ÉCOLE D'ÉTÉ INTERNATIONALE

MÉTHODES ET OUTILS DES SCIENCES DES TERRITOIRES

UNE PERSPECTIVE NORD-SUD, SUD-NORD ET SUD-SUD

ÉTAPE 2 • IRSP, Ouidah (Bénin) 27 février - 10 mars 2023

MODELISATION D'UNE VARIABLE QUANTITATIVE

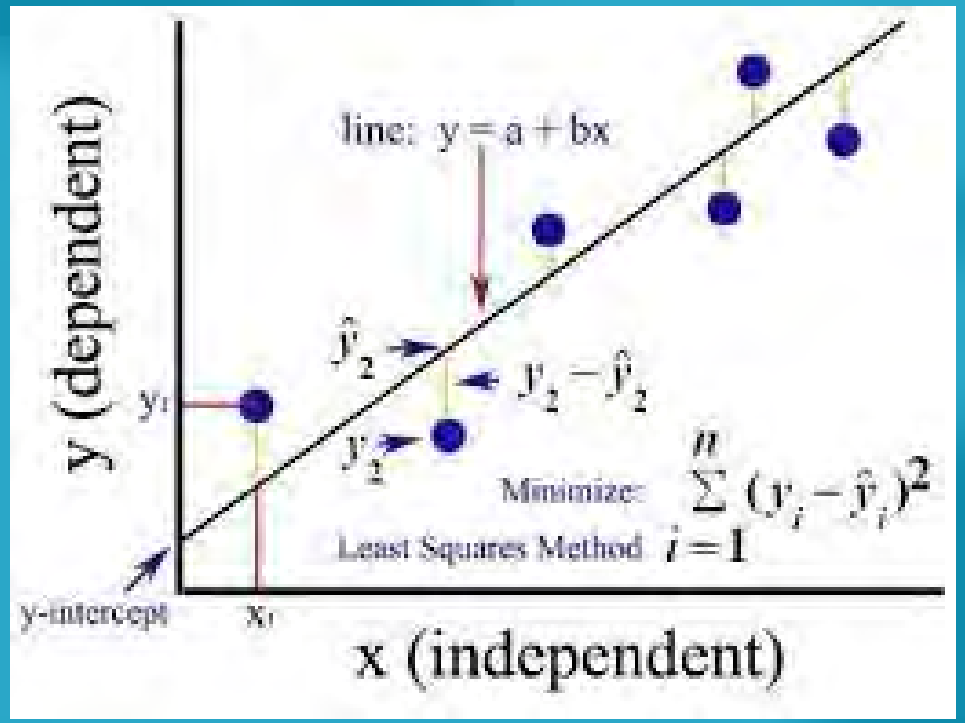
Régression linéaire,
Régression multiple &
Analyse de variance

Nadège Gbétoton Djossou

Claude Grasland

Contributeur.ice.s :

FONDEMENTS THEORIQUES DE LA REGRESSION LINEAIRE (simple ou multiple)



Analyse de la régression linéaire

- Elle s'intéresse à l'étude de la dépendance d'une variable quantitative continue(variable dépendante) par rapport à une ou plusieurs variables quantitatives ou qualitatives (variables indépendantes ou explicatives).
- Bien que très liées, l'analyse de la corrélation est conceptuellement très différente de l'analyse de la régression. L'analyse de la corrélation mesure l'intensité de la liaison entre deux variables.

  Exemple: corrélation entre le PIB et la mortalité infantile dans les pays africains

- L'analyse de la régression linéaire a pour objectif d'estimer et/ou de prédire la valeur moyenne de la variable dépendante en fonction des valeurs connues ou fixes des variables indépendantes.

  Exemple: Prédire la mortalité infantile dans les pays africains en disposant de leurs PIB

Régression linéaire simple

- On parle de modèle de régression linéaire simple lorsqu'on cherche à prédire une variable quantitative à partir d'un seul regresseur (variable indépendante quantitative)

Obersation, i	Response, Y	Variables indépendantes, X
1	y_1	x_1
2	y_2	x_2
⋮	⋮	⋮
n	y_n	x_n

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

Régression linéaire multiple

- On parle de modèle de régression linéaire multiple lorsqu'on cherche à prédire une variable quantitative à partir simultanément de plusieurs regresseurs (variables indépendantes quantitatives et qualitatives)

		Variables indépendantes			
Obersation, i	y_i	x_1	x_2	…	x_k
1	y_1	x_{11}	x_{12}	…	x_{1k}
2	y_2	x_{21}	x_{22}	…	x_{2k}
⋮	⋮	⋮	⋮		⋮
n	y_n	x_{n1}	x_{n2}	…	x_{nk}

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i$$

Signification du terme aléatoire

- Le terme aléatoire que l'on appelle ***erreur du modèle***, tient un rôle très important dans la régression.
- Il permet de résumer toute l'information qui n'est pas prise en compte dans la relation linéaire que l'on cherche à établir entre Y et X.
- Il peut s'agir:
  - ***omission de variables ou imprécision de la théorie***
  - ***erreur de mesure***
  - ***la nature intrinsèquement aléatoire du comportement humain***
  - ***erreur de fluctuation d'échantillonnage***

Fonction de régression de l'échantillon (FRE) Vs fonction de régression de la population (FRP)

- Dans la plupart des situations concrètes nous ne disposons que d'un échantillon de Y associé à quelques valeurs données de X.

- Ainsi, notre tâche est d'estimer la fonction de régression de la population (FRP) à partir des informations fournies par l'échantillon (FRE).

- En résumé, notre objectif est donc d'estimer, la FRP :

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i$$

- à partir de la fonction de régression de l'échantillon (FRE) :

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + e_i$$

Fonction de régression de l'échantillon (FRE) Vs fonction de régression de la population (FRP)

- Dans la plupart des situations concrètes nous ne disposons que d'un échantillon de Y associé à quelques valeurs données de X.

- Ainsi, notre tâche est d'estimer la fonction de régression de la population (FRP) à partir des informations fournies par l'échantillon (FRE).

- En résumé, notre objectif est donc d'estimer, la FRP :

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i$$

- à partir de la fonction de régression de l'échantillon (FRE) :

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + e_i$$

- La méthode des MCO (**Moindres Carrés Ordinaires**) est attribuée à Carl Friedrich Gauss, un mathématicien allemand.

- La méthode consiste en une prescription qui est que la fonction $f(x;\beta)$ qui décrit « le mieux » les données est celle qui minimise la somme quadratique des déviations des mesures aux prédictions de $f(x;\beta)$.

- **Hypothèse 1 : Le modèle est linéaire dans les paramètres**

  Cela ne veut pas dire que X et Y sont linéaires (elles peuvent être non linéaires), mais plutôt que les β_j sont linéaires.

- **Hypothèse 2 : L'espérance mathématique de l'erreur ε_i est nulle**.

$$E(\varepsilon_i | x_i) = 0$$

- **Hypothèse 3** : **L'homoscédasticité ou la constance de la variance de ε_i.**

$$E\left(\varepsilon_i^2 | x_i\right) = \sigma_\varepsilon^2$$

- **Hypothèse 4 : La normalité du terme d'erreur**

$$\boldsymbol{\varepsilon_i \rightsquigarrow \mathcal{N}(0, \sigma_\varepsilon^2)}$$

- **Hypothèse 5 : Absence d'autocorrélation des erreurs**

$$E(\varepsilon_i \varepsilon_j | x_i, | x_j) = 0 \quad avec\ (i \neq j)$$

- **Hypothèse 5 : Covariance nulle entre x_i et ε_i**

$$E(x_i \varepsilon_i) = 0$$

- **Hypothèse 7 : Exactitude de la variable indépendante**

  Les valeurs x_i sont fixées d'un échantillon à un autre (observées sans erreur). Ils sont supposés non stochastique (non aléatoire).

