République Algérienne 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

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

BENABDERREZAk Ahlem

Titre :

Régression Linéaire Multiple : Théorie et Applications

Membres du Comité d'Examen :

Dr.	CHERFAOUI Mouloud	UMKB	Président
Dr.	**BERKANE Hassiba**	UMKB	Encadreur
Dr.	BENBRIKA Ghozlane	UMKB	Examinateur

Juin 2018

Dédicace

Je dédie ce mémoire à :

Mes parents :

Ma mère, qui a oeuvré pour ma réussite, de par son amour, son soutien, tous les sacrifices consentis et ses précieux conseils, pour toute son assistance et sa présence dans ma vie.

Mon père, qui peut être fier et trouver ici le résultat de longues années de sacrces et de privations pour m'aider à avancer dans la vie, l'éducation et le soutient permanent venu de toi.

Mes frères et mes amis.

BENABDERREZAk Ahlem

REMERCIEMENTS

Au nom de DIEU Le Plus Clément et Le Plus Miséricordieux.

Tout d'abord, je remercie ALLAH Le Tout Puissant qui m'a accordée la volonté et le courage pour réaliser ce mémoire.

Mes plus vifs remerciements et ma profonde gratitude BERKANE HASSIBA mon promoteur de mémoire, pour sa grande patience, pour sa disponibilité, pour ses nombreux conseils, pour ses corrections, et son appréciation au cours de l'élaboration de ce travail, pour sa grande patience qui ont constitué un considérable sans lequel ce travail n'aurait pas pu etre mené au bon, il me faudrait des pages pour le remercier.

Je suis très heureuse que Madam BENBRIKA Ghozlane ait accepté examinateur de ce travail avec diligence et pour l'honneur qu'il m'a faite de présider le jury de cette thèse. Je tiens sincèrement à le remercier.

J'ai de la chance aussi parce que Melle. CHERFAOUI Mouloud ait accepté d'être d'examiner ce travail

Je tiens aussi à remercier l'ensemble de mes camarades de master, mes amis, mes proches et ma famille (au sens large) qui m'ont soutenue durant ce travail, et spécialement à celles ou ceux, elles ou ils se reconnaîtront, qui m'ont encouragée à finir ce travail et qui m'ont accompagnée dans tous les moments de joie et de tristesse.

Enfin, je ne saurais terminer cette partie sans exprimer ma gratitude à mes parents, mes frères mes amis.

Merci.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'origine du mot régression vient de Sir Francis Galton. En 1885, travaillant sur l'hérédité. Cependant, l'analyse de causalité entre plusieurs variables est plus ancienne et remonte au milieu du xviiie siècle. En 1757, R. Boscovich, né à Ragusa, l'actuelle Dubrovnik, proposa une méthode minimisant la somme des valeurs absolues entre un modèle de causalité et les observations. Ensuite Legendre, dans son célèbre article de 1805, « Nouvelles méthodes pour la détermination des orbites des comètes », introduisit la méthode d'estimation par moindres carrés des coefficients d'un modèle de causalité et donna le nom à la méthode. Parallèlement, Gauss publia en 1809 un travail sur le mouvement des corps célestes qui contenait un développement de la méthode des moindres carrés, qu'il affirmait utiliser depuis 1795 (Birkes & Dodge, 1993).

Un modèle de régression linéaire est un modèle de régression d'une variable expliquée (cas de la régression linéaire simple) ou plusieurs variables explicatives (cas de régression linéaire multiple) dans lequel on fait l'hypothèse que la fonction qui relie les variables explicatives à la variable expliquée est linéaire dans ses paramètres.

Ce modèle de régression linéaire est bien utilisé pour chercher à prédire un phénomène que pour chercher à l'expliquer. Après avoir estimé un modèle de régression linéaire, on peut prédire quel serait le niveau de Y pour des valeurs particulières de X . Il permet également d'estimer l'effet d'une ou plusieurs variables sur une autre en contrôlant par un ensemble de facteurs par exemple, dans le domaine des sciences de l'éducation, on peut évaluer l'effet de la taille des classes sur les performances scolaires des enfants en contrôlant

par la catégorie socio - professionnelle des parents ou par l'emplacement géographique de l'établissement en apprentissage statistique.

Dans ce mémoire, qui s'articule autour de trois chapitres, on essaie d'étudier :

Chapitre 1 : Régression linéaire simple

On a rappelé les formules de la régression linéaire simple et les notions d'estimation des paramètres du modèle, l'estimation de l'intervalle de confiance, tester la signification des paramètres.

Chapitre 2 : Régression linéaire multiple

La régression linéaire multiple constitue une généralisation naturelle de la régression linéaire simple (le nombre des variables explicatives supérieure ou égal 2).

Chapitre 3 : Régression linéaire multiple avec **R**

Le dernier chapitre est réservé à l'application des modèles de régression linéaire multiple, traités théoriquement dans le chapitre 2.

On mentionne que tous les travaux, présentés dans ce mémoire, sont traités à l'aide du logiciel **R**.

Chapitre 1

Régression linéaire simple

Ce chapitre introduit la notion de modèle linéaire par la version la plus élémentaire : expliquer Y par une fonction affine de X . Après avoir expliciter les hypothèses nécessaires et les termes du modèle, les notions d'estimation des paramètres du modèle, de prévision par intervalle de confiance, la signification des tests d'hypothèse sont discutées.

1.1 Modèle de régression linéaire simple

1.1.1 Modélisation statistique

Définition 1.1.1 (modèle de régression linéaire simple) *:- Un Modèle de régression linéaire simple est défini par une équation de la forme :*

$$\forall i \in \{1, ..., n\} \qquad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

n	: *est le nombre d'obsarvation.*
x_i	: *est la variable explicative et indépendante (valeur fixée).*
y_i	: *est la variable expliquer et dépendante (valeur observé et aléatoire).*
β_0 *et* β_1	: *sont les paramètres inconnue du modèle (sont les coefficients).*

La quantité ε_i : *est une erreur (viennent du fait que les points ne sont jamais parfaitement alignés sur une droite).*

La distribution des erreurs : $E(\varepsilon_i) = 0, \qquad E(\varepsilon_i^2) = \sigma_\varepsilon^2, \qquad Cov(\varepsilon_i \varepsilon_j) = \delta_{ij}\sigma_\varepsilon^2.$

Les erreurs sont donc supposées centrées, de même variance (homoscédasticité) et nom corrélées entre elles (δ_{ij} *est le symbole de Kronecker, i.e* $\delta_{ij} = 1$ *si* $i = j$, $\delta ij = 0$ *si* $i \neq j$*).*

Le modèle est linéaire en x par rapport deux paramètres.

1.1.2 Modèle linéaire simple sous forme matricèlle

Notons que le modèle de régression linéaire simple de la définition peut encore s'écrire de façon vectorielle :

$$Y = \beta_0 \mathbf{1} + \beta_1 X + \varepsilon,$$

Où :

- Le vecteur $Y = [y_1, ..., y_n]^{'}$ est aléatoire de dimension n.
- Le vecteur $\mathbf{1} = [1, ..., 1]^{'}$ est le vecteur de $\mathbb{R}^n$ dont les n composante valent toutes $\mathbf{1}$.
- Le vecteur $X = [x_1, ...x_n]^{'}$ est un vecteur de dimension n donné (non aléatoire).
- Les coefficients $\beta_0$et β_1 sont les paramètres inconnus (mais non aléatoire) du modèle.
- Le vecteur $\varepsilon = [\varepsilon_1, ..., \varepsilon_n]^{'}$ est aléatoire de dimension n.

