République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE KASDI MERBAH OUARGLA

———————————-

Faculté des Mathématiques et des Sciences de la Matière

N° d'ordre :
N° de série :

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire Présenté En Vue De L'obtention Du

DIPLOME DE MASTER

Spécialité : Mathématiques

Option : Probabilité et Statistique

Présenté par:

BERKANI Fatiha

Application de la Régression Linéaire Multiples sur la Balance Commerciale Algériènne

Soutenu publiquement le : mai 2016

Devant le jury composé de:

Mr. AMARA Abdelkader	Université de KASDI Merbah - Ouargla	Président
Melle. ARBIA Hanane	Université de KASDI Merbah - Ouargla	Examinateur
Melle. SAIDANE Hadda	Université de KASDI Merbah - Ouargla	Rapporteur

DEDICACE

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, reçois à travers ce travail aussi modeste soit-il, l'expression de mes sentiments.

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

Mon mari, mes frères et mon fils aymen qui n'ont cessé d'être pour moi de courage et de générosité.

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 vont à Melle. Saidane Hadda 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 Monsieur AMARA Abdelkader 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. Arbia Hanane ait accepté d'être d'examiner ce travail ainsi que pour l'attention qu'elle a porté à mon travail malgré son emploi du temps très chargé.

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, mon mari, mes frères et mon fils aymen. Je n'aurais jamais pu arriver ici, sans l'équilibre, la chaleur, le soutien et le bonheur dans lequel j'ai vécu. Merci !

Table des matières

Table des figures

Liste des tableaux

Introduction

En statistiques, en économétrie et en apprentissage automatique, un modèle de régression linéaire est un modèle de régression d'une variable expliquée sur une (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. On parle aussi de modèle linéaire ou de modèle de régression linéaire.

Donc, la régression linéaire est l'un des modèles statistiques les plus employés : son champ d'application s'étend de la description et de l'analyse des données expérimentales jusqu'à la prévision et il est aussi utilisé pour l'interpolation. Il est par conséquent indispensable que le praticien possède une solide connaissance des prérequis, de la portée et des limites du modèle linéaire.

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 (voir [6], [5], [23], [30]) ou plusieurs variables ([30], [14], [16], [19]) 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, la méthode de régression linéaire est considérée comme une méthode d'apprentissage supervisé utilisée pour prédire une variable quantitative.

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

Chapitre 1 : Regression linéaire simple

Dans ce chapitre, on a rappeler brièvement les formules de la régression linéaire simple qui permettent d'expliquer des hypothèses nécessaires et les termes du modèle, les notions d'estimation des paramètres du modèle par la méthode de moindres carrés ordinaire, l'estimation de l'intervalle de confiance, tester la signification des paramètres et la signification globale du modèle et finalement une attention particulière est accordée aux prévisions.

Chapitre 2 : Regression linéaire multiple

Ce chapitre 2, concernant la régression multiple, constitue une généralisation naturelle de la régression simple, où on cherche à expliquer les valeurs prises par la variable endogène Y à l'aide de p variables exogènes $X_1, ..., X_p$. La différence essentielle réside dans le formalisme qui passe par des écritures matricielles des estimateurs et de leurs variances. Pratiquement, on reccourt aux moyens informatique par l'inversion de matrices (calculs très longs).

Chapitre 3 : Application de la RLM sur la balance commerciale algériènne

Le dernier chapitre est réservé à l'application du modèles de régression linéaire multiple, traités théoriquement dans le chapitre 2 où on cherche à établir une relation entre la balance commerciale et les exportations hors hydrocarbures, les exportations hydrocarbures ainsi que l'importations.

On mentionne que tous les travaux, présentés dans ce mémoire, sont traités à l'aide du logiciel *R,* R version 3.3.0 (2016), (voir *Ihaka, R. et Gentleman, R.* [20]) qui est présenté dans l'annexe $[B]$.

Chapitre 1

Régression linéaire simple

La régression linéaire se classe parmi les méthodes d'analyses multivariées qui 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.

Ce chapitre est une introduction à la modélisation linéaire par le modèle le plus élémentaire, la régression linéaire simple (RLS) où une variable Y est expliquée, modélisée par une fonction affine d'une autre variable X.

Après avoir expliciter les hypothèses nécessaires et les termes du modèle, les notions d'estimation des paramètres du modèle par la méthode de moindres carrés ordinaire, l'estimation par intervalle de confiance, on passe à tester la signification des paramètres et la signification globale du modèle. Enfin une attention particulière est faite au prévision.

Pour une présentation assez complète du sujet, on renvoie à les ouvrages : [6], [5], [23], [30],...

1.1 Modèle de la RLS

On cherche à expliquer ou à prévoir les variations d'une variable Y (appelée une variable *endogène*, *dépendante* ou *prédire*) par celles d'une fonction linéaire d'une variable explicative$X$ (variable *exogène*, *indépendante* ou *prédictive*).

Définition 1.1.1 *(Modèle de la RLS)*

Le modèle de régression linéaire simple, noté par RLS, ou modèle linéaire simple est défini par :

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

où β_0 et β_1 sont des paramètres réels inconnus (les coefficients du modèle), et le ε_i est l'erreur du modèle.

Remarque 1.1.1

1. *Le coefficient β_0 est appelé aussi l'ordonnée à l'origine (intercept ou constante). C'est la valeur prédite de y quand $x = 0$ et le coefficient β_1 est appelé la pente. C'est le changement sur y lorsque x change d'une unité.*
2. *Le terme aléatoire ε tient un rôle très important dans la régression. Il permet de résumer toute l'information qui n'est pas prise en compte dans la relation linéaire que l'on cherche à établir entre Y et X c.à.d. résumer le rôle des variables explicatives absentes.*

Les hypothèses relatives à ce modèle sont les suivantes :

H_1 L'erreur est centrée (en moyenne les erreurs s'annulent c.à.d. le modèle est bien spécifié) et de variance constante (homoscédasticité) :

$$E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma_\varepsilon^2, \quad \forall\ i = 1, ..., n.$$

H_2 Les erreurs relatives à 2 observations sont indépendantes :

$$cov(\varepsilon_i, \varepsilon_j) = 0.$$

H_3 Les ε_i sont indépendants et identiquement distribués (iid) suit la loi normale de moyenne nulle et de variance σ_ε^2, on écrit $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Remarque 1.1.2 *Les propriétés des estimateurs des coefficients de la RLS, comme on le verra plus bas, reposent en grande partie sur les hypothèses que on formulera à propos de ε. En pratique, après avoir estimé les paramètres de la régression, les premières vérifications portent sur l'erreur calculée sur les données (on parle de "résidus").*

1.2 Moindres Carrés Ordinaires

La première chose à faire est de dessiner le nuage des points (x_i, y_i) $\forall$ $i = 1, ..., n$, pour déterminer le type de liaison pouvant exister entre x et y. A priori, n'importe quel type de liaison est possible, par exemple celles présentées dans la figure (1.1)

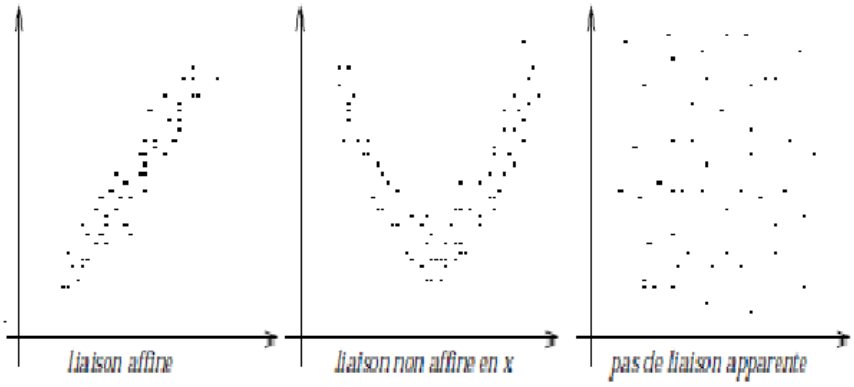

FIG. 1.1 – Exemple de différentes liaisons possibles entre x et y

Si le nuage des points est affine (linéaire), le problème est maintenant de trouver la droite "la plus proche" de ce nuage, en un certain sens. La méthode la plus couramment utilisée est la méthode des moindres carrés ordinaires, due à *Legendre* dans un article de 1805 sur la détermination des orbites des comètes (voir [25]). Cette méthode consiste a prendre la droite pour laquelle la somme des carrés des distances verticales des points à la droite est minimale (voir la figure 1.2) c.à.d. elle consiste d'abord à déterminer des valeurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$ qui minimisent la quantité de la somme des carrés des écarts (des erreurs).

Définition 1.2.1 *(Estimateurs des Moindres Carrés Ordinaires)*
On appelle estimateurs des Moindres Carrés Ordinaires (notée par MCO) $\widehat{\beta}_0$ *et* $\widehat{\beta}_1$ *les valeurs qui minimisent la quantité :*

$$S\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2, \quad où \quad \widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

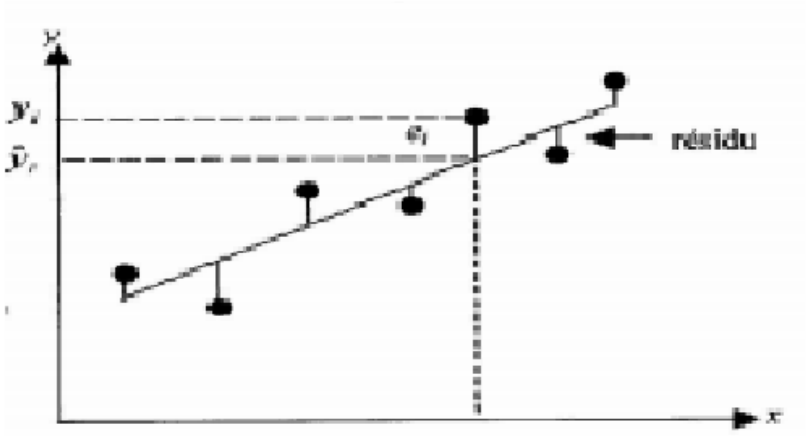

FIG. 1.2 – Droite et résidu de la régression linéaire

1.2.1 Calcul des estimateurs de β_0 et β_1

La fonction de deux variables $S\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$ est une fonction quadratique et sa minimisation ne pose aucun problème, comme on va le voir maintenant.

Proposition 1.2.1 $\left(\textit{Estimateurs } \widehat{\beta}_0 \textit{ et } \widehat{\beta}_1\right)$

Les estimateurs des MCO ont pour expressions :

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

et

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

Preuve (Voir [17], [30], [6] et [23])

La valeur de la fonction $S\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$ est minimum lorsque les dérivées de S par rapport à $\widehat{\beta}_0$ et $\widehat{\beta}_1$ s'annulent :

$$\frac{\partial S\left(\widehat{\beta}_0, \widehat{\beta}_1\right)}{\partial \widehat{\beta}_0} = 0 \quad et \quad \frac{\partial S\left(\widehat{\beta}_0, \widehat{\beta}_1\right)}{\partial \widehat{\beta}_1} = 0.$$

Les dérivées par rapport à $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont :

$$\begin{aligned}\frac{\partial S\left(\widehat{\beta}_0,\widehat{\beta}_1\right)}{\partial\widehat{\beta}_0}=0 &\Leftrightarrow -2\sum_{i=1}^{n}(y_i-\widehat{\beta}_1x_i-\widehat{\beta}_0)=0\\ &\Leftrightarrow -2\left[\sum_{i=1}^{n}y_i-\widehat{\beta}_1\sum_{i=1}^{n}x_i-n\widehat{\beta}_0\right]=0.\end{aligned} \tag{1.1}$$

$$\begin{aligned}\frac{\partial S\left(\widehat{\beta}_0,\widehat{\beta}_1\right)}{\partial\widehat{\beta}_1}=0 &\Leftrightarrow -2\sum_{i=1}^{n}x_i(y_i-\widehat{\beta}_1x_i-\widehat{\beta}_0)=0,\\ &\Leftrightarrow -2\left[\sum_{i=1}^{n}x_iy_i-\widehat{\beta}_1\sum_{i=1}^{n}x_i^2-\widehat{\beta}_0\sum_{i=1}^{n}x_i\right]=0.\end{aligned} \tag{1.2}$$

On pose :

$\overline{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i,$ La moyenne empirique des x_i.

$\overline{y}=\dfrac{1}{n}\sum_{i=1}^{n}y_i,$ La moyenne empirique des y_i.

$s_x^2=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2=\dfrac{1}{n}\sum_{i=1}^{n}x_i-\overline{x}_n^2,$ La variance empirique des x_i.

$s_y^2=\dfrac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y}_n)^2=\dfrac{1}{n}\sum_{i=1}^{n}y_i-\overline{y}_n^2,$ La variance empirique des y_i.

$\begin{aligned}s_{xy} &=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_n)(y_i-\overline{y}_n)\\ &=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{x}_n\overline{y}_n,\end{aligned}$ La covariance empirique entre les x_i et les y_i.

En multipliant l'équation (1.2) par $\dfrac{-1}{n}$, on obtient :

$$\frac{1}{n}\sum_{i=1}^{n}y_i-\widehat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}x_i-\widehat{\beta}_0=0.$$

Alors

$$\widehat{\beta}_0=\overline{y}-\widehat{\beta}_1\overline{x}. \tag{1.3}$$

En substituants l'équation (1.3) dans (1.2), on a :

$$-2\left(\sum_{i=1}^{n}x_iy_i-\widehat{\beta}_1\sum_{i=1}^{n}x_i^2-\overline{y}\sum_{i=1}^{n}x_i+\widehat{\beta}_1\overline{x}\sum_{i=1}^{n}x_i\right)=0. \tag{1.4}$$

En multipliant (1.4) par $\frac{n}{n}$, on trouve :

$$\sum_{i=1}^{n} x_i y_i - \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right) - n\overline{x}\ \overline{y} = 0.$$

Donc

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \\ &= \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}. \end{aligned}$$

■

Remarque 1.2.1

1. *On peut reécrire* $\widehat{\beta}_1$ *sous la forme :*

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_x^2}.$$

2. *Le résidu de la régression* $\widehat{\varepsilon}$ *est donné par :*

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i.$$

3. *La somme (et donc la moyenne) des résidus est nulle dans une régression avec constante :*

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0.$$

En effet :

$$\begin{aligned} \sum_{i=1}^{n} \widehat{\varepsilon}_i &= \sum_{i=1}^{n} \left[y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \right] \\ &= n\overline{y} - n\widehat{\beta}_0 - n\widehat{\beta}_1 \overline{x} \\ &= n\overline{y} - n(\overline{y} - \widehat{\beta}_1 \overline{x}) - n\widehat{\beta}_1 \overline{x} \\ &= 0. \end{aligned}$$

4. *La droite de régression avec constante passe forcément par le centre de gravité du nuage de points* $(\overline{x}, \overline{y})$. *Pour le vérifier simplement, réalise la projection pour le point* $\overline{x}$:

$$\begin{aligned}\widehat{y}(\overline{x}) &= \widehat{\beta}_0 + \widehat{\beta}_1 x_i \\ &= (\overline{y} - \widehat{\beta}_1 \overline{x}) + \widehat{\beta}_1 x_i \\ &= \overline{y}.\end{aligned}$$

5. *L'estimateur de la variance de l'erreur, noté par* s^2_ε, *est donné comme suit :*

$$\begin{aligned}s^2_{\widehat{\varepsilon}} &= \frac{1}{n-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2.\end{aligned}$$

Le $(n-2)$ *peut s'expliquer par la règle : nombre de données* (n) *moins le nombre de paramètres du modèle* (2).

