وزارة التعليم العالي والبحث العلمي

Université 20 Aout 1955 de Skikda
Faculté des Sciences
Département de Mathématiques

جامعة 20 أوت 1955 ، سكيكدة
كلـــــــية العلـــــوم
قسم الرياضيات

N0 : U.S/F.S/D.M/……/2023.

Faculté des Sciences
Département de Mathématiques

Mémoire

Présenté en vue de l'obtention du diplôme de
Master en Mathématiques

Application de la régression linéaire multiple sur l'économie nationale

Option : ANEDP

Par :

1. KHEMIS Aya

Encadré par : TILBI Djahida **MCB** U. SKIKDA

Devant le jury :

Président :LALLOUCHE Abdallah MCB U. SKIKDA
Examinateur:*KIMOUCHE Karima* MCA U. SKIKDA

Année : 2022/2023

DEDICACE

Je dédie ce travail à mes chers parents, vous avez toujours été là pour moi toutes ces années, cet effort n’aurait eu aucun sens sans vous, cet accomplissement pour vous avec tout mon amour et ma reconnaissance. A ma sœur Bouchra et mon frère Mouayadbellah à toute ma famille, je vous remercie pour votre soutien , vous m’avez été d’un réel soutien. A tous mes proches et a tous mes amis Ghada, Assala, Marwa, Kawtar, Manale et Nawale. A mes amis, un lot de bac 2018, une bonne remise des diplômes, si Dieu le veut, Meilleurs vœux de succès.

A.KHEMIS

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.*

Je tiens à remercier mon encadrant ***Dr. TILBI Djahida*** *pour sa supervision de mon mémoire et pour ses indications, ses commentaires et son temps.*

Je suis très heureuse que ***Dr. KIMOUCHE Karima*** *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'adresse mon sincère remercie ***Dr. LALLOUCHE Abdallah*** *pour l'honneur qu'il me fait en acceptant la présidence du jury.*

Je vifs remerciement vont aussi à tous les membres de la faculté des sciences, plus précisément de département de mathématique ainsi que toute ma camarades et amis, sous oublier ma parents pour leurs en courage mentes et leurs soutiens.

Enfin un grand merci à mes parents, mon frère, ma sœur et tous mes proches, je n'aurais jamais pu arriver ici, sons l'équilibre, la chaleur, la soutien et le bonheur dans lequel j'ai vécu. Merci

ملخص

سنتناول في هذه المذكرة بعض نماذج الإنحدار الخطي وهي: البسيط والمتعدد، كما سنتناول أيضا تطبيق تقنيات الإنحدار الخطي المتعدد بإستخدام برنامج R في محاولة دراستنا للعلاقة بين الإقتصاد الوطني (الناتج المحلي الإجمالي) و الزراعة و كذلك قطاعي الصناعة و التجارة لمعرفة مدى تأثير هذه المتغيرات الثلاثة على الإقتصاد الوطني.

الكلمات المفتاحية: *الإنحدار الخطي البسيط، الإنحدار الخطي المتعدد، تطبيق عددي، برنامج R ، الإقتصاد الوطني (الناتج المحلي الإجمالي)، الزراعة، الصناعة، التجارة.*

RÉSUMÉ

On va aborder dans ce mémoire quelques modèles de régression savoir: la régression linéaire simple et multiple, nous aborderons également l'application des techniques de régression linéaire multiple en utilisant logiciel R dans notre tentative d'étudier la relation entre l'économie nationale (PIB) et l'agriculture, l'industrie ainsi que le commerce, pour voir l'effet des trois variables citées sur l'économie nationale.

Mots clés : *régression linéaire simple, régression linéaire multiple, application numérique, logiciel R, l'économie national (PIB), agriculture, l'industrie, le commerce.*

ABSTRACT

We will in this graduation note approach sonne regressions models: simple and multiple regression. We will also discuss the application of multiple linear regression techniques using the R software, in our attepmt to study the relationship between national economy (gross dometice product) and the commerce, the agricultural as will as the industrial, to see the effect of the three variables mentioned on the national economy.

Key words *: simple linear regression, multiple linear regression, digital application, R software, national economy(GDP), the agricultural, the industial, the commerce*

TABLE DES MATIÈRES

INTRODUCTION GÉNÉRALE

L'objectif de la régression linéaire est comment analyser un phénomène quelconque on utilisant des méthodes statistiques dites économétriques. Elle est appliquée dans plusieurs domaines, tels que la physique, la biologie, la chimie, l'économie...etc.
La régression linéaire est une modélisation linéaire qui permet d'établir des estimations dans le futur à partir d'informations provenant du passé. Dans ce modèle de régression linéaire, on a plusieurs variables dont une qui est une variable explicative (quantitative, exogène ou dépendante) et les autres qui sont des variables expliquées (quantitatives, endogènes ou indépendantes). Elle est souvent calculée avec la méthode des moindres carrés qui permet de réduire les erreurs en ajoutant de l'information. À titre d'exemples, on peut citer : la relation entre la variable Investissement et la variable Croissance économique, il s'agit de la régression linéaire simple et d'après la théorie économique, la demande d'un produit peut être expliquée par les grandeurs Prix, Revenu et Publicité, il s'agit de la régression linéaire multiple.
La régression linéaire simple et multiple est un outil d'analyse qui fait appel à trois domaines scientifique, à savoir :
* La théorie économique .
* L'analyse statistique .
* La modélisation mathématique.

Ce mémoire est organisé comme suit :
Le premier chapitre est consacré aux rappels sur les notions fondamentales concernant les formules du modèle de la régression linéaire simple et les estimations des paramètres par la méthode des moindres carrés ordinaire, ainsi que l'estimation par intervalle de confiance du ce modèle et de tester la signification des paramètres et la signification globale.

Dans le deuxième chapitre on introduit les principales notions et propriétés du modèle de la régression linéaire multiple avec plusieurs variables explicatives. Dans ce chapitre, la différence essentielle réside dans le formalisme qui passe par des écritures matricielles des estimateurs et de leurs variances. On teste aussi la signification des paramètres et la signification globale du modèle.

Le troisième chapitre est réservé à l'application du modèle de régression linéaire multiple, traité théoriquement dans le chapitre deux où on cherche à établir une relation entre le produit intérieur de brut (PIB) de l'activité économie nationale et l'agriculture, l'industrie et le commerce.
Les données utilisées dans ce mémoire ont été recueille du Cercle d'action et de Réflexion autour de l'Entreprise CARE. Tous les calculs sont effectués à l'aide du logiciel R.

CHAPITRE 1

RÉGRESSION LINÉAIRE SIMPLE

Le présent chapitre trait 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 explicite les hypothèses fondamentales et les notions d'estimations des paramètres du modèle par la méthode de moindres carrés ordinaire, l'estimation par intervalle de confiance. Enfin, tester la signification de paramètres et la signification globale du modèle.

1.1 Modèle de la RLS (Régression linéaire simple)

1.1.1 Définition

La RLS est un acronyme pour le modèle de la régression linéaire simple est défini par :

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad \forall i = \overline{1, n}}$$

où $\boldsymbol{\beta_0}$ et $\boldsymbol{\beta_1}$ sont des paramètres réels inconnus, le $\boldsymbol{\varepsilon_i}$ est l'erreur du modèle, $\boldsymbol{y}$ est une variable de réponse et le $\boldsymbol{x}$ est une régresseur ou variable "explicative".

1.1.1 Remarque

1. Le coefficient $\boldsymbol{\beta_0}$ est appelé aussi l'ordonné à l'origine (intercept ou constant). C'est la valeur prédite de $\boldsymbol{y}$ quand $\boldsymbol{x = 0}$ et le coefficient **1** est appelé la pente. C'est le changement sur $\boldsymbol{y}$ lorsque $\boldsymbol{x}$ change d'une unité.

2. Le terme aléatoire $\boldsymbol{\varepsilon}$ 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 établir entre $\boldsymbol{Y}$ et $\boldsymbol{X}$ c.à.d. résumer le rôle des variables explicatives absentes.

1.1.2 Hypothèses fondamentales du modèle RLS

Les hypothèses relatives du modèle de RLS sont les suivantes :
$\boldsymbol{H1}$: En moyenne les erreurs s'annulent c.à.d le modèle est bien spécifié et de homoscédasticité (variance constante) :

$$\boldsymbol{E(\varepsilon_i) = 0; Var(\varepsilon_i) = \sigma_\varepsilon^2, \qquad \forall i = \overline{1,n}}$$

$\boldsymbol{H2}$: Les erreurs relatives à deux observations sont indépendantes :

$$\boldsymbol{cov(\varepsilon_i;\varepsilon_j) = 0,\ \forall i,j = \overline{1,n}}$$

$\boldsymbol{H3}$: Les $\boldsymbol{\varepsilon_i}$ sont indépendants et identiquement distribués (iid) suivent la loi normale de moyenne nulle et de variance $\boldsymbol{\sigma_\varepsilon^2}$, on écrit $\boldsymbol{\varepsilon_i \sim N(0,\sigma_\varepsilon^2)}$.

1.2 Moindres carrés ordinaires :(MCO)

Pour déterminer le type de liaison pouvant exister entre $\boldsymbol{x}$ et $\boldsymbol{y}$, la première de chose faire la dessination du nuage des points $\boldsymbol{(x_i, y_i)}$ $\boldsymbol{\forall i = \overline{1,n}}$, par exemple celles présentées dans la FIGIRE(1.1) :

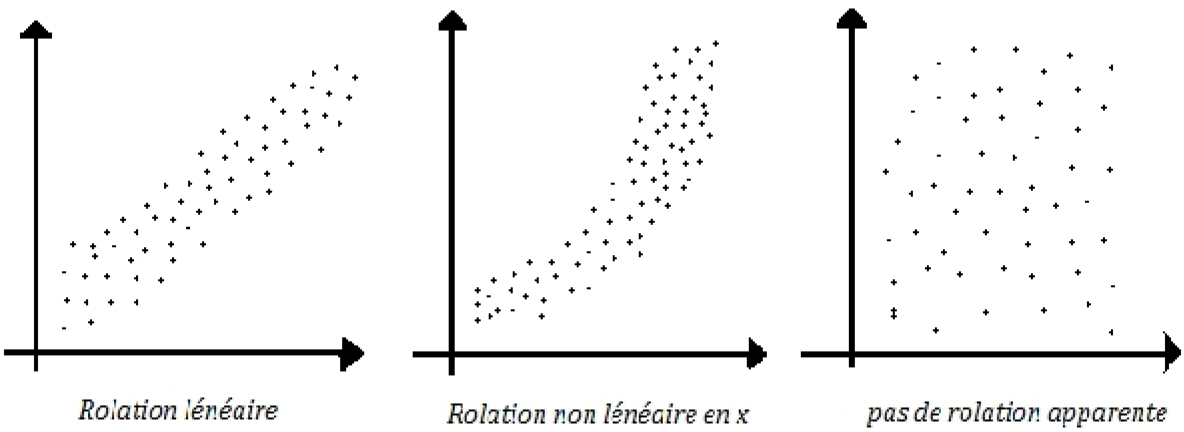


FIGURE 1.1 : Exemple de différentes relation possibles entre x et y

1.2.1 Définition

On appelle estimateurs des moindres carrés ordinaires $\widehat{\boldsymbol{\beta_0}}$ et $\widehat{\boldsymbol{\beta_1}}$ qui réalisent min $(\sum |\widehat{\boldsymbol{\varepsilon_i}}|)$:

$$\boldsymbol{S(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2}$$

où

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta_1} x_i$$

alors

$$S(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

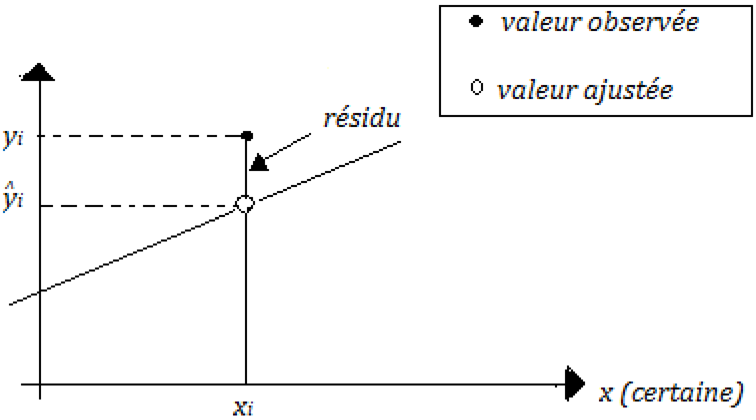


FIGURE 1.2 : Droite et résidu de la régression

1.2.2 Calcul des estimateurs de β_0 et β_1

La fonction de deux variables $S(\widehat{\beta_0}, \widehat{\beta_1})$ est une fonction quadratique et sa minimisation ne pose aucun problème, comme on va le voir maintenant.

