $\theta$ .

$$\theta_{obs} = P_{H_0} \left( F(1, n - 2) > (n - 2) \frac{r^2}{1 - r^2} \right)$$

Dans Statistica, les valeurs observées de  $F$  sont données ainsi que la p-valeur.

### 2. Tests sur les paramètres

Reprenons l'exemple de la tension en fonction de l'âge. Nous avons modélisé la tension  $Y$  par l'âge  $X$ . Il paraît raisonnable de se poser les questions suivantes : (a) est-ce-que le coefficient  $a$  est non nul, autrement dit la variable  $X$  a-t-elle réellement une influence sur  $Y$  ?

(b) est-ce-que le coefficient  $b$  est non nul, autrement dit faut-il une constante dans le modèle ?

Rappelons que le modèle utilisé est le suivant :

$$y_i = ax_i + b + \varepsilon_i$$

Nous pouvons expliciter les questions précédentes en terme de test d'hypothèse :

(a)correspond à  $H_0 : a = 0$  contre  $H_1 : a \neq 0$

(b)correspond à  $H_0 : b = 0$  contre  $H_1 : b \neq 0$

**La statistique du test** : statistique T de Student

On utilise la statistique, notée  $T$  :

pour le coefficient  $a$  définie par :  $T = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}}$

pour le coefficient  $b$  définie par :  $T = \frac{\hat{b}}{\hat{\sigma}_{\hat{b}}}$

**loi de  $T$  sous  $H_0$**  La statistique  $T$  suit la loi de Student à  $(n - 2)$  ddl.

**Région de rejet de  $H_0$**  Sous  $H_0$  , on s'attend à observer une valeur de  $T$  proche de 0.

Plus la valeur de  $|T|$  est grande et plus elle est en faveur de  $H_1$ .

La région de rejet est située à l'extrémité droite et à l'extrémité gauche du domaine (test bilatéral).

**Décision** Règle basée sur la p-valeur : si  $\theta_{obs} \leq \theta$  , on rejette  $H_0$  au risque d'erreur  $\theta$

pour le coefficient  $a$  :  $\theta_{obs} = 2P_{H_0} \left( T(n - 2) > \left| \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}} \right| \right)$

pour le coefficient  $b$  :  $\theta_{obs} = 2P_{H_0} \left( T(n - 2) > \left| \frac{\hat{b}}{\hat{\sigma}_{\hat{b}}} \right| \right)$

Dans Statistica, les valeurs observées de  $T$  sont données ainsi que la p-valeur.

**Remarque :**

On peut remarquer que dans le cas de la régression linéaire simple, il est équivalent de tester la significativité globale du modèle ou bien de tester  $H_0 : a = 0$ , contre  $H_0 : a \neq 0$ .

Effectivement, on a  $T^2 = F$  . Au niveau des lois l'égalité est aussi valable bien évidemment et nous avons que le carré d'un Student à  $(n - 2)$  ddl est une loi de Fisher à  $(1, n - 2)$  ddl.

Bien entendu le quantile  $(1 - \alpha)$  d'une loi de Fisher correspond au quantile  $(1 - \frac{\theta}{2})$  d'une loi de Student.

Pour l'exemple, on obtient :

Synthèse de la Régression; Variable Dép. : tension (donnees-boi R= ,78678955 R²= ,61903780 R² Ajusté = ,60713273 F(1,32)=51,998 p<,00000 Err-Type de l'Estim.: 9,0235						
N=34	Bêta	Err-Type de Bêta	B	Err-Type de B	t(32)	niveau p
OrdOrig.			60,39282	11,87925	5,083893	0,000016
age	0,786790	0,109110	1,57709	0,21871	7,210952	0,000000

