République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

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

MASTER en Mathématiques

Option : **Statistique**

Par

Rouahna Manal

Titre :

Analyse des résidus d'un modèle de régression

Membres du Comité d'Examen :

Mr.	DJABRANE Yahia	Prof.	UMKB	Encadreur
Mr.	BENATIA Fatah	Prof.	UMKB	Examinateur
Melle.	KHEIREDDINE Souraya	MCB.	UMKB	Examinatrice

Juin 2021

Dédicace

Après avoir terminé cette recherche, si "**Dieu**" le veut.

Je dédie ce travail à ce qui est le plus cher à tout être humain.

A tous ceux qui m'ont appris une lettre, et m'ont soutenu même d'un mot dans cette vie mortelle.

A qui je dois mon existence après "**Dieu**", mes très chers parents que "**Dieu**" leur donne longue vie.

A qui la vie nous a réunis, mes chers frères et soeurs, que "**Dieu**" les bénisse et les protège de tout mal.

Aux chers membres de la famille **ROUAHNA,** petits et grands. Et à ceux qui sont dans le coeur et absents de la langue.

A ceux qui ont su les trouver et m'ont appris à ne pas les perdre.....mes amies.

A tous ceux qui m'ont aidé de près ou de loin à accomplir cette mémoire.

Et nous demandons à "**Dieu**" d'en faire un phare pour chaque étudiant de la connaissance.

A tout la promotion 2eme Mester mathématique **2020-2021.**

Merci.

REMERCIEMENTS

*Louanges et remerciements à "**Dieu**"*

*J'adresse mes sincères remerciements à mon superviseur, Prof. **DJABRANE Yahia** pour les instructions et les directions, qu'il m'a données et qui m'a accompagné tout au long de la réalisation du mémoire. Qu'**Allah** le récompense de tout le meilleur. Il a toute mon appréciation et respect.*

Je remercie également tous les professeurs qui m'ont aidé avec leurs conseils et orientations.

*Je tiens à remercier les mombres du Jury, Prof. **Benatia F**. et Dr. **Khiereddine S**. qui ont accepté la lecture et l'évaluation du mémoire.*

Un grand merci particulier à mes collègues et mes amies, pour les sympathiques moments qu'on a passés ensemble, je les remercie pour leur confiance, et leurs soutien moral au cours de ces années.

A tous ceux qui ont contribué à la réalisation de ce mémoire.

Table des matières

Dédicace i

Remerciements ii

Table des matières iii

Liste des figures v

Liste des tableaux vi

Introduction 1

1 Modèles de régression 3

1.1 Modèle de régression linéaire simple 3

1.1.1 Estimation des paramètres 5

1.1.2 Résidus et variance résiduelle 10

1.1.3 Lois des estimateurs : 15

1.1.4 Tests d'hypothèses et intervalle de confiance 17

1.1.5 Qualité d'ajustement 19

2 Application sous R 25

2.1 Analyse des résidus d'un modèle de régression linéaire 25

2.1.1 Exemple sur la régression linéaire simple 25

2.1.2 Validation du modèle 29

Conclusion **42**

Bibliographie **43**

Annexe A : Logiciel *R* **45**

Annexe B : Abréviations et Notations **47**

Table des figures

1.1 Nuage de points de la consommation en fonction du revenu monsuelle. . . . 5

2.1 La surface de couverture ST en fonction de la surface de couverture SC. . . 27

2.2 Nuage des points et droite de régression 29

2.3 Graphe des résidus en fonction des prédictions 30

2.4 Le graphique Q-Q plot 32

2.5 Histogramme des résidus 33

2.6 Graphique des résidus studentisés en fonction des prédictions 36

2.7 Série chronologique des résidus 37

2.8 Graphiques ACF des résidus 38

2.9 Graphique des résidus en fonction des numéros d'observations 38

Liste des tableaux

1.1 Le revenu monsuelle et la consommation 5
1.2 Table d'analyse de la variance (ANOVA) 20
1.3 Table des données 21
1.4 Table d'analyse de la variance (ANOVA) 24

2.1 Table des données 26

Introduction

La statistique est la science qui consiste à collecter des données numériques, puis à les organiser, et les analyser afin d'atteindre des résultats spécifiques, pour clarifier un phénomène ou une situation ; Ces données étant classées en deux grands catégories : quantitatives et qualitatives.

Tout cela, en fait une application importante dans divers domaines des sciences tels que : la physique, la chimie, la médecine, les sciences sociales et humaines, l'économie, l'industrie, les sports et même la politique.... Les propriétés(biais, convergence,...), et l'inférence statistique (test de significativité, intervalle de confiance...) reposent en grande partie sur des hypothèses sur les résidus, dont nous devons s'assurer de la conformité aux hypothèses.

Une hypothèse statistique est un ensemble de suppositions et de suggestions (qui peuvent être vraies ou fausses), qu'il a développées à l'aide de modèles statistiques pour confirmer la relation entre les variables, et sont divisées en : l'hypothèse nulle et l'alternative. L'interprétation ou la conclusion est mise sous forme mathématique, et elle est testée selon des tests statistiques qui est une technique de base en inférence statistique, afin de prendre la décision de rejeter ou de ne pas rejeter l'hypothèse nulle.

La régression est l'une des méthodes les plus simples utilisées en statistique, car la pente de la droite est égale à la relation entre une ou plusieurs variables explicatives et une variable à expliquer, corrigée par le rapport des écarts-types de ces variables.

Dans le cadre d'un modèle de régression, nous avons d'exposer la régression linéaire dans

un cas simple afin de bien comprendre les enjeux de cette méthode, les problèmes posés et les réponses apportées. Nous traiterons ici, l'analyse des résidus qui est une étape primordiale dans cette méthode par la vérification des hypothèses, qui peut être évaluée graphiquement ou par les tests statistiques.

Le but générale est donc, de montrer la validation de cette régression. Ce mémoire est composé en deux chapitres :
Chapitre1 : nous verrons comment réaliser une régression linéaire simple, mais également comprendre le principe général ainsi que les hypothèses des résidus.
Chapitre2 : nous traitons l'application de la régression linéaire simple, et son diagnostic par l'analyse des résidus avec des données réelles sous le logiciel statistique R.

Chapitre 1

Modèles de régression

Dans ce premier chapitre, nous étudions et analysons les résidus pour un modèle de régression linéaire simple (RLS). Il s'agit bien d'une technique statistique permettant de modéliser la relation linéaire entre une variable explicative (notée X) et une variable à expliquer (notée Y).

1.1 Modèle de régression linéaire simple

Définition 1.1.1 *Le modèle de RLS est une variable endogène y (dépendante) expliquée par une seule variable exogène x (indépendante), que l'on peut écrire sous la forme suivante :*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \forall i = \overline{1, n}. \tag{1.1}$$

où

- y_i *est la variable aléatoire à expliquer (à valeur dans* $\mathbb{R}$*),*
- x_i *est la variable explicative (à valeur dans* $\mathbb{R}$*),*
- β_0 *et* β_1 *les paramètres inconnus du modèle à estimer,*
- ε_i *est l'erreur(ou bruits) aléatoire du modèle,*
- n *est nombre d'observations.*

Les hypothèses relatives à ce modèle sont :

1) $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, les ε_i sont les réalisations indépendantes identiquements distribuées (iid), et d'une variable aléatoire gaussienne ;

2) $\mathbb{E}[\varepsilon_i] = 0$, $\forall i = \overline{1,n}$, l'erreur centrée ;

3) $Var(\varepsilon_i) = \mathbb{E}[\varepsilon_i^2] = \sigma_\varepsilon^2 < +\infty$, $\forall i = \overline{1,n}$, la variance de l'erreur est constante (inconnue) (l'hypothèse d'homoscédasticité) ;

4) $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$, $\forall i \neq j$, les erreurs sont non-corrélées (indépendantes entre eux) ;

5) L'erreur ε_i est indépendante de x_i : $Cov(x_i, \varepsilon_i) = 0$.

Le modèle (1.1) prend la forme matricielle suivante :

$$Y = X\beta + \varepsilon. \tag{1.2}$$

Tel que

$$Y = \begin{bmatrix} y_1 \\ . \\ . \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_1 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}, \; \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ . \\ . \\ \varepsilon_n \end{bmatrix}.$$

- Y désigne le vecteur à expliquer de taille $n \times 1$,
- X la matrice explicative de taille $n \times 2$,
- β le vecteur des paramètres de taille 2×1,
- ε le vecteur d'erreurs de taille $n \times 1$.

Exemple 1.1.1 *Considérons l'exemple suivant représentant la relation entre le revenu monsuelle X, et la consommation Y de 10 familles :*

En deduire du nuage statistique qu'il s'agit d'un modèle de régression linéaire simple :

$$y_i = a + bx_i + \xi_i \quad , \forall i = \overline{1,10}.$$

X $(\times 10^4 DA)$	2	2.5	3.5	4	4.5	5	6	8	10	13
Y $(\times 10^4 DA)$	1.5	2.9	2.6	3.4	5.3	7.1	6.8	8.2	7.8	9.6

TAB. 1.1 – Le revenu monsuelle et la consommation

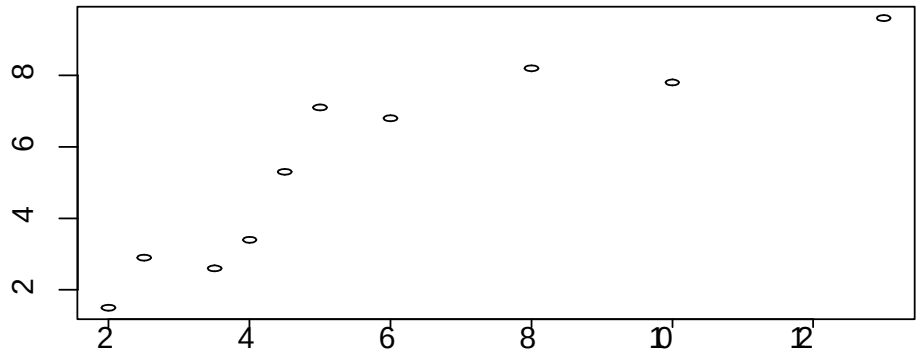

FIG. 1.1 – Nuage de points de la consommation en fonction du revenu monsuelle.

Avec y_i est la consommation et x_i est le revenu monsuelle.