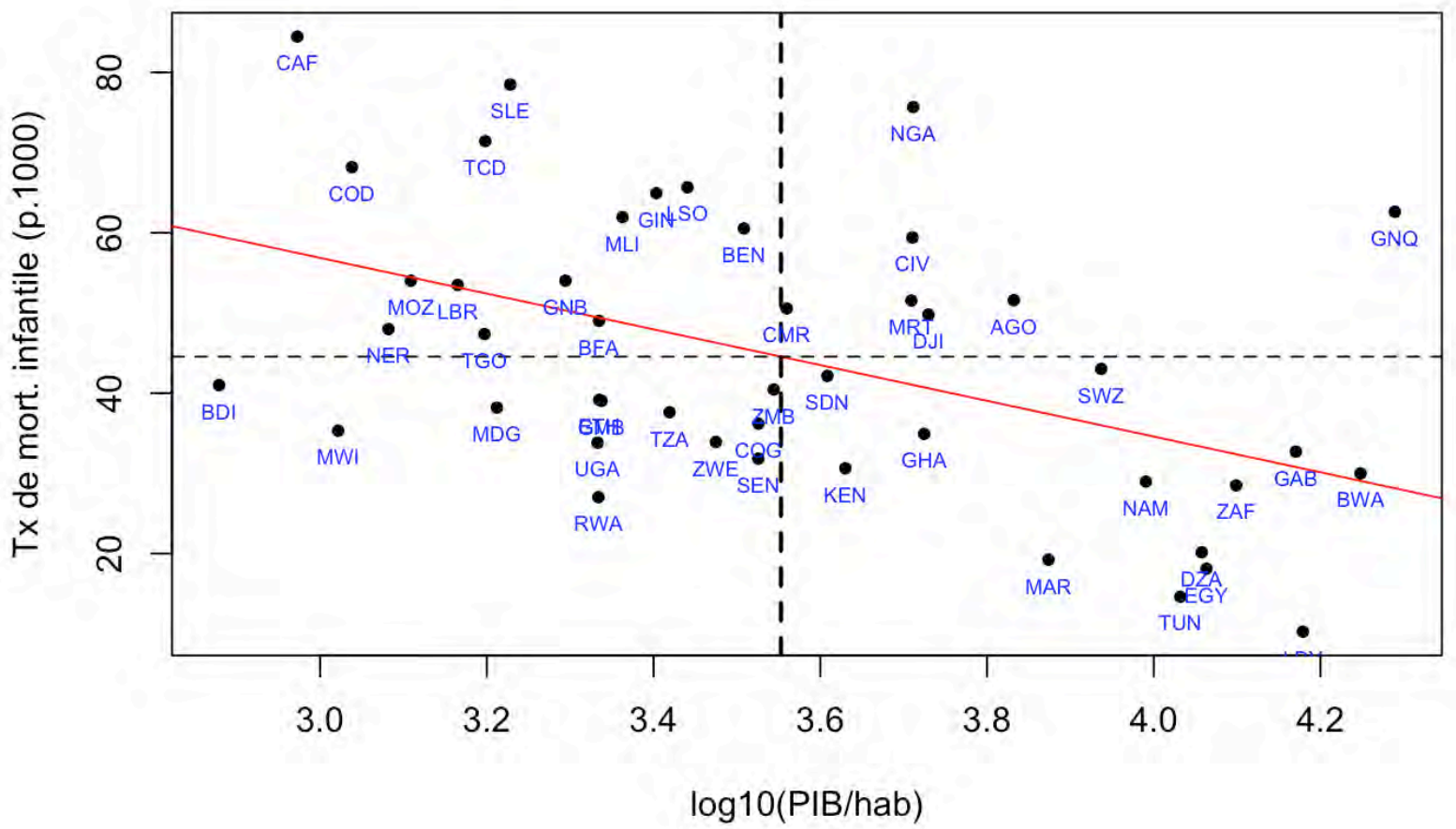
  Cependant, la variable dépendante est supposée statistique, aléatoire ou stochastique, c'est-à-dire ayant une distribution de probabilité.

- **Hypothèse 6 : $n > k$**

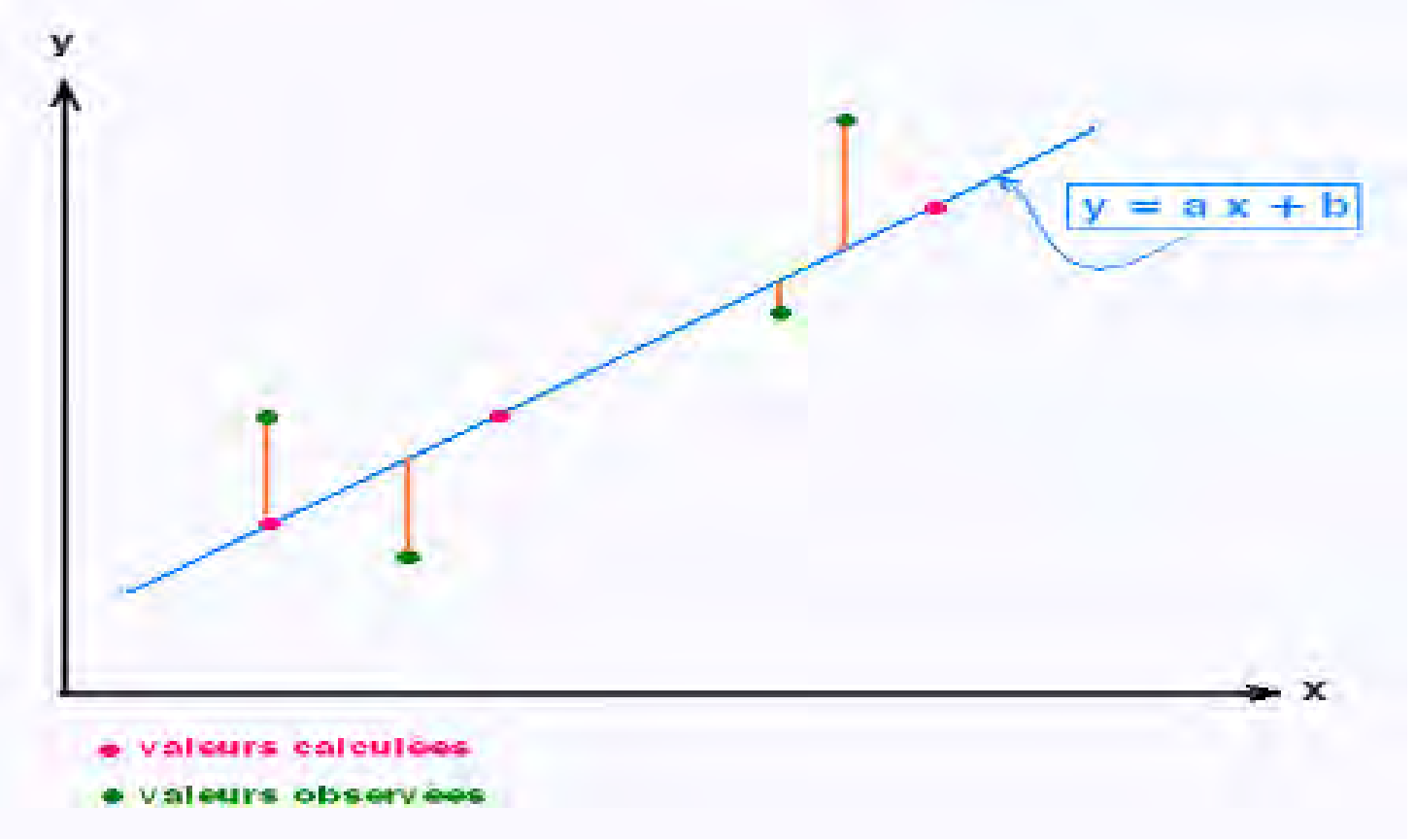
  Le nombre d'observations n doit être plus élevé que le nombre de paramètres k à estimer

- Soit la fonction la fonction de régression de la population
- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- β_0 représente l'intercept (ordonnée à l'origine); β_1 représente la pente. Les deux sont des paramètres de la population que l'on cherche à estimer à partir des données de l'échantillon.
- $\beta_0 + \beta_1 x_i$ est la partie déterministe du modèle.
- Ainsi, si l'on estime β_0 et β_1, on pourra prédire la valeur de y_i
- ε_i est le terme d'erreur qui regroupe les imperfections du modèle. C'est la partie aléatoire ou stochastique du modèle

Estimation du modèle de régression linéaire simple par les MCO



Estimation du modèle de régression linéaire simple par les MCO



- Soit $\widehat{y_i}$ la valeur prédite de y_i
- $\widehat{y_i}$ correspond aux y_i qui sont exactement sur la ligne de régression
- Cependant, pour toutes les observations il est possible de prédire l'erreur qui est sous la forme: $y_i - \widehat{y_i}$
- La méthode consiste en une prescription selon laquelle la fonction $f(x;\hat{\beta})$ qui décrit « le mieux » les données est celle qui minimise la somme quadratique des déviations des mesures aux prédictions de $f(x;\hat{\beta})$

$$Min \sum_{i=1}^{n} e_i^2 = Min \sum_{i} (y_i - \widehat{y_i})^2 = Min \left(y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)\right)^2$$

$$Min \sum_{i=1}^{n} e_i^2 = Min \sum_{i} (y_i - \hat{y_i})^2 = Min \left(y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i) \right)^2$$

- La résolution de cette équation va nous donner les paramètres estimés suivants:

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

$$\widehat{\boldsymbol{\beta_1}} = \frac{\sum_{i=1}^{n} (\boldsymbol{x_i} - \overline{\boldsymbol{x}}) (\boldsymbol{y_i} - \overline{\boldsymbol{y}})}{\sum_{i=1}^{n} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2}$$

- Le modèle de régression linéaire correspondant, se présente alors comme suit :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

