INTRODUCTION

L'analyse de régression peut être définie comme la recherche de la relation stochastique qui lie deux ou plusieurs variables. Son champ d'application recouvre de multiples domaines, parmi lesquels on peut citer la physique, l'astronomie, la biologie, la chimie, la médecine, la géographie, la sociologie, l'histoire, l'´économie, la linguistique et le droit.

La régression est l'une des méthodes les plus connues et les plus appliquées en statistique pour l'analyse de données quantitatives. Elle est utilisée pour établir une liaison entre une variable quantitative et une ou plusieurs autres variables quantitatives.

Sous la forme d'un modèle, si on s'intéresse à la relation entre deux variables, on parlera de régression simple en exprimant une variable en fonction de l'autre. Si la relation porte entre une variable et plusieurs autres variables, on parlera de régression multiple.

La régression linéaire simple et multiple est une classe particulière de modèle de règression qui sert à expliquer une variable Y, appelée variable endogène par une ou plusieurs variables explicatives dites exogènes a travers une fonction affine.

Ce travail se situe dans le cadre de l'analyse des données quantitatives, il est composé de trois chapitres. Le premier consiste à présenter des généralités sur les séries statistiques, le deuxième chapitre est consacré à étudier la régression linéaire simple où on va estimer les paramètres b et a du modèle

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

par la méthode des moindres carrés ordinaire **(MCO)** et tester sa significativité. Finalement, dans le dernier chapitre on généralise les résultats du chapitre 2 en traitant la régression multiple représentée par le modèle suivant :

$$Y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + ... + b_k x_{kt} + \varepsilon_t, \quad \text{pour} \quad t = 1, ..., n$$

où $b_0, b_1, b_2, ..., b_k$ sont les paramètres du modèle.

Chapitre 1

Généralités sur les séries statistiques

Ce chapitre sert à rappeler essentiellement des définitions, outils et notions dont aura besoin tout au long de ce travail. Pour plus de détails, on invite le lecteur à consulter les références ([2], [5], [7], et [9]).

1.1 Vocabulaire statistique

Le terme statistique désigne à la fois un ensemble de données d'observations, et l'activité qui consiste en leur traitement et leur interprétation.

* **La population Statistique** : est l'ensemble d'individus sur lequel des méthodes et des techniques de présentation et de description statistique sont appliquées**.** Par exemple :

1/ La population d'un pays.

2/ Les fonctionnaires d'un établissement.

3/ Le parc automobile français.

* **Les individus statistiques** ou **unités statistiques** : sont les éléments de la population statistique. On associe à ces éléments des propriétés appelées **"caractères"**.

Un **caractère** peut être :

1/ Qualitatif : couleur, sexe, efficacité d'un traitement,...

2/ Quantitatif : grandeur, âge, taille, poids,...

On distingue deux types de caractère quantitatif :

1/ discret.

2/ continu.

Une étude statistique est en fait l'étude de propriétés caractéristiques "caractères" de la population considérée. Généralement le nombre d'individus d'une population est très grand, pour cela l'étude statistique se restreint à une partie de cette population supposée assez représentative qu'on appelle **"échantillon"**.
Si l'échantillon est composé de n individus qui forment la population alors $x = (x_1, x_2, ..., x_n)$ est la série statistique associée au caractère étudié et n est la taille de l'échantillon.

1.2 Séries statistiques doubles

Il arrive fréquemment d'étudier deux caractères quantitatifs différents x et y d'un échantillon d'une même population pour déterminer s'il existe une relation entre eux (par exemple : les dépenses et les revenus d'une société) dans le sens que les valeurs de l'un peuvent être obtenues à partir de l'autre à laide d'une correspondance, on dit alors que x et y sont deux variables statistiques dépendantes.
En passant aux séries des centres, on pourra toujours supposer que les séries considérées sont à valeurs quantitatives discrètes (isolées).

1.2.1 Traitement des données et représentation graphique

Pour chaque individu de l'échantillon (de la population), on relève la valeur de deux caractères x et y, on obtient alors une liste de couples de nombres $(x_i, y_j)_{\substack{1\leq i\leq k \\ 1\leq j\leq l}}$, $k,\ l\ \in \mathbb{N}^*$ que l'on peut présenter sous forme d'un tableau à deux entrées

x/y	y_1	y_2	...	y_j	.	y_l	effictif marginal de x
x_1	n_{11}	n_{12}	...	n_{1j}	.	n_{1l}	$n_{1.}$
x_2	n_{21}	n_{22}	...	.	.	.	.
x_i	n_{i1}	n_{i2}	...	n_{ij}	.	n_{il}	$n.$
.	.	.	...	.	.	.	.
x_i	n_{k1}	n_{k2}	...	n_{kj}	.	n_{kl}	$n_{k.}$
effictif marginal de y	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	.	$n_{.l}$	n

* $x_1,\ x_2, ..., x_k$ sont les valeurs du caractère x.

* $y_1,\ y_2, ..., y_l$ sont les valeurs du caractère y.

* n_{ij} est l'effectif du couple (x_i, y_j) pour tout $1 \leq i \leq k$ et $1 \leq j \leq l$.

* $n_{i\ .}$ est l'effectif marginal de x_i, $n_{i\ .} = \sum_{j=1}^{l} n_{ij}$.

* $n_{.\ j}$ est l'effectif marginal de y_j, $n_{.\ j} = \sum_{i=1}^{k} n_{ij}$.

* n est l'effectif total, $n = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = \sum_{i=1}^{k} n_{i\ .} = \sum_{j=1}^{l} n_{.\ j}$.

Soit la série statistique double $(x_i, y_j, n_{ij})_{\substack{1\leq i\leq k\\ 1\leq j\leq l}}$ où $x_1,\ x_2, ..., x_k$ sont les différentes valeurs du premier caractère x et $y_1,\ y_2, ..., y_l$ sont les valeurs du second caractère y.

Définition 1.2.1 ***"Nuage de points"*** *Dans un repère orthogonal du plan, l'ensemble des points M_{ij} de coordonnées $(x_i,\ y_j)$ constitue le nuage de points associé à la série statistique double donnée ci-dessus.*

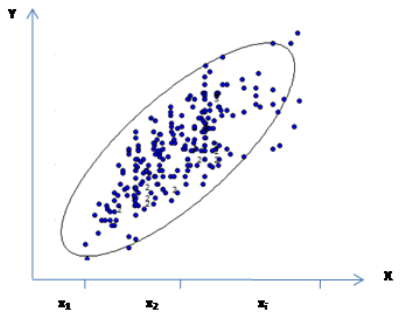

Figure 1 : nuage de points

Définition 1.2.2 ***"Le point moyen"*** *On appelle le point moyen du nuage de points associé à la série statistique double $(x_i, y_j, n_{ij})_{\substack{1\leq i\leq k\\ 1\leq j\leq l}}$ le point G de coordonnées $(\bar{x},\ \bar{y})$ où $\bar{x}$ et $\bar{y}$ sont les moyennes arithmétiques des séries statistiques $(x_i, n_{i.})_{1\leq i\leq k}$ et $(y_j, n_{.j})_{1\leq j\leq l}$ respectivement définies comme suit :*

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

et

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{l} y_j.$$

Exemple 1.2.1 *Le tableau suivant donne le chiffre d'affaire réalisé au cours des 6 derniers mois par un site de vente en ligne en fonction du nombre de commandes reçues.*

Nombre de commandes x_i	6400	8350	9125	9600	10050	12000
Chiffre d'affairemensuel y_i (DA)	250000	320000	335000	350000	370000	400000

a. Représentation du nuage de points

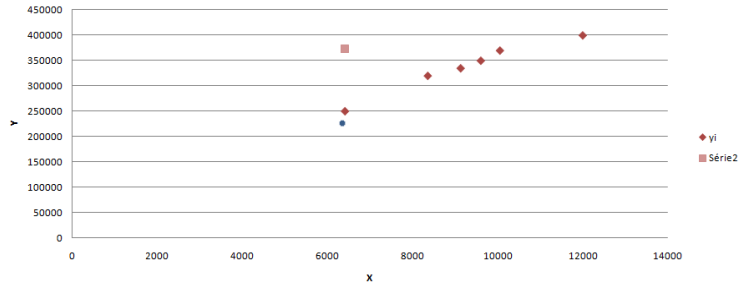

Figure 2 : Nuage de points exp 1.2.1

b. Calcul du point moyen

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{k} x_i \\ &= \frac{1}{6}\sum_{i=1}^{6} x_i \\ &= 9254,166667 \simeq 9254,17 \end{aligned}$$

et

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{j=1}^{l} y_j \\ &= \frac{1}{6}\sum_{j=1}^{6} y_j \\ &= 337500\ DA \end{aligned}$$

D'où le point moyen est $G(9254,17\ ;\ 337500)$.

1.2.2 Covariance d'une série statistique double

Définition 1.2.3 *On appelle covariance d'une série statistique double* $(x_i, y_j, n_{ij})_{\substack{1\leq i\leq k\\ 1\leq j\leq l}}$ *le nombre* $Cov(x,y)$ *donné par :*

$$\begin{aligned} Cov(x,y) &= \overline{xy} - \bar{x}\ \bar{y} \\ &= \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_i y_j\right) - \bar{x}\ \bar{y} \end{aligned}$$

où $\bar{x}$ *et* $\bar{y}$ *sont les moyennes arithmétiques des deux séries simples* x *et* y *respectivement définies précédemment.*

Remarque 1.2.1 *La covariance mesure les dispersions des deux variables* x *et* y *autour de leurs moyennes.*

1/ *Si* $Cov(x,y)$ *est nulle, on dit que les deux variables* x *et* y *se produisent indépendamment.*

2/ *Si* $Cov(x,y) > 0$, *les deux variables* x *et* y *sont liées positivement* $\Rightarrow$ *l'une augmente, l'autre augmente aussi.*

3/ *Si* $Cov(x,y) < 0$, *les deux variables* x *et* y *sont liées négativement* $\Rightarrow$ *l'une augmente, l'autre diminue.*

Proposition 1.2.1 ***"Propriétés de la covariance"*** *Soient* x *et* y *deux variables statistiques et* a, b, a' *et* b' *des constantes réelles, alors :*

1/ $Cov(x,x) = Var(x)$.

2/ $Cov(x,y) = Cov(y,x)$.

3/ $Cov(ax+b, a'y+b') = aa'Cov(x,y)$.

1.3 Ajustement affine d'une série statistique double

On considère la série statistique double $(x_i, y_j, n_{ij})_{\substack{1\leq i\leq k\\ 1\leq j\leq l}}$, avec un nuage de points $M_{ij}(x_i, y_j)$ associé.

Si les points de nuage paraissent presque alignés, on peut chercher une relation de la forme $y = ax + b$ où a et b sont deux constantes réelles telles que a est non nulle. Cette relation

exprime de façon approchée les valeurs de la variable statistique y en fonction des valeurs de la variable x. Alors, graphiquement, il s'agit de déterminer l'équation de la droite qui passe au plus prés de la majorité des points du nuage.

