# 

REPUBLIQUE DU BENIN

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







ANNEE ACADEMIQUE: 2024 -2025

**SUPPERVISEUR:**

Dr : SODJINOU

**Membres du groupe: 11**

1. **FANOUDH KORNELIA SIMONE AMAL**
2. **ATTIN AYEDJO ANGELO MERESSE**
3. **GAWA MICHEE**

**COURS:** introduction aux logiciels statistiques

**MASTER 1**

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

**THEME: REGRESSION LINEAIRE STATA**

**TRAVAUX GROUPE:11**

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

**Table des matières**

[Introduction 2](#_Toc199774248)

[1. Fondements théoriques de la régression linéaire 2](#_Toc199774249)

[1.2 Régression linéaire multiple 3](#_Toc199774250)

[1.3 Hypothèses classiques : 3](#_Toc199774251)

[1.4 Indicateurs importants : 4](#_Toc199774252)

[2. Mise en œuvre dans Stata 5](#_Toc199774253)

[3. Analyse et interpretation 6](#_Toc199774254)

[3.1 Lecture des résultats 6](#_Toc199774255)

[3.2 Validation du modèle 6](#_Toc199774256)

[3.3 Exemple : 6](#_Toc199774257)

[Conclusion 6](#_Toc199774258)

[6. Exercice d’illustration: Régression linéaire à l’Université de ParakoU 7](#_Toc199774259)

[Objectif 7](#_Toc199774260)

[Données simulées : etudiants\_parakou.dta 7](#_Toc199774261)

[Travail demande 7](#_Toc199774262)

# Introduction

Dans le domaine de l’analyse de données, la régression linéaire est l’un des outils statistiques les plus couramment utilisés pour étudier les relations entre une variable dépendante et une ou plusieurs variables explicatives. Elle permet non seulement de modéliser ces relations, mais aussi de faire des prédictions et de tester des hypothèses. Le logiciel STATA, largement utilisé en recherche et en évaluation de projet, offre une interface efficace pour réaliser ce type d’analyse. Grâce à ses commandes simples et précises, il facilite l’estimation des modèles et la lecture des résultats. Avant d’interpréter ces derniers, il est toutefois important de comprendre la logique de la régression linéaire, les étapes de sa mise en œuvre dans STATA, ainsi que les principales hypothèses sur lesquelles repose cette méthode.

# 1. Fondements théoriques de la régression linéaire

**1.1 Régression linéaire simple:** La régression linéaire simple permet d’étudier la relation entre une variable dépendante (ou à expliquer) et une seule variable indépendante (ou explicative). Elle vise à modéliser cette relation à l’aide d’une droite d’équation :

**Y = a + bX + ε,**

où Y est la variable dépendante, X la variable indépendante, a l’ordonnée à l’origine, b le coefficient de régression (pente), et ε l’erreur aléatoire.

Formule : Y = β₀ + β₁X + ε  
Y : variable dépendante  
X : variable explicative  
β₀ : constante (ordonnée à l’origine)  
β₁ : pente (variation de Y pour une unité de X)  
ε : erreur aléatoire

## 1.2 Régression linéaire multiple

La **régression linéaire multiple**, quant à elle, étend ce principe à plusieurs variables explicatives. L’objectif est d’expliquer Y à partir de plusieurs X₁, X₂, ..., Xₙ afin d’améliorer la précision des prédictions et de mieux comprendre les effets combinés de plusieurs facteurs. Le modèle prend la forme :  
**Y = a + b₁X₁ + b₂X₂ + ... + bₙXₙ + ε**.  
Formule : Y = β₀ + β₁X₁ + β₂X₂ + … + βnXn + ε  
Utilisée quand plusieurs variables influencent Y.

## 1.3 Hypothèses classiques :

Pour que les résultats d’une régression linéaire soient fiables et interprétables, certaines **hypothèses statistiques** doivent être respectées. Elles permettent de garantir la validité des estimateurs et la pertinence des tests réalisés. Les principales hypothèses sont les suivantes :

* **Linéarité** :  
  La relation entre la variable dépendante et les variables explicatives doit être linéaire. Cela signifie que l’effet de chaque variable explicative sur la variable dépendante est constant.
* **Indépendance des erreurs** :  
  Les résidus (ou erreurs aléatoires) doivent être statistiquement indépendants les uns des autres. Cette hypothèse est essentielle, surtout avec des données temporelles (séries chronologiques).
* **Homoscedasticité** :  
  La variance des erreurs doit être constante pour toutes les valeurs des variables explicatives. En cas de variances inégales (hétéroscédasticité), les tests statistiques peuvent devenir invalides.
* **Normalité des erreurs** :  
  Les erreurs doivent suivre une distribution normale, surtout si l’on veut effectuer des tests d’hypothèses et construire des intervalles de confiance précis.
* **Absence de multicolinéarité (dans la régression multiple)** :  
  Les variables explicatives ne doivent pas être fortement corrélées entre elles. Une forte corrélation rend difficile l’estimation des effets propres de chaque variable.
* **Pas d’omission de variable pertinente et pas d’inclusion de variable inutile** :  
  Le modèle doit être bien spécifié. L’omission d’une variable importante ou l’inclusion d’une variable sans lien réel avec *Y* peut fausser les résultats.

## 1.4 Indicateurs importants :

Une fois la régression estimée, plusieurs **indicateurs statistiques** permettent d’évaluer la qualité du modèle et la signification des résultats obtenus. Les plus courants sont :

* **R² (coefficient de détermination)** :  
  Il mesure la proportion de la variance de la variable dépendante expliquée par le modèle. Il varie entre 0 et 1 ; plus il est proche de 1, meilleure est l'ajustement du modèle.