- On pourra présenter le modèle sous forme matricielle

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

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{et} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- L'estimation par les MCO

$$\hat{\beta} = (X'X)^{-1}X'Y$$

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

- En se basant sur les hypothèses des MCO

$$\widehat{\beta_j} \rightsquigarrow \mathcal{N}\left(\beta_j, Var(\widehat{\beta_j})\right)$$

$$\boldsymbol{t_{\widehat{\beta_j}}} = \frac{\widehat{\beta_j} - \beta_j}{\sqrt{Var(\widehat{\beta_j})}} = \frac{\widehat{\beta_j} - \beta_j}{\sigma(\widehat{\beta_j})} \rightsquigarrow \boldsymbol{t_{(n-p)}}$$

- Soit α le seuil de significativité ou risque d'erreur. La règle de décision est comme suit :
  - Si $t_{cal} > t_{lue}(n-p)$ alors on rejette H_0
  - Si $t_{cal} < t_{lue}(n-p)$, on ne rejette pas H_0

- Sous les hypothèses du modèle de régression linéaire, nous pouvons également construire un intervalle de confiance pour un paramètre β_j de la population.
- Cet intervalle de confiance fournit l'ensemble des valeurs possibles du paramètre de la population et pas simplement une estimation ponctuelle de cette valeur.
- Ainsi, sous l'hypothèse alternative $\boldsymbol{H_1}: \boldsymbol{\beta_j \neq 0}$, l'intervalle de confiance (IC) au niveau $(\mathbf{1-\alpha})$ c'est-à-dire avec 95% de chance de contenir le paramètre inconnu $\boldsymbol{\beta_j}$ est donné par :

$$\boldsymbol{IC = [\hat{\beta}_j - t_{(\alpha/2;n-2)} \times \hat{\sigma}_{\hat{\beta}_j} \; ; \; \hat{\beta}_j - t_{(\alpha/2;n-2)} \times \hat{\sigma}_{\hat{\beta}_j}]}$$

- Avec $\hat{\sigma}_{\hat{\beta}_i}$, l'estimateur de l'écart type de $\hat{\beta}_j$.

- L'intercept $\widehat{\boldsymbol{\beta_0}}$ représente la valeur moyenne prédite $\widehat{y_i}$ lorque la variable indépendante est nulle.

- La pente $\widehat{\beta_j}$ s'interprètent comme des **effets marginaux.**

- En d'autres termes, lorsque la variable indépendante augmente d'une unité, on espère une variation moyenne de $\widehat{\beta_j}$ pour la variable dépendante

- Il est important de noter qu'il est incorrect de dire qu'une « *augmente d'une unité de la variable indépendante, entraine une variation de* $\widehat{\beta_j}$ *pour la variable dépendante* » puisque cette interprétation ne tient pas compte de la non-justesse des données mais considère uniquement la régression linéaire parfaite.

- En utilisant les expressions « espère » et « moyenne », on tient compte du fait que la prédiction n'est pas parfaite et que la droite de regréssion représente juste une prédiction des nuages de points non alignés (imparfaits)

- L'analyse de la variance encore appelé ANOVA consiste à expliquer la variance totale sur l'ensemble des échantillons:
  - en fonction de la variance due à l'interaction entre les variables du modèle (la variance expliquée par le modèle)
  - et de la variance résiduelle aléatoire (la variance non expliquée par le modèle).
- Elle est fondée sur l'orthogonalité entre le vecteur des résidus estimés et de la variable prédite.

Décomposition de la variance - Équation d'analyse de variance

$$y_i = \hat{y_i} + e_i$$

- On montre que:

$$\sum_{i=1}^{n}(y_i - \bar{y_i})^2 = \sum_{i=1}^{n}(\hat{y_i} - \bar{y_i})^2 + \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

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

- Cette équation est l'équation fondamentale de l'analyse de la variance pour les modèles de régression.
  - SCE indique la variabilité expliquée par le modèle, c'est-à-dire la variation de Y expliquée par X
  - SCR indique la variabilité non-expliquée par le modèle, c'est-à-dire l'écart entre les valeurs observées de Y et celle prédites par le modèle.

Décomposition de la variance - le coefficient de détermination R^2

- Deux situations extrêmes peuvent survenir :
  - Dans le meilleur des cas, $SCR = 0$: les variations de Y sont complétement expliquées par celles de X. On a un modèle parfait. La droite de régression passe exactement par tous les points du nuage ($\widehat{y_i} = y_i$)
  - Dans le pire des cas, $SCE = 0$: X n'apporte aucune information sur Y

- A partir de l'équation de la variance, on déduit le coefficient de détermination R^2 qui permet de mesurer la qualité de l'ajustement

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

- R^2 indique la proportion de variance de y expliquée par le modèle
- Plus R^2 est proche de 1, meilleur sera le modèle.
- La connaissance des valeurs de X permet de deviner avec précision celle de Y

- Bien que R^2 est un coefficient très populaire de la qualité de l'adjustement des modèles de régression, il présente l'inconvénient de toujours croitre avec l'ajout de nouvelles variables indépendantes dans le modèle
- Ceci suppose que ces nouvelles variables apportent une contribution au modèle
- Ce qui n'est pas toujours vrai
- Le ${R^2}_{Adj}$ corrige ce biais du coefficient R^2

$$ {R^2}_{Adj} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right) $$

Décomposition de la variance – le tableau de l'ANOVA

Source de variation	Somme des carrés	Degré de liberté	Carrés moyens
Explicatives	$SCE = \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}}_i)^2$	1	p
Résidus	$SCR = \sum_{i=1}^{n} (e_i)^2$	$n - p - 1$	$SCR/(n - p - 1)$
Total	$SCT = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$	$n - 1$	

- À partir du tableau de l'ANOVA, nous effectuons le test de la linéarité de la régression en calculant la statistique F qui suit une loi de Fisher $F(p, n - p - 1)$.
- Il revient à tester si l'ensemble des variables explicatives X contribue pas à l'explication du modèle.
- Le test d'hypothèse est le suivant :

Décomposition de la variance – le tableau de l'ANOVA

- Le test d'hypothèse est le suivant :

$$H_0 : SCE = 0$$

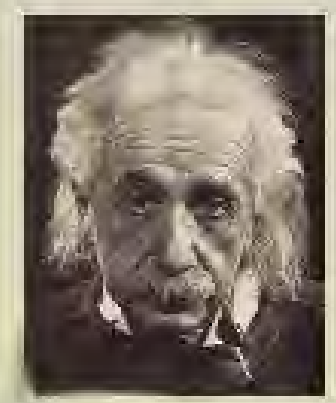
- La statistique de Ficher est donnée par :

$$\boldsymbol{F^* = \frac{SCE/p}{SCR/(n-p-1)}}$$

- La statistique F^* permet de tester la ***significativité globale de la régression ou encore d'effectuer une évaluation globale de la régression***,

DE LA THEORIE A LA PRATIQUE

Quelle est la différence entre la théorie et la pratique ?
(question posée à Albert Einstein au terme
d'une conférence donnée à Washington)



La théorie, c'est quand on sait tout
et que rien ne fonctionne.
La pratique, c'est quand tout fonctionne
et que personne ne sait pourquoi.
Mais ici, nous avons réuni théorie et
pratique : rien ne fonctionne et
personne ne sait pourquoi.

J.Piat – P.Wajsman, *Vous n'aurez pas le dernier mot !*
Albin Michel, Paris 2006, p.59

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

On se propose de modéliser la relation entre :

