République Algériennne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Départementde Mathématiques

Mémoire présenté pour obtenir le diplôme de

Master en "**Mathematiques Appliquées**"

Option :**Statistique**

par HAFFAS Nesrine

Titre :

Méthode de moindre carré et quelques estimateurs

Devant le Jury :

Mr.	Meraghni Djamal	Pr	U. Biskra	Président
Mr.	Necir Abdelhakim	Pr	U. Biskra	Rapporteur
Mme.	Soltane Louiza	Dr	U. Biskra	Examinatrice

Soutenu Publiquement le 27/06/2022

Dédicace

Je dédie ce travail à mes chers adorables parents. Merci d'avoir toujours été de mon coté pour m'aider et merci aussi pour l'amour sans faille que vous n'avez cessé de m'apporter. Vos conseils ont toujours permis d'être sur le droit chemin. Ce mémoire et un des fruits de tous vos efforts et sacrifices que vous ne cessez de me les apporter.

Je didie aussi ce travail :

A mon mari Mr ALI MESSEDEK pour son soutien et pour l'amour qui ma toujours donner.

A ma fille Jasmine, ma tante Yamina, mes soeurs Loubna, Roumissa, Iman et Hadjar, merci beaucoup pour votre soutien et votre aide.

A mes frères Abdul Malek et Muhammad Al-Amin,

A mon oncle Arabe.

Je vous remercie tous pour votre soutien. Je suis tellement chanceux que vous soyez tous autour de moi.

Remerciements

Avant tout, je tiens à remercier Allah le tout puissant de m'à avoir donné la santé, la patience, la force et le courage pour arriver là où je suis.

Je tiens à remercier vivement mon promoteur le Professeur Necir Abdelhakim pour sa disponibilité et sa sympathie tout au long de la réalisation de ce mémoire.

Je remercie tous les membres de ma famille qui m'ont accompagné tout au long de mes études par leur amour inconditionnel et leur soutien constant.

Je tiens à remercier les membres du jury qui m'ont fait l'honneur de participer à l'examen de ce travail.

Je remercie enfin tous ceux qui ont participé de près ou de loin à l'accomplissement de ce travail.

Résumé du mémoire

Dans ce mémoire, nous proposons .....

Notations et symbols

rlm	régression linéaire multiple
ε	terme d'erreur.
$\| \, . \, \|$	Norme
E	.Espérance
(Ω, A, P)	espace probabilisé
emco	Estimateur des moindres carrés ordinaire.
(O, I, J)	repère orthonormé
$H(.)$	la matrice hessienne
e_j	vecteur colonne
$C_n(.,.)$	Covariance
rls	régression linéaire simple
$N_n(\mu; \Sigma)$	la loi normale multidimensionnelle
emv	estimateur du maximum de vraisemblance
$L(\beta; z)$	la fonction de vraisemblance
L	sous-espace vectoriel de R^n
Σ	matrice de covariance
N	loi normale

Table des matières

Table des figures

Liste des tableaux

Introduction

La méthode des moindres carrés est une méthode ancienne datant du dix-huitième Siècle ; elle a été développée par le célèbre mathématicien Gauss.

Dans le cadre de ces travaux sur les planètes, Gauss a développé cette méthode pour déterminer les orbites des planètes à partir des positions observées. Gauss a eu l'idée d'approximer l'orbite d'une planète par un modèle mathématique dont les paramètres sont ajustables. En effet, cette approximation est réalisée par la minimisation de la somme des carrés des erreurs, dite "erreur d'estimation", entre les positions observées et celles générées par le modèle mathématique. Au fil du temps, cette méthode a reçu une très grande attention de la communauté Scientifique.

La méthode des moindres carrés est le nom technique de la régression mathématique en statistiques et plus particulièrement de la régression linéaire. Il s'agit d'un modèle couramment utilisé en économétrie. Elle consiste à minimiser la somme des carrés des écarts, écarts pondérés dans le cas multidimensionnel, entre chaque point du nuage de régression et son projeté, parallèlement à l'axe des ordonnées, sur la droite de régression. Il s'agit d'ajuster un nuage de points $\{Y_i, X_i\}$, $i = 1, ..., n$ selon une relation linéaire, prenant la forme de la relation matricielle $Y = X\beta + \varepsilon$, où ε certainement la plus utilisée. C'est pourquoi nous avons, dans

ce travail, rappeler les concepts fondamentaux de base utilisés dans cette méthode et avons travaillé sur une méthode de représentation permettant d'en visualiser les principaux résultats.

Ce mémoire est composé de deux chapitres :

Le premier chapitre est une synthèse sur la régression multiple et les méthodes d'estimation correspondantes.

Dans deuxième chapitre nous considérons la régression simple à fin d'appliquer la méthode des moindre carrés et les techniques d'estimation correspondantes.

Nous présentons ici des exemples avec des données réelles illustrant cette méthode. Tous mos calculs sont réalisés à l'aide des deux langages R et SPSS.

Chapitre 1

Modèle de régression linéaire multiple et estimateur des moindres carrés ordinaire :

1.1 Modèle de régression linéaire multiple(rlm)

1.1.1 forme générique :

On souhait prédire et/ou expliquer les valeurs d'une variable quantitative Y à partir des valeurs de p variables $X_1, ..., X_p$. On dit alors que l'on souhaite "expliquer Y à partir de $X_1, ..., X_p$". La variable aléatoire Y est appelée "variable à expliquer" et $X_1, ..., X_p$ sont appelées "variables explicatives".

Pour ce fait,on dispose de données qui sont n observation de $(Y, X_1, ..., X_p)$ notées $(y_1, x_{1,1}, ..., x_{p,1})$, $(y_2, x_{1,2}, ..., x_{p,2}), ..., (y_1, x_{1,n}, ..., x_{p,n})$. Elles se présentent généra-

lement sous la forme d'un tableau :

Y	X_1	$\cdots$	X_p
y_1	$x_{1,1}$	$\cdots$	$x_{p,1}$
y_2	$x_{1,2}$	$\cdots$	$x_{p,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	$x_{1,n}$	$\cdots$	$x_{p,m}$

Si une liasion linéaire entre Y et $X_1, ..., X_p$ envisageable, on peut considérer le modèle de régression linéaire multiple (rlm). Sa forme générique est :

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon, \tag{1.1}$$

où $\beta_0, ..., \beta_p$ sont des coefficients réels inconnus et ε est une variable quantitative de valeur moyenne nullle indépendante de $X_1, ..., X_p$ qui représente une somme d'erreurs aléatoires et multifactorielles (erreurs de mesures, effets non prévisibles, variables omises).

Remarque 1.1.1 *Note principal objectif est d'estimé convenablement $\beta_0, ..., \beta_p$ à l'aide des données. Entre autres, cela nous permettra de mesurer l'importance des variable $X_1, ..., X_p$ dans l'explication de Y et de prédire avec précision la valeur moyenne de Y pour une nouvelle valeur de $(X_1, ..., X_p)$.*

[3] **Loyers** :On peut considérer le jeu de données "loyers" :

Dans un quartier parisien, une étude a été menée afin de mettre en évidence une relation entre le loyer mensuel et la surface des appartements ayant exactement 3 pièces pour 30 appartements de ce type, on dispose :

– de la surface en mètres carrés (variable X_1).

– du loyer mensuel en francs (variable Y).

Y	3000, 2844, 3215, 2800, 3493, 3140, 3593, 3688, 4452, 3361 4542, 3181, 3575, 4017, 3430, 4226, 3639, 4015, 4741, 3628 4426, 3390, 4766, 4413, 4900, 5020, 4962, 5294, 4847, 5549.
X_1	44, 44, 44, 45, 45, 48, 49, 52, 53, 54 54, 54, 55, 57, 57, 59, 61, 62, 62, 63 63, 64, 67, 69, 70, 72, 74, 75, 76, 88.

TAB. 1.1 – le jeu de données "loyers"

–

– Le tableau représente les données la surface en mètres carrés (variable $X1$). et loyer mensuel en francs (variable Y)obtenues à partir d'une étude pour 30 appartements de ce type.

[4] **Fromages** : On peut considérer le jeu de données "fromages" :

Le goût d'un fromage dépend de la concentration de plusieurs composés chimiques, dont :

– la concentration de l'acide acétique (variable $X1$).

– la concentration d'hydrogène sulfuré (variable $X2$).

– la concentration d'acide lactique (variable $X3$).

Pour 30 types de fromage, on dispose du score moyen attribué par des consommateurs (variable Y).On souhaite expliquer Y à partir de $X1, X2$ et $X3$.

Table 2 : Le tableau représente les données du score moyen attribué par des consommateurs (variable Y) de formages etla concentration de l'acide acétique (variable $X1$), la concentration d'hydrogène sulfuré (variable $X2$),la concentration d'acide lactique (variable $X3$)

obtenues à partir d'une étude pour Pour 30 types de fromage.

	4.543; 5.159; 5.366; 5.759; 4.663; 5.697; 5.892; 6.078; 4.898; 5.242
X_1	5.74; 6.446; 4.477; 5.236; 6.151; 6.365; 4.787; 5.412; 5.247; 5.438
	4.564; 5.298; 5.455; 5.855; 5.366; 6.043; 6.458; 5.328; 5.802; 6.176.
	3.135; 5.043; 5.438; 7.496; 3.807; 7.601; 8.726; 7.966; 3.85; 4.174
$X2$	6.142; 7.908; 2.966; 4.942; 6.752; 9.588; 3.912; 4.7; 6.147; 9.064
	10.199; 3.664; 3.219; 6.962; 3.912; 6.685; 4.787
	0.86; 1.53; 1.57; 1.81; 0.99; 1.09; 1.29; 1.78; 1.29; 1.58
$X3$	1.68; 1.9; 1.06; 1.3; 1.52; 1.74; 1.16; 1.49; 1.63; 1.99
	1.15; 1.33; 1.44; 2.01; 1.31; 1.46; 1.72; 1.25; 1.08; 1.25
	12.3; 20.9; 39; 47.9; 5.6; 25.9; 37.3; 21.9; 18.1; 21.9
Y	34.9; 57.2; 0.7; 25.9; 54.9; 40.9; 15.9; 6.4; 18; 38.9
	14; 15.2; 32; 56.7; 16.8; 11.6; 26.5; 0.7; 13.4; 5.5

TAB. 1.2 – le jeu de données "fromages"

1.2 Modèle de rlm

On modélise les variables considérées comme des variables aléatoires réelles (var) (définies sur un espace probabilisé (Ω, A, P)) en gardant les mêmes notations par convention. À partir de celles-ci, le modèle de rlm est caractérisé pour tout $i \in \{1, ..., n\}$.

– $(x_{1,i}, ..., x_{p,i})$ est une réalisation du vecteur aléatoire réel $(X_1, ...X_P)$.

– sachant que $(X_1, ..., X_P) = (x_{1,i}, ..., x_{p,i}), y_i$ est une réalisation de

$$Y = \beta_0 + \beta_1 x_{1,i} + ... + \beta_P x_{p,i} + \varepsilon_i, \tag{1.2}$$

où ε_i est une var indépendante de $X_1, ..., X_P$ avec $E(\varepsilon_i) = 0$.

D'autres hypothèses sur $\varepsilon_1, ..., \varepsilon_n$ seront formulées ultérieurement.

Écriture matricielle du modèle de rlm

Le modèle de rlm peut alors s'écrire sous la forme matricielle : $Y = X\beta + \varepsilon$, où :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{bmatrix} 1 & x_{1,1} & ... & x_{p,1} \\ 1 & x_{1,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{bmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Estimateur des moindres carrés ordinaire :

Soient $\| \, . \, \|$ la norme euclidienne : pour tout vecteur colonne $\| \, x \, \|^2 = x^t x =$ somme des carrés des composantes de x. Partant du modèle de rlm écrit sous la forme matricielle : $Y = X\beta + \varepsilon$. pour tout, $u = \begin{pmatrix} u_0 \\ \vdots \\ u_p \end{pmatrix}$, on considère la fonction :

$$f\,(u) = \| \, Y - Xu \, \|^2 = \sum_{i=1}^{n} (Y_i - (u_0 + u_1 x_{1,i} + ... + u_p x_{p,i}))^2. \tag{1.3}$$

Alors, un estimateur des moindres carrés ordinaires (emco), noté $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$, vérifie $f\left(\hat{\beta}\right)$ est minimale. L'idée globale est "Y_I et $\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + ... + \hat{\beta}_p x_{p,i}$ aussi proche que possible".

Estimateur des moindres carrés ordinaire

Si l'on suppose que X est de rang colonnes plein : il n'existe pas de vecteur colonne x à $p+1$ composantes non nul tel que $Xx =$ le vecteur nul (cela entraîne l'existence

de $(X^tX)^{-1}$). Alors $\hat{\beta}$ est unique ; il est donné par la formule $\hat{\beta} = (X^tX)^{-1}X^tY$.

Preuve. Par définition, $\hat{\beta}$ est un extremum de $f(u)$,et $\hat{\beta}$ extremum de $f(u) \Rightarrow \frac{\partial}{\partial u_j} f(\hat{\beta}) = 0, j \in \{0, ..., p\}$. Simplifions l'écriture de $f(u)$. En utilisant les formules : $(A+B)^t = A^t + B^t$ et $(AB)^t = B^tA^t$, il vient :

$$\begin{aligned} f(u) =& \| Y - Xu \|^2 = (Y - Xu)^t(Y - Xu) = (Y^t - (Xu)^t)(Y - Xu) \\ &= (Y^t - u^tX^t)(Y - Xu) = YY^t - Y^tXu - u^tX^tY + u^tX^tXu. \end{aligned} \tag{1.4}$$

Comme Y^tXu est la multiplication d'un vecteur ligne Y^t par un vecteur colonne Xu, c'est un réel. Par conséquent, il est égal à sa transposé ; on a $Y^tXu = (Y^tXu)^t = (Xu)^t(Y^t)^t = u^tX^tY$,il vient $f(u) = Y^tY - 2u^tX^tY + u^tX^tXu$.Pour tout $j \in \{0, ..., p\}$ déterminons la dérivée partielle $\frac{\partial}{\partial u_j} f(u)$. Soit e_j le vecteur colonne à $p+1$ composantes avec p composantes nulles, sauf la $j+1-$ème qui vaut 1. ■

En utilisant la formule : $(u(x)v(x))' = u'(x)v(x) + u(x)v'(x)$ il vient :

$$\begin{aligned} \frac{\partial}{\partial u_j} f(u) &= \frac{\partial}{\partial u_j}(Y^tY - 2u^tX^tY + u^tX^tXu) \\ &= \frac{\partial}{\partial u_j}(Y^tY) - 2\frac{\partial}{\partial u_j}(u^tX^tY) + \frac{\partial}{\partial u_j}(u^tX^tXu) \\ &= 0 - 2e_j^tX^tY + e_j^tX^tXu + u^tX^tXe_j. \end{aligned} \tag{1.5}$$

Comme $e_j^tX^tXu$ est la multiplication d'un vecteur ligne $e_j^tX^t$ par un vecteur colonne Xu, c'est un réel. Pa rconséquent, il est égal à sa transposé ; on a $e_j^tX^tXu = (e_j^tX^tXu)^t = (Xu)^t(X^te_j^t)^t = u^tX^tXe_j$. Donc.$\frac{\partial}{\partial u_j} f(u) = -2e_j^tX^tY + 2e_j^tX^tXu$.Il s'ensuit $\frac{\partial}{\partial u_j} f(\hat{\beta}) = 0 \Longleftrightarrow -2e_j^tX^tY + 2e_j^tX^tX\hat{\beta} = 0 \Leftrightarrow e_j^tX^tX\hat{\beta} = e_j^tX^tY$.Comme cela est vraie pour tout $j \in \{0, ..., p\}$ et que et $e_j^tX^tX\hat{\beta}$ calcule la $j+1$-ème ligne de la matrice $X^tX\hat{\beta}$,il vient $\frac{\partial}{\partial u_j} f(\hat{\beta}) = 0, j \in \{0, ..., p\} \Leftrightarrow X^tX\hat{\beta} = X^tY$ Comme

$(X^tX)^{-1}$ existe, l'égalité $(X^tX)^{-1}X^tX = I_{p+1}$ entraîne.

$$X^tX\hat{\beta} = X^tY \Leftrightarrow (X^tX)^{-1}X^tX\hat{\beta} = (XtX)^{-1}X^tY \Longleftrightarrow \hat{\beta} = (X^tX)^{-1}X^tY. \quad (1.6)$$

Ainsi, on a $\hat{\beta}$ est un extremum de $f(u)$ si $\hat{\beta} = (X^tX)^{-1}X^tY$.Il reste à montrer que $\hat{\beta}$ est bien un minimum pour $f(u)$. Pour cela, on calcule la matrice hessienne $H(f) = (\frac{\partial^2}{\partial u_j \partial u_k} f(u))_{(j,k)\in\{0,....,p\}^2}$,et on montre qu'elle est définie positive : pour tout vecteur colonne non nul x à $p+1$ composantes, on a $x^tH(f)x > 0$. Pour tout $(j,k) \in \{0,\ldots,p\}^2$ on a :

$$\begin{aligned} \frac{\partial^2}{\partial u_j \partial u_k} f(u) &= \frac{\partial}{\partial u_k}(\frac{\partial}{\partial u_j} f(u)) = \frac{\partial}{\partial u_k}(-2e_j^tX^tY + 2e_j^tX^tXu) \quad (1.7)\\ &= -2\frac{\partial}{\partial u_k}(e_j^tX^tY) + 2\frac{\partial}{\partial u_k}(e_j^tX^tXu) \\ &= 0 + 2e_j^tX^tXe_k = 2e_j^tX^tXe_k. \quad (1.8)\end{aligned}$$

donc $H(f) = (2e_j^tX^tXe_k)_{(j,k)\in\{0,\ldots,p\}^2} = .2X^tX$.Pour tout $x = \begin{pmatrix} x_0 \\ \vdots \\ x_p \end{pmatrix}$ non nul, comme X est de rang colonnes plein, on a :

$$x^tH(f)x = x^t(2X^tX)x = 2x^tX^tXx = 2(Xx)^tXx = 2 \parallel Xx \parallel^2 > 0. \quad (1.9)$$

Ainsi,$H(f)$ est définie positive ; $\hat{\beta}$ est bien un minimum pour $f(u)$. On en déduit que $\hat{\beta} = (X^tX)^{-1}X^tY$.