1.3.1 Méthodes d'ajustement affine d'une série statistique double

Il existe de nombreuses méthodes d'obtenir un ajustement affine d'une série statistique double, les plus utilisées sont :

Méthode de Mayer

La détermination d'un ajustement affine d'une série statistique double par la méthode de Mayer se fait en trois étapes :

i/ Découper la série statistique double en deux sous-séries de même effectif, si n l'effectif total de la population est impair, on peut mettre la valeur surnuméraire du caractère dans n'importe laquelle des deux sous-séries.

ii/ Calculer les coordonnées des deux points moyens G_1 et G_2 associés à ces deux sous-séries.

Trouver l'équation de la droite qui passe par les deux points G_1 et G_2. Cette droite est dite la droite de Mayer, elle constitue un ajustement affine de la série statistique double.

Exemple 1.3.1 *On reprend l'exemple (1.2.1). On partage le nuage de points en deux groupes de même importance suivant les valeurs croissantes de x_i, et on calcule les coordonnées des points moyens G_1et G_2 de chaque groupe de points.*

Les coordonnées de G_1 $(\overline{x_1}; \overline{y_1})$ avec $\overline{x_1}$ = moyenne des valeurs x du premier groupe et $\overline{y_1}$ = moyenne des valeurs y du même groupe.

$$\overline{x_1} = \frac{6400 + 8350 + 9125}{3} = 7958,33.$$

et

$$\overline{y_1} = \frac{250000 + 320000 + 335000}{3} = 301666,67.$$

Donc G_1 $(7958,33\ ;\ 301666,67)$.

De manière similaire on trouve les coordonnées du point moyen G_2 $(\overline{x_2}; \overline{y_2})$ du deuxième groupe.

$$\overline{x_2} = \frac{9600 + 10050 + 12000}{3} = 10550$$

et

$$\overline{y_2} = \frac{350000 + 370000 + 400000}{3} = 373333,33.$$

D'où G_2 $(10550\ ; 373333,33)$.

On trace la droite d'ajustement qui passe par les deux points G_1 $(7958,33\ ;\ 301666,67)$ et G_2 $(10550\ ; 373333,33)$, elle a la forme $y = ax + b$ où

$$a = \frac{373333,33 - 301666,67}{10550 - 7958,33} = 27.07$$

Et

$$b = 301666,67 - 27.07 * 7958,33 = 86932.$$

Ainsi, la droite d'ajustement affine a pour équation : $y = 27.07x + 86932$.

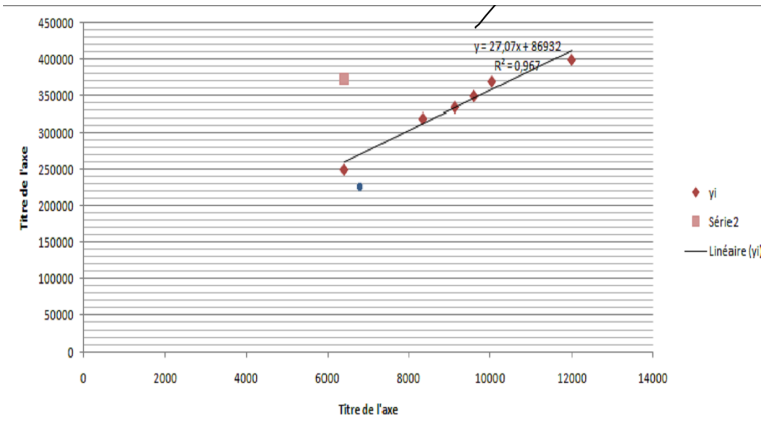

Figure 3 : Ajustement linaire par la mthode de Mayer

Méthode de moindres carrés

La méthode de moindres carrés consiste à trouver l'équation : $y = ax + b$ de la droite (D) dite la droite de régression y en x, telle que

$$a = \frac{Cov(x,y)}{Var(x)} \text{ et } b = \bar{y} - a\ \bar{x}.$$

De façon similaire on peut chercher l'équation $x = a'y + b'$ de la droite (D) dite la droite de régression x en y, où

$$a' = \frac{Cov(x,y)}{Var(y)} \text{ et } b' = \bar{x} - a'\ \bar{y}.$$

Remarque 1.3.1 *Les deux droites de régression données ci-dessous passent par le point moyen* $G\left(\bar{x},\ \bar{y}\right)$.

Exemple 1.3.2 *On considère* 5 *spécimens fossiles d'un animal disparu pour lesquels on possède les mesures de la longueur en* (cm) *de leur humérus* x *et de leur fémur* y.

1/ Le tableau suivant regroupe les résultats obtenus

t	x_t	y_t	x_t^2	y_t^2	$x_t y_t$
1	44	40	1936	1600	1760
2	65	60	4225	3600	3900
3	71	59	5041	3481	4189
4	75	65	5625	4225	4875
5	87	77	7569	5929	6699
Moyenne	68,4	60,2	4879,2	3767	4284,6

$$\begin{aligned} Var(x) &= \bar{x^2} - \bar{x}^2 = 4879,2 - (68,4)^2 = 200,64\ cm^2 \\ Var(y) &= \bar{y^2} - \bar{y}^2 = 3767 - (60,2)^2 = 142,96\ cm^2 \\ Cov(x,y) &= \overline{xy} - \bar{x}\ \bar{y} = 4284,6 - (68,4)\,(60,2) = 166,92\ cm^2 \end{aligned}$$

2/ On peut déterminer, par la méthode des moindres carrées ordinaires, l'équation $y = ax+b$ de la droite (D) de régression de x et y. Comme

$$\hat{a} = \frac{Cov(x,y)}{Var(x)} = \frac{166.92}{200.64} = 0,83194$$

et

$$\begin{aligned} b &= \bar{y} - a\ \bar{x} = 60,2 - 0,83194 \times 68,4 \\ &= 3,2955. \end{aligned}$$

Alors

$$(D) : y = 0,83194x + 0,83194.$$

3/ De façon similaire on peut chercher l'équation : $x = a'y + b'$ de la droite (D') dite la droite de régression x en y.

$$a' = \frac{Cov(x,y)}{Var(y)} = \frac{166,92}{142,96} = 1,16759$$

et

$$b' = \bar{x} - a'\ \bar{y} = 68,4 - 1,16759 \times 60,2 = 70,28892.$$

D'où

$$(D') : x = 1,16759y + 70,28892.$$

1.3.2 Coefficient de corrélation linéaire

Définition 1.3.1 *On appelle coefficient de corrélation linéaire du couple (x,y), le nombre réel $r(x,y)$ déduit de la covariance et donné par :*

$$r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}.$$

1/ *Le coefficient de corrélation ne dépend pas des unités de mesure des variables statistiques étudiées.*

2/ *La valeur du coefficient de corrélation est comprise entre -1 et 1 :*

$$|r(x,y)| \leq 1.$$

3/ *L'absence de corrélation $(r(x,y) = 0)$ implique l'absence de relation linéaire entre les variables statsitiques.*

4/ *Si $r(x,y) > 0$, alors les variables aléatoires x et y varient dans le même sens, et elles varient dans le sens contraire si $r(x,y) < 0$.*

5/ *Lorsque la corrélation est forte $r(x,y) \geq \frac{3}{4}$ les deux droites de régression sont très proches et l'ajustement linéaire du nuage de points sera possible, dans le cas contraire le nuage de points ne peut pas être ajusté par une droite, mais il se peut trouver une fonction qui donne la variabilité de x par rapport à y.*

Chapitre 2

La régression linéaire simple

Dans ce chapitre, on va introduire le modéle de la régression linéaire. Comme il s'agit d'un modéle avec une seule variable explicative, on parle de régression simple. On va ainsi définir la droite de régression à partir des données d'un échantillon par la méthode des moindres carrés.

2.1 Estimation des paramètres par la méthode des moindres carrés ordinaires (MCO)

Soit le modèle de régression suivant :

$$Y_t = b + aX_t + \varepsilon_t \tag{2.1.1}$$

L'estimation des paramètres b et a est obtenue en minimisant la somme des carrés des erreurs

$$Min \sum_{t=1}^{n} \varepsilon_t^2 = Min \sum_{t=1}^{n} (Y_t - b - aX_t)^2 = Min \sum_{t=1}^{n} S^2$$

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

$$\frac{\partial S}{\partial b} = 0 \Leftrightarrow 2\sum_{t=1}^{n} (Y_t - b - aX_t)(-1) = 0 \Rightarrow \sum_{t=1}^{n} Y_t = nb + a\sum_{t=1}^{n} X_t \tag{2.1.2}$$

et

$$\frac{\partial S}{\partial a} = 0 \Leftrightarrow 2\sum_{t=1}^{n} (Y_t - b - aX_t)(-X_t) = 0 \Rightarrow \sum_{t=1}^{n} Y_t X_t = b\sum_{t=1}^{n} X_t + a\sum_{t=1}^{n} X_t^2 \tag{2.1.3}$$

En notant $\hat{b}$ et $\hat{a}$ les solutions des équations (2.1.2) et (2.1.3), on obtient

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

ou

$$\hat{b} = \bar{Y} - \hat{a}\bar{X} \tag{2.1.4}$$

car

$$\bar{Y} = \frac{\sum_{t=1}^{n} Y_t}{n} \text{ et } \bar{X} = \frac{\sum_{t=1}^{n} X_t}{n}$$

sont les moyennes arithmétiques de Y et X respectivement.

En remlaçant la valeur de $\hat{b}$ dans l'équation (2.1.3), on obtient

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

ce qui donne

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

Les estimateurs des MCO du modèle de régression linéaire simple (2.1.1) sont :

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

et

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

2.2 Les modèles de la régression linéaire simple

Définition 2.2.1 *Le modèle de régression linéaire simple est une variable endogène (dépendante) expliquée par une seule variable exogène (indépendante) mise sous forme mathématique suivante :*

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

avec

*/ Y_t : La variable endogène (dépendante, à expliquer) à la date t.

*/ X_t : La variable exogène (indépendante, explicative) à la date t.

*/ $b,\ a$: Les paramètres inconnus du modèle.

*/ ε_t : L'erreur aléatoire du modèle.

*/ n : Nombre d'observations.

Remarque 2.2.1 *La régression linéaire simple a deux modèles :*

1/ Le modèle théorique *(modèle non ajusté)*

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

2/ Le modèle estimé *(modèle ajusté)*

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

avec

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

et

$$\begin{aligned} e_t &= Y_t - \hat{Y}_t \\ e_t &= Y_t - \hat{b} - \hat{a}X_t \end{aligned}$$

e_t : est le résidu du modèle.

2.3 Les tests d'hypothèses dans une régression linéaire simple

Les tests d'hypothèses statistiques sur les paramètres (les coefficients) du modèle de la régression simple et l'établissement des intervalles de confiance qui permettent de vérifier la validité du modèle.