Cette notation vectorielle sera commode notamment pour l'interprétation géométrique du problème.

Exemple 1.1.1 *:- La concentration d'ozone* O_3 *dans l'aire (en microgrammes par millilitre). En particulier, on cherche à savoir s'il est possible d'expliquer le taux maximal d'ozone de la journée par la température* T_{12} *à midi.Les données sont :*

D'un point de vue pratique, le but de cette régression est double :

Température à12 h	23.8	16.3	72.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
O_3max	115.4	76.8	113.8	81.6	115.4	125	83.6	75.2	136.8	302.8

TAB. 1.1 – données journalières de température et d'ozone

Ajuster un modèle pour expliquer O_3 en fonction de T_{12} ; prédire les valeur d'O_3 pour de nouvelle valeur de T_{12}.

Avant toute analyse, il est intéressant de représenter les données, comme sur la figure 1

Pour analyser la relation entre les x_i (température) et les y_i(ozone) ,nous allons chercher une fonction f telle que :

$$y_i \approx f(x_i)$$

En code R :

tab=read.table("ozone-2.txt",header= T)

plot(O_3~T_{12},data=tab,xlab="T_{12}",ylab="O_3")

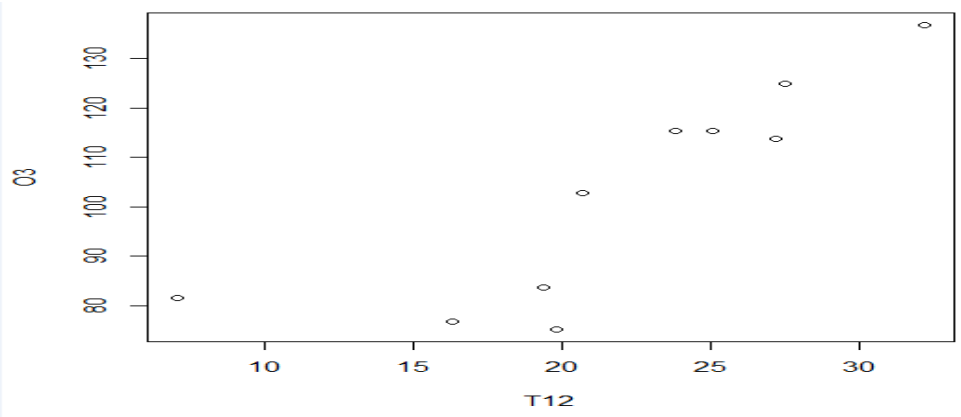

FIG. 1.1 – 10 données journalières de température et d'ozone.

1.2 Estimation

1.2.1 Méthode de Moinndres carrés Ordinaires

Définition 1.2.1 *(estimateurs des MCO) On appelle estimateurs des moindres carrés (MCO) de β_0 et β_1, les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ obtenus par minimisation de la quantité :*

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \|Y - \beta_0 \mathbf{1l} - \beta_1 X\|^2,$$

où $\mathbf{1l}$ est le vecteur de $\mathbb{R}^n$ dont tous les coefficients valent 1. Les estimateurs peuvent également s'écrire sous la forme suivante :

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0,\beta_1)\in\mathbb{R}\times\mathbb{R}}{\arg\min}\, S(\beta_0, \beta_1)$$

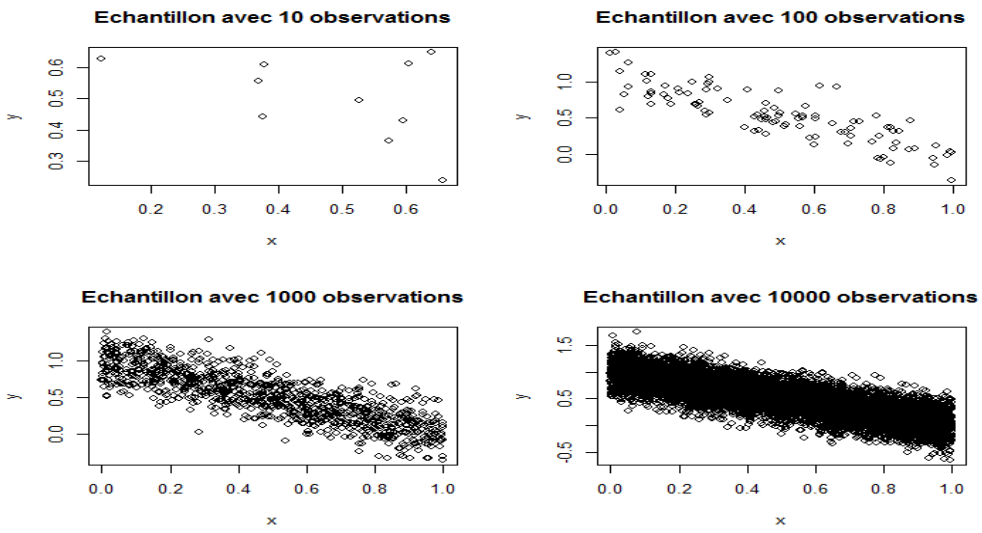

FIG. 1.2 – Exemple de différentes liaisons possibles entre x et y.

1.2.2 Calcul des estimateurs de β_0 et β_1

Les estimateurs du maximum de vraisemblance

La fonction de vraisemblance des trois paramètres, associée aux réalisations $y_1, y_2, ..., y_n$, est :

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[y_i - (\beta_0 + \beta_1 x_i)]^2\right\}.$$

D'où la log vraisemblance :

$$ln\ L(\beta_0, \beta_1, \sigma^2) = -n(\ln\sqrt{2\pi} + \frac{1}{2}\ln\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2.$$

En annulant les dérivées partielles par rapport β_0, β_1, on obtient les equations de vraisemblance :

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

Les deux équations étant linéaires de β_0 et β_1 peuvent être résolues.De la première équation on déduit

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

puis en remplaçant dans :

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

Où

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

et de même façon :

$$\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

D'où finalement, en substituant les Y_i aux y_i les estimateurs du MV de β_0 et β_1 :

$$\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \\ \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{cases}$$

Propriétés des estimateurs

Théorème 1.2.1 (*biais des estimateurs*) *:- $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 et β_1 c'est-à-dire que*

$$E\left(\hat{\beta}_0\right) = \beta_0 \quad et \quad E\left(\hat{\beta}_1\right) = \beta_1.$$

Preuve. On a :

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

dans cette expression, seuls les bruits ε_i sont aléatoires, et puisqu'ils sont centrés, on en déduit bien que :

$$E[\hat{\beta}_1] = \beta_1.$$

Pour $\hat{\beta}_0$ on part de l'expression : $\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$ d'où l'on tire :

$$E[\hat{\beta}_0] = E[\bar{y}] - \bar{x}E[\hat{\beta}_1] = \bar{y} - \bar{x}\beta_1 = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0.$$