$$\hat{\sigma} = \sqrt{\frac{2605.569}{32}} = 9.0235$$

(a) :  $H_0 : a = 0$ , contre  $H_1 : a \neq 0$

$$t_{obs}^2 = (7.211)^2 = 51998 = f_{obs}$$

$\theta_{obs} = 0$ , on rejette donc  $H_0$  au risque 5%

(a) :  $H_0 : b = 0$ , contre  $H_1 : b \neq 0$

$$t_{obs} = 5.084$$

$\theta_{obs} \approx 0$ , on rejette donc  $H_0$  au risque 5%

### 3) Intervalles de confiance et intervalles de prévision

#### 1. Intervalles de confiance

On supposera dans la suite que les  $\varepsilon_i$  en plus d'être indépendants, de même loi, centrées et de même variance, sont distribuées suivant une loi  $\mathcal{N}(0, \sigma^2)$

La valeur ponctuelle d'un estimateur est en général insuffisante et il est nécessaire de lui adjoindre un intervalle de confiance.

(i) Un IC de  $a$  au niveau  $1 - \theta$  est donné par :

$$\left[ \hat{a} - t\hat{\sigma}_{\hat{a}}, \hat{a} + t\hat{\sigma}_{\hat{a}} \right]$$

où  $t$  représente le quantile de niveau  $(1 - \frac{\theta}{2})$  d'une loi de Student de  $(n - 2)$  ddl.

(ii) Un IC de  $b$  au niveau  $1 - \theta$  est donné par :

$$\left[ \hat{b} - t\hat{\sigma}_{\hat{b}}, \hat{b} + t\hat{\sigma}_{\hat{b}} \right]$$

où  $t$  représente le quantile de niveau  $(1 - \frac{\theta}{2})$  d'une loi de Student de  $(n - 2)$  ddl.

Nous pouvons également donner un intervalle de confiance de la droite de régression.

Un IC de  $y_i$  au niveau  $(1 - \theta)$  est donné par :

$$\hat{y}_j - t\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{y}_j + t\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

En calculant les IC pour tous les points de la droite, nous obtenons une hyperbole de confiance. En effet, lorsque  $x_j$  est proche de  $\bar{x}$ , le terme dominant de la variance est  $\frac{1}{n}$ , mais dès que  $x_j$  s'éloigne de  $\bar{x}$ , le terme dominant est le terme au carré.

2. **Intervalles de prédiction** Un des buts de la régression est de proposer des prédictions pour la variable à expliquer  $Y$ . Soit  $x_{n+1}$  une nouvelle valeur de la variable  $X$ , nous voulons prédire  $y_{n+1}$ . Le modèle indique que :

$$y_{n+1} = ax_{n+1} + b + \varepsilon_{n+1}$$

Nous pouvons prédire la valeur correspondante grâce au modèle estimé

$$\hat{y}_{n+1}^p = \hat{a}x_{n+1} + \hat{b}$$

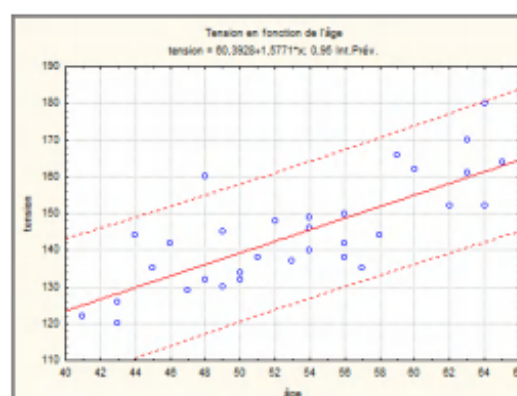
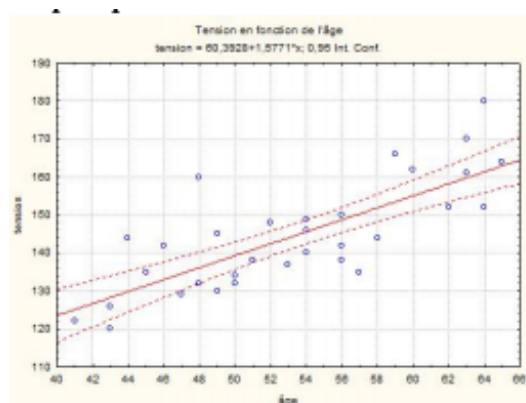
En utilisant la notation  $\hat{y}_{n+1}^p$ , nous souhaitons insister sur la notion de prévision : la valeur pour laquelle nous effectuons la prévision ici la  $(n + 1)^{me}$  n'a pas servi dans le calcul des