1.2.2 Propriétés des estimateurs

Sous l'hypothèse H_1 et H_2 de centrages, homoscédasticités et décorrélations des erreurs ε_i du modèle, on peut donner certaines propriétés des estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$ des moindres carrés (de base pour les calculs ci-dessous voir [30], [6], [17], [23] et [11]).

Théorème 1.2.1 *(Estimateurs sans biais)*
$\widehat{\beta}_0$ *et* $\widehat{\beta}_1$ *sont des estimateurs sans biais de* β_0 *et* β_1 :

$$E\left[\widehat{\beta}_0\right] = \beta_0 \quad et \quad E\left[\widehat{\beta}_1\right] = \beta_1.$$

Preuve Soit le modèle de la RLS :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1.5}$$

on peut calculer :

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \beta_1\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \frac{1}{n}(n\beta_0) + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i.$$

$$\overline{y} = \beta_0 + \beta_1 \overline{x} + \overline{\varepsilon}. \tag{1.6}$$

Par (1.5) − (1.6) , on obtient

$$y_i - \overline{y} = \beta_1 (x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon}).$$

Et on a :

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Ainsi

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})[\beta_1 (x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})]}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \\
&= \frac{\beta_1 \sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.
\end{aligned}$$

On a $\overline{\varepsilon} \sum_{i=1}^{n} (x_i - \overline{x}) = 0$, donc :

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

Alors

$$\begin{aligned}
E(\widehat{\beta}_1) &= E(\beta_1) + E\left[\frac{\sum_{i=1}^{n} (x_i - \overline{x})\varepsilon_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right] \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^{n} (x_i - \overline{x})\varepsilon_i}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right].
\end{aligned}$$

On a :

$$E(\widehat{\beta}_1) = E(\beta_1) + E\left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right].$$

La variable exogène X n'est pas aléatoire. Donc :

$$E(\widehat{\beta}_1) = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \times E(\varepsilon_i),$$

D'après l'hypothèse que $E(\varepsilon_i) = 0$. On obtient :

$$E(\widehat{\beta}_1) = \beta_1.$$

Pour $\widehat{\beta}_0$, on part de l'expression

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x},$$

d'où l'on tire :

$$\begin{aligned} E\left[\widehat{\beta}_0\right] &= E\left[\overline{y}\right] - \overline{x}E\left[\widehat{\beta}_1\right] \\ &= E\left[\beta_0 + \beta_1\overline{x} + \overline{\varepsilon}\right] - \overline{x}E\left[\widehat{\beta}_1\right] \\ &= \beta_0 + \beta_1\overline{x} + \overline{\varepsilon} - \overline{x}E\left[\widehat{\beta}_1\right], \\ &= \beta_0 + \overline{\varepsilon} - (\beta_1 - \beta_1)\overline{x}. \end{aligned}$$

Donc,

$$E(\widehat{\beta}_0) = \beta_0.$$

■

On peut également exprimer la variances et la covariance de ces estimateurs.

Théorème 1.2.2 *(Variances et covariance)*

Les variances des estimateurs sont :

$$s^2_{\widehat{\beta}_0} = \frac{s^2_{\widehat{\varepsilon}} \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \overline{x})^2} = s^2_{\widehat{\varepsilon}} \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right),$$

et

$$s^2_{\widehat{\beta}_1} = \frac{s^2_{\widehat{\varepsilon}}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

tandis que leur covariance vaut :

$$cov(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{s^2_{\widehat{\varepsilon}} \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Preuve (Voir [30] et [5])

1. Preuve que : $s^2_{\widehat{\beta}_1} = \dfrac{s^2_{\widehat{\varepsilon}}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$

   On a

$$\begin{aligned} s^2_{\widehat{\beta}_1} &= s^2 \left(\beta_1 + \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \varepsilon_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \\ &= s^2 (\beta_1) + s^2 \left(\sum_{i=1}^{n} \frac{(x_i - \overline{x})}{\sum_{j=1}^{n} (x_i - \overline{x})^2} \varepsilon_i \right), \text{ on pose } w_i = \frac{(x_i - \overline{x})}{\sum_{j=1}^{n} (x_i - \overline{x})^2} \\ &= s^2_{\widehat{\varepsilon}} \left(\sum_{i=1}^{n} w_i \varepsilon_i \right), \quad car\ \beta_1 \text{ est constante.} \\ &= \sum_{i=1}^{n} w_i^2 s^2 (\varepsilon_i) + 2 \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j cov(\varepsilon_i, \varepsilon_j), \text{ où } cov(\varepsilon_i, \varepsilon_j) = 0, \end{aligned}$$

   alors

$$s^2_{\widehat{\beta}_1} = \sum_{i=1}^{n} w_i^2 s^2 (\varepsilon_i), \text{ avec } s^2 (\varepsilon_i) = s^2_{\widehat{\varepsilon}} \text{ d'après l'hypothèse } H_1.$$

Donc

$$\begin{aligned} s^2_{\widehat{\beta}_1} &= s^2_\varepsilon \sum_{i=1}^{n} w_i^2 \\ &= s^2_\varepsilon \sum_{i=1}^{n} \left(\frac{(x_i - \overline{x})}{\sum_{j=1}^{n} (x_i - \overline{x})^2} \right)^2 \\ &= s^2_\varepsilon \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\left[\sum_{j=1}^{n} (x_i - \overline{x})^2 \right]^2} \\ &= s^2_\varepsilon \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\left[\sum_{i=1}^{n} (x_i - \overline{x})^2 \right]^2} \\ &= \frac{s^2_\varepsilon}{\sum_{i=1}^{n} (x_i - \overline{x})^2}. \end{aligned}$$

2. Preuve que $s^2_{\widehat{\beta}_0} = s^2_\varepsilon \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)$

Ona :

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x},$$

donc :

$$\begin{aligned} s^2_{\widehat{\beta}_0} &= s^2 \left[\overline{y} - \widehat{\beta}_1 \overline{x} \right] \\ &= s^2 \left(\overline{y} \right) + \overline{x}^2 s^2 \left(\widehat{\beta}_1 \right) - 2\overline{x} cov(\overline{y}, \widehat{\beta}_1) \end{aligned}$$

Où $cov(\overline{y}, \widehat{\beta}_1) = 0$ (voir [5]), et

$$\begin{aligned} s^2 \left(\overline{y} \right) &= s^2 \left(\frac{1}{n} \sum_{i=1}^{n} y_i \right) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} s^2 \left(y_i \right) \\ &= \frac{1}{n^2} n s^2_\varepsilon \\ &= \frac{s^2_\varepsilon}{n}. \end{aligned}$$

Donc

$$\begin{aligned} s^2_{\widehat{\beta}_0} &= s^2(\overline{y}) + \overline{x}^2 s^2\left(\widehat{\beta}_1\right) \\ &= s^2_\varepsilon \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right). \end{aligned}$$

3. Montre que $cov(\widehat{\beta}_0, \widehat{\beta}_1) = -\dfrac{s^2_\varepsilon \overline{x}}{\sum (x_i - \overline{x})^2}$

   On a, d'après l'Equation 1.6, que :

$$\overline{y} = \beta_0 + \beta_1 \overline{x} + \overline{\varepsilon}, \tag{1.7}$$

et

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{1.8}$$

En substitution l'Equation 1.7 dans l'Equation 1.8, on obtient :

$$\begin{aligned} \widehat{\beta}_0 &= (\beta_0 + \beta_1 \overline{x} + \overline{\varepsilon}) - \widehat{\beta}_1 \overline{x}, \\ &= \beta_0 - \overline{x}\left(\widehat{\beta}_1 - \beta_1\right) + \overline{\varepsilon}, \end{aligned}$$

donc,

$$\widehat{\beta}_0 - \beta_0 = -\overline{x}\left(\widehat{\beta}_1 - \beta_1\right) + \overline{\varepsilon}. \tag{1.9}$$

D'autre part, on a :

$$\begin{aligned} cov(\widehat{\beta}_0, \widehat{\beta}_1) &= E\left[(\widehat{\beta}_0 - E(\widehat{\beta}_0))(\widehat{\beta}_1 - E(\widehat{\beta}_1))\right] \\ &= E\left[(\widehat{\beta}_0 - \beta_0)(\widehat{\beta}_1 - \beta_1)\right], \quad \textit{car } \widehat{\beta}_0 \textit{ et } \widehat{\beta}_1 \textit{ : sont des} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{estimateurs sans biais} \\ &= E\left[(-\overline{x}(\widehat{\beta}_1 - \beta_1) + \overline{\varepsilon})((\widehat{\beta}_1 - \beta_1))\right], \textit{ en utilisant l'Equation 1.9} \\ &= E\left[-\overline{x}\left((\widehat{\beta}_1 - \beta_1)\right)^2 + \overline{\varepsilon}((\widehat{\beta}_1 - \beta_1))\right] \\ &= -\overline{x}E\left[\left((\widehat{\beta}_1 - \beta_1)\right)^2\right] + \underbrace{\overline{\varepsilon}E\left[((\widehat{\beta}_1 - \beta_1))\right]}, \text{ car } \overline{\varepsilon} = 0 \\ &= -\overline{x}s^2_{\widehat{\beta}_1}, \textit{ avec } s^2_{\widehat{\beta}_1} = s^2_{\widehat{\varepsilon}} / \sum (x_i - \overline{x})^2, \end{aligned}$$

donc,

$$cov(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{-s^2_{\widehat{\varepsilon}}\overline{x}}{\sum (x_i - \overline{x})^2}.$$

■

1.3 Lois des estimateurs et régions de confiance

Pour étudier les coefficients estimés $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$, il importe d'en calculer les paramètres (l'espérance et la variance essentiellement) et de déterminer la loi de distribution. Dès lors, on peut mettre en oeuvre les outils usuels de la statistique inférentielle : la définition des intervalles de variation à un niveau de confiance donné et des tests d'hypothèses, notamment les tests de significativité (voir [29]).

Proposition 1.3.1 *(Lois des estimateurs avec variance connue)*
Les lois des estimateurs des MCO avec variance σ^2_ε connue sont :

1. $\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} \sim \mathcal{N}(\beta, \Gamma)$ *où* $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ *et*

$$\Gamma = \sigma^2_\varepsilon \begin{bmatrix} \sum_{i=1}^n x_i^2 / \left(n\sum_{i=1}^n (x_i - \overline{x})^2\right) & -\overline{x}/\sum_{i=1}^n (x_i - \overline{x})^2 \\ -\overline{x}/\sum_{i=1}^n (x_i - \overline{x})^2 & 1/\sum_{i=1}^n (x_i - \overline{x})^2 \end{bmatrix}$$

$$= \begin{bmatrix} s^2_{\widehat{\beta}_0} & cov(\widehat{\beta}_0, \widehat{\beta}_1) \\ cov(\widehat{\beta}_0, \widehat{\beta}_1) & s^2_{\widehat{\beta}_1} \end{bmatrix}.$$

   où $cov(\widehat{\beta}_0, \widehat{\beta}_1)$ *: est la covariance entre* $\widehat{\beta}_0$ *et* $\widehat{\beta}_1$*.*

2. $\dfrac{(n-2)}{\sigma^2_\varepsilon} s^2_\varepsilon \sim \mathcal{X}^2_{n-2}$*,loi du* $\mathcal{X}^2$ *à* $(n-2)$ *degrés de liberté (ddl).*
3. $\widehat{\beta}$ *et* s^2_ε *sont indépendants.*

Remarque 1.3.1 *Le problème des propriétés ci-dessus vient de ce qu'elles font intervenir la variance théorique σ^2_ε, généralement inconnue. La façon naturelle de procéder est de la remplacer par son estimateur s^2_ε.*

Proposition 1.3.2 *(Lois des estimateurs avec variance inconnue)*

Les lois des estimateurs des MCO avec variance σ_ε^2 inconnue sont :

1. $\dfrac{\widehat{\beta}_0 - \beta_0}{s_{\widehat{\beta}_0}} \sim \mathcal{T}_{n-2}$ *où* $\mathcal{T}_{n-2}$ *est une loi de Student à* $(n-2)$ *ddl.*
2. $\dfrac{\widehat{\beta}_1 - \beta_1}{s_{\widehat{\beta}_1}} \sim \mathcal{T}_{n-2}$.
3. $\dfrac{1}{2s^2_{\widehat{\varepsilon}}}(\widehat{\beta}_0 - \beta_0)^t \Gamma^{-1} (\widehat{\beta}_1 - \beta_1) \sim \mathcal{F}_{(2,n-2)}$, *où* $\mathcal{F}_{(2,n-2)}$ *est la loi de Fisher de paramètres* $(2, n-2)$.

Ces dernières propriétés on permettent de donner des intervalles de confiance (IC) ou des régions de confiance (RC) des estimateurs. En effet, la valeur ponctuelle d'un estimateur est de peu d'intérêt en général et il est intéressant de lui associer un intervalle de confiance. Les résultats sont donnés pour un seuil α petit où $\alpha \in [0, 1]$.

Proposition 1.3.3 *(Intervalles et régions de confiance)*

1. *L'intervalle de confiance pour β_0, noté par $IC(\beta_0)$, est :*

$$\left[\widehat{\beta}_0 - t_{n-2}^{1-\alpha/2} s_{\widehat{\beta}_0}, \widehat{\beta}_0 + t_{n-2}^{1-\alpha/2} s_{\widehat{\beta}_0}\right],$$

   où $t_{n-2}^{1-\alpha/2}$ est le quantile de niveau $(1-\alpha/2)$ d'une loi de Student $\mathcal{T}_{n-2}$.