■

Théorème 1.2.2 (*Variances de* $\hat{\beta}_0$ *et* $\hat{\beta}_1$) *:- Les variances et covariance des estimateurs des paramètres valent :*

$$V\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum x_i^2}{n \sum \left(x_i - \bar{x}\right)^2}, \qquad V\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum \left(x_i - \bar{x}\right)^2}$$

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\sigma^2 \bar{x}}{\sum \left(x_i - \bar{x}\right)^2}$$

Preuve. :- On part à nouveau de l'expression de $\hat{\beta}_1$ utilisée dans la preuve du non - biais :

$$\hat{\beta}_{1=}\beta_1 + \frac{\sum \left(x_i - \bar{x}\right) \varepsilon_i}{\sum \left(x_i - \bar{x}\right)^2}.$$

Or les ε_i sont décorrélées et de même variance σ^2donc la variance de la somme est la somme des variances :

$$Var\left(\hat{\beta}_1\right) = \frac{\sum \left(x_i - \bar{x}\right)^2 \sigma^2}{\left(\sum \left(x_i - \bar{x}\right)^2\right)^2} = \frac{\sigma^2}{\sum \left(x_i - \bar{x}\right)^2}.$$

Par ailleurs la covariances entre$\bar{y}$ et $\hat{\beta}_1$ s'ecrit :

$$Cov\left(\bar{y}, \hat{\beta}_1\right) = Cov\left(\sum \frac{y_i}{n}, \frac{\sum \left(x_i - \bar{x}\right) \varepsilon_i}{\sum \left(x_i - \bar{x}\right)^2}\right) = 0$$

d'où il vient pour la variance de $\hat{\beta}_1$:

$$Var\left(\hat{\beta}_0\right) = Var\left(\frac{\sum y_i}{n} - \hat{\beta}_1 \bar{x}\right) = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum \left(x_i - \bar{x}\right)^2} - 2\bar{x} Cov\left(\bar{y}, \hat{\beta}_1\right)$$

$$Var\left(\hat{\beta}_0\right) = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum \left(x_i - \bar{x}\right)^2} = \frac{\sigma^2 \sum x_i^2}{n \sum \left(x_i - \bar{x}\right)^2}$$

Enfin, pour la des deux estimateurs :

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = Cov\left(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1\right) = \ Cov\left(\bar{y}, \hat{\beta}_1\right) - \bar{x} Var(\hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{\sum \left(x_i - \bar{x}\right)^2}$$

■

Remarque 1.2.1 *A chaque valeur de x correspond la valeur estimée ou ajustée de y :*

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

1.3 Analyse des résidus

Résidus et variance résiduelle

Nous avons estimé β_0 et β_1.La variance σ^2 des ε_i est le dernier paramètre inconnu à estimer. Pour cela, nous allons utiliser les résidus, ce sont des estimateurs des erreurs iconnues .

Définition 1.3.1 *:- Les résidus sont définis par :*

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

Remarque 1.3.1 *:- Dans un modèle de régression linéaire simple, la somme des résidus est nulle.*

Proposition 1.3.1 (*Estimateur de la variance du bruit*) *:- La stastique*

$$S^2 = \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

est un estimateur sans biais de σ^2.

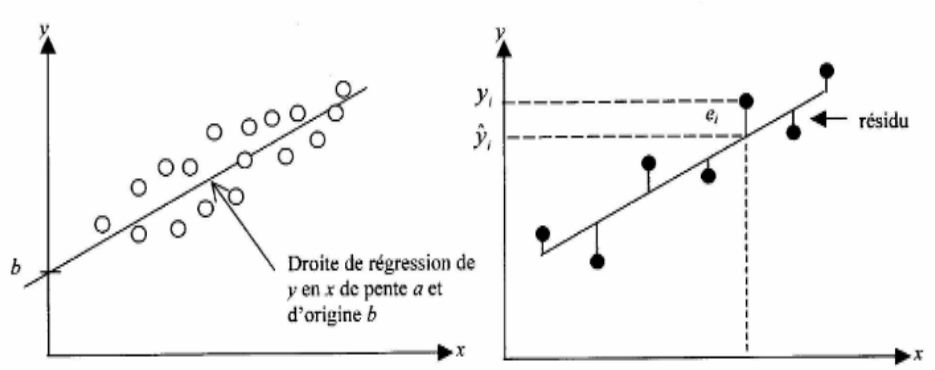

FIG. 1.3 – Droite et résidu de la régression linéaire.

1.4 Qualité d'ajustement

1.4.1 Coefficient de détermination R^2

Proprieté des moindres carées

$$\sum \left(Y_i - \bar{Y}\right)^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum \left(Y_i - \hat{Y}\right)^2$$

$$SC_{tot} = SC_{reg} + SC_{res}$$

Le coefficient de détermination est défini par :

$$R^2 = \frac{SC_{reg}}{SC_{tot}}$$

Intuitivement ce coefficient de détermination quantifie la capacité du modèle à expliquer les variations de Y.

Si R^2 est proche de 1 alors le modèle est proche de la réalité.

Si R^2 est proche de 0 alors le modèle expliquer très mal la réalité.

1.4.2 Lois des estimateurs

Les estimateurs $\hat{\beta}_0$et $\hat{\beta}_1$ sont des variables aléatoires réelles de matrice de covariance :

$$M = \sigma^2 \begin{pmatrix} \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2} & \frac{-\bar{x}}{(n-1)S_x^2} \\ \frac{-\bar{x}}{(n-1)S_x^2} & \frac{1}{(n-1)S_x^2} \end{pmatrix}$$

qui est estimée en remplaçant σ^2 par son estimation S^2. Sous l'hypothèse que les résidus sont gaussiens, on montre que :

$$\frac{(n-2)\,S^2}{\sigma^2} \sim X^2_{(n-2)}$$

et donc que les statistiques :

$$\frac{(\hat{\beta}_0 - \beta_0)}{S\left(\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\right)^{\frac{1}{2}}}$$

et

$$\frac{(\hat{\beta}_1 - \beta_1)}{S\left(\frac{1}{(n-1)S_x^2}\right)^{\frac{1}{2}}}$$

suivent des lois de student à $(n-2)$ degrés de liberté. Ceci permet de tester l'hypothèse de nullité d'un de ces paramètres.

