Modèle linéaire et extension
Régression linéaire simple

M1 Math et Interactions – UEVE/ENSIIE

semestre d'automne 2016

http://julien.cremeriefamily.info/teachings_M1MINT_Reg.html

Laboratoire de Mathématiques et Modélisation d'Évry
LaMME

université evry val-d'essonne

Plan

Régression simple
Objectif général

Idée

Expliquer les variations d'une variable **quantitative** Y à partir des valeurs observées d'une variable quantitative x.

Exemples

- Tension artérielle = f(age)
- Rendement de blé = f(dose de fertilisant)
- Concentration ozone = f(température)
- Effet d'un traitement = f(dose)
- Taux de DDT = f(age du brochet)

Régression simple
Précision sur les variables en jeu

Vocabulaire

Les rôles de Y et x ne sont **pas symétriques**:

- Y est la variable **réponse**, ou **à expliquer**
- x est la variable **explicative**, **covariable**, ou **prédictive**

Remarques

- Y est une variable aléatoire
- la covariable peut être aléatoire X ou contrôlée x
  - on la considère fixe ici (d'où le x)
- **attention** à la différence de notation majuscule/minuscule

Régression linéaire simple

Le modèle

On suppose que la vraie relation entre Y et x est linéaire:

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

- β_0 est la **constante** (**intercept**)
- β_1 est la **pente** (ou **slope**)
- ε est appelé terme d'erreur ou **résidu**
  - représente une erreur de mesure,
  - la variabilité individuelle
  - un/des facteur(s) non expliqué par le modèle

$\rightsquigarrow$ En pratique, β_0, β_1 et ε sont inconnus

Régression linéaire simple

Hypothèses statistiques

$\rightsquigarrow$ Nécessaire pour faire de l'inférence (tests, ...) !

Hypothèses sur les résidus

- $\mathbb{E}(\varepsilon) = 0$
- $\mathbb{V}(\varepsilon) = \sigma^2$
- $\varepsilon \sim N(0, \sigma^2)$

Collecte de données / échantillonnage aléatoire

Soit $\{(Y_i, x_i)\}_{i=1}^n$ un n-échantillon. On a

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

avec $\{\varepsilon_i\}_{i=1}^n$ indépendants, identiquement distribués.

Régression linéaire simple

Linéarité en les paramètres

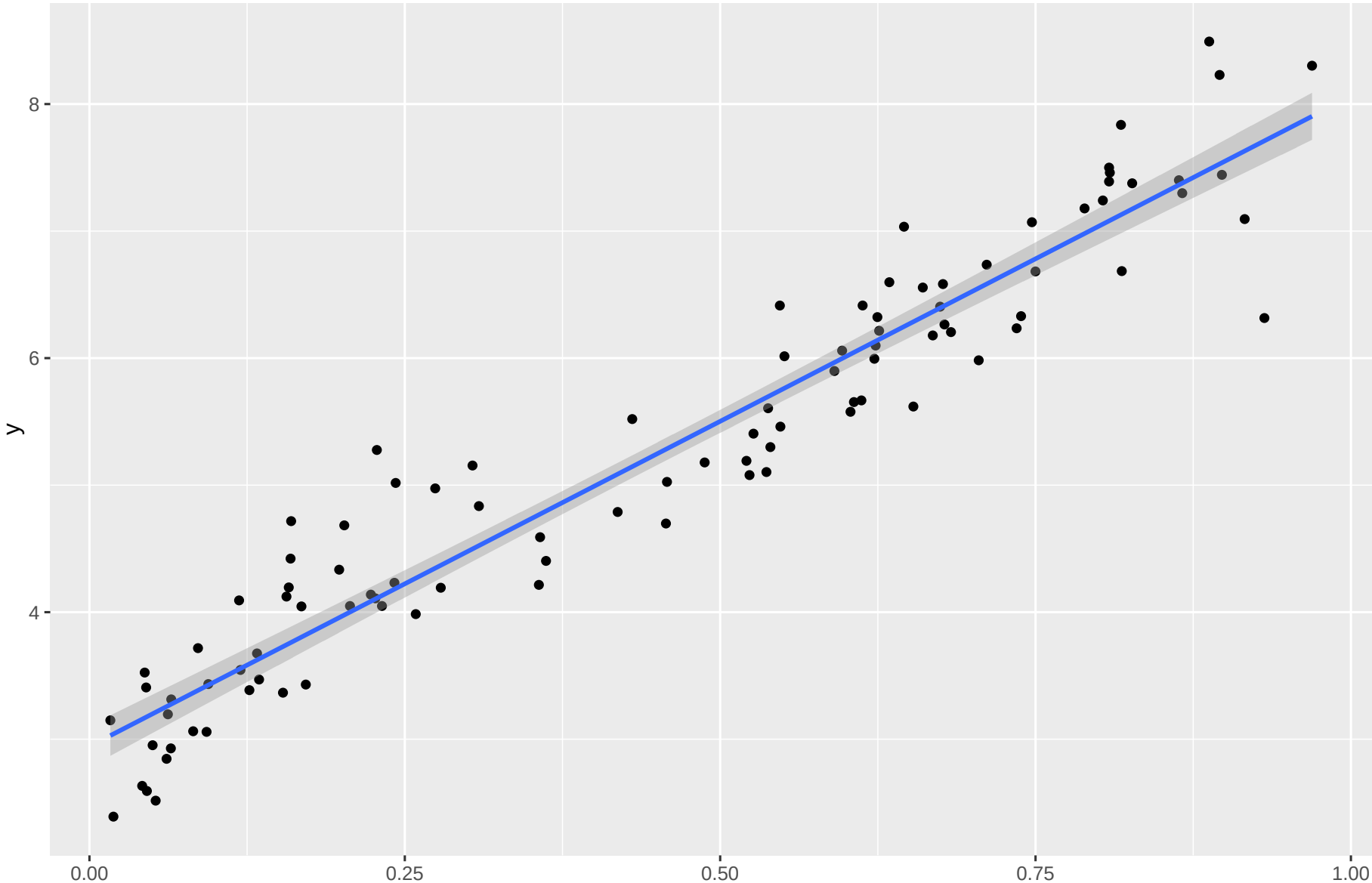
Le modèle est **linéaire en ses paramètres** (pas nécessairement en x)

```r
## true parameters
beta0 <- 3; beta1 <- 5; sigma <- .5

## simulation parameters
n <- 100
x <- runif(n,0,1)
epsilon <- rnorm(n,0,sigma)

## data generation
## linear in x and (beta0,beta1)
d1 <- data.frame(x=x,y=beta0 + beta1 * x + epsilon)
## linear in  (beta0,beta1)
d2 <- data.frame(x=x,y=beta0 + beta1 * x^2 + epsilon)
## linear in  (beta0,beta1)
d3 <- data.frame(x=x,y=beta0 + beta1 * log(x) + epsilon)
## linear in  (beta0,beta1) (after log transoform)
d4 <- data.frame(x=x,y= beta0 *exp(beta1 * x) + epsilon)
## not linear in  (beta0,beta1)
d5 <- data.frame(x=x,y= beta0 *exp(sin(beta1 * x)) + epsilon)
```

Régression linéaire simple

Linéarité en les paramètres (modèle 1)

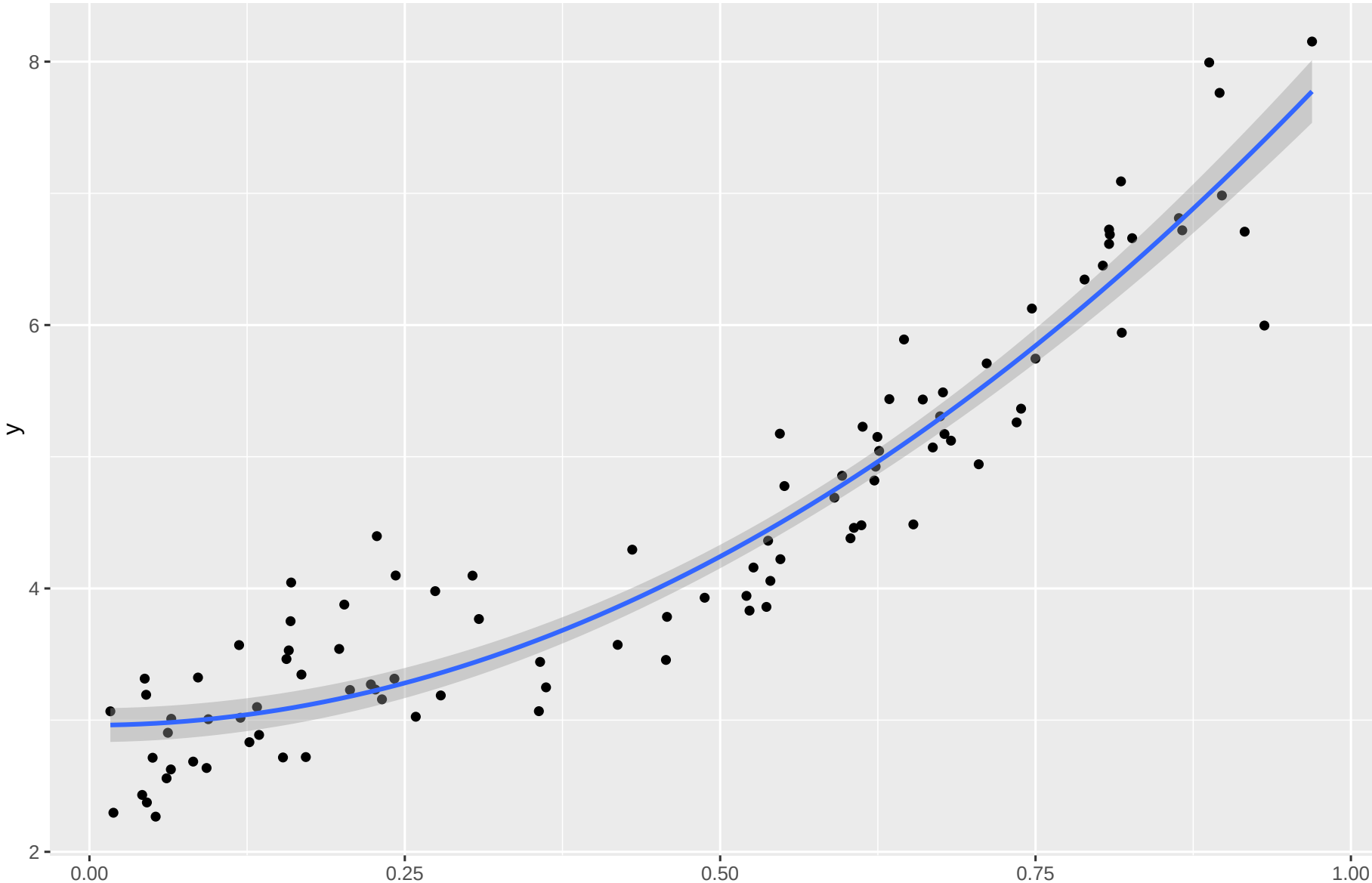
```
ggplot(d1,aes(x,y)) + geom_point() + stat_smooth(method="lm", formula=y~x)
```



Régression linéaire simple

Linéarité en les paramètres (modèle 2)

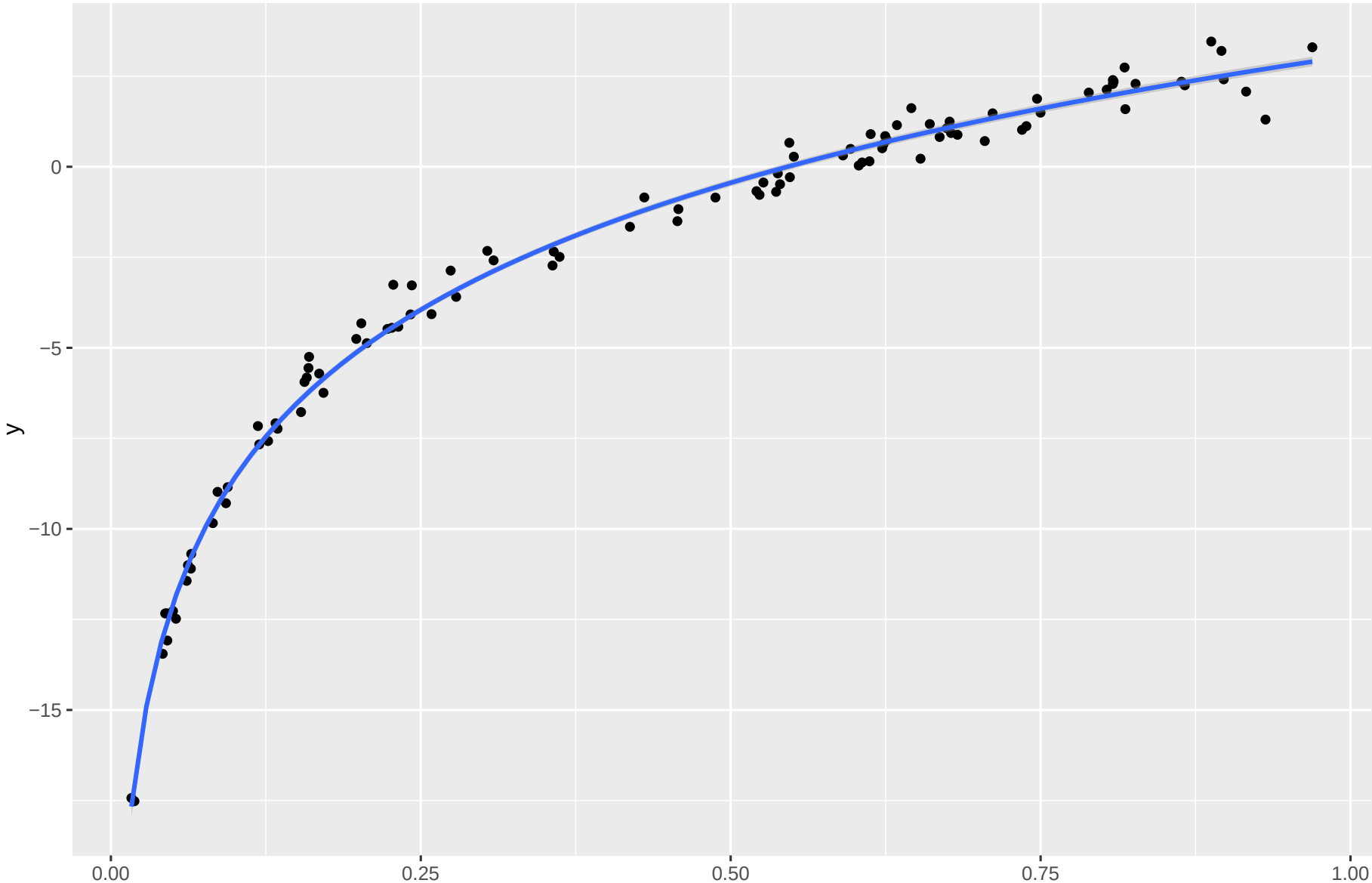
```
ggplot(d2,aes(x,y)) + geom_point() + stat_smooth(method="lm", formula=y~I(x^2))
```