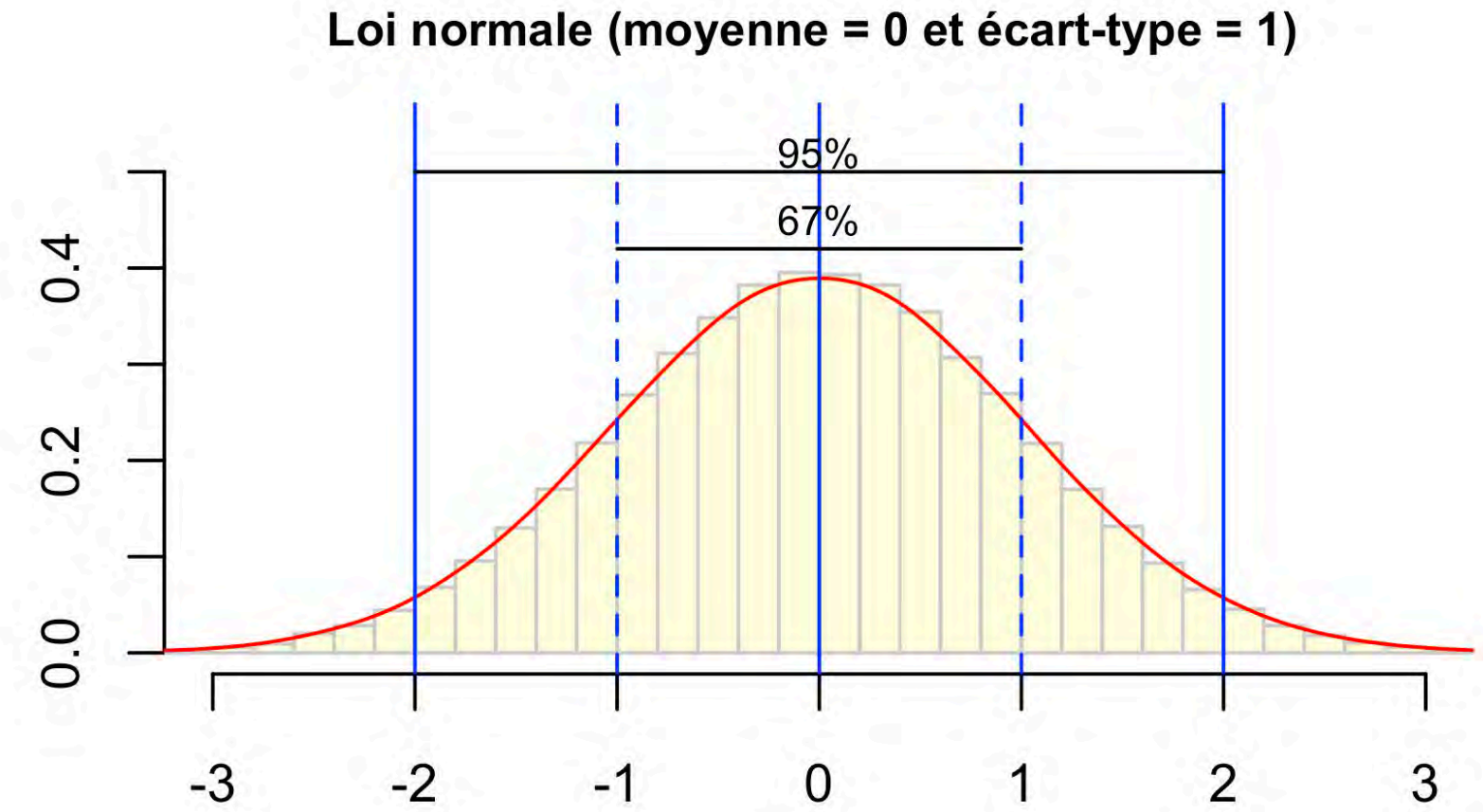
X = PIB par habitant en 2018 (ppa $/hab)

Y = émissions de CO2 en 2018 (tonnes/hab.)

Pour 46 pays africains

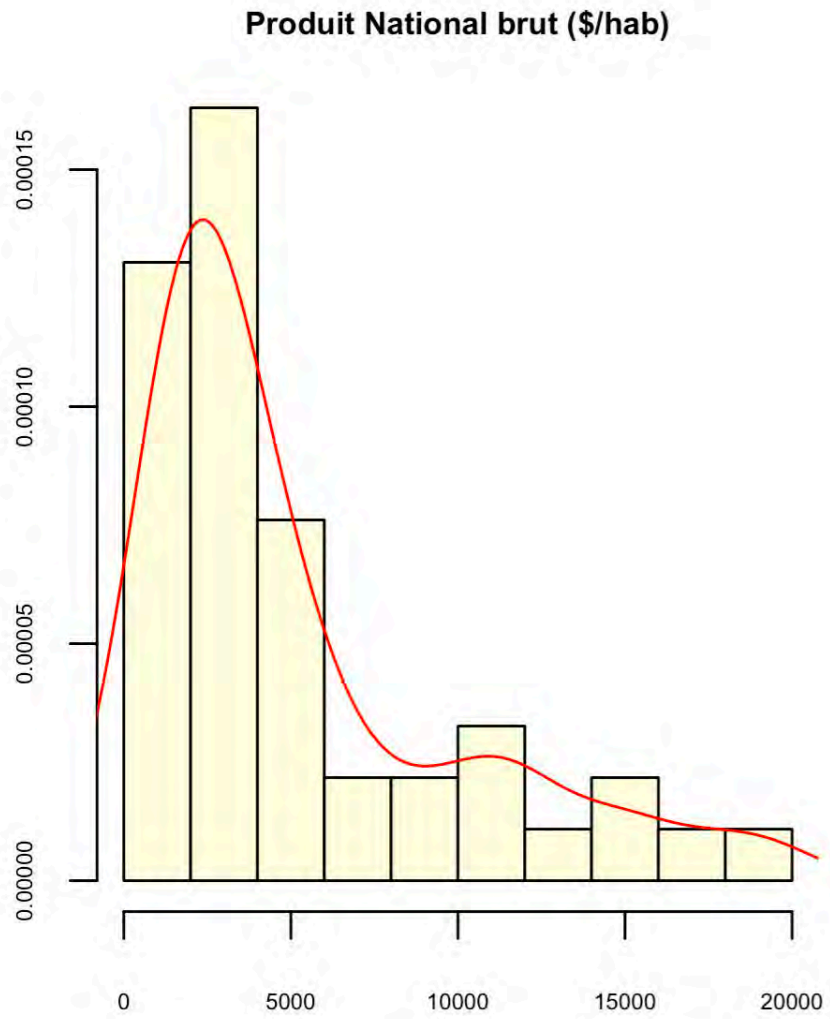
CODE	PAYS	X	Y
AGO	Angola	6794	1.12
BDI	Burundi	757	0.05
BEN	Bénin	3224	0.62
BFA	Burkina Faso	2161	0.20
BWA	Botswana	17700	2.96
CAF	Rep. Centrafricaine	939	0.07
TUN	Tunisie	10760	2.73
TZA	Tanzanie	2625	0.22
UGA	Ouganda	2152	0.14
ZAF	Afrique du Sud	12556	8.09
ZMB	Zambie	3501	0.30
ZWE	Zimbabwe	2983	0.85

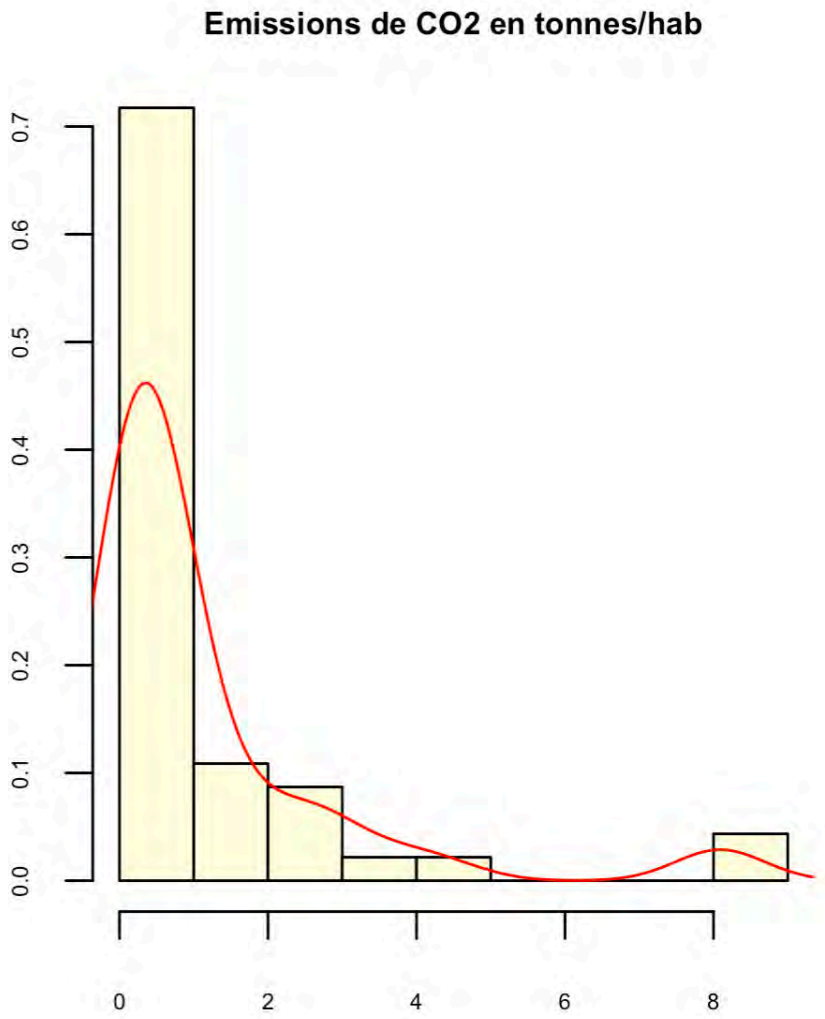
En théorie … X et Y sont des variables gaussiennes