Pour tester des hypothèses sur les paramètres a et b, on suppose que les erreurs e_t sont indépendantes et normalement distibuées. C'est-à-dire les erreurs suivent la loi normale $N(0,\ \sigma^2)$.

Soit le modèle de régression simple

$$Y_t = b + aX_t + \varepsilon_t \tag{2.3.1}$$

Les hypothèses du modèle : Le modèle (2.3.1) possède les propriétés suivantes :

1. $E(\varepsilon_t) = 0\ \forall t$: L'erreur est centrée.
2. $E(\varepsilon_t^2) = \sigma_\varepsilon^2,\ \forall t$: La variance de l'erreur est constante.
3. $Cov(\varepsilon_t, \varepsilon_{t'}) = 0,\ \forall \varepsilon_t \neq \varepsilon_{t'}$: Les erreurs ne sont pas auto-correlées.
4. $Cov(X_t, \varepsilon_{t'}) = 0$: L'erreur n'est pas corrélée avec la variable X .
5. La variable X n'est pas aléatoire.
6. Le modèle est linéaire en X par rapport aux paramètres.

2.3.1 Les espérances mathématiques, les variances et les covariances des estimateurs

On pose les nouvelles variables (changement de variables)

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

1/ L'espérance mathématique de $\hat{a}$:

Soit le modèle suivant :

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

D'après la méthode des MCO, on a :

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

ce qui implique

$$\hat{a} = \frac{\sum_{t=1}^{n} y_t x_t}{\sum_{t=1}^{n} x_t^2} \tag{2.3.2}$$

On remplace la valeur y_t dans (2.3.2), on obtient :

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

Alors

$$\hat{a} = \frac{\sum_{t=1}^{n} x_t Y_t}{\sum_{t=1}^{n} x_t^2}. \tag{2.3.3}$$

Car

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

On remplace maintenant $Y_t = b + aX_t + \varepsilon_t$ dans l'équation (2.3.3), on obtient

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

Il résulte alors

$$\hat{a} = a + \frac{\sum_{t=1}^{n} x_t \varepsilon_t}{\sum_{t=1}^{n} x_t^2} \tag{2.3.4}$$

En passant à l'espérance mathématique, on trouve

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

Or, d'après l'hypothèse (1)

$$E(\varepsilon_t) = 0$$

alors,

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

Finalemet a est un estimateur sans biais.

2/ L'espérance mathématique de $\hat{b}$:

On a

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

D'ou

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

Or que

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

donc

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

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

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

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

$$\begin{aligned} E\left(\hat{b}\right) &= E(b) + E\left(\sum_{t=1}^{n}\left(\frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^{n} x_t^2}\right)\varepsilon_t\right) \\ &= E(b) + \sum_{t=1}^{n}\left(\frac{1}{n} - \frac{\bar{X}x_t}{\sum_{t=1}^{n} x_t^2}\right)\underbrace{E(\varepsilon_t)}_{=0} \\ E\left(\hat{b}\right) &= E(b) = b. \end{aligned}$$

D'où b est aussi un estimateur sans biais.

3/ <u>La variance de $\hat{a}$:</u>

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

$$Var(\hat{a}) = E\left[\hat{a} - E(\hat{a})\right]^2 \quad \text{et} \quad E(\hat{a}) = a.$$

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

$$\begin{aligned} Var(\hat{a}) &= E(\hat{a}-a)^2 = E\left(\frac{\sum_{t=1}^{n} x_t\varepsilon_t}{\sum_{t=1}^{n} x_t^2}\right)^2 \\ &= \frac{1}{\left(\sum_{t=1}^{n} x_t^2\right)^2} E\left(\sum_{t=1}^{n} x_t\varepsilon_t\right)^2 \\ &= \frac{1}{\left(\sum_{t=1}^{n} x_t^2\right)^2} E\left(x_1^2\varepsilon_1^2 + x_2^2\varepsilon_2^2 + .. + x_n^2\varepsilon_n^2 + 2x_1x_2\varepsilon_1\varepsilon_2 + .. + 2x_{n-1}x_n\varepsilon_{n-1}\varepsilon_n\right). \end{aligned}$$

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

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

D'où

$$\begin{aligned} Var(\hat{a}) &= \frac{\sigma_\varepsilon^2 \sum_{t=1}^{n} x_t^2}{\left(\sum_{t=1}^{n} x_t^2\right)^2} \\ &= \frac{\sigma_\varepsilon^2}{\sum_{t=1}^{n} x_t^2}. \end{aligned}$$

Finalement

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

3/ La variance mathématique de $\hat{b}$:

On a

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

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

$$\begin{aligned}Var\left(\hat{b}\right) &= E\left[\hat{b}-E\left(\hat{b}\right)\right]^2 \\ &= E\left(\hat{b}-b\right)^2.\end{aligned}$$

Puisque $E\left(\hat{b}\right)=b$, on obtient

$$\begin{aligned}Var\left(\hat{b}\right) &= E\left(\hat{b}-b\right)^2 \\ &= E\left[\sum_{t=1}^{n}\left(\frac{1}{n}-\frac{\bar{X}x_t}{\sum_{t=1}^{n}x_t^2}\right)\varepsilon_t\right]^2 \\ &= E\left[\left(\frac{1}{n}-\frac{\bar{X}x_1}{\sum_{t=1}^{n}x_t^2}\right)\varepsilon_1+...+\left(\frac{1}{n}-\frac{\bar{X}x_n}{\sum_{t=1}^{n}x_n^2}\right)\varepsilon_n\right]^2 \\ &= E\left[\begin{array}{l}\left(\frac{1}{n}-\frac{\bar{X}x_1}{\sum_{t=1}^{n}x_t^2}\right)^2\varepsilon_1^2+...+\left(\frac{1}{n}-\frac{\bar{X}x_n}{\sum_{t=1}^{n}x_n^2}\right)^2\varepsilon_n^2+ \\ 2\left(\left(\frac{1}{n}-\frac{\bar{X}x_1}{\sum_{t=1}^{n}x_t^2}\right)\varepsilon_1\left(\frac{1}{n}-\frac{\bar{X}x_2}{\sum_{t=1}^{n}x_t^2}\right)\varepsilon_2+...+\left(\frac{1}{n}-\frac{\bar{X}x_{n-1}}{\sum_{t=1}^{n}x_t^2}\right)\varepsilon_{n-1}\left(\frac{1}{n}-\frac{\bar{X}x_n}{\sum_{t=1}^{n}x_t^2}\right)\varepsilon_n\right)\end{array}\right]\end{aligned}$$

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

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

Il résulte, alors

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

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

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

et comme $X_t = x_t + \bar{X}$, donc on déduit que

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

5/ La covariance mathématique $\left(\hat{b}; \hat{a}\right)$

Par définition, la covariance ente $\left(\hat{b}; \hat{a}\right)$ se calcule comme suit :

$$\begin{aligned} Cov\left(\hat{b}; \hat{a}\right) &= E\left[\left(\hat{b} - E\left(\hat{b}\right) \right) \left(\hat{a} - E\left(\hat{a}\right) \right) \right] \\ &= E\left[\left(\hat{b} - b \right) \left(\hat{a} - a \right) \right] \end{aligned}$$

Comme $\left(\hat{b};\hat{a}\right)$ sont sans biais, alors

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

D'après les hypothèses (1), (2) et (3) du modèle de régression simple, il résulte que

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

Théorème 2.3.1 *(Gaus Markov) : Les estimateurs sans biais du modèle de la régression simple (2.1.1) sont linéaires et de variances minimales.*

Preuve. L' estimateur des MCO s'écrit $\hat{a}=\sum_{t=1}^{n}p_ty_t$, avec

$$p_t = \frac{(X_t-\bar{X})}{\sum_{t=1}^{n}(X_t-\bar{X})^2.}$$

On considère un autre estimateur β linéaire en y_t et sans biais, c'est-à-dire

$$\beta=\sum_{t=1}^{n}\lambda_t y_t$$

et on montre que $\sum_{t=1}^{n}\lambda_t=0$ et $\sum_{t=1}^{n}\lambda_t x_t=1$ et $E(\varepsilon_{_t})=0$. L'égalité

$$\begin{aligned}
E(\beta) &= b\sum_{t=1}^{n}\lambda_t + a\sum_{t=1}^{n}\lambda_t x_t + \sum_{t=1}^{n}\lambda_t E(\varepsilon_{_t}) \\
&= b\sum_{t=1}^{n}\lambda_t + a\sum_{t=1}^{n}\lambda_t x_t
\end{aligned}$$

est vraie pour tout a. L'estimateur β est sans biais donc $E(\beta) = a$ pour tout a, c'est-à-dire que $\sum_{t=1}^{n} \lambda_t = 0$ et $\sum_{t=1}^{n} \lambda_t x_t = 1$. On montre maintenant que $Var(\beta) \geq Var(\hat{a})$. On a

$$\begin{aligned} Var(\beta) &= Var(\beta - \hat{a} + \hat{a}) \\ &= Var(\beta - \hat{a}) + Var(\hat{a}) + 2Cov(\beta - \hat{a}, \hat{a}). \end{aligned}$$

Or

$$\begin{aligned} Cov(\beta - \hat{a}, \hat{a}) &= Cov(\beta, \hat{a}) - Var(\hat{a}) \\ &= \frac{\sigma^2 \sum_{t=1}^{n} \lambda_t (X_t - \bar{X})}{\sum (X_t - \bar{X})^2} - \frac{\sigma^2}{\sum (X_t - \bar{X})^2} \\ &= 0. \end{aligned}$$

Cette dernière égalité étant due aux relations $\sum_{t=1}^{n} \lambda_t = 0$ et $\sum_{t=1}^{n} \lambda_t x_t = 1$. Ainsi,

$$Var(\beta) = Var(\beta - \hat{a}) + Var(\hat{a}).$$

Comme la variance est toujours positive, alors

$$Var(\beta) \geq Var(\hat{a}).$$

De façon similaire, on montre que $\hat{b}$ est linéaire et de variance minimale. □

2.3.2 Analyse de la variance

Pour définir l'équation de l'analyse de la variance des paramètres du modèle (2.1.1), on a besoin des propriétés données dans le lemme suivant :

Lemme 2.3.1 *Soit le modèle de régression simple (2.1.1) vérifiant les hypothèses citées ci-dessus, alors*

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/ $\sum_{t=1}^{n} Y_t = \sum_{t=1}^{n} \hat{Y}_t$: égalité entre la moyenne de la série à expliquer et la moyenne de la série ajustée.