2. *Un intervalle de confiance pour β_1, noté par $IC(\beta_1)$, est :*

$$\left[\widehat{\beta}_1 - t_{n-2}^{1-\alpha/2} s_{\widehat{\beta}_1}, \widehat{\beta}_1 + t_{n-2}^{1-\alpha/2} s_{\widehat{\beta}_1}\right].$$

3. *Une région de confiance simultanée pour β_0 et β_1, notée $RC(\beta)$, au niveau $(1-\alpha)$ est :*

$$\frac{1}{2s_\varepsilon^2}[n(\widehat{\beta}_0 - \beta_0) - 2n\overline{x}(\widehat{\beta}_0 - \beta_0)(\widehat{\beta}_1 - \beta_1) + \sum_{i=1}^{n} x_i^2 (\widehat{\beta}_1 - \beta_1)^2] \leq \mathcal{F}_{(2,n-2)}^{1-\alpha},$$

   où $\mathcal{F}_{(2,n-2)}^{1-\alpha}$: le quantile de niveau $(1-\alpha)$ d'une loi $\mathcal{F}_{(2,n-2)}$.

4. *Un intervalle de confiance de σ_ε^2 est donné par :*

$$\left[\frac{(n-2)s_\varepsilon^2}{c_{n-2}^{1-\alpha/2}}, \frac{(n-2)s_\varepsilon^2}{c_{n-2}^{\alpha/2}}\right].$$

   où $c_{n-2}^{1-\alpha/2}$ est le quantile de niveau $(1-\alpha/2)$ d'une loi Khi-2 $\mathcal{X}_{n-2}^2$.

1.4 Analyse de la variance et coefficient de détermination

1.4.1 Décomposition de la variance et tableau d'ANOVA

En un point d'observation (x_i, y_i) on décompose l'écart entre y_i et la moyenne des y_i en ajoutant puis retranchant $\widehat{y}$ la valeur estimée de y par la droite de régression.

Cette procédure fait apparaître une somme de deux écarts :

$$y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})\,.$$

Ainsi l'écart total $(y_i - \overline{y})$ peut être vu comme la somme de deux écarts :

- Un écart entre y_i observé et $\widehat{y}_i$ la valeur estimée par le modèle.
- Un écart entre $\widehat{y}_i$ la valeur estimée par le modèle et la moyenne $\overline{y}$.

On élève les deux membres au carré et on somme sur les observations i :

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}\left[(y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})\right]^2 \\ &\sum_{i=1}^{n}\left[(\widehat{y}_i - \overline{y}) + \widehat{\varepsilon}_i\right]^2 \\ &= \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}\widehat{\varepsilon}_i^2 + 2\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})\widehat{\varepsilon}_i.\end{aligned}$$

Dans la régression avec constante et uniquement dans ce cas, on montre que :

$$2\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})\widehat{\varepsilon}_i = 0.$$

En effet, on a :

$$\begin{aligned}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})\widehat{\varepsilon}_i &= \sum_{i=1}^{n}(\widehat{\beta}_1 x_i + \widehat{\beta}_0 - \overline{y})\widehat{\varepsilon}_i \\ &= \widehat{\beta}_1\sum_{i=1}^{n}x_i\widehat{\varepsilon}_i + \widehat{\beta}_0\sum_{i=1}^{n}\widehat{\varepsilon}_i - \overline{y}\sum_{i=1}^{n}\widehat{\varepsilon}_i.\end{aligned}$$

On a $\sum_{i=1}^{n}\widehat{\varepsilon}_i = 0$, donc on obtient que $\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})\widehat{\varepsilon}_i = 0$.
On aboutit enfin à l'égalité fondamentale (l'équation d'analyse de la variance) :

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{SCR}. \tag{1.10}$$

Alors

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}_{SCR}.$$

Où

✓ SCT : est la somme des carrés totaux où elle indique la variabilité totale de Y c.à.d. l'information disponible dans les données.

✓ SCE : est la somme des carrés expliquées, cette quantité est la variabilité expliquée par le modèle c.à.d. la variation de Y expliquée par X.

✓ SCR : est la somme des carrés résiduels. Elle écrite la variabilité non-expliquée (résiduelle) par le modèle (l'écart entre y et $\widehat{y}$).

Remarque 1.4.1 *Deux situations extrêmes peuvent survenir :*

- *Dans le meilleur des cas, $SCR = 0$ et donc $SCT = SCE$: les variations de Y sont complètement expliquées par celles de X. On a un modèle parfait, la droite de régression passe exactement par tous les points du nuage ($\widehat{y}_i = y_i$).*
- *Dans le pire des cas, $SCE = 0$: X n'apporte aucune information sur Y.*

On produire du tableau d'analyse de variance (voir le tab 1.1) à partir de la décomposition de la variance (voir [11]), comme suit :

Source de variation	ddl	Somme des carrés	Carrés moyens
Expliquée	1	SCE	$CME = \dfrac{SCE}{1}$
Résiduelle	$n-2$	SCR	$CMR = \dfrac{SCR}{n-2}$
Totale	$n-1$	SCT	

TAB. 1.1 – Tableau d'analyse de variance de la régression linéaire simple

1.4.2 Coefficient de détermination R^2

Le coefficient de détermination R^2 est défini par :

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$
$$= 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

En effet, d'après l'équation d'ANOVA (1.10), on a :

$$\frac{SCT}{SCT} = \frac{SCE}{SCT} + \frac{SCR}{SCT}$$
$$\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} + \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

La quantité $\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 / \sum_{i=1}^{n}(y_i - \overline{y})^2 = R^2$ est appelée le coefficient de détermination, alors :

$$1 = R^2 + \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$
$$\Rightarrow \quad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Remarque 1.4.2

1. *Ce coefficient R^2 qui varie entre 0 et 1, mesure la proportion de variation totale de Y autour de la moyenne expliquée par la régression, c.à.d. prise en compte par le modèle.*
2. *Plus R^2 se rapproche de la valeur 1, meilleure est l'adéquation du modèle aux données et un R^2 faible (proche de 0) signifie que le modèle a un faible pouvoir explicatif.*

1.5 Test de signification

1.5.1 Test de signification globale du modèle

Ce test permet de connaître l'apport global de la variable X à la détermination de Y.

On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0. \end{cases}$$

Pour tester cette hypothèse, on a basé sur la statistique de Fisher, notée par F :

$$F = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}}. \tag{1.11}$$

Cette statistique indique si la variance expliquée est significativement supérieure à la variance résiduelle. Dans ce cas, on peut considérer que l'explication emmenée par la régression traduit une relation qui existe réellement dans la population.

Sous H_0, SCE est distribué selon un $\mathcal{X}^2(1)$ et SCR selon un $\mathcal{X}^2(n-2)$, de fait pour F, on a

$$F \equiv \frac{\frac{\mathcal{X}^2(1)}{1}}{\frac{\mathcal{X}^2(n-2)}{n-2}} = f_{1,n-2}^{1-\alpha}. \tag{1.12}$$

Alors, sous H_0, F est donc distribué selon une loi de Fisher à $(1, n-2)$ degrés de liberté, où on rejette H_0 si :

$$F \geq f_{1,n-2}^{1-\alpha},$$

avec $f_{1,n-2}^{1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi de Fisher à $(1, n-2)$ ddl.

Remarque 1.5.1

1. *On peut réecrire la statistique F en fonction de R^2 comme suit:*

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

2. *Dans la plupart des logiciels statistiques, on fournit directement la probabilité critique (p-value). Elle correspond à la probabilité que la loi de Fisher dépasse la statistique calculée F. Ainsi, la règle de décision (rejette H_0) au risque α devient :*

$$p - value \leq \alpha.$$

1.5.2 Test de signification des paramètres

Test de signification de β_0

On veut tester l'hypothèse :

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

Pour tester cette hypothèse, on forme la statistique de test :

$$T_{\widehat{\beta}_0} = \frac{\widehat{\beta}_0}{s_{\widehat{\beta}_0}}.$$

On rejette H_0 si l'observation de la statistique de test, notée $T_{\widehat{\beta}_0}$, est telle que :

$$\left|T_{\widehat{\beta}_0}\right| \geq t_{n-2}^{1-\frac{\alpha}{2}},$$

où $t_{n-2}^{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi Student à $(n-2)$ ddl.

Test de signification de la pente β_1

Le test de significativité de la pente consiste à vérifier l'exogène X sur l'endogène Y. L'hypothèse à confronter s'écrive :

$$\begin{cases} H_0 : \beta_1 = 0, \\ H_1 : \beta_1 \neq 0. \end{cases}$$

Pour tester cette hypothèse, on forme la statistique de test :

$$T_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1}{\widehat{\sigma}_{\widehat{\beta}_1}}.$$

On rejette H_0 si l'observation de la statistique de test, notée T_{β_1}, est telle que :

$$\left| T_{\widehat{\beta}_1} \right| \geq t_{n-2}^{1-\frac{\alpha}{2}},$$

où $t_{n-2}^{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi Student à $(n-2)$ ddl.

Remarque 1.5.2 *Si $n \geq 30$, loi de student tend vers la loi normale.*

1.6 Prévision

Un des buts de la régression est de faire la prévision, c.à.d. de prévoir la variable à expliquer y en présence d'une nouvelle valeur de la variable explicative x.

Soit x_{n+1} une nouvelle valeur, pour laquelle on veut prédire y_{n+1}. Le modèle est toujours le même :

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}.$$

Avec $E[\varepsilon_{n+1}] = 0, Var(\varepsilon_{n+1}) = \sigma_\varepsilon^2$ et $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, ..., n$. Il est naturel de prédire la valeur correspondante via le modèle ajusté :

$$\widehat{y}_{n+1} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{n+1}. \tag{1.13}$$

En matière de prévision dans le cas d'erreurs gaussiennes, les résultats obtenus c'est pour l'espérance et la variance sont toujours valables. De plus, puisque $\widehat{y}_{n+1}$ est linéaire en $\widehat{\beta}_0$ et $\widehat{\beta}_1$, ε_{n+1} on peut préciser sa loi :

$$\widehat{\varepsilon}_{n+1} = y_{n+1} - \widehat{y}_{n+1} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\right) \tag{1.14}$$

On ne connaît pas σ_ε^2,on l'estime donc par $s_{\widehat{\varepsilon}}^2$. Comme $(y_{n+1} - \widehat{y}_{n+1})$ et $s_{\widehat{\varepsilon}}^2(n-2)/\sigma_\varepsilon^2$ sont indépendants, on peut énoncer un résultat donnant des intervalles de confiance pour y_{n+1}. Avec les notations et hypothèses précédentes, on a :

$$\frac{y_{n+1} - \widehat{y}_{n+1}}{s_{\widehat{\varepsilon}} \sqrt{\dfrac{1}{n} + \dfrac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} \sim T_{n-2},$$

d'où l'on déduit l'intervalle de confiance suivant pour y_{n+1} :

$$\left[\widehat{y}_{n+1} - t_{n-2}^{1-\alpha/2} s_{\widehat{\varepsilon}} \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}, \widehat{y}_{n+1} + t_{n-2}^{1-\alpha/2} s_{\widehat{\varepsilon}} \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \right] \tag{1.15}$$

Chapitre 2

Régréssion linéaire multiple

La régression linéaire est une des méthodes statistiques les plus utilisées dans de nombreux domaines pour l'étude de données multidimensionnelles, il constitue la généralisation naturelle de la régression simple, où on cherche à expliquer les valeurs prises par la variable endogène Y à l'aide de p variables exogènes $X_1, ..., X_p$. La différence essentielle réside dans le formalisme qui passe par des écritures matricielles des estimateurs et de leurs variances.

Sur le plan pratique, la mise en oeuvre des méthodes statistiques obtenues nécessite donc, sauf dans des situations particulières, de recourir à des moyens informatiques (logiciels) pour l'inversion de matrices.

On retrouve dans ce chapitre la plupart des éléments de base présentés dans le précédent et la justification de certains d'entre eux : hypothèses du modèle, étude des résidus, méthode des moindres carrés, propriétés des estimateurs, l'estimation par intervalle de confiance et aussi on teste la signification des paramètres et la signification globale du modèle et finalement la prévision.

Pour plus détails sur la régression linéaire multiple, se référer : [30], [14], [16], [19],...

2.1 Modèle de la RLM

On cherche à décrire la relation existant entre une variable quantitative Y appelée variable à expliquer (ou encore, réponse, exogène, dépendante) et plusieurs variables quantitatives $X_1, ..., X_p$ dites variables explicatives (ou encore de contrôle, endogènes, indépendantes, régresseurs) (pour plus détails voir [30]).

Définition 2.1.1 *(Modèle de la RLM)*

Le modèle de régression linéaire multiple, noté par RLM, ou modèle linéaire multiple est défini par :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \quad \forall i \in \{1, ..., n\}, \tag{2.1}$$

où $\beta_0, \beta_1, \beta_2, ..., \beta_p$ sont appelés les paramètres ou les coefficients inconnus du modèle que l'on veut estimer à partir des données et .

On remarque dans ce modèle de RLM (*Equation* 2.1) :

- $i = 1, ..., n$ correspond au numéro des observations.
- y_i est la $i - ème$ observation de la variable Y.
- x_{ij} est la $i - ème$ observation de la $j - ème$ variable.
- ε_i est l'erreur du modèle (bruit). Il représente la déviation entre ce que le modèle prédit et la réalité.

2.2 Notation matricielle

Ce modèle s'écrit sous la forme des équations comme suit :

$$\begin{cases} y_1 = & \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_p x_{1P} + \varepsilon_1 \\ y_2 = & \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_p x_{2P} + \varepsilon_2 \\ . & . \\ . & . \\ . & . \\ y_n = & \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_p x_{nP} + \varepsilon_n \end{cases}$$

Ou sous la forme matricielle :

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & . & . & . & x_{1p} \\ 1 & x_{21} & . & . & . & x_{2p} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ 1 & x_{n1} & . & . & . & x_{np} \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}. \tag{2.2}$$

De façon équivalente, on écrit :

$$Y = \beta X + \varepsilon,$$

où

- Y : est le vecteur à expliquer de taille $n \times 1$.
- X : est la matrice, de taille $n \times (p+1)$, qui contient l'ensemble des observations sur les exogènes , avec une première colonne formée par la valeur 1 indiquant que l'on intègre la constante β_0 dans l'équation où p étant le nombre de variables explicatives réelles.
- ε : est le vecteur des erreurs de taille $n \times 1$.