1.3 Emco de β_J

Pour tout $j \in \{0, ..., p\}$, l'emco de β_j est $\hat{\beta}_j$. Dorénavant, $\hat{\beta}$ désignera l'emco de $\hat{\beta}$ et $\hat{\beta}_j$ l'emco de β_j.

1.4 Estimateur de la valeur moyenne

On appelle valeur moyenne de Y quand $(X_1, ..., X_P) = (x_1, ..., x_p) = x$ le réel inconnu : $y_x = E(Y_j \{(X1, ..., Xp) = x\}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$.Un estimateur de y_x est :

$$\hat{Y}x = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p. \tag{1.10}$$

En posant $x_\bullet = (1, x_1, ..., x_p)$ on a $y_x = x_\bullet \beta$ et $\hat{Y}_x = x_\bullet \hat{\beta}$.

1.4.1 Estimations ponctuelles

Dorénavant, l'expression "la réalisation" fera référénce à celle correspondante aux données.Une estimation ponctuelle de β est la réalisation b de $\hat{\beta}$:

$$b = (X^t X)^{-1} X^t y. \tag{1.11}$$

Avec $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ et $b = \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix}$ Ainsi, b minimise la function $f_*(u) = \| y - Xu \|^2$ .L'idée globale est "y_i et $b_0 + b_1 x_{1,i} + ... + b_p x_{p,i}$ aussi proche que possible".Pour tout $j \in \{0, .., p\}$,b_j est une estimation ponctuelle de β_j.On dit que b est l'emco ponctuel de β et b_j est l'emco ponctuel de β_j.Soit $x_\bullet = (1, x_1, ..., x_p)$.Une estima-

tion ponctuelle de $y_x = x_\bullet \beta$ est la réalisation d_x de $\hat{Y}_x = x_\bullet \hat{\beta}$:

$$dx = x_\bullet b = b_0 + b_1 x_1 + ... + b_p x_p. \tag{1.12}$$

On dit que dx est la valeur prédite de Y quand $(X_1, .., X_p) = x$.

1.4.2 Coefficient de détermination

Soit 1_n le vecteur colonne à n composantes égales à 1. On pose $\hat{Y} = X\hat{\beta}$ et $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ On appelle coefficient de détermination la réalisation R^2 de $\hat{R}^2 = 1 - \frac{\|\hat{Y}-Y\|^2}{\|\bar{Y}1_n - Y\|^2}$

Avec les notations déjà introduites et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, on peut écrire :

$$R^2 = 1 - \frac{\| X\hat{\beta} - y \|^2}{\| \bar{Y}1_n - y \|^2}. \tag{1.13}$$

On a $R^2 \in [0, 1]$. Plus R^2 est proche de 1,plus la liaison linéaire entre Y et $X_1, ..., X_p$ est forte.En effet, plus R^2 est proche de 1, plus $\| Xb - y \|^2$ est proche de 0, plus y est proche de Xb, plus le modèle de rlm est pertinent, plus la liaison linéaire entre Y et $X_1, ..., X_p$ est forte.En remarquant que $\| \bar{y}1_n - y \|^2 = \| \bar{y}1_n - Xb \|^2 + \| Xb - y \|^2$,on a aussi :$R^2 = \frac{\|\bar{y}1_n - Xb\|^2}{\|\bar{y}1_n - y\|^2}$ .Coefficient de détermination ajusté.Une version améliorée du R^2 est le coefficient de détermination ajustée défini par :$\bar{R}^2 = 1 - \frac{n-1}{n-(p+1)}(1 - R^2)$.Il s'interprète comme le R^2.

Chapitre 2

Modèle de régression linéaire simple et estimateur des moindres carrés ordinaire

2.1 Modèle de régression linéaire simple (rls)

Le modèle de régression linéaire simple (rls) est le modèle de rlm avec $p = 1$.

2.1.1 Contexte

On souhaite expliquer une variable quantitative Y à partir d'une variable X_1. Pour ce faire, on dispose de données qui sont n observations de (Y, X_1) notées $(y_1, x_{1,1})$, $(y_2, x_{1,2})$, ..., $(y_n, x_{1,n})$.

Ces observations peuvent être représentées sur le repère orthonormé (O, I, J) par les points de coordonnées $(y_1, x_{1,1})$, $(y_2, x_{1,2})$, ..., $(y_n, x_{1,n})$. L'ensemble de ces points est appelé : nuage de points. Si la silhouette de ce nuage de points est allongée

dans une direction, une liaison linéaire entre Y et X_1 est envisageable. On peut alors considérer le modèle de rls. :Sa forme générique est $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, où β_0 et β_1 sont des coefficients réels inconnus et ε est une variable quantitative de valeur moyenne nulle, indépendante de X_1, qui représente une somme d'erreurs aléatoires et multifactorielles.

Notre principal objectif est d'estimer convenablement β_0 et β_1 à l'aide des données. On pourra alors prédire avec précision la valeur moyenne de Y pour une nouvelle valeur de X_1. Cela revient à ajuster du mieux possible le nuage de points par une droite (on parle alors d'ajustement affine).

Exemple 2.1.1 ***Scores*** *:On peut considérer le jeu de données "scores" :*

[5]

$X1$	$4; 9; 10; 14; 4; 7; 12; 1; 3; 8; 11; 5; 6; 10; 11; 16; 13; 13; 10$
Y	$390; 580; 650; 730; 410; 530; 600; 350; 400; 590; 640; 450$
	$520; 690; 690; 770; 700; 730; 640$
Table 3	

TAB. 2.1 – le jeu de données "scores" :

Le tableau représente les données du score obtenu sur 800 points (variable Y) et temps de révision en heures (variable $X1$) obtenues à partir d'une étude de 19 étudiants.

Une étude a été menée auprès de 19 étudiants afin de mettre en évidence une relation entre le score (note) final à un examen de mathématiques et le temps consacré à la préparation de cet examen. Pour chaque étudiant, on dispose :

– du temps de révision en heures (variable $X1$).

– du score obtenu sur 800 points (variable Y).

Fibres : On peut considérer le jeu de données "fibres" :

[6]

$X1$	2.3; 3; 3.6; 4.3; 5; 5.7; 6.7; 6.8; 8; 8.8; 9.7; 11; 12.4; 13.4; 14.3; 14.7
Y	16; 12; 18; 28; 28; 38; 30; 44; 50; 54; 54; 72; 56; 76; 72; 76

TAB. 2.2 – le jeu de données "fibres"

Le tableau représente les données du vitesse de l'influx nerveux en m/s (variable Y). et diamètre en microns (variable $X1$). obtenues à partir d'une étude 16 fibres nerveuses différentes.

Une étude s'intéresse à la vitesse de propagation de l'influx nerveux dans une fibre nerveuse. Pour 16 fibres nerveuses différentes,on considère :

– le diamètre en microns (variable $X1$).

– la vitesse de l'influx nerveux en m/s (variable Y).

On souhaite expliquer Y à partir de $X1$.

Toluca : On peut considérer le jeu de données "toluca" :

[7]

$X1$	80; 30; 50; 90; 70; 60; 120; 80; 100; 50; 40; 70; 90
	20; 110; 100; 30; 50; 90; 110; 30; 90; 40; 80; 70
Y	399; 121; 221; 376; 361; 224; 546; 352; 353; 157; 160; 252
	389; 113; 435; 420; 212; 268; 377; 421; 273; 468; 244; 342; 323

TAB. 2.3 – le jeu de données toluca

Le tableau représente les données du la taille du lot (variable $X1$) et nombre total d'heures de travail (variable Y) obtenues à partir d'une étude 25 lots représentatifs de taille variable.

L'entreprise Toluca fabrique des pièces de rechange pour l'équipement de réfrigération. Pour une pièce particulière, le processus de production prend un certain

temps Dans le cadre d'un programme d'amélioration des coûts, l'entreprise souhaite mieux comprendre la relation entre :

– la taille du lot (variable $X1$),

– nombre total d'heures de travail (variable Y).

Les données ont été rapportées pour 25 lots représentatifs de taille variable.

2.2 Écriture matricielle du modèle de rls

On modélise les variables considérées comme des var (définies sur un espace probabilisé $(\Omega; A; P)$),en gardant les mêmes notations par convention. À partir de celles-ci, le modèle de rls est caractérisé par : pour tout $\in \{1, .., n\}$.

– $x_{1,i}$ est une réalisation de X_1,

sachant que $X_1 = x_{1,i}$, y_i est une réalisation de $Y = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i$, où ε_i est une var modélisant une somme d'erreurs aléatoires et multifactorielles.Notons que le modèle de rls peut s'écrire sous la forme matricielle $Y = X\beta + \varepsilon$, où

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, X = \begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{1,2} \\ \vdots & \vdots \\ 1 & x_{1,n} \end{bmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

2.2.1 Emco et modèle de rls

À l'instar du modèle de rlm, on peut déterminer les emco de β_0 et β_1. Le résultat suivant présentent des expressions analytiques des estimateurs obtenus.

On pose $\bar{x}_1 = \frac{1}{n}\sum_{i=1}^n x_{1,i}$ et $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.On rappelle que $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$ est l'emco de $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ Alors on a :

$$\hat{\beta}_1 = \frac{1}{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2} \sum_{i=1}^n (x_{1,i} - \bar{x}_1)(Y_i - \bar{Y}),\ \hat{\beta}_0 = \bar{Y} - \bar{x}_1\hat{\beta}_1 \tag{2.1}$$

On rappelle que le modèle de rls s'écrit sous la forme matricielle :$Y = X\beta + \varepsilon$, où

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{1,2} \\ \vdots & \vdots \\ 1 & x_{1,n} \end{bmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

L'emco $\hat{\beta}$ de β est donné par la formule : $\hat{\beta} = (X^tX)^{-1}X^tY$ Calcul de X^tX. On a :

$$\begin{aligned} X^tX &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{1,1} & x_{1,2} & \cdots & x_{1,n} \end{pmatrix} \begin{pmatrix} 1 & x_{1,1} \\ 1 & x_{1,2} \\ \vdots & \vdots \\ 1 & x_{1,n} \end{pmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_{1,i} \\ \sum_{i=1}^n x_{1,i} & \sum_{i=1}^n x_{1,i}^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x}_1 \\ n\bar{x}_1 & \sum_{i=1}^n x_{1,i}^2 \end{bmatrix}. \end{aligned} \tag{2.2}$$

Calcul de $(X^tX)^{-1}$. En utilisant la formule matricielle : si $ad - bc \neq 0$,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \Leftrightarrow A^{-1} = \frac{1}{ab-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix},$$

$$(X^tX)^{-1} = \frac{1}{n\sum_{i=1}^n x_{1,i}^2 - (n\bar{x}_1)^2} \begin{bmatrix} \sum_{i=1}^n x_{1,i}^2 & -n\bar{x}_1 \\ -n\bar{x}_1 & n \end{bmatrix} \tag{2.3}$$

$$= \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\bar{x}_1^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_{1,i}^2 & -\bar{x}_1 \\ -\bar{x}_1 & 1 \end{bmatrix} \tag{2.4}$$

Calcul de X^tY. On a :

$$X^tY. = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{1,1} & x_{1,2} & \cdots & x_{1,n} \end{bmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_{1,i}Y_i \end{bmatrix} = \begin{bmatrix} n\bar{Y} \\ \sum_{i=1}^n x_{1,i}Y_i \end{bmatrix}.$$

Calcul de $\hat{\beta} = (X^tX)^{-1}X^tY$. En mettant bout à bout les égalités précédentes, il

vient :

$$\hat{\beta} = (X^tX)^{-1}X^tY = \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\bar{x}_1^2}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_{1,i}^2 & -\bar{x}_1 \\ -\bar{x}_1 & 1 \end{bmatrix}\begin{bmatrix} n\bar{Y} \\ \sum_{i=1}^n x_{1,i}Y_i \end{bmatrix}$$
$$= \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\bar{x}_1^2}\begin{bmatrix} (\frac{1}{n}\sum_{i=1}^n x_{1,i}^2)nY_i - \bar{x}_1\sum_{i=1}^n x_{1,i}Y_i \\ -\bar{x}_1 \times n\bar{Y} + \sum_{i=1}^n x_{1,i}Y_i \end{bmatrix}$$
$$= \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\bar{x}_1^2}\begin{bmatrix} \bar{Y}\sum_{i=1}^n x_{1,i}^2 - \bar{x}_1\sum_{i=1}^n x_{1,i}Y_i \\ \sum_{i=1}^n x_{1,i}Y_i - n\bar{x}_1\bar{Y} \end{bmatrix}. \tag{2.5}$$

$$\hat{\beta}_0 = \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n{\bar{x}_1}^2}\left(\sum_{i=1}^n x_{1,i}^2\bar{Y} - \bar{x}_1\sum_{i=1}^n x_{1,i}Y_i\right), \tag{2.6}$$
$$\hat{\beta}_1 = \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n{\bar{x}_1}^2}\left(\sum_{i=1}^n x_{1,i}Y_i - n\bar{x}_1\bar{Y}\right).$$

Réécriture de $\hat{\beta}_1$On a :

$$\sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2 = \sum_{i=1}^n (x_{1,i}^2 - 2\bar{x}_1x_{1,i} + \bar{x}_1^2) = \sum_{i=1}^n x_{1,i}^2 - 2\bar{x}_1\sum_{i=1}^n x_{1,i} + \bar{x}_1^2\sum_{i=1}^n 1$$
$$= \sum_{i=1}^n x_{1,i}^2 - 2\bar{x}_1 \times n\bar{x}_1 + \bar{x}_1^2 n = \sum_{i=1}^n x_{1,i}^2 - 2n\bar{x}_1^2 + n\bar{x}^2$$
$$= \sum_{i=1}^n x_{1,i}^2 - n\bar{x}_1^2. \tag{2.7}$$