Preuve. Soit le modèle de régression simple :$Y_t = b + aX_t + \varepsilon_t$, sachant que le résidu est :

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

où

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

1/ On sait que

$$\begin{aligned} Y_t &= \hat{Y}_t + e_t \\ &= \hat{b} + \hat{a}X_t + e_t. \end{aligned}$$

Donc

$$\sum_{t=1}^{n} Y_t = n\hat{b} + \hat{a}\sum_{t=1}^{n} X_t + \sum_{t=1}^{n} e_t$$

ce qui imlique

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

En remplaçant $\hat{b}$ par sa valeur, on obtient alors

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

Par conséquent

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

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

$$\begin{aligned} \sum_{t=1}^{n} e_t &= \sum_{t=1}^{n} \left(Y_t - \hat{Y}_t\right) = 0 \\ &\Longrightarrow \sum_{t=1}^{n} Y_t - \sum_{t=1}^{n} \hat{Y}_t = 0 \\ &\Longrightarrow \sum_{t=1}^{n} Y_t = \sum_{t=1}^{n} \hat{Y}_t. \end{aligned}$$

D'où $\bar{Y} = \overline{\hat{Y}}$. □

A partir des deux propriétés du Lemme (2.3.1), on peut déduire la fonction fondamentale d'analyse de la variance. On a

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

D'où

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

Cela implique

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

Passant aux sommes, on trouve

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

Comme

$$\sum_{t=1}^{n} e_t = 0 \text{ et } \sum_{t=1}^{n} Y_t = \sum_{t=1}^{n} \hat{Y}_t,$$

alors

$$\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right) = 0.$$

On déduit donc

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

D'où

$$SCT = SCE + SCR. \tag{2.3.5}$$

Définition 2.3.1 *L équation (2.3.5) est appelée l'équation de l'analyse de la variance, telle que*

1/ $SCT = \sum_{t=1}^{n}\left(Y_t - \bar{Y}\right)^2$: *désigne la variabilité totale.*

2/ $SCE = \sum_{t=1}^{n}\left(\hat{Y}_t - \bar{Y}\right)^2$: *désigne la variabilité expliquée.*

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

2.3.3 Coefficient de détermination

De l'équation (2.3.5), on peut déduire le coefficient de détermination

$$\frac{SCT}{SCT} = \frac{SCE}{SCT} + \frac{SCR}{SCT}$$

ce qui implique

$$1 = \frac{SCE}{SCT} + \frac{SCR}{SCT}.$$

D'où

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

Alors

$$r^2 = \frac{SCE}{SCT}.$$

Remarque 2.3.1 *Le coefficient de détermination permet de déterminer la qualité de l'ajustement linéaire en mesurant la part de la variation total de y expliquée par le modèle de régression sur x. Sa valeur est comprise entre 0 et 1 $(0 \leq r^2 \leq 1)$. Plus la valeur de r^2 est proche de 1, plus le modèle est plus significatif.*

Exemple 2.3.1 *On dispose des données qui sont représentés dans le tableau suivant :*

X_t (kg)	100	200	300	400	500	600	700
Y_t (DA)	40	50	50	70	65	65	80

où Y_t : Désigne les quantités consommées (en kg) et X_t : Désigne le prix des quantités consommées (en DA).

1/ *On trace un graphique des couples de données liant le prix et les quantités consommées. On obtient le nuage de points suivant :*

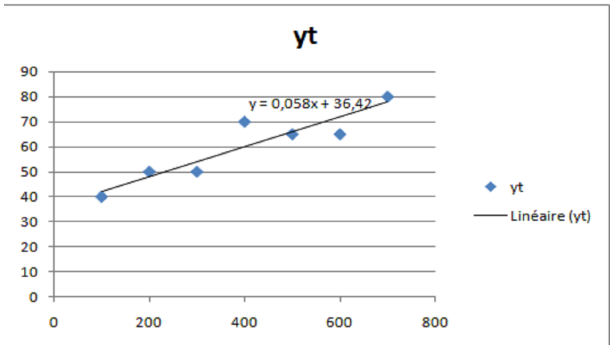

Figure 4 : Droite de rgression simple exp

2.2.1

2/ *Les estimateurs par la méthodes des moindres carrés de la pente* $\hat{a}$ *et de l'ordonnée à l'origine* $\hat{b}$ *de la droite de régression estimée d'équation* $\hat{Y}_t = \hat{b} + \hat{a}X_t$ *sont :*

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

et

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

On a $n = 7$, $\bar{X} = 400\ kg$ *et* $\bar{Y} = 60\ DA$.

$$\begin{aligned} Var\left(X\right) &= \overline{X^2} - \left(\bar{X}\right)^2 \\ Var\left(X\right) &= \frac{1}{7}\sum_{t=1}^{7}X^2 - \left(\bar{X}\right)^2 \\ Var\left(X\right) &= 200000 - \left(400\right)^2 \\ Var\left(X\right) &= 40000\ \left(kg\right)^2 \end{aligned}$$

et

$$\begin{aligned} Cov\,(X,Y) &= \overline{XY} - \bar{X}\ \bar{Y} \\ Cov\,(X,Y) &= \frac{1}{7}\sum_{t=1}^{7} X_t Y_t - \bar{X}\ \bar{Y} \\ Cov\,(X,Y) &= 26357,14286 - 24000 \\ Cov\,(X,Y) &= 2357,142857\ kg.DA \end{aligned}$$

On trouve, donc

$$\begin{aligned} \hat{a} &= \frac{2357,142857}{40000} \\ &= 0,058928571 \end{aligned}$$

et

$$\begin{aligned} \hat{b} &= \bar{Y} - \hat{a}\bar{X} \\ \hat{b} &= 60 - (0,0589 * 400) \\ \hat{b} &= 36,42857143 \end{aligned}$$

Par conséquent, la droite qui ajuste le nuage de point est :

$$(D) : \hat{Y}_t = 36,42 + 0,058X_t.$$

2.3.4 Tests des coefficients et les intervalles de confiance

Estimation de la variance des erreurs

Soit le modèle de régression simple :

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

Le résidu est donné par

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

où

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

et

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

On a

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

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

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

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

$$\begin{aligned} e_t &= (a - \hat{a})\, X_t - (a - \hat{a})\, \bar{X} + \varepsilon_t - \bar{\varepsilon} \\ &= (a - \hat{a}) \left(X_t - \bar{X}\right) + (\varepsilon_t - \bar{\varepsilon})\,. \end{aligned}$$

D'où

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

En élevant ce terme au carré et effectuant la somme sur les n observations, on trouve

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

Or d'après l'expression (2.3.4), on a

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

Cela conduit à

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

En passant à l'espérance mathématique, on obtient

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

On examine les deux membres de cette équation

a/

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

Comme $E\left(\varepsilon_t^2\right)=\sigma_\varepsilon^2$ (hypothèse 2)

$$E\left[\sum_{t=1}^{n}(\varepsilon_t-\bar{\varepsilon}\ )^2\right]=n\sigma_\varepsilon^2-\frac{1}{n}E(\varepsilon_1^2+\varepsilon_2^2+...+\varepsilon_n^2).$$

Or, d'après l'hypothèse 3 d'indépendance des erreurs, les doubles produits sont donc tous nuls. Par conséquent

$$\begin{aligned}
E\left[\sum_{t=1}^{n}(\varepsilon_t-\bar{\varepsilon}\ )^2\right] &= n\sigma_\varepsilon^2-\frac{1}{n}n\sigma_\varepsilon^2=n\sigma_\varepsilon^2-\sigma_\varepsilon^2\\
&= \sigma_\varepsilon^2(n-1)
\end{aligned}$$

b/

$$E\left[(a-\hat{a})^2\right]\sum_{t=1}^{n}x_t^2=\sigma_\varepsilon^2.$$

En effet

$$E\left[(a-\hat{a})^2\right] = \frac{\sigma_\varepsilon^2}{\sum_{t=1}^{n} x_t^2}.$$

D'où

$$\begin{aligned} E\left(\sum_{t=1}^{n} e_t^2\right) &= \sigma_\varepsilon^2(n-1) - \sigma_\varepsilon^2 \\ &= (n-2)\,\sigma_\varepsilon^2. \end{aligned}$$

Alors

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

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

Test des coefficients

En considérant le modèle de régression simple

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

On sait que

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{t=1}^{n} e_t^2}{n-2} \tag{2.3.6}$$

et ε_i suit la loi normale de moyenne nulle et de variance σ_ε^2, $\mathcal{N}(0, \sigma_\varepsilon^2)$.

De (2.3.6) on a

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

où χ_{n-2}^2 est la loi de Khi-deux de $(n-2)$ degrés de liberté.

D'autre part, on a

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

Par un changement de variables, on obtient les variables centrées réduites Z_1 et Z_2 suivant toutes les deux la loi normale centrée réduite $\mathcal{N}(0,1)$

$$Z_1 = \frac{\hat{a} - a}{\sqrt{\frac{\sigma_\varepsilon^2}{\sum_{t=1}^{n} x_t^2}}} \longrightarrow \mathcal{N}(0,1) \quad \text{et} \quad Z_2 = \frac{\hat{b} - b}{\sqrt{\sigma_\varepsilon^2\left(\frac{\sum_{t=1}^{n} X_t^2}{n\sum_{t=1}^{n}\left(X_t - \bar{X}\right)^2}\right)}} \longrightarrow \mathcal{N}(0,1)$$

D'après la définition de la loi de Student T qui est : le rapport d'une loi centrée réduite et la racine carrée d'une loi de khi-deux divisée par le nombre de ses degrés de liberté. On a

$$\begin{aligned} T_C &= \frac{Z_1}{\sqrt{(n-2)\,\hat{\sigma}_\varepsilon^2/\sigma_\varepsilon^2}/\sqrt{n-2}} \\ T_C &= \frac{\sigma_\varepsilon Z_1}{\hat{\sigma}_\varepsilon} = \frac{\sigma_\varepsilon}{\hat{\sigma}_\varepsilon}.\frac{\hat{a} - a}{\sigma_\varepsilon/\sqrt{\sum_{t=1}^{n} x_t^2}} \to T_{n-2}. \end{aligned}$$

Finalement,

$$T_C = \frac{\hat{a} - a}{\hat{\sigma}_{\hat{a}}^2} \to \left(T_{n-2}, \frac{\alpha\%}{2}\right) \tag{2.3.7}$$

où T_{n-2} est la loi de Student de $(n-2)$ degré de liberté et α est le taux (seuil) de risque. Suivant la même procédure pour $\hat{b}$, on obtient

$$T_C = \frac{\hat{b} - b}{\hat{\sigma}_{\hat{b}}^2} \to \left(T_{n-2}, \frac{\alpha\%}{2}\right) \tag{2.3.8}$$

A partir des résultats (2.3.7) et (2.3.8), on peut effectuer les tests d'hypothèses suivants :

$$\begin{cases} H_0 : b = 0 \\ contre \\ H_1 : b \neq 0 \end{cases} \quad \text{et} \quad \begin{cases} H_0 : a = 0 \\ contre \\ H_1 : a \neq 0 \end{cases}$$

Règle de décision et la construction du test des paramètres

*On considère un test bilatéral et on commence par le paramètre a.*Tester *a consiste à tester l'hypothèse suivante :*

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

Sous l'hypothèse $H_0 : a = 0$ *, on obtient la valeur critique* (T_C) *calculée à partir de l'expression*