Remarque 2.2.1

1. *Le coefficient β_0 est un paramètre appelé intercepte qui représente la moyenne des y_i lorsque la valeur de chaque variable explicative est égale à 0.*
2. *Les coefficients β_j $(j = 1, ..., p)$ représentent le changement subi par $E(y_i)$ correspondant à un changement unitaire dans la valeur de la $j-ième$ variable explicative, lorsque les autres variables explicatives demeurent inchangées.*

2.3 Hypothèses relatives au modèle de la RLM

Comme en régression simple, les hypothèses permettent de déterminer les propriétés des estimateurs (biais, convergence) et leurs lois de distributions (pour les estimations par in-

tervalle et les tests d'hypothèses) (voir *Bourbonnais* [6], *Labrousse* [23] et *Giraud et Chaix* [17]).

Il existe principalement deux catégories d'hypothèses :

1. **Hypothèses stochastiques**

   H_1 Les erreurs sont centrées (le modèle est bien spécifié en moyenne), c.à.d l'ensenble des déterminants de y qui n'ont pas été retenus dans le modèle est d'espérance nulle :

   $$E(\varepsilon_i) = 0, \quad \forall i \in \{1, ..., n\}.$$

   H_2 La variance des erreurs est constante, on parle d'homogénéité des variances ou encore d'homoscédasticité :

   $$Var(\varepsilon_i) = \sigma_\varepsilon^2, \quad \forall i \in \{1, ..., n\}.$$

   H_3 Les erreurs relative à deux terme aléatoires ne sont pas corrélé, on dit n'y a pas de corrélation sérielle :

   $$cov(\varepsilon_i, \varepsilon_j) = 0, \quad \forall\ i \neq j.$$

   H_4 Les ε_i sont indépendants et identiquement distribués ($i.i.d$) suit la loi normale de moyenne nulle et de variance σ_ε^2, on écrit :

   $$\varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right).$$

2. **Hypothèses structurelles**

   H_1 La matrice (X^tX) (X^t est la matrice transposée de X), est non singulière de rang p, c.à.d $det(X^tX) \neq 0$ et $(X^tX)^{-1}$ existe. Cette hypothèse implique l'absence de colinéarité entre les variables exogènes $(X_1, ..., X_p)$, c.à.d. les différents vecteurs X_j sont linéairement indépendants. En cas de multicolinéarité, la méthode des MCO devient défaillante.

   H_2 $\dfrac{(X^tX)}{n}$ tend vers une matrice finie non singulière lorsque $n \to +\infty$.

   H_3 $n > p + 1$ le nombre d'observations est supérieur au nombre des paramètres du modèle β_j $(j = 0, ..., p)$.

2.4 Estimation des paramètres par MCO

Conditionnellement à la connaissance des valeurs des X_j $(j = 1, ..., p)$, les paramètres inconnus du modèle : le vecteur $\beta = (\beta_0, ..., \beta_p)^t$ et σ_ε^2, sont estimés par minimisation du critère des moindres carrés ordinaire (MCO).

Le principe des moindres carrés choisit le vecteur $\widehat{\beta}$ minimisant la fonction de la somme des carrés des résidus, notée par SCR.

2.4.1 Dérivation des estimateurs

Soit $\widehat{\varepsilon} = (\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n)^t$ le vecteur de dimension $n \times 1$ des résidus définie par :

$$\widehat{\varepsilon} = Y - \widehat{Y} = Y - X\widehat{\beta},$$

où $\widehat{Y} = X\widehat{\beta}$ représente les valeurs estimées par le modèle, on les appelle aussi valeurs ajustées.
La somme des carrés des résidus est donnée par (voir [9]) :

$$\begin{aligned} SCR &= \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \widehat{\varepsilon}^t \widehat{\varepsilon} \\ &= \left(Y - X\widehat{\beta}\right)^t \left(Y - X\widehat{\beta}\right). \end{aligned}$$

Proposition 2.4.1 *(Estimateur par MCO de β)*
L'estimateur par moindres carrés ordinaires (MCO) $\widehat{\beta}$ de β dans le modèle de régression linéaire multiple est :

$$\widehat{\beta} = (X^t X)^{-1} X^t Y \tag{2.3}$$

Preuve

On voit que SCR est une fonction de $\widehat{\beta}$. On le choisira de façon à minimiser SCR. Le

minimum de SCR est atteint lorsque la dérivée de SCR par rapport à $\widehat{\beta}$ s'annule, donc :

$$\begin{aligned}
\widehat{\beta} &= Argmin \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 \\
&= Argmin(\widehat{\varepsilon}^t \widehat{\varepsilon}) \\
&= (Y - X\widehat{\beta})^t (Y - X\widehat{\beta}) \\
&= (Y^t - \widehat{\beta}^t X^t)(Y - X\widehat{\beta}), \qquad \text{car } \left(X\widehat{\beta}\right)^t = \widehat{\beta}^t X^t \\
&= Y^t Y - Y^t X\widehat{\beta} - X^t \widehat{\beta}^t Y + X^t \widehat{\beta}^t X\widehat{\beta} \\
&= Y^t Y - 2X^t\ \widehat{\beta}^t Y + X^t\ \widehat{\beta}^t X\widehat{\beta}, \qquad \textit{car } X^t\widehat{\beta}^t Y \textit{ est un scalaire,} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{il est égal à sa transposée} \\
&= Y^t Y - 2X^t \widehat{\beta}^t Y + X^t \widehat{\beta}^t X\widehat{\beta}.
\end{aligned}$$

Pour déterminer le minimum de SCR, on réalise la dérivation matricielle :

$$\begin{aligned}
\frac{\partial SCR}{\partial \widehat{\beta}} = 0 &\iff -2X^t Y + 2X^t X\widehat{\beta} = 0. \\
&\iff \widehat{\beta}(X^t X) = X^t Y.
\end{aligned}$$

D'où le résultat. ■

2.4.2 Propriétés des estimateurs MCO

Deux questions reviennent toujours lorsque l'on souhaite étudier les propriétés d'un estimateur :

est-il sans biais ? est-il convergent ?

Le passage du modèle de régression de 2 à p variables explicatives ne modifie en rien les propriétés statistiques de l'estimateur MCO. De même, l'interprétation de ces propriétés reste inchangée (voir [2]).

Théorème 2.4.1 *(Propriétés des estimateurs MCO)*

L'estimateur $\widehat{\beta}$ des moindres carrés ordinaire est sans biais, c.à.d $E(\widehat{\beta}) = \beta$, et sa matrice de variance covariance, notée par $varcov(\widehat{\beta})$ ou par $s^2(\widehat{\beta})$, est :

$$varcov(\widehat{\beta}) = \sigma_\varepsilon^2 (X^t X)^{-1}.$$

Remarque 2.4.1 *La matrice de variance-covariance $\left(varcov(\widehat{\beta})\right)$ des coefficients, de dimension $(p+1, p+1)$, est donnée par :*

$$varcov(\widehat{\beta}) = \begin{pmatrix} var\left(\widehat{\beta}_0\right) & cov\left(\widehat{\beta}_0, \widehat{\beta}_1\right) & \cdots & cov\left(\widehat{\beta}_0, \widehat{\beta}_p\right) \\ \cdots & var\left(\widehat{\beta}_1\right) & \cdots & cov\left(\widehat{\beta}_1, \widehat{\beta}_p\right) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdots & \cdots & \cdots & var\left(\widehat{\beta}_p\right) \end{pmatrix}$$

Cette matrice est symétrique, sur sa diagonale principale on observe les variances des coefficients estimés ($var\left(\widehat{\beta}_0\right), ..., var\left(\widehat{\beta}_p\right)$).

Preuve

1. Montre que $E(\widehat{\beta}) = \beta$

   Soit

$$\begin{aligned} \widehat{\beta} &= (X^t X)^{-1} X^t Y \\ &= (X^t X)^{-1} X^t [X\beta + \varepsilon] \\ &= (X^t X)^{-1} \left(X^t X\right) \beta + (X^t X)^{-1} X^t \varepsilon \\ &= \beta + (X^t X)^{-1} X^t \varepsilon \end{aligned}$$

   Alors, l'espérance mathématique de $\widehat{\beta}$ est :

$$\begin{aligned} E(\widehat{\beta}) &= E[\beta + (X^t X)^{-1} X^t \varepsilon] \\ &= \beta + E\left[(X^t X)^{-1} X^t \varepsilon\right] \\ &= \beta + (X^t X)^{-1} X^t E[\varepsilon], \qquad \text{car } X \text{ est non aléatoire.} \end{aligned}$$

Et sous l'hypothèse que $E(\varepsilon) = 0$. Il vient que :

$$E(\widehat{\beta}) = \beta.$$

2. Montre que $Varcov(\widehat{\beta}) = \sigma_\varepsilon^2 (X^t X)^{-1}$

   On procède de même, on a $\widehat{\beta} = \beta + (X^t X)^{-1} X^t \varepsilon$, donc :

$$\begin{aligned} Varcov(\widehat{\beta}) &= Var(\beta + (X^t X)^{-1} X^t \varepsilon) \\ &= (X^t X)^{-1} X^t Var(\varepsilon) X (X^t X)^{-1} \\ &= (X^t X)^{-1} X^t \sigma_\varepsilon^2 \mathbb{1}_n X (X^t X)^{-1} \\ &= \sigma_\varepsilon^2 (X^t X)^{-1} X^t X (X^t X)^{-1} \\ &= \sigma_\varepsilon^2 (X^t X)^{-1} \end{aligned}$$

■

On trouve :

$$Varcov(\widehat{\beta}) = \sigma_\varepsilon^2 (X^t X)^{-1}. \tag{2.4}$$

Pour estimer cette matrice de variance-covariance de $\widehat{\beta}$ donnée par l'*Equation* 2.4. Il suffit de remplacer la variance théorique des résidus, σ_ε^2, par son estimateur.

2.5 Estimation de la variance du résidu σ_ε^2

Lemme 2.5.1 *Soit un vecteur Z composé de n variables aléatoires d'espérances nulles, et tel que var $(Z) = \sigma_Z^2 \mathbb{1}_n$ et A une matrice symétrique non aléatoire, alors*

$$E\left[Z^t A Z\right] = \sigma_Z^2 tr(A),$$

où $tr(A)$: est la trace de la matrice A.

Grâce au Lemme 2.5.1, on peut calculer l'espérance de $\widehat{\varepsilon}^t \widehat{\varepsilon}$.

Théorème 2.5.1 *(voir [34])*
Soit $\widehat{\varepsilon} = Y - X\widehat{\beta}$, alors

$$E\left[\widehat{\varepsilon}^t \widehat{\varepsilon}\right] = (n - p - 1)\sigma_\varepsilon^2.$$

Preuve (voir [34])

Soit

$$\begin{aligned}\widehat{Y} &= X\widehat{\beta} \\ &= \left[X(X^{t}X)^{-1}X^{t}\right]Y \\ &= HY, \quad avec\ H = X(X^{t}X)^{-1}X^{t}.\end{aligned}$$

La différence entre les valeurs observées Y et les valeurs ajustées $\widehat{Y}$ porte le nom vecteurs des résidus, noté par $\widehat{\varepsilon}$, est donné par la relation :

$$\widehat{\varepsilon} = Y - \widehat{Y} = Y - HY = (\mathbb{1}_n - H)Y, \tag{2.5}$$

où $\mathbb{1}_n$ est la matrice unité de dimension (n, n).

La matrice H est appelée la matrice chapeau (en anglais hat matrix) de taille $n \times n$. Elle vérifie les deux propriétés (voir [19]) :

$$\begin{cases} Symétrique : & H^t = H \\ Idempotente : & H^2 = H \end{cases} \tag{2.6}$$

La matrice $\mathbb{1}_n - H$ a les mêmes propriétés :

$$\begin{cases} (\mathbb{1}_n - H)^t = \mathbb{1}_n - H \\ (\mathbb{1}_n - H)^2 = \mathbb{1}_n - H \end{cases}$$

$$\begin{aligned}\widehat{\varepsilon} &= Y - \widehat{Y} = Y - HY \\ &= (\mathbb{1}_n - H)\,(X\beta + \varepsilon) \\ &= X\beta - HX\beta + \varepsilon - H\varepsilon,\end{aligned}$$

où $HX = \left(X(X^{t}X)^{-1}X^{t}\right)X = X$ et $(I - H)X = 0$, ce qui donne

$$\widehat{\varepsilon} = (\mathbb{1}_n - H)\varepsilon$$

On obtient :

$$\begin{aligned}
\widehat{\varepsilon}^t\widehat{\varepsilon} &= ((\mathbb{1}_n - H)\varepsilon)^t\,(\mathbb{1}_n - H)\varepsilon \\
&= \varepsilon^t(\mathbb{1}_n - H)^t(\mathbb{1}_n - H)\varepsilon \\
&= \varepsilon^t(\mathbb{1}_n - H)\varepsilon, \text{ d'après l'Equation 2.6} \\
&= \varepsilon^t\mathbb{1}_n\varepsilon - \varepsilon^t H\varepsilon
\end{aligned}$$

Donc, par le lemme 2.5.1, on obtient :

$$\begin{aligned}
E\left[\widehat{\varepsilon}^t\widehat{\varepsilon}\right] &= E\left[\varepsilon^t\mathbb{1}_n\varepsilon - \varepsilon^t H\varepsilon\right] \\
&= E\left[\varepsilon^t\mathbb{1}_n\varepsilon\right] - E\left[\varepsilon^t H\varepsilon\right] \\
&= \sigma_\varepsilon^2 tr(\mathbb{1}_n) - \sigma_\varepsilon^2 tr(H) \\
&= \left[tr(\mathbb{1}_n) - tr(H)\right]\sigma_\varepsilon^2
\end{aligned}$$

Où $tr(\mathbb{1}_n) = n$ et

$$\begin{aligned}
tr(H) &= tr(X(X^tX)^{-1}X^t) \\
&= tr(X^tX(X^tX)^{-1}), && \text{puisque } tr(AB) = tr(BA) \\
&= tr(\mathbb{1}_{p+1}), && \text{car } X^tX \text{ est une matrice} \\
& && \text{carrée de taille } (p+1) \\
&= p+1
\end{aligned}$$

Donc

$$E\left[\widehat{\varepsilon}^t\widehat{\varepsilon}\right] = (n - (p+1))\,\sigma_\varepsilon^2.$$

■

D'après le théorème 2.5.1, on peut construire l'estimateur sans biais pour σ_ε^2 qui est :

$$\begin{aligned}
s_{\widehat{\varepsilon}}^2 = \widehat{\sigma}_{\widehat{\varepsilon}}^2 &= \frac{\widehat{\varepsilon}^t\widehat{\varepsilon}}{n-p-1} \\
&= \frac{\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}{n-p-1} \\
&= \frac{SCR}{n-p-1}, \quad \text{où } \widehat{\varepsilon} = Y - X\widehat{\beta}.
\end{aligned}$$

Remarque 2.5.1 *Géométriquement, H est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace vectoriel $vect(X)$ engendré par les vecteurs colonnes de X (vect $\{1, X_1, ..., X_p\}$) (voir la Figure 2.1).*

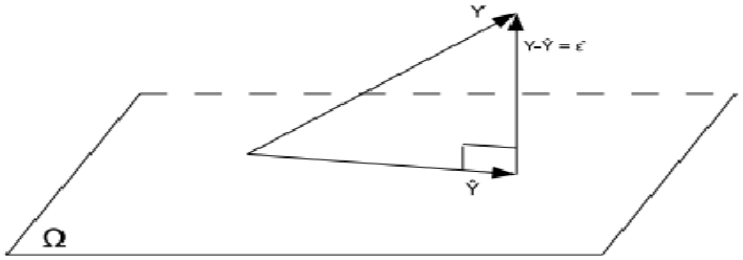

FIG. 2.1 – Géométriquement, la régression est la projection $\widehat{Y}$ de Y sur l'espace vectoriel Vect $\{1, X_1, ..., X_p\}$

2.6 Lois des estimateurs et intervalles de confiance

Après avoir obtenu l'estimateur, son espérance et une estimation de sa variance, il ne reste plus qu'a calculer sa loi de distribution pour construire des intervalles de confiance ou des tests d'hypothèses sur β.