De plus, on a :

$$\begin{aligned}\sum_{i=1}^{n}(x_{1,i}-\bar{x}_1)(Y_i-\bar{Y}) &= \sum_{i=1}^{n}(x_{1,i}Y_i - x_{1,i}\bar{Y} - \bar{x}_1 Y_i + \bar{x}_1\bar{Y}) \\ &= \sum_{i=1}^{n} x_{1,i}Y_i - \bar{Y}\sum_{i=1}^{n} x_{1,i} - \bar{x}_1\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n}\bar{x}_1\bar{Y} \\ &= \sum_{i=1}^{n} x_{1,i}Y_i - \bar{Y}\times n\bar{x}_1 - \bar{x}_1\times n\bar{Y} + \bar{x}_1\bar{Y}\times n \\ &= \sum_{i=1}^{n} x_{1,i}Y_i - n\bar{x}_1\bar{Y} - n\bar{x}_1\bar{Y} + n\bar{x}_1\bar{Y} \\ &= \sum_{i=1}^{n} x_{1,i}Y_i - n\bar{x}_1\bar{Y}\end{aligned} \tag{2.8}$$

Par conséquent, on peut réécrire $\hat{\beta}_1$ comme :

$$\hat{\beta}_1 = \frac{1}{\sum_{i=1}^{n} x_{1,i}^2 - n\bar{x}_1^2}\left(\sum_{i=1}^{n} x_{1,i}Y_i - n\bar{x}_1\bar{Y}\right) = \frac{1}{\sum_{i=1}^{n}(x_{1,i}-\bar{x}_1)^2}\sum_{i=1}^{n}(x_{1,i}-\bar{x}_1)(Y_i-\bar{Y}). \tag{2.9}$$

Réécriture de $\hat{\beta}_0$. En introduisant $0 = -n\bar{x}_1^2\bar{Y} + n\bar{x}_1^2\bar{Y}$, on obtient :

$$\begin{aligned}\bar{Y}\sum_{i=1}^{n} x_{1,i}^2 - \bar{x}_1\sum_{i=1}^{n} x_{1,i}Y_i &= \bar{Y}\sum_{i=1}^{n} x_{1,i}^2 - n\bar{x}_1^2\bar{Y} + n\bar{x}_1^2\bar{Y} - \bar{x}_1\sum_{i=1}^{n} x_{1,i}Y_i \\ &= \bar{Y}\left(\sum_{i=1}^{n} x_{1,i}^2 - n\bar{x}_1^2\right) - \bar{x}_1\left(\sum_{i=1}^{n} x_{1,i}Y_i - n\bar{x}_1\bar{Y}\right).\end{aligned} \tag{2.10}$$

Il vient :

$$\begin{aligned}\hat{\beta}_0 &= \frac{1}{\sum_{i=1}^{n} x_{1,i}^2 - n\bar{x}_1^2}\left(\bar{Y}\sum_{i=1}^{n} x_{1,i}^2 - \bar{x}_1\sum_{i=1}^{n} x_{1,i}Y_i\right) \\ &= \frac{1}{\sum_{i=1}^{n} x_{1,i}^2 - n\bar{x}_1^2}\left(\bar{Y}\left(\sum_{i=1}^{n} x_{1,i}^2 - n\bar{x}_1^2\right) - \bar{x}_1\left(\sum_{i=1}^{n} x_{1,i}Y_i - n\bar{x}_1\bar{Y}\right)\right) \\ &= \bar{Y} - \bar{x}_1\frac{1}{\sum_{i=1}^{n} x_{1,i}^2 - n\bar{x}_1^2}(\sum_{i=1}^{n} x_{1,i}Y_i - n\bar{x}_1\bar{Y}) = \bar{Y} - \bar{x}_1\hat{\beta}_1.\end{aligned} \tag{2.11}$$

Au final. L'emco $\hat{\beta}$ de β a pour composantes :

$$\hat{\beta}_1 = \frac{1}{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2} \sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)(Y_i - \bar{Y}), \tag{2.12}$$

$$\hat{\beta}_0 = \bar{Y} - \bar{x}_1\hat{\beta}_1. \tag{2.13}$$

Estimateur de la prédiction

Soit y_x la valeur moyenne de Y quand $X_1 = x_1 = x : y_x = \beta_0 + \beta_1 x_1$.Un estimateur de y_x est $\hat{Y}_x = \hat{\beta}_0 + \hat{\beta}_1 x_1$.

2.2.2 Quantités utilisées

Partant des données, on considère les quantités suivantes :

Moyennes

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1,i}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{2.14}$$

Écarts-type

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2}, \sqrt{(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2)}. \tag{2.15}$$

Sommes des carrés des écarts

$$\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2 = \sum_{i=1}^{n} x_{1,i}^2 - n\bar{x}_1^2, \tag{2.16}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2. \tag{2.17}$$

Somme des produits des écarts

$$\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)(y_i - \bar{y}) = \sum_{i=1}^{n} x_{1,i}y_i - n\bar{x}_1\bar{y}. \tag{2.18}$$

2.2.3 Estimations ponctuelles

Les formules analytiques de $\hat{\beta}_1$ et $\hat{\beta}_0$ donnent les estimations ponctuelles suivantes.Une estimation ponctuelle de β_1 est la réalisation de $\hat{\beta}_1$:

$$b_1 = \frac{1}{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2} \sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)(y_i - \bar{y}). \tag{2.19}$$

- On dit que b_1 est l'emco ponctuel de β_1.Une estimation ponctuelle de β_0 est la réalisation de $\hat{\beta}_0$: $b_0 = \bar{y} - b_1\bar{x}_1$.
- On dit que b_0 est l'emco ponctuel de β_0.Une estimation ponctuelle de $y_x = \beta_0 + \beta_1 x_1$ est la réalisation de $\hat{Y}_x = \hat{\beta}_0 + \hat{\beta}_1 x_1$: $d_x = b_0 + b_1 x_1$.
- On dit que d_x est la valeur prédite de Y quand $X_1 = x_1$.

2.2.4 Coefficient de corrélation linéaire

On appelle coefficient de corrélation linéaire le réel $r_{x;y}$ défini par $r_{x;y} = \frac{\sum_{i=1}^{n}(x_{1,i}-\bar{x}_1)(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{1,i}-\bar{x}_1)^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$.on a $r_{x;y} \in [-1, 1]$.

2.2.5 Droite de régression et coefficient de corrélation linéaire

On a $b_1 = \frac{\sqrt{(\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2)}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{1,i}-\bar{x}_1)^2}} r_{x;y}$.Comme $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2} > 0$ et $\sqrt{(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2)} > 0$, le coefficient directeur b_1 de la droite de régression et $r_{x,y}$ sont de même signe (à une droite de régression croissante correspond un $r_{x;y}$ positif). Dès lors, on peut deviner lesigne de $r_{x;y}$ avec la silhouette du nuage de points.Plus $\mid r_{x;y} \mid$ est proche de 1, plus la liaison linéaire entre Y et X_1 est forte.En effet, plus$\mid r_{x;y} \mid$est proche 1, plus b_1 diffère de 0, plus 1 diffère de 0, plus la liaison linéaire entre Y et X_1 est forte.Aussi, plus $\mid r_{x;y} \mid$ est proche 1, plus X_1 influe sur/est corrélée avec Y.

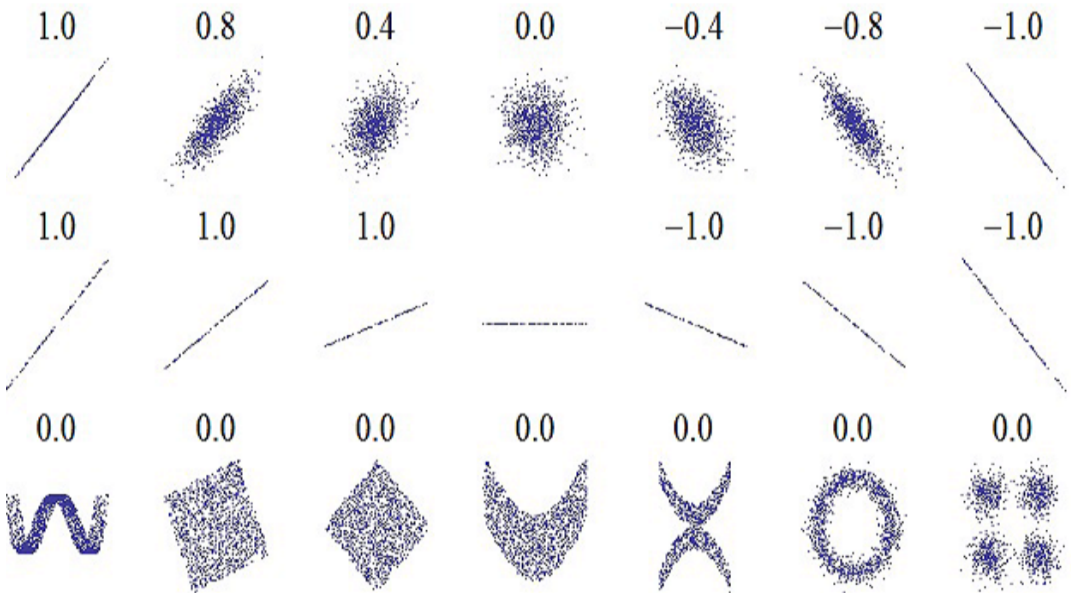

FIG. 2.1 – Le graphique suivant illustre le lien existant entre la pertinence del'ajustement d'un nuage depoints par une droite, caractérisée par la corrélation linéaire entre Y et $X1$, et la valeur associée de $r_{x;y}$

Source du graphique :

https : *//en.wikipedia.org/wiki/Pearson_product−moment_correlation_coefficient*

2.2.6 Coefficient de détermination et coefficient de corrélation linéaire

Dans le cas du modèle de rls, on peut montrer que $R^2 = r^2_{x;y}$.Dans ce cas, l'interprétation des valeurs de $R^2 = r^2_{x;y}$ est donc identique.

2.3 Loi normale multidimensionelle

2.3.1 Vecteur gaussien

Soient $n \in \mathbb{N}$ et $U_1, ..., U_n$ n var. On dit que $U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$ est un vecteur gaussien si et seulement si toute combinaison linéaire de $U_1, ..., U_n$ suit une loi normale : pour tout $(a_1, ..., a_n) \in \mathbb{R}^n$,$a_1U_1 + ... + a_nU_n \rightsquigarrow N$.

Un vecteur gaussien est caractérisé par son espérance et sa matrice de covariance .La loi de U est la loi normale multidimensionnelle notée $N_n(\mu; \sum)$.

2.3.2 Critère d'indépendance

Soient $U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$ un vecteur gaussien et $(j, k) \in \{1, ..., n\}^2$avec $j \neq k$. Alors U_J et U_k sont indépendantes si et seulement si $C(U_j, U_k) = 0$.

2.3.3 Loi normale multidimensionnelle

Soient $\mu \in \mathbb{R}^n$ et une matrice de dimension $n \times n$ symétrique définie positive vérifiant $\det(\sum) > 0$. Alors $U \rightsquigarrow N_n(\mu, \sum)$ si et seulement si U possède la densité :

$$f(x) = \frac{1}{(2\pi)^{n/2}\,(\sqrt{\det(\sum)})} \exp(-1/2(x-\mu)^t \sum^{-1}(x-\mu)), x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in R^n. \tag{2.20}$$

Représentation graphique

Une densité associée à la loi $N_2(0_2, \Sigma)$, avec $\sum = \begin{bmatrix} 0.5 & 1 \\ 1 & 0.5 \end{bmatrix}$ est présentée ci-dessous :

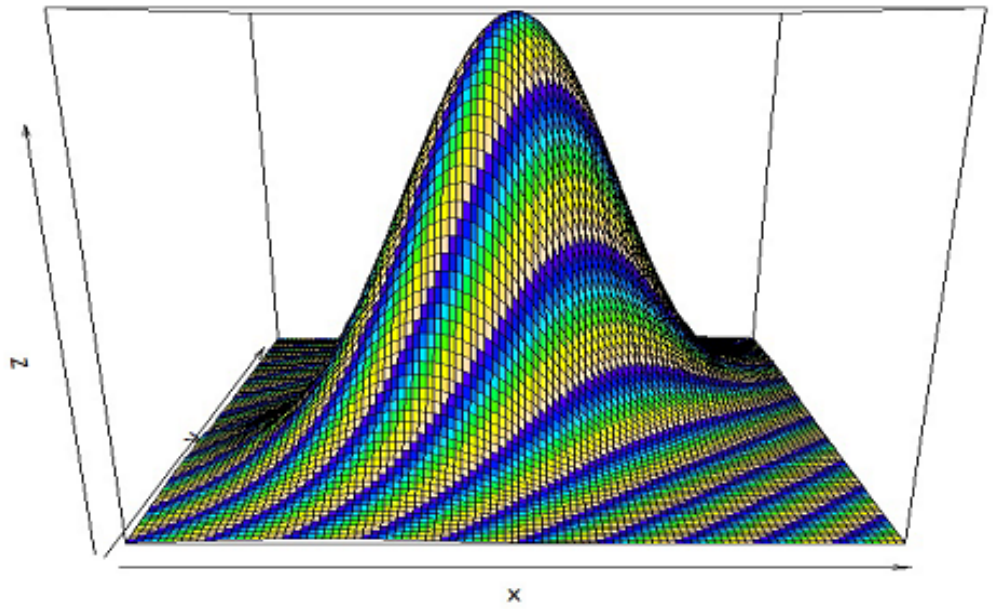

FIG. 2.2 – densité associée à la loi normale multidimensionnelle

Notation 2.3.1 $U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$ *et* $V = \begin{pmatrix} V_1 \\ \vdots \\ V_n \end{pmatrix}$ *des vecteurs de var,*

et $W = \begin{bmatrix} W_{1,1} & W_{2,1} & \cdots & W_{n,1} \\ W_{1,2} & W_{2,2} & \cdots & W_{n,2} \\ \vdots & \vdots & \vdots & \vdots \\ W_{1,n} & W_{2,1} & \cdots & W_{n,n} \end{bmatrix}$ matrice de var. On adopte les notations :

Espérance de U : $E(U) = \begin{pmatrix} E(U_1) \\ \vdots \\ E(U_n) \end{pmatrix}$.

Espérance de W :$E(W) = \begin{bmatrix} E(W_{1,1}) & E(W_{2,1}) & \cdots & E(W_{n,1}) \\ E(W_{1,2}) & E(W_{2,2}) & \cdots & E(W_{n,2}) \\ \vdots & \vdots & \vdots & \vdots \\ E(W_{1,n}) & E(W_{2,1}) & \cdots & E(W_{n,n}) \end{bmatrix}$.

Covariance de U et V :

$$C_n(U,V) = E_{n,n}((U - E_n(U))(V - E_n(V))^t)$$
$$= \begin{bmatrix} C(U_1,V_1) & C(U_1,V_2) & \cdots & C(U_1,Vn) \\ C(U_1,V_2) & C(U_2,V_2) & \cdots & C(U_2,Vn) \\ \vdots & \vdots & \vdots & \vdots \\ C(Un,V_1) & C(U_n,V_2) & .\cdots & C(Un,Vn) \end{bmatrix}.$$

Matrice de covariance de

$$U : V_n(U) = C_n(U,U) = E_{n,n}((U - En(U))(V - En(V))^t). \tag{2.21}$$

Paramètres

. Si $U \rightsquigarrow N_n(\mu;\Sigma)$, alors $En(U) = \mu$ et $Vn(U) = \Sigma$.

Forme linéaire

Soient $X \rightsquigarrow N_n(\mu,\Sigma)$, A une matrice à n colonnes et q lignes, et a un vecteur colonne à q composantes. Alors on a $AX+a \rightsquigarrow Nq(A\mu+a, A\Sigma A^t)$.En particulier, si $X \rightsquigarrow N_n(\mu;\Sigma)$, on a $E_q(AX+a) = A\mu + a$ et $V_q(AX+a) = A\Sigma A^t$.

Vecteurs gaussiens et indépendance

Soient $X \rightsquigarrow N_n(\mu, \Sigma)$, A une matrice à n colonnes et p lignes et B une matrice à n colonnes et q lignes. Alors AX et BX sont indépendantes si et seulement si $A\Sigma B^t = 0_{p,q}$.Ainsi, les composantes de tout sous vecteur de X sont indépendantes si et seulement si leurs covariances sont nulles.