Proposition 1.2.1(Estimateurs de MCO)

Les estimateurs β_0 et β_1 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 nous avons

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1 \overline{x}}.$$

preuve (voir[11],[10],[8] et[15])
La valeur de la fonction $S(\widehat{\beta_0}, \widehat{\beta_1})$ et minimum lorsque les dèrivées de S par rapport à $\widehat{\beta_0}$ et $\widehat{\beta_1}$ s'annulent

$$\frac{\partial S(\widehat{\beta_0}, \widehat{\beta_1})}{\partial \widehat{\beta_0}} = 0,$$

et

$$\frac{\partial S(\widehat{\beta_0}, \widehat{\beta_1})}{\partial \widehat{\beta_1}} = 0.$$

Les dérivées par rapport à $\widehat{\beta_0}$ et $\widehat{\beta_1}$ sont :

$$\begin{aligned} \frac{\partial S(\widehat{\beta_0},\widehat{\beta_1})}{\partial \widehat{\beta_0}} = 0 &\Longleftrightarrow -2\sum_{i=1}^{n}(y_i - \widehat{\beta_1}x_i - \widehat{\beta_0}) = 0 \\ &\Longleftrightarrow -2\left[\sum_{i=1}^{n} y_i - \widehat{\beta_1}\sum_{i=1}^{n} x_i - n\widehat{\beta_0}\right] = 0 \end{aligned} \tag{1.1}$$

$$\begin{aligned} \frac{\partial S(\widehat{\beta_0},\widehat{\beta_1})}{\partial \widehat{\beta_1}} = 0 &\Longleftrightarrow -2\sum_{i=1}^{n} x_i(y_i - \widehat{\beta_1}x_i - \widehat{\beta_0}) = 0 \\ &\Longleftrightarrow -2[\sum_{i=1}^{n} x_i y_i - \widehat{\beta_1}\sum_{i=1}^{n} x_i^2 - \widehat{\beta_0}\sum_{i=1}^{n} x_i] = 0. \end{aligned} \tag{1.2}$$

On pose :

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

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

$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x_n}^2$ La variance empirique des x_i.

$S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_n})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i - \overline{y_n}^2$ La variance empirique des y_i.

$S_{xy}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})(y_i - \overline{y_n})$ La covariance empirique entre
$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \overline{x_n y_n}$ les x_i et y_i.

En multipliant l'équation(1.2) par $-\frac{1}{n}$, 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 substituant l'équation(1.3) dans(1.2) ,on a :

$$-2\left(\sum_{i=1}^{n} x_i y_i - \widehat{\beta_1}\sum_{i=1}^{n} x_i^2 - \overline{y}\sum_{i=1}^{n} x_i + \widehat{\beta_1}\overline{x}\sum_{i=1}^{n} x_i\right) = 0. \tag{1.4}$$

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

$$\sum_{i=1}^{n} x_i y_i - \overline{\beta_1}(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2) - n\overline{xy} = 0,$$

donc

$$\begin{aligned} \widehat{\beta_1} &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \\ &= \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \end{aligned}$$

1.2.1 Remarque

1. Le $\widehat{\beta_1}$ il peut être écrit sous la forme :

$$\widehat{\beta_1} = \frac{S_{xy}}{S_x}$$

2. L'erreur de la régression $\widehat{\varepsilon}$ est donnée par :

$$\widehat{\varepsilon_i} = y_i - \widehat{y_i}$$

3. La somme des erreur est nulle dans une régression avec constante :

$$\sum_{i=1}^{n} \widehat{\varepsilon_i} = \sum_{i=1}^{n} [y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)]$$

$$= n\overline{y} - n\widehat{\beta_0} - n\widehat{\beta_1}\overline{x}$$

$$= n\overline{y} - n(\overline{y} - \widehat{\beta_1}\overline{x}) - n\widehat{\beta_1}\overline{x}$$

$$= 0,$$

donc

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

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

$$\widehat{y}(\overline{x}) = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

$$= (\overline{y} - \widehat{\beta_1}\overline{x}) + \widehat{\beta_1} x_i$$

$$= \overline{y}$$

5. le S_ε^2 est l'encodage de l'estimateur de la variance de l'erreur, est donné comme suit :

$$S_\varepsilon^2 = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{\varepsilon_i}^2$$

$$= \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

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

1.2.3 Propriétés des estimateurs

Théorème 1.2.1(Estimateurs sans biais)

Les estimateurs sans biais de β_0 et β_1 sont $\widehat{\beta_0}$ et $\widehat{\beta_1}$:
$E[\widehat{\beta_0}] = \beta_0$ et

$$E[\widehat{\beta_1}] = \beta_1$$

preuve
Soit le modèle de la RLS

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

On peut calculer :

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

Alors

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

En soustraction on trouve

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

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

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})[\beta_1(x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon})]}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

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

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

On a

$$\overline{\varepsilon}\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

Donc

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

Alors

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

$\boldsymbol{x}$ (La variable exogène) n'est pas aléatoire.
Donc :

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

D'après l'hypothèse que $\boldsymbol{E(\varepsilon_i) = 0}$ on obtient

$$\boldsymbol{E(\widehat{\beta_1}) = \beta_1,}$$

pour $\boldsymbol{\widehat{\beta_0}}$, nous avons

$$\boldsymbol{\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x},}$$

d'où l'on tire $\boldsymbol{E[\widehat{\beta_0}] = E[\overline{y}] - \overline{x}E[\widehat{\beta_1}]}$

$$\boldsymbol{= E[\beta_0 + \beta_1\overline{x} + \overline{\varepsilon}] - \overline{x}E[\widehat{\beta_1}]}$$

$$\boldsymbol{= \beta_0 + \beta_1\overline{x} + E[\overline{\varepsilon}] - \overline{x}\beta_1}$$

$$\boldsymbol{= \beta_0 + \overline{\varepsilon} - (\beta_1 - \beta_1)\overline{x}.}$$

Donc

$$\boldsymbol{E(\widehat{\beta_0}) = \beta_0,}$$

on peut aussi exprimer la variance et la covariance entre ces estimateurs.

Théorème 1.2.2 (variances)

Les variances des estimateurs sont :

$$\boldsymbol{var[\widehat{\beta_0}] = s^2_{\widehat{\beta_0}} = \frac{s^2_{\widehat{\varepsilon}}\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2},}$$

et

$$\boldsymbol{var[\widehat{\beta_1}] = s^2_{\widehat{\beta_1}} = \frac{s^2_{\widehat{\varepsilon}}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.}$$

Théorème 1.2.3 (covariance)

Le covariance vaut :

$$\boldsymbol{cov(\widehat{\beta_0, \beta_1}) = -\frac{s^2_{\widehat{\varepsilon}}\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.}$$

1.3 Régions de confiance

Proposition 1.3.1 (Régions de confiance et intervalles)
1/ IC($\boldsymbol{\beta_0}$) : L'intervalle de confiance pour $\boldsymbol{\beta_0}$ est :

$$\left[\widehat{\beta_0} - t_{n-2}^{1-\frac{\alpha}{2}} s_{\widehat{\beta_0}}, \widehat{\beta_0} + t_{n-2}^{1-\frac{\alpha}{2}} s_{\widehat{\beta_0}}\right]$$

ou $t_{n-2}^{1-\frac{\alpha}{7}2}$ le quantile de niveau $(1-\frac{\alpha}{2})$ d'une loi de student T_{n-2}
2/ TC($\boldsymbol{\beta_1}$) : L'intervalle de confiance pour $\boldsymbol{\beta_1}$ est :

$$\left[\widehat{\beta_1} - t_{n-2}^{1-\frac{\alpha}{2}} s_{\widehat{\beta_1}}, \widehat{\beta_1} + t_{n-2}^{1-\frac{\alpha}{2}} s_{\widehat{\beta_1}}\right]$$

3/ RC($\boldsymbol{\beta}$) : Une région de confiance simultanée pour $\boldsymbol{\beta_0}$ et $\boldsymbol{\beta_1}$,au niveau $(\mathbf{1}-\boldsymbol{\alpha})$ est :

$$\frac{1}{2s_{\varepsilon}^2}\left[n(\widehat{\beta_0}-\beta_0) - 2n\overline{x}(\widehat{\beta_0}-\beta_0)(\widehat{\beta_1}-\beta_1) + \sum_{i=1}^{n} x_i^2(\widehat{\beta_1}-\beta_1)^2\right]$$

$$\leqslant F_{2,n-2}^{1-\alpha}$$

où $F_{2,n-2}^{1-\alpha}$ le quantile de niveau $(\mathbf{1}-\boldsymbol{\alpha})$ d'un loi $F_{2,n-2}$.
4/ Un intervalle de confiance de $\boldsymbol{\sigma_{\varepsilon}^2}$ est donné par :

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

où $C_{n-2}^{1-\frac{\alpha}{2}}$ est la quantile de niveau $(\mathbf{1}-\frac{\alpha}{2})$ d'un loi Chi-2 χ_{n-2}^2 .

1.4 Analyse de la variance et coefficient de détermination

1.4.1 Décomposition de la variance et tableau d'ANOVA

On décompose l'écart entre $\boldsymbol{y_i}$ et la moyenne des $\boldsymbol{y_i}$ en un point d'observation $(\boldsymbol{x_i}, \boldsymbol{y_i})$, en ajoutant puis retranchant $\widehat{\boldsymbol{y}}$ la valeur estimée de $\boldsymbol{y}$ par la droite de régression. Cette procédure fait apparaitre une somme de deux écarts :

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

Ainsi l'écart total $(\boldsymbol{y_i} - \overline{\boldsymbol{y}})$ peut être vu comme la somme de deux écarts : un écart entre $\boldsymbol{y_i}$ observé et $\widehat{\boldsymbol{y_i}}$ la valeur estimée par le modèle et un écart entre $\widehat{\boldsymbol{y_i}}$ la valeur estimé par le modèle et la moyenne $\overline{\boldsymbol{y}}$.