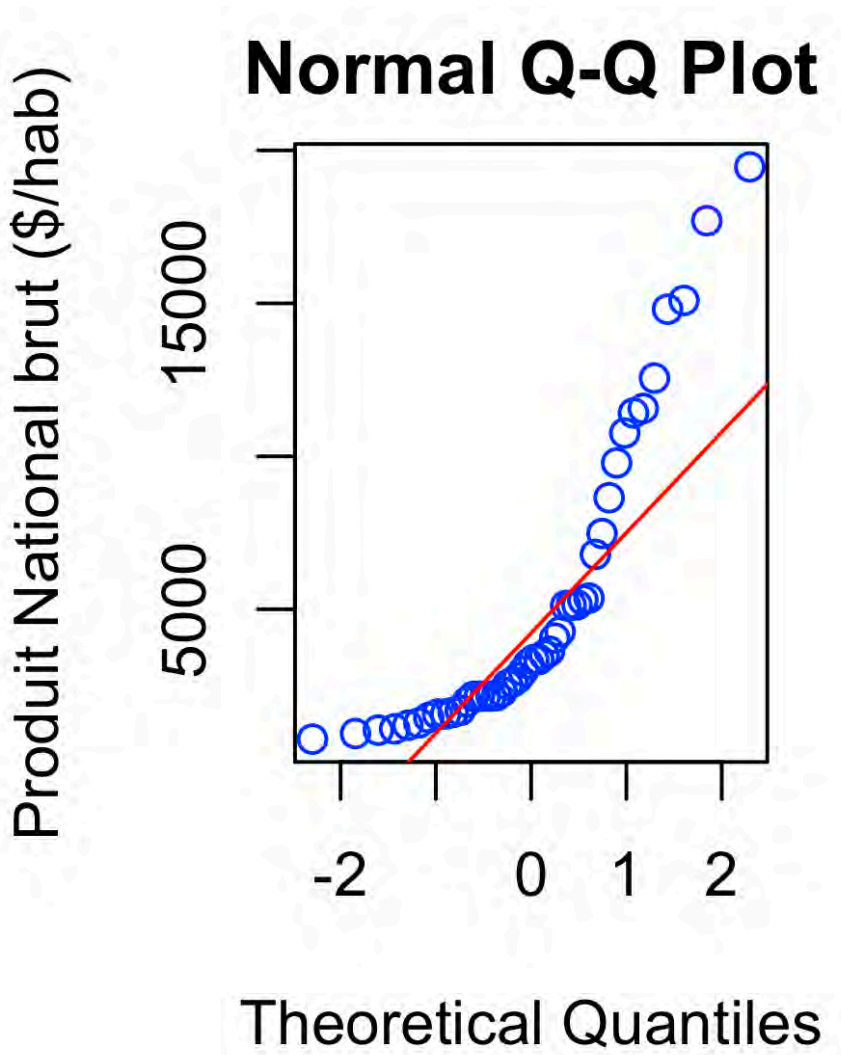
PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

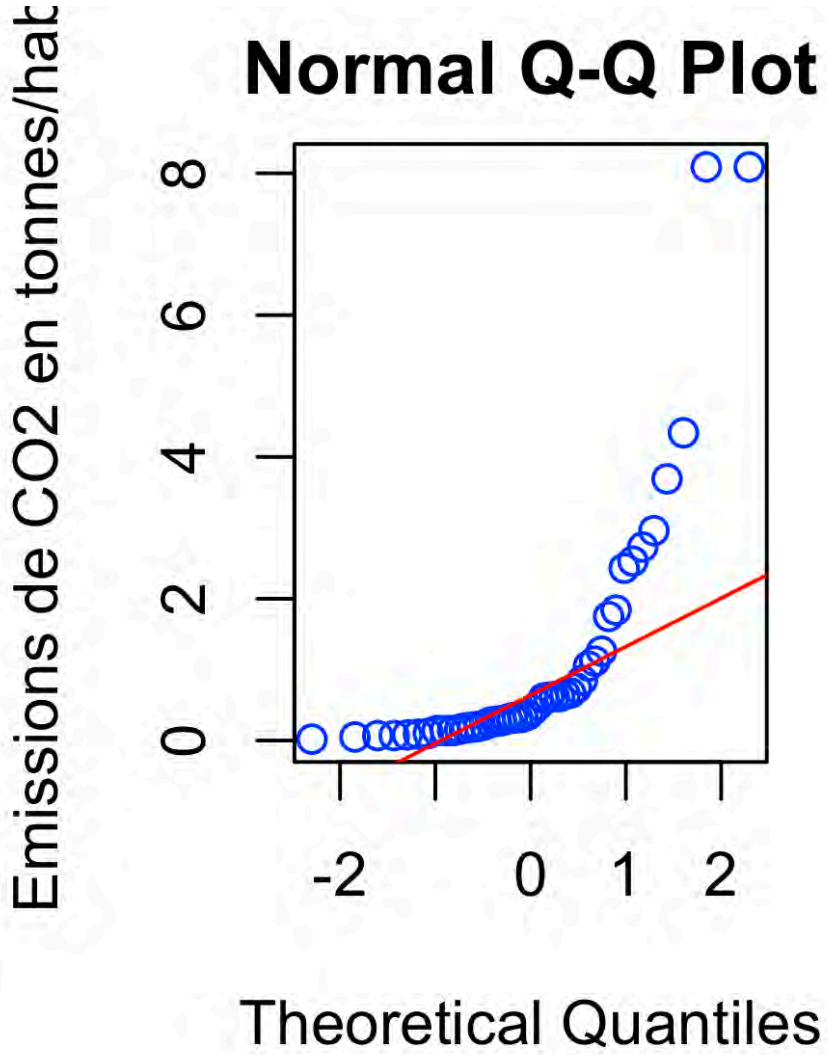
En pratique ...





En pratique ...


Normal Q-Q Plot
Produit National brut ($/hab)
5000
15000
-2
0
1
2
Theoretical Quantiles


Normal Q-Q Plot
Emissions de CO2 en tonnes/hat
0
2
4
6
8
-2
0
1
2
Theoretical Quantiles

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE


```
        Shapiro-Wilk normality test

data:  eur$X
W = 0.79601, p-value = 1.584e-06
```

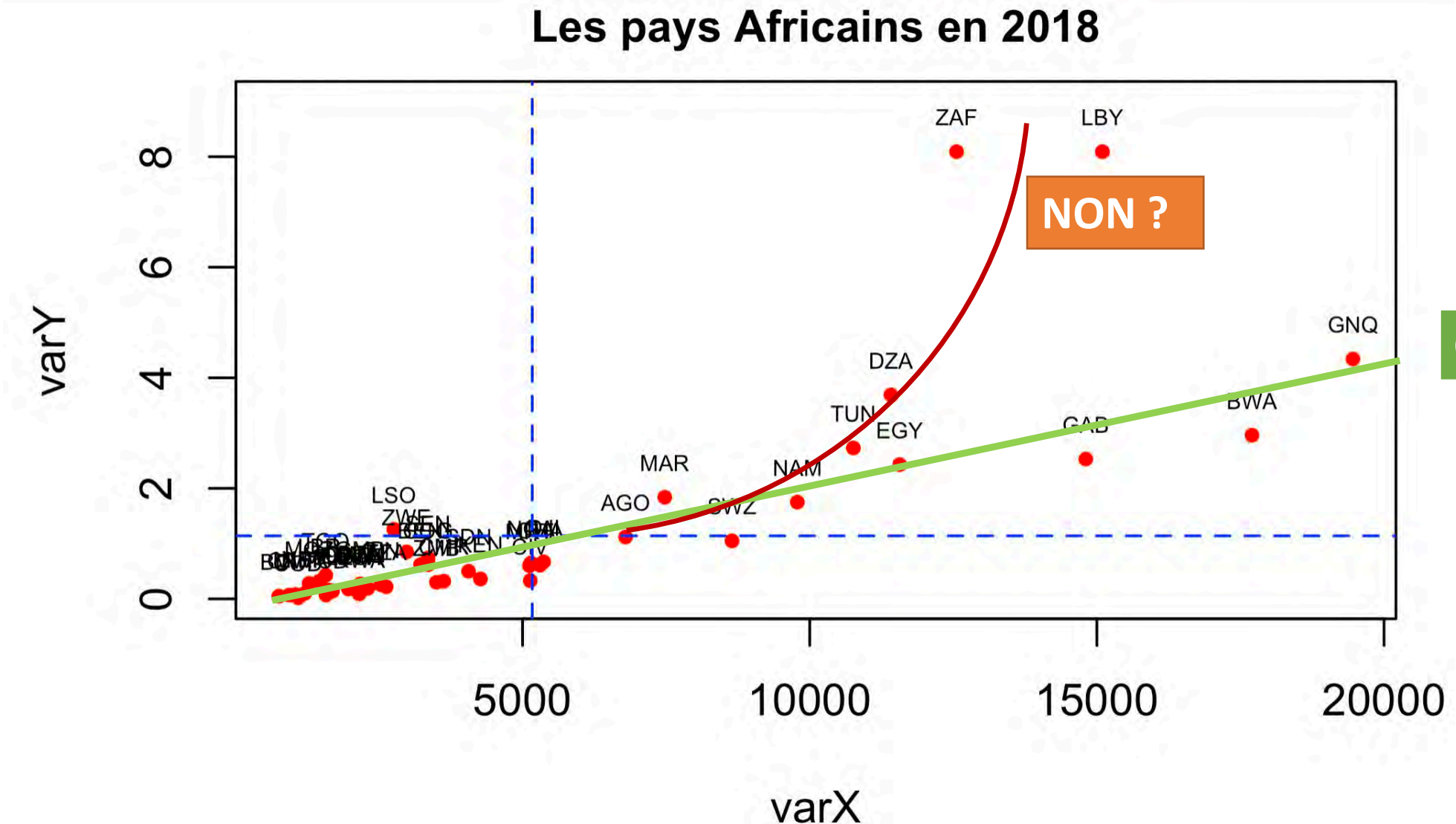
```
        Shapiro-Wilk normality test

data:  eur$Y
W = 0.60755, p-value = 7.326e-10
```