1.1.1 Estimation des paramètres

On cherche les valeurs $\hat{\beta}_0$, $\hat{\beta}_1$et $\hat{\sigma}^2_\varepsilon$ qui sont les estimateurs des paramètres β_0, β_1 et σ^2_ε (respectivement), ils définissant la droite de régression estimé

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

On peut utiliser la méthode du maximum de vraisemblance (MV) pour estimer les paramètres β_0 et β_1. Posons :
$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ les moyennes empiriques des x_i et des y_i (respectivement),

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

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

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

L'hypothèse de normalité de $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, le modèle (1.1) permet d'endéduire que $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma_\varepsilon^2)$. La densité de y_i est donc donnée par :

$$f(y_i) = \frac{1}{\sqrt[2]{2\pi\sigma_\varepsilon^2}} \exp\left\{\frac{-1}{2\sigma_\varepsilon^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\}, \quad \forall i = \overline{1, n}.$$

Du fait de l'indépendance des y_i la densité jointe de $y_1, ..., y_n$, peut s'écrire comme le produit des densités marginales :

$$f(y_1, ..., y_n) = \prod_{i=1}^{n} f(y_i) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{n}{2}}} \exp\left\{\frac{-1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}.$$

La fonction de vraisemblance est égale à la fonction de densité, mais est vue comme une fonction des paramètres en considérent que les observations $y_1, ..., y_n$ sont fixé :

$$L(\beta_0, \beta_1, \sigma_\varepsilon^2) = (2\pi\sigma_\varepsilon^2)^{\frac{-n}{2}} \exp\left\{\frac{-1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}.$$

La fonction de log-vraisemblance vaut

$$\log L(\beta_0, \beta_1, \sigma_\varepsilon^2) = \frac{-n}{2}\log(2\pi\sigma_\varepsilon^2) - \frac{-1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

En annulant les dérivées de log-vraisemblance par rapport à β_0, β_1 et σ_ε^2, on obtient :

$$\begin{cases} \frac{\partial \log L(\beta_0,\beta_1,\sigma_\varepsilon^2)}{\partial \beta_0} = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)}{\sigma_\varepsilon^2} = 0, \\ \frac{\partial \log L(\beta_0,\beta_1,\sigma_\varepsilon^2)}{\partial \beta_1} = \frac{\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)}{\sigma_\varepsilon^2} = 0, \\ \frac{\partial \log L(\beta_0,\beta_1,\sigma_\varepsilon^2)}{\partial \sigma_\varepsilon^2} = \frac{-n}{2\sigma_\varepsilon^2} + \frac{1}{2\sigma_\varepsilon^4}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = 0. \end{cases}$$

Le système devient

$$\begin{cases} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0, \\ \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0, \\ -n + \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = 0. \end{cases}$$

Finalement, le solution est

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_x}, \\ \hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i). \end{cases}$$

Propriétés des estimateurs

Théorème 1.1.1 *$\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 et β_1.*

Preuve. $\hat{\beta}_1$ est un estimateurs sans biais de $\beta_1 \Longleftrightarrow \mathbb{E}\left[\hat{\beta}_1\right] = \beta_1$. En effet,

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

Tel que

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \\ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \beta_0 + \beta_1 \overline{x} + \overline{\varepsilon}, \end{cases} \Longrightarrow y_i - \overline{y} = \beta_1(x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon}).$$

Alors

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

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

Car l'erreurs ε_i aléatoires $\mathbb{E}[\varepsilon_i] = 0$.

Pour $\hat{\beta}_0$: on a

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \Longrightarrow \mathbb{E}\left[\hat{\beta}_0\right] = \mathbb{E}\left[\overline{y} - \hat{\beta}_1\overline{x}\right] = \mathbb{E}[\overline{y}] - \mathbb{E}\left[\hat{\beta}_1\right]\overline{x} = \beta_0 + \beta_1\overline{x} - \beta_1\overline{x} = \beta_0.$$

■

Théorème 1.1.2 *Les variances des estimateurs sont :*

$$Var(\hat{\beta}_0) = \sigma^2_{\hat{\beta}_0} = \frac{\sigma^2_\varepsilon \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sigma^2_\varepsilon \sum_{i=1}^{n} x_i^2}{nS_x},$$

$$Var(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_\varepsilon}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sigma^2_\varepsilon}{S_x}.$$

Tendis que leurs covariance est :

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2_\varepsilon \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = -\frac{\sigma^2_\varepsilon \overline{x}}{S_x}.$$

Preuve. 1)

$$\begin{aligned} Var(\hat{\beta}_1) &= Var\left(\beta_1 + \frac{\sum_{i=1}^{n}(x_i-\overline{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right), \\ &= Var\left(\frac{\sum_{i=1}^{n}(x_i-\overline{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^2} Var(\varepsilon_i), \\ &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sigma_\varepsilon^2}{S_x}. \end{aligned}$$

2)

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\overline{y} - \hat{\beta}_1\overline{x}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1\overline{x}\right), \\ &= Var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + Var\left(\hat{\beta}_1\right)\overline{x} - 2\overline{x}Cov\left(\overline{y}, \hat{\beta}_1\right). \end{aligned}$$

Où

$$\begin{aligned} Cov\left(\overline{y}, \hat{\beta}_1\right) &= Cov\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \beta_1 + \frac{\sum_{i=1}^{n}(x_i-\overline{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right), \\ &= \frac{1}{n}Cov\left(\sum_{i=1}^{n} y_i, \beta_1\right) + \frac{\sum_{i=1}^{n}(x_i-\overline{x})}{n\sum_{i=1}^{n}(x_i-\overline{x})^2} Cov\left(\sum_{i=1}^{n} y_i, \varepsilon_i\right), \end{aligned}$$

$$Cov\left(\overline{y},\hat{\beta}_1\right)=0 \text{ car } \begin{cases} Cov\left(\sum_{i=1}^{n} y_i,\beta_1\right)=0, \\ \sum_{i=1}^{n}(x_i-\overline{x})=0. \end{cases}$$

Donc

$$Var(\hat{\beta}_0)=\frac{\sigma_\varepsilon^2}{n}+\frac{\sigma_\varepsilon^2}{S_x}\overline{x}=\frac{\sigma_\varepsilon^2\sum_{i=1}^{n}x_i^2}{nS_x}.$$

3)

$$\begin{aligned} Cov(\hat{\beta}_0,\hat{\beta}_1)&=Cov(\overline{y}-\hat{\beta}_1\overline{x},\hat{\beta}_1)=Cov(\overline{y},\hat{\beta}_1)-\overline{x}Cov(\hat{\beta}_1,\hat{\beta}_1), \\ &=-\overline{x}Var(\hat{\beta}_1)=-\frac{\sigma_\varepsilon^2\overline{x}}{S_x}. \end{aligned}$$

■

1.1.2 Résidus et variance résiduelle

Définition 1.1.2 *Les résidus $\hat{\varepsilon}_i$ sont les différences entre les valeurs observées y_i et les valeurs ajustées $\hat{y}_i$, définis par :*

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

Où $\hat{y}_i$ est la valeur ajustée de y_i par le modèle, c'est-à-dire

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i\,;\ \ \forall i=\overline{1,n}.$$

Les $\hat{\varepsilon}_i$ sont les variables aléatoires observés.

Proposition 1.1.1 *1) Le paramètre $\sigma_\varepsilon^2=Var(\varepsilon_i)=Var(y_i)$ est la variance résiduelle estimer par :*

$$S^2=\hat{\sigma}_\varepsilon^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

2)

$$S^2 \text{ est un estimateur sans biais} \Longleftrightarrow \mathbb{E}\left[S^2\right] = \sigma_\varepsilon^2.$$

3)

$$\mathbb{E}\left[\hat{\varepsilon}_i\right] = 0 \text{ et } Var(\hat{\varepsilon}_i) = \sigma_\varepsilon^2\left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_x}\right).$$

4) La somme des résidus est nulle

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

5) La moyenne empirique de $\hat{\varepsilon}_i$ est nulle

$$\overline{\hat{\varepsilon}_i} = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$$

6) $\frac{(n-2)S^2}{\sigma_\varepsilon^2} \sim \chi^2_{n-2}$, où χ^2_{n-2} est la loi de Khi-2 à $(n-2)$ ddl.

7) β et $\hat{\sigma}_\varepsilon^2$ sont indépendants.

Preuve. 1) On a d'après la méthode du maximum de vraisemblance estimer la variance σ_ε^2 par

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Mais c'est un estimateur biaisé, d'espérance égale à $\frac{n-2}{n}\sigma_\varepsilon^2$. On choisit donc plutôt :

$$S^2 = \hat{\sigma}_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Qui est un estimateur sans biais de σ_ε^2.

2) On a

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

Premièrement, le calcul de $\sum_{i=1}^{n}\hat{\varepsilon}_i^2$, donne

$$\begin{aligned}\hat{\varepsilon}_i = y_i - \hat{y}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = y_i - \bar{y} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i,\\ &= \beta_0 + \beta_1 x_i + \varepsilon_i - \beta_0 - \beta_1\bar{x} - \bar{\varepsilon} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i,\\ &= (\beta_1 - \hat{\beta}_1)(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}).\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 &= \sum_{i=1}^{n}\Big[(\beta_1 - \hat{\beta}_1)(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})\Big]^2,\\ &= (\beta_1 - \hat{\beta}_1)^2\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\beta_1 - \hat{\beta}_1)\sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) + \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2.\end{aligned}$$

D'après la formule

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

$$\Longrightarrow (\beta_1 - \hat{\beta}_1)\sum_{i=1}^{n}(x_i - \bar{x})^2 = -\sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}).$$

D'où

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 - (\beta_1 - \hat{\beta}_1)^2\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Donc

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2\right] - \mathbb{E}\left[(\beta_1 - \hat{\beta}_1)^2\right]\sum_{i=1}^{n}(x_i - \bar{x})^2,\\ &= \sum_{i=1}^{n}\mathbb{E}\left[\varepsilon_i^2\right] - n\mathbb{E}\left[\bar{\varepsilon}^2\right] - Var(\hat{\beta}_1)\sum_{i=1}^{n}(x_i - \bar{x})^2,\\ &= n\sigma_\varepsilon^2 - n(\frac{\sigma_\varepsilon^2}{n}) - \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = (n-2)\sigma_\varepsilon^2,\end{aligned}$$

$$\Longrightarrow \mathbb{E}\left[\hat{\sigma}_{\varepsilon}^2\right] = \frac{1}{(n-2)}\mathbb{E}\left[\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right] = \sigma_{\varepsilon}^2.$$