2.4 Propriétés standards et lois associées

Proposition 2.4.1 *Hypothèses standards*

On considère le modèle de rlm sous la forme matricielle :$Y = X\beta + \varepsilon$.On suppose que :

– X est de rang colonnes plein,

– ε et $X_1, ..., X_p$ sont indépendantes,

– $\varepsilon \rightsquigarrow N_n(0_n, \sigma^2 I_n)$ où $\sigma > 0$ est un paramètre inconnu.

L'hypothèse $N_n(0_n, \sigma^2 I_n)$ entraîne que $\varepsilon_1, ..., \varepsilon_n$ sont indépendantes et identiquement distribuées de loi commune la loi normale $N(0, \sigma^2)$.Les hypothèses standards sont à la base d'une analyse statistique avancée avec le modèle de rlm.

Dorénavant, on suppose que les hypothèses standards sont satisfaites.

Proposition 2.4.2 *Loi de Y :*

$Y \rightsquigarrow N_n(X\beta, \sigma^2 I_n)$

Preuve. On peut écrire $Y = a + \varepsilon$,où $a = X\beta$ est un vecteur colonne à n composantes constantes. Comme $\varepsilon \rightsquigarrow N_n(0_n, \sigma^2 I_n)$, on a $\varepsilon + a \rightsquigarrow N_n(a + 0_n, \sigma^2 I_n)$, ce qui entraîne

$$Y \rightsquigarrow N_n(X\beta, \sigma^2 I_n).$$

■

Proposition 2.4.3 *Loi de $\hat{\beta}$:*

On a $\hat{\beta} \rightsquigarrow N_{p+1}(\beta, \sigma^2(X^tX)^{-1})$ les conséquences immédiates de ce résultats sont :

- $\hat{\beta}$ est un estimateur sans biais de β : $\mathbf{E}_{p+1}(\hat{\beta}) = \beta$.
- la matrice de covariance de $\hat{\beta}$ est $\sigma^2(X^tX)^{-1}$: $\mathbf{V}_{p+1}(\hat{\beta}) = \sigma^2(X^tX)^{-1}$.
- en notant $[(X^tX)^{-1}]_{j+1,j+1}$ la $j+1$-ème composante diagonale de $(X^tX)^{-1}$, on a $\hat{\beta}_j \rightsquigarrow N(\beta_j, \sigma^2[(X^tX)^{-1}]_{j+1,j+1})$.
- si la $(j+1, k+1)$-ème composante de $(X^tX)^{-1}$ est nulle, alors $\hat{\beta}_j$ et $\hat{\beta}_k$ sont indépendantes.

Preuve. On a $\hat{\beta} = (X^tX)^{-1}X^tY$. Ainsi, on peut écrire $\hat{\beta} = AY$, où $A = (X^tX)^{-1}X^t$ est une matrice à composantes constantes. Comme $Y \rightsquigarrow N_n(X\beta, \sigma^2 I_n)$,il vient $\hat{\beta} \rightsquigarrow N_{p+1}(AX\beta, A\sigma^2 I_n A^t)$. En remarquant que $(X^tX)^{-1}X^tX = In$, on a :

$$AX\beta = (X^tX)^{-1}X^tX\beta = \beta In = \beta. \tag{2.22}$$

D'autre part, en utilisant les opérations matricielles : $(CD)^t = D^tC^t$, $(C^t)^t = C$ et $(C^{-1})^t = (C^t)^{-1}$, on a :

$$\begin{aligned} A(\sigma^2 In)A^t &= \sigma^2 AA^t = \sigma^2(X^tX)^{-1}X^t((X^tX)^{-1}X^t)^t \qquad (2.23)\\ &= \sigma^2(X^tX)^{-1}X^t(X^t)^t(X^t(X^t)^t)^{-1} \\ &= \sigma^2(X^tX)^{-1}X^tX(X^tX)^{-1} = \sigma^2(X^tX)^{-1}In = \sigma^2(X^tX)^{-1}. \qquad (2.24) \end{aligned}$$

D'où $\hat{\beta} \rightsquigarrow N_{p+1}(\beta, \sigma^2(X^tX)^{-1})$. ■

Conséquence du théorème de Gauss-Markov

L'emco $\hat{\beta}$ est le meilleur estimateur linéaire sans biais de β; c'est le BLUE (Best Linear Unbiased Estimator) : aucun autre estimateur linéaire sans biais de β n'a une variance plus petite.

Lien avec l'estimateur du maximum de vraisemblance (emv)

L'emco $\hat{\beta}$ est l'emv de β. Dès lors, il est fortement consistant et asymptotiquement efficace.

Comme $Y \rightsquigarrow N_n(X\beta, \sigma^2 I_n)$ la fonction de vraisemblance associée à $(Y_1, ..., Y_n)$ est donnée par : $L(\beta, z) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|z-X\beta\|^2}{2\sigma^2}\right),\ z \in \mathbb{R}^n$ Soit $\hat{\beta}$ l'emv de β :$\tilde{\beta}$ $= argmax_{\beta\in\mathbb{R}^{p+1}} L(\beta; Y)$ Par croissance de la fonction exponentielle, on a :

$$\tilde{\beta} = argmax_{\beta\in\mathbb{R}^{p+1}} L(\beta; Y) = argmax_{\beta\in\mathbb{R}^{p+1}} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right) \tag{2.25}$$

$$= argmax_{\beta\in\mathbb{R}^{p+1}} \left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right) = argmax_{\beta\in\mathbb{R}^{p+1}} \|Y - X\beta\|^2 = \hat{\beta}. \tag{2.26}$$

Un estimateur de σ^2 est $\hat{\sigma}^2 = \frac{1}{n-(p+1)}\|Y - X\hat{\beta}\|^2$.On a :

- $\hat{\sigma}^2$ est sans biais pour $\hat{\sigma}^2$: $\mathbf{E}(\hat{\sigma}^2) = \sigma^2$.
- $\hat{\sigma}^2$ et β sont indépendantes.
- $(n - (p + 1))\frac{\hat{\sigma}^2}{\sigma^2} \rightsquigarrow X^2(v)$, $v = n - (p + 1)$.

Preuve. Soit L le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de X. On peut montrer que $I_n - X(X^tX)^{-1}X^t$ est la matrice de projection sur l'orthogonal de L noté $L^\perp$ Ce sous-espace est de dimension $n-(p+1)$: $Dim(L^\perp) = n-(p+1)$.

■

On peut montrer que $Y - X\beta \rightsquigarrow N_n(X\beta, \sigma^2(I_n - X(X^tX)^{-1}X^t))$ Comme la trace d'une matrice de projection est égale à la dimension de l'image de la projection on :

$$\begin{aligned}\mathbf{E}(\|Y - X\hat{\beta}\|) &= \mathbf{E}(\mathbf{Trace}((Y - X\hat{\beta})(Y - X\hat{\beta})^t)) \quad (2.27)\\ &= \mathbf{Trace}(\mathbf{E}_{n,n}((Y - X\hat{\beta})(Y - X\hat{\beta})^t))\\ &= \sigma^2\mathbf{Trace}(I_n - X(X^tX)^{-1}X^t) = \sigma^2 Dim(L^{\perp})\\ &= \sigma^2(n - (p+1)). \quad (2.28)\end{aligned}$$

Donc $\mathbf{E}(\hat{\sigma}^2) = \sigma^2$.On peut montrer que le vecteur aléatoire réel $(Y - X\hat{\beta}, \hat{\beta})$ est gaussien et que toutes les covariances d'une composante de $Y - X\hat{\beta}$ et d'une composante de $\hat{\beta}$ sont nulles. Cela entraîne l'indépendance de $Y - X\hat{\beta}$ et $\hat{\beta}$. Comme $\hat{\sigma}^2$ est uniquement fonction de $\hat{\varepsilon}$, on a aussi l'indépendance de $\hat{\sigma}^2$ et $\hat{\beta}$ On peut écrire :

$$(n - (p+1))\frac{\hat{\sigma}^2}{\sigma^2} = \|(I_n - X(X^tX)^{-1}X^t)(\frac{\varepsilon}{\sigma})\|^2. \quad (2.29)$$

comme $\frac{\varepsilon}{\sigma} \rightsquigarrow N_n(0_n, I_n)$ et $I_n - X(X^tX)^{-1}X^t$ est la matrice de projection sur $L^{\perp}$ avec $Dim(L^{\perp}) = n - (p+1)$, le théorème de Cochran entraîne :

$$(n - (p+1))\frac{\hat{\sigma}^2}{\sigma^2} \rightsquigarrow X^2(v), \ v = Dim(L^{\perp}) = n - (p+1). \quad (2.30)$$

Degrés de liberté : Dorénavant, désigne le nombre de degrés de liberté :$v = n - (p+1)$.

Proposition 2.4.4 *Emco et loi de Student :*

Pour tout vecteur ligne c à $p+1$ composantes, on a

$$\frac{c\hat{\beta}-c\beta}{\hat{\sigma}(c(X^tX)^{-1}c^t)^{1/2}} \rightsquigarrow T(v).$$

Preuve. Dans un premier temps, rappelons une caractérisation de la loi de Student. Soient A et B deux var indépendantes avec $A \rightsquigarrow N(0,1)$ et $B \rightsquigarrow X^2(v)$, alors $T = \frac{A}{(\frac{B}{v})^{1/2}} \rightsquigarrow T(v)$.On pose alors : $A = \frac{c\hat{\beta}-c\beta}{\sigma(c(X^tX)^{-1}c^t)^{1/2}}$, $B = (n-(p+1))\frac{\hat{\sigma}^2}{\sigma^2}$.Comme $\hat{\beta}$ et σ^2 sont indépendantes, il en est de même pour A et B. Comme $\hat{\beta} \rightsquigarrow N_{P+1}(\beta, \sigma^2(X^tX)^{-1})$, on a $c\hat{\beta} \rightsquigarrow N(c\beta, \sigma^2(X^tX)^{-1}c^t)$ ce qui entraîne $A \rightsquigarrow N(0,1)$.De plus, on a $B \rightsquigarrow X^2(v)$. Par la caractérisation de la loi de Student, il s'ensuit

$$\frac{c\hat{\beta}-c\beta}{\hat{\sigma}(c(X^tX)^{-1}c^t)^{1/2}} = \frac{A}{(\frac{B}{n-(p+1)})^{1/2}} \rightsquigarrow T(v). \tag{2.31}$$

■

Proposition 2.4.5 *Pour tout $j \in \{0,...,p\}$,on a $\frac{\hat{\beta}_j-\beta_j}{\hat{\sigma}([(X^tX)^{-1}]_{j+1,j+1})^{1/2}} \rightsquigarrow T(v)$.Soient $x_\bullet = (1, x_1, ..., x_p)$, $y_x = x_\bullet\beta$ et $\hat{Y}_x = x_\bullet\hat{\beta}$. On a*

$$\frac{\hat{Y}_x - y_x}{\hat{\sigma}(x_\bullet(X^tX)^{-1}x_\bullet^t)^{1/2}} \rightsquigarrow T(v).$$

Preuve. On peut utiliser le résultat :$\frac{c\hat{\beta}-c\beta}{\hat{\sigma}(c(X^tX)^{-1}c^t)^{1/2}} \rightsquigarrow T(v)$.On obtient le premier point en prenant $c = c_j$ le vecteur ligne à $p+1$ composantes, toutes nulles sauf la $j+1$-ème qui vaut 1. On obtient le deuxième point en prenant $c = x_\bullet$. ■

Proposition 2.4.6 *Emco et loi de Fisher*

Soit Q une matrice de réels à $p+1$ colonnes et k lignes de rang colonnes plein.Alors on a

$$\frac{(Q\hat{\beta}-Q\beta)^t(Q(X^tX)^{-1}Q^t)^{-1}(Q\hat{\beta}-Q\beta)}{k\hat{\sigma}^2} \rightsquigarrow F(k,v).$$

Par exemple, avec $\hat{\beta}=\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}$ et $Q=\begin{bmatrix}4 & 1 & 0\\ 0 & 2 & -5\end{bmatrix}$, on a $Q\hat{\beta}=\begin{bmatrix}4\hat{\beta}_0+\hat{\beta}_1\\ 2\hat{\beta}_1-5\hat{\beta}_2\end{bmatrix}$.

Dans un premier temps, rappelons une caractérisation de la loi de Fisher. Soient A et B deux var indépendantes avec $A \rightsquigarrow X^2(v_1)$ et $B \rightsquigarrow X^2(v_2)$,alors $F=\frac{v_1A}{v_2B} \rightsquigarrow F(v_1,v_2)$.On pose alors :

$$A=\frac{(Q\hat{\beta}-Q\beta)^t(Q(X^tX)^{-1}Q^t)^{-1}(Q\hat{\beta}-Q\beta)}{\hat{\sigma}^2}, B=v\frac{\hat{\sigma}^2}{\sigma^2}\ . \tag{2.32}$$

En utilisant le théorème de Cochran, on peut montrer que A et B sont indépendantes avec $A \rightsquigarrow X^2(v_1)$ et $B \rightsquigarrow X^2(v_2)$.Par la caractérisation de la loi de Fisher, il s'ensuit

$$\frac{(Q\hat{\beta}-Q\beta)^t(Q(X^tX)^{-1}Q^t)^{-1}(Q\hat{\beta}-Q\beta)}{k\hat{\sigma}^2}=\frac{vA}{kB} \rightsquigarrow F(k,v). \tag{2.33}$$

Une estimation ponctuelle de σ est la réalisation de $\hat{\sigma}$:$s=(\frac{1}{n-(p+1)}\|y-Xb\|^2)^{1/2}=(\frac{1}{n-(p+1)}sce_y\,(1-R^2)$ .Pour tout $j\in\{0,...,p\}$ une estimation ponctuelle de l'écart-type de $\hat{\beta}_j$ est $ete_j=s([(X^tX)^{-1}]_{j+1,j+1})^{1/2}$.Soit $x_\bullet=(1,x_1,...,x_p)$. Une estimation ponctuelle de l'écart-type de $\hat{Y}_x=x_\bullet\hat{\beta}$ est $ete_x=s(x_\bullet(X^tX)^{-1}x_\bullet^t)^{1/2}$.

Résidus : Pour tout $i\in\{1,...,n\}$, on appelle i-ème résidu la réalisation e_i de $\hat{\varepsilon}_i=Y_i-\hat{Y}_i$,où $\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1x_{1,i}+...+\hat{\beta}_px_{p,i}$.On appelle résidus les réels $e_1,...,e_n$. Avec les notations déjà introduites, on peut écrire :$e_i=y_i-d_{x_i}$, avec $x_i=(x_{1,i},...,x_{p,i})$.

Graphique des résidus : On appelle graphique des résidus le graphique du nuage de points : $N_e = \{(1, e_1), (2, e_2), ..., (n, e_n)\}$.

Interprétation : première approche : Ainsi, $\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$ est la réalisation de $\begin{pmatrix} \hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}$, lequel est un estimateur grossier de ε. Donc, sous les hypothèses standards, $\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$ devrait avoir les caractéristiques grossières d'une réalisation de $N_n(0_n; \sigma^2 I_n)$.

Si le nuage de points n'a aucune structure particulière, avec une relative symétrie dans la répartition des points par rapport à l'axe des abscisses, alors on admet que $\varepsilon \rightsquigarrow N_n(0_n; \sigma^2 I_n)$.

Exemples de graphiques des résidus : Des exemples de graphiques des résidus sont proposés ci-dessous ; seul le premier colle avec le shypothèses standards.