Régression linéaire simple

Linéarité en les paramètres (modèle 3)

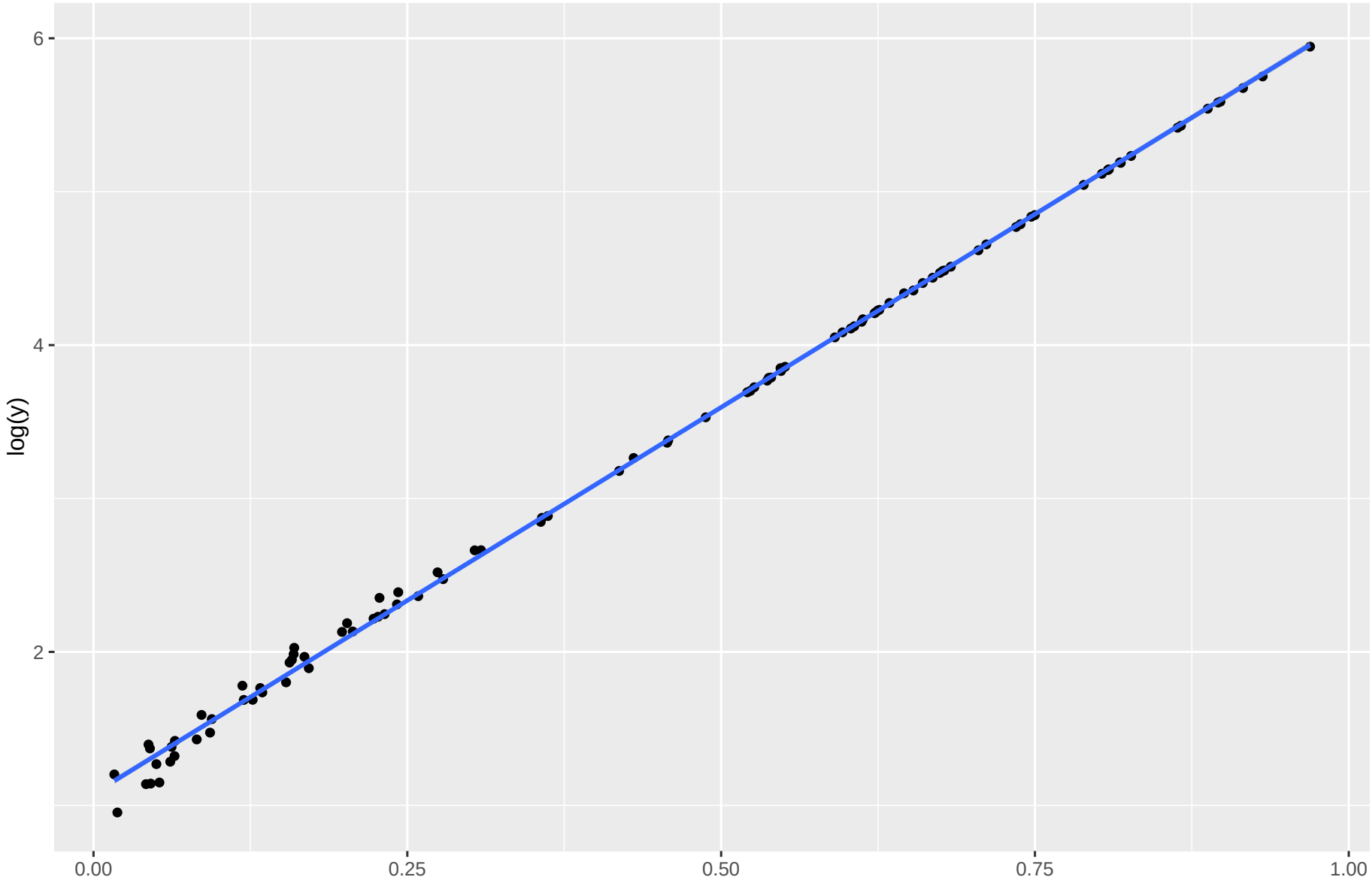
```
ggplot(d3,aes(x,y)) + geom_point() + stat_smooth(method="lm", formula=y~I(log(x)))
```



Régression linéaire simple

Linéarité en les paramètres (modèle 4)

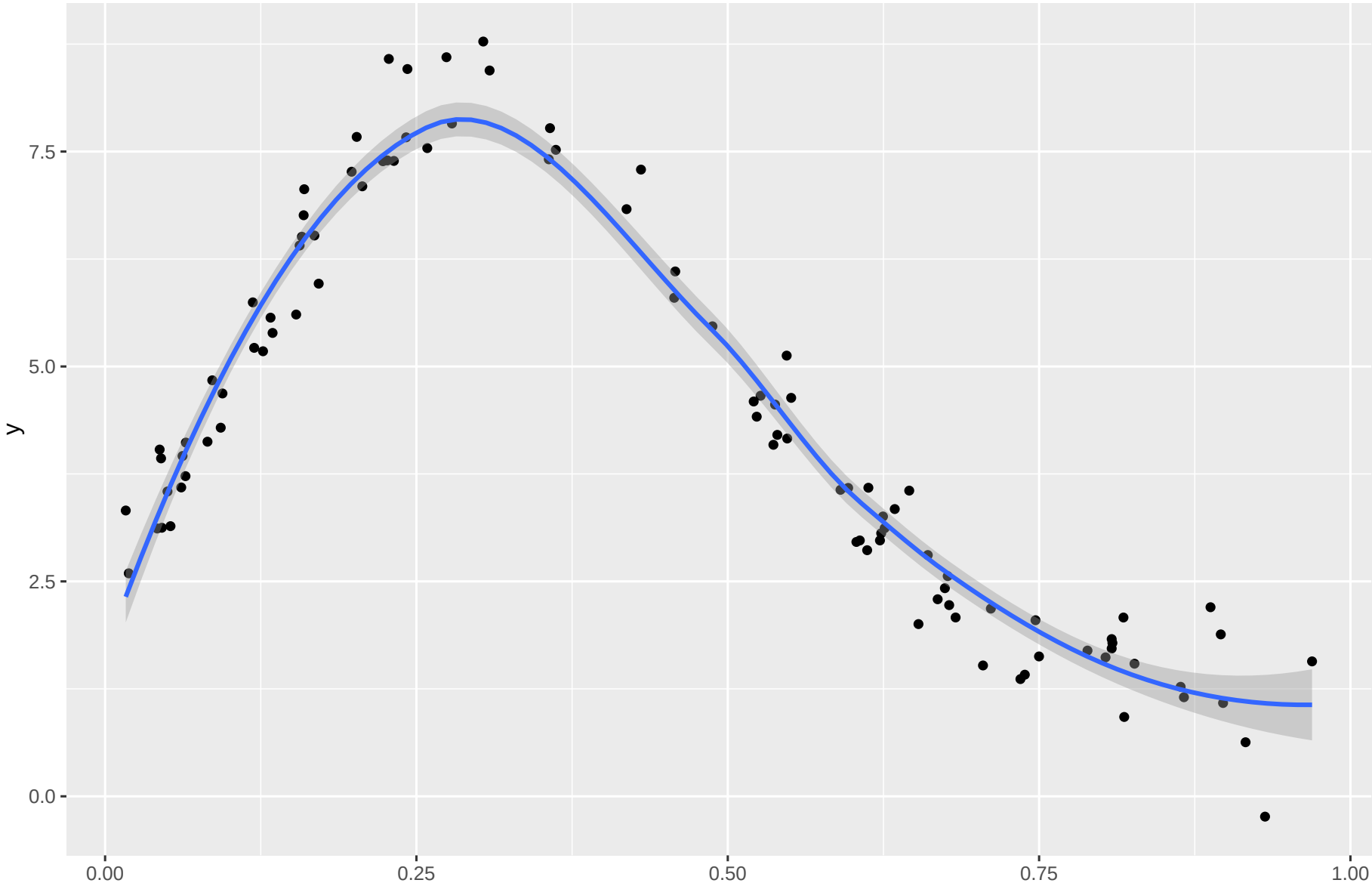
```
ggplot(d4,aes(x,log(y))) + geom_point() + stat_smooth(method="lm", formula=y ~ x)
```



Régression linéaire simple

Linéarité en les paramètres (modèle 5)

```
ggplot(d5,aes(x,y)) + geom_point() + stat_smooth()
```



Régression linéaire simple
En résumé

Objectifs statistiques

1. Estimer les paramètres β_0, β_1 et σ^2
2. Tester la nullité des paramètres β_0, β_1
3. Prédire Y pour une nouvelle observation x_0
4. Tester la pertinence générale du modèle

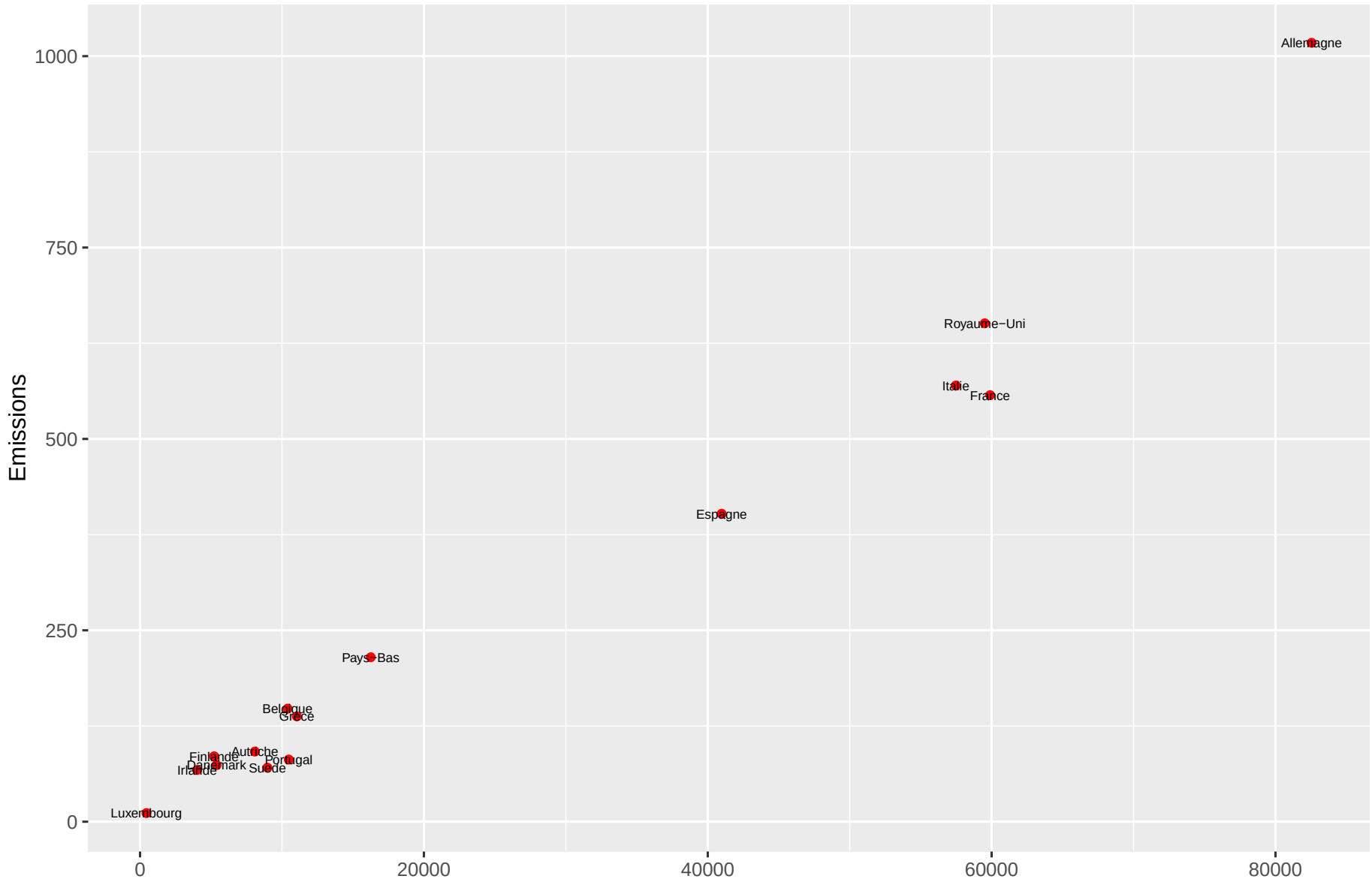
Exemple récurrent
Données Kyoto (I)

```
#### Infos
# European contries
# Population: Thousands
# Emissions:  Mil. tons CO2
# US population for prediction: 291049
Kyoto <- read.table(file='Emissions.txt',header=F)
colnames(Kyoto) <- c("Country","Population","Emissions")

head(Kyoto)
```

```
##     Country Population Emissions
## 1 Allemagne    82545.1    1017.5
## 2  Autriche     8091.9      91.6
## 3  Belgique    10396.7     147.7
## 4  Danemark     5397.6      74.0
## 5   Espagne    40977.6     402.3
## 6  Finlande     5220.2      85.5
```

Exemple récurrent
Données Kyoto (II)

```
ggplot(Kyoto, aes(Population,Emissions,label=Country)) + geom_point(colour="red") +
```