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

Hypothèse 1 : Le modèle est linéaire …



PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE


```
        Pearson's product-moment correlation

data:  eur$X and eur$Y
t = 8.971, df = 44, p-value = 1.703e-11
alternative hypothesis: true correlation is not
equal to 0
95 percent confidence interval:
 0.6702106 0.8872622
sample estimates:
      cor
0.8040684
```

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE


```
        Spearman's rank correlation rho

data:  eur$X and eur$Y
S = 1629.4, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to
0
sample estimates:
      rho
0.8995127
```

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

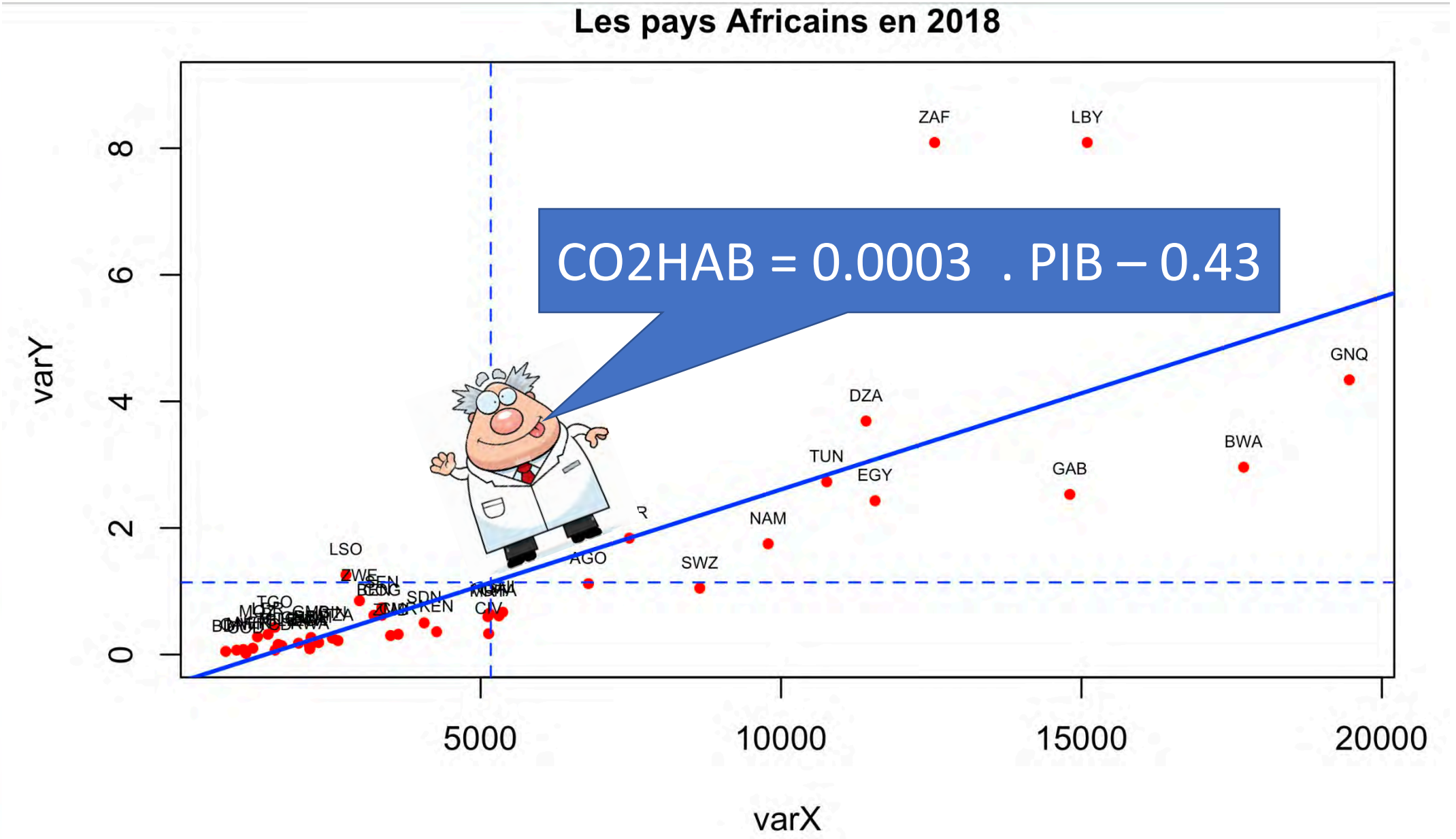
PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE


```
Residuals:
     Min       1Q   Median       3Q      Max
-1.9921 -0.5004 -0.0518  0.1611  4.7026

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.318e-01  2.377e-01  -1.817   0.0761 .
eur$X        3.042e-04  3.391e-05   8.971  1.7e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.089 on 44 degrees of freedom
Multiple R-squared:  0.6465,    Adjusted R-squared:  0.6385
F-statistic: 80.48 on 1 and 44 DF,  p-value: 1.703e-11
```