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

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}\left[(y_i - \widehat{y_i}) + (\widehat{y_i} - \overline{y})\right]^2$$

$$= \left[\sum_{i=1}^{n}(\widehat{y_i} - \overline{y}) + \widehat{\varepsilon_i}\right]^2$$

$$\sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2 + \sum_{i=1}^{n}\widehat{\varepsilon_i}^2 + 2\sum_{i=1}^{n}(\widehat{y_i} - \overline{y})\widehat{\varepsilon_i}.$$

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

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

A l'origine, on a :

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

$$= \widehat{\beta}_1\sum_{i=1}^{n}x_i\widehat{\varepsilon}_i + \widehat{\beta}_0\sum_{i=1}^{n}\widehat{\varepsilon}_i - \overline{y}\sum_{i=1}^{n}\widehat{\varepsilon}_i$$

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 obtient enfin à l'égalité fondamentale (l'équation d'analyse de la variance) :

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

Alors

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

D'où :

$$SCT = SCE + SCR$$

* $\sum_{i=1}^{n}(y_i - \overline{y})^2$: est la somme des carré totaux où elle indique la variabilité totale de y c.à.d l'information disponible dans les données(noté par SCT)
* $\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$: 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 (noté par SCE)
* $\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$: est la somme des carrés résiduels. Elle écrite la variabilité non expliquée (résiduelle) par le modèle (l'écart entre y et $\widehat{y}$)(noté par SCR)

1.4.1 Remarque

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 : $\boldsymbol{X}$ n'apporte aucune information sur $\boldsymbol{Y}$.
On produire du tableau danalyse de variance (voir TABLE(1.1)) à partir de la décomposition de la variance, comme suit :

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

TABLE 1.1 : Tableau danalyse de variance de la régression linéaire simple

1.4.2 Coefficient de détermination $\boldsymbol{R^2}$

Le coefficient de détermination, noté $\boldsymbol{R^2}$ donné par :

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

Théorème 1.4.1

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

preuve
d'après l'équation, on a :

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

$$\boldsymbol{\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} = \frac{\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} + \frac{\sum_{i=1}^{n}(y_i-\widehat{y_i})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

La quantité

$$\boldsymbol{R^2 = \frac{\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2},}$$

est appelée la coefficient de détermination, alors :

$$\boldsymbol{1 = R^2 + \frac{\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}\widehat{\varepsilon_i}^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

- Relation entre les degrés de liberté l'additivité des degrés de liberté nous donne la relation suivante :

$$\boldsymbol{dlpourSCT = dlpourSCE + dlpourSCR}$$

$$\boldsymbol{(n-1) = 1 + (n+2)}$$

1.4.2 Remarque

1. Ce coefficient$\boldsymbol{R^2}$ qui varie entre 0 et 1, mesure la proportion de variation totale de $\boldsymbol{Y}$ autour de la moyenne expliquée par la régression, c.à.d. prise en compte par le modèle.
2. Plus $\boldsymbol{R^2}$ se rapproche de la valeur 1, meilleure est l'adéquation du modèle aux données et un $\boldsymbol{R^2}$ 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 globale de la variable $\boldsymbol{X}$ à la détermination de $\boldsymbol{Y}$. On veut tester l'hypothèse :

$$\begin{cases} \boldsymbol{H_0 : \beta_1 = 0} \\ \boldsymbol{H_1 : \beta_1 \neq 0} \end{cases} \tag{1.5}$$

pour tester cette hypothèse, on a basé sur la statistique de Fisher, notée par $\boldsymbol{F}$:

$$\frac{\frac{SCE}{1}}{\frac{SCR}{(n-2)}} \tag{1.6}$$

Cette statistique indique si la variance expliquée est significativement supérieure 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 $\boldsymbol{H_0}$, SCE est distribué selon un $\boldsymbol{\chi^2(1)}$ et SCR selon $\boldsymbol{\chi^2(n-2)}$, de fait pour F, on a :

$$\boldsymbol{F \equiv \frac{\frac{\chi^2(1)}{1}}{\frac{\chi^2(n-2)}{(n-2)}} = f_{1,n-2}^{1-\alpha}} \tag{1.7}$$

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

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

avec $\boldsymbol{F \geq f_{1,n-2}^{1-\alpha}}$ est le quantile d'ordre $\boldsymbol{1-\alpha}$ d'un loi de Fisher à (1,n-2)

1.5.1 Remarque

*/ On peut réécrire la statistique F en fonction de $\boldsymbol{R^2}$ comme suit :

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

*/ Dans la plupart des logiciels statistique, on fournit directement la probabilité critique. Elle correspond à la probabilité que la loi de Fisher dépasse la statistique calculée $\boldsymbol{F}$. Ainsi,

la règle de décision (rejette $\boldsymbol{H_0}$) au risque devient :

$$\boldsymbol{P - value < \alpha}$$

1.5.2 Test de signification des paramètres

Test de signification de $\boldsymbol{\beta_0}$

On veut tester l'hypothèse :

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

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

$$\boldsymbol{T_{\widehat{\beta_0}} = \frac{\widehat{\beta_0}}{\delta_{\widehat{\beta_0}}}},$$

on rejette $\boldsymbol{H_0}$ si l'observation de la statistique de test notée $\boldsymbol{T_{\widehat{\beta_0}}}$, est telle que :

$$\boldsymbol{|T_{\widehat{\beta_0}}| \geqslant t_{n-2}^{1-\frac{\alpha}{2}}},$$

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

Test de signification de $\boldsymbol{\beta_1}$

Le test de significativité de la pente consiste à vérifier l'exogène $\boldsymbol{X}$ sur l'endogène $\boldsymbol{Y}$, l'hypothèse à confronter s'écrive :

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

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

$$\boldsymbol{T_{\widehat{\beta_1}} = \frac{\widehat{\beta_1}}{\sigma_{\widehat{\beta_1}}}},$$

on rejette $\boldsymbol{H_0}$ si l'observation de la statistique de teste, notée $\boldsymbol{T_{\widehat{\beta_1}}}$, est telle que :

$$\boldsymbol{|T_{\widehat{\beta_1}}| \geqslant t_{n-2}^{1-\frac{\alpha}{2}}},$$

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

1.6 Prévision

Les buts de la régression est de faire la prévision, c.à.d. de prévoir la variable expliquer $\boldsymbol{y}$ en présence d'une nouvelle valeur de la variable explicative $\boldsymbol{y}$.
Soit $\boldsymbol{x_{n+1}}$ une nouvelle valeur, pour laquelle on veut prédire $\boldsymbol{y_{n+1}}$. Le modèle est toujours le même :

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

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

$$\boldsymbol{\widehat{y}_{n+1} = \widehat{\beta_0} + \widehat{\beta_1} x_{n+1}.}$$

En matière de prévision dans la cas d'erreurs gaussiennes, les résultats obtenus c'est pour l'espérance et la variance sont toujours valables. De plus, puisque $\boldsymbol{\widehat{y}_{n+1}}$ est linéaire en $\boldsymbol{\beta_0}$ est $\boldsymbol{\beta_1}$, $\boldsymbol{\varepsilon_{n+1}}$ on peut préciser sa loi :

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

on ne connaît pas $\boldsymbol{\sigma^2_\varepsilon}$, on l'estime par $\boldsymbol{s^2_{\widehat{\varepsilon}}}$. Comme $\boldsymbol{(y_{n+1} - \widehat{y}_{n+1})}$ et $\boldsymbol{\frac{s^2_{\widehat{\varepsilon}}(n-2)}{\sigma^2_\varepsilon}}$ sont indépendants, on peut énoncer un résultat donnant des intervalles de confiance pour $\boldsymbol{y_{n+1}}$. Avec les notation et hypothèse précédentes, on a :

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

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

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

CHAPITRE 2

RÉGRESSION LINÉAIRE MULTIPLE (RLM)

Ce chapitre traite le modèle de la régression linéaire multiple (RLM) qui est une des méthodes statistique les plus habituellement mise en œuvre pour l'étude des données multidimensionnelles, où c'est expliqué les valeurs prises par la variable endogène Y à l'aide de $\boldsymbol{p}$ variables exogènes $(\boldsymbol{X_1, ..., X_p})$. Comme nous trouvons dans ce chapitre la plupart des éléments de basé présentés dans le précédent et la justification de certains d'entre eux : hypothèses du modèle, la forme matricielle de la régression linéaire par la méthode des moindres carrés ordinaire, l'estimation par intervalle de confiance et aussi on teste la signification des paramètres et enfin la signification globale du modèle la prévision.

2.1 Modèle de la RLM (Régression linéaire multiple)

2.1.1 Définition (modèle de la RLM)

La RLM est un acronyme pour le modèle de régression linéaire multiple est défini par :

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \qquad \forall i = \overline{1, n}} \tag{2.1}$$

où $\boldsymbol{\beta_0, \beta_1, \cdots, \beta_p}$ sont des coefficients réels inconnus, $\boldsymbol{\varepsilon_i}$ est l'erreur du modèle (bruit), $\boldsymbol{y_i}$ est la i-ème observation de la variable $\boldsymbol{y}$ et le $\boldsymbol{x_{ij}}$ est la j-ème observation de la i-ème variable.

2.2 Notation matricielle

La forme matricielle de cet modèle est :

$$\begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_n} \end{pmatrix} = \begin{pmatrix} \boldsymbol{1} & \boldsymbol{x_{11}} & . & . & . & \boldsymbol{x_{1p}} \\ \boldsymbol{1} & \boldsymbol{x_{21}} & . & . & . & \boldsymbol{x_{2p}} \\ \vdots & & & & & \\ \boldsymbol{1} & \boldsymbol{x_{n1}} & . & . & . & \boldsymbol{x_{np}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta_0} \\ \boldsymbol{\beta_1} \\ \vdots \\ \boldsymbol{\beta_p} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_n} \end{pmatrix} \tag{2.2}$$

on écrit de façon équivalente :

$$\boldsymbol{Y = X\beta + \varepsilon}$$

il est tiré de la forme des équations suivants :

$$\begin{cases} \boldsymbol{y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_p x_{1p} + \varepsilon_1} \\ \boldsymbol{y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_p x_{2p} + \varepsilon_2} \\ \vdots \qquad \vdots \qquad \qquad \qquad \qquad \vdots \\ \boldsymbol{y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_p x_{np} + \varepsilon_n} \end{cases}$$

Avec :
* Y est le vecteur à expliquer de taille $\boldsymbol{(n \times 1)}$
* X est la matrice de taille $\boldsymbol{n \times (p+1)}$ qui contient l'ensemble de observation sur les exogènes
* $\boldsymbol{\varepsilon}$ le vecteur des erreurs de taille $\boldsymbol{(n \times 1)}$

Remarque peur X : la première colonne formée par la valeur 1 indiquant que l'on intègre la constante $\boldsymbol{\beta_0}$ dans l'équation en pétant le nombre des variables explicatives réelles.

2.2.1 Remarque

1/ Le coefficient $\boldsymbol{\beta_0}$ est un paramètre appelé intercepte qui représente la moyenne des $\boldsymbol{y_i}$ lorsque la valeur de chaque variable explicative est égale à 0.

2/ Les coefficients $\boldsymbol{\beta_j}$, $\forall \boldsymbol{j = \overline{1, p}}$ représentent le changement subi par $\boldsymbol{E(y_i)}$ correspondant à un changement unitaire dans la valeur de la j-ème variable explicative, lorsque les autres variables explicatives demeurent inchangées.

2.3 Hypothèses relatives du modèle de la RLM

le modèle existe principalement deux catégories d'hypothèses (voir Bourbonnais[4], Labrousse[19]).