1.4.3 Intervalles de confiance

Intervalles de confiance pour les cœfficients de régression

Théorème 1.4.1 *:- Soit $\alpha \in]0,1[$. On note $t_{n-2}(u)$ le u-quantile de la loi $T(n-2)$. Un intervalle de confiance de niveau de confiance$(1-\alpha)$ pour β_0 est donné par :*

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

Un intervalle de confiance de niveau de confiance $(1-\alpha)$*pour* β_1 *est donné par :*

$$\left[\hat{\beta}_1 - t_{n-2}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \hat{\beta}_1 + t_{n-2}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right]$$

Un intervalle de confiance de Y

Un intervalle de confiance de niveau de confiance $(1-\alpha)$pour $Y_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ est donné par :

$$\left[\hat{Y}_0 - t_{n-2}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \hat{Y}_0 + t_{n-2}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right]$$

1.4.4 Test sur les paramètres du modèle

Test de student de signification d'un coefficient β_j $(j=0,1)$

On souhate de tester :

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

qui est un test bilatérale de signification de β_j.

p-valeur : On considère des hypothèses de la forme :

$$H_0 : A \qquad \text{contre} \qquad H_1 : \text{ contraire de } A$$

La p-valeur est le plus petit réel $\alpha \in]0,1[$ calculé à partir des données tel que l'on puisse se permettre degrés de significativité : Le rejet de H_0 sera :

1. Significatif si p-valeur $\in]0.01, 0.05]$.
2. Très significatif si p-valeur $\in]0.001, 0.01]$.

3. Hautement significatif si p-valeur < 0.001.

4. presque significatif si p-valeur $\in]0.05, 0.1]$.

On considère une variable $T \sim T(n-2)$.

Alors la p-valeur associée est :

$$p - valeur = p\left(|T| \geq \left|t_{\beta_j}\right|\right)$$

Si

1. l'influence de X_j sur Y est significative.

2. l'influence de X_j sur Y est très significative.

3. l'influence de X_j sur Y est hautement significative.

ou :

Régle de décision au niveau α : on rejette H_0 si $t_{\beta_j} = \left|\dfrac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}\right| > T$

Exemple 1.4.1 *:- Pour un enquête on a eu recours à* 54 *enquêteurs. Pour chacun d'entre eux on dispose du nombre d'entretiens qu'il a effectués et de la durée médiane .On cherche à vérifier si le nombre d'entretiens effectués X est un facteur explicatif de la durée de l'entretien Y. On calculé initialement :*

$$\bar{x} = 53, \quad \bar{y} = 30.535, \quad \sum_{i=1}^{54} x_i^2 = 4274.8, \quad \sum_{i=1}^{54} y_i^2 = 957.32, \quad \sum_{i=1}^{54} x_i y_j = 1531.7$$

On déduit les estimation suivantes :

$$\hat{\beta}_0 = 33.668, \quad \hat{\beta}_1 = -0.05911, \quad \hat{\sigma}^2 = 20.437$$

$$\hat{\sigma}^2_{\hat{\beta}_0} = 1.1057, \quad \hat{\sigma}^2_{\beta_1} = 0.0002589, \quad Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -0.01371.$$

Pour tester $H_0 : \beta_1 = 0$

La statistique de test prend la valeur $t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2_{\beta_1}}} = -3.68$, qui suit la loi de student $t(52)$, à p-valeur de l'ordre de 0.001.

Alors X est un facteur explicatif trés significatif.Pour $x = 50$ entretiens on une estimation de la durée médiane des entetients égale à :

$$\hat{\beta}_0 + \hat{\beta}_1.50 = 30.713.$$

et un intrvalle de confiance de niveau 0.95 associé :

$$IC^{(52)}_{0.975}\left(\beta_0 + \beta_1 50\right) = \left[30.713 \begin{array}{c} + \\ - \end{array} t^{(52)}_{0.975}.0.618\right]$$

$t^{(52)}_{0.975} = 2.007$, l'intervalle est : [29.47, 31.95].

Chapitre 2

Régression linéaire multiple

Dans ce chapitre nous généralisons et étendons les résultats précédents au cas plus intéressant où l'on cherche à expliquer une variable Y par un ensemble de variables X. De façon à simplifier la notation, on utilisera la notation matricielle. Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple.

2.1 Modèle de régression linéaire multiple.

2.1.1 Ecriture du modèle

Le modèle de régression linéaire multiple s'écrit :

$p = 2$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad i = 1 : n$$

$p > 2$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \varepsilon_i \qquad i = 1 : n$$

– X_{ij} est le j-ème variable explicative (indépendante) pour l'individu i $(j = 1 : p)$.

– ε_i est indépendante de X et suit une loi normale $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$.

– $\beta_0, \beta_1 ... \beta_p$ sont les paramètres inconnus du modèle.

2.1.2 Le modèle sous forme matricielle

Ce modèle peuvent s'écrire sous forme matricielle :

$$Y = XB + \varepsilon$$

telle que :

$$Y = \begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{pmatrix}; \qquad X = \begin{pmatrix} 1 & X_{11} & X_{12} & ... & X_{1p} \\ . & . & & & . \\ . & . & & & . \\ . & . & & & . \\ 1 & X_{n1} & X_{n2} & ... & X_{np} \end{pmatrix}$$

$$B = \begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_p \end{pmatrix}; \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

Exemple 2.1.1 *:- Dans une maternité, on a recensé les données suivantes sur plusieurs femmes qui viennent d'accoucher : poids, âge et temps d'aménorrhée. On cherche à savoir s'il y a une relation entre le poids du bébé à la naissance et chacune de ces variables.*

Le modèle est :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

telle que : Y poids du bébé, X_1 poid de la mère, X_2 âge de la mère, X_3 temps d'aménorrhée.

2.2 Estimation

2.2.1 Estimateurs des moindres carrés

Définition 2.2.1 *(Estimateur des MCO) :- On appelle estimateur des moindres carrés (noté MCO) $\hat{\beta}$ de β la valeur suivante*

$$\hat{\beta} = \underset{\beta_1,\ldots\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(Y - X\beta\right)^{'} \left(Y - X\beta\right).$$

Théorème 2.2.1 *(Expression de l'estimateur des MCO) :- Si l'hypothèse H_1 est vérifiée, l'estimateur des MCO $\hat{\beta}$ de β vaut :*

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

2.2.2 Calcul des estimateurs de B

$$B = (\beta_0, \ldots \beta_p)^t \qquad B \in \mathbb{R}^p$$

Les estimateurs du maximum de vraisemblance

La log-vraisemblance s'écrit comme en chapitre 1 en remplaçant $\beta_0 + \beta_1 x_i$ par $\beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{p1}$. Les équations de vraisemblance obtenues en d´érivant par rapport à chacun des $p+1$ paramètres forment un système linéaire de $p+1$ équations dont l'´ecriture matricielle est $(X^t X)B = X^t Y$.

On suppose que les vecteurs colonnes de X sont linéairement indépendants (i.e. $n > p$ et pas de redondance d'information dans les vecteurs prédicteurs ni de Combinaison linéaire entre eux donnant un vecteur constant, ce qui signifierait une sur paramétrisation du modèle). Alors la matrice X^tX est inversible et on a la solution unique :

$$\hat{B} = (X^tX)^{-1}X^tY$$

Remarque 2.2.1 *:- $\hat{B}$ est également le vecteur des estimateurs des moindres carrées, c'est à dire :*

$$\|Y - XB\|^2 = \min_B \|Y - XB\|^2$$

$\|.\|^2$ *représente la norme euclidienne usuelles d'un vecteurs de* $\mathbb{R}^n$.