Proposition 2.6.1 *(Lois des estimateurs avec variance connue)*
Les lois des estimateurs des MCO avec variance connue sont :

1. *$\widehat{\beta}$ est un vecteur gaussien de moyenne β et de variance $\sigma_\varepsilon^2(X^tX)^{-1}$:*

$$\widehat{\beta} \sim \mathcal{N}(\beta, \sigma_\varepsilon^2(X^tX)^{-1}).$$

2. *$\widehat{\beta}$ et s_ε^2 sont indépendants.*

3. $(n-p-1)\dfrac{S_\varepsilon^2}{\sigma_\varepsilon^2} \sim \mathcal{X}^2_{n-p-1}$.

Remarque 2.6.1 *Bien entendu le premier point du résultat précédent n'est pas satisfaisant pour obtenir des région de confiance sur β car il suppose la variance σ_ε^2 connue, ce qui n'est pas le cas en général. La proposition suivante pallie cette insuffisance.*

Proposition 2.6.2 *(Lois des estimateurs avec variance inconnue)*

Les lois des estimateurs des MCO avec variance inconnue sont :

1. *Pour $j = 1, ..., p$, on a t_j suit une loi de Student à $(n-p-1)$ degrés de liberté (ddl), on écrit :*
$$t_j = \frac{\widehat{\beta}_j - \beta_j}{s_{\widehat{\beta}_j}} \sim \mathcal{T}^{1-\frac{\alpha}{2}}_{n-p-1}.$$
2. *Soit R une matrice de taille $q \times (p+1)$ de rang q [1] $(q \leq (p+1))$ alors :*
$$\frac{1}{q s^2_{\widehat{\varepsilon}}}(R(\widehat{\beta}-\beta))^t[R(X^tX)^{-1}R^t]^{-1}R(\widehat{\beta}-\beta) \sim \mathcal{F}^{1-\alpha}_{(q,n-p)},$$
où $\mathcal{F}^{1-\alpha}_{(q,n-p-1)}$ est le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à $(q, n-p-1)$ ddl.

Les logiciels et certains ouvrages donnent des intervalles de confiance (IC) pour les paramètres pris séparément. Cependant ces intervalles de confiance ne tiennent pas compte de la dépendance des paramètres, ce qui conduirait à construire plutôt des régions de confiance (RC). On a donc traiter les deux cas.

Proposition 2.6.3 *(Intervalles et régions de confiance)*

1. *Pour tout $j \in \{1, ..., p\}$, un intervalle de confiance de niveau $(1-\alpha)$ pour β_j est :*
$$\left[\widehat{\beta}_j - t^{1-\alpha/2}_{n-p-1}s_{\widehat{\beta}_j}, \widehat{\beta}_j + t^{1-\alpha/2}_{n-p-1}s_{\widehat{\beta}_j}\right],$$
où $t^{1-\alpha/2}_{n-p-1}$ est le quantile de niveau $(1-\alpha/2)$ d'une loi de Student $\mathcal{T}_{n-p-1}$.
2. *Un intervalle de confiance de niveau $(1-\alpha)$ pour σ_ε^2 est :*
$$\left[\frac{(n-p-1)s^2_\varepsilon}{c^{1-\alpha/2}_{n-p-1}}, \frac{(n-p-1)s^2_\varepsilon}{c^{(\alpha/2)}_{n-p-1}}\right],$$
où $c^{1-\alpha/2}_{n-p-1}$ est le quantile d'ordre $(1-\alpha/2)$ d'une loi $\mathcal{X}^2_{n-p-1}$.

[1] Le rang d'une matrice est le nombre maximum de lignes (ou de colonnes) linéairement indépendantes.

3. Pour q $(q \leq (p+1))$, une région de confiance de niveau $(1-\alpha)$ pour $R\beta$ est :

$$\frac{1}{qs^2_{\widehat{\varepsilon}}}(R(\widehat{\beta}-\beta))^t[R(X'X)^{-1}R^t]^{-1}(R(\widehat{\beta}-\beta)) \leq \mathcal{F}^{1-\alpha}_{(q,n-p-1)},$$

où $\mathcal{F}^{1-\alpha}_{(q,n-p-1)}$ est le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à $(q, n-p-1)$ ddl.

2.7 Analyse de la variance et coefficient de détermination

2.7.1 Décomposition de la variance et tableau d'ANOVA

On peut aisément vérifier que, comme dans le modèle de régression simple, l'écart entre y_i et la moyenne des y_i en ajoutant puis retranchant $\widehat{y}$ la valeur estimée de y par la droite de régression.

Cette procédure fait apparaître une somme de deux écarts :

$$y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})\,.$$

On peut obtenir la décomposition :

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{SCR}$$
$$= \underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}_{SCR}.$$

Où

✓ SCT : désigne la somme des carrés totaux (centrés).

✓ SCE : la somme des carrés expliqués (centrés).

✓ SCR : la somme des carrés des résidus.

Le tableau "d'analyse de la la variance" se présenté sous la forme suivante (voir le tableau 2.1) :

Source de variation	ddl	Somme des carrés	Carrés moyens
Expliquée	p	SCE	$CME = \dfrac{SCE}{p}$
Résiduelle	$n-p-1$	SCR	$CMR = \dfrac{SCR}{n-p-1}$
Totale	$n-1$	SCT	

TAB. 2.1 – Tableau d'analyse de la variance de la régression linéaire multiple

Remarque 2.7.1 *On peut reécrire l'équation d'ANOVA matriciellement comme suit :*

$$\underbrace{(y-\overline{y})^t (y-\overline{y})}_{SCT} = \underbrace{(\widehat{y}-\overline{y})^t (\widehat{y}-\overline{y})}_{SCE} + \underbrace{\widehat{\varepsilon}^t\widehat{\varepsilon}}_{SCR},$$

où $\overline{y}$ *le vecteur de* $\mathbb{R}^n$ *contenant* n *fois la moyenne de la variable* y*, c.à.d*

$$\overline{y} = (\overline{y}, ...\overline{y})^t .$$

2.7.2 Coefficient de détermination R^2

Le rapport entre SCE et SCT représente la proportion de variance expliquée et porte le nom de coefficient de détermination, noté par R^2, (voir [10]) :

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$
$$= 1 - \frac{\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{SCR}{SCT}.$$

Ce coefficient R^2 est compris entre 0 et 1 : plus il est proche de 1 et plus grande est la part expliquée, autrement dit meilleure est la régression. Inversement, un coefficient R^2 proche de 0 indique que la quantité SCR est élevée.

2.7.3 Coefficient de détermination ajusté R^2_{adj}

Le coefficient R^2 est un indicateur de la qualité de l'ajustement des valeurs observées par le modèle mais il a le défaut de ne pas tenir compte du nombre de variables explicatives utilisés dans le modèle. On ne peut pas l'utiliser pour comparer plusieurs modèles entre eux car, si on ajoute une variable explicative à un modèle, la part des erreurs diminue forcément et donc le coefficient R^2 augmente : cela signifie que plus il y a de variables explicatives et plus le R^2 est élevé. Or un modèle n'est pas nécessairement meilleur parce qu'il a plus de variables explicatives.

On définit donc un coefficient R^2 ajusté qui tient compte des degrés de liberté. Ce coefficient, noté par R^2_{adj}, est défini comme suit :

$$R^2_{adj} = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)} = 1 - \frac{(n-1)}{(n-p-1)}(1-R^2), \tag{2.7}$$

Remarque 2.7.2 *On a R^2_{adj} est toujours inférieur à R^2, et ceci d'autant plus que le modèle contient un grand nombre de prédicteurs (variables explicatives).*

2.8 Test de signification

2.8.1 Test de signification globale du modèle

L'objectif du test global de Fisher est d'étudier la liaison globale entre Y *et les variables explicatives* X_j $(j = 1, ..., p)$ (significative ou non) (voir [14]).

On considère les hypothèses :

$$\begin{cases} H_0 : \beta_0 = \beta_1 = ... = \beta_p = 0 \\ H_1 : \exists!\ \beta_j \neq 0,\ (j = 1, ..., p). \end{cases}$$

Pour tester l'hypothèse, on a vu que l'on peut utiliser la statistique de Fisher F :

$$F = \frac{SCE/p}{SCR/(n-p-1)} = \frac{CMR}{CME},$$

où F suit une loi de Fisher avec p et $(n-p-1)$ degrés de liberté.

On rejette H_0 si :

$$F \geq f_{p,n-p-1}^{1-\alpha},$$

avec $f_{1,n-2}^{1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à $(p, n-p-1)$ ddl (voir [1]).

Remarque 2.8.1

1. *Dans la plupart des logiciels statistiques, on fournit directement la probabilité critique (p-value). Elle correspond à la probabilité que la loi de Fisher dépasse la statistique calculée F. Ainsi, la règle de décision (rejette H_0) au risque α devient :*

$$p - value \leq \alpha.$$

2. *Il existe une relation mathématique entre le R^2 et la statistique de test de signification globale (du F de Fisher) comme suit :*

$$F = \frac{(n-p-1)}{p} \frac{R^2}{(1-R^2)}.$$

2.8.2 Test de Student de signification du paramètre du modèle

L'objectif du test de Student est d'évaluer l'influence de la variable X_j sur Y $(j = 1, ..., p)$, on considère les hypothèses :

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

où β_j est le paramètre associé à la variable explicative X_j.

L'hypothèse H_0 de nullité d'un paramètre du modèle peut être testée au moyen de la statistique de Student :

$$T_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j}{s_{\widehat{\beta}_j}}. \tag{2.8}$$

à comparer $t_{n-p-1}^{1-\frac{\alpha}{2}}$, où $t_{n-p-1}^{1-\frac{\alpha}{2}}$ est le quantile d'ordre $\left(1-\frac{\alpha}{2}\right)$ d'une loi Student à $(n-p-1)$ ddl.

1. Si $\left|T_{\widehat{\beta}_j}\right| \geq t_{n-p-1}^{1-\frac{\alpha}{2}}$, on rejette H_0.

2. Si $\left|T_{\widehat{\beta}j}\right| < t_{n-p-1}^{1-\frac{\alpha}{2}}$, on ne peut pas rejeter H_0,est le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi student à $(n-p-1)$ *ddl*.

2.9 Prévision

Comme dans le cadre du modèle de régression linéaire multiple, les buts de la régression est de proposer des prédictions pour la variable à expliquer Y lorsque on a de nouvelles valeurs $x_{n+1} = (x_{n+11}, ..., x_{n+1p})^t$ (voir [16]).

Soit donc $x_{n+1} = (x_{n+11}, ..., x_{n+1p})^t$ une nouvelle valeur pour laquelle, on veut prédire y_{n+1} qui est définie par :

$$y_{n+1} = x_{n+1}^{t}\beta + \varepsilon_{n+1},$$

avec $\varepsilon_{n+1} \sim N(0, \sigma_\varepsilon^2)$ indépendant des $(\varepsilon_i)_{1\leq i\leq n}$.

A partir des n observations précédentes, on a pu calculer un estimateur $\widehat{\beta}$ de β. Donc, il est naturel de prédire la valeur correspondante via le modèle ajusté $\widehat{y}_{n+1}$ définie par :

$$\widehat{y}_{n+1} = x_{n+1}^{t}\widehat{\beta}. \tag{2.9}$$

Pour quantifier l'erreur de prévision $(y_{n+1} - \widehat{y}_{n+1})$, on utilise la décomposition :

$$y_{n+1} - \widehat{y}_{n+1} = x_{n+1}^{t}(\beta - \widehat{\beta}) + \varepsilon_{n+1},$$

$(y_{n+1} - \widehat{y}_{n+1})$ est une variable gaussienne, dont moyenne 0 et variance $\sigma_\varepsilon^2(1 + x_{n+1}^t(X^tX)^{-1}x_{n+1})$ ce qui donne :

$$y_{n+1} - \widehat{y}_{n+1} \sim \mathcal{N}(0, \sigma_\varepsilon^2(1 + x_{n+1}^t(X^tX)^{-1}x_{n+1})) \tag{2.10}$$

Comme dans le cas de la régression linéaire simple, on obtient que :

$$\frac{y_{n+1} - \widehat{y}_{n+1}}{s_{\widehat{\varepsilon}}\sqrt{x_{n+1}^t(X^tX)^{-1}x_{n+1}}} \sim T_{(n-p-1)},$$

ce qui permet de construire l'intervalle de confiance pour y_{n+1} :

$$\left[\widehat{y}_{n+1} - t_{(n-p-1)}^{1-\alpha/2}s_{\widehat{\varepsilon}}\sqrt{x_{n+1}^t(X^tX)^{-1}x_{n+1}}, \widehat{y}_{n+1} + t_{(n-p-1)}^{1-\alpha/2}s_{\widehat{\varepsilon}}\sqrt{x_{n+1}^t(X^tX)^{-1}x_{n+1}}\right]. \tag{2.11}$$

Chapitre 3

Application de la RLM sur la balance commerciale algérièn ne

Le commerce est un facteur d'évolution et d'équilibre des états pour la sécurisation de leurs besoins par l'importation et en même temps par l'exportation, qui introduisent et soulignent l'importance de la balance commerciale. Cette balance commerciale est l'indice qui mesure la différence globale entre la valeur des exportations et des importations de biens et services.