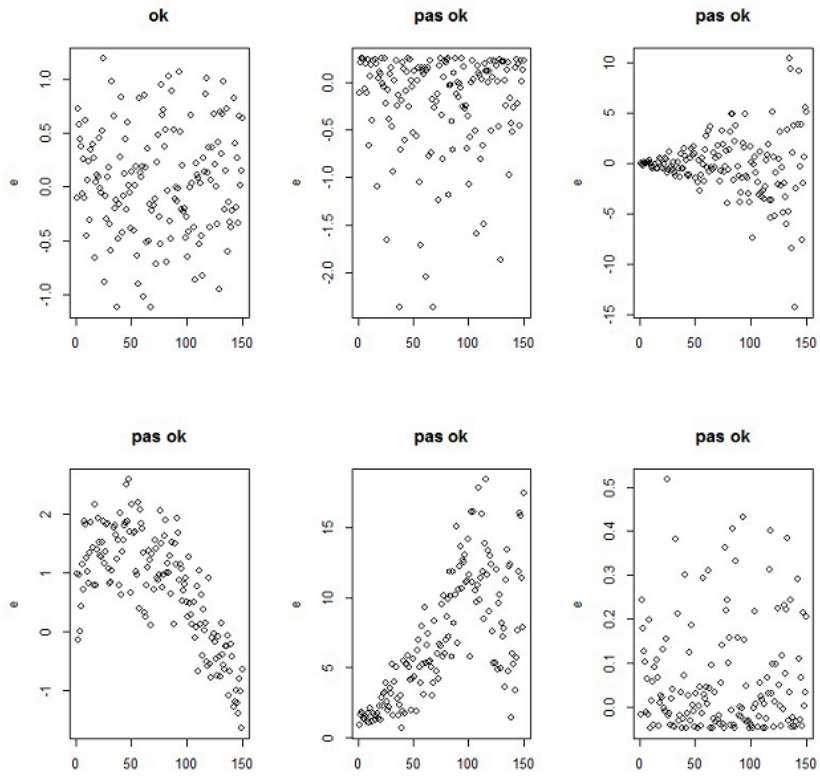

FIG. 2.3 – le premier colle avec le shypothéses standards.

Chapitre 3

Implementation numérique

3.1 Mise en oeuvre avec le logiciel spss

3.1.1 Modèle de régression linéaire multiple

Nous étudions les exemples précédents dans le programme Spss pour les clarifier davantage, et ils nous donnent les résultats suivants :

loyers :

	Moyenne	Ecart type	N
Y	$4023,2667$	$761,72024$	30
$X1$	$59,2000$	$11,37268$	30

TAB. 3.1 – statistiques descriptive

Table de statistiques Le tableau précédent présente des statistiques descriptives (la taille de l'échantillon est de 30.moyennes lorsque la de moyenne Y est de $4023,2667$, la moyenne de $X1$ est de $59,2000$ Il s'agit des variables introduites dans le modèle de régression dépendante et autonome.

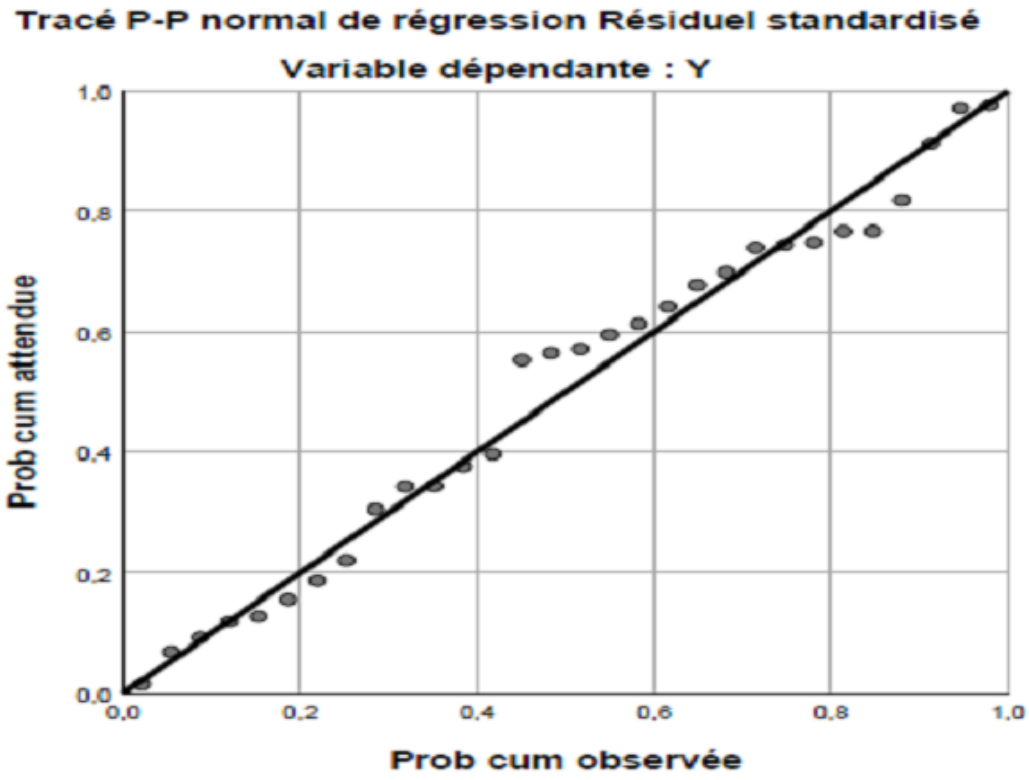

FIG. 3.1 – Trace p-p normal de régression résiduel standardisé

Le tableau précédent montre la matrice de corrélation entre les variables du modèle de régression.

- Où le coefficient de corrélation entre Y et $X4$ était de $0,535$ avec une signification inférieure à 0.01.
- .Variable dépendante : Y.
- Toutes les variables demandées ont été introduites.

Le tableau montre les noms des variables qui sont entrées dans l'équation de régression en tant que variable dépendante et les variables indépendantes (Y et$X1$) et l'analyse n'a exclu aucune variable. Il montre également la méthode utilisée, qui est la régression standard

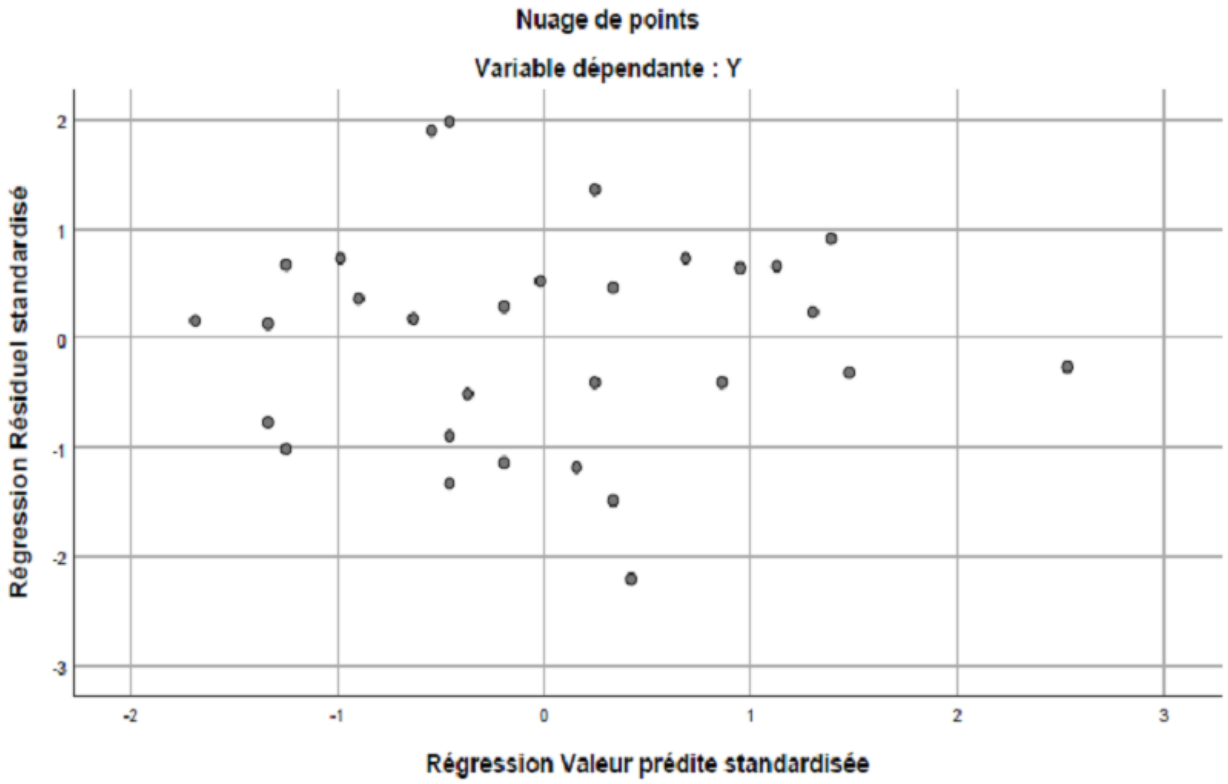

FIG. 3.2 – Nuage de points

– Prédicteurs : (Constante), $X1$.

– Variable dépendante :Y.

Le tableau précédent montre le coefficient de corrélation de Pearson entre la variable dépendante et les variables indépendantes, où il atteint une valeur élevée de 0.847 en divisant le coefficient de détermination par 0.718 et la valeur du coefficient de détermination modifié 0.708, c'est-à-dire que les variables indépendantes expliquent 70% de la variance de la satisfaction au travai.

– Variable dépendante : Y.

– Prédicteurs : (Constante), $X1$.

Le tableau précédent montre les résultats de l'analyse ANOVA pour tester la signification de la régression, et nous notons que la valeur de Sig est 0.00, qui est

		Y	$X1$
Corrélation de Pearson	Y	$1,000$	0.847
	$X1$	0.847	$1,000$
Sig. (unilatéral)	Y	000	.
	$X1$	.	000
N	Y	30	30
	$X1$	30	30

TAB. 3.2 – C
orrélations

Modèle	Variables introduites	Variables éliminées	Méthode
1	$X1^b$	.	Introduire

TAB. 3.3 – Variables introduites/éliminéesa

inférieure à 0.01. Ainsi, nous rejetons l'hypothèse nulle et acceptons l'hypothèse alternative, qui est que la régression est significative et donc il y a un effet des variables indépendantes sur la variable dépendante, et nous pouvons prédire la variable dépendante à travers ces variables indépendantes.

– Coeff non stad : Coefficients non standardisés.

– Coeff stad : Coefficients standardisés.

Le tableau précédent présente les facteurs de régression standard et non standard, la ligne standard, la valeur du test avec la valeur de probabilité des tests (indication statistique) Le tableau précédent aide également à écrire l'équation de la droite de régression.

Predicted(JS)= $0.847 * X1 + e$.

Formages :

Variables introduites/éliminéesa :

– Variable dépendante : Y.

– Toutes les variables demandées ont été introduites.

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	0.847	0.718	0.708	411, 87128

TAB. 3.4 – Récapitulatif des modèlesb

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig
1	Régression	12076451, 292	1	12076451, 292	71, 190	000^b
	de Student	4749862, 575	28	169637, 949		
	Total	16826313, 867	29			

TAB. 3.5 – ANOVA

Le tableau montre les noms des variables qui sont entrées dans l'équation de régression en tant que variable dépendante et les variables indépendantes (Y, $X1, X2$ et $X3$) et l'analyse n'a exclu aucune variable. Il montre également la méthode utilisée, qui est la régression standard.Variable dépendante : Y

– Prédicteurs : (Constante),$X3, X1, X2$.

Le tableau précédent montre le coefficient de corrélation de Pearson entre la variable dépendante et la variable indépendante, atteignant une valeur élevée de $0, 815$ avec une valeur du coefficient de détermination sur $0, 665$ et la valeur du coefficient de détermination ajusté $0, 626$, signifiant que la variable indépendante explique 62% de la variance de la satisfaction au travail.

– Variable dépendante : Y

– Prédicteurs : (Constante), $X3, X1, X2$

Le tableau précédent montre les résultats de l'analyse ANOVA pour tester la signification de la régression, et nous notons que la valeur de Sig est 000, qui est inférieure à 0.01. Ainsi, nous rejetons l'hypothèse nulle et acceptons l'hypothèse alternative, qui est que la régression est significative et donc il y a un effet des variables indépendantes sur la variable dépendante, et nous pouvons prédire la variable dépendante à travers ces variables indépendantes.

– Coeff non stad : Coefficients non standardisés.

		Coeff non stad		Coeff stad		
Modèle		B	Erreur standard	Bêta	t	Sig
	$(Constante)$	$664,116$	$405,166$		$1,639$	112
	$X1$	$56,742$	$6,725$	0.847	$8,437$	000

TAB. 3.6 – Coefficients

	Minimum	Maximum	Moyenne	Ecart type	N
Valeur prédite	$2933,8125$	$5657,4478$	$4023,2667$	$645,31338$	30
de Student	$905,63019$	$813,79382$	00000	$404,70775$	30
Valeur prévue standard	$-1,688$	$2,532$	000	$1,000$	30
Résidu standard	$-2,199$	$1,976$	000	983	30

TAB. 3.7 – Statistiques des résidusa

– Coeff stad : Coefficients standardisés.

Le tableau précédent présente les facteurs de régression standard et non standard, la ligne standard, la valeur du test avec la valeur de probabilité des tests (indication statistique) Le tableau précédent aide également à écrire l'équation de la droite de régression

Predicted(JS)$= -0.012 * X1 + 0.512 * X2 + 0.393 * X3 + e.$

Les figures précédentes montrent la modération de la distribution des résidus et collectent des données autour de la ligne droite.Par conséquent, les résidus suivent la distribution normale, qui est l'une des conditions de validité de l'analyse de régression.

PROF :

Pour illustrer les notions précédentes avec le logiciel spss on peut considérer le jeu de données "profs".

Dans une étude statistique, 23 professeurs sont évalués quant à la qualité de leur enseignement. Pour chacun d'entre eux, on dispose :

Modèle	Variables introduites	Variables éliminées	Méthode
1	$X3, X1, X2^b$	.	Introduire

TAB. 3.8 – Variables introduites/éliminéesa

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	0.815	0.665	0.626	$10,37921$

TAB. 3.9 – Récapitulatif des modèlesb

d'un indice de performance globale donné par les étudiants (variable Y), des résultats de 4 tests écrits donnés à chaque professeur (variables X_1, X_2, X_3 et X_4),du sexe (variableX_5, avec $X_5 = 0$ pour femme, $X_5 = 1$ pour homme).L'objectif est d'expliquer Y à partir de X_1, X_2, X_3, X_4 et X_5.Le jeu de données est disponible ici :

Le tableau représente les données obtenues à partir d'une étude d'un indice de performance globale donné par les étudiants (variable Y),de 23 professeurs et,des résultats de 4 tests écrits donnés à chaque professeur (variables X_1, X_2, X_3 et X_4),du sexe (variableX_5, avec $X_5 = 0$ pour femme, $X_5 = 1$ pour homme)

En entrant des données dans le programme, nous obtenons les résultats suivants : [8]

Le tableau précédent présente des statistiques descriptives (la taille de l'échantillon est de 23.moyennes lorsque la de moyenne Y est de $444,3913$, la moyenne de $X1$ est de $80,9130$ la moyenne de $X2$ est de $143,0870$ la moyenne de $X3$ est de $56,04$ la moyenne de $X4$ est de $50,1304$ la moyenne de $X5$ est de $0,5217$) Il s'agit des variables introduites dans le modèle de régression dépendante et autonome.

Le tableau précédent montre la matrice de corrélation entre les variables du modèle de régression,

– Où le coefficient de corrélation le plus élevé entre Y et $X4$ était de $0,535$ avec

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig
1	Régression	$5558,888$	3	$1852,963$	$17,200$	000^b
	de Student	$2800,927$	26	$107,728$		
	Total	$8359,815$	29			

TAB. 3.10 – ANOVA

		Coeff non stad		Coeff stad		
Modèle		B	Erreur standard	Bêta	t	Sig
1	$(Constante)$	$-29,216$	$19,657$		$-1,486$	149
	$X1$	-0.342	$4,524$	-0.012	-0.076	940
	$X2$	$4,090$	$1,240$	0.512	3.297	003
	$X3$	$21,984$	$9,174$	0.393	$2,396$	024

TAB. 3.11 – Coefficients

une signification inférieure à 0.01.

– Le coefficient de corrélation entre Y et $X1$ est de $0,440$.

– Le coefficient de corrélation entre Y et $X2$ est de $0,405$.

– Coefficient de corrélation minimal entre Y et $X3$ est de $-0,243$.

– Le coefficient de corrélation entre Y et $X5$ est de $0,165$.

– Variable dépendante : Y.

– Toutes les variables demandées ont été introduites.

.Le tableau montre les noms des variables qui sont entrées dans l'équation de régression en tant que variable dépendante et les variables indépendantes (Y et $X5, X4, X3, X2, X1$) et l'analyse n'a exclu aucune variable. Il montre également la méthode utilisée, qui est la régression standard.