2.3.1 Hypothèses stochastiques

,

* $\boldsymbol{H_1 : E(\varepsilon_i) = 0)} \quad \forall \boldsymbol{i} = \overline{\boldsymbol{1, n}}$: l'ensemble des déterminants de $\boldsymbol{y}$ qui sont pas été retenus dans le modèle est d'espérance nulle .
* $\boldsymbol{H_2}$ **:** les valeurs $\boldsymbol{x_{ij}}$ sont observées sans erreurs.
* $\boldsymbol{H_3 : E(\varepsilon_i) = var(\varepsilon_i) = \sigma_\varepsilon^2,} \quad \forall \boldsymbol{i} = \overline{\boldsymbol{1, n}}$: la variance des erreurs est constante.
* $\boldsymbol{H_4 : cov(\varepsilon_i, \varepsilon_j) = 0} \quad \forall \boldsymbol{i \neq j}$: on dit n'y a pas de corrélation sérielle et les erreurs relatives termes aléatoires ne sont pas corrélés.
* $\boldsymbol{H_5 : \varepsilon_i \sim N(0, \sigma_\varepsilon^2)}$: les erreurs suivent une loi normale multidimensionnelle.
* $\boldsymbol{H_6}$ implique les hypothèses $\boldsymbol{H_1, H_3}$ et $\boldsymbol{H_4}$

2.4 Estimateur par MCO

2.4.1 L'estimation par MCO de $\boldsymbol{\beta}$

Proposition 2.4.1

L'estimation de paramètre $\boldsymbol{\beta}$ est donnée par :

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

preuve

On voit que SCR est une fonction de $\boldsymbol{\beta_0}$. On le choisira de façon à minimise SCR. Le minimum de SCR est atteint lorsque la dérivée de SCR par rapport à $\widehat{\boldsymbol{\beta}}$ s'annule, donc :

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{argmin} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$$

$$= \boldsymbol{argmin}(\widehat{\varepsilon}^t \widehat{\varepsilon})$$

$$= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^t)(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})$$

$$= (\boldsymbol{Y}^t - \widehat{\beta}^t \boldsymbol{X}^t)(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}),$$

car : $(\boldsymbol{X}\widehat{\beta})^t = \widehat{\beta}\boldsymbol{X}^t$

$$= \boldsymbol{Y^t Y} - \boldsymbol{Y^t X}\widehat{\beta} - \boldsymbol{X}^t\widehat{\beta}^t\boldsymbol{Y} + \boldsymbol{X}^t\widehat{\beta}^t\boldsymbol{X}\widehat{\beta}$$

$$= \boldsymbol{Y^t Y} - \boldsymbol{2X}^t\widehat{\beta}^t\boldsymbol{Y} + \boldsymbol{X}^t\widehat{\beta}^t\boldsymbol{X}\widehat{\beta},$$

car : $\boldsymbol{X}^t\widehat{\beta}^t\boldsymbol{Y}$est un scalaire, il est égal à sa transposée

$$= \boldsymbol{Y^t Y} - \boldsymbol{2X}^t\widehat{\beta}^t\boldsymbol{Y} + \boldsymbol{X}^t\widehat{\beta}^t\boldsymbol{X}\widehat{\beta},$$

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

$$\frac{\partial \boldsymbol{SCR}}{\partial \boldsymbol{\beta}} = \mathbf{0.}$$

2.4.2 Propriétés des estimateurs MCO

Nous présentons quelques propriétés des estimateurs :

Estimateur sans biais

Théorème(Propriétés des estimateurs MCO)

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

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

Remarque 2.4.1

La matrice de la variance-covariance ($\boldsymbol{varcov}(\widehat{\boldsymbol{\beta}})$) est symétrique sur sa diagonale principale on observe les variance des coefficients estimés $(\boldsymbol{var}(\widehat{\beta}_0), ..., \boldsymbol{var}(\widehat{\beta}_0))$

preuve (voir[1]**)**

Montre que $\boldsymbol{E}(\widehat{\beta}) = \boldsymbol{\beta}$ et $\boldsymbol{varcov}(\widehat{\beta}) = \boldsymbol{\sigma_{\varepsilon}^2}(\boldsymbol{X^tX})^{-1}$:

1/ $\boldsymbol{E}(\widehat{\beta}) = \boldsymbol{\beta}$

soit :

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

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

$$\begin{aligned}\boldsymbol{E}(\widehat{\beta}) &= \boldsymbol{E}[\beta + (\boldsymbol{X^tX})^{-1}\boldsymbol{X^t}\varepsilon]\\ &= \beta + \boldsymbol{E}[(\boldsymbol{X^tX})^{-1}\boldsymbol{X^t}\varepsilon]\\ &= \beta + (\boldsymbol{X^tX})^{-1}\boldsymbol{X^tE}[\varepsilon],\end{aligned}$$

car X est non aléatoire

et sous l'hypothèse que $\boldsymbol{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, il vient que :

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

2/ $\boldsymbol{varcov(\widehat{\beta}) = \sigma_\varepsilon^2(X^tX)^{-1}}$
on procède de même, on a $\boldsymbol{\widehat{\beta} = \beta + (X^tX)^{-1}X^t}$, donc :

$$\begin{aligned}\boldsymbol{varcov(\widehat{\beta})} &= \boldsymbol{var(\beta + (X^tX)^{-1}X^t)\varepsilon} \\ &= \boldsymbol{(X^tX)^{-1}X^t var(\varepsilon) X(X^tX)^{-1}} \\ &= \boldsymbol{(X^tX)^{-1}X^t\sigma_\varepsilon^2 I_n X(X^tX)^{-1}} \\ &= \boldsymbol{\sigma_\varepsilon^2(X^tX)^{-1}X^tX(X^tX)^{-1}} \\ &= \boldsymbol{\sigma_\varepsilon^2(X^tX)^{-1}},\end{aligned}$$

on trouve :

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

Estimateur convergent
Théorème(Gauss-Markov)
L'estimateure $\boldsymbol{\widehat{a} = (X'X)^{-1}X'Y}$ des moindres carrés est qualifié de Best Linear Unbiaised Estimator (BLUE), car il s'agit du meilleur estimateur linéaire sans biais (au sens qu'il fournit les variance les plus faibles pour les estimateurs).

2.5 L'intervalle de confiance

Proposition 2.5.1 (Régions de confiance et intervalles)
1/ Un intervalle de confiance de niveau $(\boldsymbol{1-\alpha})$ pour $\boldsymbol{\beta_j}$ $\forall \boldsymbol{j = \overline{1,p}}$ est :

$$\left[\boldsymbol{\widehat{\beta_j} - t_{n-p-1}^{1-\frac{\alpha}{2}} s_{\widehat{\beta_j}}, \widehat{\beta_j} + t_{n-p-1}^{1-\frac{\alpha}{2}} s_{\widehat{\beta_j}}}\right],$$

où $\boldsymbol{t_{n-p-1}^{1-\frac{\alpha}{2}}}$ est le quantile d'ordre $(\boldsymbol{1-\frac{\alpha}{2}})$ d'une loi Student $\boldsymbol{T_{n-p-1}}$.
2/ Un intervalle de confiance de niveau $(\boldsymbol{1-\alpha})$ pour $\boldsymbol{\sigma_\varepsilon^2}$ est :

$$\left[\boldsymbol{\frac{(n-p-1)s_\varepsilon^2}{C_{n-p-1}^{1-\frac{\alpha}{2}}}, \frac{(n-p-1)s_\varepsilon^2}{C_{n-p-1}^{\frac{\alpha}{2}}}}\right],$$

où $\boldsymbol{C_{n-p-1}^{1-\frac{\alpha}{2}}}$ est le quantile d'ordre $\boldsymbol{1-\frac{\alpha}{2}}$ d'une loi $\boldsymbol{\chi_{n-p-1}}$.
3/ Soit R une matrice de taille $\boldsymbol{q \times (p+1)}$ de rang $\boldsymbol{q^1(q < (p+1))}$ alors :

$$\boldsymbol{\frac{1}{qs_{\widehat{\varepsilon}}^2}(R(\widehat{\beta} - \beta))^t[R(X^tX)^{-1}R^t]^{-1}R(X^tX)^{-1}F_{(q,n-p-1)}^{1-\alpha}},$$

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

2.6 Analyse de la variance et coefficient de détermination

2.6.1 Décomposition de la variance et tableau d'ANOVA

Nous pouvons facilement vérifier cette décomposition d'une manière similaire à celle effectuée pour la régression linéaire simple, l'écart entre $\boldsymbol{y_i}$ et la moyenne des $\boldsymbol{y_i}$ en ajoutant puis retranchant $\widehat{\boldsymbol{y}}$ la valeur estimée de y par la droit de régression. Cette équation fait apparaître une somme de deux écarts :

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

on peut obtenir la décomposition :

$$\boldsymbol{\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

$$\boldsymbol{= \sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2 + \sum_{i=1}^{n}\widehat{\varepsilon_i}}$$

$$\boldsymbol{SCT = SCE + SCR,}$$

où

* $\boldsymbol{\sum_{i=1}^{n}(y_i - \overline{y})^2}$: désigne la somme de carrés totaux (centrés)(noté par SCT)
* $\boldsymbol{\sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2}$: la somme des carré expliquée(centrés) (noté par SCE)
* $\boldsymbol{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$: $(\boldsymbol{\sum_{i=1}^{n}\widehat{\varepsilon_i}})$ la somme des carrés des résidus (noté par SCR)

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

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

TABLE 2.1 : tableau d'analyse de la variance de la régression linéaire multiple

2.6.1 Remarque

On peut récrire l'équation d'ANOVA matricielle comme suit

$$\boldsymbol{(y - \overline{y})^t(y - \overline{y}) = (\widehat{y} - \overline{y})^t(\widehat{y} - \overline{y}) + \widehat{\varepsilon}^t\widehat{\varepsilon},}$$

SCT : $\boldsymbol{(y - \overline{y})^t(y - \overline{y})}$
SCE : $\boldsymbol{(\widehat{y} - \overline{y})^t(\widehat{y} - \overline{y})}$
SCR : $\boldsymbol{\widehat{\varepsilon}^t\widehat{\varepsilon}}$

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

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

2.6.2 Coefficient de détermination R^2

Le coefficient de détermination est un rapport entre SCE et SCT, noté par R^2 (voir[12])

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(\widehat{y_i} - \hat{y})^2}{\sum_{i=1}^{n}(y_i - overliney)^2}$$

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

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.

2.6.2 Remarque

Le R^2 ne permet de comparer que des modèles ayant le même nombre de variables explicatives, le même nombre d'observation et la même forme (on ne peut pas comparer un modèle simple avec un modèle en log).

2.6.3 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 définie comme suit :

$$R^2_{adj} = 1 - \frac{\frac{SCR}{(n-p-1)}}{\frac{SCT}{(n-1)}} = 1 - \frac{(n-1)}{(n-p-1)}(1 - R^2). \tag{2.3}$$

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

2.7 Test de signification

2.7.1 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 $\boldsymbol{X_j} \quad \forall \boldsymbol{j} = \overline{\boldsymbol{1,p}}$ (voir[11]). On considère les hypothèses :

$$\begin{cases} \boldsymbol{H_0} : \boldsymbol{\beta_0 = \beta_1 = ... = \beta_p = 0} \\ \boldsymbol{H_1} : \boldsymbol{\exists \beta_j \neq 0, (\forall j = \overline{1,p})} \end{cases} \tag{2.4}$$

Pour tester l'hypothèse, on a vu que l'on peut utiliser la statistique de Fisher $\boldsymbol{F}$:

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

où $\boldsymbol{F}$ suit une loi de Fisher avec p et $\boldsymbol{(n-p-1)}$ dégrée de liberté, on rejette $\boldsymbol{H_0}$ si :

$$\boldsymbol{F \geq f^{1-\alpha}_{p,n-p-1}}$$

avec $\boldsymbol{f^{1-\alpha}_{p,n-p-1}}$ est le quantile d'ordre $\boldsymbol{(1-\alpha)}$ d'un loi de Fisher à $\boldsymbol{(p, n-p-1)}$ ddl (voir[2])
2.7.1 Remarque :
Il existe une relation mathématique entre le $\boldsymbol{R^2}$ et la statistique de test de signification globale (du $\boldsymbol{F}$ de Fisher) comme suit :

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

2.7.2 Test T de signification du paramètre du modèle (Test de Student)

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

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

où $\boldsymbol{\beta_j}$ est le paramètre associé à la variable explicative $\boldsymbol{X_j}$.
* L'hypothèse $\boldsymbol{H_0}$ de nullité d'un paramètre du modèle peut être testée au moyen de la statistique du Student :

$$\boldsymbol{T_{\widehat{\beta}_j}} = \frac{\boldsymbol{\widehat{\beta}_j}}{\boldsymbol{s_{\widehat{\beta}_j}}} \tag{2.5}$$

à comparer $\boldsymbol{t_{n-p-}^{1-\frac{\alpha}{2}}}$ où $\boldsymbol{t_{n-p-1}^{1-\frac{\alpha}{2}}}$ est le quantile d'ordre $\boldsymbol{(1-\frac{\alpha}{2})}$ d'une loi Student à $\boldsymbol{(n-p-1)}$ ddl.
1 * Si $\boldsymbol{|T_{\widehat{\beta}_j}| \geq t_{n-p-1}^{1-\frac{\alpha}{2}}}$, on rejette $\boldsymbol{H_0}$.
2 * Si $\boldsymbol{|T_{\widehat{\beta}_j}| < t_{n-p-1}^{1-\frac{\alpha}{2}}}$, on ne peut pas rejeter $\boldsymbol{H_0}$.