L'Algérie est considérée parmi les pays qui sont les processus d'importation et d'exportation, comme elle a suivi les stratégies suivies pour son développement où la stratégie de substitution des importations d'abord, puis le développement des exportations et les encourager.

C'est grâce à l'analyse de chacune des structures des produits de base des importations et des exportations, que l'on constate que la chose la plus importante est d'importer les fournitures industrielles, matériaux et produits semi-finis. Comme pour les exportations, généralement accompagnés par le secteur des hydrocarbures en dépit des efforts pour développer les exportations de carburant vers l'extérieur et tous les matériaux et matières premières semi transformées. En termes de répartition géographique, on constate que l'Union européenne est la plus en contact avec l'Algérie, soit dans le processus d'importation ou d'exportation.

Dans ce chapitre, on cherche à établir une relation entre la balance commerciale et les expor-

tations hors hydrocarbures, les exportations hydrocarbures ainsi que l'importations.

On mentionne que tous les travaux, présentés dans ce chapitre, sont traités à l'aide du logiciel *R,* R version 3.3.0 (2016), (voir *Ihaka, R. et Gentleman, R.* [20]) qui est présenté dans l'annexe $[B]$.

3.1 Définitions économiques

3.1.1 Importations et Exportations

Il s'agit de l'ensemble des opérations sur les biens et services entre les unités institutionnelles résidentes et le reste du monde. Il s'agit de tous les biens et services qui sortent de manière définitive du territoire économique vers le reste du monde.

1. **Importations** : le terme "importations" désigne en économie l'ensemble des achats de marchandises à l'extérieur d'un pays, qu'il s'agisse de biens destinés à la consommation (biens de consommation) ou de biens destinés à servir à l'investissement, en d'autre terme est une entrée dans un pays de biens ou services provenant d'un autre pays.
2. **Exportations** : le terme "exportations" désigne l'ensemble des ventes de marchandises à l'extérieur d'un pays. Elles représentent une injection dans le circuit économique et leur variation positive entraîne une augmentation du revenu national et de l'emploi.

3.1.2 Balance commerciale

La balance commerciale des biens et services retrace la différence entre les exportations de biens et services et les importations de biens et services. Une balance positive indique que les exportations dépassent en valeur les importations (excédent commercial). Une balance négative indique au contraire que les importations dépassent en valeur les exportations (déficit commercial).

3.2 Présentation des données

On souhaite expliquer une variable (ou caractère) quantitative Y (balance commerciale) en fonction de 3 autres variables X_1 (exportations hors hydrocarbures), X_2 (exportations hydrocarbures) et X_3 (importations).

On dispose 30 données relevées durant la période (de 1985 à 2014), concernant l'économie algérienne (où la source est le Centre National de l'Informatique et des Statistiques des Douanes Algériennes). Ces données sont présentées sous forme d'un tableau des données (voir le Figure 3.1).

	Année	exp.Hhydro.usd.	exp.hydro.usd.	imp.usd.	balance.comm
1	1985	252	9893	9840	305
2	1986	199.0	7621	9213	−1393
3	1987	214.0	8019	7056	1177
4	1988	420.0	7685	7324	781
5	1989	396.0	8572	9208	−240
6	1990	439.0	10865	9684	1620
7	1991	375.0	11726	7681	4420
8	1992	449.0	10388	8406	2431
9	1993	479.0	9612	8788	1303
10	1994	287.0	8053	9365	−1025
11	1995	509.0	9731	10761	−521
12	1996	881.0	12494	9098	4277
13	1997	511.0	13378	8687	5202
14	1998	358.0	9855	9403	810
15	1999	438.0	12084	9164	3358
16	2000	947.0	21084	9173	12858
17	2001	648.0	18484	9940	9192
18	2002	734.0	18091	12009	6816
19	2003	673.0	23939	13534	11078
20	2004	781.0	31302	18308	13775
21	2005	907.0	45094	20357	25644
22	2006	1184.0	53429	21456	33157
23	2007	1332.0	58831	27631	32532
24	2008	1937.0	77361	39479	39819
25	2009	1066.0	44128	39294	5900
26	2010	1526.0	55527	40473	16580
27	2011	2062.0	71427	47247	26242
28	2012	2187.0	71794	46801	21490
29	2013	2014.0	62960	55028	9946
30	2014	2582.0	60304	58580	4306

FIG. 3.1 – Table des données (Source : le Centre National de l'Informatique et des Statistiques des Douanes Algériennes)

3.3 Modèle de la régression linéaire multiple

Soit le modèle de la régression linéaire multiple (RLM) avec 3 variables explicatives :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$

où $\beta_0, \beta_1, \beta_2$ et β_3 : sont des paramètres réels inconnus (les coefficients du modèle).

x_{i1} : la i-ème valeur de la variable X_1 (correspond aux exportations hors hydrocarbures en \$).

x_{i2} : la i-ème valeur de la variable X_2 (représente les exportations hydrocarbures en \$).

x_{i3} : la i-ème valeur de la variable X_3 (représente les importations en \$).

y_i : la i-ème valeur de la variable Y (correspond la balance commerciale).

Le dernier terme ε_i représente la déviation entre ce que le modèle prédit et la réalité (l'erreur du modèle).

Donc, l'objet essentiel est d'expliquer la balance commerciale en l'algérie en fonction de variables explicatives : exportations hors hydrocarbures en \$ (X_1), les exportations hydrocarbures en \$ (X_2) et les importations en \$ (X_3).

On peut reécrire ce modèle sous la forme matricielle suivante :

$$Y = \mathbf{X}\beta + \varepsilon.$$

On modélise, sous R, le modèle de RLM par la fonction lm en faisant :

$\boxed{> reg < -lm(Y \sim X_1 + X_2 + X_3)}$

Comme précédemment notre but ici va être de déterminer :

1. La valeur de la constante β_0 et des différents coefficients β_1, β_2 et β_3 qui permettent de minimiser l'erreur entre la droite de régression linéaire estimée et les valeurs réelles de Y.
2. Les variables significatives, c.à.d voir si ces différents coefficients sont différents de 0 ou non.
3. La précision de notre modèle, en utilisant entre autre, le coefficient de détermination "*R-squared*".

3.4 Hypothèses relatives au modèle de la RLM

3.4.1 Résidus

Pour tout $i \in \{1, ..., 30\}$, on appelle $i - ème$ résidu la réalisation e_i de :

$$e_i = \widehat{\varepsilon}_i = y_i - \widehat{y}_i.$$

Ces résidus vont nous permettre de valider ou non les hypothèses initiales du modèle de la RLM. Sous R, on écrit :

```
>residuals(reg)
```

On obtient, le résultat suivant (voir Figure 3.2) :

```
> residuals(reg)
        1         2         3         4         5         6
-265.8869 -309.4565 -250.5442  -62.7702 -121.0634  -83.8020
        7         8         9        10        11        12
-101.1674  -49.8307  -31.4513 -228.4038  -42.7940  349.1204
       13        14        15        16        17        18
  11.2219 -157.2537  -70.8269  433.6855  129.1577  167.2205
       19        20        21        22        23        24
  95.1232  120.7213  236.3295  498.2945  527.7701  909.5623
       25        26        27        28        29        30
   0.0875  441.4072  856.3550 -4708.5644  624.5121 1083.2471
```

FIG. 3.2 – Résidus de la RLM

3.4.2 Indépendance de $\varepsilon_1, ..., \varepsilon_n$

Pour étudier l'indépendance de $\varepsilon_1, ..., \varepsilon_n$ la première approche consiste à tracer le corrélogramme[1]. Celui-ci représente les estimations ponctuelles de la fonction d'autocorrélation

[1]Le corrélogramme est une représentation graphique mettant en évidence une ou plusieurs corrélations entre des séries de données

(acf) définie par :

$$\rho(h) = \frac{cov(\varepsilon_i, \varepsilon_{i+h})}{\sigma(\varepsilon_i)\,\sigma(\varepsilon_{i+h})}, \quad i \in \{1, ..., n-h\}, \quad h \in \{1, ..., n-1\},$$

sous forme de bâtons.

On peut aussi calculer un intervalle de confiance pour (h). Si les bâtons sont de tailles et de signes alternés (ou presque) et qu'aucun d'entre eux ne dépassent les bornes de l'intervalle de confiance (ou presque), on admet l'indépendance de $\varepsilon_1, ..., \varepsilon_n$.

Sous R, avec la commande :

`> acf(residuals(reg), lag=30)`, on obtient la Figure 3.3

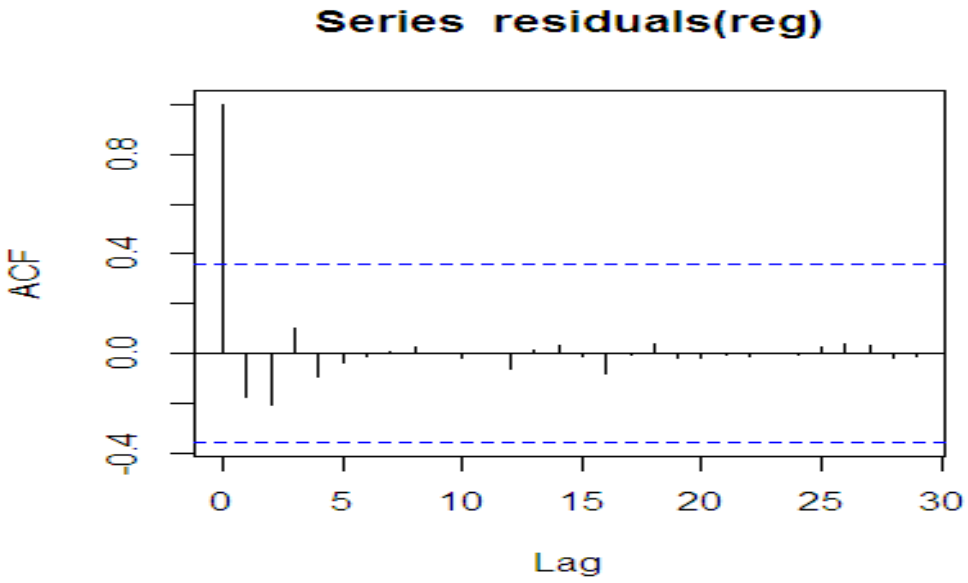

FIG. 3.3 – Fonction d'autocorélation (acf) des résidus

D'après, cette figure (voir la Figure 3.3) on observe que $\varepsilon_1, ..., \varepsilon_n$ sont indépendantes.

Test de Ljung-Box (ou du portemanteau)

On considère les hypothèses :

$$H_0 : \rho(1) = ... = \rho(n) = 0 \quad contre \quad H_1 : au\ moins\ une\ corrélation\ n'est\ pas\ nulle.$$

Partant des résidus $e_1, ..., e_n$, on peut utiliser le test de *Ljung-Box*[2] : si $p - value < \alpha = 0.05$, on admet qu'au moins une corrélation n'est pas nulle, donc que $\varepsilon_1, ..., \varepsilon_n$ ne sont pas indépendantes.

Sous R, on écrit les commandes :

```
> library(lawstat)
> Box.test(residuals(reg), type = "Ljung")
```

Box-Ljung test

data : residuals(reg)

$X - squared = 1.0313$, $df = 1$, $p - value = 0.3099$.

$p - value = 0.3099 > \alpha = 0.05$, on accepte $H_0 : \rho(1) = ... = \rho(n) = 0$, donc les résidus $\varepsilon_1, ..., \varepsilon_n$ ne sont pas corrélés (c.à.d ils sont indépendantes).

3.4.3 Teste d'hétéroscédasticité (égalité des variances des erreurs)

Test de White / Test de Breusch-Pagan

Les tests d'hétéroscédasticité impliquent les deux hypothèses suivantes :

H_0 : *homoscédasticité* (les residus ont tous la mème variance σ^2)

H_1 : *hétéroscédasticité* (les residus n'ont pas tous la mème variance)

Par conséquent, si la $p - value$ associée à un test d'hétéroscédasticité se trouve en-dessous d'un certain seuil (où $\alpha = 0.05$ dans notre application), on pourra dire que les données s'écartent significativement de l'homoscédasticité.

Donc, on a $\varepsilon_1, ..., \varepsilon_n$ sont indépendantes. Pour étudier l'égalité $\sigma^2(\varepsilon_1) = ... = \sigma^2(\varepsilon_n)$, on peut utiliser le test de White ou le test de *Breusch-Pagan* (voir [35])[3].

[2]La statistique de Ljung-Box permet de tester l'hypothèse que les n coefficients d'autocorrélation sont nuls. Elle est basée sur la somme des autocorrélations de la série et elle est distribuée selon une loi Chi-carrée avec n degrés de liberté.

[3]En statistiques, le test de *Breusch-Pagan* permet de tester l'hypothèse d'homoscédasticité du terme d'erreur d'un modèle de régression linéaire. Il a été proposé par *Trevor Breusch* et *Adrian Pagan* dans un article publié en 1979 dans la revue Econometrica.

Ce test s'écrit, sous R, comme suit :

```
> library(lmtest)
> bptest(reg)
```

Studentized Breusch-Pagan test test for homoscedasticity

data : reg

$BP = 5.6577$, $df = 3$, $p - value = 0.1295$.

On remarque que $p - value = 0.1295 > \alpha = 0.05$, nous décidons de garder l'hypothèse nulle d'homoscédasticité donc on admet l'égalité des variances des erreurs (c.à.d $\sigma^2(\varepsilon_1) = ... = \sigma^2(\varepsilon_n)$).

3.4.4 Test de normalité des erreurs

On admet que $\varepsilon_1, ..., \varepsilon_n$ soient indépendantes et $\sigma^2(\varepsilon_1) = ... = \sigma^2(\varepsilon_n)$.