– Prédicteurs : (Constante), $X5, X4, X3, X2$,$X1$.

– Variable dépendante :Y

Le tableau précédent montre le coefficient de corrélation de Pearson entre la variable dépendante et les variables indépendantes, où il atteint une valeur élevée de 0.824 en divisant le coefficient de détermination par 0.679 et la valeur du coeffi-

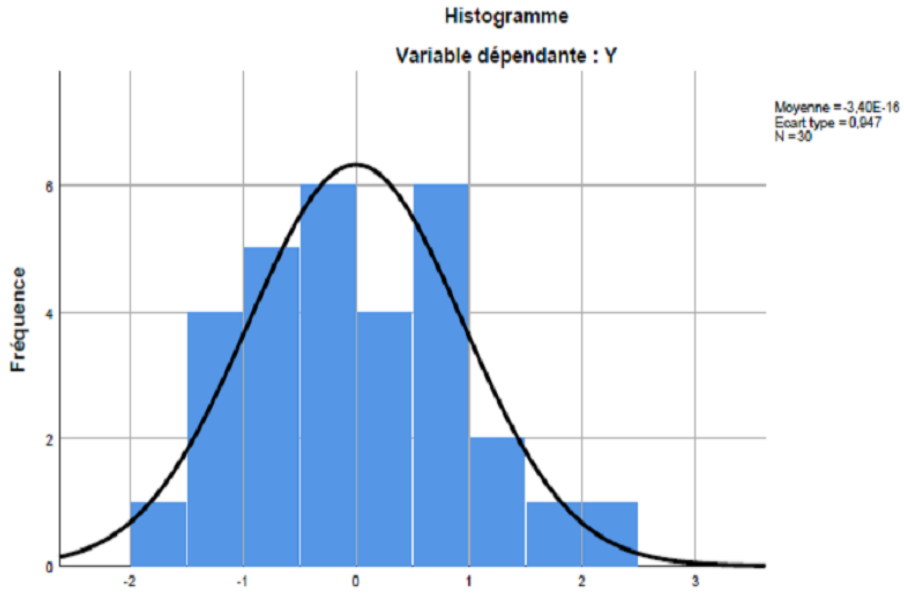

FIG. 3.3 – Régression Résiduel standardisé

cient de détermination modifié 0.584, c'est-à-dire que les variables indépendantes expliquent 58% de la variance de la satisfaction au travai.

Le tableau précédent montre l'analyse de l'ANOVA pour tester la signification de la régression, et nous notons que la valeur du sig 1 sur 1000 est inférieure à 5%. Il existe des différences statistiquement significatives.

Le tableau précédent présente les facteurs de régression standard et non standard, la ligne standard, la valeur du test avec la valeur de probabilité des tests (indication statistique) et la valeur des transactions d'inflation de variabilité (variance inflation factor)VIF.Opérations de tolérance :VIF=1/ Tolérance

Ce qui montre qu'il n'y a pas de problème de multiplicité linéaire entre les variables où les facteurs d'inflation étaient inférieurs à 3

Le tableau précédent aide également à écrire l'équation de pente

Predicted(JS)=$0,266 * X1 + 0,443 * X2 + -0,258 * X3 + 0,571 * X4 + 0,086 + e$.

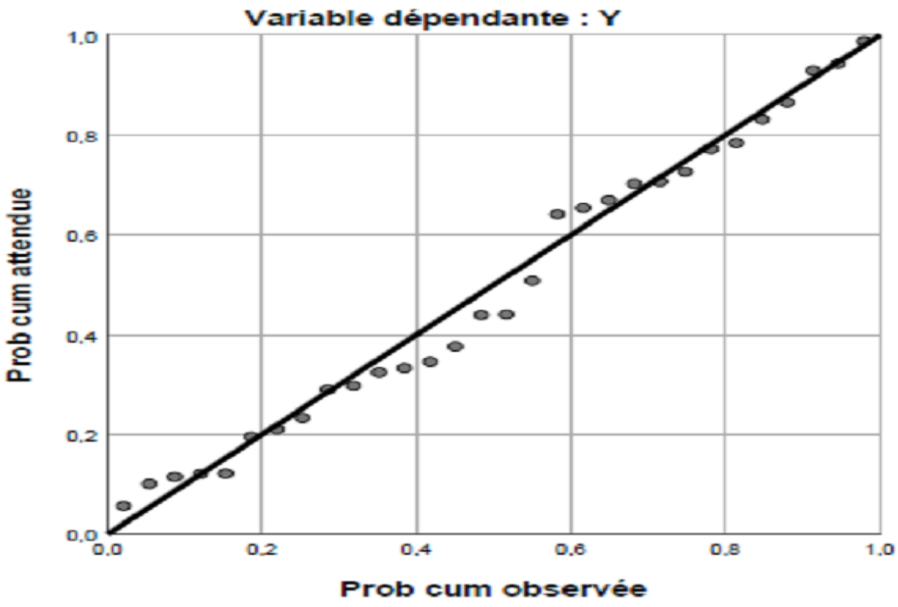

FIG. 3.4 – Trace p-p normal de régression résiduel standardisé

3.1.2 Modèle de régression linéaire simple

Droite de régression :

On appelle droite de régression la droite qui ajuste au mieux le nuage de points. Cet ajustement se fait en termes de distance euclidienne, les points de la droite étant pris aux mêmes abscisses que ceux des points du nuage. La droite de régression est donnée par l'équation : $y = b_0 + b_1 x$..Comme $b_0 = \bar{y} - b_1 \bar{x}_1$, notons que la droite de régression passe par le point G de coordonnée $(x1; y)$,appelé point moyen, centre d'inertie ou centre de gravité du nuage de points

Nuage de points :

Les nuages de points associées aux exemples introduits précédents sont présentés (scores fibres toluca) ci-dessous :

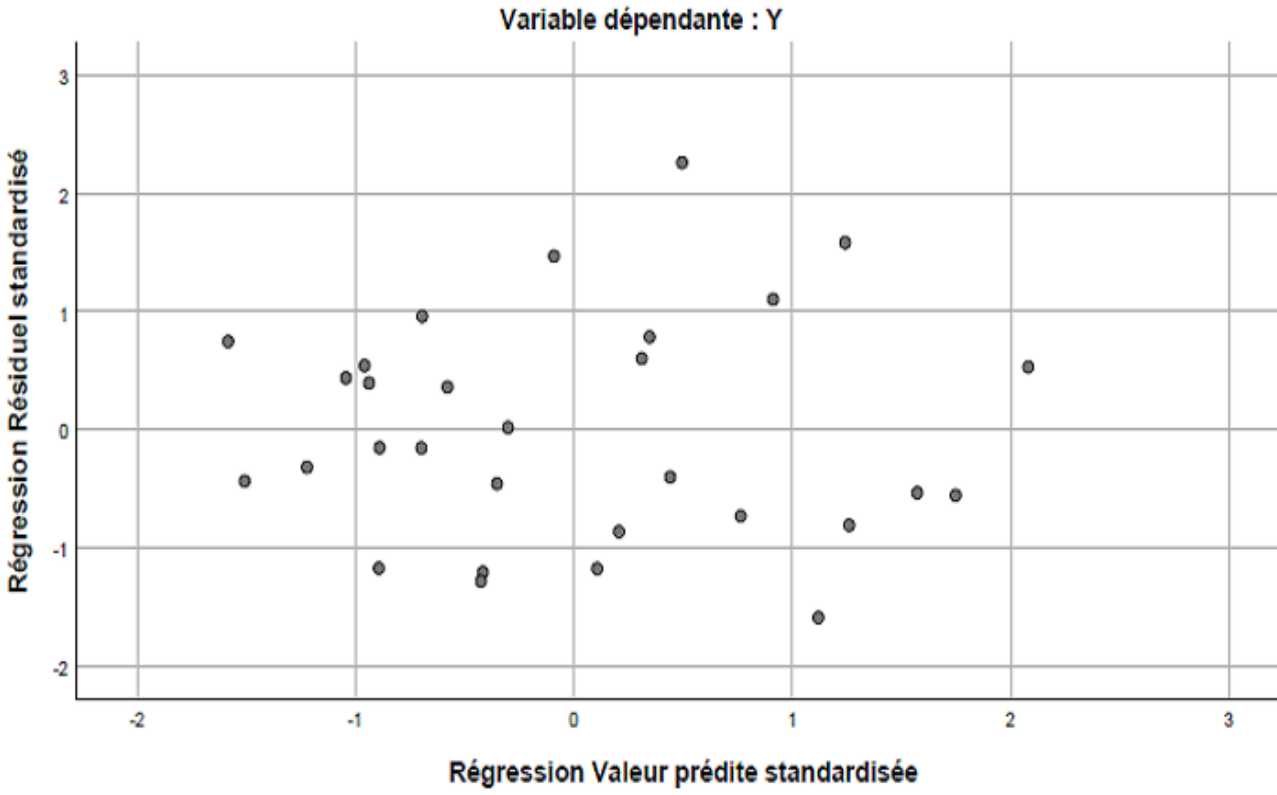

FIG. 3.5 – Nuage de points

3.1.3 Modèle de régression linéaire simple :

Pour illustrer le résultat théorique précédent, on peut considérer le jeu de données "loyers". Dans un quartier parisien, une étude a été menée afin de mettre en évidence une relation entre le loyer mensuel et la surface des appartements ayant exactement 3 pièces.Pour 30 appartements de ce type, on dispose :

– de la surface en mètres carrés (variable X_1),

– du loyer mensuel en francs (variable Y).

L'objectif est d'expliquer Y à partir de X_1.Le jeu de données est disponible ici :

Écrire dans une fenêtre R :

w = read.table("https ://chesneau.users.lmno.cnrs.fr/loyers.txt", header=T)