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

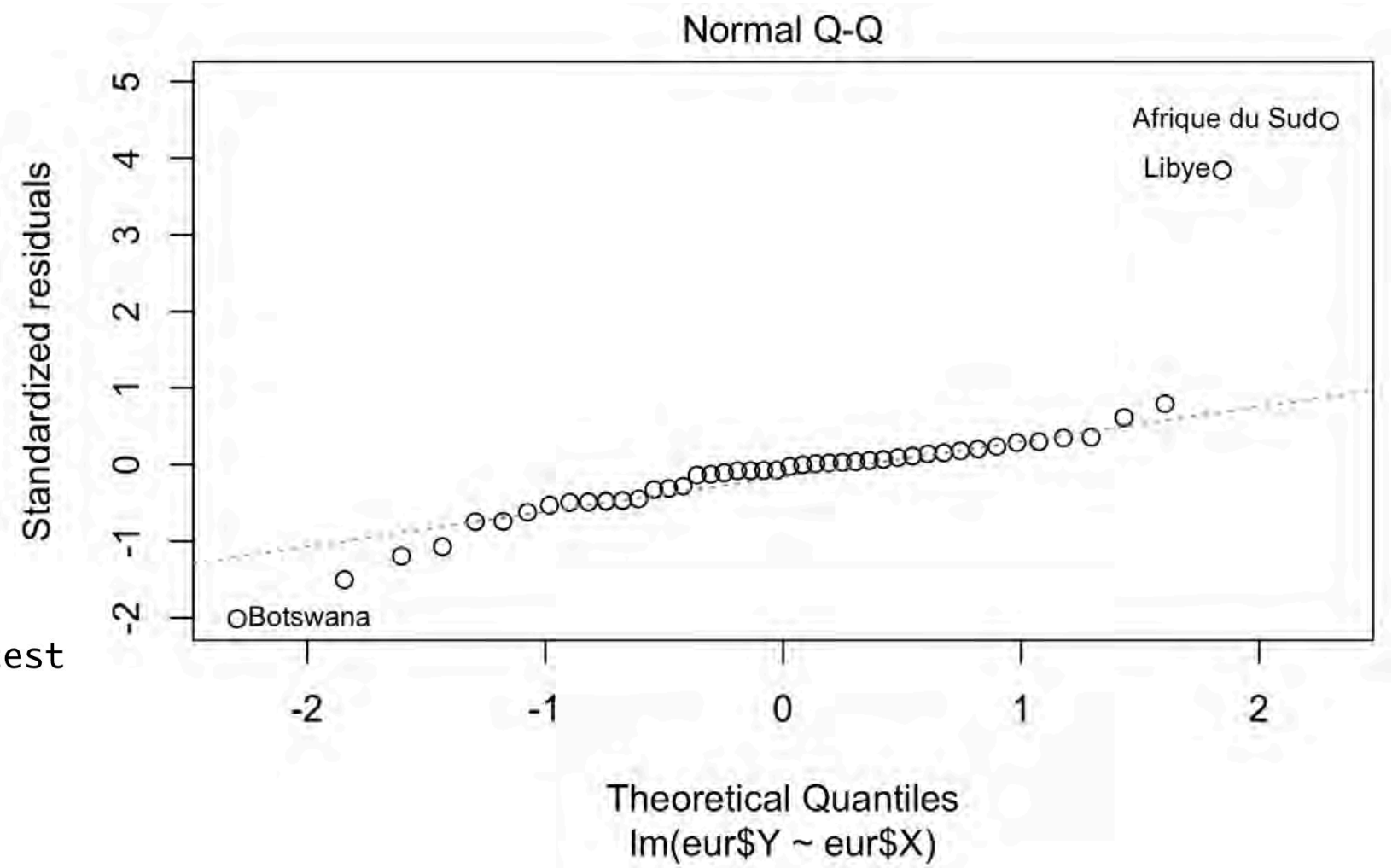


PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

- **Hypothèse 1 : Le modèle est linéaire**
- **Hypothèse 2 : L'espérance de l'erreur ε_i est nulle.**
- **Hypothèse 3 : Constance de la variance de ε_i.**
- **Hypothèse 4 : La normalité du terme d'erreur de ε_i.**
- **Hypothèse 5 : Absence d'autocorrélation des erreurs**
- **Hypothèse 5 : Covariance nulle entre x_{i} et ε_{i}**
- **Hypothèse 7 : Exactitude de la variable indépendante**
- **Hypothèse 8 : $n > k$**

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

Hypothèse 4 : La normalité du terme d'erreur de ε_i.

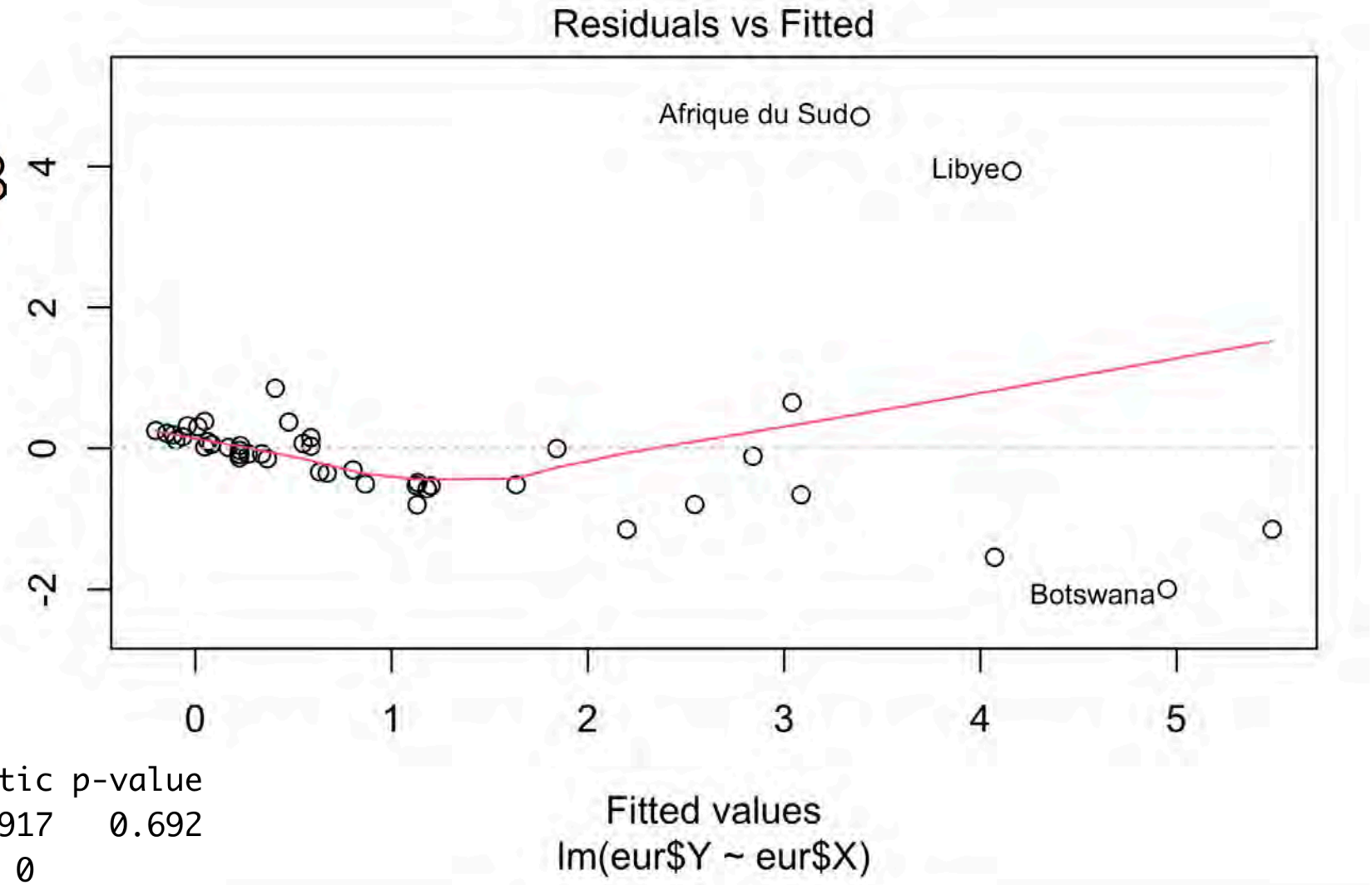



```
Shapiro-Wilk normality test

data:  monmodel$residuals
W = 0.68454, p-value = 1.17e-08
```

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

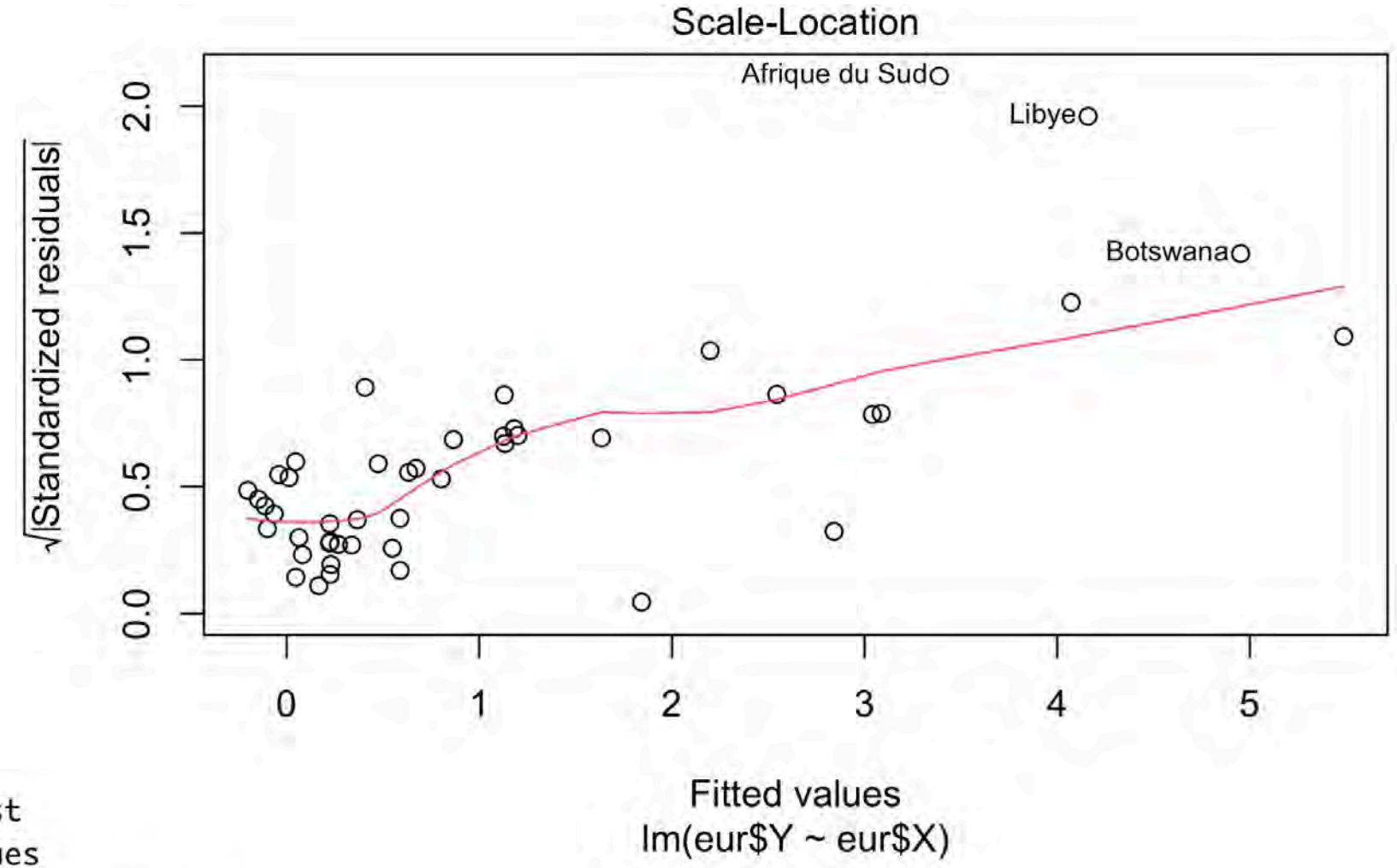
Hypothèse 5 : Absence d'autocorrélation des erreurs




```
lag Autocorrelation D-W Statistic p-value
  1      0.04015674      1.911917   0.692
Alternative hypothesis: rho != 0
```