3)

$$\mathbb{E}\left[\hat{\varepsilon}_i\right] = \mathbb{E}\left[y_i - \hat{y}_i\right] = \mathbb{E}\left[y_i\right] - \mathbb{E}\left[\hat{y}_i\right] = \beta_0 + \beta_1 x_i - \beta_0 + \beta_1 x_i = 0.$$

$$\begin{aligned}
Var(\hat{\varepsilon}_i) &= Var(y_i - \hat{y}_i) = Var(y_i) + Var(\hat{y}_i) = Var(\varepsilon_i) + Var(\hat{\beta}_0 - \hat{\beta}_1 x_i),\\
&= \sigma_{\varepsilon}^2 + Var(\hat{\beta}_0) + x_i^2 Var(\hat{\beta}_1) + 2x_i Cov(\hat{\beta}_0, \hat{\beta}_1),\\
&= \sigma_{\varepsilon}^2 + \frac{\sigma_{\varepsilon}^2 \sum_{i=1}^{n} x_i^2}{nS_x} + x_i^2 \frac{\sigma_{\varepsilon}^2}{S_x} - 2x_i \frac{\sigma_{\varepsilon}^2 \overline{x}}{S_x},\\
&= \sigma_{\varepsilon}^2 (1 + \frac{(S_x + n\overline{x}^2)}{nS_x} + \frac{x_i^2}{S_x} - 2x_i \frac{\overline{x}}{S_x}),\\
&= \sigma_{\varepsilon}^2 (1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_x}).
\end{aligned}$$

4)

$$\begin{aligned}
\sum_{i=1}^{n}\hat{\varepsilon}_i &= \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i),\\
&= \sum_{i=1}^{n}(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i),\\
&= \sum_{i=1}^{n}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x}) = 0.
\end{aligned}$$

Car

$$\sum_{i=1}^{n}(y_i - \bar{y}) = 0 \text{ et } \sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

5)

$$\bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})\right] = 0.$$

6) On a $\varepsilon_i \sim N(0,\sigma_\varepsilon^2) \Longrightarrow \frac{\varepsilon_i}{\sigma_\varepsilon} \sim N(0,1)$ et $\frac{\hat{\varepsilon}_i}{\sigma_\varepsilon} \sim N(0,1) \Longrightarrow \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sigma_\varepsilon^2} \sim \chi_n^2$
$\Longrightarrow \frac{(n-2)S^2}{\sigma_\varepsilon^2} \sim \chi_{n-2}^2$.

7)

$$Cov(\beta,\hat{\sigma}_\varepsilon^2) = Cov(\beta,\frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2) = \frac{1}{n-2}Cov(\beta,\sum_{i=1}^{n}\hat{\varepsilon}_i^2),$$

$$\Longrightarrow Cov(\beta,\sum_{i=1}^{n}\hat{\varepsilon}_i^2) = \mathbb{E}\left[\beta\right]\mathbb{E}\left[\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right] - \mathbb{E}\left[\beta\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right],$$
$$= \beta(n-2)\sigma_\varepsilon^2 - \beta(n-2)\sigma_\varepsilon^2 = 0\,;$$

$Cov(\beta,\hat{\sigma}_\varepsilon^2) = 0$ donc β et $\hat{\sigma}_\varepsilon^2$ sont indépendants. ■

Les différents résidus

Résidu standardisé Le résidu standardisé, appelé également résidu studentisé interne dans certains ouvrages, s'intéresse à l'importance du résidus observé

$$\hat{\varepsilon} = Y - \hat{Y} = (1-H)Y.$$

S'il est anormalement élevé, en valeur absolue, le point a été mal reconstitué par le modèle : il s'écarte ostensiblement de la relation modélisée entre les exogènes et l'endogène ; Si par hypothèse, la variance de l'erreur $\sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2$ est constante, il en va autrement du résidu

$$\sigma_{\hat{\varepsilon}_i}^2 = \sigma_\varepsilon^2(1-h_i).$$

Où h_i est l'élément (i) de la matrice

$$H = X(X^tX)^{-1}X^t.$$

Nous devons donc normaliser le résidu par son écart-type, pour rendre les écarts comparables d'une observation à l'autre. Lorsque nous travaillons sur un échantillon, nous ne dispose pas de la vraie valeur de σ_ε^2, nous estimons la variance des résidus avec

$$\hat{\sigma}^2_{\hat{\varepsilon}_i} = \hat{\sigma}^2_\varepsilon(1 - h_i).$$

Où

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

Est l'estimateur de la variance résiduelle. Donc le résidu standardisé est défini par :

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{\hat{\varepsilon}_i}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_\varepsilon \sqrt[2]{(1-h_i)}}.$$

Résidu studentisé : La statistique t_i a cependant un problème : $\hat{\varepsilon}_i$ n'est pas indépendant de $\hat{\sigma}_\varepsilon$; Il n'est donc pas possible de déduire la distribution de probabilité des t_i. On préférera calculer des résidus studentisés défini par :

$$t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_\varepsilon(i) \sqrt[2]{(1-h_i)}}.$$

Où $\hat{\sigma}^2_\varepsilon(i)$ est la variance estimée des erreurs en réalisant une régression sans l'individu i. En procédant ainsi, le numérateur et le dénominateur de t_i^* sont indépendants ; Les t_i^* ont donc une distribution de Student à $n-p-1$ degrés de liberté.

1.1.3 Lois des estimateurs :

Théorème 1.1.3 ***(Lois des estimateurs avec variance résiduelle connue).***

Sous l'hpothèse de normalité des résidus et si σ_ε^2 est connue, les estimateurs $\hat{\beta}_0$, $\hat{\beta}_1$ et le vecteur $(\hat{\beta}_0, \hat{\beta}_1)^t$ suivent respectivements les lois Normales, comme suit :

i)

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0}).$$

ii)

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1}).$$

iii)

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \Gamma(\hat{\beta}_0, \hat{\beta}_1) \right).$$

Où

$$\Gamma(\hat{\beta}_0, \hat{\beta}_1) = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix} = \frac{\sigma^2_\varepsilon}{S_x} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix}.$$

Est la matrice de variance-covariance de $(\hat{\beta}_0, \hat{\beta}_1)$.

Théorème 1.1.4 ***(Lois des estimateurs avec variance résiduelle estimée).***

Si σ^2_ε *est inconnue, dans ce cas* σ^2_ε *est estimé par* S^2*, nous avons :*

i)

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim T_{n-2}.$$

Où T_{n-2} *la loi de Student à* $(n-2)$ *degrés de liberté (ddl).*

ii)

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim T_{n-2}.$$

iii)

$$\frac{1}{2} \begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix}^t \hat{\Gamma}^{-1}(\hat{\beta}_0, \hat{\beta}_1) \begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix} \sim F(2, n-2).$$

Où $F(2, n-2)$ *est la loi de Fisher à* 2 *et* $(n-2)$ *ddl.*

1.1.4 Tests d'hypothèses et intervalle de confiance

Tests d'hypothèses sur β_0 :

On souhaite tester l'hypothèse nulle $(H_0) : \beta_0 = b$ contre l'alternative $(H_1) : \beta_0 \neq b$ (test bilatère) ou $\beta_0 < b$ ou $\beta_0 > b$ (tests unilatères) où $b \in \mathbb{R}^*$,

On utilise alors comme statistique de test :

$$T_0 = \frac{\hat{\beta}_0 - b}{\hat{\sigma}_{\hat{\beta}_0}} \sim T_{n-2}.$$

Sous (H_0).

On peut ensuite prendre, pour un niveau $\alpha \in]0, 1[$, comme région de rejet ou région critique dans le cas d'un test bilatère :

$$R_{(H_0)} = \left\{ |T_0| \geq t_{n-2}(1 - \frac{\alpha}{2}) \right\}.$$

Le test de significativité

$$(H_0) : \beta_0 = 0 \text{ contre } (H_1) : \beta_0 \neq 0.$$

Permet de tester l'utilité de la constante β_0 dans le modèle.

Tests d'hypothèses sur β_1

On souhaite tester l'hypothèse nulle $(H_0) : \beta_1 = b$ contre l'alternative $(H_1) : \beta_1 \neq b$ (test bilatère) ou $\beta_1 < b$ ou $\beta_1 > b$ (tests unilatères) où $b \in \mathbb{R}^*$,

On utilise alors comme statistique de test :

$$T_1 = \frac{\hat{\beta}_1 - b}{\hat{\sigma}_{\hat{\beta}_1}} \sim T_{n-2}.$$

Sous (H_0).

On peut prendre, pour un niveau $\alpha \in]0,1[$, comme région de rejet ou région critique dans le cas d'un test bilatère :

$$R_{(H_0)} = \left\{ |T_1| \geq t_{n-2}(1-\frac{\alpha}{2}) \right\}.$$

Le test de significativité

$$(H_0) : \beta_1 = 0 \text{ contre } (H_1) : \beta_1 \neq 0.$$

Permet de tester l'utilité du modèle de régression.