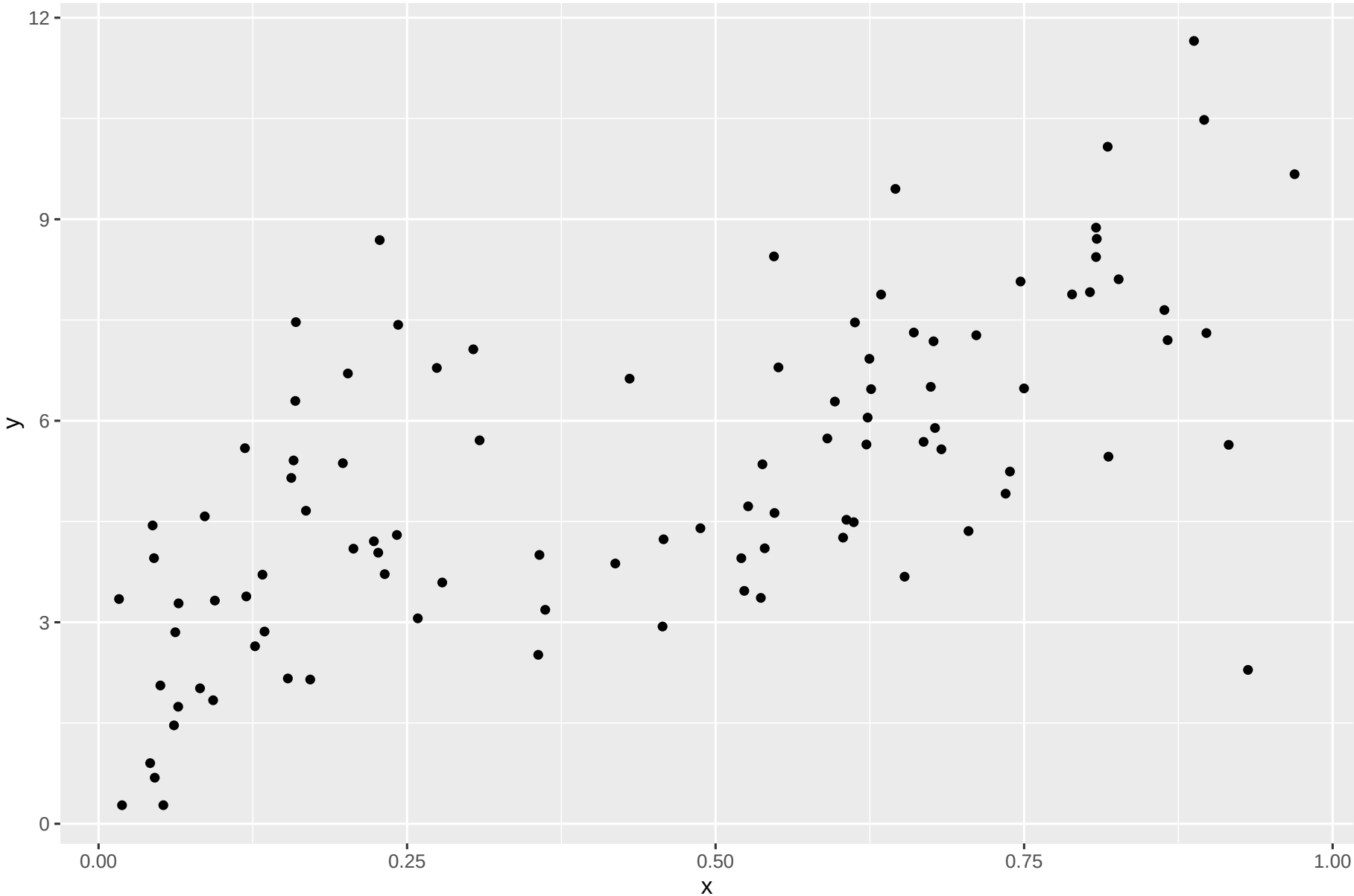
Plan

Plan

Moindres carrés ordinaires

Intuition

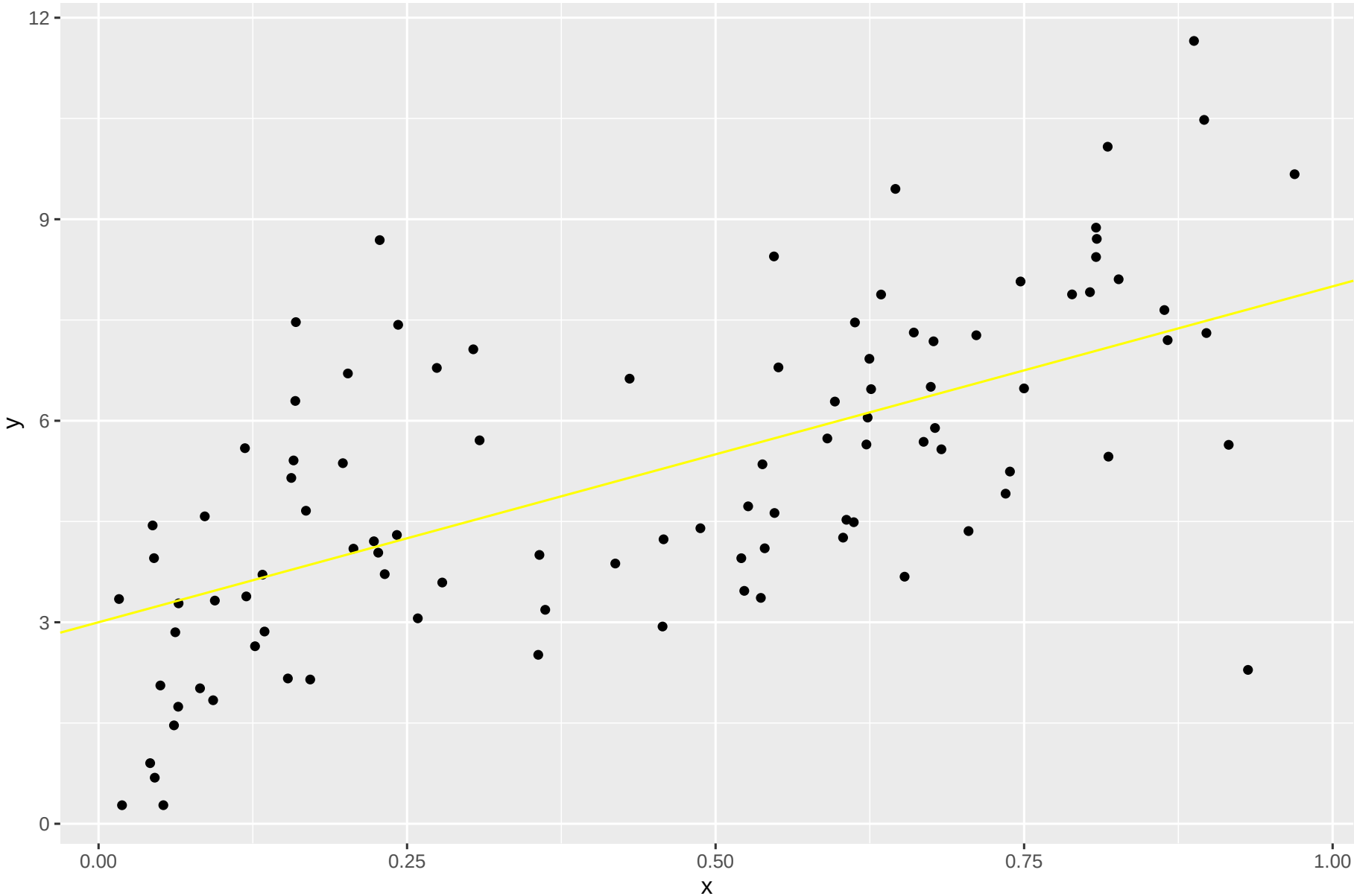
On dispose de points de l'**échantillon**.



Moindres carrés ordinaires

Idée

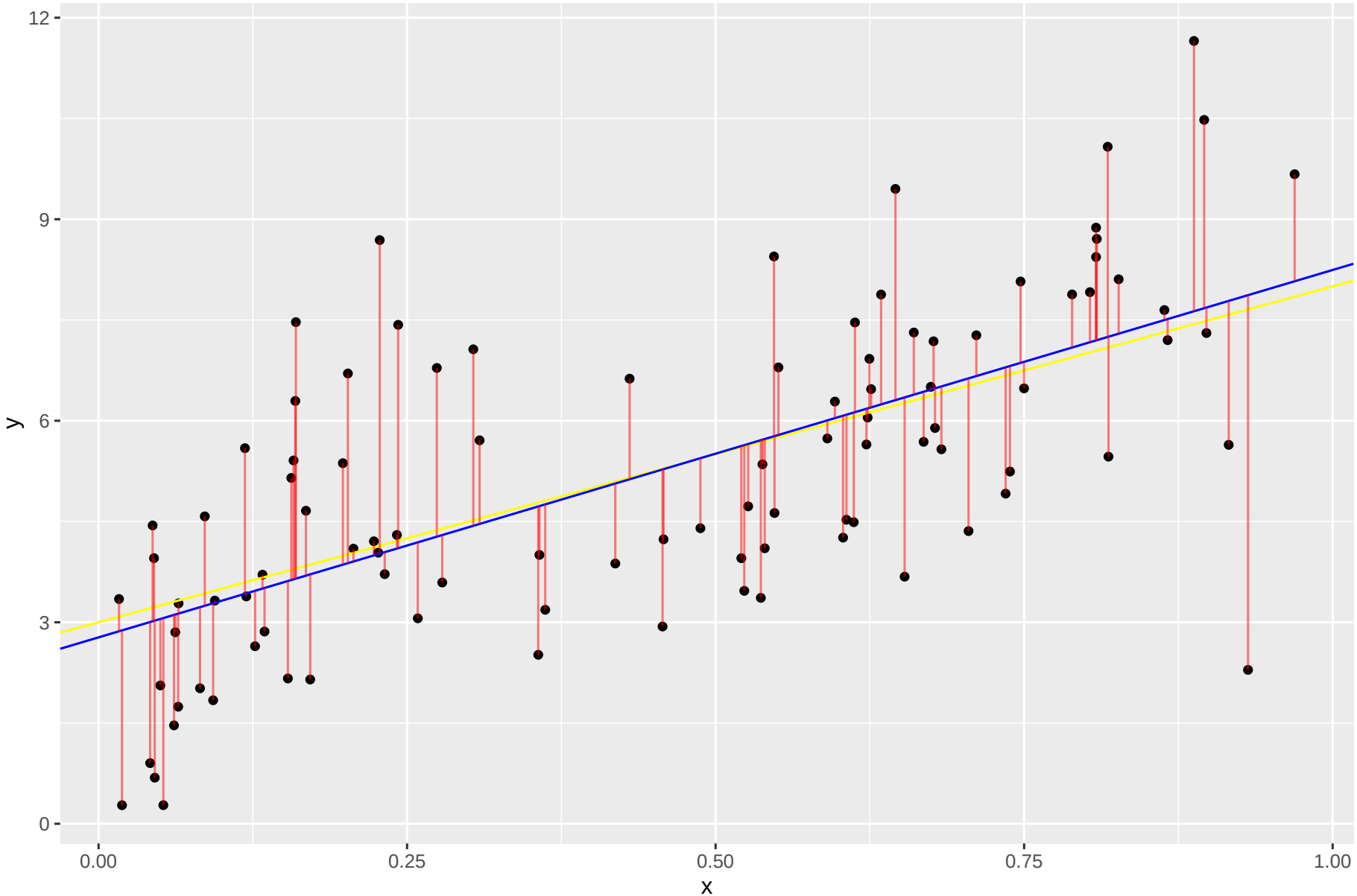
La **"vrai"** droite passe au plus près des points de la **population**.



Moindres carrés ordinaires

Idée

On cherche celle passant **au plus près** des point de l'**échantillon**



Moindres carrés ordinaires

Le critère

Formalisme

- distance à un point de l'échantillon: $(y_i - x_i\beta_1 - \beta_0)^2$
- distance à l'ensemble des points: $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$

$\rightsquigarrow$ Meilleure droite: constante $\hat{\beta}_0$ et pente $\hat{\beta}_1$ tels que
$\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$ soit minimum parmi tous les β_0, β_1 possible.

Estimateurs OLS

Les valeurs estimées (estimations) de β_0 et β_1 par OLS vérifient

Moindres carrés ordinaires

Le critère

Formalisme

- distance à un point de l'échantillon: $(y_i - x_i\beta_1 - \beta_0)^2$
- distance à l'ensemble des points: $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$

$\rightsquigarrow$ Meilleure droite: constante $\hat{\beta}_0$ et pente $\hat{\beta}_1$ tels que $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$ soit minimum parmi tous les β_0, β_1 possible.

Estimateurs OLS

Les valeurs estimées (estimations) de β_0 et β_1 par OLS vérifient

Moindres carrés ordinaires

Le critère

Formalisme

- distance à un point de l'échantillon: $(y_i - x_i\beta_1 - \beta_0)^2$
- distance à l'ensemble des points: $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$

$\rightsquigarrow$ Meilleure droite: constante $\hat{\beta}_0$ et pente $\hat{\beta}_1$ tels que $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$ soit minimum parmi tous les β_0, β_1 possible.

Estimateurs OLS

Les valeurs estimées (estimations) de β_0 et β_1 par OLS vérifient

Moindres carrés ordinaires

Le critère

Formalisme

- distance à un point de l'échantillon: $(y_i - x_i\beta_1 - \beta_0)^2$
- distance à l'ensemble des points: $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$

$\rightsquigarrow$ Meilleure droite: constante $\hat{\beta}_0$ et pente $\hat{\beta}_1$ tels que $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$ soit minimum parmi tous les β_0, β_1 possible.