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

où

* T_C : *Déigne la valeur critique de la statistique T (dite calculée).*

* $\hat{a}$: *Désigne la valeur estimée du paramètre a.*

* $\hat{\sigma}_{\hat{a}}$: *Désigne la valeur de l'écart-type du paramètre a.*

* α : *Le seuil donné.(En général, on prend* $\alpha = 5\%$*).*

* $n - 2$: *Degré de liberté.*

* T_t : *Désigne la valeur de la statistique Student T lue à partir de la table statistique.*

Règle de décision

- *Si* $|T_C| < T_t^{\alpha=0.05}$ *on accepte l'hypothèse* H_0 *: la variable* x_i *n'est pas contributive à l'explication de* Y*.*
- *Si* $|T_C| > T_t^{\alpha=0.05}$ *on accepte l'hypothèse* H_1 *: la variable* x_i *est contributive à l'explication de* Y*.*

De même, le test du paramètre b consiste à tester l'hypothèse suivante :

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

Sous l'hypothèse $H_0 : b = 0$ *, on obtient la valeur critique* T_C *tel que*

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

avec

* T_C : *Désigne la valeur critique de la statistique T (dite calculée).*

* $\hat{b}$: *Désigne la valeur estimée du paramètre b.*

* $\hat{\sigma}_{\hat{b}}$: *Désigne la valeur de l'écart-type du paramètre b.*

* α : *Le seuil donné.(Engénéral $\alpha = 5\%$).*

* $n-2$: *Degré de liberté.*

* T_t : *Désigne la valeur de la statistique Student T lue à partir de la table statistique.*

Règle de décision

– *Si* $|T_C| < T_t^{\alpha=0.05}$ *on accepte l'hypothèse* H_0 : *le modèle ne contient pas de constante.*
– *Si* $|T_C| > T_t^{\alpha=0.05}$ *on accepte l'hypothèse* H_1 : *le modèle contient la constante.*

2.3.5 Estmation des paramètres par des intervalles de confiance

Définition 2.3.2 *L'estimation par intervalle d'un paramètre inconnu θ consiste à calculer à partir d'un estimateur choisi $\hat{\theta}$, un intervalle $I_c = [LI,\ LS]$ dans lequel il est vraisemblable que la valeur correspondante du paramètre s'y trouve. La probabilité liée à cet intervalle de confiance est exprimée en pourcentage noté $1-\alpha$ où α est le taux de risque. La probabilité $1-\alpha$ est appelée le niveau de confiance ou le seuil de confiance*

$$\Pr\left(LI \leq \theta \leq LS\right) = 1-\alpha$$

où $1-\alpha$ est la probabilité associée à l'intervalle d'encadrer la vraie valeur du paramètre et LI (resp. LS) est la borne inférieure (resp. supérieure) de l'intervalle de confiance.

Remarque 2.3.2 *Le niveau de confiance est toujours associé à l'intervalle et non au paramètre inconnu θ.*

Intervalles de confiances des paramètres a et b

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

$\surd$ **Intervalle de confiance du paramètre a :**

$$\Pr\left[\hat{a} - T_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{a}} < a < \hat{a} + T_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{a}}\right] = 1 - \alpha$$

$\surd$ **Intervalle de confiance du paramètre** b :

$$\Pr\left[\hat{b} - T_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{b}} < b < \hat{b} + T_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{b}}\right] = 1 - \alpha$$

$\surd$ **Intervalle de confiance de** σ^2 : On a

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum_{t=1}^{n} e_t^2}{n-2} = S^2 \\ &\Longrightarrow \frac{(n-2)\,S^2}{\sigma^2} \to \chi^2_{n-2} \end{aligned}$$

D'où

$$\Pr\left[\chi_1^2 < \frac{(n-2)\,S^2}{\sigma^2} < \chi_2^2\right] = 1 - \alpha$$

$$\Pr\left[\frac{\chi_1^2}{(n-2)\,S^2} < \frac{(n-2)\,S^2}{\sigma^2} < \frac{\chi_2^2}{(n-2)\,S^2}\right] = 1 - \alpha.$$

Finalement

$$\Pr\left[\frac{(n-2)\,S^2}{\chi_2^2} < \sigma^2 < \frac{(n-2)\,S^2}{\chi_1^2}\right] = 1 - \alpha$$

2.3.6 Analyse de la variance et test de Fisher

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

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

Sous l'hypothèse $H_0 : b = 0$, On a

$$E\,(SCT) = (n-1)\sigma^2,\ E\,(SCE) = (1)\sigma^2 \text{ et } E(SCR) = (n-2)\,\sigma^2$$

avec $(n-1),\ (1)$ et $(n-2)$ sont des degrés de libertés de $SCT,\ SCE$, et SCR respectivement. D'autre part, lorsque H_0 est vérifiée on a

$$\frac{SCT}{n-1} \longrightarrow \chi^2_{n-1},\ \frac{SCE}{1} \longrightarrow \chi^2_1 \text{ et } \frac{SCR}{n-2} \longrightarrow \chi^2_{n-2}$$

Du moment que$\frac{SCE}{\sigma^2}$ et $\frac{SCR}{\sigma^2}$ sont indépendants, on peut déduire donc la valeur critique de Fisher F qui se définit comme suit : le rapport entre deux Chi-deux (χ) indépendants et leurs degrés de libertés.

Alors, on obtient :

$$\begin{aligned} F_C &= \frac{\frac{\frac{SCE}{\sigma^2}}{1}}{\frac{\frac{SCR}{\sigma^2}}{n-2}} = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} \\ &= \frac{SCE \times (n-2)}{SCR \times 1} \longrightarrow F_T(1, n-2, \alpha\%) \end{aligned}$$

avec

*/ F_C : désigne la valeur critique de Fisher calculée.

*/ $(1, n-2, \alpha\%)$: se sont des degrés de libertés.

*/ F_T : désigne la valeur de Fisher lue à partir de la table statistique de Fisher aux degrés de liberté $1,\ n-2$.

*/ $\alpha\%$: Le seuil donné.

Règle de décision

1. Si $|F_C| > F_T(1, n-2, \alpha\%)$ on rejette l'hypothèse H_0, cela signifie que, H_1 est acceptée : c'est-à-dire le modèle est globalement significatif.
2. Si $|F_C| < F_T(1, n-2, \alpha\%)$ on rejette l'hypothèse H_1, cela signifie que, H_0 est acceptée : c'est-à-dire le modèle n'est pas globalement significatif.

Tableau d'analyse de la variance

Source de la variance	Degré de liberté	Somme des carrés	Moyenne des carrés	F calculé
Régression	$n=1$	$SCE = \sum_{t=1}^{n} \left(\hat{Y}_t - \bar{Y}\right)^2$	$MC_{reg} = \frac{SCE}{1}$	$F_C = \frac{MC_{reg}}{MC_{res}}$
Résiduelle	$n-2$	$SCR = \sum_{t=1}^{n} \left(Y_t - \hat{Y}_t\right)^2$	$MC_{res} = \frac{SCR}{n-2}$	
Total	$n-1$	$SCT = \sum_{t=1}^{n} \left(Y_t - \bar{Y}\right)^2$		

Exemple 2.3.2 *On veut expliquer le rendement de mais Y (en quintal) à partir de la quantité d'angrais utilisé (en kg) sur des parcelles de terrain similaires.*

i	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
10	34	41
Moyenne	26.1	30.4

Le tableau ci-dessous représente les données "Rendements Agricoles"

i	Y	X	$\left(Y_t-\bar{Y}\right)$	$\left(X_t-\bar{X}\right)$	$\left(Y_t-\bar{Y}\right)\left(X_t-\bar{X}\right)$	$\left(X_t-\bar{X}\right)^2$
1	16	20	-10,1	-10,4	105,4	108,160
2	18	24	-8,1	-6,4	51,4	40,960
3	23	28	-3,1	-2,4	7,44	5,765
4	24	22	-2,1	-8,4	17,64	70,560
5	28	32	1,9	1,6	3,04	2,560
6	29	28	2,9	-2,4	-6,96	5,760
7	26	32	-0,1	1,6	-0,6	2,560
8	31	36	4,9	5,6	27,44	31.360
9	32	41	5,9	10.6	62.54	112.360
10	34	41	7,9	10.6	38.74	112.360
Moyenne	26,1	30,4		Somme	351,6	492,4

Estimation des coefficients "Rendements agricoles " voici les prancipales étapes :

1/ On calcule les moyennes des variables : $\bar{Y} = 26,1\ Q$ et $\bar{X} = 30,4\ kg$.

2/ On forme les valeurs de $\left(Y_t-\bar{Y}\right)$, $\left(X_t-\bar{X}\right)$, $\left(Y_t-\bar{Y}\right)\left(X_t-\bar{X}\right)$, $\left(X_t-\bar{X}\right)^2$.

3/ On réalise les sommes $\sum_{i=1}^{n}\left(Y_t-\bar{Y}\right)\left(X_t-\bar{X}\right) = 351,6\ Q\ Kg$ et $\sum_{i=1}^{n}\left(X_t-\bar{X}\right)^2 = 492,4\ kg^2$.

4/ On déduit enfin les estimations des parmètres du modèle de la régression simple Y en x

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(Y_t-\bar{Y}\right)\left(X_t-\bar{X}\right)}{\sum_{i=1}^{n}\left(X_t-\bar{X}\right)^2} = \frac{351,6\ Q}{492,4\ kg} = 0,7141\ Q/kg$$

et

$$\hat{b} = \bar{Y} - \hat{a}\bar{X} = 26,1 - 0,7141 \times 30,4 = 4,3928\ Q$$

La droite de régression peut être représentée dans le ghraphique nuage de points.

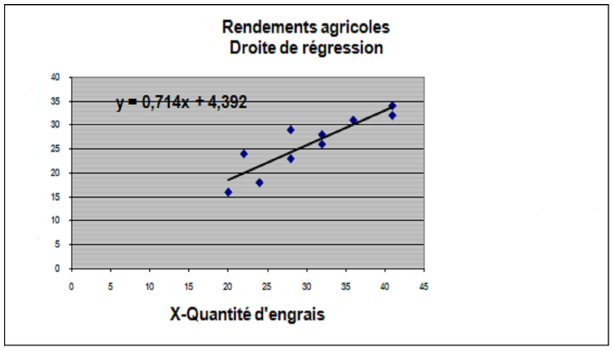

Figure 5 : Rendement agricole

Calculer la colonne des valeurs prédites $\hat{Y}$,en déduire le résidu $\hat{\varepsilon}$ et finalement obtenir les sommes des carrés .

i	Y	X	$(Y_t - \bar{Y})$	$(X_t - \bar{X})$	$(Y_t - \bar{Y})^2$	$\hat{Y}$	$\hat{\varepsilon}$	$\hat{\varepsilon}^2$
1	16	20	-10,1	-10,4	102,010	18,674	-2,674	7,149
2	18	24	-8,1	-6,4	65,610	21,530	-3,530	12,461
3	23	28	-3,1	-2,4	9,610	24,386	-1,386	1,922
4	24	22	-2,1	-8,4	4,410	20,102	3,898	15,195
5	28	32	1,9	1,6	3,610	27,242	0,758	0,574
6	29	28	2,9	-2,4	8,410	24,386	4,614	21,286
7	26	32	-0,1	1,6	0,010	27,242	-1,242	1,544
8	31	36	4,9	5,6	24,010	30,099	0,901	0,812
9	32	41	5,9	10,6	34,810	33,669	-1,669	2,785
10	34	41	7,9	10,6	62,410	33,669	0,331	0,110
Moyenne	$\underbrace{26,1}_{\bar{\mathbf{Y}}}$	$\underbrace{30,4}_{\bar{\mathbf{X}}}$		Somme	$\underbrace{314,9}_{\mathbf{SCT}}$		Somme	$\underbrace{63,838749}_{\mathbf{SCR}}$

On a

$$SCE = SCT - SCR = 251,061251,$$

le coefficient de détermination $R^2 = 0.79727295$ et le coéfficient de corrélation $R = 0.89290142$.