Intervalles et régions de confiance pour les coefficients de régression

Théorème 1.1.5 *Soit $\alpha \in]0,1[$, on note $t_{n-2}(u)$ et $f_{2,n-2}(u)$ les $u-$quantiles respectifs des lois T_{n-2} et $F_{2,n-2}$.*

1) Un intervalle de confiance pour β_0 au seuil de confiance $\alpha\%$, est donné par :

$$\beta_0 \in \left[\hat{\beta}_0 - t_{n-2}(1-\frac{\alpha}{2})\hat{\sigma}_{\hat{\beta}_0}, \hat{\beta}_0 + t_{n-2}(1-\frac{\alpha}{2})\hat{\sigma}_{\hat{\beta}_0} \right].$$

2) Un intervalle de confiance pour β_1 au seuil de confiance $\alpha\%$, est donné par :

$$\beta_1 \in \left[\hat{\beta}_1 - t_{n-2}(1-\frac{\alpha}{2})\hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{n-2}(1-\frac{\alpha}{2})\hat{\sigma}_{\hat{\beta}_1} \right].$$

3) Une région de confiance simultanée pour $\beta = (\beta_0, \beta_1)$ au seuil de confiance $\alpha\%$, est donné par :

$$\left\{ \beta, \frac{1}{2S^2} \left(n(\hat{\beta}_0 - \beta_0)^2 + 2n\bar{x}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} x_i^2 \right) \leq f_{2,n-2}(1-\alpha) \right\}.$$

Intervalles de confiance et tests d'hypothèses sur la variance

Théorème 1.1.6 *Soit $\alpha \in]0,1[$, on note $c_{n-2}(u)$ le $u-$quantile de loi χ^2_{n-2}.*
Un intervalle de confiance de niveau de $(1-\alpha)\%$ pour σ^2_ε, est donné par :

$$\sigma^2_\varepsilon \in \left[\frac{(n-2)S^2}{c_{n-2}(1-\frac{\alpha}{2})}, \frac{(n-2)S^2}{c_{n-2}(\frac{\alpha}{2})}\right].$$

Si l'on souhaite tester l'hypothèse nulle $(H_0) : \sigma^2_\varepsilon = k^2$ contre l'alternative $(H_1) : \sigma^2_\varepsilon \neq k^2$ (test bilatère) ou $\sigma^2_\varepsilon < k^2$ ou $\sigma^2_\varepsilon > k^2$ (tests unilatères). On utilise comme statistique de test tel que $S^2 = \hat{\sigma}^2_\varepsilon$,

$$H^2 = \frac{n-2}{k^2}S^2.$$

Qui suit sous l'hypothèse (H_0) la loi χ^2_{n-2}.

1.1.5 Qualité d'ajustement

Pour juger la qualité d'ajustement du modèle, nous utilisons la formule de l'analyse de la variance suivante :

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

SCT : somme des carrés totales (ou variabilité totale),

SCR : somme des carrés résiduelles (ou variabilité résiduelle),

SCE : somme des carrés expliques (ou variabilité expliquée).

Coefficient de détermination

Définition 1.1.3 *Le coefficient de détermination R^2, est défini par :*

$$R^2 = \frac{S^2_{xy}}{S_x S_y} = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \leq 1.$$

Remarque 1.1.1 *Le coefficient R^2 donne la proportion de variabilité de y qui est expliquée par le modèle. Plus R^2 proche de* 1*, meilleure est l'adéquation du modèle aux données.*

Table d'analyse de la variance (ANOVA)

Source de variation	Somme des carrés	ddl	carré moyen	F
régression (expliquée)	SCE	1	SCE	$\frac{SCE}{SCR/n-2}$
Résiduelle	SCR	$n-2$	$\frac{SCR}{n-2}$	
Totale	SCT	$n-1$	$\frac{SCT}{n-1}$	

TAB. 1.2 – Table d'analyse de la variance (ANOVA)

Le statistique F dite statistique de Fisher permet de tester

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

On rejette H_0 si : $F > f_{1,n-2}(1-\alpha)$, où $f_{1,n-2}(1-\alpha)$ est le fractile d'ordre $1-\alpha$ d'une loi $F(1, n-2)$.

Exemple 1.1.2 *La distribution suivante montre une relation linéaire simple :*

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad , \quad i = \overline{1,10} \quad et, \ \varepsilon \sim N(0, \sigma_\varepsilon^2).$$

Entre l'évolution du taux de croissance économique Y (en %), et le prix de pétrole X (en \$) pendant 10 *ans. Les données sont présentées dans le tableaux suivante :*

Pour estimer les paramètres du modèle, nous utilisant les résultats du tableau précédent qui compléte par les formules suivantes :

	$x(\$)$	$y(\%)$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$	$\hat{y}$	$\hat{\varepsilon}$	$\hat{\varepsilon}^2$
1	111	1.9	57.76	0.44	5.016	2.99	-1.09	1.18
2	120	4.4	275.56	9.99	52.456	5.06	-0.66	0.43
3	122	5.6	345.96	19.01	81.096	5.52	0.08	0.01
4	105	4.1	2.56	8.18	4.576	1.61	2.49	6.2
5	101	3.2	5.76	3.84	-4.704	0.69	2.51	6.3
6	95	-3	70.56	17.98	35.616	-0.69	-2.31	5.33
7	92	-2	129.96	10.50	36.936	-1.38	-0.62	0.38
8	90	0	179.56	1.54	16.616	-1.84	1.84	3.38
9	96	-2.9	54.76	17.14	30.636	-0.46	-2.44	5.95
10	102	1.1	1.96	0.02	0.196	0.92	0.18	0.03
$\sum$	1034	12.4	1124.4	88.62	258.44	12.4	0	29.22

TAB. 1.3 – Table des données

Les moyennes empriques de x et y :

$$\begin{cases} \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(1034) = 103.4, \\ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10}(12.4) = 1.24. \end{cases}$$

La droite de régression estimé : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$; Avec

$$\begin{cases} \hat{\beta}_1 = \frac{S_{xy}}{S_x} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sum_{i=1}^{n}(x-\bar{x})^2} = \frac{258.44}{1124.4} = 0.23, \\ \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 1.24 - (0.23)(103.4) = -22.53. \end{cases}$$

Donc

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 0.23x_i - 22.53.$$

Et les résidus :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \quad , \forall i = \overline{1, 10}.$$

La variance résiduelle estimer :

$$S^2 = \hat{\sigma}_{\varepsilon}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{8}(29.22) = 3.65\,;$$

Les estimateurs des variances de β_0 et β_1 :

$$\begin{cases} \widehat{Var(\hat{\beta}_0)} = \hat{\sigma}_{\hat{\beta}_0}^2 = \dfrac{\hat{\sigma}_{\varepsilon}^2\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{(3.65)(108040)}{10(1124.4)} = 35.07, \\ \widehat{Var(\hat{\beta}_1)} = \hat{\sigma}_{\hat{\beta}_1}^2 = \dfrac{\hat{\sigma}_{\varepsilon}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{3.65}{1124.4} = 0.003. \end{cases}$$

Testez la signification des paramètres du modèle au niveau $(1-\alpha)\% = 95\%$:

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

$$\Longrightarrow |T_0| = \frac{\left|\hat{\beta}_0 - 0\right|}{\hat{\sigma}_{\hat{\beta}_0}} = \frac{22.53}{\sqrt[2]{35.07}},$$

$$= 3.8 > t_{n-2}(1-\frac{\alpha}{2}) = t_8(0.975) = 2.306\,;$$

Alors on rejet H_0, et par consequent β_0 est significatif.

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

$$\Longrightarrow |T_1| = \frac{\left|\hat{\beta}_1 - 0\right|}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0.23}{\sqrt[2]{0.003}},$$

$$= 4.2 > t_{n-2}(1-\frac{\alpha}{2}) = t_8(0.975) = 2.306\,;$$

Alors on rejet H_0, et par consequent β_1 est significatif.

1) Un intervalle de confiance pour β_0 au seul de confiance $(1-\alpha)\% = 90\%$, est donné par :

$$\begin{aligned}
\beta_0 &\in \left[\hat{\beta}_0 - t_{n-2}(1-\frac{\alpha}{2})\hat{\sigma}_{\hat{\beta}_0}, \hat{\beta}_0 + t_{n-2}(1-\frac{\alpha}{2})\hat{\sigma}_{\hat{\beta}_0}\right], \\
&= \left[(-22.53) - t_8(0.95)\sqrt[2]{35.07}, (-22.53) + t_8(0.95)\sqrt[2]{35.07}\right], \\
\beta_0 &\in [-33.5428880\ ,\ -11.5094777]\,.
\end{aligned}$$

2) Un intervalle de confiance pour β_1 au seul de confiance $(1-\alpha)\% = 90\%$, est donné par :

$$\begin{aligned}
\beta_1 &\in \left[\hat{\beta}_1 - t_{n-2}(1-\frac{\alpha}{2})\hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{n-2}(1-\frac{\alpha}{2})\hat{\sigma}_{\hat{\beta}_1}\right], \\
&= \left[0.23 - t_8(0.95)\sqrt[2]{0.003}, 0.23 + t_8(0.95)\sqrt[2]{0.003}\right], \\
\beta_1 &\in [0.1238584\ ,\ 0.3358357]\,.
\end{aligned}$$

La qualité qui mesure la bonne explication de la variable à expliquer par la variable explicative, est la coefficient de détermination :

$$\begin{aligned}
R^2 = \frac{SCE}{SCT} = \frac{S_{xy}^2}{S_x S_y} &= \frac{\left(\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})\right)^2}{\sum_{i=1}^{n}(x-\bar{x})^2 \sum_{i=1}^{n}(y-\overline{y})^2}, \\
&= \frac{(258.44)^2}{(1124.4)(88.62)} = 0.67 = 67\%.
\end{aligned}$$

C'est une valeur peu satisfaisante ($\sim \frac{1}{3}$ y et x sont indépendantes).

Table d'analyse de la variance (ANOVA) :

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 88.62.$$

Et on a

$$SCE = R^2 SCT = 0.67(88.62) = 59.37.$$

Nous utilisons la formule de l'analyse de la variance

$$SCT = SCE + SCR.$$

On trouve

$$SCR = SCT - SCE = 88.62 - 59.37 = 29.25.$$

Alors

Source de variation	Somme des carrés	ddl	carré moyen	F
régression (expliquée)	$SCE = 59.37$	1	59.37	$\frac{59.37}{3.65} = 16.26$
Résiduelle	$SCR = 29.25$	8	$\frac{29.245}{8} = 3.65$	
Totale	$SCT = 88.62$	9	$\frac{88.62}{9} = 9.85$	

TAB. 1.4 – Table d'analyse de la variance (ANOVA)

Tester la validité du modèle :

$$\begin{cases} H_0 : \beta_0 = 0 \text{ et } \beta_1 = 0, \\ H_1 : \beta_0 \neq 0 \text{ et } \beta_1 \neq 0. \end{cases}$$

$F = 16.26 > f_{1-\alpha}(1, n-2) = f_{0.95}(1, 8) = 5.32$, alors le modèle est valide à 95%.

Chapitre 2

Application sous R

2.1 Analyse des résidus d'un modèle de régression linéaire

2.1.1 Exemple sur la régression linéaire simple

On a regroupé dans le tableau ci-dessous, des mesures de la surface de couverture (mesurée depuis le ciel) notée SC (variable explicative x), par rapport à la surface de couverture mesurée à $1m$ du sol (obtenues par des mesures au sol) notée ST (variable à expliquer y). Les mesures ont été realisées pour un échantillon de $n = 50$ arbres différents d'une forêt tropicale.