Par la méthode du maximum de vraisemblance, on obtient l'estimateur de σ_ε^2 *:*

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{i=n} \varepsilon_i^2}{n - p - 1}$$

Propriétés des estimateurs

– L'estimateur $\hat{B}$ est le meilleur estimateur non-biaisé de B :

$$E\left(\hat{B}\right) = B$$

Preuve. :- En exprimer $\hat{B}$ en fonction de B :

$$\hat{B} = (X^tX)^{-1}X^tY$$
$$\hat{B} = (X^tX)^{-1}X^t(XB + \varepsilon)$$
$$\hat{B} = B + (X^tX)^{-1}X^t\varepsilon$$

Alors :

$$E\left[\hat{B}\right] = B + E\left[(X^tX)^{-1}X^t\varepsilon\right]$$
$$E\left[\hat{B}\right] = B + (X^tX)^{-1}X^tE\left[\varepsilon\right]$$
$$E\left[\hat{B}\right] = B$$

■

– La variance de $\hat{B}$ est :

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

Preuve.

$$\hat{B} - B = (X^tX)^{-1}X^t\varepsilon$$

$$E\left[\left(\hat{B} - B\right)\left(\hat{B} - B\right)^t\right] = (X^tX)^{-1}X^tE\left[\varepsilon\varepsilon^t\right]X(X^tX)^{-1}$$

$$E\left[\left(\hat{B} - B\right)\left(\hat{B} - B\right)^t\right] = \sigma_\varepsilon^2(X^tX)^{-1}$$

Car :

$$E\left[\varepsilon\varepsilon^t\right] = Var\left[\varepsilon\varepsilon^t\right] = \sigma_\varepsilon^2 1_n$$

Alors :

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

■

– L'estimateur de la variance résiduelle est donné par :

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{i=n}\varepsilon_i^2}{n-p-1}$$

Preuve. On a :

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

Alors :

$$\hat{\sigma}_{\hat{B}} = \hat{\sigma}_{\varepsilon}^2 (X^t X)^{-1}$$

et on a :

$$\hat{\varepsilon} = Y - \hat{Y} = (XB + \varepsilon) - X\hat{B}$$

$$= (XB + \varepsilon) - X(B + (X^t X)^{-1} X^t \varepsilon)$$

$$= (I - X(X^t X)^{-1} X^t)\varepsilon$$

On pose

$$(I - X(X^t X)^{-1} X^t) = A$$

telle que : A est symétrique $(A^t = A)$ et idempotente $(A^2 = A)$, de taille (n, n).

Alors :

$$\hat{\varepsilon}^t \hat{\varepsilon} = \varepsilon^t A \varepsilon$$

$$E\left[\hat{\varepsilon}^t \hat{\varepsilon}\right] = \sigma_{\varepsilon}^2 tr(A)$$

Car $tr(A)$ degrés de liberté est égal $n - (p + 1) = n - p - 1$.

$$\hat{\sigma}_{\varepsilon}^2 = \frac{\hat{\varepsilon}^t \hat{\varepsilon}}{tr(A)} = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-p-1}$$

■

2.3 Qualité d'ajustement

2.3.1 Coefficient de détermination R^2

Coefficient de détermination R^2

On appelle coefficient de détermination R^2 de :

$$R^2 = 1 - \frac{\left\|\hat{Y} - Y\right\|^2}{\left\|\bar{Y}_{1_n} - Y\right\|^2}$$

où

$$\hat{Y} = X\hat{B}, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

et 1_n désigne le vecteur colonne à n composantes égales à 1. Ce R^2 est un coefficient réel toujours compris entre 0 et 1.

Il mesure de la qualité du modèle de rlm ; plus R^2 est proche de 1, meilleur est le modèle.

Comme le R^2 dépend fortement de p, on ne peut pas l'utiliser pour comparer la qualité de 2 modèles de rlm qui diffèrent quant au nombre de variables explicatives. C'est pourquoi on lui préfère sa version ajustée présentée ci-dessous.

Coefficient de détermination R^2 ajusté

On appelle coefficient de détermination ajusté $\bar{R}^2$ de :

$$\bar{R}^2 = 1 - \frac{\left\|\hat{Y} - Y\right\|^2 / (n - (p+1))}{\left\|\bar{Y}_{1_n} - Y\right\|^2 / (n-1)} = 1 - \frac{(n-1)}{(n-(p+1))}\left(1 - R^2\right)$$

2.3.2 Lois des estimateurs

Loi de $\hat{\beta}_j$: pour tout $j \in \{0...p\}$, en notant $\left[(X^tX)^{-1}\right]_{j+1,j+1}$ la $j+1$-ème composante diagonale de $(X^tX)^{-1}$, on a :

$$\hat{\beta}_j \backsim N(\beta_j, \sigma^2 \left[\left(X^tX\right)^{-1}\right]_{j+1,j+1})] \qquad \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{\left[(X^tX)^{-1}\right]_{j+1,j+1}}} \backsim N(0,1).$$

Degrés de liberté : Dans ce qui suit, on travaillera avec le nombre de degrés de liberté :

$$\vartheta = n - (p+1)$$

Loi associée à $\hat{\sigma}^2$:

$$n - (p+1)\frac{\hat{\sigma}^2}{\sigma^2} \backsim X^2(\vartheta)$$

Apparition de la loi de Student : Pour tout $j \in \{0...p\}$, en posant :

$$\hat{\sigma}(\hat{\beta}_j) = \hat{\sigma}\sqrt{\left[(X^tX)^{-1}\right]_{j+1,j+1})]}$$

On a :

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}(\hat{\beta}_j)} \backsim T(\vartheta)$$

2.3.3 Intervalles de confiance

Intervalles de confiance pour β_j

Intervalles de confiance pour β_j : Pour tout $j \in \{0...p\}$, un intervalle de confiance pour β_j au niveau $(1-\alpha)$, $\alpha \in]0,1[$ est :

$$I_{\beta_j} = \left[\hat{\beta}_j - t_\alpha(\vartheta)\hat{\sigma}\sqrt{\left[(X^tX)^{-1}\right]_{j+1,j+1})}], \hat{\beta}_j + t_\alpha(\vartheta)\hat{\sigma}\sqrt{\left[(X^tX)^{-1}\right]_{j+1,j+1})}]\right]$$

où $t_\alpha(\vartheta)$ est le réel vérifiant :

$$P(|T| \geq t_\alpha(\vartheta)) = \alpha$$

avec $T \backsim T(\vartheta)$

Intervalles de confiance pour y_x

Soient y_x la valeur moyenne de Y quand $(X_1,...,X_p) = (x_1,...,x_p) = x$ et $x_\cdot = (1, x_1, ..., x_p)$.

Un intervalle de confiance pour y_x au niveau $(1-\alpha)$, $\alpha \in]0;1[$ est :

$$I_{y_x} = \left[\hat{y}_x - t_\alpha(\vartheta)\hat{\sigma}\sqrt{x_\cdot \left(X^t X\right)^{-1} x_\cdot^t}, \hat{y}_x + t_\alpha(\vartheta)\hat{\sigma}\sqrt{x_\cdot \left(X^t X\right)^{-1} x_\cdot^t}\right]$$

2.3.4 Test sur les paramètres du modèle

Test de student :

Soit $j \in \{0...p\}$.Le test de Student permet d'évaluer l'influence de X_j sur Y.