Estimateurs OLS

Les valeurs estimées (estimations) de β_0 et β_1 par OLS vérifient

$$(\hat{\beta}_0^{\text{ols}}, \hat{\beta}_1^{\text{ols}}) = \underset{\beta_0,\beta_1\in\mathbb{R}}{\arg\min} \left\{ \sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2 \right\}$$

Moindres carrés ordinaires

Le critère

Formalisme

- distance à un point de l'échantillon: $(y_i - x_i\beta_1 - \beta_0)^2$
- distance à l'ensemble des points: $\sum_{i=1}^n (y_i - x_i\beta_1 - \beta_0)^2$

$\rightsquigarrow$ Meilleure droite: constante $\hat{\beta}_0$ et pente $\hat{\beta}_1$ tels que $\sum_{i=1}^n (y_i - x_i\beta_1 - \beta_0)^2$ soit minimum parmi tous les β_0, β_1 possible.

Estimateurs OLS

Les valeurs estimées (estimations) de β_0 et β_1 par OLS vérifient

$$(\hat{\beta}_0^{\text{ols}}, \hat{\beta}_1^{\text{ols}}) = \underset{\beta_0,\beta_1\in\mathbb{R}}{\arg\min} \left\| \mathbf{y} - \mathbf{x}\beta_1 - \mathbf{1}_n\beta_0 \right\|_2^2$$

Moindres carrés ordinaires

Estimateurs

Théorème

Les estimateurs des moindres carrés ordinaires ont pour expressions :

$$\hat{B}_0^{\text{ols}} = \overline{Y} - \hat{\beta}_1 \overline{x}$$

$$\hat{B}_1^{\text{ols}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{S_{xY}}{S_x x}.$$

Preuve: en annulant les dérivées de la fonction objectif, qui est convexe.

Remarques

- ne repose pas sur l'hypothèse gaussienne des résidus
- attention à la différence estimateur/estimation (v.a./réalisation)
- ne dit **rien sur** σ^2...

Moindres carrés ordinaires

Estimateurs

Théorème

Les estimateurs des moindres carrés ordinaires ont pour expressions :

$$\hat{B}_0^{\text{ols}} = \overline{Y} - \hat{\beta}_1 \overline{x}$$

$$\hat{B}_1^{\text{ols}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{S_{xY}}{S_x x}.$$

Preuve: en annulant les dérivées de la fonction objectif, qui est convexe.

Remarques

- ne repose pas sur l'hypothèse gaussienne des résidus
- attention à la différence estimateur/estimation (v.a./réalisation)
- ne dit **rien sur σ^2...**

Moindres carrés ordinaires

Application aux données Kyoto

```
x <- Kyoto$Population
y <- Kyoto$Emissions
beta1.ols <- cov(x,y) / var(x)
beta0.ols <- mean(y) - beta1.ols * mean(x)
beta1.ols
```

```
## [1] 0.01082331
```

```
beta0.ols
```

```
## [1] 3.915303
```

```
coefficients(lm(y~x)) ## sanity check
```

```
## (Intercept)           x
##  3.91530293  0.01082331
```

Plan

Maximum de vraisemblance

critère

Formalisme

- vraisemblance d'un point de l'échantillon: $L(y_i) = f(y_i)$
- vraisemblance du n-échantillon: $L(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i)$
- log-vraisemblance : $\log L(y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i)$

Estimateurs du MV

Les valeurs estimées (estimations) de β_0, β_1 et σ vérifient

Maximum de vraisemblance

critère

Formalisme

- vraisemblance d'un point de l'échantillon: $L(y_i) = f(y_i)$
- vraisemblance du n-échantillon: $L(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i)$
- log-vraisemblance : $\log L(y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i)$

Estimateurs du MV

Les valeurs estimées (estimations) de β_0, β_1 et σ vérifient

Maximum de vraisemblance

critère

Formalisme

- vraisemblance d'un point de l'échantillon: $L(y_i) = f(y_i)$
- vraisemblance du n-échantillon: $L(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i)$
- log-vraisemblance : $\log L(y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i)$

$\rightsquigarrow$ Meilleurs estimateurs: $(\beta_0, \beta_1, \sigma)$ maximisant L ou $\log L$, indiquant à quel point les valeurs courantes des paramètres sont **vraisemblables** au vu des données (fixées)

Estimateurs du MV

Les valeurs estimées (estimations) de β_0, β_1 et σ vérifient

Maximum de vraisemblance

critère

Formalisme

- vraisemblance d'un point de l'échantillon: $L(y_i) = f(y_i)$
- vraisemblance du n-échantillon: $L(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i)$
- log-vraisemblance : $\log L(y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i)$

Estimateurs du MV

Les valeurs estimées (estimations) de β_0, β_1 et σ vérifient

$$(\hat{\beta}_0^{\mathsf{mv}}, \hat{\beta}_1^{\mathsf{mv}}, \hat{\sigma}^{\mathsf{mv}}) = \underset{\beta_0, \beta_1 \in \mathbb{R}, \sigma > 0}{\arg\max} \log L(y_1, \ldots, y_n)$$

Maximum de vraisemblance

critère

Formalisme

- vraisemblance d'un point de l'échantillon: $L(y_i) = f(y_i)$
- vraisemblance du n-échantillon: $L(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i)$
- log-vraisemblance : $\log L(y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i)$

Estimateurs du MV

Les valeurs estimées (estimations) de β_0, β_1 et σ vérifient

$$(\hat{\beta}_0^{\mathsf{mv}}, \hat{\beta}_1^{\mathsf{mv}}, \hat{\sigma}^{\mathsf{mv}}) = \underset{\beta_0, \beta_1 \in \mathbb{R}}{\arg\min} \left\{ -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{x}\beta_1 - \mathbf{1}_n\beta_0\right\|_2^2 \right\}$$

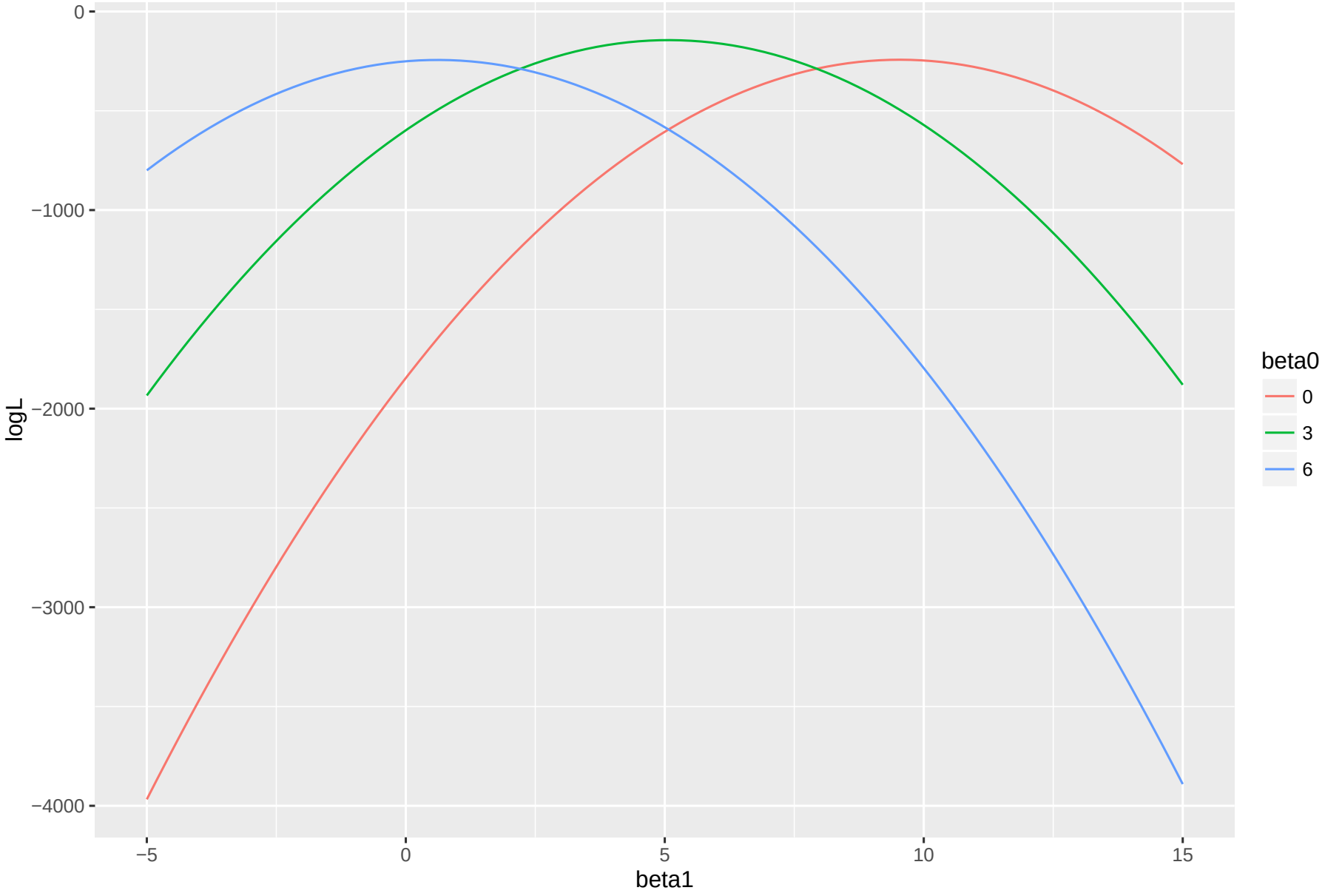
Maximum de vraisemblance I

Intuition

```
loglik <- function(beta1,x,y,beta0=3,sigma=1) {
    -n*log(2*pi)/2 -n*log(sigma) - sum((y-beta0-x*beta1)^2)/(2*sigma^2)
}
n <- 100
x <- runif(n,0,1);
beta0 <- 3; beta1 <- 5; sigma <- 1
y <- beta0 + x*beta1 + sigma*rnorm(n)

beta1 <- seq(-5,15,len=100)
logL.1 <- sapply(beta1, loglik, x=x, y=y , beta0=0,sigma=sigma)
logL.2 <- sapply(beta1, loglik, x=x, y=y , beta0=3,sigma=sigma)
logL.3 <- sapply(beta1, loglik, x=x, y=y , beta0=6,sigma=sigma)
```

Maximum de vraisemblance II

Intuition



Maximum de vraisemblance

Estimateurs

Théorème

Les estimateurs du maximum de vraisemblance ont pour expression :

$$\hat{B}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

$$\hat{B}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$S^2 = \frac{1}{n} \|\mathbf{y} - \mathbf{x}\beta_1 - \mathbf{1}_n \beta_0\|_2^2$$

Preuve:

En annulant les dérivées de la fonction objectif, qui est concave.

Maximum de vraisemblance

Estimation pratique de la variance des résidus

On ne connaît pas β_1 et β_0 ! Si on remplace par les valeurs estimées

$$\frac{1}{n}\|\mathbf{y} - \mathbf{x}\hat{\beta}_1 - \mathbf{1}_n\hat{\beta}_0\|_2^2,$$

on obtient un estimateur *biasé*. En pratique, on utilise

$$S^{\star 2} = \frac{1}{n-2}\|\mathbf{y} - \mathbf{x}\hat{\beta}_1 - \mathbf{1}_n\hat{\beta}_0\|_2^2$$

Remarque

Le "-2" provient des 2 degrés de liberté perdus en estimant β_0, β_1.

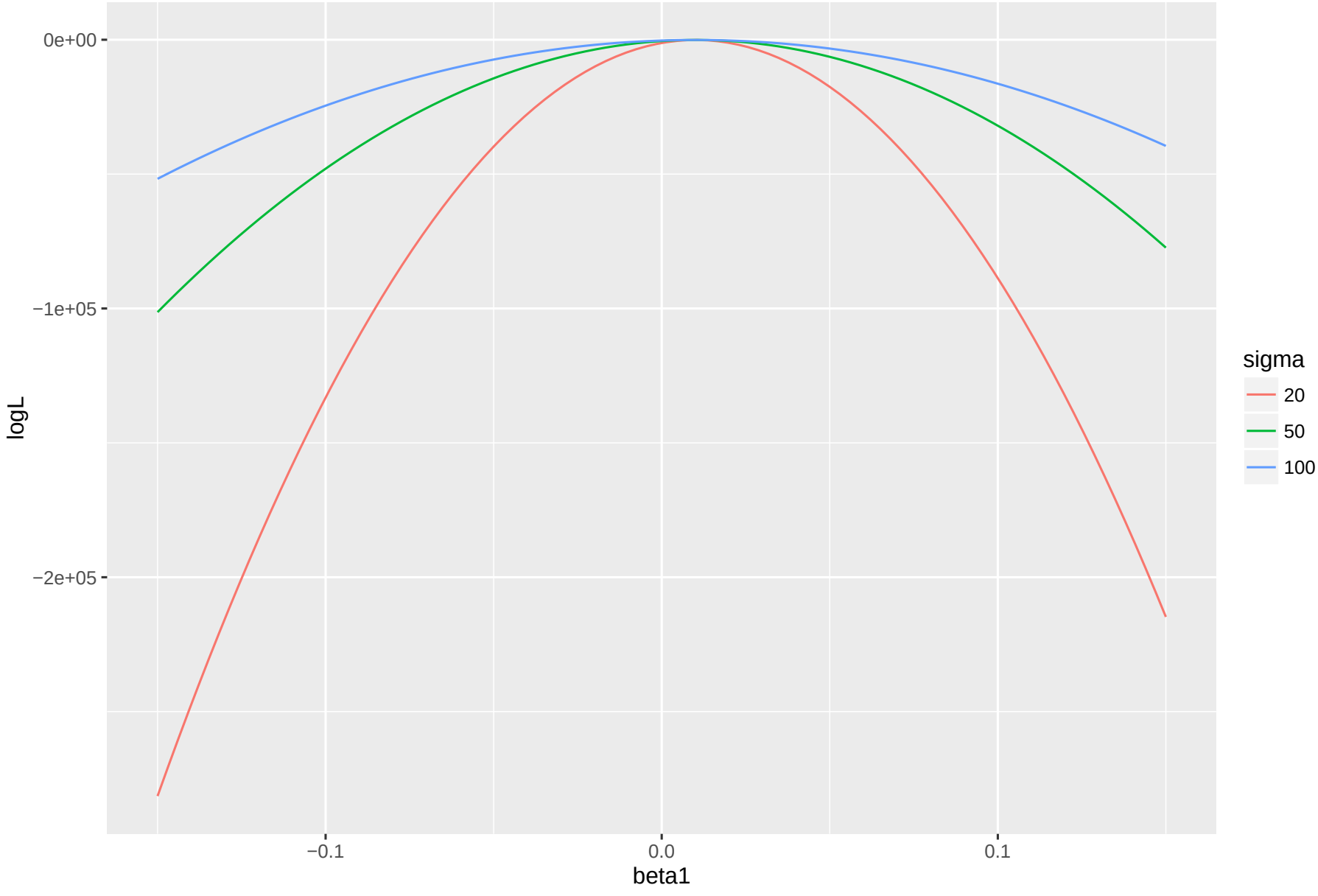
Maximum de vraisemblance I

Application aux données Kyoto

```
x <- Kyoto$Population
y <- Kyoto$Emissions
n <- length(y)
beta1 <- seq(-0.15,0.15,len=100)
logL.1 <- sapply(beta1, loglik, x=x, y=y , beta0=40,sigma=30)
logL.2 <- sapply(beta1, loglik, x=x, y=y , beta0=40,sigma=50)
logL.3 <- sapply(beta1, loglik, x=x, y=y , beta0=40,sigma=70)

sigma.hat <- sqrt(sum(residuals(lm(y~x))^2)/(n-2))
sigma.hat

## [1] 51.50069
```