Modèle

On cherche à modaliser la relation entre la variable $y = ST$ et la variable $x = SC$; Le modèle la plus simple est la régression linéaire simple qui s'écrit : $\forall i = \overline{1, 50}$ tel que :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{2.1}$$

Nr	ST	SC	Nr	ST	SC	Nr	ST	SC
1	6.34	2.43	18	12.46	2.80	35	21.27	7.04
2	6.03	1.73	19	20.15	6.09	36	9.24	3.23
3	12.61	3.29	20	13.43	3.78	37	8.66	2.78
4	17.34	6.80	21	3.73	0.97	38	9.22	2.83
5	15.25	4.41	22	9.80	3.45	39	7.99	2.53
6	15.57	4	23	5.46	1.74	40	6.55	2.40
7	15.80	4.46	24	12.24	3.68	41	12.71	3.49
8	12.93	3.36	25	16.44	6.01	42	10.97	3.17
9	13.82	4.06	26	18.86	4.23	43	9.83	2.69
10	11.28	3.34	27	12.25	4.63	44	5.02	1.35
11	12	3.67	28	17.52	4.72	45	7.88	2.63
12	11.52	3.72	29	5.07	1.59	46	8.57	2.39
13	8.05	2.65	30	11.88	4.78	47	9.75	2.92
14	5.05	1.80	31	12.65	3.60	48	13.27	3.99
15	10.63	3.17	32	8.63	2.62	49	10.05	3.13
16	15.83	5.48	33	25.94	7.84	50	8.85	3.67
17	12.71	4.20	34	7.22	2.56			

TAB. 2.1 – Table des données

Lecture des données $> x < -c(2.43, 1.73, 3.29, 6.80, 4.41, 4, 4.46, 3.36, 4.06, 3.34, 3.67,$

$3.72, 2.65, 1.80, 3.17, 5.48, 4.20, 2.8, 6.09, 3.78, 0.97, 3.45, 1.74, 3.68,$

$6.01, 4.23, 4.63, 4.72, 1.59, 4.78, 3.6, 2.62, 7.84, 2.56, 7.04, 3.23, 2.78,$

$2.83, 2.53, 2.4, 3.49, 3.17, 2.69, 1.35, 2.63, 2.39, 2.92, 3.99, 3.13, 3.67)$

$> y < -c(6.34, 6.03, 12.61, 17.34, 15.25, 15.57, 15.80, 12.93, 13.82,$

$11.28, 12, 11.52, 8.05, 5.05, 10.63, 15.83, 12.71, 12.46, 20.15, 13.43,$

$3.73, 9.80, 5.46, 12.24, 16.44, 18.86, 12.25, 17.52, 5.07, 11.88, 12.65,$

$8.63, 25.94, 7.22, 21.27, 9.24, 8.66, 9.22, 7.99, 6.55, 12.71, 10.97,$

$9.83, 5.02, 7.88, 8.57, 9.75, 13.27, 10.05, 8.85)$

Inspection graphique : On trace le nuage de points $\{(x_i, y_i), i \in \{1, ..., n\}\}$ par les commandes :

$> plot(x, y, xlab = "SC", ylab = "ST")$

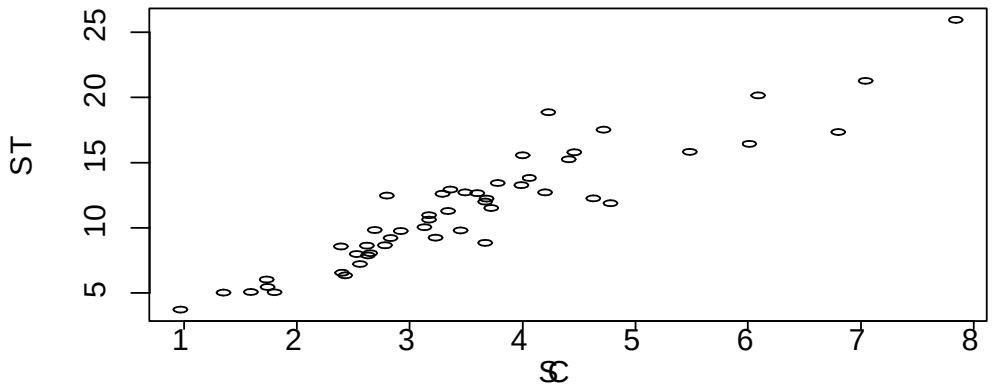

FIG. 2.1 – La surface de couverture ST en fonction de la surface de couverture SC.

Estimation des paramètres La modélisation de la RLS et les estimations des paramètres par la méthode des MCO, s'obtiennent par les commandes :

$> linreg < -lm(y\tilde{}x)$

$> summary(linreg)$

Appel :

$lm(formula = y\tilde{}x)$

Résidus :

Min	$1Q$	$Médian$	$3Q$	Max
-3.7815	-0.8722	-0.1032	0.9676	5.3447

Coefficients :

	$Estimations$	$\hat{\sigma}_{\hat{\beta}}$	T	$Pr(>\|t\|)$
β_0	0.9961	0.6492	1.534	0.132
β_1	2.9596	0.1694	17.474	$< 2e-16 * **$

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1,

Erreur standard résiduelle : 1.708 au 48 degrés de libertée,

R-carré multiple : 0.8642 , R-carré ajusté : 0.8613,

F-statistique : 305.3 au 1 et 48 ddl, p-value : $< 2.2e-16$,

Estimations ponctuelles de β_0 et β_1 : $\hat{\beta}_0 = 0.9961$ et $\hat{\beta}_1 = 2.9596$; Et l'écart-type résiduelle estimer $\hat{\sigma}_\varepsilon = 1.708$;

Estimations ponctuelles des l'écart-types de $\hat{\beta}_0$ et $\hat{\beta}_1$: $\hat{\sigma}_{\hat{\beta}_0} = 0.6492$ et $\hat{\sigma}_{\hat{\beta}_1} = 0.1694$.

T : $\begin{array}{lll} H_1 & \beta_0 \neq 0 & \beta_1 \neq 0 \\ T & 1.534 & 17.474 \end{array}$ avec $t_{n-2}(1-\frac{\alpha}{2}) = t_{48}(0.975) = 2.011$,

Test de Student pour β_0 : influence de X sur Y : p-value < 1, n'est pas significativement différent de 0.

Test de Student pour β_1 : influence de X sur Y : p-value < 0.001, $***$: hautement significative,

$R^2 = 0.8642$ et $\bar{R}^2 = 0.8613$: cela est satisfaisant,

Test de Fisher : $p - value = 2.2e - 16 < 0.001$, $***$: l'utilisation du modèle de RLS est pertinente.

Aux intervalles de confiance (IC) pour β_0 et β_1 au niveau 95%, les commandes sont :

$> confint(linreg, level = 0.95)$

	2.5%	97.5%
β_0	-0.309249	2.301392
β_1	2.619078	3.300162

Cela renvoie :

$$\begin{cases} \beta_0 \in [-0.309249\ ,\ 2.301392]\,, \\ \beta_1 \in [2.619078\ ,\ 3.300162]\,. \end{cases}$$

L'équation de la droite de régression est :

$$\hat{y}_i = 0.9961 + 2.9596 x_i \quad \forall i = \overline{1, 50}.$$

Calculer par la fonction $fitted(linreg)$; On la visualise avec les commandes :

$> plot(x, y, xlab = "SC", ylab = "ST")$

$> abline(linreg, col = "red")$

Tableau d'analyse de la variance On peut utilise la fonction $anova()$;

$> anova(linreg)$

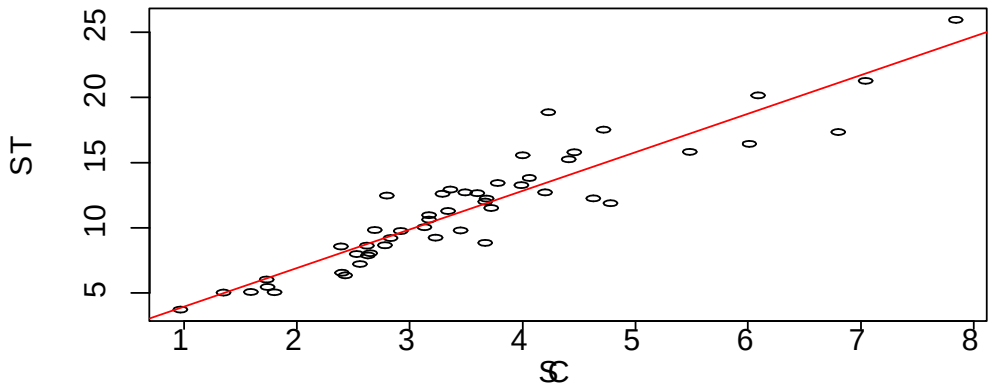

FIG. 2.2 – Nuage des points et droite de régression

Table d'analyse de la variance :

Réponse : y

$Variation$	ddl	SC	MC	F	$Pr(>F)$
$Expliquée$	1	890.40	890.40	305.35	$< 2.2e-16 * **$
$Résiduelle$	48	139.97	2.92		

—

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Remarque 2.1.1 *La table d'analyse de la variance donne les résultats du test de la validité du modèle ; On observe que* $F = 305.35 > f_{1,48}(0.95) = 4.04$*, alors le modèle est valide (signicatif globalement).*

2.1.2 Validation du modèle

Le modèle de RLS 2.1suppose que la régression est linéaire, les termes d'erreurs ont même variance, qu'ils sont indépendants, et enfin issues d'une loi gaussienne. Avant d'effectuer l'analyse de la régression, il est indispensable de vérifier ces l'hypothèses. Cette étude se généralisé à tous les modèles que nous verrons dans la suite, sauf l'hypothèse de la linéarité qui est spécifique au modèle de RLS.