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

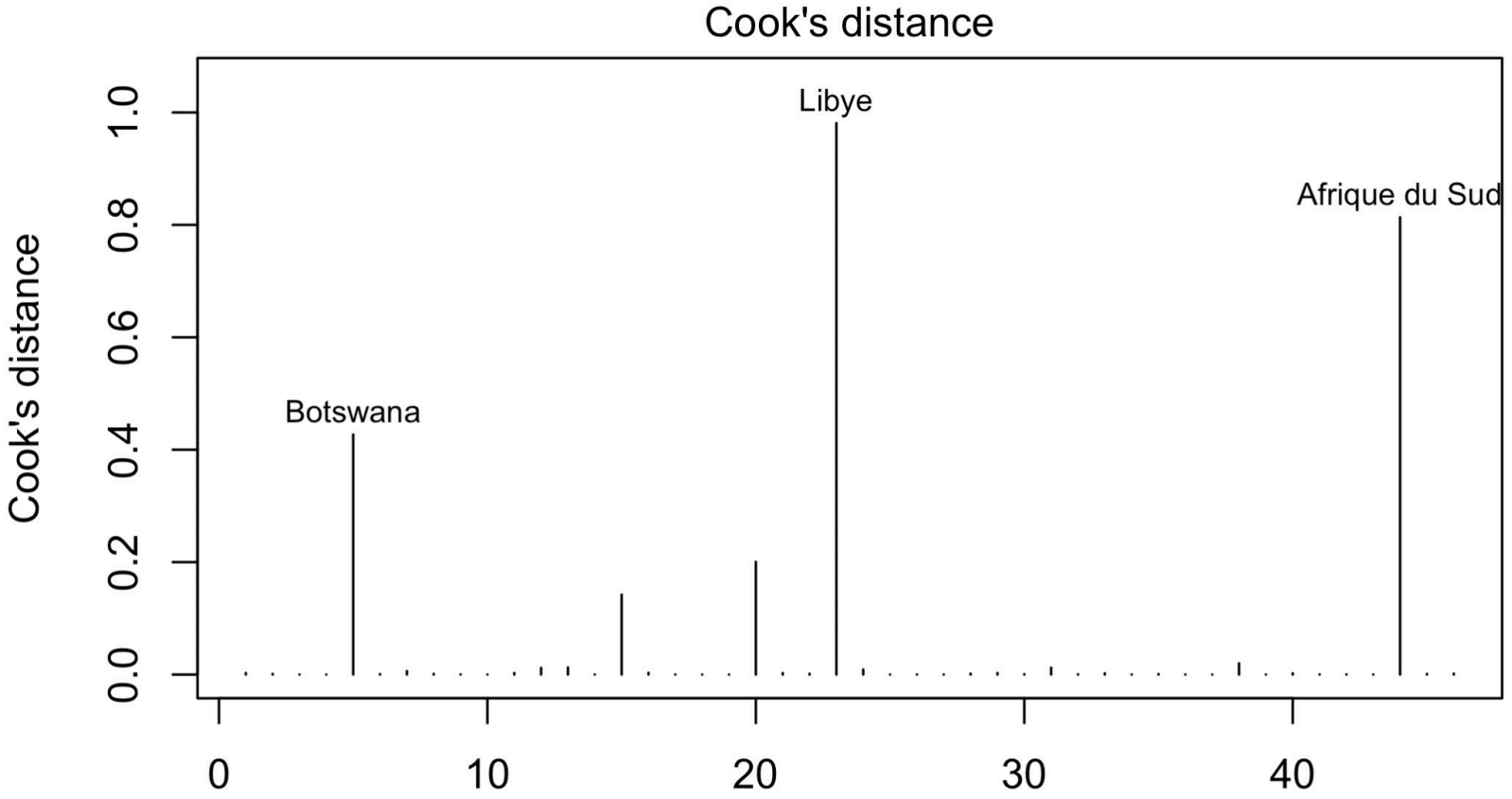
Hypothèse 3 : Constance de la variance des erreurs




```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 65.43689, Df = 1, p = 6.0005e-16
```

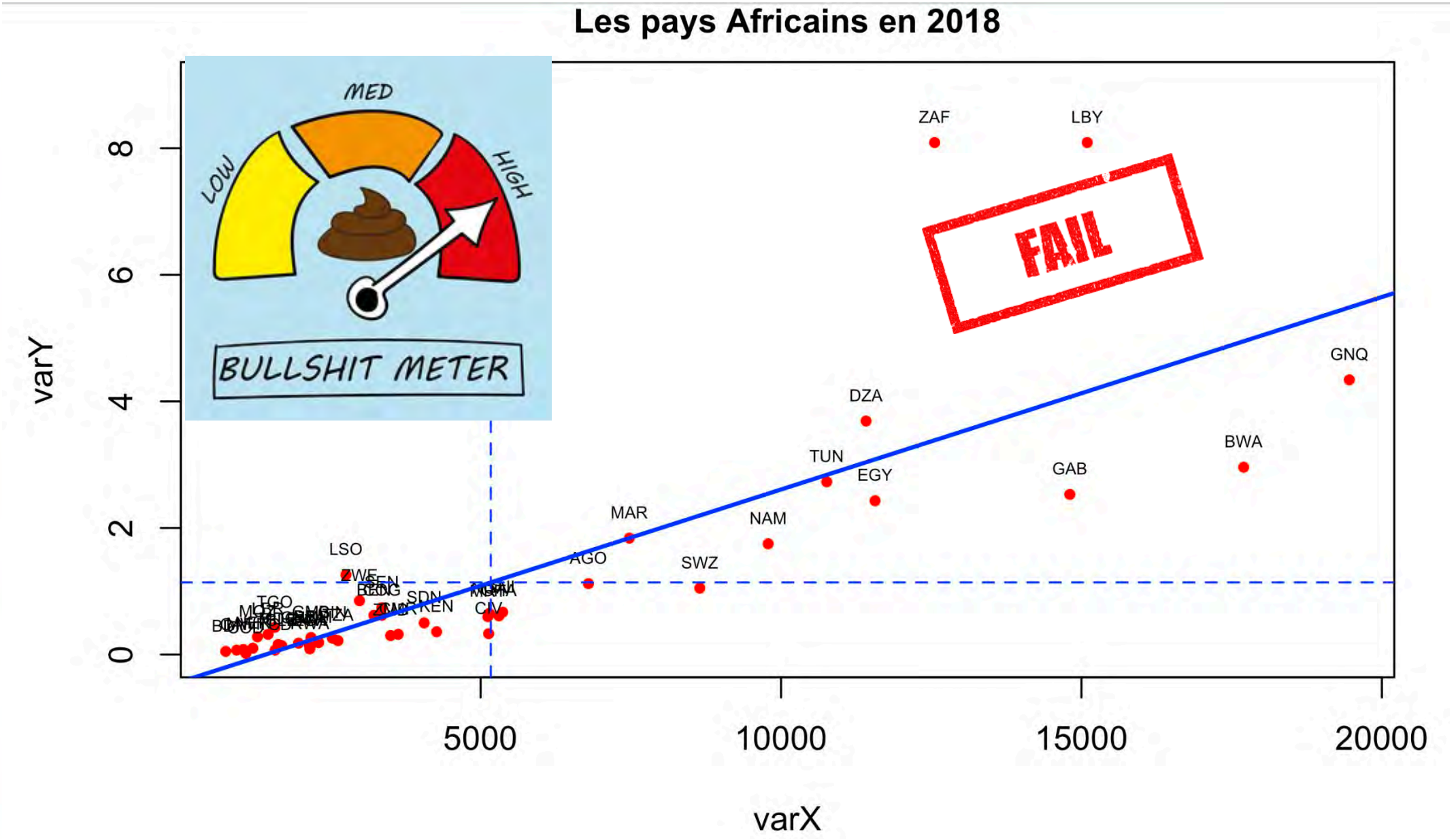
PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

Modèle non liée à quelques valeurs très influentes



	rstudent <dbl>	unadjusted p-value <dbl>	Bonferroni p <dbl>
Afrique du Sud	6.030960	3.2963e-07	1.5163e-05
Libye	4.659547	3.0612e-05	1.4081e-03

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE



PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE

PRATIQUE DE LA REGRESSION LINEAIRE SIMPLE