Maximum de vraisemblance II

Application aux données Kyoto



Plan

Estimation des paramètres

Propriétés des estimateurs de β_0 et β_1 (I)

Cas général

$\hat{B}_0$ et $\hat{B}_1$ sont des estimateurs sans biais de β_0 et β_1 de variance

$$\mathbb{V}(\hat{B}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right),$$

$$\mathbb{V}(\hat{B}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

et de covariance $\mathrm{cov}(\hat{B}_0, \hat{B}_1) = -\dfrac{\sigma^2 \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}$

Cas gaussien

Si les résidus sont gaussien, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, alors

- $\hat{B}_0 \sim \mathcal{N}(\beta_0, \mathbb{V}(\hat{B}_0))$
- $\hat{B}_1 \sim \mathcal{N}(\beta_1, \mathbb{V}(\hat{B}_1))$

Estimation des paramètres

Propriétés des estimateurs de β_0 et β_1 (I)

Cas général

$\hat{B}_0$ et $\hat{B}_1$ sont des estimateurs sans biais de β_0 et β_1 de variance

$$\mathbb{V}(\hat{B}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right),$$

$$\mathbb{V}(\hat{B}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

et de covariance $\mathrm{cov}(\hat{B}_0, \hat{B}_1) = -\dfrac{\sigma^2 \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}$

Cas gaussien

Si les résidus sont gaussien, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, alors

- $\hat{B}_0 \sim \mathcal{N}(\beta_0, \mathbb{V}(\hat{B}_0))$
- $\hat{B}_1 \sim \mathcal{N}(\beta_1, \mathbb{V}(\hat{B}_1))$

Estimation des paramètres

Propriétés des estimateurs de B_0 et B_1 (II)

Théorème de Gauss-Markov

- Cas gaussien $\hat{B}_0$ et $\hat{B}_1$ sont les meilleurs estimateurs sans biais (i.e. de variance minimale).
- Cas non gaussien $\hat{B}_0$ et $\hat{B}_1$ sont les meilleurs estimateurs **linéaires** sans biais (i.e. de variance minimale).

Théorème

- La variance σ^2 est estimée sans biais par :

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

- Si les résidus sont normalement distribués, on a de plus

$$(n-2)S^{\star 2} \sim \sigma^2 \chi^2_{n-2}$$

Estimation des paramètres

Propriétés des estimateurs de B_0 et B_1 (II)

Théorème de Gauss-Markov

- Cas gaussien $\hat{B}_0$ et $\hat{B}_1$ sont les meilleurs estimateurs sans biais (i.e. de variance minimale).
- Cas non gaussien $\hat{B}_0$ et $\hat{B}_1$ sont les meilleurs estimateurs **linéaires** sans biais (i.e. de variance minimale).

Théorème

- La variance σ^2 est estimée sans biais par :

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

- Si les résidus sont normalement distribués, on a de plus

$$(n-2) S^{\star 2} \sim \sigma^2 \chi^2_{n-2}$$

Plan

Tests sur les paramètres du modèle: pente

Sous hypothèse de normalité des résidus

Hypothèse testée: nullité de β_1 (la pente)

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

Statistique de test et règle de décision

$$T_{\beta_1} = \frac{\hat{\beta}_1}{\sqrt{\frac{S^{\star 2}}{\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2}}} \underset{H_0}{\sim} \mathcal{T}_{n-2}, \text{ on rejette } H_0 \text{ si } |T_{\beta_1}| \geq t_{n-2,1-\frac{\alpha}{2}}$$

$p-$valeur (degré de significativité)

$$\P_{H_0}\left(|\mathcal{T}_{n-2}| \geq t_{\beta_1}(\text{obs})\right)$$

Tests sur les paramètres du modèle: pente

Sous hypothèse de normalité des résidus

Hypothèse testée: nullité de β_1 (la pente)

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

Statistique de test et règle de décision

$$T_{\beta_1} = \frac{\hat{\beta}_1}{\sqrt{\frac{S^{\star 2}}{\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2}}} \underset{H_0}{\sim} \mathcal{T}_{n-2}, \text{ on rejette } H_0 \text{ si } |T_{\beta_1}| \geq t_{n-2,1-\frac{\alpha}{2}}$$

$p-$valeur (degré de significativité)

$$\P_{H_0}\left(|\mathcal{T}_{n-2}| \geq t_{\beta_1}(\text{obs})\right)$$

Tests sur les paramètres du modèle: pente

Sous hypothèse de normalité des résidus

Hypothèse testée: nullité de β_1 (la pente)

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

Statistique de test et règle de décision

$$T_{\beta_1} = \frac{\hat{\beta}_1}{\sqrt{\frac{S^{\star 2}}{\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2}}} \underset{H_0}{\sim} \mathcal{T}_{n-2}, \text{ on rejette } H_0 \text{ si } |T_{\beta_1}| \geq t_{n-2,1-\frac{\alpha}{2}}$$

$p-$valeur (degré de significativité)

$$¶_{H_0}\left(|\mathcal{T}_{n-2}| \geq t_{\beta_1}(\text{obs})\right)$$

Tests sur les paramètres du modèle: constante

Sous hypothèse de normalité des résidus

Hypothèse testée: nullité de β_0 (la constante)

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

Statistique de test et règle de décision

$$T_{\beta_0} = \frac{\hat{\beta}_0}{\sqrt{s^{\star 2}\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}} \underset{H_0}{\sim} \mathcal{T}_{n-2}, \text{ on rejette } H_0 \text{ si } |T_{\beta_0}| \geq t_{n-2,1-\frac{\alpha}{2}}$$

$p-$valeur (degré de significativité)

$$\P_{H_0}\left(|\mathcal{T}_{n-2}| \geq t_{\beta_0}(\text{obs})\right)$$

Tests sur les paramètres du modèle: constante

Sous hypothèse de normalité des résidus

Hypothèse testée: nullité de β_0 (la constante)

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

Statistique de test et règle de décision

$$T_{\beta_0} = \frac{\hat{\beta}_0}{\sqrt{s^{\star 2}\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}} \underset{H_0}{\sim} \mathcal{T}_{n-2}, \text{ on rejette } H_0 \text{ si } |T_{\beta_0}| \geq t_{n-2,1-\frac{\alpha}{2}}$$

$p-$valeur (degré de significativité)

$$\P_{H_0}\left(|\mathcal{T}_{n-2}| \geq t_{\beta_0}(\text{obs})\right)$$

Tests sur les paramètres du modèle: constante

Sous hypothèse de normalité des résidus

Hypothèse testée: nullité de β_0 (la constante)

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

Statistique de test et règle de décision

$$T_{\beta_0} = \frac{\hat{\beta}_0}{\sqrt{s^{\star 2}\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}} \underset{H_0}{\sim} \mathcal{T}_{n-2}, \text{ on rejette } H_0 \text{ si } |T_{\beta_0}| \geq t_{n-2,1-\frac{\alpha}{2}}$$

$p-$valeur (degré de significativité)

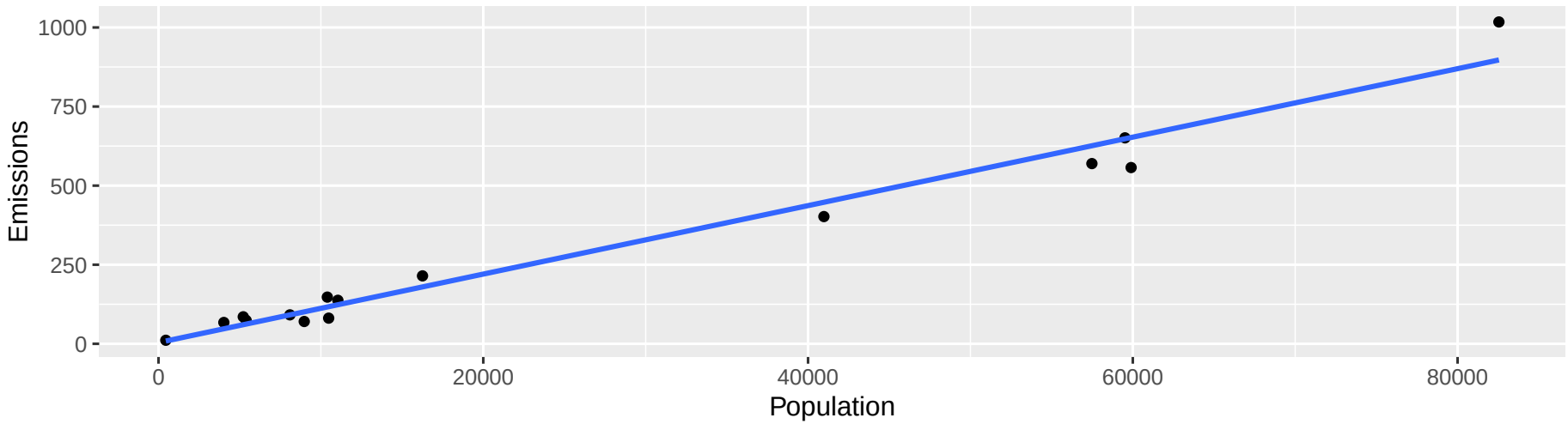
$$¶_{H_0}\left(|\mathcal{T}_{n-2}| \geq t_{\beta_0}(\text{obs})\right)$$

Test sur les paramètres
Application aux données Kyoto (I)

```
model <- lm(Emissions~Population,data=Kyoto)
summary(model)


##
## Call:
## lm(formula = Emissions ~ Population, data = Kyoto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.983 -33.297   3.004  22.605 120.173
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.915e+00  1.861e+01    0.21    0.837
## Population  1.082e-02  5.128e-04   21.11 1.93e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.5 on 13 degrees of freedom
## Multiple R-squared:  0.9716,Adjusted R-squared:  0.9695
## F-statistic: 445.4 on 1 and 13 DF,  p-value: 1.925e-11
```

Test sur les paramètres
Application aux données Kyoto (II)

```
ggplot(Kyoto, aes(Population,Emissions)) + geom_point() + geom_smooth(method=lm,se=
```



Plan

Prédiction, prédicteur

Problème

La valeur prédite par le modèle pour le i^{e} individu est

$$Y_i = \beta_0 + \beta_1 X_0 + \varepsilon_i,$$

mais β_0, β_1 et ε_i sont inconnus.

Idée

Les estimateurs et estimations de β_0 et β_1 permettent de définir

- un **prédicteur**: $\hat{Y}_i = \hat{B}_0 + \hat{B}_1 x_i$
- une **prédiction**: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Estimation des résidus

Proposition

Soit $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ l'erreur de prévision au i^{e} point. On a :

$$\mathbb{E}(\hat{\varepsilon}_i) = 0$$

$$\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_i - \bar{\mathbf{x}})^2}{\sum_{i=1}^{n} (x_i - \bar{\mathbf{x}})^2} \right)$$

Remarques

- On a $\sum \hat{\varepsilon}_i = 0$
- Contrairement à ε_i, les résidus estimés $\hat{\varepsilon}_i$ ne sont pas indépendants
- La variance de l'erreur de prédiction est d'autant plus grande que x_i est éloigné de la moyenne $\bar{\mathbf{x}}$

Prédiction d'une nouvelle observation

Valeur prédite

Soit x_0 une nouvelle observation. La valeur prédite par le modèle est $Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0$. Cette valeur peut être approchée par :

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

Remarques

Deux types d'erreurs entâchent cette prédiction :

- La non connaissance de ε_0
- L'incertitude sur l'estimation des paramètres β_0 et β_1

Prédiction: intervalle de confiance

Soit x_0 une nouvelle observation et $\hat{Y}_0$ le prédicteur associé.

Proposition

Loi de $\hat{Y}_0$ Sous l'hypothèse gaussienne, on déduite de la loi jointe de (B_0, B_1) que

$$\hat{Y}_0 \sim \mathcal{N}\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{\mathbf{x}})^2}{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2}\right)\right)$$

Remarques

- $\mathbb{V}(\hat{Y}_0) = \mathbb{V}(B_0 + B_1 x) \neq \mathbb{V}(B_0) + \mathbb{V}(B_1 x)$ car $\mathrm{cov}(B_0, B_1) \neq 0$.
- $\mathbb{V}(\hat{Y}_0)$ tient compte de l'erreur faite en estimant $\beta_0 + \beta_1 x$.
- *Plus l'on cherche à estimer $\mathbb{E}(Y_0)$ d'un point x_0 proche (resp.) éloigné de $\bar{\mathbf{x}}$, plus la variance est petite (resp. grande).*

Prédiction: intervalle de confiance

Soit x_0 une nouvelle observation et $\hat{Y}_0$ le prédicteur associé.