Pour étudier la normalité de $\varepsilon_1, ..., \varepsilon_n$, on trace le nuage de points QQ plot associé (ou diagramme Quantile-Quantile). Si le nuage de points est bien ajusté par la droite $y = x$, alors on admet la normalité de $\varepsilon_1, ..., \varepsilon_n$.

```
> qqnorm(residuals(reg)),
 > qqline(residuals(reg)).
```

On obtient, la Figure 3.4. D'après cette figure, on remarque que le nuage de points est bien ajusté par la droite $y = x$, alors on admet la normalité de $\varepsilon_1, ..., \varepsilon_n$

3.5 Estimation des paramètres par MCO

On va estimer les paramètres et obtiendra :

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1.x_{i1} + \widehat{\beta}_2.x_{i2} + \widehat{\beta}_3.x_{i3}.$$

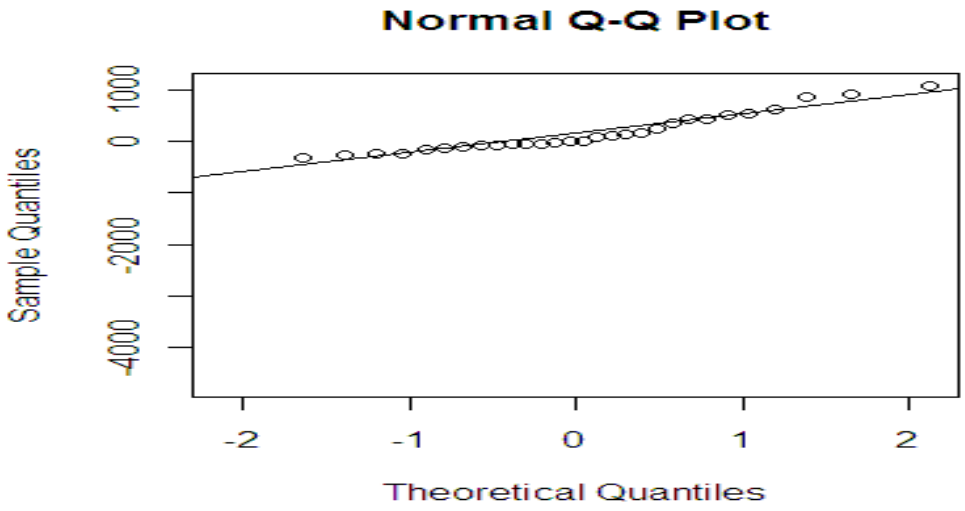

FIG. 3.4 – Normalité des résidus

Il s'agit de calculer le vecteur des estimateurs $\widehat{\beta} = \left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3\right)^t$ défini par l'égalité suivante :

$$\widehat{\beta} = (X^t X)^{-1} X^t Y. \tag{3.1}$$

Les estimateurs $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$ et $\widehat{\beta}_3$ sont donnés directement, sous R, par la commande :
$\boxed{> lm(Y \tilde{} X1 + X2 + X3)}$, on obtient

Call :

$lm(formula = Y \tilde{} X1 + X2 + X3)$

$Coefficients$:

$(Intercept)$	$X1$	$X2$	$X3$
331.6573	0.0576	0.9972	-0.9797

Ou, on écrit l'Equation 3.1, sous R, par :
$\boxed{> hatbeta < -solve(t(X)\%*\%X)\%*\%(t(X)\%*\%Y).}$

où X est la matrice explicative de taille 30×4 (voir l'Equation 2.2).

Sous R, la matrice X est écrite par la commande :
$\boxed{X < -matrix(c(rep(1, 30), X1, X2, X3), ncol = 4)}$, on obtient (voir la Figure 3.5) :

```
> x
      [,1]  [,2]   [,3]   [,4]
 [1,]    1   252   9893   9840
 [2,]    1   199   7621   9213
 [3,]    1   214   8019   7056
 [4,]    1   420   7685   7324
 [5,]    1   396   8572   9208
 [6,]    1   439  10865   9684
 [7,]    1   375  11726   7681
 [8,]    1   449  10388   8406
 [9,]    1   479   9612   8788
[10,]    1   287   8053   9365
[11,]    1   509   9731  10761
[12,]    1   881  12494   9098
[13,]    1   511  13378   8687
[14,]    1   358   9855   9403
[15,]    1   430  12004   9164
[16,]    1   947  21084   9173
[17,]    1   648  18484   9940
[18,]    1   734  18091  12009
[19,]    1   673  23939  13534
[20,]    1   781  31302  18308
[21,]    1   907  45094  20357
[22,]    1  1184  53429  21456
[23,]    1  1332  58831  27631
[24,]    1  1937  77361  39479
[25,]    1  1066  41128  39291
[26,]    1  1526  55527  40473
[27,]    1  2014  71427  47247
[28,]    1  2187  71794  46801
[29,]    1  2062  62960  55028
[30,]    1  2582  60304  58580
```

FIG. 3.5 – Matrice des variables explicatives

Alors, on calcule :

$$X^tX = \begin{bmatrix} 30 & 26787 & 863731 & 592988 \\ 26787 & 36837067 & 1201034462 & 826032749 \\ 863731 & 1201034462 & 41433963591 & 27288580079 \\ 592988 & 826032749 & 27288580079 & 19311461542 \end{bmatrix}$$

Donc,

$$(X^tX)^{-1} = \begin{bmatrix} 9.5892 \times 10^{-2} & -9.6952 \times 10^{-5} & 2.7925 \times 10^{-7} & 8.0795 \times 10^{-7} \\ -9.6952 \times 10^{-5} & 1.0015 \times 10^{-6} & -1.0907 \times 10^{-8} & -2.4449 \times 10^{-8} \\ 2.7925 \times 10^{-7} & -1.0907 \times 10^{-8} & 4.7381 \times 10^{-10} & -2.1157 \times 10^{-10} \\ 8.0795 \times 10^{-7} & -2.4449 \times 10^{-8} & -2.1157 \times 10^{-10} & 1.3717 \times 10^{-9} \end{bmatrix}$$

et

$$X^tY = \begin{bmatrix} 291840 \\ 399392446 \\ 14938316530 \\ 8536480108 \end{bmatrix}$$

Ainsi, on obtient :

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 331.6573 \\ 0.0576 \\ 0.9972 \\ -0.9797 \end{bmatrix}.$$

Donc, l'équation de l'hyperplan des moindres carrés est donc donnée par :

$$\widehat{y_i} = 331.6573 + 0.0576x_{i1} + 0.9972x_{i2} - 0.9797x_{i3}.$$

Le signe du coefficient nous indique le sens de la relation. D'après cette équation, on remarque que les coefficients de régression estimer ($\widehat{\beta}_1$ et $\widehat{\beta}_2$), associée à les variables des exportations hydrocarbures et des exportations hors hydrocarbures, sont positifs cela signifie que l'augmentation des exportations hydrocarbures et les exportations hors hydrocarbures influe positivement sur la balance commerciale. Mais les importations ont un impact négatif sur la balance commerciale car leur coefficient $\widehat{\beta}_3$ est négatif.

3.6 Evaluation

3.6.1 Estimation de la matrice de variance-covariance de $\widehat{\beta}$

La matrice de variance-covariance, notée par $varcov(\widehat{\beta})$ ou par $s^2(\widehat{\beta})$, des coefficients est importante car elle renseigne sur la variance de chaque coefficient estimé et permet de faire les tests des hypothèses, notamment de voir si chaque coefficient est significativement différent de zéro. Elle est définie par :

$$varcov(\widehat{\beta}) = s^2(\widehat{\beta}) = \sigma_\varepsilon^2(X^tX)^{-1},$$

où $s^2_{\widehat{\varepsilon}}$: est l'estimateur sans biais de la variance des résidus donné par :

$$s^2_{\widehat{\varepsilon}} = \frac{SCR}{n-(p+1)} = \frac{\sum(y_i - \widehat{y}_i)^2}{n-(p+1)} = \frac{\sum \widehat{\varepsilon}_i^2}{n-(p+1)}.$$

Sous R, on écrit :

$> SCR <- sum(residuals(reg)\hat{}2)$

$> hatsigma2\_eps <- SCR/(n-p-1)$ (où p : est le nombre des variables explicatives).

Donc, on obtient le résultat suivant :

$$s^2_{\widehat{\varepsilon}} = \frac{SCR}{n-p-1} = \frac{26802860}{26} = 1030879, \text{ où } n = 30 \text{ et } p = 3.$$

On peut alors calculer la matrice de variance-covariance sous R comme suit :

$> varcov <- -hatsigma2\_u * solve(t(X)\%*\%X)$ (où varcov : est la matrice de variance covariance), ou on peut la matrice de variance-covariance, sous R, directement par la commande :

$> vcov(reg)$, on obtient le résultat suivant :

	$(Intercept)$	X_1	X_2	X_3
$(Intercept)$	98853.2354	−99.9464	0.2879	0.8329
X_1	−99.9464	1.0324	−0.0112	−0.0252
X_2	0.2879	−0.0112	0.0005	−0.0002
X_3	0.8329	−0.0252	−0.0002	0.0014

Les écart-types $s(\widehat{\beta}_j)$ des estimateurs $\widehat{\beta}_j$ $(j = 0, 1, 2, 3)$ sont alors donnés par les racines carrées des éléments diagonaux de cette matrice de variance-covariance :

Sous R, on a :

$> hatsigmabetas <- sqrt(diag(vcov(reg)))$:

$$s(\widehat{\beta}_0) = 314.4093,$$

$$s(\widehat{\beta}_1) = 1.0161,$$

$$s(\widehat{\beta}_2) = 0.0221,$$

$$s(\widehat{\beta}_3) = 0.0376.$$

3.6.2 Estimation de β_j par intervalle de confiance

L'intervalle de confiance pour estimer β_j $(j = 0, 1, 2, 3)$, au niveau de confiance $100(1-\alpha)\%$, est donnée par :

$$\left[\widehat{\beta}_j - t_{n-2}^{1-\alpha/2} s_{\widehat{\beta}_j};\ \widehat{\beta}_j + t_{n-2}^{1-\alpha/2} s_{\widehat{\beta}_j}\right].$$

On calcule les intervalles de confiance au niveau 95% pour les 4 paramètres β_0, β_1, β_2 et β_3. Sous R, pour trouver les intervalles de confiance pour β_j $(j = 0, 1, 2, 3)$, on écrit la commande suivante :

$\boxed{> confint\ (reg,\ conf.level = 0.95).}$

On obtient :

$$\beta_0 \in [-314.6203;\ 977.9350],$$
$$\beta_1 \in [-2.0310;\ 2.1462],$$
$$\beta_2 \in [0.9518;\ 1.0426],$$
$$\beta_3 \in [-1.0570;\ -0.9025].$$

3.7 Evaluation globale de la régression

3.7.1 Tableau d'analyse de la variance

L'évaluation globale de la pertinence du modèle de prédiction, s'appuie sur l'équation fondamentale d'analyse de la variance qui est donnée par :

$$SCT = SCE + SCR.$$

Pour donner le tableau d'analyse de variance. Il s'agit de calculer les quantités suivantes :

$$SCE = \widehat{\beta} X^t Y - n\overline{y}^2,$$
$$SCT = Y^t Y - n\overline{y}^2,$$
$$SCR = \sum (y_i - \widehat{y}_i)^2 = \sum \widehat{\varepsilon}_i{}^2.$$

Sous R, on a :

$> SCE < -t(X)\% * \%Y - n * (mean(Y)\hat{}2),$

$> SCT < -t(Y)\% * \%Y - n * (mean(Y)\hat{}2),$

$> SCR < -sum(residuals(reg)\hat{}2).$

On obtient alors les résultats suivants :

$$SCE = 3813910596,$$

$$SCT = 3840713456,$$

$$SCR = 26802860.$$

Alors, le tableau d'analyse de la variance est donné par le tableau (3.1)

Source de variation	ddl	Somme des carrés	Carrés moyens	F_{obs}
Expliquée	3	3813910596	1271303532	1233.223
Résiduelle	26	26802860	1030879	
Totale	29	3840713456		

TAB. 3.1 – Tableau d'analyse de la variance

3.7.2 Coefficient de détermination

Un des usages de la RLM consiste à prédire la valeur d'un Y pour un ensemble des valeurs $x_1, x_2,...,x_p$ donné. La mesure de l'ajustement du modèle aux données est donc importante. La proportion de variabilité expliquée par les 3 régresseurs est calculé par le coefficient de détermination R^2 qui est donnée par la relation suivante :

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

Donc, ce coefficient est un indicateur spécifique permet de traduire la variance expliquée par le modèle.

La commande R associée à R^2 est :

$> R2 < -SCE/SCT.$

On obtient $R^2 = 99.3\%$, ce qui montre un ajustement très fort. Mais, ce coefficient ne prend pas en compte le nombre de variables explicatives. C'est pourquoi, il est nécessaire de s'intéresser au cofficient de détermination ajusté R^2_{adj}, qui lui représente une mesure de l'ajustement corrigée par le nombre de variables du modèle. Ici, ce coefficient vaut :

$$R^2_{adj} = 1 - \frac{n-1}{n-(p+1)}(1-R^2) = 99.22\%,$$

qui reste un ajustement très fort.

3.8 Tests de signification

3.8.1 Test globale de Fisher

Ce test permet de répondre à la question suivante :

"*Est ce que la liaison globale entre Y et les X_j est-elle significative ?*"

On veut tester l'hypothèse nulle :

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0.$$

Contre l'hypothèse alternative :

$$H_1 : \exists\ j \in \{1, 2, 3\} \text{ tel que } \beta_j \neq 0.$$

C'est-à-dire que Y dépend d'au moins une variable X_j.

On calcule la statistique de test :

$$F_{obs} = \frac{R^2/p}{(1-R2)/(n-p-1)} = 1233.223.$$

L'écriture sous R est :

$> R2 < -SCE/SCT$

$> Fobs < -(R2/p)/((1-R2)/(n-p-1))$

$> F\_tab < -qf(p = 1-0.05, df1 = p, df2 = n-p-1)$

Règle de décision

Comme la statistique $F_{obs} = 1233.223$ est supèrieure à la valeur critique $f^{0.95}_{(3,26)} = 2.9751$ (valeur théorique). On conclut que le test est significatif, on rejette l'hypothèse nulle H_0 au seuil de significativité $\alpha = 5\%$. Ce résultat montre qu'au moins une des variables contribue à expliquer la balance commerciale. Mais, il est global et ne nous indique pas si plusieurs variables y contribuent et lesquelles.