On considère les hypothèses :

$$H_0 : \beta_j = 0 \qquad contre \qquad H_1 : \beta_j \neq 0$$

On calcule T_{β_j} :

$$T_{\beta_j} = \frac{|\beta_j|}{\hat{\sigma}_{\beta_j}}$$

On considère une variable $T \sim T(\vartheta)$.

Alors la p-valeur associée est :

$$p - valeur = p\left(|T| \geq \left|T_{\beta_j}\right|\right)$$

La règele de décision est comme la chapitre 1.

Exemple 2.3.1 *:- On s'intéresse à l'effet du nombre d'années d'études des parents (M : mère et P : père) sur le nombre d'années d'études de leurs enfants noté Y .On dispose du nombre d'années d'études de* 26 *familles (enfant, mère et père). On décide d'ajuster ces données par le modèle linéaire suivant :*

$$Y_i = \beta_1 M_i + \beta_2 P_i + \varepsilon_i \qquad ;$$

telle que :

$$\varepsilon \backsim N(0, \sigma_\varepsilon^2) \qquad i = 1 : 26$$

On vous donne les sommes suivantes :

$$\sum_{i=1}^{26} M_i^2 = 288, \qquad \sum_{i=1}^{26} P_i^2 = 202, \qquad \sum_{i=1}^{26} M_i P_i = 144, \qquad \sum_{i=1}^{26} Y_i M_i = 184$$

$$\sum_{i=1}^{26} Y_i P_i = 158$$

Le modèle sous forme matricielle :

$$Y = XB + \varepsilon$$

$$\begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_{26} \end{pmatrix} = \begin{pmatrix} M_1 & P_1 \\ . & . \\ . & . \\ . & . \\ M_{26} & P_{26} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \\ \\ \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_{26} \end{pmatrix}$$

et ona

$$\left(X^tX\right) = \begin{pmatrix} \sum_{i=1}^{26} M_i^2 & \sum_{i=1}^{26} M_iP_i \\ \\ \sum_{i=1}^{26} M_iP_i & \sum_{i=1}^{26} P_i^2 \end{pmatrix} = \begin{pmatrix} 288 & 144 \\ \\ 144 & 202 \end{pmatrix}$$

$$X^tY = \begin{pmatrix} \sum_{i=1}^{26} Y_iM_i \\ \\ \sum_{i=1}^{26} Y_iP_i \end{pmatrix} = \begin{pmatrix} 184 \\ \\ 184 \end{pmatrix}$$

L'estimateur $\hat{B}$ *de* B *:*

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

$$\left(X^tX\right)^{-1} = \frac{1}{37440} \begin{pmatrix} 202 & -144 \\ -144 & 288 \end{pmatrix}$$

$$\hat{B} = \begin{pmatrix} 0.39 \\ 0.51 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

L'estimateur $\hat{\sigma}_\varepsilon^2$ *de* σ_ε^2 *sachant que* $\sum_{i=1}^{26} \hat{\varepsilon}_i = 36$ *:*

$p = 2$

$$\hat{\sigma}_\varepsilon = \frac{1}{n-p} \sum_{i=1}^{26} \hat{\varepsilon}_i = \frac{36}{24} = 1.5$$

Les variances estimées de chacun des coefficients $\hat{\sigma}^2_{\beta_1}$ *et* $\hat{\sigma}^2_{\beta_2}$ *:*

$$Var(\hat{B}) = \hat{\sigma}^2_{\varepsilon}\left(X^t X\right)^{-1} = \begin{pmatrix} 0.008 & -0.0057 \\ -0.0057 & 0.01 \end{pmatrix}$$

$$\hat{\sigma}^2_{\beta_1} = 0.008 \qquad \hat{\sigma}^2_{\beta_2} = 0.01$$

On teste la signification des paramètres β_1 *et* β_2 *à* 90 *:*

$$\begin{cases} H_0 : \beta_j = 0 \qquad j = 1, 2 \\ \\ H_1 : \beta_j \neq 0 \qquad j = 1, 2 \end{cases}$$

$$T_{\beta_1} = 4.36; \qquad T_{\beta_2} = 5.1; \qquad t_{0.95}(24)$$

$$T_{\beta_j} > t_{0.95}(24) \qquad j = 1, 2$$

Alors On rejet H_0 *et on accepte* H_1

Un intervalle de confiance des paramètres β_1 *et* β_2 *à* 90 *.*

Pour β_1 *:*

$$\beta_1 \in \left[\hat{\beta}_1 - \hat{\sigma}_{\beta_1} t_{0.95}(24), \hat{\beta}_1 + \hat{\sigma}_{\beta_1} t_{0.95}(24)\right]$$

$$\beta_1 \in [0.236, 0.540]$$

Pour β_2 *:*

$$\beta_2 \in \left[\hat{\beta}_2 - \hat{\sigma}_{\beta_2} t_{0.95}(24), \hat{\beta}_2 + \hat{\sigma}_{\beta_2} t_{0.95}(24)\right]$$

$$\beta_2 \in [0.33, 0.688]$$

La variable M a une contribution dans l'explication de Y ainsi que le nombre d'étude des mères a un effet sur ce lui de leur enfant car $\beta_1 \neq 0$.

Et le même pour le nombre d'année d'étude des pères car $\beta_2 \neq 0$.

Chapitre 3

Régression linéaire multiple avec R

R est un logiciel libre et gratuit et un langage de traitement statistique qui a notamment été présenté ici. Le langage R (R Développent Core Team, 2013) est dit orienté objet comme Python ou Ruby. Un des avantages de R réside dans la possibilité de communiquer des scripts par écrit, car le plus souvent on l'utilise en mode console. Cela évite de se soucier des problèmes techniques produits par la variété de versions des systèmes d'exploitation et d'avoir recours à des copies d'écrans, mais permet surtout de communiquer des calculs et des analyses statistiques en quelques lignes de texte.

3.1 Initiation à la régression linéaire avec R

3.1.1 Génération d'une variable aléatoire

Méthode de la fonction inverse :

On veut simuler une variable aléatoire continue X de fonction de répartition F.

Théorème 3.1.1 *: Soit X une variable aléatoire de fonction de répartition F strictement croissante, on a*

$$F(X) \backsim U_{[0;1]}$$

Preuve : On pose

$$u = F(x) \Longleftrightarrow x = F^{-1}(u),$$

par définition, on a

$$F(x) = P(X \leq x)$$

Donc $F(F^{-1}(u))$ par définition de la réciproque et

$$P\left\{X \leq F^{-1}(u)\right\} = P\left\{F(X) \leq u\right\}$$

car F est strictement croissante. On a donc

$$u = P\left\{F(X) \leq u\right\}$$

et on reconnait la fonction de répartition de la *loi uniforme*.

Méthode : si on connait la fonction de répartition F^{-1}, réciproque de F il suffit de tirer

$$X = F^{-1}(U)$$

3.1.2 Génération d'un vecteur aléatoire

Soit $X = (X_1, ..., X_n)$ ou X_i est une v.a d'un loi $F_i(x) = P(X_i \leq x)$ loi marginale et

$$F(x_1, ..., x_n) = P(X_1 \leq x, ..., X_n \leq x_n),$$

loi conjointe de vecteur X.