Proposition

Loi de $\hat{Y}_0$ Sous l'hypothèse gaussienne, on déduite de la loi jointe de (B_0, B_1) que

$$\hat{Y}_0 \sim \mathcal{N}\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{\mathbf{x}})^2}{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2}\right)\right)$$

Remarques

- $\mathbb{V}(\hat{Y}_0) = \mathbb{V}(B_0 + B_1 x) \neq \mathbb{V}(B_0) + \mathbb{V}(B_1 x)$ car $\mathrm{cov}(B_0, B_1) \neq 0$.
- $\mathbb{V}(\hat{Y}_0)$ tient compte de l'erreur faite en estimant $\beta_0 + \beta_1 x$.
- *Plus l'on cherche à estimer $\mathbb{E}(Y_0)$ d'un point x_0 proche (resp.) éloigné de $\bar{\mathbf{x}}$, plus la variance est petite (resp. grande).*

Prédiction: intervalle de prédiction

Pour l'intervalle de confiance de **prédiction**, il faut rajouter le bruit incompressible estimé, i.e., $\hat{\sigma}^2$.

Intervalle de confiance de la prédiction

$$IC_{1-\alpha}(y_0) = \left[\hat{Y}_0 \pm q_{t_{n-2},1-\frac{\alpha}{2}} \sqrt{s^{\star 2} \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)} \right]$$

$\rightsquigarrow$ pour une nouvelle valeur observée, s'ajoute l'aléa du tirage.

Prédiction, résidus

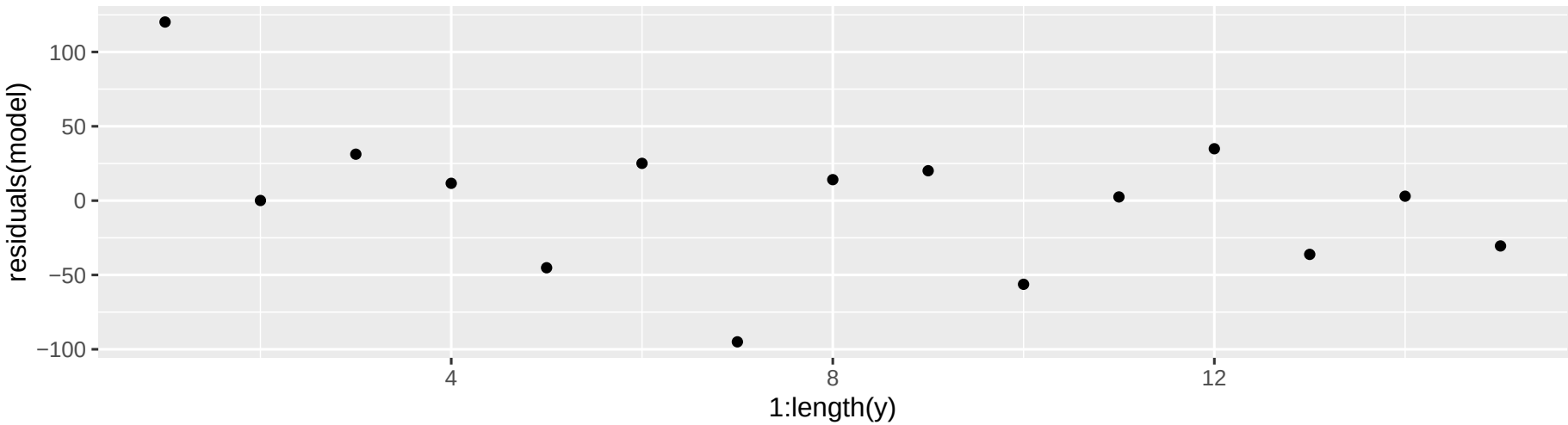
Application aux données Kyoto (I)

```
model <- lm(Emissions~Population,data=Kyoto)
## résidus estimés
head(residuals(model))

##           1           2           3           4           5           6
## 120.1732717   0.1035334  31.2579624  11.6647847 -45.1286796  25.0848403

sum(residuals(model))

## [1] -7.81597e-14

## valeurs estimés
head(fitted(model))

##         1         2         3         4         5         6
## 897.32673  91.49647 116.44204  62.33522 447.42868  60.41516
```

Prédiction, résidus

Application aux données Kyoto (II)

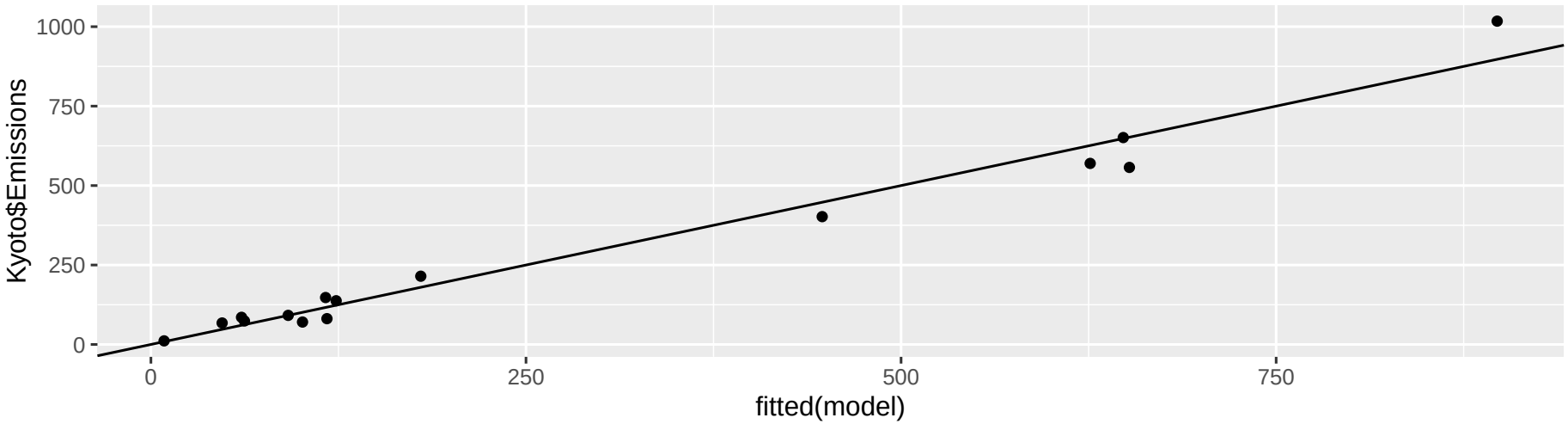
```
qplot(1:length(y),residuals(model), geom='point')
```



Prédiction, résidus

Application aux données Kyoto (III)

```
qplot(fitted(model),Kyoto$Emissions, geom='point') + geom_abline(intercept=0,slope=
```



Prédiction, résidus
Application aux données Kyoto (IV)

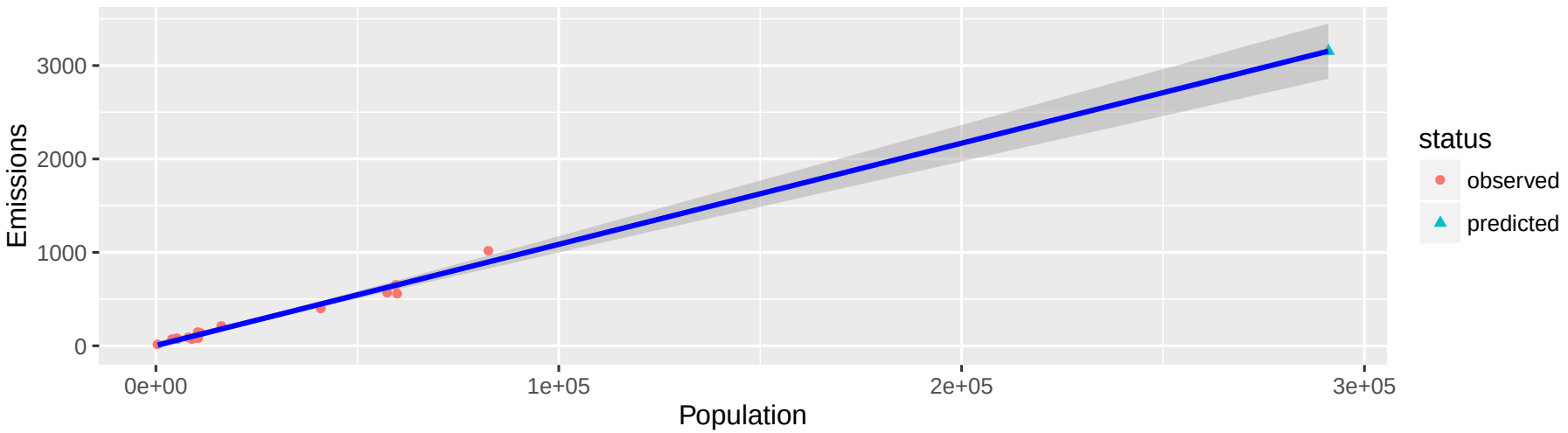
Population US pour la prédiction: 291049

```
Emission.US <- predict(model, newdata=data.frame(Population=291049), interval="conf
Emission.US

##       fit      lwr      upr
## 1 3154.03 2858.296 3449.764

Emission.US <- predict(model, newdata=data.frame(Population=291049), interval="pred
Emission.US

##       fit      lwr  upr
## 1 3154.03 2838.059 3470

Kyoto2 <- data.frame(
    Country    = c(Kyoto$Country   , "US"),
    Population = c(Kyoto$Population, 291049),
    Emissions  = c(Kyoto$Emissions , Emission.US[1]),
    status     = factor(c(rep("observed",nrow(Kyoto)),"predicted")))
```

Prédiction, résidus

Application aux données Kyoto (V)

```
ggplot(Kyoto2,aes(x=Population,y=Emissions,colour=status,shape=status)) + geom_poir
      stat_smooth(method=lm,colour="blue",fullrange=TRUE)
```

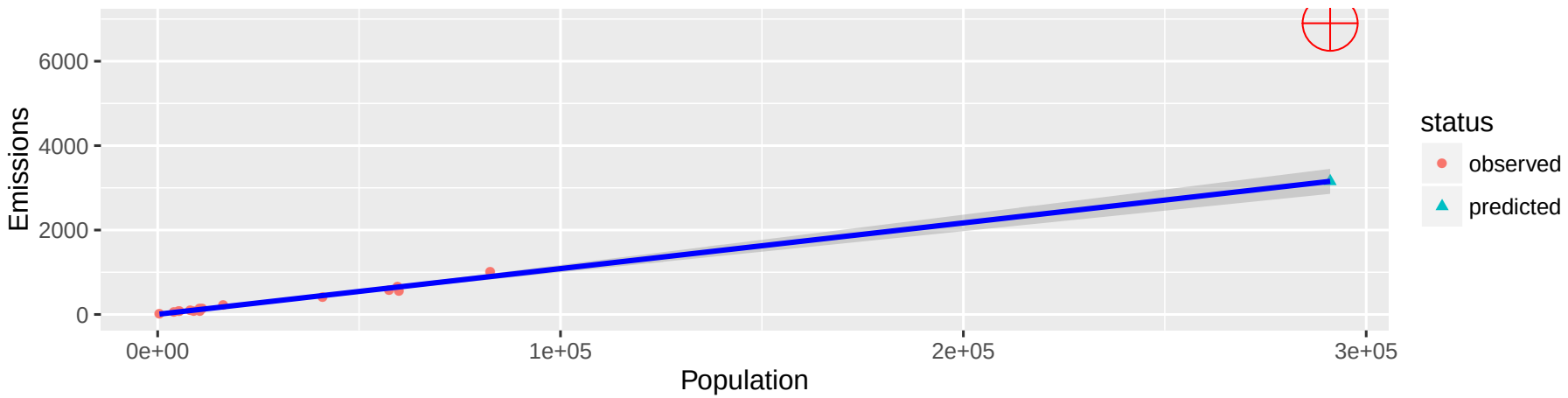


Prédiction: attention!

Application aux données Kyoto (VI)

Si l'individu prédit ne suit pas le même modèle que les autres...

```
ggplot(Kyoto2,aes(x=Population,y=Emissions,colour=status,shape=status)) +
    geom_point() + stat_smooth(method=lm,colour="blue",fullrange=TRUE) +
        annotate("point", 291049, 6900, colour="red", size=10, shape=10)
```



Plan

Décomposition de la variance

Théorème fondamental

$$\underbrace{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}_{SCT} = \underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}_{SCR} + \underbrace{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}_{SCM}$$

Vocabulaire

- SCT = Somme des carrés totale
  $\rightsquigarrow$ variabilité totale à expliquer
- SCM = Somme des carrés due au modèle
  $\rightsquigarrow$ variabilité expliquée par le modèle
- SCR = Somme des carrés résiduelle
  $\rightsquigarrow$ variabilité non expliquée par le modèle

Décomposition de la variance
Interprétation

Théorème fondamental (Pythagore)

Avec $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ et $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\top$, on

$$SCT = SCR + SCM$$
$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|_2^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2 + \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|_2^2$$

Et ainsi

$$(\mathbf{Y} - \hat{\mathbf{Y}}) = \hat{\boldsymbol{\varepsilon}} \perp\!\!\!\perp (\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) \Leftrightarrow SCR \perp\!\!\!\perp SCM,$$

- la variabilité expliquée par le modèle est **indépendente** de la résiduelle.
- géométriquement, $\hat{\mathbf{Y}}$ est la **projection orthogonale** $\mathbf{Y}$ sur le sous espace de $\mathbb{R}^n$ engendré par $\mathbf{x}$.

Décomposition de la variance
Interprétation

Théorème fondamental (Pythagore)

Avec $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ et $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\top$, on

$$SCT = SCR + SCM$$
$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|_2^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2 + \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|_2^2$$

Et ainsi

$$(\mathbf{Y} - \hat{\mathbf{Y}}) = \hat{\boldsymbol{\varepsilon}} \perp\!\!\!\perp (\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) \Leftrightarrow SCR \perp\!\!\!\perp SCM,$$

- la variabilité expliquée par le modèle est **indépendente** de la résiduelle.
- géométriquement, $\hat{\mathbf{Y}}$ est la **projection orthogonale** $\mathbf{Y}$ sur le sous espace de $\mathbb{R}^n$ engendré par $\mathbf{x}$.