2.8 Prévision

Comme dans le cadre du modèle de régression linéaire multiple, le but de la régression est de proposer des prédictions pour la variable à expliquer Y lorsque on a de nouvelles valeurs (voir[9])

$$\boldsymbol{x_{n+1} = (x_{n+1,1}, .., x_{n+1,p})^t}.$$

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

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

avec $\boldsymbol{\varepsilon_{n+1} \sim N(0, \sigma_\varepsilon^2)}$ indépendant des $\boldsymbol{(\varepsilon_i)_{(1<i<n)}}$
A partir des n observations précédentes, on a pu calculer un estimateur $\boldsymbol{\widehat{\beta}}$ de $\boldsymbol{\beta}$, donc, il est naturel de prédire la valeur correspondante via le modèle ajusté $\boldsymbol{\widehat{y}_{n+1}}$ définie par :

$$\boldsymbol{\widehat{y}_{n+1} = \widehat{x}_{n+1}^t \widehat{\beta}}.$$

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

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

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

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

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

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

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

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

CHAPITRE 3

APPLICATION DE LA RLM SUR L'ÉCONOMIE NATIONALE

Dans ce chapitre, on cherche à établir une relation entre l'économie nationale et l'agriculture, l'industrie ainsi que la commerce. on mentionne que tous les travaux, présentés dans ce chapitre, sont traités à l'aide du logiciel R.

3.1 Définitions

3.1.1 Économie nationale

L'économie algérienne est devenue l'une des économies les plus importantes du monde parce que l'Algérie est riches en richesses telles que le pétrole, le gaz, les minéraux et d'autres ressources naturelles indépendamment d'être rentier, il dépend en grande partie des hydrocarbures. La valeur totale annuelle de l'économie nationale est déterminée en calculant le produit intérieur brut en attribuant les valeurs de l'agriculture, l'industrie, le commerce et d'autres ressources.

3.1.2 PIB : Produit intérieur brut

Le produit intérieur brut est l'indicateur économique qui permet de quantifier la valeur de la "production de richesse" annuelle effectuée par les agents économiques résidant à l'intérieur d'un territoire.

3.1.3 Agriculture

L'agriculture où l'élevage est le processus de production d'aliment, de fourrage de fibre et d'autre produits grâce à la culture systématique de plante. L'activité du secteur agricole algérien est également importante dans l'économie.

3.1.4 Industrie

C'est le processus de transformation de la forme des matières dans la nature pour augmenter leur valeur, le secteur industriel est un pilier important du développement long terme de l'économie. C'est l'un des secteurs les plus importants pour diversifier les sources de revenu national.

3.1.5 Commerce

C'est un ensemble d'opérations commerciales qui comprennent l'achat et la vente de services et de biens. Le commerce est défini comme l'échange d'une transaction commerciale basée sur la vente où l'achat des produits ou matériaux spécifiques. La balance commerciale est un indicateur économique important c'est l'une des entrées du PIB (le produit intérieur brut) des pays.

3.2 Présentation des données

On souhaite expliquer une variable quantitative Y (PIB) en fonction de 3 autres variables X $\boldsymbol{(X_1, X_2, X_3)}$ dans lequel $\boldsymbol{X_1}$ (Agriculture), $\boldsymbol{X_2}$ (Industrie) et $\boldsymbol{X_3}$ (Commerce).
On dispose 22 données relevées durant la période (de 2000 à 2021), concernant l'économie algériènne. Ces données sont présentées sous forme d'un tableau des données (voir TABLE 3.1).

année	agriculture	industrie	commerce	PIB
2000	346171.4	290749.6	436292.1	3698683.6
2001	412119.5	315230.5	476208.7	3754870.7
2002	417225.2	33755602	509285.7	4023413.8
2003	515281.7	355370.6	552179.9	4700040.5
2004	580505.6	388193.4	607052.6	5545851.4
2005	581615.8	418294.9	668130.0	6930153.4
2006	641285.0	449581.0	728366.7	7823794.6
2007	708072.5	479791.1	863197.3	8554266.0
2008	727413.1	519631.6	1003199.4	9968908.7
2009	931349.1	570673.2	1160160.0	8770806.4
2010	1015258.8	617404.9	1283227.7	10404470.8
2011	1183216.1	664194.5	1446331.4	12211018.0
2012	1421693.3	729514.8	1649969.8	13561457.2
2013	1640006.1	771787.4	1870581.0	14096723.4
2014	1772202.4	837716.8	2067543.0	13597697.5
2015	1935113.0	919370.4	2259343.2	13812762.7
2016	2140304.7	979303.0	2341306.0	14455022.2
2017	2219064.4	1044920.1	2116090.1	15503775.1
2018	2426906.9	1127981.6	2349598.7	17252567.0
2019	2529100.0	1198400.0	2481200.0	17267700.0
2020	2598511.9	1153521.0	1987214.2	15024828.8
2021	2869600.0	1272599.0	2301500.0	18549099.0

Table 3.1 : PIB en Millions de Dinars Courants

3.3 Modèle de la RLM

L'objectif principal est d'expliquer l'économie nationale en l'Algérie (PIB) en fonction des variables explicatives : agriculture en DA ($\boldsymbol{X_1}$), l'industrie en DA ($\boldsymbol{X_2}$) et le commerce en DA ($\boldsymbol{X_3}$).

Alors, le modèle de la RLM avec 3 variables explicatives est :

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

où
* $\boldsymbol{\beta_0, \beta_1, \beta_2}$ et $\boldsymbol{\beta_3}$ sont des coefficients du cette modèle.
* $\boldsymbol{x_{i1}, x_{i2}}$ et $\boldsymbol{x_{i3}}$ les i-ème valeurs de les variables $\boldsymbol{X_1, X_2}$ et $\boldsymbol{X_3}$ successivement ($\boldsymbol{X_1}$ (Agriculture), $\boldsymbol{X_2}$ (Industrie), $\boldsymbol{X_3}$ (Commerce)).
* $\boldsymbol{y_i}$: la i-ème valeur de la variable Y (Économie nationale).
* $\boldsymbol{\varepsilon_i}$: l'erreur du modèle.
notre but ici va être de déterminer :

1 * La valeur de la constante $\boldsymbol{\beta_0}$ et des différents coefficients $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, et $\boldsymbol{\beta_3}$ qui permettent de minimiser l'erreur entre la droite de régression linéaire estimée et les valeurs réelles de Y.
2 * Les variables significative, c.à.d voir si ces différents coefficient sont différents de 0 ou non et la précision de notre modèle, en utilisant entre autre, le coefficient de détermination "R-squared".
Donc : on modélise, sous R, le modèle de RLM par la fonction suivante :

$\boxed{\boldsymbol{> Reg < -lm(Y \sim X1 + X2 + X3)}}$

3.4 Hypothèses relatives de modèle

Résidus

On appelle i-ème résidu pour tout $\boldsymbol{i \in \{1, 22\}}$. La réalisation $\boldsymbol{\varepsilon_i}$ de :

$$\boldsymbol{\varepsilon_i = \widehat{\varepsilon}_i = y_i - \widehat{y}_i}$$

Sous R, avec la commande suivante :

$\boxed{> \text{residuals(Reg)}}$

On obtient, le résultat suivant (voir FIGURE 3.1)

```
> residuals(Reg)
          1           2           3           4           5           6 
 -533754.51  -762191.09 -1133810.55  -356723.14   -62773.92   358370.65 
          7           8           9          10          11          12 
  679818.57   649354.50   705355.01  -875757.88  -252172.58   999210.27 
         13          14          15          16          17          18 
 1689088.91  1935874.28    26994.34 -1323641.94 -1009969.54  -236481.92 
         19          20          21          22 
   82843.18 -1404677.73  -265692.44  1090737.53 
```

FIGURE 3.1 : Résidus de la RLM

3.4.1 Remarque ces résidus vont nous permettre de valider ou non les hypothèses initiales du modèle de la RLM.

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

Pour étudier l'indépendance de $\varepsilon_1, ..., \varepsilon_n$, on calcule l'intervalle de confiance pour (h). Si les bâtons sont de taille et de signes alternés et qu'aucun d'entre eux ne dépassent les bornes de l'intervalle de confiance, on admet l'indépendance de $\varepsilon_1, ..., \varepsilon_n$.
Sous R,(voir FIGURE(3.2))

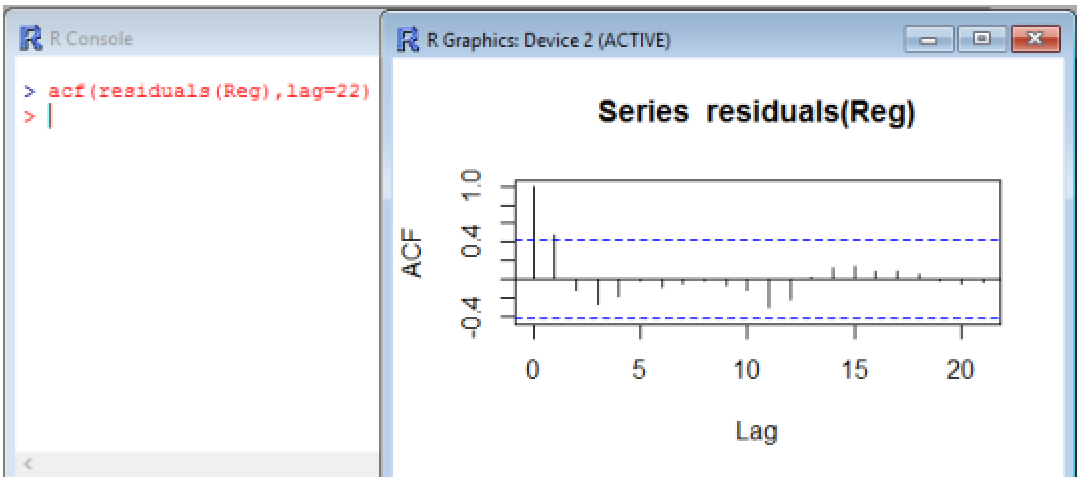


FIGURE 3.2 : Fonction acf des résidus

D'après, cette figure(voir la FIGURE(3.2)) on observe que $\varepsilon_1, ..., \varepsilon_n$ sont indépendants.

Test Ljung-Box

La statistique de Ljung-Box permet de tester l'hypothèse que les n coefficients d'auto-corrélation 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.
Alors, on considère les hypothèses :
$\boldsymbol{H_0 : \rho(1) = ... = \rho(n) = 0}$
contre
$\boldsymbol{H_1}$ **:** au moins une corrélation n'est pas nulle.
On peut utiliser le teste de Ljung-Box partant des résidus $\varepsilon_1, ..., \varepsilon_n$: si $\boldsymbol{p - value > \alpha = 0.05}$, on admet que $\boldsymbol{H_0 : \rho(1) = ... = \rho(n) = 0}$, donc que $\varepsilon_1, ..., \varepsilon_n$ sont indépendants.
Sous R, on écrit les commandes :

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

Box-Ljung test
data : residuals(Reg)
X-squared = 5.6451, df = 1, p-value = 0.0175
p-value = 0.0175 < $\boldsymbol{\alpha}$ = 0.05, on accepte $\boldsymbol{H_1}$ **:** ou moins une corrélation n'est pas nulle, donc les résidus$\varepsilon_1, ..., \varepsilon_n$ sent corrélés(c.à.d ils ne sont pas indépendants).