estimateurs. Remarquons que cette quantité serait différente de la valeur ajustée, notée  $\hat{y}_i$ , qui elle fait intervenir la  $i^{me}$  observation. Deux types d'erreurs vont entacher notre prévision, la première due à la non connaissance de  $\varepsilon_{n+1}$  et l'autre due à l'estimation des paramètres. La variance augmente lorsque  $x_{n+1}$  s'éloigne du centre de gravité du nuage. Faire de la prévision lorsque  $x_{n+1}$  est loin de  $\bar{x}$  est donc périlleux, la variance de l'erreur de prévision peut alors être très grande.

Un IC de  $y_{n+1}$  au niveau  $1 - \theta$  est donné par :

$$\hat{y}_{n+1}^p - t\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{y}_{n+1}^p + t\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Cette formule exprime que plus le point à prévoir est éloigné de  $\bar{x}$ , plus la variance de la prévision et donc l'IC seront grands. Une approche intuitive consiste à remarquer que plus une observation est éloignée du centre de gravité, moins nous avons d'information sur elle. Lorsque la valeur à prévoir est à l'intérieur de l'étendue des  $x_i$ , le terme dominant de la variance est la valeur 1 et donc la variance est relativement constante. Lorsque  $x_{n+1}$  est en dehors de l'étendue des  $x_i$ , le terme dominant peut être le terme au carré, et la forme de l'intervalle sera à nouveau une hyperbole.



## 3.2 Conclusion

Les différentes phases d'un régression peuvent se résumer par trois étapes successives.

1. La première est la modélisation : nous avons supposé que la variable  $Y$  est expliquée de manière linéaire par la variable  $X$  via le modèle de régression  $Y = aX + b + \varepsilon$ .
2. La seconde est l'étape d'estimation : nous avons ensuite estimé les paramètres grâce aux données récoltées. Les hypothèses sur le résidu  $\varepsilon$  ont permis d'établir des propriétés statistiques des estimateurs obtenus.
3. Enfin la troisième étape est celle de validation . aborder le problème de la validation des hypothèses sur les résidus et la qualité de l'ajustement observation par observation.



# Bibliographie

- [1] Adrien-Marie Legendre. (1805). Nouvelles méthodes pour la détermination des orbites des comètes, *Paris, F. Didot*, **80**.
- [2] Bresson, G., Pirotte, A. (1995). Econometrie des séries temporelles : Théorie et applications.
- [3] Chavent, M. Régression linéaire simple, *Université de Bordeaux*.
- [4] Chouquet, C. (2009-2010). Modeles Lineaire, *Université Paul Sabatier - Toulouse*.
- [5] (2010). Dodge , 451-452
- [6] (2010). Dodge , **217**
- [7] Fermin Rodriguez, A.K. (2009-2010). PMP STA 21 Méthodes statistiques pour l'analyse des données en psychologie, *Université Paris Ouest Nanterre La Défense*.
- [8] Genin, M. (2015). Régression linéaire multiple, *Université de Lille 2*.
- [9] Guyader, A. (2012-2013). Régression linéaire, *Université Rennes 2*.
- [10] Palm, R et F, A. (1995). Lemme (Quelques alternatives à la régression classique dans le cadre de la colinéarité), *Revue de statistique appliquée*, **43**, 5-33.
- [11] Saporta, G. (2006). Probabilités Analyse des données et statistique, second Edition.
- [12] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, **58**, 267-288.
- [13] Zabell, S. L. (2008). The probable error of the mean, *Journal of the American Statistical Association*, **103**, 1-7.