- Si $X_1, ..., X_n$ sont indépendants,

$$F(x_1, ..., x_n) = F_{1(x_1)}.F_{2(x_2)}....F_{n(x_n)}.$$

On a

$$f_{i(x_i)} = \frac{dF\left(x_i\right)}{dx_i}, \qquad F\left(x\right) = \frac{dF^n\left(x_i\right)}{dx_1 ... dx_n};$$

et la densité conditionnelle ;

$$f_{X/Y}\left(X/Y\right) = \frac{f_{X,Y}\left(x,y\right)}{f_Y\left(y\right)}$$

Donc

$$f_{X,Y}\left(x,y\right) = f_Y\left(y\right).f_{X/Y}\left(x/y\right).$$

Ainsi

$$F_{X/Y}\left(x/y\right) = \int_{-\infty}^{x} f_{X/Y}\left(x/y\right) dx$$

Théorème 3.1.2 *Soit* $X = (X_1, ..., X_n)$ *un vecteur aléatoire de loi conjointe* F. *Soit* $U_1, ..., U_n$ *une suite de v.a indépendants uniforme sur* $[0, 1]$.

$$\begin{cases} F_{X_1}\left(x_1\right) = U_1 \\ F_{X2/X_1}\left(x_2/x_1\right) = U_2 \\ . \\ . \\ F_{X_2/X_1,.....,X_{n-1}}\left(x_n/x_1, ..., x_{n-1}\right) = U_n \end{cases}$$

3.1.3 Les fonctions simples dans R

Les fonctions de R étant nombreuses, nous 'en citons que quelques-unes qui peuvent s'avérer utiles.

Fonction	Description
sum(x), prod(x)	Sommes et produit des composantes de x.
max(x), min(x)	Maximum et minimum des composantes de x.
which.max(x)	Retourne l'indice du maximum des composantes de x.
range(x)	Idem que $c(min(x), max(x))$.
length(x)	Nombre d'éléments dans x.
mean(x), median(x)	La moyenne et la médiane des éléments dans x.
var(x), cov(x,y)	La variance des éléments dans x, la covariance entre x et y.
t(),diag()	La transposée, la diagonale.
det(), solve()	Le déterminant , l'inverse d'une matrice.
round(x,n)	Arrondi les éléments de x à n chiffres aprés la virgule.
rank	Rangs des éléments de x.
sort(x)	Trie les éléments de x dans ordre croissante.
cumsum(x)	Idem pour le somme.
cumprod(x)	Idem pour le produit.
rev(x)	Inverse l'ordre de x.
table(x)	Retourne un tableau avec les effectifs de différentes valeurs de x.
cbind	Regrouper toutes les variables explicatives sous forme d'une matrice.
lm	Estimation du modèle linéaire.
read.table	A privilégier pour des jeux de données organisés sous la forme de tableaux, comme cela est souvent le cas en statistique.
summary	Description des résultats du modèle.
confint	Intervalle de confiance des paramètres de régression.

TAB. 3.1 – Les principales fonctions à utiliser afin d'effectuer une régression linéaire.

Les graphiques dans R

Possibilité de voir des exemples de graphiques avec demo(graphics) ou demo(persp). Lorsqu'une fonction graphique est tapée sur la console, une fenêtre graphique va s'ouvrir avec le graphe demandé.

Voici un aperçu des fonctions graphiques principales de R.

$plot(x)$	Graphe des valeurs de x en fonction des valeurs de x.
$plot(x, y)$	Graphe des valeurs de y en fonction des valeurs de x.
$boxplot(x)$	Boxplot de x.
$hist(x)$	Histogramme de x (pour x quantitative).
$borplot(x)$	Diagramme en colonnes (pour x qualitative) pour chaque
	fonction, on aplusieurs options mais certaines sont communes.
$type$	"p" points, "l" : lignes, "b" les deux, "h" : lignes verticales.
$xlab, ylab$	Noms des axes, variables caractères entre "".
$main$	Title, variable de type caractère sub : sous-titre.
$points(x, y)$	Ajoute des points.
$Ilines(x, y)$	Idem mais avec des lignes.
$abline(a, b)$	Trace une ligne de pente b et ordonnée à l'origine a.
$legend(x, y, legend)$	Ajoute une légende au point de coordonnées (x, y), avec les symboles donnés par légende.

TAB. 3.2 – Fonctions graphiques dans R

3.1.4 Les lois de probabilité

Loi	*Fonction*
Gauss (normale)	*rnorm(n, mean=0, sd=1)*
Exponentielle	*rexp(n, rate=1)*
Gamma	*rgamma(n, shape, scale=1)*
Poisson	*rpois(n, lambda)*
Cauchy	*rcauchy(n, location=0, scale=1)*
Binomiale	*rbinom(n, size, prob)*
Uniforme	*runif(n, min=0, max=1)*
Beta	*rbeta(n, shape1, shape2)*

TAB. 3.3 – Lois usuelles..

d : correspond a la densité de probabilité.

p : correspond a la fonction de répartition (valeur inverse du quantile).

q : correspond au quantile.

r : correspond a la génération aléatoire de nombre.

Le premier argument est 'x' (une valeur) pour dnom loi, 'q' (un quantile) pour pnom loi, 'p' (une probabilité) pour qnom loi et 'n' (un nombre) pour rnom loi.

3.1.5 Le modèle linéaire sous R

La fonction lm

Pour « linear model » : s'utilise avec le concept de formule :

– 1 variable explicative avec intercept (régression linéaire)

$$lm(y\tilde{}X+1),$$

– 2 variables explicatives sans intercept

$$lm(y\tilde{}X1+X2-1).$$

3.2 Partie pratique

Exemple 3.2.1 $X_1 = runif(20, min = -2, max = 2)$

$X_2 = runif(20, min = -2, max = 2)$

$X_3 = rexp(20,1)$

$X_4 = rexp(20,1)$

$M = cbind(X_1, X_2, X_3, X_4)$

$Z = 0.2 * X_1 + 0.4 * X_2 + 0.3 * X_3 + 0.5 * X_4 + 1.5 + rnorm(20, 0, 1)$

$modèle = lm(Z\tilde{}M)$

```
Call:
lm(formula = Z ~ M)

Coefficients:
(Intercept)          MX1          MX2          MX3          MX4
     1.0789       0.3668       0.6613       0.3262       0.5681
```

summary(modèle)

```
Call:
lm(formula = Z ~ M)

Residuals:
     Min       1Q   Median       3Q      Max
-1.14796 -0.43698  0.00976  0.43367  1.55996

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8966     0.3548   5.345 8.17e-05
MX1           0.1360     0.1608   0.846   0.4110
MX2           0.4070     0.1583   2.572   0.0213
MX3           0.1531     0.2451   0.625   0.5415
MX4           0.4524     0.1617   2.798   0.0135
```

confint(modèle,level= 0.99)

```
                 0.5 %    99.5 %
(Intercept)  0.15450330 2.843381
MX1         -0.09162643 1.107098
MX2          0.20054837 1.579346
MX3         -0.73292163 1.005333
MX4         -0.32474638 1.254587
```

pairs(M)