3.4.2 Test de normalité des erreurs

pour étudier la normalité de $\varepsilon_1, ..., \varepsilon_n$, on admet d'abord que $\varepsilon_1, ..., \varepsilon_n$ soient indépendantes et $\sigma^2(\varepsilon_1), ..., \sigma^2(\varepsilon_n)$, on trace le nuage de points QQ plot(Quantile-Quantile) associé. Si le nuage de points est bien ajusté par la droite $y = x$ alors an admet la normalité de $\varepsilon_1, ..., \varepsilon_n$.
Sous R, écrire :

```
> qqnorm(residuals(Reg))
> qqline(residuals(Reg))
```

(voir la FIGURE(3.3))

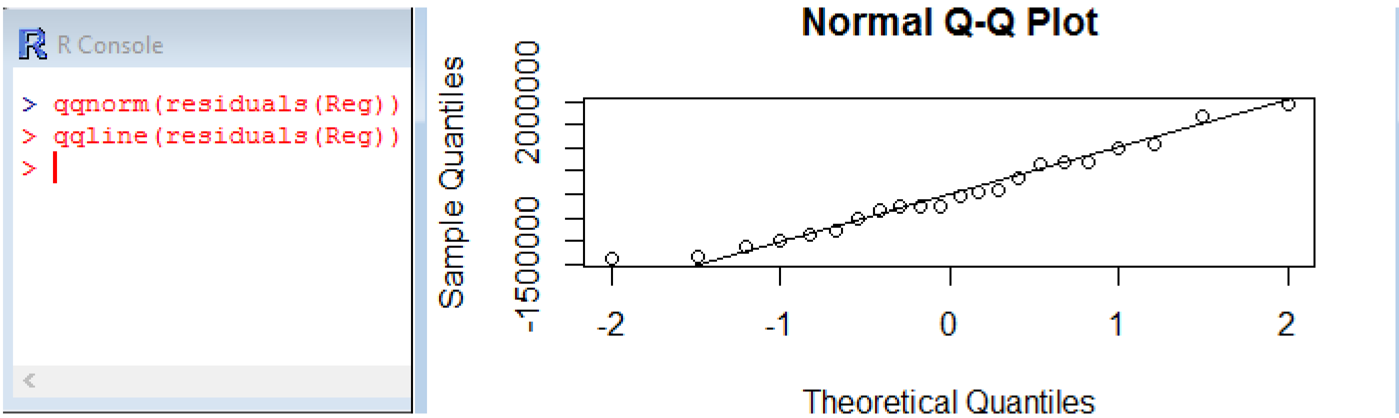


FIGURE 3.3 : Normalité des résidus

D'après cette figure, nous pouvons affiner le nuage de points par la ligne $y = x$, alors on admet la normalité de $\varepsilon_1, ..., \varepsilon_n$

3.5 Estimation de paramètres par MCO

On va estimer d'abord le vecteur des estimateurs $\widehat{\beta}$ défini par :

$$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^t = (X^t X)^{-1}(X^t Y). \quad (3.1)$$

Sous R, les estimateurs $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ sont donnés directement par la commande :

$$\boxed{> lm(Y \sim X1 + X2 + X3)}$$

ou

$$\boldsymbol{> Reg < -lm(Y \sim X1 + X2 + X3)}$$
$$\boldsymbol{> Reg} \quad .$$

(voir la FIGURE(3.4))

```
> lm(Y~X1+X2+X3)

Call:
lm(formula = Y ~ X1 + X2 + X3)

Coefficients:
(Intercept)           X1           X2           X3
 -2.130e+06   -7.131e+00    2.501e+01    3.572e+00

> Reg<-lm(Y~X1+X2+X3)
> Reg

Call:
lm(formula = Y ~ X1 + X2 + X3)

Coefficients:
(Intercept)           X1           X2           X3
 -2.130e+06   -7.131e+00    2.501e+01    3.572e+00
```

FIGURE 3.4 : Estimation des paramètres $\widehat{\boldsymbol{\beta}}$

Ou une autre façon d'entrer, sous R, les étapes suivantes :

1 * La matrice X est écrite par la commande :

> X < -matix(c(rep(1,22),X1,X2,X3,ncol=4)) (voir la FIGURE(3.5))

```
> X<-matrix(c(rep(1,22),X1,X2,X3),ncol=4)
> X
      [,1]      [,2]      [,3]      [,4]
 [1,]    1  346171.4  290749.6  436292.1
 [2,]    1  412119.5  315230.5  476208.7
 [3,]    1  417225.2  337556.2  509285.7
 [4,]    1  515281.7  355370.6  552179.9
 [5,]    1  580505.6  388193.4  607052.6
 [6,]    1  581615.8  418294.9  668130.0
 [7,]    1  641285.0  449581.0  728366.7
 [8,]    1  708072.5  479791.1  863197.3
 [9,]    1  727413.1  519631.6 1003199.4
[10,]    1  931349.1  570673.2 1160160.0
[11,]    1 1015258.8  617404.9 1283227.7
[12,]    1 1183216.1  664194.5 1446331.4
[13,]    1 1421693.3  729514.8 1649969.8
[14,]    1 1640006.1  771787.4 1870581.0
[15,]    1 1772202.4  837716.8 2067543.0
[16,]    1 1935113.0  919370.4 2259343.2
[17,]    1 2140304.7  979303.0 2341306.0
[18,]    1 2219064.4 1044920.1 2116090.1
[19,]    1 2426906.9 1127981.6 2349598.7
[20,]    1 2529100.0 1198400.0 2481200.0
[21,]    1 2598511.9 1153521.0 1987214.2
[22,]    1 2869600.0 1272599.0 2301500.0
```

FIGURE 3.5 : Matrice des variables explicatives

2 * on calcule $(\boldsymbol{X^tX})$ et $(\boldsymbol{X^tX})^{-1}$ par :

$$\boxed{\begin{array}{l} > \boldsymbol{X^tX <- t(X)\%*\%X} \quad . \\ > \boldsymbol{(X^tX)^{-1} <- solve(X^tX)} \end{array}}$$

(voir la FIGURE(3.6))

```
R Console
> XtX<-t(X)%*%X
> XtX
         [,1]         [,2]         [,3]         [,4]
[1,]       22 2.961202e+07 1.544179e+07 3.115798e+07
[2,] 29612016 5.451888e+13 2.637237e+13 5.439017e+13
[3,] 15441786 2.637237e+13 1.298453e+13 2.664364e+13
[4,] 31157978 5.439017e+13 2.664364e+13 5.564615e+13
> solve(XtX)
              [,1]          [,2]          [,3]          [,4]
[1,]  2.358067e+00  4.306679e-06 -1.167563e-05  6.051874e-08
[2,]  4.306679e-06  9.005525e-12 -2.297852e-11 -2.114828e-13
[3,] -1.167563e-05 -2.297852e-11  6.460876e-11 -1.937566e-12
[4,]  6.051874e-08 -2.114828e-13 -1.937566e-12  1.118510e-12
> 
```

FIGURE 3.6 : Matrices des $(\boldsymbol{X^tX})$ et $(\boldsymbol{X^tX})^{-1}$

3 * on calcule $(\boldsymbol{X^tY})$ par :

$\boxed{> \boldsymbol{X^tY} < -\boldsymbol{t(X)\% * \%Y}}$ (voir la FIGURE(3.7))

```
> XtY<-t(X)%*%Y
> XtY
             [,1]
[1,] 2.395079e+08
[2,] 4.020654e+14
[3,] 1.989908e+14
[4,] 4.109616e+14
```

FIGURE 3.7 : Vecteur des $\boldsymbol{X^tY}$

4 * Ainsi on écrit l'équation(3.1) par :

$\boxed{> \text{Hatbeta} < -\boldsymbol{solve}(\boldsymbol{X}^t\boldsymbol{X})^{-1})\% * \%(\boldsymbol{X}^t\boldsymbol{Y})}$

(voir la FIGURE(3.8))

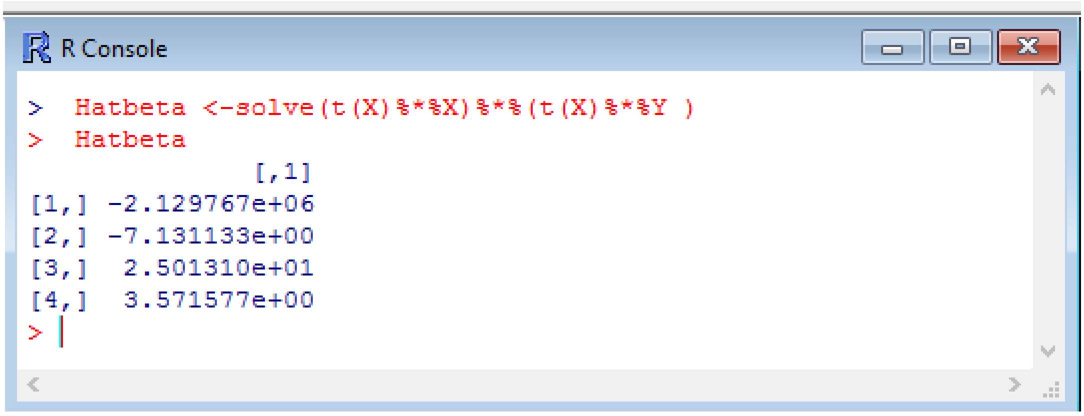


FIGURE 3.8 : Estimation des paramètres $\widehat{\beta}$

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

$$\widehat{\boldsymbol{y}}_i = -\mathbf{2130000} - \mathbf{7.131}\boldsymbol{x}_{i1} + \mathbf{2501}\boldsymbol{x}_{i2} + \mathbf{3.572}\boldsymbol{x}_{i3}.$$

Le signe du coefficient nous indique le sens de sens de la relation. D'après cette équation, on remarque que les coefficient de régression estimer $(\widehat{\beta}_2$ et $\widehat{\beta}_3)$, associée à les variables des industrie et commerce, sont positifs cela signifie que l'augmentation des industrie et commerce ce influe positivement sur l'économie nationale car leur coefficient $\widehat{\beta}_1$ est négatif.

3.6 Évaluation

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

La matrice de variance-covariance est défini par :

$$\boldsymbol{varcov}(\widehat{\beta}) = \boldsymbol{s}^2 = \boldsymbol{\sigma}_{\varepsilon}^2(\boldsymbol{X}^t\boldsymbol{X})^t$$

où les estimateur sans biais de la variance des résidus $\boldsymbol{s}_{\widehat{\varepsilon}}^2$ donnée par :

$$\boldsymbol{s}_{\widehat{\varepsilon}}^2 = \frac{\boldsymbol{SCR}}{\boldsymbol{n} - (\boldsymbol{p} - 1)} = \frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}{\boldsymbol{n} - (\boldsymbol{p} - 1)}$$

ou p : est le nombre des variables explicatives.

On peut alors calculer la matrice de variance-covariance, sous R, directement par la commande :

```
> vcov(Reg)
```

Ou en calculant s_ε^2, sous R, écrire par :

$$\boxed{\begin{array}{l} > \boldsymbol{SCR} < -\boldsymbol{sum}(\boldsymbol{residuals}(\boldsymbol{Reg})^2) \\ > \boldsymbol{Hatsigma2} - \boldsymbol{eps} < -\boldsymbol{SCR}/(\boldsymbol{n}-\boldsymbol{p}-1) \end{array}}$$ où p=3 et n=22 .