L'examen de la validité des hypothèses du modèles, se fait à partir du graphe des résidus

$\hat{\varepsilon}_i$ en fonction des prédictions $\hat{y}_i$. Rappelons que la prédiction est :

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

Et les résidus estimés sont définis par :

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

On trace le nuage des points $(\hat{y}_i, \hat{\varepsilon}_i)$ en faisant :

$> linreg < -lm(y\tilde{}x)$

$> plot(linreg, 1)$

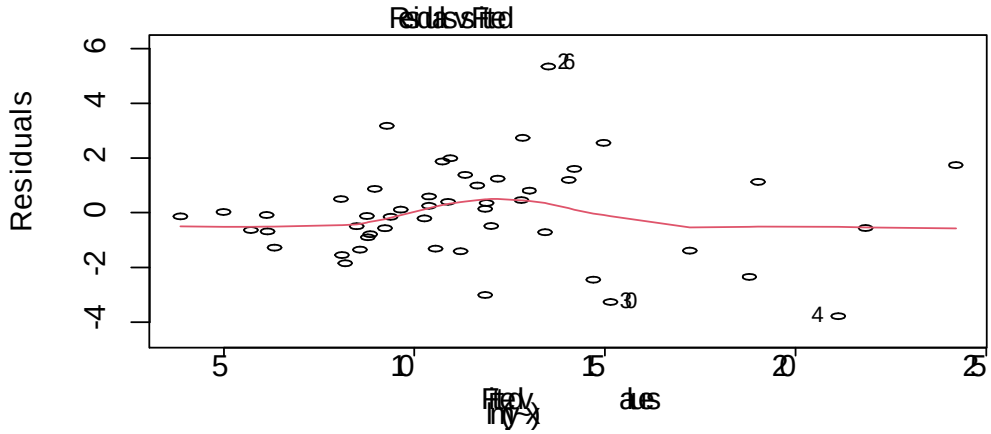

FIG. 2.3 – Graphe des résidus en fonction des prédictions

On constate que le nuage des points obtenu n'est pas ajustable par une "ligne", et la moyenne des valeurs de la ligne rouge est nulle ; On admet que ε et x sont indépendantes.

Linéarité de la relation

On peut juger de la linéarité entre x et y en visualisant le graphique des couples (x_i, y_i). Par exemple, prenons la figure 2.1 elle montre clairement que la relation entre la surface de couverture ST et SC est linéaire. (quand on parle de modèle linéaire, cela signifie linéaire

par rapport aux paramètres β_0 et β_1, une relation entre x et y qui n'est pas linéaire ne dit pas que le modèle proposé n'est pas linéaire).

Test d'existence d'une relation linéaire : on souhaite réaliser le test d'hypothèse suivant :

$$H_0 : \beta_1 = 0 \text{ contre } H_1 : \beta_1 \neq 0.$$

Ce test permet de savoir s'il existe une relation linéaire statistiquement significative entre x et y. La règle de décision est :

Si

$$z_n = \sqrt[2]{\frac{(n-2)R^2}{1-R^2}} \text{ vérifie } |z_n| > t_{1-\frac{\alpha}{2}}(n-2).$$

Alors on peut rejeter H_0. D'après cette exemple, $R^2 = 0.86$ donc on trouve $|z_n| = 17.17 > t_{1-\frac{\alpha}{2}}(n-2) = t_{0.975}(48) = 2.011$ rejet H_0 avec un risque $\alpha = 0.05$, cela signifie qu'il existe bien une relation linéaire entre x et y, puisque la pente de la droite est significativement non nulle.

Normalité des résidus

C'est l'hypothèse la moins importante, car d'une part le modèle linéaire est robuste à la normalité, et d'autre part les résidus suivent asymptotiquement une loi normale (i.e pour des grands échantillons). Nous pouvons vérifier l'hypothèse de normalité en utilisant des méthodes graphiques comme : Graphique Q-Q plot ou Histogramme des résidus, et par des tests comme : Jarque-Bera ou Shapiro-Wilk ou Kolmogorov-Smirnov.

a) Graphique Q-Q plot : (quantile-quantile plot) est appelé "Droite de Henry", est un graphique "nuage de points" qui vise à confronter les quantiles de la distribution empirique, et les quantiles d'une distribution théorique normale, de moyenne et l'écart-type estimés sur les valeurs observées ; Si la distribution est compatiple avec la loi normale, les points forment une droite. On trace le QQ plot associé la modèle par les commondes suivantes :

$> linreg < -lm(y\tilde{}x)$

$> qqnorm(residuals(linreg), xlab = "", ylab = "", main = "")$

$> qqline(residuals(linreg))$

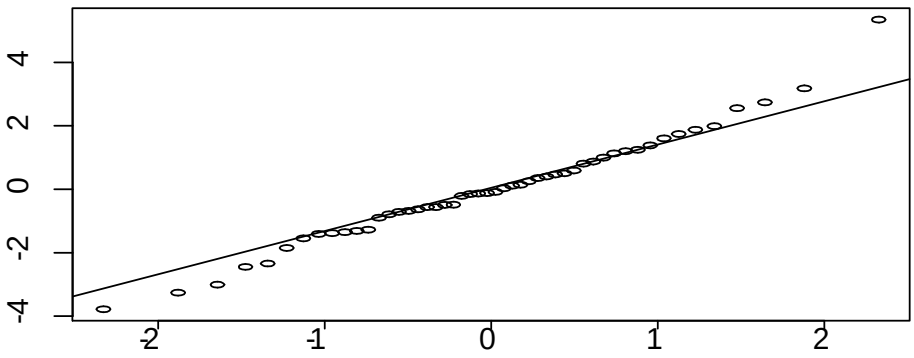

FIG. 2.4 – Le graphique Q-Q plot

On constate que les points sont à peu près alignés, sauf quelque des points aberrants ; Ce graphique traduit la normalité des résidus.

b) Histogramme des résidus : on obtient également une représentation des résidus du modèle, dont on peut vérifier la compatibilité avec la distribution Gaussienne. En utilisant les commandes :

$> hist(residuals(linreg), breaks = 7, freq = FALSE, xlab = "", main = "", ylab = "", col = "blue")$

$> plot(function(w)dnorm(w, mean(residuals(linreg)), sd(residuals(linreg))), -4, 6, col = "red", lwd = 2, add = TRUE)$

Cet histogramme semble indiquer que la loi est proche d'une loi normale ; Donc l'hypothèse de normalité est vérifiée.

c) Test de Jarque-Bera : Ce test est basé sur le coefficient d'asymétrie

$$\gamma_1 = \frac{\mu_3}{\sigma^3}.$$

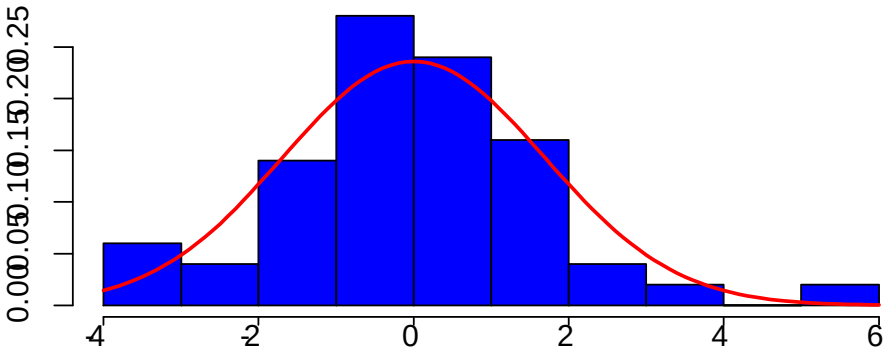

FIG. 2.5 – Histogramme des résidus

Et le coefficient d'aplatissement

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3.$$

Où μ_3 et μ_4 sont les moments centrés d'ordre 3 et 4 resp, et σ l'écart-type ; Le test d'hypothèse s'écrit de la manière suivante :

$$\begin{cases} H_0 : \varepsilon \text{ suit une loi normale, par conséquent } \gamma_1 = 0 \text{ et } \gamma_2 = 0, \\ H_1 : \varepsilon \text{ ne suit pas une loi normale, par conséquent } \gamma_1 \neq 0 \text{ et } \gamma_2 \neq 0. \end{cases}$$

La statistique de Jarque-Bera est :

$$T = \frac{n-p-1}{6}(g_1^2 + \frac{g_2^2}{4}) \sim \chi_2^2.$$

Avec

$$\begin{cases} g_1 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right)^{\frac{3}{2}}} & \text{Le coefficient d'asymétrie empirique.} \\ g_2 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right)^{2}} - 3 & \text{Estimateur de } \gamma_2. \end{cases}$$

La valeur $(n-p-1)$ ddl, nous disposons d'un échantillon de taille n et $(p+1)$ coefficients à estimer dans la régression avec constante. La région critique du test au risque α, s'écrit :

$$RC : T > \chi^2_{1-\alpha}(2).$$

Sous H_0. Par ailleurs, en utilisant les commandes suivantes pour calculer la statistique de Jarque-Bera :

$> (1/50 * sum(residuals(linreg)\hat{}3))/(1/50 * sum(residuals(linreg)\hat{}2))\hat{}(3/2)$

[1]0.3742548

$> (1/50 * sum(residuals(linreg)\hat{}4))/(1/50 * sum(residuals(linreg)\hat{}2))\hat{}2 - 3$

[1]1.017827

$> 48/6 * (0.37\hat{}2 + (1.02\hat{}2)/4)$

[1]3.176

Pour un risque $\alpha = 0.05$, le seuil critique est $\chi^2_{0.95}(2) = 5.99$; Dans cette exemple, la statistique $T = 3.176$ est largement inférieure à $\chi^2_{0.95}(2) = 5.99$; La distribution observée est compatible avec une distribution normale.

d) Test de Shapiro-Wilk : Très populaire, en comparaison des autres tests, il est particu-

lièrement puissant pour les petits effectifs ($n \leq 50$). La statistique du test s'écrit :

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i (x_{(n-i+1)} - x_{(i)})\right]^2}{\sum_i (x_i - \bar{x})^2}.$$

Où $x_{(i)}$ correspond à la série des données triées ;

$\left[\frac{n}{2}\right]$ est la partie entière du rapport $\frac{n}{2}$;

a_i sont des constantes générées à partir de la moyenne, et de la matrice de variance covariance des quantiles d'un échantillon de taille n suivant la loi normale. La région critique, rejet de la normalité, s'écrit :

$$RC : W < W_{crit}.$$

Avec les valeurs seuils W_{crit} pour différents risques α, et effectifs n sont lues dans la table de Shapiro-Wilk. Dans cette exemple, en utilisant les commandes suivantes pour calculer la statistique de Shapiro-Wilk :

$> linreg < -lm(y \tilde{} x)$

$> e = residuals(linreg)$

$> shapiro.test(e)$

Test de normalité Shapiro-Wilk

Les données : e

$W = 0.98032 \quad , \quad p - value = 0.5657$

Ici on a $p - value = 0.5657 > 0.05$, donc les résidus sont normaux.