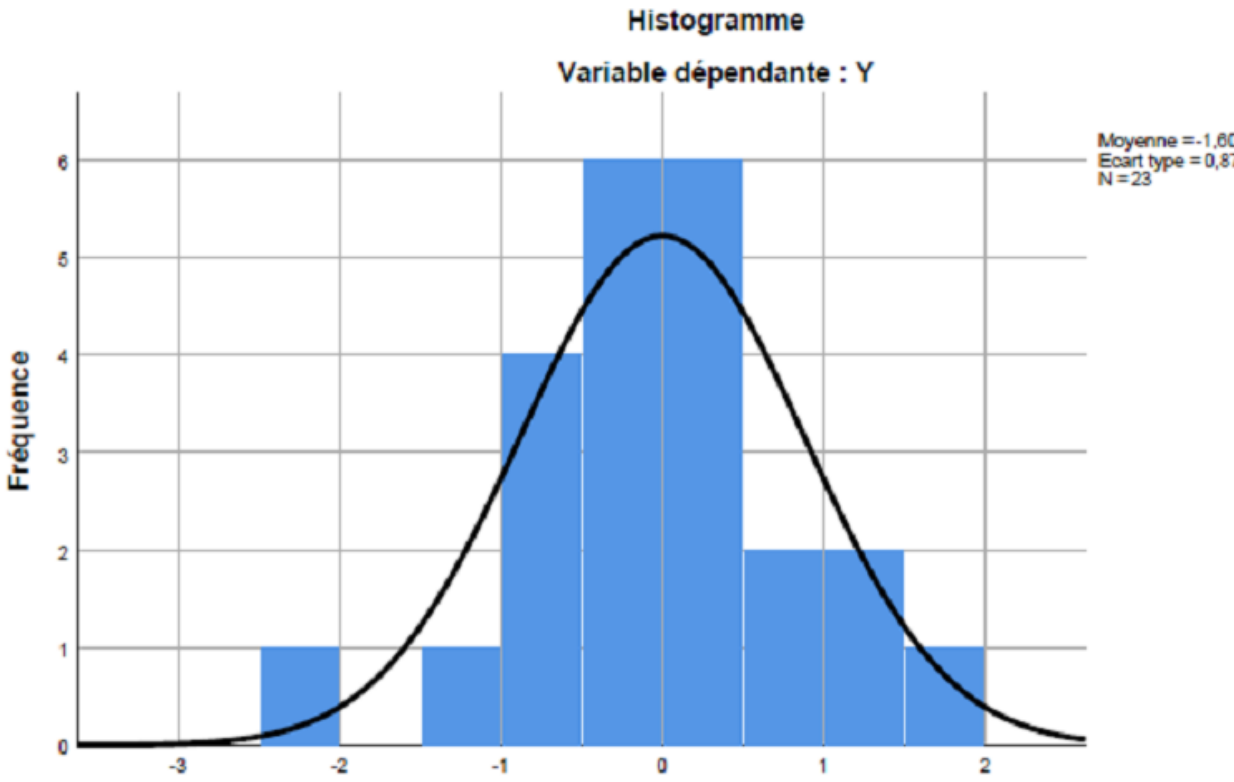

FIG. 3.6 – Régression Résiduel standardisé

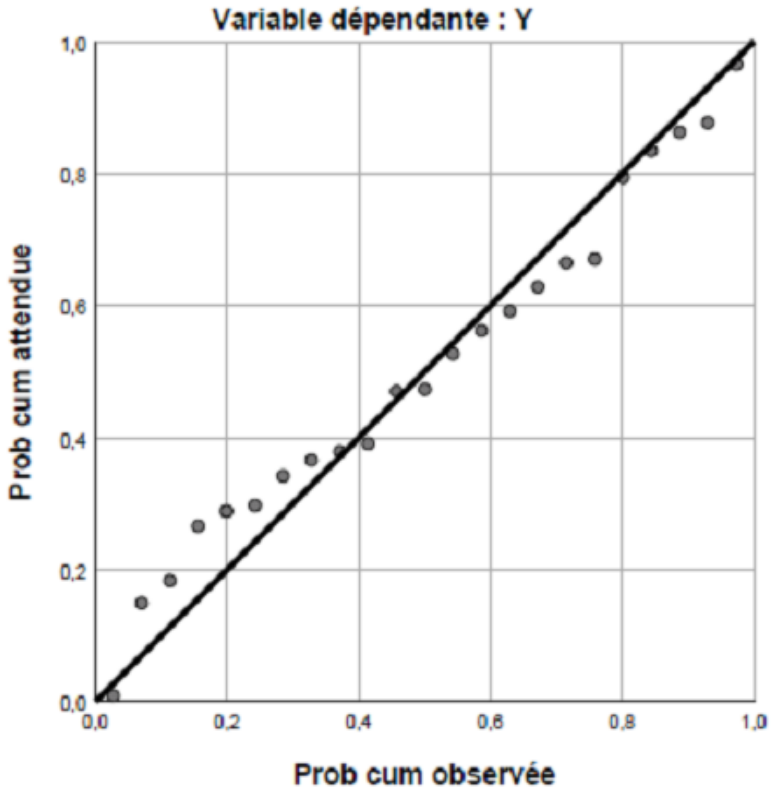

FIG. 3.7 – Trace p-p normal de régression résiduel standardisé

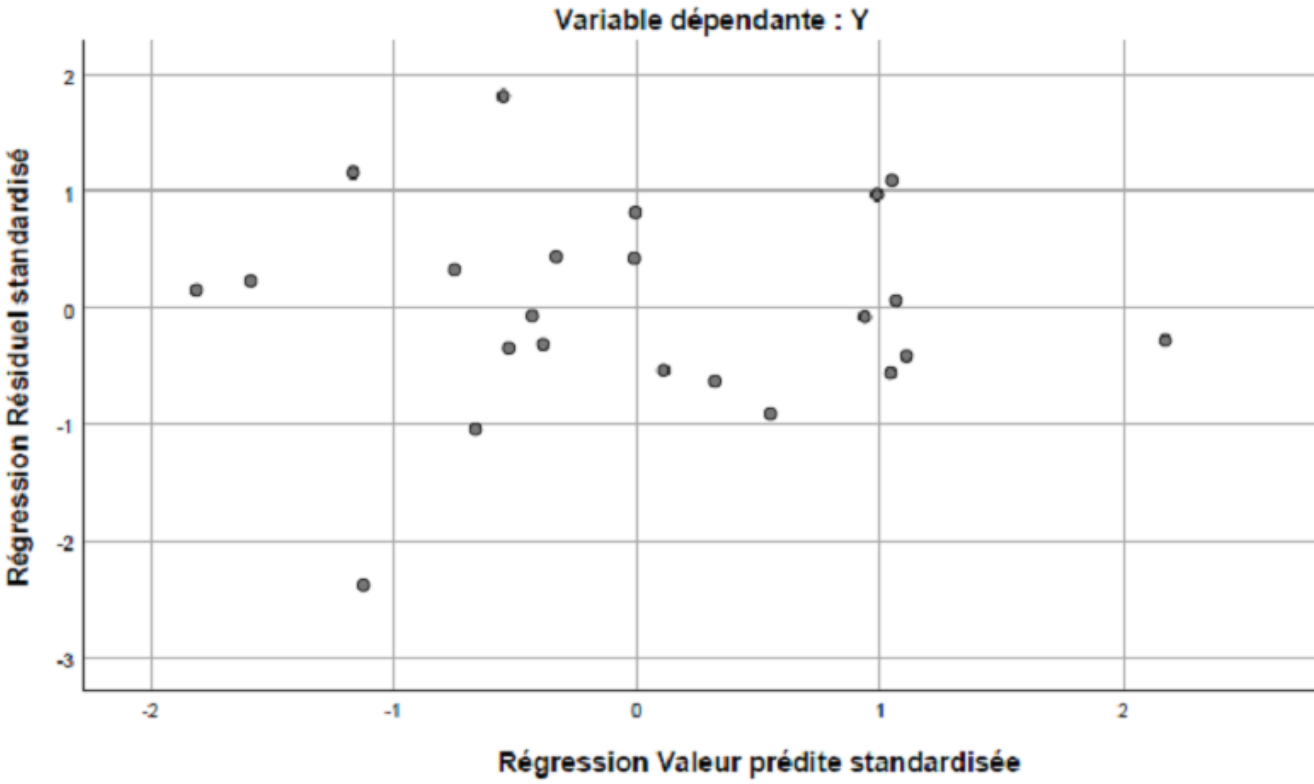

FIG. 3.8 – Nuage de pointes

Y	489; 423; 507; 467; 340; 524; 488; 445; 388; 579; 433; 409 410; 568; 425; 344; 324; 505; 234; 501; 400; 584; 434.
	81; 68; 80; 107; 43; 129; 139; 88; 99; 121; 91; 87.
$X1$	69; 57; 77; 81; 0; 53; 77; 76; 65; 97; 76.
	151; 156; 165; 149; 134; 163; 159; 135; 141; 145; 129; 115
$X2$	125; 131; 141; 122; 141; 152; 141; 132; 157; 166; 141.
	45.5; 46.45; 76.5; 55.5; 49.4; 72.0; 86.2; 64.0; 44.15; 42.5; 79.25; 59
$X3$	31.75; 80.5; 75.0; 49.0; 49.35; 60.75; 41.25; 50.75; 32.25; 54.52.
	43.61; 44.69; 54.57; 43.27; 49.21; 49.96; 39.57; 51.89; 53.77; 56.32
$X4$	55.66; 63.97; 45.32; 46.67; 41.21; 43.83; 41.61; 64.57; 42.41; 57.95; 57.90.
$X5$	1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1 0; 0; 0; 0; 0; 0; 0; 0; 0; 0.

TAB. 3.12 – le jeu de données "profs"

attach(w))

head(w

Cela renvoie l'entête du jeu de données :

	Y	$X1$
1	3000	40
2	2844	44
3	3215	44
4	2800	45
5	3493	45
6	3140	48

Table 19

Le nuage de points associé est donné par les commandes R :

plot($X1, Y$)

Y	489; 423; 507; 467; 340; 524; 488; 445; 388; 579; 433; 409 410; 568; 425; 344; 324; 505; 234; 501; 400; 584; 434.
	81; 68; 80; 107; 43; 129; 139; 88; 99; 121; 91; 87.
$X1$	69; 57; 77; 81; 0; 53; 77; 76; 65; 97; 76.
	151; 156; 165; 149; 134; 163; 159; 135; 141; 145; 129; 115
$X2$	125; 131; 141; 122; 141; 152; 141; 132; 157; 166; 141.
	45.5; 46.45; 76.5; 55.5; 49.4; 72.0; 86.2; 64.0; 44.15; 42.5; 79.25; 59
$X3$	31.75; 80.5; 75.0; 49.0; 49.35; 60.75; 41.25; 50.75; 32.25; 54.52.
	43.61; 44.69; 54.57; 43.27; 49.21; 49.96; 39.57; 51.89; 53.77; 56.32
$X4$	55.66; 63.97; 45.32; 46.67; 41.21; 43.83; 41.61; 64.57; 42.41; 57.95; 57.90.
$X5$	1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1 0; 0; 0; 0; 0; 0; 0; 0; 0; 0.

TAB. 3.13 – Statistiques descriptives

Le nuage de points étant étiré dans une direction, le modèle de rls est envisageable.

Sa forme générique est

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon,$$

ou β_0 et β_1 sont des coefficients réels inconnus.Pour estimer ponctuellement β_0 et β_1 nous allons utiliser les formules analytiques de $b0$ et $b1$:

$b1 = (1/(sum((X1 - mean(X1))\hat{}2))) * sum((X1 - mean(X1)) * (Y - mean(Y)))$

$b0 = mean(Y) - mean(X1) * b1$

$b0$

$b1$

Cela renvoie :$b0 = 548.9782$ et $b1 = 58.37875$.On peut calculer le R^2 en utilisant l'égalité : $R^2 = r^2_{x;y}$:

$R2 = cor(Y, X1)\hat{}2$

Cela renvoie : 0 : 7311242.De même pour le R^2ajusté :

		Y	$X1$	$X2$	$X3$	$X4$	$X5$
Corrélation de Pearson	Y	$1,000$	$0,440$	$0,405$	$-0,243$	$0,535$	$0,165$
	$X1$	$0,440$	$1,000$	$0,277$	$0,321$	$0,160$	$0,494$
	$X2$	$0,405$	$0,277$	$1,000$	$-0,021$	$-0,230$	$0,157$
	$X3$	$-0,243$	$0,321$	$-0,021$	$1,000$	$-0,135$	$0,190$
	$X4$	$0,535$	$0,160$	$-0,230$	$-0,135$	1.000	$-0,129$
	$X5$	$0,165$	$0,494$	$0,157$	$0,190$	$-0,129$	$1,000$
Sig. (unilatéral)	Y		$0,018$	$0,028$	$0,132$	$0,004$	$0,226$
	$X1$	$.0,018$		$0,100$	$0,067$	$0,233$	$0,008$
	$X2$	$0,028$	$0,100$		$0,461$	$0,146$	$0,237$
	$X3$	$0,132$	$0,067$	$0,461$		$0,270$	$0,193$
	$X4$	$,004$	$,233$	$,146$	$,270$		$,278$
	$X5$	$0,226$	$0,008$	$0,237$	$0,1930$	$0,278$	
N	Y	23	23	23	23	23	23
	$X1$	23	23	23	23	23	23
	$X2$	23	23	23	23	23	23
	$X3$	23	23	23	23	23	23
	$X4$	23	23	23	23	23	23
	$X5$	23	23	23	23	23	23

TAB. 3.14 – Corrélations

Modèle	Variables introduites	Variables éliminées	Méthode
1	$X5, X4, X3, X2, X1^b$	.	Introduire

TAB. 3.15 – Variables introduites/éliminéesa

$R2aj = 1 - ((30-1)/(30-(1+1))) * (1-R2)$

$R2aj$

Cela renvoie : 0.7215215.Le R^2(et $\bar{R}^2$) étant proche de 1, le modèle de rls semble être pertinent avec les données traitées

Commande summary : On retrouve plus simplement ces estimations (et beaucoup plus) avec la commande summary :

$reg = lm(Y \tilde{} X1)$

$summary(reg)$

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	0.824^a	0.679	0.584	$55,47270$

TAB. 3.16 – Récapitulatif des modèlesb

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig
1	Régression	$110506,732$	5	$22101,346$	$7,182$	$0,001^b$
	de Student	$52312,747$	17	$3077,220$		
	Total	$162819,478$	22			

TAB. 3.17 – ANOVA

Cela renvoie :

– Residual standard error : 409.7 on 28 degrees of freedom

– Multiple R-squared : 0.7311, Adjusted R-squared : 0.7215

– F-statistic : 76.14 on 1 and 28 DF, p-value : $1.783e-09$.

On retrouve b_0 et b_1 dans la colonne Estimate du tableau.On retrouve également : $R^2 = 0.7311$ et $\bar{R}^2 = 0.7215$.D'autre part, la droite de régression est donnée par l'équation :

$$y = b_0 + b_1 x = 548.9782 + 58.3787x.$$

On peut la visualiser en faisant :

plot $(X1, Y)$

abline(reg, col = "red")

		Coeff non stad		Coeff stad				
Modèle		B	Erreur standard	Bêta	t	Sig	Modèle	
1	$(Constante)$	$-271,666$	$185,887$		$-1,461$	$,162$	1	$(Constante)$
	$X1$	$0,784$	$0,542$	$0,266$	$1,448$	$0,166$		$X1$
	$X2$	$2,698$	$0,928$	$0,443$	$2,907$	$0,010$		$X2$
	$X3$	$-1,416$	$0,828$	$-0,258$	$-1,710$	$0,105$		$X3$
	$X4$	$6,749$	$1,831$	$0,571$	$3,687$	$0,002$		$X4$
	$X5$	$14,564$	$27,500$	$0,086$	$0,530$	$0,603$		$X5$

TAB. 3.18 – Coefficients

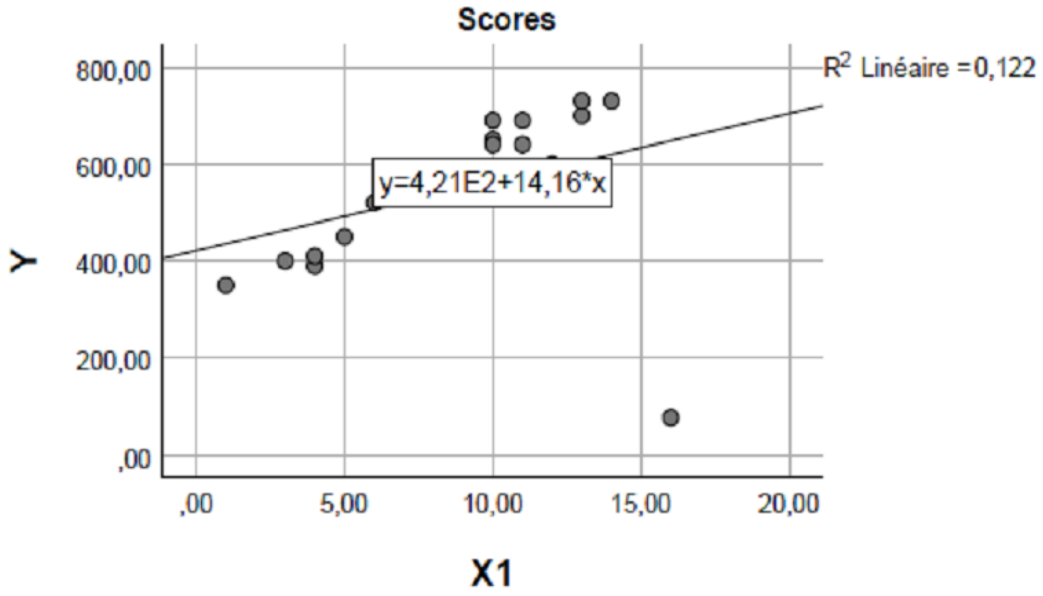

FIG. 3.9 – nuages de points de scores

3.2 Mise en oeuvre avec le logiciel R

3.2.1 Modèle de régression linéaire multiple

Pour illustrer les notions précédentes avec le logiciel R on peut considérer le jeu de données "profs".

Dans une étude statistique,23 professeurs sont évalués quant à la qualité de leur enseignement. Pour chacun d'entre eux, on dispose :

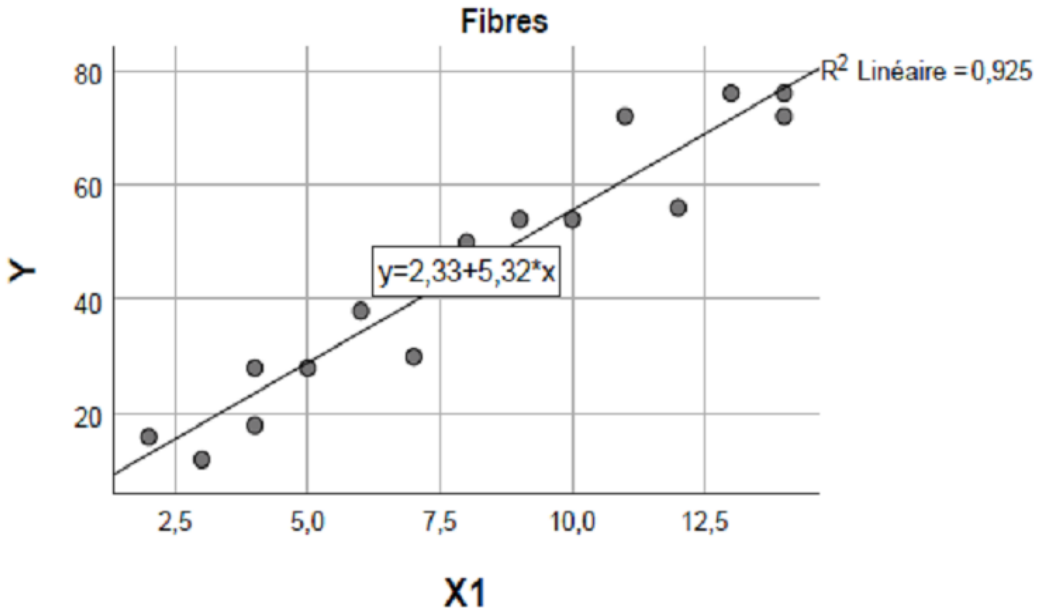

FIG. 3.10 – nuages de points de Fibres

	3000; 2844; 3215; 2800; 3493; 3140; 3593; 3688; 4452; 3361
Y	4542; 3181; 3575; 4017; 3430; 4226; 3639; 4015; 4741; 3628
	4429; 3390; 4766; 4413; 4900; 5020; 4962; 5294; 4847; 5549.
	40; 44; 44; 45; 45; 48; 49; 52; 53; 54
$X1$	54; 54; 55; 57; 57; 59; 61; 62; 62; 63;
	63; 64; 67; 69; 70; 72; 74; 75; 76; 88.