On calcule $\hat{Y}_i$. pour $(i = 1)$ **:** $\hat{Y}_1 = \hat{a} \times X_1 + \hat{b} = 0,17405 \times 20 + 4,39277 = 18,674$. On en déduit le résidu $\hat{\varepsilon}$: $\hat{\varepsilon}_1 = Y_1 - \hat{Y}_1 = 16 - 18,674 = -2,674$. Le coefficient de détermination est obtenu avec sa forme usuelle

$$R^2 = \frac{SCE}{SCT} = \frac{251,061}{314,900} = 0,797273.$$

Puis, le coefficient de corrélation linéaire simple

$$R = \sqrt{0,797273} = 0,892901.$$

Le tableau d'analyse de variance complet et le test F de significativité globale

Source de la variance	Degré de liberté	Somme des carrés	Moyenne des carrés	F calculé
Régression	$n = 1$	$SCE = \sum_{t=1}^{n} \left(\hat{Y}_t - \bar{Y}\right)^2 = 251,061251$	$MC_{reg} = \frac{SCE}{1} = 251,061251$	$F_C = \frac{MC_{reg}}{MC_{res}} = 31,46192618$
Résiduelle	$n - 2 = 10 - 2 = 8$	$SCR = \sum_{t=1}^{n} \left(Y_t - \hat{Y}_t\right)^2 = 63,83874898$	$MC_{res} = \frac{SCR}{n-2} = 7,979843623$	
Total	$n - 1 = 10 - 1 = 9$	$SCT = \sum_{t=1}^{n} \left(Y_t - \bar{Y}\right)^2 = 314,9$		

A ce stade , nous obtenons l'estimation de la variance de l'erreur, soit

$$\hat{\sigma}_\varepsilon^2 = \frac{SCR}{n-2} = \frac{63,839}{8} = 7,980.$$

L'écart -type estimé de l'erreur correspond à la racine carrée

$$\hat{\sigma}_\varepsilon = \sqrt{7,980} = 2,825.$$

Pour obtenir l'estimation de l'écart -type de la pente , nous avons besoin de la somme des écarts à la moyenne au carré des x

$$\begin{aligned}\hat{\sigma}_{\hat{a}} &= \sqrt{\frac{\hat{\sigma}_{\varepsilon}^2}{\sum_{t=1}^{n}\left(X_t-\bar{X}\right)^2}} \\ &= \sqrt{\frac{7,980}{492,4}} \\ &= \sqrt{0,01621} \\ &= 0,12730.\end{aligned}$$

On forme la statistique de test

$$T_{\hat{a}} = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}} = \frac{0,71405}{0,12730} = 5,60909.$$

Au risque $\alpha = 5\%$, le seuil critique pour la loi de Student à $(n-2)$ degrés de liberté pour un test bilatéral est $T_{1-\frac{\alpha}{2}} = 2,30600$. Puisque $|5,60909| > 2,30600$, on calcule que la pente est significativement nou nulle au risque 5%. On souhaite mettre en oeuvre le test d'hypothéses suivant pour les "Rendements agricoles "

$$\begin{cases} H_0 : a = 0.5 \\ H_1 : a > 0.5 \end{cases}$$

Il s'agait d'un test de conformité à un standard unilatéral. La région critique au risque α du test s'écrit

$$R.C : \frac{\hat{a}-0,5}{\hat{\sigma}_{\hat{a}}} > T_{1-\alpha}$$

Voyons ce qu'il en sur nos données

$$\frac{\hat{a}-0,5}{\hat{\sigma}_{\hat{a}}} = \frac{0,71405-0,5}{0,12730} = 1,68145$$

A comparer avec $T_{0,95}(8) = 1,85955$ pour un test à 5% .Nous sommes dans la région d'acceptation c-à-d .nous ne pouvons pas rejeter l'hypothèse nulle .

La valeur du paramétre a n'est pas significativement supérieur à la référence $0,5$ au risque 5%.

Intevalle de confiance au niveau 95% :

$$\left[\hat{a} - T_{\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{a}} \ ; \ \hat{a} + T_{\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{a}}\right]$$

$$[0,71405 - 2,30600 \times 0,12730 \ ; 0,71405 + 2,30600 \times 0,12730]$$

$$[0,42046 \ ; \ 1,00761]$$

Le résultat est cohérent avec le teste de significativité de la pente , l'intervalle de confiance ne contient pas la valeur 0.

Chapitre 3

Régression linéaire multiple

En pratique la variable expliquée ne dépend pas que d'une variable explicative. On introduit dans ce chapitre des modèles de régression linéaire avec plusieurs variables explicatives. On parle alors de la régression linéaire multiple.
La droite de régression que l'on avait en régression simple est ici remplacée par un hyperplan des régressions. Afin d'estimer les paramètres de cet hyperplan, on peut comme en régression simple, utiliser la méthode des moindres carrés.

3.1 Modèle de la régression multiple

Le modèle général de la régrssion multiple est une généralisation du modèle simple dans lequel figurent plusieurs variables explicatives

$$Y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + ... + b_k x_{kt} + \varepsilon_t, \quad \text{pour} \quad t = 1, ..., n \tag{3.1.1}$$

avec

1. Y_t = variable à expliquer a la date t.
2. x_{1t} = variable explicative 1 à la date t.
3. x_{kt} = variable explicative k à la date t.
4. $b_0, b_1, b_2, ..., b_k$: les paramètres du modèle.
5. ε_t = erreur de spécification, elle est inconnue.
6. n = nombre d'observations.

3.1.1 La forme matricielle du modèle de la régression multiple

L'écriture du modèle (3.1.1) est n'est pas très pratique. Afin de faciliter l'expression de certains résultats, on a habituellement recours aux notations matricielles. En écrivant le modèle (3.1.1), observation par observation, on obtient

$$\begin{cases} Y_1 = b_0 + b_1 x_{11} + b_2 x_{21} + ... + b_k x_{k1} + \varepsilon_1 \\ Y_2 = b_0 + b_1 x_{12} + b_2 x_{22} + ... + b_k x_{k2} + \varepsilon_2 \\ \qquad\qquad ... \\ Y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + ... + b_k x_{kt} + \varepsilon_t \\ Y_n = b_0 + b_1 x_{1n} + b_2 x_{2n} + ... + b_k x_{kn} + \varepsilon_n \end{cases}$$

On peut réécrire le modèle (3.1.1) sous la forme matricielle suivante :

$$\mathbf{Y}_{(n,1)} = \mathbf{X}_{(n,k+1)}\mathbf{B}_{(k+1,1)} + \boldsymbol{\varepsilon}_{(n,1)} \tag{3.1.2}$$

telle que

$$\mathbf{Y}_{(n,1)} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ . \\ . \\ \mathbf{Y}_t \\ . \\ \mathbf{Y}_n \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} b_0 \\ b_1 \\ . \\ . \\ b_t \\ . \\ b_k \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & . & . & x_{k1} \\ 1 & x_{12} & x_{22} & . & . & x_{k2} \\ . & . & . & . & . & x \\ . & . & . & . & . & . \\ 1 & x_{1t} & x_{2t} & . & . & x_{kt} \\ 1 & x_{1n} & x_{2n} & . & . & x_{kn} \end{pmatrix} \text{ et } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_t \\ \varepsilon_n \end{pmatrix}$$

On remarque que la première colonne de la matrice X, est le vecteur $\mathbf{1}_n$ (tous les termes sont 1), qui correspond au coefficient b_0 (coefficient du terme constant).

La matrice X est de n lignes et $k+1$ colonnes (k est le nombre de variables explicatives réelles, c'est-à-dire constante exclue) et n est le nombre d'observations.

Les hypothèses du modèle le modèle **(3.1.1)** repose sur les hypothèses suivantes :

1/ H_0 : Les valeurs x_{it} sont observées sans erreurs.

2/ $H_1 : E(\varepsilon_t) = 0$ espérance des erreurs est nulle.

3/ $H_2 : E(\varepsilon_t^2) = \sigma_\varepsilon^2$ la variance de l'erreur est constante $\forall(t)$.

4/ $H_3 : E(\varepsilon_t \ \varepsilon_t') = 0 \ si \ t \neq t'$ (indépendance des erreurs).

5/ $H_4 : Cov\,(x_{ij}, \varepsilon_i) = 0$, l'erreur est indépendante de x_{ij}.

6/ H_5 : Absence de colinéarité entre les variables explicatives $\Longrightarrow$ $(X\ ´X)$ régulière et $(X\ ´X)^{-1}$ existe.

7/ H_6 : $\left(\frac{X\ ' X}{n}\right)$: Tend vers une matrice finie non singulière.

8/ H_7 : $n > k+1$: nombre d'observations est supérieur aux nombre des explicatives.

3.2 Estimation des coefficients de régression

Soit le modèle sous forme matricielle à k variables explicatives et n observations

$$Y_t = X_t B + \varepsilon_t \tag{3.2.1}$$

Afin d'estimer le vecteur (B) composé des coefficients $b_0, b_1, b_2, ..., b_k$, on applique la méthode des moindres carrées ordinaire (MCO) qui consiste à minimiser la somme des carrées des erreurs. L'expression à minimiser sur B s'écrit

$$\begin{aligned}\sum_{i=1}^{n} \varepsilon_i^2 &= \sum_{i=1}^{n} \left[y_i - b_0 - b_1 x_{t1} - ... - b_k x_{tk}\right]^2 \\ &= \|Y - XB\|^2 \\ &= (Y - XB)'(Y - XB)\end{aligned}$$

où $(Y - XB)'$ est la matrice transposée de $(Y - XB)$. On pose

$$\begin{aligned}S &= (Y - XB)'(Y - XB) \\ &= Y'Y - 2B'X'Y + B'X'XB\end{aligned}$$

Par dérivation matricielle de S

$$\frac{\partial S}{\partial B} = -2X'Y + 2X'XB.$$

Comme $\min \sum_{i=1}^{n} \varepsilon_i^2$ est la valeur qui annule $\frac{\partial S}{\partial B}$, alors

$$\frac{\partial S}{\partial B} = 0 \Rightarrow -2X'Y + 2X'XB = 0$$

d'où

$$B = (X'X)^{-1} X'Y$$

car de l'hypothèse H_6, $(X'X)$ est inversible.