Homoscédasticité / Hétéroscédasticité

$$\begin{cases} \text{Homoscédasticité : la variance résiduelle } \sigma_\varepsilon^2 \text{ est constante.} \\ \text{Hétéroscédasticité : la variance résiduelle } \sigma_\varepsilon^2 \text{ n'est pas constante.} \end{cases}$$

Il n'existe pas de procédure précise pour vérifier l'hypothèse d'homoscédasticité. Nous proposons plusieurs graphiques possibles pour détecter une hétéroscédasticité ; Il est recommandé de tracer les résidus studentisés t_i^* en fonction des $\hat{y}_i$. Si une structure apparaît (tendance, cône, vagues), l'hypothèse d'homoscédasticité risque fort de ne pas être vérifiée. Voyons cela sur un graphique :

$> linreg < -lm(y\tilde{}x)$

$> rstudent(linreg)$

$> fitted(linreg)$

$> plot(fitted(linreg), rstudent(linreg), xlab = "", ylab = "")$

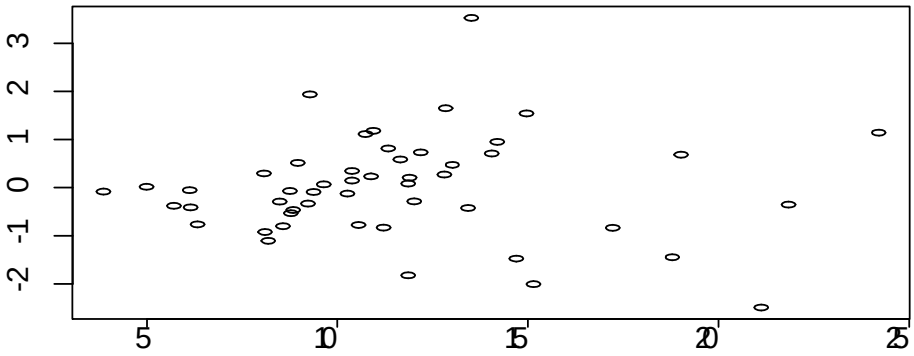

FIG. 2.6 – Graphique des résidus studentisés en fonction des prédictions

Sur cette figure, notons que les points doivent former un nuage homogène ; Donc on a homoscédasticité des résidus.

Ou bien par l'examen de la série chronologique tracée, après avoir divisé les résidus en deux parties et calculé la variance de chaque partie, nous arrivons pour conclure l'homoscédasticité ; En utilisant les commandes :

$> plot.ts(e)$

$> sd(e[1:25])$

[1] 1.624821

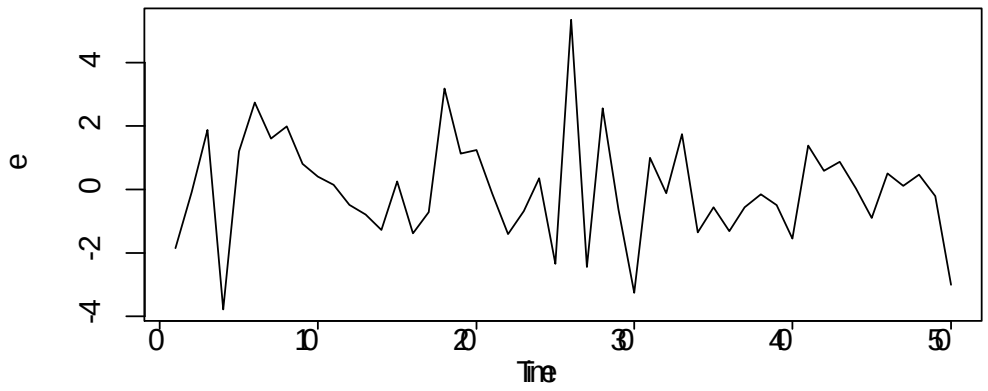

FIG. 2.7 – Série chronologique des résidus

$> sd(e[26:50])$

[1] 1.782959

$> sd(e[1:25])\hat{}2$

[1] 2.640043

$> sd(e[26:50])\hat{}2$

[1] 3.178943

Nous remarquons qu'une égalité des variances de deux parties des résidus, soit graphiquement (les vagues sont à peu près de même longueur), soit arithmétiquement (l'écart-type des deux parties des résidus est également à 1.7), donc l'hypothèse d'homoscédasticité est vérifiée.

Indépendance :(Absence d'autocorrélation)

La détection de l'autocorrélation des résidus, peut s'effectuer visuellement à l'aide du graphique des résidus ou par les tests ; On vérifie cela avec les graphiques acf :

$> e = residuals(linreg)$

$> acf(e, xlab = "", main = "", ylab = "")$

On ne constate aucune structure particulière, et peu de bâtons dépassent les bornes limites ; On admet l'indépendance des résidus. Ou bien par le graphique des résidus en fonction

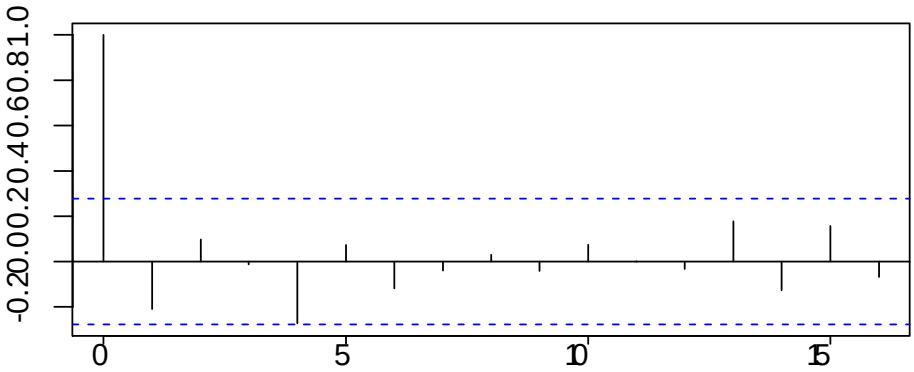

FIG. 2.8 – Graphiques ACF des résidus

des numéros d'observations t, en utilisant les commandes :

$> e = residuals(linreg)$

$> t = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25,$

$26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50)$

$> plot(t, e, xlab = "", ylab = "")$

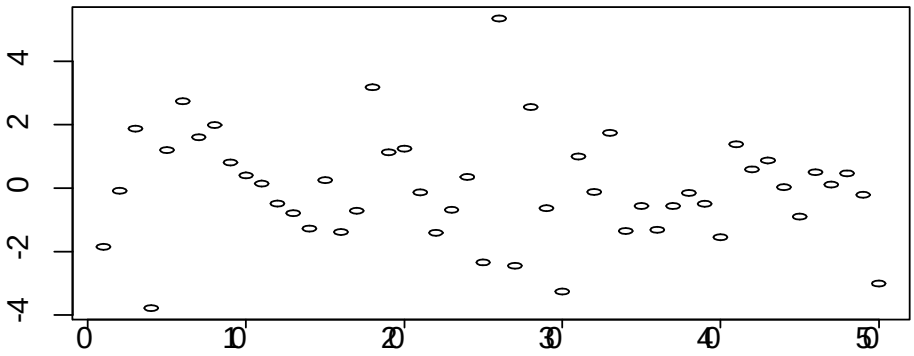

FIG. 2.9 – Graphique des résidus en fonction des numéros d'observations

On remarque que les points de la graphe sont répartis, séparés et espacés les uns des autres et ne forment pas de clusters ; Donc on admet l'absence d'auto-corrélation des résidus.

Par ailleurs, nous pouvons aussi utiliser des tests comme :

Le test de Durbin-Watson : test spécifique à une forme de l'erreur, puissant pour cette forme :

$$\varepsilon_i = \rho.\varepsilon_{i-1} + \nu_i \quad \text{, avec } \nu_i \sim N(0, \sigma_\nu).$$

Le test d'hypothèses s'écrit :

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

On utilise la statistique de Durbin-Watson :

$$d = \frac{\sum_{i=2}^{n} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}.$$

Par construction, $0 \leq d \leq 4$, $d = 2$ lorsque $\hat{\rho} = 0$. Le règle de décision résume de la manière suivante :

Acceptation de H_0 si $d_U < d < 4 - d_U$,

Rejet de H_0 si $d < d_L$ $(\rho > 0)$ ou $d > 4 - d_L$ $(\rho < 0)$,

Incertitude si $d_L < d < d_U$ ou $4 - d_U < d < 4 - d_L$.

Nous pouvons alors calculer la statistique de Durbin-Watson, en faisant :

$> linreg < -lm(y\tilde{}x)$

$> e = residuals(linreg)$

$> d = (sum((e[2:50] - e[1:49])\hat{}2))/(sum(e[1:50]\hat{}2))$

$> d$

[1] 2.330058

Nous obtenons $d =$ 2.33, pour un test bilatéral à 10%, nous récupérons les valeurs critiques dans la table de Durbin-Watson ; Pour $n = 50$ et $K = 1$, $d_L = 1.50$, et $d_U = 1.59$. Nous constatons que $d_U < d < 4 - d_U$ acceptation de H_0, les résidus ne sont pas autocorrélées.

Le test des séquences : (test de Wald Wolfowitz) est plus générique que le précédent ;

Le test repose sur la détection des séquences de valeurs positives '+' ou négatives '−' des résidus. La statistique du test r est le nombre total de séquences dans la série d'observations.

Posons n_+ (resp, n_-) le nombre des résidus positifs (resp, négatifs) dans la série des résidus. Sous l'hypothèse H_0 le processus de génération des données est aléatoire, la statistique r suit asymptotiquement une loi normale de paramètres :

$$\begin{cases} \mu_r = \frac{2n_+n_-}{n} + 1, \\ \sigma_r = \sqrt[2]{\frac{(\mu_r-1)(\mu_r-2)}{n-1}}. \end{cases}$$

Nous pouvons former la statistique centrée et réduite $z = \frac{r-\mu_r}{\sigma_r}$; La région critique du test rejet de l'hypothèse de génération aléatoire des résidus s'écrit :

$$RC : |z| > u_{1-\frac{\alpha}{2}}.$$

Où $u_{1-\frac{\alpha}{2}}$ est le fractile d'ordre $_{1-\frac{\alpha}{2}}$ de la loi normale centrée et réduite $N(0,1)$.

