REPUBLIQUE DU BENIN

\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**MINISTERE DE L’ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE**







***ECOLE NATIONAL DE STATISTIQUE DE LA PLANIFIQUATION ET DE LA DEMOGRAPHIQUE (ENSPD)***

**Membres du groupe : 12**

1. **TAWAYE DAN DOSSA ISSOUFOU**
2. **GOUNOU SERO SEKE**
3. **BIO SIKA ABDIAS**
4. **MOUTOUMA YATTE SIMON**

**SOUS LA SUPPERVISION DE:**

**Dr Ir SODJINOU**

**COURS:** Echantillonnage et Réechantillonnage

**MASTER 1**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**THEME: Conditions d’application des regressions lineaires**

**TRAVAUX GROUPE:27**

ANNEE ACADEMIQUE: 2024 -2025

# Sommaire

1. Introduction

2. Les conditions théoriques d'application de la régression linéaire

- 2.1 Caractère aléatoire et indépendance des observations

- 2.2 Indépendance des résidus

- 2.3 Normalité des résidus

- 2.4 Homoscédasticité

- 2.5 Multicolinéarité (en régression multiple)

3. Outils de diagnostic dans R

- 3.1 Ajustement du modèle

- 3.2 Vérification graphique

- 3.3 Tests statistiques

4. Exemple pratique

5. Conclusion

6. Références

7. Exercice d’application

# 1. Introduction

La régression linéaire est une méthode statistique puissante pour modéliser la relation entre une variable dépendante Y et une ou plusieurs variables explicatives X. Toutefois, pour que les résultats soient fiables et interprétables, certaines hypothèses doivent être vérifiées. Le logiciel R propose des outils simples pour diagnostiquer ces conditions.

# 2. Les conditions théoriques d’application

## 2.1 Caractère aléatoire et indépendance des observations

Les données doivent provenir d’un échantillon aléatoire simple. Chaque observation doit être indépendante, c’est-à-dire que la valeur d’un individu ne dépend pas d’un autre.

## 2.2 Indépendance des résidus

Les erreurs doivent être non corrélées entre elles. C’est essentiel pour éviter les problèmes d’autocorrélation, notamment dans les séries chronologiques ou données spatiales.

## 2.3 Normalité des résidus

Les résidus doivent suivre une loi normale. Cette hypothèse est importante pour la validité des tests statistiques (comme les p-values et les intervalles de confiance).

## 2.4 Homoscédasticité

La variance des résidus doit rester constante pour toutes les valeurs de X. En cas de variances non constantes (hétéroscédasticité), les estimations peuvent être biaisées.

## 2.5 Multicolinéarité (en cas de régression multiple)

En présence de plusieurs variables explicatives, il faut éviter une forte corrélation entre elles, ce qui fausse l’interprétation des coefficients.

# 3. Outils de diagnostic dans R

## 3.1 Ajustement du modèle

R  
modele <- lm(Y ~ X, data = donnees)

## 3.2 Vérification graphique

### a. Graphiques de diagnostic de base

R  
par(mfrow = c(2, 2))  
plot(modele)

### b. Histogramme et QQ-plot des résidus

R  
residus <- residuals(modele)  
hist(residus)  
qqnorm(residus); qqline(residus)

## 3.3 Tests statistiques

### a. Normalité des résidus

R  
shapiro.test(residus)

### b. Homoscédasticité (test de Breusch-Pagan)

R  
library(lmtest)  
bptest(modele)

### c. Indépendance des résidus (test de Durbin-Watson)

R  
library(car)  
durbinWatsonTest(modele)

### d. Multicolinéarité (Variance Inflation Factor)

R  
library(car)  
vif(modele) # VIF > 5 indique une colinéarité problématique

# 4. Exemple pratique

R  
modele <- lm (mpg ~ wt + hp, data = mtcars)  
summary(modele)  
plot(modele)

# 5. Conclusion

La validité d’un modèle de régression linéaire repose sur le respect de certaines hypothèses statistiques. Grâce aux outils intégrés dans R, il est possible de diagnostiquer rapidement et efficacement les éventuelles violations de ces hypothèses. Cela permet d’ajuster le modèle, ou de choisir une méthode alternative si nécessaire (régression robuste, transformation, etc.).

# 6. Références

James et al. (2013). An Introduction to Statistical Learning.

R documentation : lm, bptest, shapiro.test, vif

# 7. Exercice d’application

Vérification des conditions d’application de la régression linéaire à l’ENSP Parakou

Contexte :  
Une étude est menée à l’ENSP de Parakou pour analyser l'effet de certaines variables académiques sur la moyenne générale des étudiants en fin de semestre. Les variables suivantes sont collectées sur un échantillon de 60 étudiants :  
- moyenne : Moyenne générale (variable dépendante Y)  
- heures\_revision : Nombre moyen d'heures de révision par jour  
- taux\_absence : Taux d’absence aux cours  
- age : Âge de l’étudiant  
  
On souhaite construire un modèle de régression linéaire multiple et vérifier la validité des hypothèses classiques.

**1. Ajustement du modèle**

modele <- lm(moyenne ~ heures\_revision + taux\_absence + age, data = enspp)  
summary(modele)

**2. Vérification des hypothèses**

**a. Indépendance des résidus**

library(car)  
durbinWatsonTest(modele)  
Résultat attendu : Si la statistique est proche de 2 (et p-value > 0.05), on conclut à l’indépendance des résidus.

**b. Normalité des résidus**

residus <- residuals(modele)  
shapiro.test(residus)  
Résultat attendu : Si p-value > 0.05, les résidus suivent une loi normale.

**c. Homoscédasticité**

library(lmtest)  
bptest(modele)  
Résultat attendu : Si p-value > 0.05, l’hypothèse d’homoscédasticité est respectée.

**d. Multicolinéarité**

library(car)  
vif(modele)  
Résultat attendu : Si tous les VIF < 5, il n’y a pas de multicolinéarité sérieuse.

**3. Vérification graphique**

par(mfrow = c(2,2))  
plot(modele)  
Interprétations attendues :  
- Residuals vs Fitted : pas de motif régulier → OK  
- Normal Q-Q : les points suivent la droite → OK  
- Scale-Location : dispersion homogène → OK  
- Residuals vs Leverage : pas d’influence excessive → OK

**4. Conclusion**

Si tous les tests statistiques et diagnostics graphiques sont conformes (p-values > 0.05, graphiques sans anomalies majeures), alors le modèle est bien spécifié et peut être utilisé pour des prévisions ou des analyses.  
Sinon, il faut envisager des corrections : transformation des variables, suppression des valeurs influentes ou utilisation d’un autre modèle (régression robuste).

**5. Jeu de données**

|  |  |  |  |
| --- | --- | --- | --- |
| moyenne | heures\_revision | taux\_absence | age |
| 12.5 | 2.0 | 5.0 | 21 |
| 13.2 | 2.5 | 3.0 | 22 |
| 14.1 | 3.0 | 2.0 | 21 |
| 11.5 | 1.0 | 6.0 | 23 |
| 15.0 | 3.5 | 1.0 | 22 |
| 13.8 | 2.8 | 3.5 | 20 |
| 12.0 | 1.5 | 4.0 | 24 |
| 14.5 | 3.2 | 2.2 | 21 |
| 13.0 | 2.3 | 3.1 | 23 |
| 11.8 | 1.2 | 5.8 | 22 |