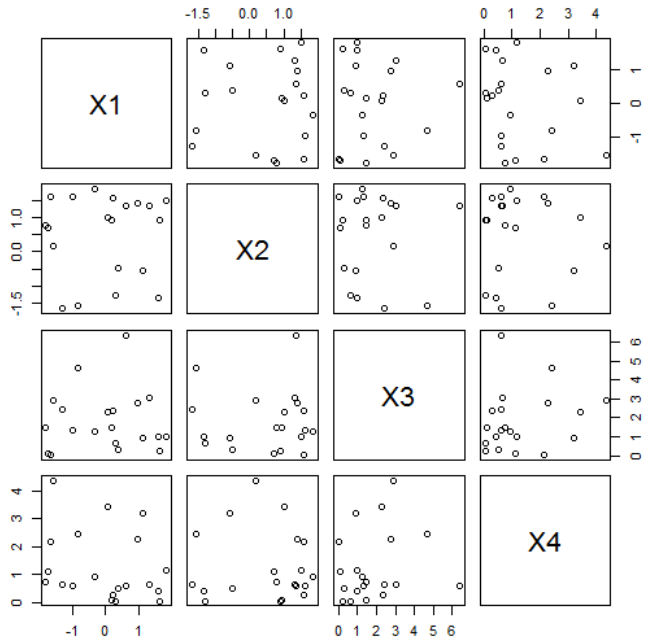

FIG. 3.1 – Matrice des nuages de points.

Exemple 3.2.2 *L'entreprise CITRON fabrique un matériau en matière plastique qui est utilisé dans la fabrication de jouets. Le département de contrôle de qualité de l'entreprise a effectué une étude qui a pour but d'établir dans quelle mesure la résistance à la rupture (en kg/cm2) de cette matière plastique pouvait être affectée par l'épaisseur du matériau ainsi que la densité de ce matériau. Douze essais ont été effectués et les résultats sont présentés dans le tableau ci-dessous :*

Essai numéro	Résistance à la rupture Y	Epaisseur du matériau X_1	Densité X_2
1	37.8	4	4
2	22.5	4	3.6
3	17.1	3	3.1
4	10.8	2	3.2
5	7.2	1	3
6	42.3	6	3.8
7	30.2	4	3.8
8	19.4	4	2.9
9	14.8	1	3.8
10	9.5	1	2.8
11	32.4	3	3.4
12	21.6	4	2.8

TAB. 3.4 – 12 données de (Résistance à la rupture,Epaisseur du matériau,Densité)de la matière plastique

Données=read.table("Données.txt",header=T)

RégMulti=lm(Y˜X$_1$ *+ X*$_2$*,data=Données)*

```
Call:
lm(formula = Y ~ X1 + X2, data = Données)

Coefficients:
(Intercept)           X1           X2
    -30.081        4.905       11.072
```

summary(RégMulti)

```
Call:
lm(formula = Y ~ X1 + X2, data = Données)

Residuals:
   Min     1Q Median     3Q    Max
-6.897 -2.135 -1.126  1.714 10.122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -30.081     11.455  -2.626 0.027542
X1             4.905      1.014   4.838 0.000923
X2            11.072      3.621   3.058 0.013617
```

confint(RégMulti,level= 0.99)

```
                   0.5 %    99.5 %
(Intercept) -67.3090320  7.147188
X1            1.6101788  8.199246
X2           -0.6948777 22.838992
```

Conclusion

La régression linéaire multiple traitent des données quantitatives. C'est une méthode d'investigation sur données d'observations, ou d'expérimentations, où l'objectif principal est de rechercher une liaison linéaire entre une variable Y quantitative et une ou plusieurs variables X également quantitatives.

Dans ce mémoire, on explique les hypothèses nécessaires et les termes du modèle, les notions d'estimation des paramètres du modèle, il existe de nombreuses méthodes pour estimer ce modèle, telle que dans ce travaille en utilise la méthode de maximum de vraisemblance, avec les tests de signification du paramètres du modèle et l'estimation des intervalles de confiances.

Bibliographie

[1] André Cornillon,P Lber. Eric Matzner. (2007). Régression :Théorie et application.Springer, 317 p., Statistique et probabilités appliquées, 978-2287396922

[2] Bernard G, Catherine P (2011). Introduction à la méthode statistique. ISBN 978-2-10-055892-6. Dunod, Paris.

[3] Christophe Chesneau. 9 Jan 2017. Modèles de régression. Université de Caen.

[4] Cornillon, Pierre-André.and Eric Matzner-Lober (2010).Régression avec R. Springer Science & Business Media.

[5] Dodge Y, & Rousson V. (2004).Analyse de régression appliquée. Dunod.

[6] Frédéric Bertrand et Myriam Maumy[1]. Master 1ère Année 23-03-2009, Régression linéaire multiple. [1]IRMA, Université Louis Pasteur Strasbourg, France.

[7] Guyader, A.(2011).Régression linéaire.Université Rennes,2,60-61.

[8] Lafaye de Micheaux, P. Drouilhet,R. Liquet, B. (2011). Le logiciel R : Maitriser le langage-Effectuer des analyses statistiques. 978-2-8178-0114-8. Springer-Verlag Paris.

[9] Laurent Ferrara, Mars 2013. Modèle de régression linéaire multivarié.

[10] Lejeune, M (2010). Statistique La théorie et ses applications. Springer-Verlag Paris.

[11] M. BONNEU, E. LECONTE. Modèle linéaire. Université des Sciences Sociales. Place Anatole France.

[12] Ricco Racotomalala. Régression linéaire multiple Université Lumière Lyon 2.

[13] Tenenhaus,M. (1998).La régression PLS : théorie et pratique. Technip.

Notations

$\mu,\ \overline{x}$	La moyenne.
$\sigma^2,\ \ Var(X)$	La variance.
IC	Intervale de confiance.
$\hat{\sigma}_{\varepsilon}$	Estimation de la variance et degrés de liberté.
Cov	Covariance.
$\mathcal{N}(\mu, \sigma^2)$	Loi normale d'espérance $\mu \in \mathbb{R}$ et de variance $\sigma^2\ \in \mathbb{R}$.
$T(\vartheta)$	Loi de Student.
$\mathcal{X}^2(n)$	Loi du chi-deux à $n \in \mathbb{N}^{\circ}$ degrés de liberté.
Y	Est un matrice de taille $n\,(p+1)$ comme appelée matrice du plan d'expérince.
RLM	Régression Linéaire Multiple
β	Est le vecteur de dimension $p+1$.des paramètres.
MCO	Estimateurs des Moindres Carrés Ordinaires.
MV	Maximum De Vraisemblance.
R^2	Le coefficient de détermination.
ε	L'erreure.
$t_{n-p,1-\frac{\alpha}{2}}$	Quantile d'order $\left(1-\frac{\alpha}{2}\right)$ de la loi de student à $(n-p)$ ddl
SC_{tot}	Somme des carrés totale.
SC_{reg}	Somme des carrés régression.
SC_{res}	Somme des carrés résiduelle.