Nous pouvons alors calculer la statistique des séquences, en faisant :

$> fitted(linreg)$

$> e = residuals(linreg)$

Nous comptons le nombre des valeurs positives et négatives, $n_+ = n1$ et $n_- = n2$; Et le nombre de séquences r ; Nous vérifions que $n = n_+ + n_- = 50$;

$> n = n1 + n2$; $n1 = 24$; $n2 = 26$; $r = 23$

Nous pouvons calculer la moyenne $\mu_r = \mu r$, et l'écart-type $\sigma_r = c$ de la statistique de test sous l'hypothèse nulle :

$> \mu r = ((2 * n1 * n2)/n) + 1$

$> \mu r$

[1] 25.96

$> c = sqrt(((\mu r - 1) * (\mu r - 2))/(n - 1))$

$> c$

[1] 3.493556

Nous calculons enfin la statistique centrée et réduite z :

$> z = (r - \mu r)/c$

$> z$

[1] -0.8472743

Que nous comparons au fractile d'ordre 0.95 (pour un test bilatéral à 10%) de la loi normal centrée et réduite $u_{0.95} = 1.64 > |z| = 0.8472743$, donc nous sommes dans la région d'acceptation de H_0. On conclure que les résidus sont indépendantes.

Conclusion

Dans ce mémoire, nous avons discuté de l'analyse des résidus d'un modèle de régression. Nous consacrons l'étude au cas de la régression linéaire simple, car cette étude se généralise à tous les modèles que nous traitions avant comme dans le cas de la régression linéaire multiple. Sauf l'hypothèse de la linéarité qui est spécifique au modèle de régression linéaire simple.

L'examen de la validité des hypothèses du modèle (linéarité de la relation, normalité, indépendance et homoscédasticité des résidus,...), se fait à partir du graphe des résidus ou bien sous les différentes tests qui sont disponibles dans logiciel R.

En conclusion, l'analyse des résidus consiste à examiner si les hypothèses de base du modèle linéaire simple est violées. Vérifier ces hypothèses semblent incontournable pour obtenir des résultats exacts. En effet, si les hypothèses sont vérifiées alors on peut interpréter les résultats de la régression linéaire simple. Si non, il faut transformer les données ou supprimer les points aberrants, de façon à se ramener à un modèle dont les hypothèses sont valides.

Bibliographie

[1] Caussinus, H. (1980). Sur l'analyse des résidus dans le modèle linéaire. Statistique et analyse des données, 5(3), 29 – 39.

[2] Cornillon, P. A., Matzner-Lober, E. (2007). Régression : théorie et applications (pp. 302-p). Springer.

[3] Cornillon, P. A., Hengartner, N., Matzner-Lober, E., Rouvière, L. (2019). Régression avec R-2e édition. EDP sciences.

[4] Chesneau, C.(2020) Etudes ; modeles de régression.
https ://chesneau.users.lmno.cnrs.fr/etudes-reg.pdf.

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

[6] Guyader, A. (2013). Régression linéaire. Université Rennes, 2
http ://www.lpsm.paris/pageperso/guyader/files/teaching/Regression.pdf.

[7] Lejeune, M. (2004). Statistique : La théorie et ses applications. Springer Science & Business Media.

[8] Magalie, F. Modèles de régression linéaire. UNIVERSITE RENNES 2.
http ://pageperso.lif.univ-mrs.fr/~alexis.nasr/Ens/IAAAM2/SlidesModStat_C1_print.pdf.

[9] Marie, C. Régression linéaire simple, Chapitre 1 Licence 3 MIASHS-Université de Bordeaux.
http ://pageperso.lif.univ-mrs.fr/~alexis.nasr/Ens/IAAAM2/SlidesModStat_C1_print.pdf.

[10] Olivier, M. (2018). Statistiques appliquées avec introduction au logiciel R. Ellipses Marketing.

[11] Palm, R. (1986). Etude des résidus de régression : principes et application. Notes de Statistique et d'Informatique, 1, 1-13.

[12] Rakotomalala, R. (2011). Tests de normalité. Université Lumière Lyon.
http ://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf.

[13] Rakotomalala, R. (2015). Pratique de la regression lineaire multiple. Diagnostic et selection de variables.
https ://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf.

[14] Yves, T. (2011). Resumé du Cours de Modèles de Régression. Institut de statistique, Université de Neuchâtel Suisse.
https ://www.unine.ch/files/live/sites/statistics/files/shared/documents/cours_modeles_regressio

Annexe A : Logiciel *R*

Aujourd'hui beaucoup des logiciels peuvent être utilisés à des fins statistiques : Execel, SAS, SPSS, et R figurant parmi les plus utilisés ; Le système R présente l'avantage d'être un logiciel libre et gratuit spécialement conçu pour l'analyse statistique, qui permet de faire les calculs ainsi que les représentations graphiques de manière automatique. Il a été initialement crée, en 1996 par Robert Gentleman et Ross Ihaka. Depuis 1997, R est développé par une équipe" R Core Team ". Enfin, le logiciel est disponible en téléchargement gratuit pour les principaux système d'exploitation à l'adresse : https ://www.r-project.org/.
Les différentes commandes utilisées tout au long de ce mémoire sont expliquées ci-dessous :

c	Concaténation (fonction pour créer des vecteurs).
$plot(.,.)$	Trace la graphique des couples (.,.).
$linreg$	Régression linéaire.
$lm()$	La fonction qui sert à produire un régression linéaire.
$summary(lm())$	Statistique descriptives d'un modèle linéaire.
$confint()$	Intervalle de confiance des paramètres de régression.
$fitted()$	Donne les valeurs des prédictions calculées par la régression.
$abline$	Trace la droite estimée de régression à un graphe.
$anova$	La sortie de la table d'analyse de la variance.
$qqnorm(e)$	Crée un graphique qui compare e à une distribution normale théorique.
$qqline$	Trace la droite de Henry.

$residuals()$	Donne le vecteur des résidus.
$rstudent()$	Donne le vecteur des résidus studentisés.
$hist()$	Histogramme.
$sum()$	Somme des éléments d'un vecteur.
$library()$	Pour charger le package en mémoire.
$shapiro.test$	Test de normalité de Shapiro-Wilk.
$acf()$	Le corrélogramme des résidus.
$sqrt()$	Racine carrée des éléments d'un vecteur.
$sd()$	L'écart-type des éléments d'un vecteur.
$plot.ts()$	Utilisez le format des données d'origine directement pour la série chronologique.

Annexe B : Abréviations et Notations

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

y	La variable à expliquer.
$\hat{y}$	La prédiction.
x	La variable explicative.
β	Le vecteur des paramètres.
$\hat{\beta}$	Le vecteur des estimateurs.
ε	Le vecteur d'erreurs.
$\hat{\varepsilon}$	Le vecteur des résidus.
n	La taille de l'échantillon.
$\mathbb{R}$	L'ensemble des nombres réels.
$\mathbb{E}[.]$	Espérance mathématique.
$Var()$	Variance.
$Cov()$	Covariance.
σ_{ε}^2	La variance de l'erreur.
$\hat{\sigma}_{\varepsilon}^2$	La variance résiduelle estimer.
$\bar{x}, \bar{y}$	Les moyennes empiriques de x et y (respectivement).
S_x, S_y, S_{xy}	Moments empiriques.

$N(\mu, \sigma^2)$	Loi Normale d'espérance $\mu \in \mathbb{R}$ et de variance $\sigma^2 \in \mathbb{R}_+^*$.
S^2	La variance résiduelle estimer.
$\chi^2_{(v)}$	Loi du Khi-deux à v ddl.
$T_{(v)}$	Loi de Student à v ddl.
$F(v_1, v_2)$	Loi de Fisher à (v_1, v_2) ddl.
$\Gamma(.,.)$	La matrice de variance-covariance.
α	$\alpha \in]0, 1[$ un niveau du signification.
$t_{(v)}(1 - \frac{\alpha}{2})$	Quantile d'ordre $(1 - \frac{\alpha}{2})$ de la loi de Student.
$f_{(v_1,v_2)}(1 - \alpha)$	Quantile d'ordre $(1 - \alpha)$ de la loi de Fisher.
$c_{(v)}(\frac{\alpha}{2})$	Quantile d'ordre $(\frac{\alpha}{2})$ de la loi de Khi-deux.
H_0, H_1	L'hypothèse nulle et l'alternative (resp).
$R_{(H)}$	Région critique (Région de rejet).
R^2	Le coefficient de détermination.
F	La statistique de Fisher.
RLS	Régression linéaire simple.
iid	indépendantes identiquements distribuées.
MV	Maximum de vraisemblance.
ddl	degrés de liberté.
SCT	Somme des carrés totales.
SCR	Somme des carrés résiduelles.
SCE	Somme des carrés expliques.
ANOVA	Analyse de la variance.
RC	Région critique.
IC	Intervalle du confiance.
i.e	C'est-à-dire.

الملخص

يعد تحليل الأخطاء أحد أكثر الطرق موثوقية لتقييم جودة الإنحدار. نقدم في هذه المذكرة لمحة عامة عن الإنحدار الخطي البسيط باستخدام تحليل الأخطاء، و الذي يمكن إجراؤه بالإختبارات الإحصائية أو بأدوات رسومية بسيطة. باستخدام البرنامج الإحصائي R، نقدم أيضا بعض الأمثلة و التطبيقات.

Résumé

L'analyse des résidus est l'un des moyens les plus sûrs d'évaluer la qualité d'une régression. Dans ce mémoire nous donnons un aperçu général sur la régression linéaire simple à l'aide de l'analyse des résidus, qui peut être réalisé avec les tests statistiques ou bien par des outils graphiques simples. A l'aide du logiciel statistique R, nous donnons aussi quelques exemples et des applications.

Abstract

Residual analysis is one of the most reliable ways to assess the quality of a regression. In this memory, we give a general overview of simple linear regression using residual analysis, which can be performed with statistical tests or by simple graphical tools. Using the statistical software R, we also give some examples and applications.