TAB. 3.19 – le jeu de données "loyers"

D'un indice de performance globale donné par les étudiants (variable Y),des résultats de 4 tests écrits donnés à chaque professeur (variables X_1, X_2, X_3 et X_4), du sexe (variableX_5, avec $X_5 = 0$ pour femme, $X_5 = 1$ pour homme).L'objectif est d'expliquer Y à partir de X_1, X_2, X_3, X_4 et X_5.Le jeu de données est disponible ici :

Le tableau représente les données obtenues à partir d'une étude d'un indice de performance globale donné par les étudiants (variable Y),de 23 professeurs et,des résultats de 4 tests écrits donnés à chaque professeur (variables X_1, X_2, X_3 et X_4),du sexe (variableX_5, avec $X_5 = 0$ pour femme, $X_5 = 1$ pour homme)

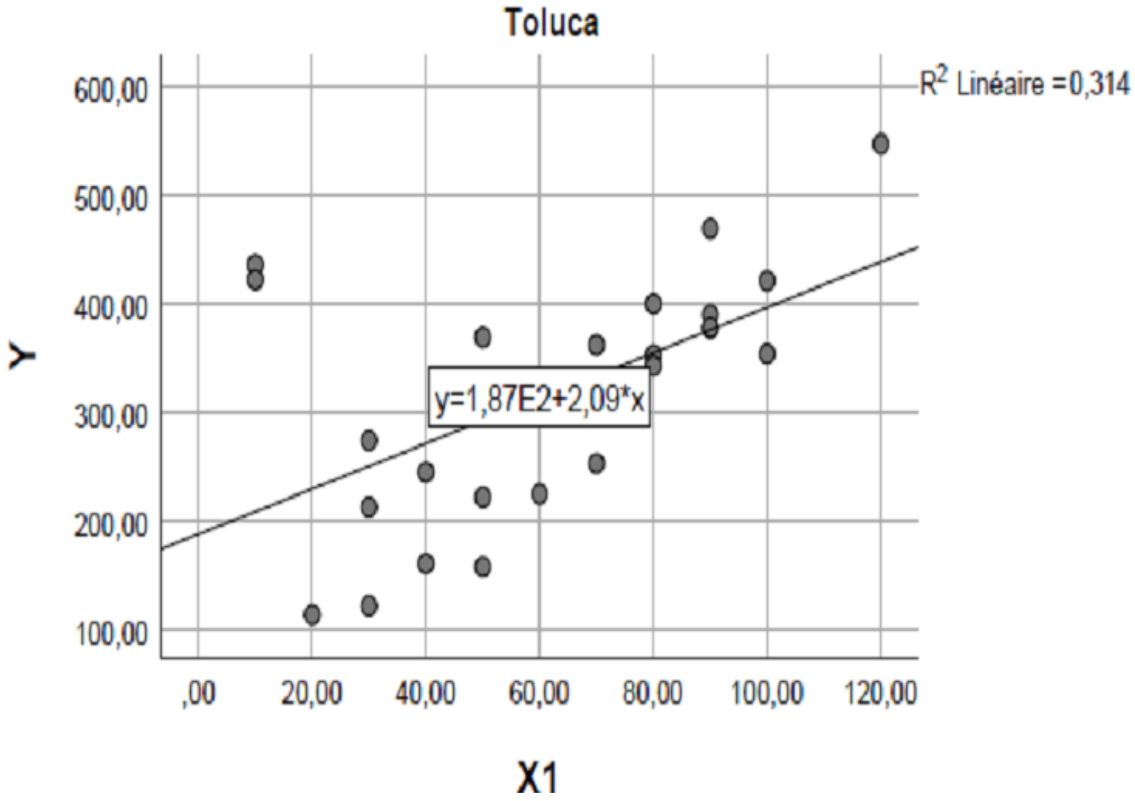

FIG. 3.11 – nuage de points toluca

	Estimate	Std. Error	t value	Pr(>| t |)
(Intercept)	**548.9782**	403.0783	1.36	0.1841
$X1$	**58.3787**	6.6905	8.73	***

TAB. 3.20 – résultat commande summary

w = read.table("https ://chesneau.users.lmno.cnrs.fr/profs.txt", header = T)

attach(w)

head(w)

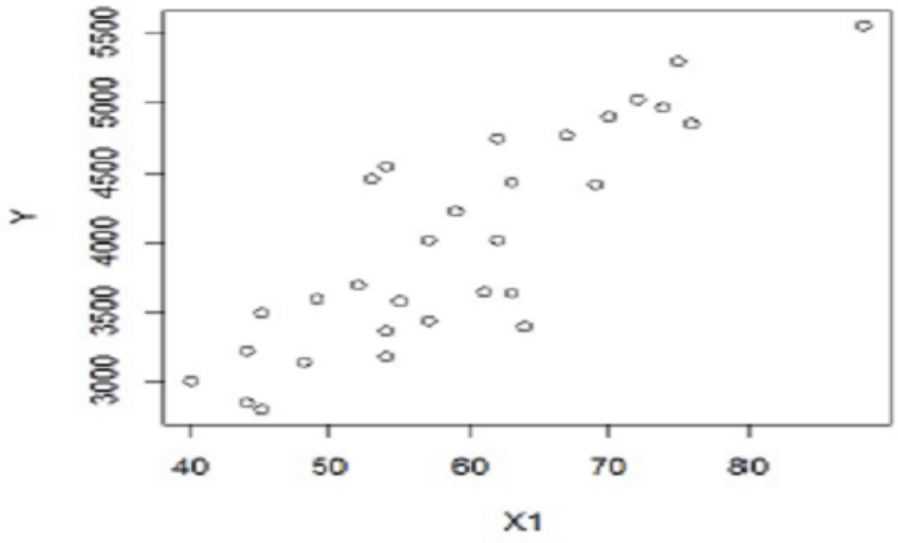

FIG. 3.12 – Nuage de points de loyers

Cela renvoie l'entête du jeu de données :

	Y	$X1$	$X2$	$X3$	$X4$	$X5$
1	489	81	151	45.50	43.61	1
2	423	68	156	46.45	44.69	1
3	507	80	165	76.50	54.57	1
4	467	107	149	55.50	43.27	1
5	340	43	134	49.40	49.21	1
6	524	129	163	72.00	49.96	1

Le modèle de rlm est envisageable. Sa forme générique est :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

où β_0, $\beta_1, \beta_2, \beta_3, \beta_4$,et β_5 sont des coefficients réels inconnus. On le considère sous

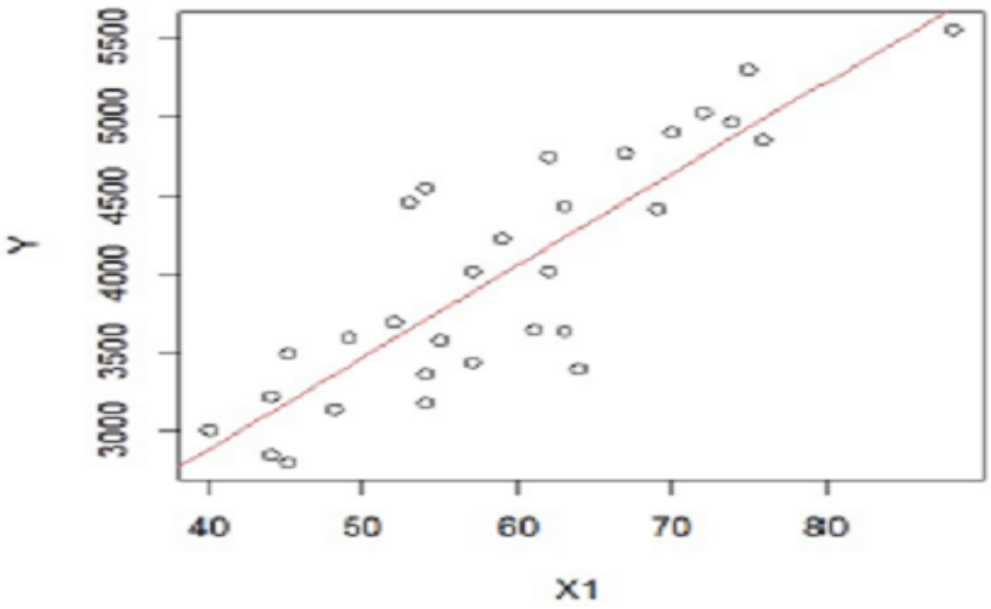

FIG. 3.13 – droite de régression de nuge de points "loyers"

sa forrme matricielle :$Y = X\beta + \varepsilon$, où :

$$X = \begin{bmatrix} 1 & 81 & 151 & 45.50 & 43.61 & 1 \\ 1 & 68 & 156 & 46.45 & 44.69 & 1 \\ 1 & 80 & 165 & 76.50 & 54.57 & 1 \\ 1 & 107 & 149 & 55.50 & 43.27 & 1 \\ 1 & 43 & 134 & 49.40 & 49.21 & 1 \\ 1 & 129 & 163 & 72.00 & 49.96 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}, Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_{23} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \vdots \\ \varepsilon_{23} \end{pmatrix}$$

Nous allons maintenant étudier l'emco de β correspondant aux données.Il s'agit

Y	489; 423; 507; 467; 340; 524; 488; 445; 388; 579; 433; 409 410; 568; 425; 344; 324; 505; 234; 501; 400; 584; 434.
	81; 68; 80; 107; 43; 129; 139; 88; 99; 121; 91; 87.
$X1$	69; 57; 77; 81; 0; 53; 77; 76; 65; 97; 76.
	151; 156; 165; 149; 134; 163; 159; 135; 141; 145; 129; 115
$X2$	125; 131; 141; 122; 141; 152; 141; 132; 157; 166; 141.
	45.5; 46.45; 76.5; 55.5; 49.4; 72.0; 86.2; 64.0; 44.15; 42.5; 79.25; 59
$X3$	31.75; 80.5; 75.0; 49.0; 49.35; 60.75; 41.25; 50.75; 32.25; 54.52.
	43.61; 44.69; 54.57; 43.27; 49.21; 49.96; 39.57; 51.89; 53.77; 56.32
$X4$	55.66; 63.97; 45.32; 46.67; 41.21; 43.83; 41.61; 64.57; 42.41; 57.95; 57.90.
$X5$	1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1 0; 0; 0; 0; 0; 0; 0; 0; 0; 0.

TAB. 3.21 – le jeu de données "profs"

donc de calculer l'emco ponctuel β défini par :

$$b = (X^t X)^{-1} X^t y, \qquad y = \begin{pmatrix} 489 \\ 423 \\ 507 \\ \vdots \end{pmatrix}$$

Introduisons la matrice X composée des colonnes "que des 1", X_1, X_2, X_3, X_4 et X_5 :

$X = cbind(1, X_1, X_2, X_3, X_4, X_5.)$

En utilisant les commandes R : $\% * \%$ = produit matriciel, $t(A) = A^t$ et solve $(A) = A^{-1}$.Calculons $b = (X^t X)^{-1} Xty$:

b=solve(t(X)%*%X)%*%t(X)%*%Y Cela renvoie :

b

$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix} = \begin{pmatrix} -272.04 \\ 0.79 \\ 2.68 \\ -1.44 \\ 6.83 \\ 14.90 \end{pmatrix}$$

Entre autre, ces estimations nous permettent de faire des prédictions sur Y pour de nouvelles valeurs de $(X_1, X_2, X_3, X_4, X_5)$.

Par exemple, pour $(X_1, X_2, X_3, , X_4, X_5) = (82, 158, 47, 49, 1) = x$, en posant $x_\bullet = (1, 82, 158, 47, 49, 1)$,

la valeur prédite de Y est $dx = x_\bullet b$. Cela s'obtient en faisant :

x=c(1,82,158,47,49,1)
d=x%*%b
d

Cela renvoie : 498.5063.Ainsi, pour de tels critères, l'indice de performance globale moyen est de 498.5063.Le R^2 peut se calculer en faisant :

R2 = 1 - sum((X %*% b - Y)^2) / sum((mean(Y) - Y)^2)

R2

R2aj=1-((length(Y)-1)/(length(Y)-(5+1)))*(1-R2)

R2aj

Cela renvoie :0.5903106.

Le R^2(et $\bar{R}^2$) étant relativement proche de 1, le modèle de rlm semble être perti-

nent avec les données traitées.On constate que cette droite ajuste correctement le nuage de points ; les prédictions issues dumodèle sont alors relativement fiables.Par exemple, pour $X_1 = 56 = x$, la valeur prédite de Y est :

$$dx = b_0 + b_1 \times 56 = 548.9782 + 58.378756 = 3818.185$$

Ainsi, pour une surface de 56 mètres carrés, le loyer mensuel moyen est de 3818.185 francs.On aurait aussi pu utiliser les commandes R :

predict (reg,data.frame(X1=56))

Dorénavant, dès que possible, on utilisera la commande summary dans les analyses.

Commande summary :

On retrouve plus simplement ces estimations (et beaucoup plus) avec la commande summary :

reg=lm(Y~X1+X2+X3+X4+X5)

summary(reg)

Cela renvoie :

	Estimate	Std. Error	t value	Pr(>jtj)
(Intercept)	-272.0388	184.3865	-1.48	0.1584
$X1$	0.7913	0.5363	1.48	0.1583
$X2$	2.6828	0.9216	2.91	0.0097
$X3$	-1.4434	0.8217	-1.76	0.0970
$X4$	6.8308	1.8192	3.75	0.0016
$X5$	14.9008	27.3134	0.55	0.5925
Table 21				

– Residual standard error : 55.06 on 17 degrees of freedom.

– Multiple R-squared : 0.6834, Adjusted R-squared : 0.5903.

– F-statistic : 7.34 on 5 and 17 DF, p-value : 0.0007887.

On retrouve b dans colonne Estimate du tableau.On retrouve également : $R^2 = 0.6834$ et $\bar{R}^2 = 0.5903$.Pour la valeur prédite de Y quand $(X_1, X_2, X_3, X_4, X_5) = (82, 158, 47, 49, 1)$,on peut faire :

predict(reg,data.frame($X1 = 82, X2 = 158, X3 = 47, X4 = 49, X5 = 1$))

Chapitre 4

Conclusion

Dans ce mémoire, nous nous sommes intéressés aux problèmes d'identification des-systèmes linéaires à une entrée et une sortie. Ces systèmes peuvent avoir un modèle mathématique de regréssion linéare simple ou bien multiple. Nous nous somme plus exactement intéressé au problème d'identification au sens des moindres carrés, où nous nous sommes appliqués à démontrer un certain nombre de propriétés reliées à ce sujet. moindres carrés dans les deux cas de modélisations.

Plus précisément, nous avons présenté Les définitions principaux des modéles de regréssion linéare et quelques notions de bases, la Coefficient de détermination et leLoi normale multidimensionelle et présenté estimateurs de moindres carrés de deux modéle et propriétés standards et lois associées de cette estimateurs

A la fin. Nous avons présenté une application numérique des modéles de regréssion linéare Et nous comparons entre le modèle sur logiciel R et spss Et aussi présenté une application numérique des estimateurs de moindres carrés.

Bibliographie

[1] A. Monfort, "Cours de statistique mathématique", Économica, 1997.

[2] C.R. Rao, "Estimation of variance and covariance components in linear models", Journal of the American Statistical Association, 67 (337), 112-115, 1972

[3] C Chesneau .cours Statistique 2 du M1 orienté statistique de l'université de Caen 22 Février 2020 (https ://chesneau.users.lmno.cnrs.fr/loyers.txt)

[4] C Chesneau .cours Statistique 2 du M1 orienté statistique de l'université de Caen 22 Février 2020 (https ://chesneau.users.lmno.cnrs.fr/fromages.txt)

[5] C Chesneau .cours Statistique 2 du M1 orienté statistique de l'université de Caen 22 Février 2020(https ://chesneau.users.lmno.cnrs.fr/scores.txt)

[6] C Chesneau .cours Statistique 2 du M1 orienté statistique de l'université de Caen 22 Février 2020(https ://chesneau.users.lmno.cnrs.fr/fibres.txt)

[7] C Chesneau .cours Statistique 2 du M1 orienté statistique de l'université de Caen 22 Février 2020(https ://chesneau.users.lmno.cnrs.fr/toluca.txt)

[8] C Chesneau .cours Statistique 2 du M1 orienté statistique de l'université de Caen 22 Février 2020(https ://chesneau.users.lmno.cnrs.fr/profs.txt)

[9] G. Verbeke & G. Molenberghs, "Linear mixed models for longitudinaldata", Springer, 2000.

[10] J.-M. Azaïs & J.-M. Bardet, “Le modèle linéaire par l’exemple”,Dunod,2005.

[11] R.A. Fisher & F. Yates, “Statistical tables”, Oliver and Boyd,1963.

Résumé.

Ce mémoire porte sur la méthode de moindre carre et ces applications aux modèles linéaires. Nous avons applique cette méthode à l'estimation des régressions simple et multiple. Cette méthode offre des résultats satisfaisants en termes d'erreur moyenne quadratiques des estimateurs des coefficients des modèles évoqués. Nous avons appliquées cette dernière aux données réelles en utilisant les deux langages de statistique, R et SPSS.

Abstract.

This memory focuses on the least squares method and its applications to linear models. We applied this method to the estimation of simple and multiple regressions. This latter offers satisfactory results in terms of mean square error to the coefficients estimators of the aforementioned models. We applied this method to real data using the two statistical languages, R and SPSS.

ملخص.

تركز هذه الرسالة على طريقة المربعات الصغرى وتطبيقاتها على النماذج الخطية. طبقنا هذه الطريقة لتقدير الانحدارات البسيطة والمتعددة. تقدم هذه الطريقة نتائج مرضية من حيث متوسط الخطأ التربيعي لمقدرات معاملات النماذج المذكورة. قمنا بتطبيق الأخير على البيانات الحقيقية باستخدام اللغتين الإحصائيتين ، R و SPSS.