Coefficient d'ajustement
Définition

Coefficient d'ajustement

Le coefficient d'ajustement (ou de détermination) est défini par :

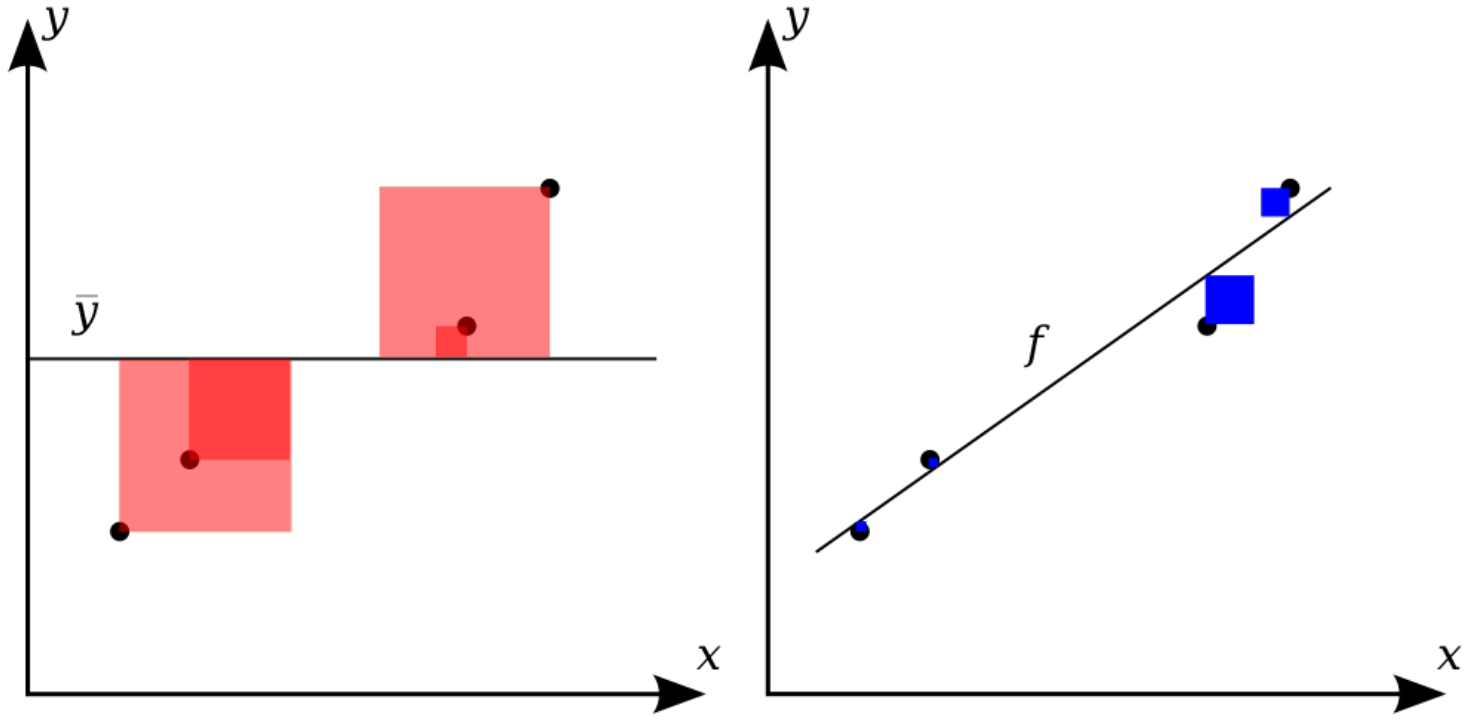
$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT}$$

Remarque

Le coefficient d'ajustement peut être interprété comme le pourcentage de variance expliquée par le modèle

Coefficient d'ajustement

Interprétation pour la régression simple

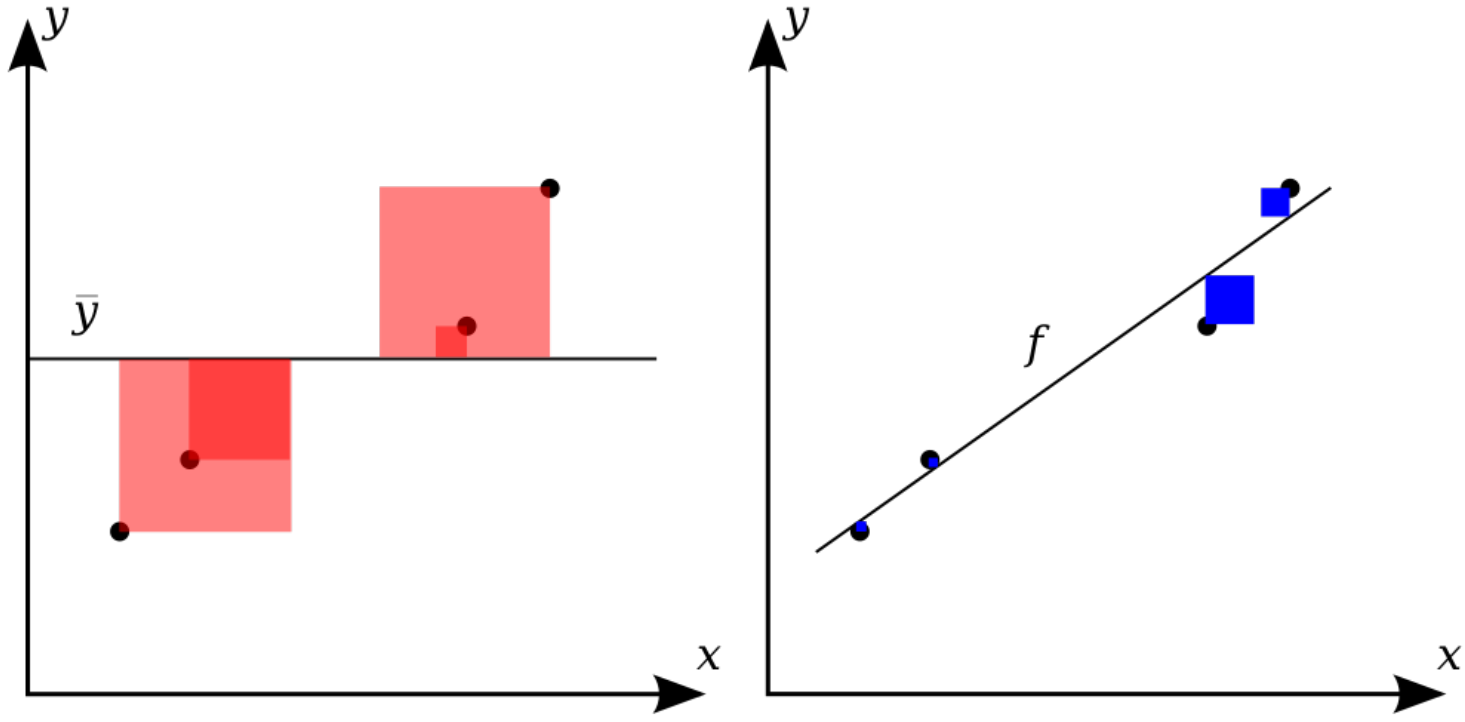


Modèle avec juste la constante

$$\arg\min_{\beta_0} \sum_{i \in \mathcal{D}} (y_i - \beta_0)^2 = \bar{\mathbf{y}}.$$

Coefficient d'ajustement

Interprétation pour la régression simple

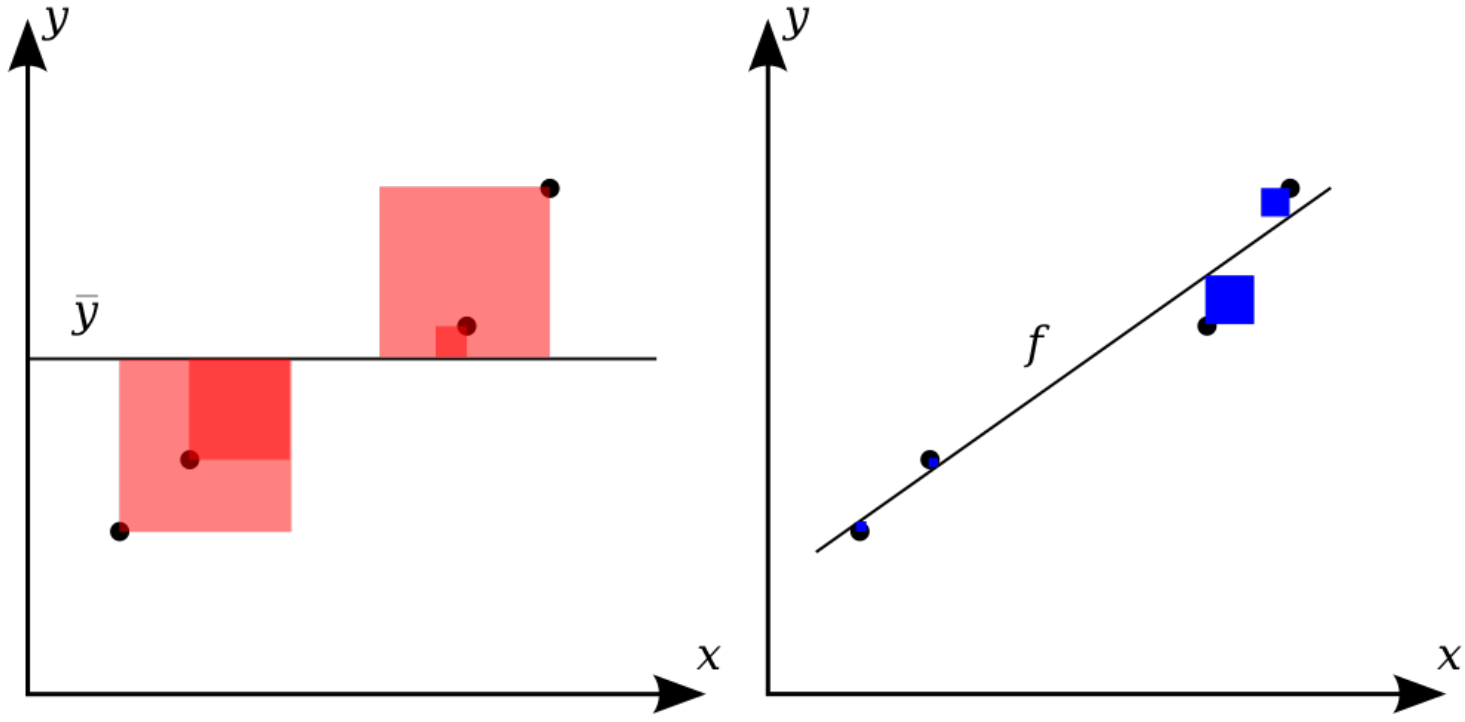


Modèle avec la constante et la pente

$$\arg\min_{\beta_0,\beta_1} \sum_{i\in\mathcal{D}} (y_i - \underbrace{\beta_0 - \beta_1 x_{i1}}_{f_i})^2.$$

Coefficient d'ajustement

Interprétation pour la régression simple



Coefficient d'ajustement

$$R^2 = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{\mathbf{y}})^2} = 1 - \frac{SCR}{SCT}$$

Test du modèle (I)

Hypothèse testée: nullité de β_1 (la pente)

$$\begin{cases} \mathcal{M}_0: & \text{modèle le plus simple} \\ \mathcal{M}_1: & \text{modèle le plus complexe} \end{cases} \Leftrightarrow \begin{cases} \mathcal{M}_0: & Y_i = \beta_0 + \varepsilon_i \\ \mathcal{M}_1: & Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \end{cases}$$

Loi des sommes de carrés sous H_0

- $SCR = (n-2)S^{\star 2} \sim \sigma^2 \chi^2_{n-2}$
- Sous l'hypothèse $\{H_0 : \beta_1 = 0\}$: $SCT \underset{H_0}{\sim} \sigma^2 \chi^2_{n-1}$
- Sous l'hypothèse $\{H_0 : \beta_1 = 0\}$: $SCM \underset{H_0}{\sim} \sigma^2 \chi^2_1$

De plus, $SCR \perp\!\!\!\perp SCM$

Test du modèle (I)

Hypothèse testée: nullité de β_1 (la pente)

$$\begin{cases} \mathcal{M}_0 : & \text{modèle le plus simple} \\ \mathcal{M}_1 : & \text{modèle le plus complexe} \end{cases} \Leftrightarrow \begin{cases} \mathcal{M}_0 : & Y_i = \beta_0 + \varepsilon_i \\ \mathcal{M}_1 : & Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \end{cases}$$

Loi des sommes de carrés sous H_0

- $SCR = (n-2)S^{\star 2} \sim \sigma^2 \chi^2_{n-2}$
- Sous l'hypothèse $\{H_0 : \beta_1 = 0\}$: $SCT \underset{H_0}{\sim} \sigma^2 \chi^2_{n-1}$
- Sous l'hypothèse $\{H_0 : \beta_1 = 0\}$: $SCM \underset{H_0}{\sim} \sigma^2 \chi^2_1$

De plus, $SCR \perp\!\!\!\perp SCM$

Test du modèle (II)

Statistique de test: Fisher

Intuitivement, on rejette lorsque la valeur observée de la statistique F est "grande":

$$F = \frac{SCM/1}{SCR/(n-2)} \underset{H_0}{\sim} \mathcal{F}_{1,n-2}$$

Preuve...

Règle de décision et $p-$valeur

On rejette H_0 si $F \geq f_{1,n-2;1-\alpha}$ $p-\text{val} = \P_{H_0}\left(\mathcal{F}_{1,n-2} \geq f(\text{obs})\right)$

Test du modèle (II)

Statistique de test: Fisher

Intuitivement, on rejette lorsque la valeur observée de la statistique F est "grande":

$$F = \frac{SCM/1}{SCR/(n-2)} \underset{H_0}{\sim} \mathcal{F}_{1,n-2}$$

Preuve...