3.8.2 Test de Student sur le paramètre β_j

L'objectif du test de Student est d'évaluer l'influence de la variable X_j sur Y. Donc, il permet de répondre à la question suivante :

"*L'apport marginal d'une variable X_j est-il significatif ?*"

Pour $j \in \{0, 1, 2, 3\}$, on considère les hypothèses :

$$H_0 : \beta_j = 0 \quad contre \quad H_1 : \beta_j \neq 0.$$

On calcul la statistique de test T_{obs} :

$$Pour\ \beta_0 : \quad T_{obs} = \frac{\widehat{\beta}_0}{s\left(\widehat{\beta}_0\right)} = \frac{331.6573}{314.4093} = 1.055,$$

$$Pour\ \beta_1 : \quad T_{obs} = \frac{\widehat{\beta}_1}{s\left(\widehat{\beta}_1\right)} = \frac{0.0576}{1.0161} = 0.057,$$

$$Pour\ \beta_2 : \quad T_{obs} = \frac{\widehat{\beta}_2}{s\left(\widehat{\beta}_2\right)} = \frac{0.9972}{0.0221} = 45.122,$$

$$Pour\ \beta_3 : \quad T_{obs} = \frac{\widehat{\beta}_3}{s\left(\widehat{\beta}_3\right)} = \frac{-0.9797}{0.0376} = -26.054.$$

L'écriture sous R :

```
> tobs <- -hatbeta/hatsigmabetas
> T_tab <- -qt(p = 1 - 0.05/2, df = n - p - 1)
> ifelse(abs(tobs) > T_tab,"rejet H0","non rejetH0")
```

Règle de décision

Pour respectivement $j = 0$, $j = 1$, $j = 2$ et $j = 3$. Comme la valeur critique est donnée par $t_{0.975,26} = 2.0555$, on décide d'accepter l'hypothèse nulle H_0 au seuil de significativité $\alpha = 5\%$ pour $j = 0$ et $j = 1$ et on la rejette pour $j = 2$ et $j = 3$.

Donc, cela veut dire que la variable X_1 n'est pas significative dans le modèle (elle n'explique pas les valeurs prises par Y) et que les variables (exportation hydrocarbures (X_2) et importation (X_3)) ont des influences très important sur la balance commerciale. Cela montre que la balance commerciale globale de l'Algérie est basée essentiellement sur l'exportation des hydrocarbures et l'importation de divers produits.

3.9 Prévision

Pour faire une prediction à partir d'une ou plusieurs nouvelles observations, il suffit de créer un *data.frame* contenant exactement les memes noms de colonnes que les données initial. En effet, sous R, on fait la prévision comme suit :

```
x1 <- c(2602, 2610)
x2 <- c(77380, 77388)
x3 <- c(58600, 58608)
new <- data.frame(X1, X2, X3)
print(new) ## c.à.d afficher new ##
```

On obtient le résultat :

	$x1$	$x2$	$x3$
1	2602	77380	58600
2	2610	77388	58608

La valeur prédite moyenne de Y et l'intervalle de confiance pour y_p au niveau $100(1-\alpha)\%$ sont donnés par la commande R suivante :

$\boxed{> predict(reg, new, interval = "prediction")}$, on obtient le résultat :

	fit	lwr	upr
1	20232.83	17861.79	22603.86
2	20233.43	17858.97	22607.88

Où fit : sont les valeurs prédites estimer de Y. lwr : sont les bornes inférieurs des intervalles de confiance et upr : sont les bornes supérieurs des intervalles de confiance.

Conclusion

La régression linéaire se classe parmi les méthodes d'analyses multivariées qui 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 par la méthode de moindres carrés ordinaire (mais il existe aussi de nombreuses autres méthodes pour estimer ce modèle, telle la méthode de maximum de vraisemblance ou encore par inférence bayésienne), l'estimation par intervalle de confiance, on teste la signification des paramètres et la signification globale du modèle. Enfin, une attention particulière est faite aux prévisions.

Bibliographie

[1] Antoniadis A., Berruyer J., Carmona R. (1992). Régression non linéaire et applications, Economica.

[2] Arnauad Guyader (2012/2013). Régression linéaire. Université Rennes 2 Master de statistique.

[3] Belsley, D.A., (1991). Conditioning diagnostics. Collinearity and weak data in regression. ed. Wiley & Sons, p 396.

[4] Bergonzini, J.-Cl. & C. Duby (1995). Analyse et planification des expériences. Les dispositifs en blocs, ed. Masson, Paris, Milan, Barcelone, p 353.

[5] Bertrand F. (02/06/2010). Compléments sur la régression linéaire simple Anova et inférence sur les paramètres. Master 1, IRMA, Université de Strasbourg, Strasbourg, France

[6] Bourbonnais, R. (1998). Econométrie, Manuel et exercices corrigés. Dunod, 2-ème édition.

[7] Chatterjee, S. & B. Price (1991). Regression analysis by example, 2nd edition, J. Wiley & Sons ed., p 278.

[8] Christophe Chesneau (16/03/2016). Modèles de régression. Université de Caen (http ://www.math.unicaen.fr/~chesneau/).

[9] Corinne Perraudin (2004/2005). Le modèle de régression linèaire économetrie. Univ paris Paris I Panthéon Sorbonne DESS Conseil en Organisation et Stratégie.

[10] Desgraupes, M. (2014/2015). Le modèle linéaire. UNIVERSITÉ PARIS OUEST NANTERRE LA DÉFENSE U.F.R. MIASHS L3.

[11] Dodge, Y, Rousson, V. (2004). Analyse de régression appliquée, Dunod, $2^{ième}$ édition.

[12] Faraway, J. ($July, 2002$). Practical Regression and ANOVA using R, http :// cran.r-project. org/doc/ contrib/ Faraway-PRA. pdf.

[13] Fredéric Bertrand & Myriam Maumy-Bertrand (2010). Initiation à la statistique avec R. Dunod.

[14] Frédéric Bestrand & Myriam Maumy-Bertrand (08/03/2012). Régression linéaire multiple, IRMA,univer-de Strasbourg France.

[15] Fourcassié V. & Jost, C. (2012). Introduction aux modeles linéaires généraux (General linear model - GLM). Cours Modules Statistiques Master 2 NCC.

[16] Giorgio Russolillo (2016). La Régression Linéaire Multiple, Département IMATH-CNAM, giorgio.russolillo@com.fr.

[17] Giraud, R., Chaix, N. (1989). Econométrie. Presses Universitaires de France (PUF).

[18] Haurie, A., Modèle de régression linéaire, sur http ://ecolu-info. unige.ch /~haurie /mba 05/.

[19] Hoaglin, D. C. & Welsch, R. E. (1978). The hat matrix in regression and anova. The American Statistician.

[20] Ihaka, R. and Gentleman, R. (1996). R : A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics $\mathbf{5} : 299 - 314$.

[21] Josiane Confais & Monique Le Guen (2006) . Premiers pas en Régression linéaire avec SAS. Revue MODULAD, n°35.

[22] Kmenta, J. (1986). Elements of Econometrics, 2nd. Edit., Macmillan Publishing CO.

[23] Labrousse, C. (1983) . Introduction à l'économétrie. Maîtrise d'économétrie, Dunod.

[24] Lebarbier, E. & Robin, S. (2007).Exemples d'application du modèle linèaire. Institut National Agronomique Paris-Grignon.

[25] Legendre, A.M. (1805) . Nouvelles méthodes pour la détermination des orbites des comètes, Paris, Firmin Didot.

[26] Mc Cullagh, P., Nelder, J. A. (1989). Generalized linear models, Chapman and Hall.

[27] Montgomery, D.C., & Peck, E.A. (1992). Introduction to linear regression analysis, 2nd edition, ed. J. Wiley & Sons, p 527.

[28] Nicolas Jung (2015). Regression linèaire avec R avec l'utilisation de *ggplot*2. Dunod.

[29] Pierre-André Cornillon & Eric Matzner-LØber (2011) . Régression avec R. Springer-Verlag - France.

[30] Rakotomalala, R. (07/06/2011). Econométrie - La régression linéaire simple et multiple. http :// eric.univ-lyon2.fr/~ricco/publications.html.

[31] Rakotomalala, R. (11/03/2016). Analyse de corrélation, étude des dépendances - Variables quantitatives, http ://eric.univ-lyon2.fr/~ricco/publications.html.

[32] Ratkowsky, D.A. (1983). Nonlinear Regression Modeling. A unified practical approach, ed. M. Dekker, p 276.

[33] Scheffe, H. (1959). The Analysis of Variance, J. Wiley & Sons ed., p 477.

[34] Tillé Yves (2011). Resumé du Cours de Modèles de Régression. Institut de statistique, Université de Neuchâtel, Suisse.

[35] Trevor Breusch et Adrian Pagan (1979). Simple test for heteroscedasticity and random coefficient variation, Econometrica. The Econometric Society, vol. 47, no°5, p. 1287 − 1294.

Annexe A : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

Symbole	**Description**
$ANOVA$	Analyse de la variance.
$c_{n-p-1}^{1-\alpha/2}$	quantile d'ordre $(1-\alpha/2)$ d'une loi de $\mathcal{X}^2$ à $(n-p-1)$ ddl.
$c.à.d$	C'est à dire.
CME	carée moyenne expliquée par le modèle.
CMR	carée moyenne résiduelle.
$cov(x,y)$	covariance entre x et y.
ddl	degrés de liberté.
$\mathbb{E}[X]$	espérance mathématique ou moyenne du v.a. X.
$f_{(p,q)}^{(1-\alpha)}$	quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à (p,q) ddl.
Γ	matrice de variance covariance.
$\mathbb{1}_n$	matrice unité de dimension (n,n).
IC	Intervalle de Confiance.
iid	indépendantes et identiquement distribuées.
$\mathbb{R}$	ensemble des valeurs réelles.
$\mathbb{R}_+$	ensemble des valeurs réelles positives.
$\mathcal{X}_p^2$	loi de $\mathcal{X}^2$ à p ddl.
MCO	Moindres Carrés Ordinaires.

$\mathcal{N}(0,1)$	loi normale standard (centrée réduite).
R^2	coefficient de détermination.
R^2_{adj}	coefficient de détermination ajusté.
RC	région de confiance
$resp.$	respectivement.
RLS	régression linéaire simple.
RLM	régression linéaire multiple.
σ^2_x	variance d'une v.a. X.
s^2_y	La variance empirique de la v.a y.
s_{xy}	covariance empirique entre les variables X et Y.
s^2_x	variance empirique de la v.a. X.
SCE	somme des carrés expliquée par le modèle.
SCR	somme des carrés résiduelle.
SCT	somme des carrés totale.
$t^{1-\alpha/2}_{n-p-1}$	qunatile d'ordre $(1-\frac{\alpha}{2})$ de la loi de Student à $(n-p-1)$ ddl.
$Tr(A)$	trace de matrice A.
$v.a.$	variable aléatoire.
$var(X)$	variance mathématique du v.a. X.
X^t	transposée de X.
$(X_1, X_2, ..., X_n)$	échantillon de taille n de X.

Annexe B : Logiciel *R*

R est un système, communément appelé langage et logiciel, qui permet de réaliser des analyses statistiques. Plus particulièrement, il comporte des moyens qui rendent possible la manipulation des données, les calculs et les représentations graphiques. *R* a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes et comporte un grand nombre de procédures statistiques appelées paquets. Ces derniers permettent de traiter assez rapidement des sujets aussi variés que les modèles linéaires (simples et généralisés), la régression (linéaire et non linéaire), les séries chronologiques, les tests paramétriques et non paramétriques classiques, les différentes méthodes d'analyse des données,...

Il a été initialement crée, en 1996, par *Robert Gentleman* et *Ross Ihaka* du département de statistique de l'Université d'Auckland en Nouvelle Zélande. Depuis 1997, il s'est formé une équipe "*R Core Team*" qui développe *R*. Il est conçu pour pouvoir être utilisé avec les systèmes d'exploitation *Unix*, *Linux*, *Windows* et *MacOS*.

Un élément clé dans la mission de développement de *R* est le *Comprehensive R Archive Network* (CRAN) qui est un ensemble de sites qui fournit tout ce qui est nécessaire à la distribution de R, ses extensions, sa documentation, ses fichiers sources et ses fichiers binaires. Le site maître du CRAN est situé en Autriche à Vienne, on peut y accéder par l'URL : "*http ://cran.r-project.org/*". Les autres sites du CRAN, appelés sites miroirs, sont répandus partout dans le monde.

R est un logiciel libre distribué sous les termes de la "GNU Public Licence". Il fait partie intégrante du projet GNU et possède un site officiel à l'adresse "*http ://www.R-project.org*".

ملخص

هذه المذكرة مركزة على تطبيق تقنيات الانحدار الخطي المتعدد باستخدام برنامج R، حيث حاولنا دراسة العلاقة بين الميزان التجاري الصادرات غير النفطية و النفطية وكذلك الواردات، يعني ونحن مهتمون لمعرفة مدى تأثير المتغيرات الثلاثة المذكورة في الميزان التجاري والتنبؤ بالسنوات القادمة.

التطبيق العملي في برنامج R يظهر لنا معامل تحديد يساوي 99.3٪ والمعدل يساوي 99.22٪، و هذه النتيجة تثبت أن الصادرات والواردات تؤثر إلى حد كبير في الميزان التجاري.

الكلمات المفتاحية: معامل التحديد، الميزان التجاري؛ الصادرات النفطية؛ الصادرات الغير نفطية؛ الواردات؛ الانحدار الخطي البسيط؛ الانحدار الخطي المتعدد.

Résumé

Ce mémoire est centré sur l'application des techniques de la régression linéaire multiple en utilisant le logiciels R, où on a essayé d'étudier de la relation entre la balance commerciale et les exportations hors hydrocarbures, les exportations hydrocarbures ainsi que l'importations, c.à.d on s'intéresse à connaitre la manière dont influent les trois variables citées sur la balance commerciale et à prédir les années à venir.

L'application pratique sur le logiciel R nous affiche un coefficient de détermination égal à 99.3% et celui l'ajusté égal à 99.22%, et ce résultat montre que les exportations et l'importation influent largement sur la balance .commerciale

Mots clés: coefficient de détermination, balance commerciale, exportation hydrocarbures, exportation hors hydrocarbure, importation, régression linéaire simple, régression linéaire multiple.

Abstract

This memory focuses on the application of multiple linear regression techniques using the R software, where we tried to study the relationship between the trade balance and the non-hydrocarbon exports, hydrocarbon exports as well as imports, i.e we are interested to know how influential the three variables mentioned on the trade balance and pred years to come.

The practical application in the R software shows us a coefficient of determination equal to 99.3% and Adjusted equal to 99.22%, and this result shows that exports and imports largely affect the trade balance.

Key words: coefficient of determination; trade balance; non-hydrocarbon exports; hydrocarbon exports; imports; simple linear regression; multiple linear regression.