Le modéle estimé s'écrit :

$$\hat{Y}_t = \hat{b}_0 + \hat{b}_1 x_{1t} + \hat{b}_2 x_{2t} + ... + \hat{b}_k x_{kt} + e_t$$

avec

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

où e_t : est le résidu, c'est-à-dire l'écart entre la valeur observée de la variable à expliquer et sa valeur estimée (ajustée).

3.2.1 Propriétés des estimateurs du modèle de la régression multiple

1/ **$\hat{\mathbf{B}}$ est un estimateur sans biais**

Soit le modèle

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

On a

$$\begin{cases} Y = X\ \hat{B} + \varepsilon \\ \hat{Y} = X\ \hat{B} \end{cases} \Longrightarrow\ e = Y - \hat{Y}\ .$$

Alors

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

$$\hat{B} - B = (X'X)^{-1}X'\varepsilon \tag{3.2.2}$$

D'où

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

car $E(\varepsilon) = 0$ (hypothèse H_2). L'estimateur $\hat{B}$ est donc sans biais.

2/ **Estimateur de la variance de l'erreur**

On a

$$Var(\ \hat{B}) = E\left[(\ \hat{B} - B)(\ \hat{B} - B)^{'}\right] \tag{3.2.3}$$

Or d'après (3.2.2)

$$\hat{B} - B = (X^{'}X)^{-1}X'\varepsilon \ \text{ et } \ \left(\hat{B} - B\right)^{'} = \varepsilon' X(X'X)^{-1}.$$

ce qui conduit à

$$\left(\hat{B} - B\right)\left(\hat{B} - B\right)^{'} = (X^{'}X)^{-1}X'\varepsilon\varepsilon' X(X'X)^{-1}.$$

D'où

$$Var(\ \hat{B}) = (X^{'}X)^{-1}X'E\left(\varepsilon\varepsilon'\right)X(X'X)^{-1}$$

car $E\left(\varepsilon\varepsilon'\right) = \sigma_{\varepsilon}^2 I$: matrice des variances de l'erreur ε. En effet, d'après les hypothèses H_3 et H_4, on a

$$\begin{aligned} E\left(\varepsilon\varepsilon'\right) &= \begin{pmatrix} [E\left(\varepsilon_1\varepsilon_1\right)] & E\left(\varepsilon_1\varepsilon_2\right) & ... & E\left(\varepsilon_1\varepsilon_n\right) \\ E\left(\varepsilon_2\varepsilon_1\right) & E\left(\varepsilon_2\varepsilon_2\right) & ... & E\left(\varepsilon_2\varepsilon_n\right) \\ ... & & & ... \\ E\left(\varepsilon_n\varepsilon_1\right) & E\left(\varepsilon_n\varepsilon_2\right) & ... & E\left(\varepsilon_n\varepsilon_n\right) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{\varepsilon}^2 & 0 & ... & 0 \\ 0 & \sigma_{\varepsilon}^2 & ... & 0 \\ ... & & & ... \\ 0 & 0 & 0 & \sigma_{\varepsilon}^2 \end{pmatrix} \end{aligned}$$

Alors, $Var(\ \hat{B})$ devient

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

Enfin,

$$Var(\ \hat{B}) = \sigma_{\varepsilon}^2(X^{'}X)^{-1}. \tag{3.2.4}$$

Puisque σ_{ε}^2 est inconnu, on l'estime . Soit le modèle

$$Y_t = X\ B + \varepsilon_t$$

On a

$$e = Y - \hat{Y}$$

et

$$\hat{Y} = X\ \hat{B}.$$

D'où

$$e_t = Y - X\hat{B} \tag{3.2.5}$$

En remplaçant $\hat{B} = (X'X)^{-1}X'Y$ dans l'équation (3.2.5), on aura

$$e_t = Y - X(X'X)^{-1}X'Y = \left[I - X(X'X)^{-1}X'\right] Y.$$

On pose

$$M = I - X(X'X)^{-1}X'.$$

Alors

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

et

$$E(e'e) = \sigma_\varepsilon^2 I_n Tr\left[I - X(X'X)^{-1}X'\right] = \sigma_\varepsilon^2 I_n\left(n-k-1\right).$$

On obtient alors

$$\hat{\sigma}_\varepsilon^2 = \frac{e'e}{n-k-1}, n > k+1\ \text{(hypothsèse } H_8).$$

L'estimateur sans biais de la variance $\hat{\sigma}_\varepsilon^2$ est donné par

$$\begin{aligned} s^2 &= \frac{\sum_{t=1}^{n}\left(\hat{Y}_t - \bar{Y}\right)^2}{n-k-1} \\ &= \frac{SCE}{n-k-1} \end{aligned}$$

où SCE est la somme des carrés des résidus

$$SCE = \left\|\hat{Y}_t - \bar{Y}In\right\|^2$$

On définit également la somme totale des carrés

$$SCT = \left\|Y_t - \bar{Y}In\right\|^2 = Y'Y - n\left(\bar{Y}\right)^2,$$

et la somme des carrés de la régression

$$SCR = \|\hat{\varepsilon}\|^2 .$$

On peut aisément vérifier que

$$SCT = SCR + SCE.$$

3.2.2 Coefficient de détermination

Définition 3.2.1 *On appelle coefficient de détermination la quantité*

$$\begin{aligned} R^2 &= 1 - \frac{SCR}{SCT} \\ &= 1 - \frac{\sum_{t=1}^{n} \hat{\varepsilon}^2}{\sum_{t=1}^{n} \left(Y_t - \bar{Y}\right)^2} \\ &= 1 - \frac{\sum_{t=1}^{n} \left(\hat{Y}_t - \bar{Y}\right)^2}{\sum_{t=1}^{n} \left(Y_t - \bar{Y}\right)^2} \end{aligned}$$

Qui est donc la part de variation de Y expliquée par le modèle de régression.

Définition 3.2.2 *Le coefficient de détermination corrigé $\bar{R}^2$ est défini par :*

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{\frac{SCR}{(n-k-1)}}{\frac{SCT}{n-1}} \\ \bar{R}^2 &= 1 - \left(1 - R^2\right)\left(\frac{n-k-1}{n-1}\right) \end{aligned}$$

où n est le nombre d'observations et $(k+1)$ est nombre de séries explicatives.

Remarque 3.2.1 *Le coefficient de détermination corrigé $\bar{R}^2$ vérifie les propriétés suivantes*

1/ $\bar{R}^2 < R^2$ si $k > 1$.

2/ $R^2 = \bar{R}^2$ si $k = 1$.

3/ $\bar{R}^2$ peut prendre des valeurs négatives.

3.3 Les tests statistiques

3.3.1 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 on prend $\alpha = 0,05$. Le test bilatéral d'hypothèse est le suivant :

$$\begin{cases} H_0 : b_t = 0, \\ \quad contre \\ H_1 : b_t \neq 0. \end{cases}$$

La statistique de student est la suivante :

$$T_C = \left| \frac{\hat{b}_t - b_t}{\hat{\sigma}_{\hat{b}_t}} \right| \rightarrow S_t \left(n - k - 1, \frac{\alpha\%}{2} \right) \tag{3.3.1}$$

avec

*/ Tc : la valeur calculée,

*/ T : la valeur de la table de Student lue à partir de la table statistique,

*/ $n - k - 1$: degrés de liberté,

*/ $\frac{\alpha\%}{2}$: seuil de risque.

Règle de décision

Si $|T_C| < T^{\alpha=0,05}$ on accepte l'hypothèse H_0 : la variable x_t n'est pas contributive à l'explication de Y. Dans le cas contraire, on rejette H_0 et on accepte H_1.

3.3.2 Test de Fisher

Le test de Fisher est test de signification globale du modèle de regression. Pour tester si l'ensemble des variables explicatives ont une infuence sur la variable à expliquer, on fait le test d'hypothèse suivant :

$$\text{L'hypothèse nulle } \ H_0 : b_1, b_2, ..., b_k = 0, \ \text{ pour un seuil } \alpha\%$$

Contre

$$\text{L'hypothèse alternative } H_1 : \text{il existe au moins } b_i \neq 0$$

Sous H_0, on a

$$F_C = \frac{\frac{\sum_{t=1}^{n}\left(\hat{Y}_t-\bar{Y}\right)^2}{k}}{\frac{\sum_{t=1}^{n}\hat{\varepsilon}^2}{(n-k-1)}} = \frac{\frac{\sum_{t=1}^{n}\hat{y}_t^2}{k}}{\frac{\sum_{t=1}^{n}\hat{\varepsilon}^2}{(n-k-1)}}$$

$$F_C = \frac{\frac{R^2}{k}}{\frac{(1-R^2)}{(n-k-1)}} \longrightarrow F_\alpha\left(k; n-k-1, \alpha\%\right) \tag{3.3.2}$$

avec

*/ F_C : la valeur calculeé,

*/ F : la valeur lue de la table de Fisher,

*/ $(k, n-k-1)$: degrés de liberté .

*/ $\alpha\%$: seuil de risque .

Régle de décision

Si $|F_C| > F_t$ on rejette l'hypothèse H_0 et on accepte $H_1 \longrightarrow$ le modèle est globalement significatif. Dans le cas contraire, on accepte l'hypothèse $H_0 \longrightarrow$ le modèle n'est plus significatif.