Donc, on obtient la résultat suivent : (voir la FIGURE(3.9))

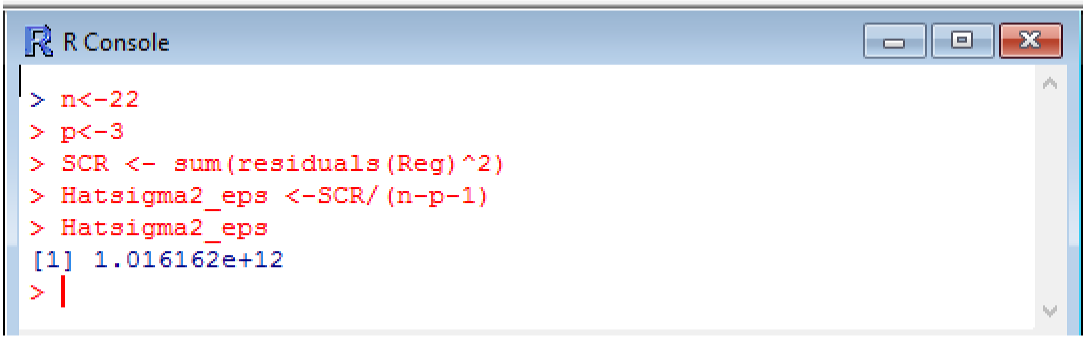


FIGURE 3.9 : L'estimation sans biais de la variance des résidus s_ε^2

Alors :

$$\boldsymbol{s_\varepsilon^2 = 1.016162 \times 10^{12}.}$$

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

> varcov < -$\boldsymbol{Hatsigma2 - eps * solve(t(X)\% * \%X)}$

On obtient le résultat suivant (voir la FIGURE(3.10)) :

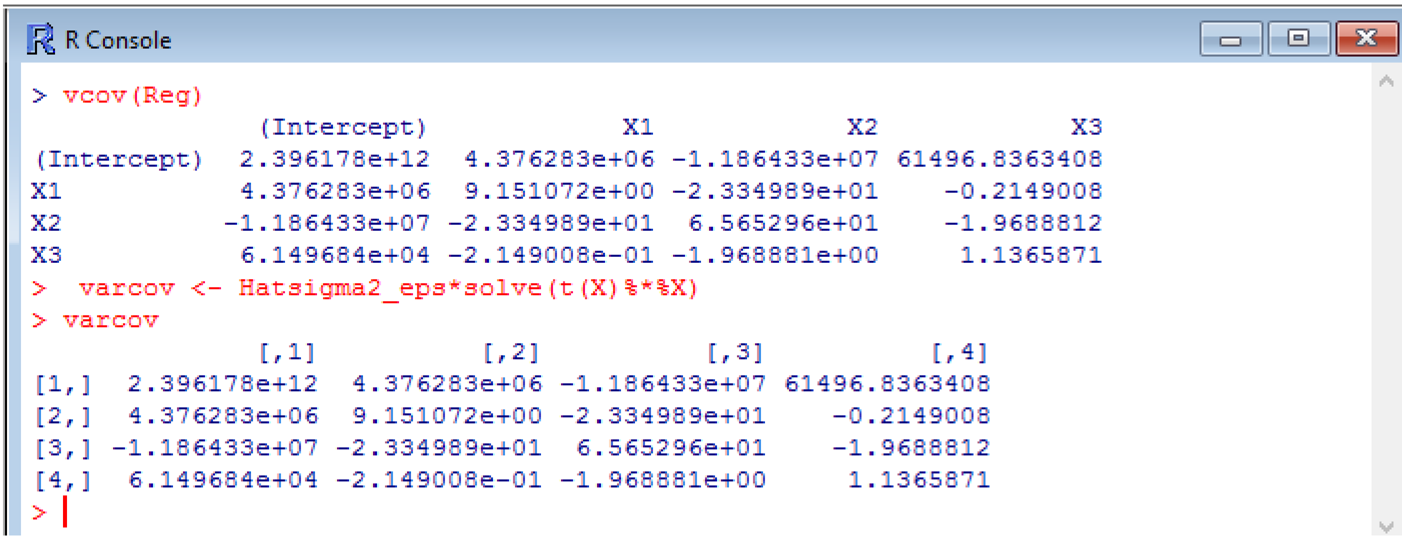


FIGURE 3.10 : La matrice $\boldsymbol{varcov}$

En suit, nous calculons les écart-types $\boldsymbol{s(\widehat{\beta}_j)}$ sont alors donnés par les racines carrées des éléments diagonaux de cette matrice (var cov), sous R, on a :

> Hatsigmabetas <- sqrt(diag(vcov(Reg)))

On obtient (voir la FIGURE(3.11))

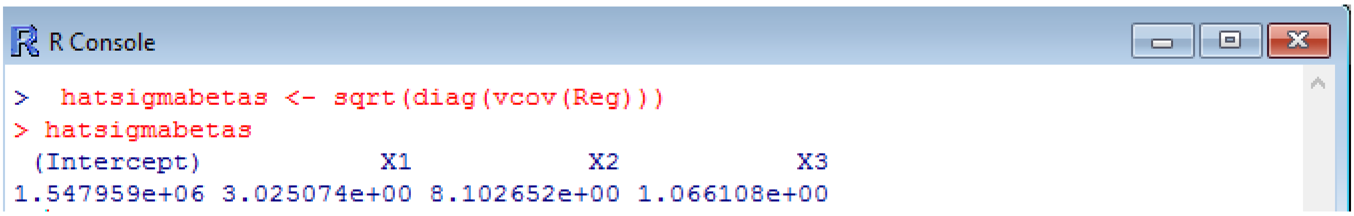


FIGURE 3.11 : Les écart-types $s(\widehat{\beta}_j)$

3.6.2 Estimation de β_j par intervalle de confiance

L'intervalle de confiance pour estimer $\beta_0, \beta_1, \beta_2$ et β_3 est donnée par :

$$\left[\widehat{\beta}_j - t_{n-2}^{1-\frac{\alpha}{2}} s_{\widehat{\beta}_j}, \widehat{\beta}_j + t_{n-2}^{1-\frac{\alpha}{2}} s_{\widehat{\beta}_j}\right] \forall j = \overline{0.3}$$

On calcule les intervalles de confiance au niveau $100(1-\alpha)\%$ (95%) pour les paramètres $\beta_0, \beta_1, \beta_2$ et β_3.
Sous R on écrit la commande suivant :

```
> confint (Reg,conf,level=0.95)
```

On obtient :

$$\beta_0 \in [-5381908, 1122375]$$

$$\beta_1 \in [-13.48658, -0.7756890]$$

$$\beta_2 \in [7.990060, 42.03614]$$

$$\beta_3 \in [1.331766, 5.811387]$$

3.7 Évaluation globale de la régression

3.7.1 Tableau d'analyse de la variance

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

$$SCT = Y^tY - n\overline{y}^2$$

$$SCR = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum \widehat{\varepsilon}_i^2$$

$$SCE = SCT - SCR$$

L'écriture sous R :

$> \boldsymbol{SCT < -t(y)\% * \%Y - n * (mean(Y)^2)}$
$> \boldsymbol{SCR < -sum(residuals(Reg)^2)}$.
$> \boldsymbol{SCE < -SCT - SCR}$.

On obtient les résultats suivants :

$$\boldsymbol{SCT = 4.787144 \times 10^{14}}$$

$$\boldsymbol{SCR = 1.829091 \times 10^{13}}$$

$$\boldsymbol{SCE = 4.604235 \times 10^{14}}$$

Donc le tableau d'analyse de la variance est donnée par le tableau (3.2) :

source de variation	ddl	somme des carrés	carrés moyens	$\boldsymbol{F_{obs}}$
Expliquée	3	$\boldsymbol{4604235 \times 10^8}$	$\boldsymbol{1534745 \times 10^8}$	151.0335
Résiduelle	18	$\boldsymbol{1829091 \times 10^7}$	$\boldsymbol{1016161667 \times 10^3}$	
Totale	21	$\boldsymbol{4787144 \times 10^8}$		

TABLE 3.2 : Tableau d'analyse de la variance .

3.7.2 coefficient de détermination

La proportion de variabilité par les trois régresseurs est calculé par le coefficient de détermination $\boldsymbol{R^2}$ qui est donnée par la relation suivante :

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

Sous R, on a :

$> \boldsymbol{R2 < -SCE/SCT}$

On obtient $\boldsymbol{R^2 = 96.2\%}$ ce qui montre un ajustement très fort. Ce coefficient ne prend pas en compte le nombre de variables explicatives. Il est nécessaire de s'intéresser au coefficient de détermination ajusté $\boldsymbol{R^2_{adj}}$ qui est donnée par la relation suivante :

$$\boldsymbol{R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1 - R^2) = 95.57\%}$$

qui reste un ajustement très fort.

3.8 Tests de signification

3.8.1 Test de Student sur le paramètre β

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

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

On calcul la statistique de test T_{obs} :

, pour β_0 : $T_{obs} = \frac{\widehat{\beta_0}}{s(\widehat{\beta_0}} = \frac{-2129767}{1547959} = -1.376$

, pour β_1 : $T_{obs} = \frac{\widehat{\beta_1}}{s(\widehat{\beta_1}} = \frac{-7.131133}{3.025074} = -2.357$

, pour β_2 : $T_{obs} = \frac{\widehat{\beta_2}}{s(\widehat{\beta_2}} = \frac{25.01310}{81.02652} = 0.309$

, pour β_3 : $T_{obs} = \frac{\widehat{\beta_3}}{s(\widehat{\beta_3}} = \frac{3.571577}{1.066108} = 3.350$

Sous R, on écrire par :

$> Tobs < -Hatbeta/Hatsigmabetas$.
$> T - tab < -qt(p = 1 - 0.05/2, df = n - p - 1)$.
$> ifelse(abs(tobs) > T - tab"rejetH0", "nonrejetH0$

On obtient le résultat suivant :

, [.1]
, [1.] "non rejet H0"
, [2.] "rejet H0"
, [3.] "rejet H0"
, [4.] "rejet H0"

Règle de décision

Comme la valeur critique est donnée par $t_{18}^{0.975} = 2.100922$, on décide l'accepter l'hypothèse nulle H_0 au seuil de significativité $\alpha = 0.05$ pour j=0 et on la rejette pour j=1,j=2 et j=3 pour respectivement.
Alors, cela veut dine que les variables X_1, X_2 et X_3 (agriculture, industrie, Commerce) ont des influences très important sur l'économie nationale. Cela montre que l'économie nationale globale de l'Algérie est basée essentiellement sur l'agriculture,l'industrie et le commerce.

3.8.2 Test globale de Fisher

Est ce que la liaison globale entre les X_i et Y est-elle significative ?
Nous répondrons de cette question par le test de Fisher.
On veut tester les hypothèses suivantes : * l'hypothèse nulle : $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

contre
* l'hypothèse alternative : $\boldsymbol{H_1 : \exists j \in \{1,2,3\}}$ tel que $\boldsymbol{\beta_j \neq 0}$ (Y dépend d'au moins une variable $\boldsymbol{X_j}$.
on calcule :

$$\boldsymbol{F_{obs} = \frac{\frac{R^2}{p}}{\frac{(1-R^2)}{(n-p-1)}} = 151.0335}$$

Sous R, on a :

$$\boxed{\begin{array}{l} \boldsymbol{> R2 <- SCE/SCT} \quad . \\ \boldsymbol{> Fabs <- (R2/p)/((1-R2)/(n-p-1))} \quad . \\ \boldsymbol{> F_tab <- qf(p = 1-0.05, df1 = p, df2 = n-p-1)} \end{array}}$$

Règle de décision

La statistique $\boldsymbol{F_{obs} = 151.0335}$ est supérieure à la valeur critique $\boldsymbol{f_{0.95}^{(3,18)} = 3.159908}$, on conclut que le test est significatif, on rejette l'hypothèse nulle $\boldsymbol{H_0}$ au seuil de significativité $\boldsymbol{\alpha = 0.05}$. Ce résultat montre qu'au moins une des variables contribue à expliquer l'économie nationale.

CONCLUSION

Dans ce mémoire, on a présenté des modèles de la régression linéaire simple et multiple, ainsi qu'une application faite dans le domaine économique pour le modèle multiple. Ces modèles sont des solutions qui existent pour observer et expliquer les liens entre une variable à explicative et d'autre variables indépendants. Enfin, une attention particulière est accordée à la prospective.

BIBLIOGRAPHIE

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

[2] Bo Régis économétrie, Cours et exercices corrigés, Dunod, (2015).

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

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

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

[6] CARE. TABLEAU DE BORD DE L'ÉCONOMIE NATIONALE, https ://tbn.care.dz/.

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

[8] Dominok Salvatore. économie et statistique appliquée ; série SHAUM, M C GAW-HILL, NEWYORK, USA ;1982.