Exemple : un R² de 0,85 signifie que 85 % des variations de Y sont expliquées par les variables X.

* **R² ajusté** :  
  Il tient compte du nombre de variables explicatives dans le modèle. Contrairement au R² simple, il peut diminuer si l’ajout d’une variable n’améliore pas réellement la qualité du modèle.
* **Statistique F (test global du modèle)** :  
  Elle permet de tester si l’ensemble des variables explicatives a un effet significatif sur la variable dépendante. Un p-value < 0,05 indique que le modèle est globalement significatif.
* **Coefficients de régression (b)** :  
  Ils indiquent l’effet d’une unité de variation de chaque variable explicative sur la variable dépendante. Dans STATA, chaque coefficient est accompagné d’un écart-type, d’une statistique *t* et d’une *p-value*.
* **P-value** :  
  Pour chaque variable, elle permet de tester l’hypothèse selon laquelle le coefficient est nul (absence d’effet). Si *p < 0,05*, on considère généralement l’effet comme significatif.
* **Erreur standard des résidus (Root MSE)** :  
  Elle donne une idée de la précision des prédictions. Plus cette erreur est faible, plus le modèle est performant.

# 2. Mise en œuvre dans Stata

* **Importer les données**
  + use fichier.dta, clear (pour un fichier STATA)
  + import excel "fichier.xlsx", firstrow clear (pour Excel)
* **Vérifier les données**
  + describe (description des variables)
  + list in 1/10 (afficher les 10 premières observations)
* **Lancer une régression linéaire simple**
  + Exemple : reg revenu niveau\_étude
* **Lancer une régression linéaire multiple**
  + Exemple : reg revenu niveau\_étude experience age
* **Examiner les résultats**
  + Coefficients, p-values, R², R² ajusté, test F
* **Vérifier les hypothèses**
  + Linéarité : twoway (scatter revenu niveau\_étude) (lfit revenu niveau\_étude)
  + Normalité des résidus :
    - predict residu, residual
    - histogram residu, normal ou swilk residu
  + Homoscédasticité : estat hettest
  + Indépendance des erreurs : estat durbinalt
  + Multicolinéarité : estat vif

# 3. Analyse et interpretation

## 3.1 Lecture des résultats

Les coefficients indiquent l’effet des variables sur la variable expliquée. La p-value montre si cet effet est statistiquement significatif (p < 0,05). Le R² mesure la qualité du modèle, c’est-à-dire la part de variance expliquée. La statistique F teste si le modèle est globalement pertinent.

## 3.2 Validation du modèle

La validation du modèle consiste à vérifier que les hypothèses de la régression sont respectées : linéarité, indépendance, homoscédasticité, normalité des résidus, et absence de multicolinéarité. Ces contrôles garantissent la fiabilité des résultats.

On utilise plusieurs tests dans STATA, par exemple :

* test de Breusch-Pagan pour l’homoscédasticité,
* test de Durbin-Watson pour l’indépendance des erreurs,
* test de Shapiro-Wilk pour la normalité des résidus,
* VIF pour la multicolinéarité

## 3.3 Exemple :

Si β₁ = 2.5 pour weight, chaque kg augmente le prix de 2.5 unités, toutes choses égales par ailleurs.

# Conclusion

La régression linéaire demeure un outil incontournable pour analyser les relations entre variables continues, et son implémentation dans le logiciel STATA permet une application rigoureuse, rapide et reproductible. Toutefois, la fiabilité des résultats repose sur le respect de certaines conditions fondamentales, telles que la normalité des résidus, l’homoscédasticité, l’indépendance des erreurs ou encore l’absence de multicolinéarité. La vérification systématique de ces hypothèses est donc essentielle pour éviter les interprétations erronées.

Au-delà de la régression linéaire classique, il est important de rester attentif à la nature des données et aux objectifs de l’analyse. D'autres approches, comme la régression logistique pour les variables dépendantes binaires ou les modèles de panel pour les données longitudinales, offrent des solutions adaptées à des contextes plus complexes. Ainsi, maîtriser les outils de régression dans STATA, tout en développant un regard critique sur leurs conditions d’utilisation, constitue un atout majeur pour toute analyse statistique rigoureuse et pertinente.

# 6. Exercice d’illustration: Régression linéaire à l’Université de ParakoU

# Objectif

Utiliser une régression linéaire pour analyser les facteurs influençant la moyenne générale des étudiants à l’Université de Parakou, à partir de données simulées réalistes.

# Données simulées : etudiants\_parakou.dta

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | sexe | age | heures\_revision | distance\_domicile | moyenne |
| 1 | 1 | 21 | 12 | 1.2 | 13.5 |
| 2 | 0 | 19 | 20 | 0.5 | 15.2 |
| 3 | 1 | 23 | 8 | 2.5 | 11.8 |
| 4 | 0 | 20 | 16 | 1.0 | 14.0 |

Variables :  
- moyenne : moyenne générale (dépendante)  
- heures\_revision : heures de révision hebdomadaires  
- age : âge de l’étudiant  
- sexe : 1 = homme, 0 = femme  
- distance\_domicile : distance domicile-campus (km)

# Travail demande

**1. Importer le fichier dans Stata**

use etudiants\_parakou.dta, clear

**2. Résumé des données**

summarize

**3. Régression linéaire simple**

reg moyenne heures\_revision

**4. Régression linéaire multiple**

reg moyenne heures\_revision age sexe distance\_domicile

**5. Tester les hypothèses**

vif  
hettest  
sktest resid

**6. Prédictions**

predict yhat, xb  
predict resid, residuals

**7. Visualisation des résultats**

twoway (scatter moyenne heures\_revision) (lfit moyenne heures\_revision)  
rvfplot.