Tableau d'analyse de la variance

Source de la variance	Degré de liberté	Somme des carrés	Moyenne des carrés	F calculé
Régression	k	$SCE = \sum_{t=1}^{n}\left(\hat{Y}_t - \bar{Y}\right)^2$	$MC_{reg} = \frac{SCE}{k}$	$F_C = \frac{MC_{reg}}{MC_{res}}$
Résiduelle	$n-k-1$	$SCR = \sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2$	$MC_{res} = \frac{SCR}{n-k-1}$	
Total	$n-1$	$SCT = \sum_{t=1}^{n}\left(Y_t - \bar{Y}\right)^2$		

3.4 Application

Soit le modéle à trois variables explicatives $x_1,\ x_2$ et x_3

$$Y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + b_3 x_{3t} + \varepsilon_t.$$

On dispose des données du tableau

1/

t	Y	x_1	x_2	x_3
1	12	2	45	121
2	14	1	43	132
3	10	3	43	154
4	16	6	47	145
5	14	7	42	129
6	19	8	41	156
7	21	8	32	132
8	19	5	33	147
9	21	5	41	128
10	16	8	38	163
11	19	4	32	161
12	21	9	31	172
13	25	12	35	174
14	21	7	29	180

1/ **Forme matricielle du modèle :** on dispose de 14 observations ($n = 14$) et trois variables explicatives, le modèle peut donc s'écrire : $Y = XB + \varepsilon$, avec

$$Y = \begin{pmatrix} 12 \\ 14 \\ 10 \\ \dots \\ 21 \end{pmatrix},\ X = \begin{pmatrix} 1 & 2 & 45 & 121 \\ 1 & 1 & 43 & 132 \\ 1 & 3 & 43 & 145 \\ \dots & \dots & \dots & \dots \\ 1 & 7 & 29 & 180 \end{pmatrix},\ B = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_{14} \end{pmatrix}.$$

2/ **Estimation des paramètre du modèle :** $\hat{B} = (X'X)^{-1}X'Y$, on calcule $X'X$ et $(X'X)^{-1}$

$$X'X = \begin{pmatrix} 1 & 1 & 1 & ... & 1 \\ 2 & 1 & 3 & ... & 7 \\ 45 & 43 & 43 & ... & 29 \\ 121 & 132 & 154 & ... & 180 \end{pmatrix} \begin{pmatrix} 1 & 2 & 45 & 121 \\ 1 & 1 & 43 & 132 \\ 1 & 3 & 43 & 145 \\ ... & ... & ... & ... \\ 1 & 7 & 29 & 180 \end{pmatrix}$$
$$= \begin{pmatrix} 14 & 85 & 532 & 2094 \\ 85 & 631 & 3126 & 13132 \\ 532 & 3126 & 20666 & 78683 \\ 2094 & 13132 & 78683 & 317950 \end{pmatrix}$$

et

$$(X'X)^{-1} = \begin{pmatrix} 20,1864 & 0,015065 & -0,23145 & -0,07617 \\ 0,015065 & 0,013204 & 0,001194 & -0,00094 \\ -0,23145 & 0,001194 & 0,003635 & 0,000575 \\ -0,07617 & -0,00094 & 0,000575 & 0,000401 \end{pmatrix}.$$

On a aussi

$$X'Y = \begin{pmatrix} 1 & 1 & 1 & ... & 1 \\ 2 & 1 & 3 & ... & 7 \\ 45 & 43 & 43 & ... & 29 \\ 121 & 132 & 154 & ... & 180 \end{pmatrix} \begin{pmatrix} 12 \\ 14 \\ 10 \\ ... \\ 21 \end{pmatrix} = \begin{pmatrix} 248 \\ 1622 \\ 9202 \\ 37592 \end{pmatrix}.$$

Donc

$$\hat{B} = \begin{pmatrix} 20,1864 & 0,015065 & -0,23145 & -0,07617 \\ 0,015065 & 0,013204 & 0,001194 & -0,00094 \\ -0,23145 & 0,001194 & 0,003635 & 0,000575 \\ -0,07617 & -0,00094 & 0,000575 & 0,000401 \end{pmatrix} \begin{pmatrix} 248 \\ 1622 \\ 9202 \\ 37592 \end{pmatrix}$$
$$= \begin{pmatrix} 32,89132 \\ 0,801900 \\ -0,38136 \\ -0,03713 \end{pmatrix}.$$

On prend $\hat{b}_0 = 32,89$, $\hat{b}_1 = 0,80$, $\hat{b}_2 = -0,38$, et $\hat{b}_3 = -0,03$.

3/ Calcul de $\hat{\sigma}^2_\varepsilon$ et de $\hat{\sigma}^2_{\hat{B}}$: On commence tout d'abord par calculer le résidu e

$$e = Y - \hat{Y} = Y - X\ \hat{B}$$
$$= Y_t - \left(\hat{b}_0 + \hat{b}_1 x_{1t} + \hat{b}_2 x_{2t} + \hat{b}_3 x_{3t}\right).$$

En remplçant les paramètres $\hat{b}_t$ par leurs valeurs trouvées ci-dessus, on trouve

$$e_t = \ Y_t - \ 32,89 - 0,80x_{1t} + 0,38x_{2t} + 0,03x_{3t}).$$

Les résultats numériques sont consignés dans le tableau ci-dessous

t	Y	$\hat{Y}_t$	e_t	e_t^2
1	12	12,84	-0,80	0,70
2	14	12,39	1,61	2,58
3	10	13,18	-3,18	10,11
4	16	13,39	1,61	2,58
5	14	17,70	-3,70	13,67
6	19	17,88	1,12	1,26
7	21	22,20	-1,20	1,44
8	19	18,86	0,14	0,02
9	21	16,51	4,49	20,14
10	16	18,76	-2,70	7,63
11	19	17,92	1,08	1,17
12	21	21,90	-0,90	0,81
13	25	22,71	2,90	5,27
14	21	20,76	0,24	0,06
/	/	/	$\sum_{t=1}^{14} e_t = 0$	$\sum_{t=1}^{14} e_t^2 = 67,45$

Il est clair que la somme des résidus est bien nulle et

$$\hat{\sigma}_\varepsilon^2 = \frac{e'e}{n-k-1} = \frac{\sum_{t=1}^{14} e_t^2}{14-3-1} = \frac{67,45}{10} = 6,745.$$

La matrice des variances donnée par (3.2.4) est :

$$\begin{aligned} Var(\ \hat{B}) &= \hat{\sigma}_\varepsilon^2 \left((X'X)^{-1}\right) \\ &= 6,745 \times \begin{pmatrix} 20,1864 & 0,015065 & -0,23145 & -0,07617 \\ 0,015065 & 0,013204 & 0,001194 & -0,00094 \\ -0,23145 & 0,001194 & 0,003635 & 0,000575 \\ -0,07617 & -0,00094 & 0,000575 & 0,000401 \end{pmatrix}. \end{aligned}$$

Les variances des coefficient de régression se trouvent sur la diagonale

$$\begin{aligned}
\hat{\sigma}^2_{\hat{b}_0} &= 6,745 \times 20,17 = 136,04 \Rightarrow \hat{\sigma}_{\hat{b}_0} = 11,66 \\
\hat{\sigma}^2_{\hat{b}_1} &= 6,745 \times 0,013 = 0,087 \Rightarrow \hat{\sigma}_{\hat{b}_1} = 0,29 \\
\hat{\sigma}^2_{\hat{b}_2} &= 6,745 \times 0,0036 = 0,024 \Rightarrow \hat{\sigma}_{\hat{b}_2} = 0,15 \\
\hat{\sigma}^2_{\hat{b}_3} &= 6,745 \times 0,0004 = 0,0026 \Rightarrow \hat{\sigma}_{\hat{b}_3} = 0,05
\end{aligned}$$

4/ Le coefficient de détermination R^2 : d'après ce qui précède, on a $\sum_{t=1}^{14} e_t^2 = 67,45$, et $\sum_{t=1}^{14} \left(Y_t - \bar{Y}\right)^2 = 226,86$. Alors

$$R^2 = 1 - \frac{\sum_{t=1}^{14} e_t^2}{\sum_{t=1}^{14} \left(Y_t - \bar{Y}\right)^2} = 1 - \frac{67,45}{226,86} = 0,702$$

Le $\bar{R}^2$corrigé est donné par :

$$\begin{aligned}
\bar{R}^2 &= 1 - \left(1 - R^2\right)\left(\frac{n-k-1}{n-1}\right) \\
&= 1 - (1 - 0,702)\left(\frac{14-1}{14-4}\right) = 0,613
\end{aligned}$$

On remarque la baisse du coefficient de détermination lorsque on le corrige par le degré de liberté.

5/ Test Student au seuil de 5% (pour la valeur $\alpha = 5\%$) , soit :

$$T_{\hat{b}_1} = \frac{\hat{b}_1}{\hat{\sigma}_{\hat{b}_1}} = \frac{0,80}{0,29} = 2,75$$

et la valeur lue de la table de Student est $T_{10}^{0,05} = 2,228$. Comme $T_{\hat{b}_1} > T_{10}^{0,05}$, alors l'hypothèse H_o est rejetée. Par conséquent H_1 est acceptée $b_1 \neq 0$. On déduit donc que la variable explicative x_1 est contributive à l'explication de Y, de même pour les autres paramètres du modèle

*/ $\frac{\hat{b}_2}{\hat{\sigma}_{\hat{b}_2}} = \left|\frac{-0,38}{0,15}\right| = T_{\hat{b}_2} = 2,53 > T_{10}^{0,05} = 2,228 \rightarrow b_2 \neq 0$ est l'hypothèse acceptée.

*/ $\frac{\hat{b}_3}{\hat{\sigma}_{\hat{b}_3}} = \left|\frac{-0,03}{0,05}\right| = T_{\hat{b}_3} = 0,60 < T_{10}^{0,05} = 2,228 \rightarrow b_3 \neq 0$ est l'hypothèse rejetée $\Rightarrow b_3 = 0$ est l'hypothèse acceptée.

Remarque 3.4.1 *La variable x_2 est explicative de Y alors que la variable x_3 n'est pas contributive à l'explication de Y, il convient donc de la retirer de ce modèle et de procéder à une nouvelle estimation .*

6/ Les intervalles de confiance de chacun des coefficients :

$$\begin{aligned} IC_{b_1} &= \left[\hat{b}_1 - \hat{\sigma}_{\hat{b}_1} T_{n-k-1}^{0.05};\ \hat{b}_1 + \hat{\sigma}_{\hat{b}_1} T_{n-k-1}^{0.05}\right] = [0,14; 1.45]\,. \\ IC_{b_2} &= \left[\hat{b}_2 - \hat{\sigma}_{\hat{b}_2} T_{n-k-1}^{0.05};\ \hat{b}_2 + \hat{\sigma}_{\hat{b}_2} T_{n-k-1}^{0.05}\right] = [-0,17; -0,04]\,. \\ IC_{b_3} &= \left[\hat{b}_3 - \hat{\sigma}_{\hat{b}_3} T_{n-k-1}^{0.05};\ \hat{b}_3 + \hat{\sigma}_{\hat{b}_3} T_{n-k-1}^{0.05}\right] = [-0,14; 0,08]\,. \end{aligned}$$

La valeur 0 n'appartient pas à l'intervalle de confiance à 95% de b_1 et b_2, donc ces deux coefficients sont significativement différents de 0, en revanche, 0 appartient à l'intervalle de confiance de b_3, ce coefficient n'est pas significativement différent de 0.

7/ Le test de Fisher de la signification globale de la régression

$$F_C = \frac{\frac{R^2}{k}}{\frac{(1-R^2)}{(n-k-1)}} = \frac{\frac{0,702}{3}}{\frac{(1-0,702)}{10}} = 7,878 > F_{(3,10)}^{0.05} = 3,71$$

$F_{(3,10)}^{0.05}$ est la valeur lue de la table de Fisher à $(3, 10)$ degrés de liberté et au seuil $\alpha = 5\%$. On rejette l'hypothèse H_0 de nullité de tous les coefficients. Alors, la régression est globalement significative.