[9] Dodge, Y, Rousson, V. (2004). Analyse de régression appliquée, Dunod, 2-ème édition

[10] Francis Galton Régression toward mediocrity in herditary stature, journal of the Anthropological Institute 15 : 246-1886

[11] Frédéric Bestrand, Myriam Maumy-Bertrand Initiation à la statistique avec R. Dunod.(2010).

[12] Fourcassié V. et Jost C. Inroduction aux modèle linéaire génénraux (General Linear model -GLM)

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

[14] Francis Galton, «Regression Towards Mediocrity in Hereditary stature», Journal of the Anthropological Institute, vol. 15, 1886, p. 246-263.

[15] Giorgio Russolillo (2016). La Régression Linéaire Multiple, Département IMATHC-NAM, giorgio.russolillo@com.fr.

[16] Ihaka, R. and Gentleman, R. (1996). R : A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5 : 299 314.

[17] Kalbeichj. probabilité and statistical , Springer-verlage, NEW YORK USA ;1986.

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

[19] Labrousse, C. (1983) . Introduction à léconométrie. Maîtrise déconométrie, Dunod

[20] Mémoire de master de l'étudiant BERKANI F, intitulé :Application de la Régression Linéaire Multiples sur la Balance Commerciale Algérienne.mai/2016.

[21] R Palm et A.F. lemma, «Quelques alternatives à la régression classique dans le cadre de la colinéarité», Revue de statistique appliquée, vol. 43,$\boldsymbol{n^0}$ 2, 1995, p. 5-33.

[22] Yadolah Dodge. analyse de régression appliquée, Dunod, PARIS ;1999.

[23] Yadolah Dodge, The Concise Encyclopaedia of Statistics, New York, Springer, 2010, 622 p. ISBN 978-0-387-31742-7.

ANNEXE :A

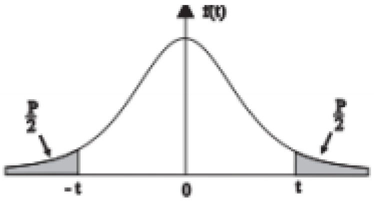
TABLEAU DES ABRÉVIATION

Symbole	Description
ANOVA	Analyse de la variance.
$c_{np1}^{1-\frac{\alpha}{2}}$	quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi de X_2 à $(np1)$ ddl.
c.à.d	C'est à dire.
CME	carrée moyenne expliquée par le modèle
CMR	carrée moyenne résiduelle.
cov(x,y)	covariance entre x et y
ddl	degrés de liberté.
E[X]	espérance mathématique ou moyenne du v.a. X.
$f_{(p,q)}^{(1\alpha)}$	quantile d'ordre (1α) d'une loi de Fisher à (p,q) ddl.
iid	indépendantes et identiquement distribuées.
Γ	matrice de variance covariance.
MCO	Moindres Carrés Ordinaires.
$\mathbb{R}$	ensemble des valeurs réelles.
$\mathbb{R}_+$	ensemble des valeurs réelles positives.
χ_p^2	Loi de χ^2 à p ddl
$N(0,1)$	loi normale standard (centrée réduite).
PIB	Produit intérieur brut
R^2	coefficient de détermination.

$\boldsymbol{R^2_{adj}}$	coefficient de détermination ajusté
RC	région de confiance
resp	respectivement.
RLS	régression linéaire simple.
RLM	régression linéaire multiple.
$\boldsymbol{\sigma^2_x}$	variance d'une v.a. X
$\boldsymbol{s^2_y}$	La variance empirique de la v.a Y
$\boldsymbol{s^2_{xy}}$	covariance empirique entre les variables X et Y
$\boldsymbol{s^2_x}$	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.
$\boldsymbol{t^{1\frac{\alpha}{2}}_{np1}}$	quantile d'ordre de la loi de Student à $\boldsymbol{(np1)}$ ddl.
Tr(A)	trace de matrice A
v.a	variable aléatoire.
var(X)	variance mathématique du v.a. X.
$\boldsymbol{X^t}$	transposée de X
$\boldsymbol{(X_1, X_2, ..., X_n)}$	échantillon de taille n de X.

ANNEXE : B

Tableau de la loi de Student



ν	P =0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,128	0,260	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,96	2,326	2,576

ANNEXE :C

Table de la loi de Fisher-Seconder $\alpha = 0.05$

num	den 1	2	3	4	5	6	7	8	9	10
1	161.4476	18.5128	10.1280	7.7086	6.6079	5.9874	5.5914	5.3177	5.1174	4.9646
2	199.5000	19.0000	9.5521	6.9443	5.7861	5.1433	4.7374	4.4590	4.2565	4.1028
3	215.7073	19.1643	9.2766	6.5914	5.4095	4.7571	4.3468	4.0662	3.8625	3.7083
4	224.5832	19.2468	9.1172	6.3882	5.1922	4.5337	4.1203	3.8379	3.6331	3.4780
5	230.1619	19.2964	9.0135	6.2561	5.0503	4.3874	3.9715	3.6875	3.4817	3.3258
6	233.9860	19.3295	8.9406	6.1631	4.9503	4.2839	3.8660	3.5806	3.3738	3.2172
7	236.7684	19.3532	8.8867	6.0942	4.8759	4.2067	3.7870	3.5005	3.2927	3.1355
8	238.8827	19.3710	8.8452	6.0410	4.8183	4.1468	3.7257	3.4381	3.2296	3.0717
9	240.5433	19.3848	8.8123	5.9988	4.7725	4.0990	3.6767	3.3881	3.1789	3.0204
10	241.8817	19.3959	8.7855	5.9644	4.7351	4.0600	3.6365	3.3472	3.1373	2.9782
11	242.9835	19.4050	8.7633	5.9358	4.7040	4.0274	3.6030	3.3130	3.1025	2.9430
12	243.9060	19.4125	8.7446	5.9117	4.6777	3.9999	3.5747	3.2839	3.0729	2.9130
13	244.6898	19.4189	8.7287	5.8911	4.6552	3.9764	3.5503	3.2590	3.0475	2.8872
14	245.3640	19.4244	8.7149	5.8733	4.6358	3.9559	3.5292	3.2374	3.0255	2.8647
15	245.9499	19.4291	8.7029	5.8578	4.6188	3.9381	3.5107	3.2184	3.0061	2.8450
16	246.4639	19.4333	8.6923	5.8441	4.6038	3.9223	3.4944	3.2016	2.9890	2.8276
17	246.9184	19.4370	8.6829	5.8320	4.5904	3.9083	3.4799	3.1867	2.9737	2.8120
18	247.3232	19.4402	8.6745	5.8211	4.5785	3.8957	3.4669	3.1733	2.9600	2.7980
19	247.6861	19.4431	8.6670	5.8114	4.5678	3.8844	3.4551	3.1613	2.9477	2.7854
20	248.0131	19.4458	8.6602	5.8025	4.5581	3.8742	3.4445	3.1503	2.9365	2.7740
21	248.3094	19.4481	8.6540	5.7945	4.5493	3.8649	3.4349	3.1404	2.9263	2.7636
22	248.5791	19.4503	8.6484	5.7872	4.5413	3.8564	3.4260	3.1313	2.9169	2.7541
23	248.8256	19.4523	8.6432	5.7805	4.5339	3.8486	3.4179	3.1229	2.9084	2.7453
24	249.0518	19.4541	8.6385	5.7744	4.5272	3.8415	3.4105	3.1152	2.9005	2.7372
25	249.2601	19.4558	8.6341	5.7687	4.5209	3.8348	3.4036	3.1081	2.8932	2.7298
26	249.4525	19.4573	8.6301	5.7635	4.5151	3.8287	3.3972	3.1015	2.8864	2.7229
27	249.6309	19.4587	8.6263	5.7586	4.5097	3.8230	3.3913	3.0954	2.8801	2.7164
28	249.7966	19.4600	8.6229	5.7541	4.5047	3.8177	3.3858	3.0897	2.8743	2.7104
29	249.9510	19.4613	8.6196	5.7498	4.5001	3.8128	3.3806	3.0844	2.8688	2.7048
30	250.0951	19.4624	8.6166	5.7459	4.4957	3.8082	3.3758	3.0794	2.8637	2.6996

num	den 11	12	13	14	15	16	17	18	19	20
1	4.8443	4.7472	4.6672	4.6001	4.5431	4.4940	4.4513	4.4139	4.3807	4.3512
2	3.9823	3.8853	3.8056	3.7389	3.6823	3.6337	3.5915	3.5546	3.5219	3.4928
3	3.5874	3.4903	3.4105	3.3439	3.2874	3.2389	3.1968	3.1599	3.1274	3.0984
4	3.3567	3.2592	3.1791	3.1122	3.0556	3.0069	2.9647	2.9277	2.8951	2.8661
5	3.2039	3.1059	3.0254	2.9582	2.9013	2.8524	2.8100	2.7729	2.7401	2.7109
6	3.0946	2.9961	2.9153	2.8477	2.7905	2.7413	2.6987	2.6613	2.6283	2.5990
7	3.0123	2.9134	2.8321	2.7642	2.7066	2.6572	2.6143	2.5767	2.5435	2.5140
8	2.9480	2.8486	2.7669	2.6987	2.6408	2.5911	2.5480	2.5102	2.4768	2.4471
9	2.8962	2.7964	2.7144	2.6458	2.5876	2.5377	2.4943	2.4563	2.4227	2.3928
10	2.8536	2.7534	2.6710	2.6022	2.5437	2.4935	2.4499	2.4117	2.3779	2.3479
11	2.8179	2.7173	2.6347	2.5655	2.5068	2.4564	2.4126	2.3742	2.3402	2.3100
12	2.7876	2.6866	2.6037	2.5342	2.4753	2.4247	2.3807	2.3421	2.3080	2.2776
13	2.7614	2.6602	2.5769	2.5073	2.4481	2.3973	2.3531	2.3143	2.2800	2.2495
14	2.7386	2.6371	2.5536	2.4837	2.4244	2.3733	2.3290	2.2900	2.2556	2.2250
15	2.7186	2.6169	2.5331	2.4630	2.4034	2.3522	2.3077	2.2686	2.2341	2.2033
16	2.7009	2.5989	2.5149	2.4446	2.3849	2.3335	2.2888	2.2496	2.2149	2.1840
17	2.6851	2.5828	2.4987	2.4282	2.3683	2.3167	2.2719	2.2325	2.1977	2.1667
18	2.6709	2.5684	2.4841	2.4134	2.3533	2.3016	2.2567	2.2172	2.1823	2.1511
19	2.6581	2.5554	2.4709	2.4000	2.3398	2.2880	2.2429	2.2033	2.1683	2.1370
20	2.6464	2.5436	2.4589	2.3879	2.3275	2.2756	2.2304	2.1906	2.1555	2.1242
21	2.6358	2.5328	2.4479	2.3768	2.3163	2.2642	2.2189	2.1791	2.1438	2.1124
22	2.6261	2.5229	2.4379	2.3667	2.3060	2.2538	2.2084	2.1685	2.1331	2.1016
23	2.6172	2.5139	2.4287	2.3573	2.2966	2.2443	2.1987	2.1587	2.1233	2.0917
24	2.6090	2.5055	2.4202	2.3487	2.2878	2.2354	2.1898	2.1497	2.1141	2.0825
25	2.6014	2.4977	2.4123	2.3407	2.2797	2.2272	2.1815	2.1413	2.1057	2.0739
26	2.5943	2.4905	2.4050	2.3333	2.2722	2.2196	2.1738	2.1335	2.0978	2.0660
27	2.5877	2.4838	2.3982	2.3264	2.2652	2.2125	2.1666	2.1262	2.0905	2.0586
28	2.5816	2.4776	2.3918	2.3199	2.2587	2.2059	2.1599	2.1195	2.0836	2.0517
29	2.5759	2.4718	2.3859	2.3139	2.2525	2.1997	2.1536	2.1131	2.0772	2.0452
30	2.5705	2.4663	2.3803	2.3082	2.2468	2.1938	2.1477	2.1071	2.0712	2.0391