Règle de décision et $p-$valeur

On rejette H_0 si $F \geq f_{1,n-2;1-\alpha}$ $\quad p-\text{val} = \P_{H_0}\left(\mathcal{F}_{1,n-2} \geq f(\text{obs})\right)$

Analyse de la variance

Tableau de synthèse

Source	Degrés de liberté	Sommes des carrés	Carrés moyens	F
Modèle	1	SCM	SCM	$F = \frac{(n-2)SCM}{SCR}$
Résiduelle	$n-2$	SCR	$\frac{SCR}{(n-2)}$	
Total	$n-1$	SCT		

Analyse de la variance du modèle de régression simple I

Application aux données Kyoto

```
M0 <- lm(Emissions~1,Kyoto)
M1 <- lm(Emissions~Population,Kyoto)
anova(M0,M1)

## Analysis of Variance Table
##
## Model 1: Emissions ~ 1
## Model 2: Emissions ~ Population
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     14 1215852
## 2     13   34480  1   1181371 445.41 1.925e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analyse de la variance du modèle de régression simple II

Application aux données Kyoto

```
anova(M1)

## Analysis of Variance Table
##
## Response: Emissions
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Population   1 1181371 1181371  445.41 1.925e-11 ***
## Residuals   13   34480    2652
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analyse de la variance du modèle de régression simple III

Application aux données Kyoto

```
summary(M1)


##
## Call:
## lm(formula = Emissions ~ Population, data = Kyoto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.983 -33.297   3.004  22.605 120.173
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.915e+00  1.861e+01    0.21    0.837
## Population  1.082e-02  5.128e-04   21.11 1.93e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.5 on 13 degrees of freedom
## Multiple R-squared:  0.9716,Adjusted R-squared:  0.9695
## F-statistic: 445.4 on 1 and 13 DF,  p-value: 1.925e-11
```

Plan

Rappels des hypothèses

Liées aux **variables résiduelles**

1. Résidus centré: $\mathbb{E}(Y) = \beta_0 + \beta_1 x$, soit $\mathbb{E}(\varepsilon_i) = 0$
2. Résidus homoscédastiques : $\mathbb{V}(\varepsilon_i) = \sigma^2$ pour tout i,
3. Résidus indépendents, $\mathbb{V}(\varepsilon_i) \perp\!\!\!\perp \mathbb{V}(\varepsilon_j)$, pour tout $i \neq j$
4. Résidus gaussiens: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Analyse des résidus

Diagnostic et solutions envisageables

À défaut de diposer de ε_i, on diagnostique $\hat{\varepsilon}_i$

1. analyse du **graphe des résidus**
   - recherche d'une tendance, hétéroscédasticité, perte de centrage
   - transformation des Y_i et/ou des x_i
2. Test d'indépendance (Durbin-Watson)
3. Test de normalité (Shapiro, Kolmogorov, χ^2)

Tolérance

- écart à la loi normale: **assez peu d'impact**, d'autant moins que la distributions des résidus est symétrique
- indépendance des résidus: **importante** pour les résultas d'estimation et de test.

Analyse des résidus

Diagnostic et solutions envisageables

À défaut de diposer de ε_i, on diagnostique $\hat{\varepsilon}_i$

1. analyse du **graphe des résidus**
   - recherche d'une tendance, hétéroscédasticité, perte de centrage
   - transformation des Y_i et/ou des x_i
2. Test d'indépendance (Durbin-Watson)
3. Test de normalité (Shapiro, Kolmogorov, χ^2)

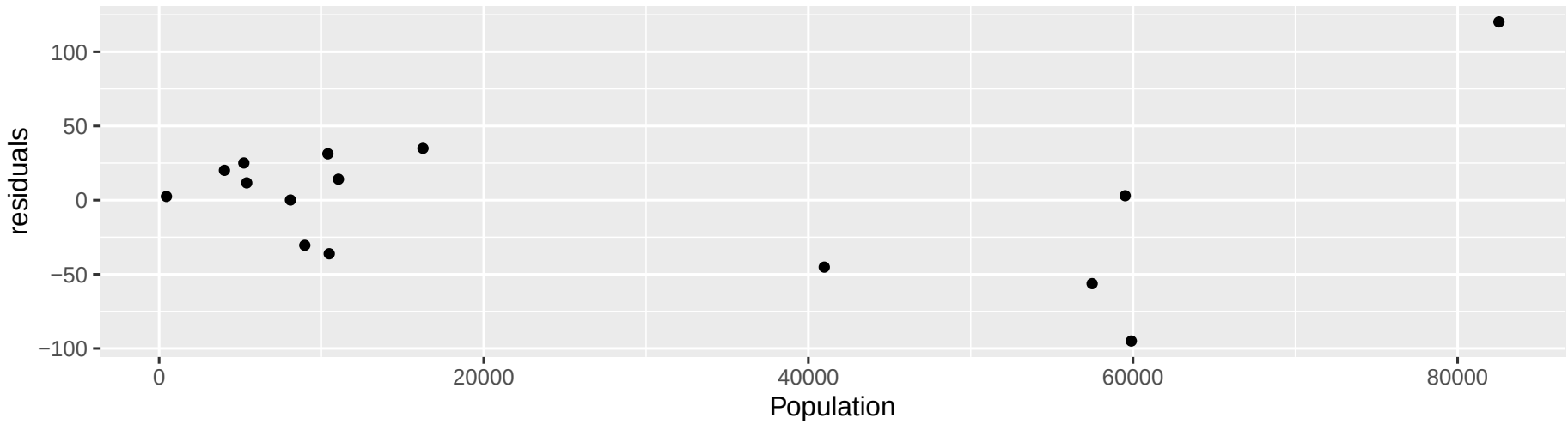
Tolérance

- écart à la loi normale: **assez peu d'impact**, d'autant moins que la distributions des résidus est symétrique
- indépendance des résidus: **importante** pour les résultas d'estimation et de test.

Diagnostic
Application aux données Kyoto (I)

Homoscédasticité ? résidus centrés ? hum...

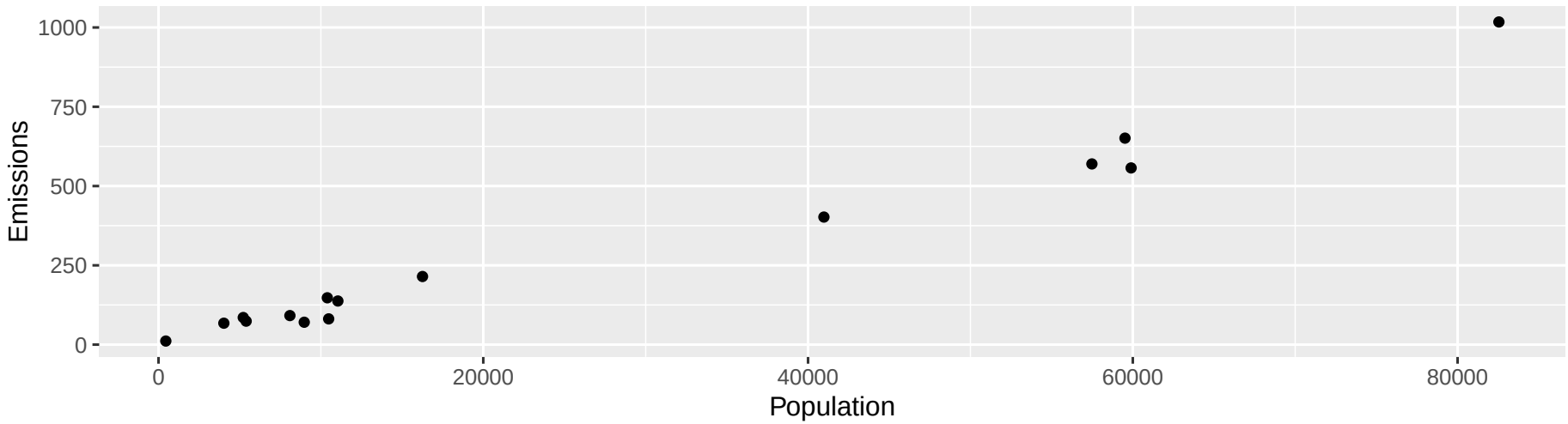
```
M1 <- lm(Emissions~Population,Kyoto)
Kyoto <- cbind(Kyoto, residuals=residuals(M1))
ggplot(Kyoto, aes(Population,residuals)) + geom_point()
```



Diagnostic: données originales

Application aux données Kyoto (II)

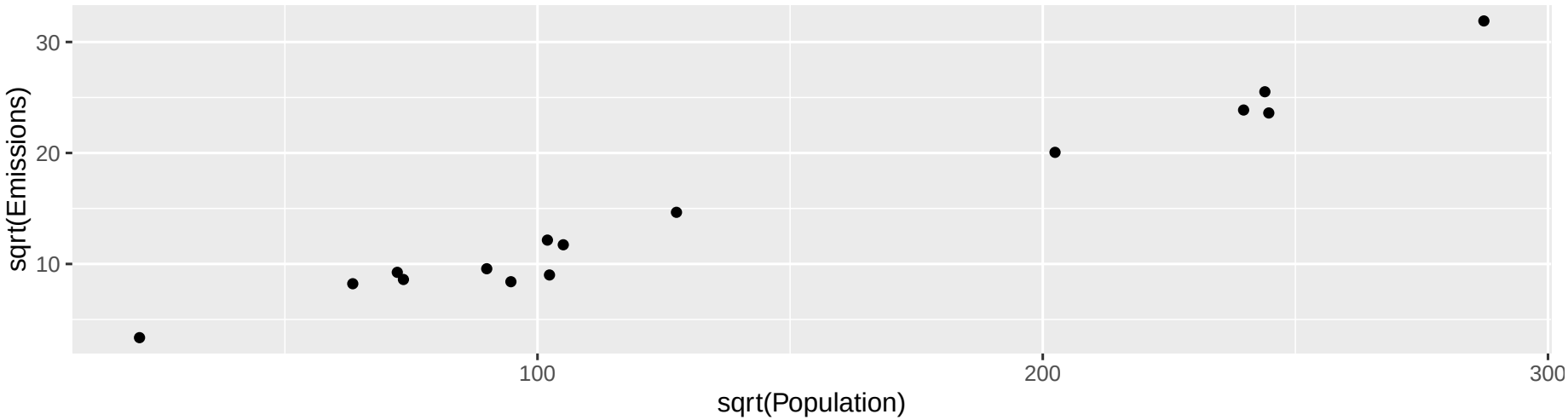
```
ggplot(Kyoto, aes(Population,Emissions)) + geom_point()
```



Diagnostic: transformation racine carrée

Application aux données Kyoto (III)

```
ggplot(Kyoto, aes(sqrt(Population),sqrt(Emissions))) + geom_point()
```



Diagnostic: transformation racine carrée

Application aux données Kyoto (IV)

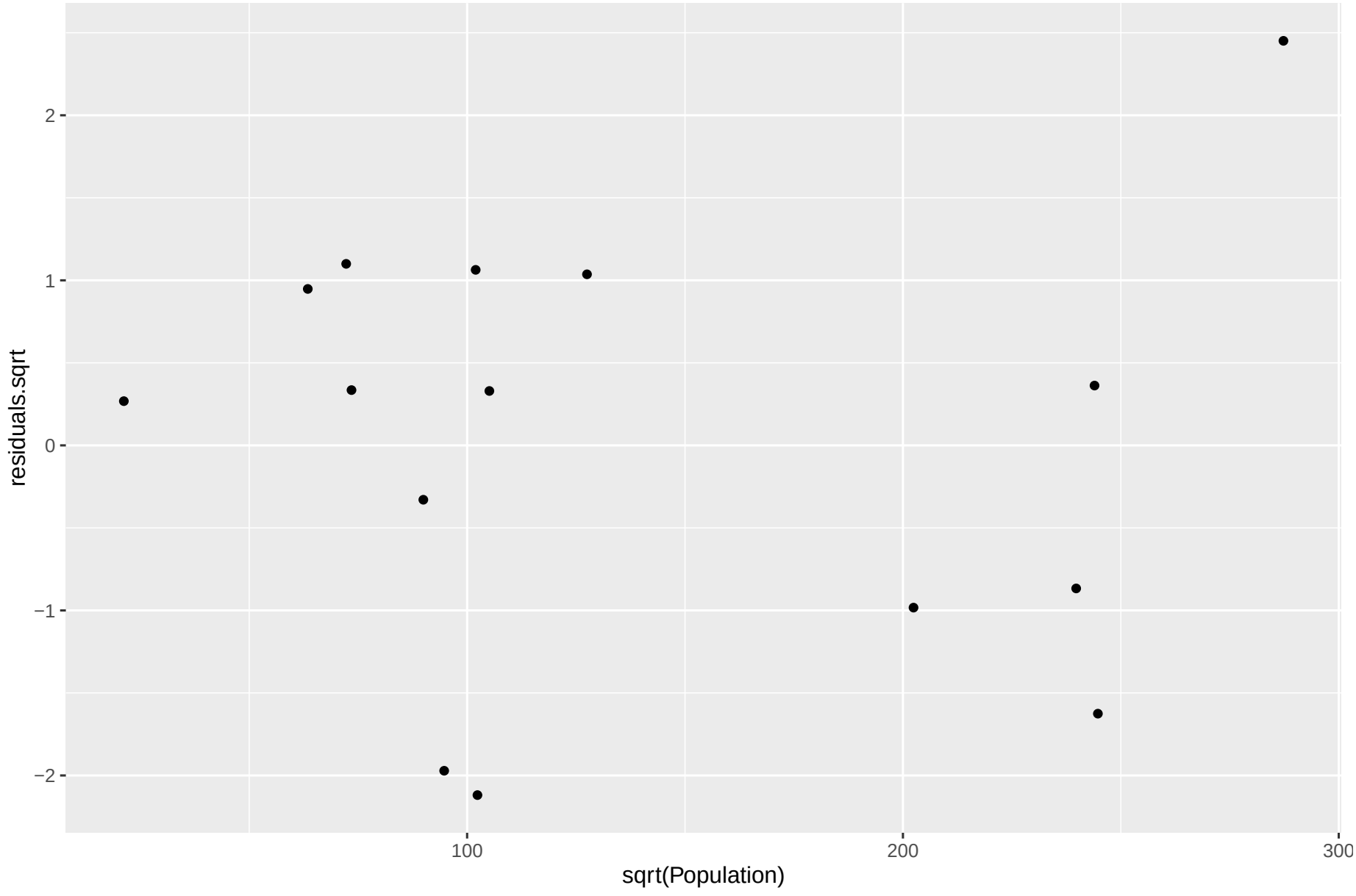
```
M1.sqrt <- lm(sqrt(Emissions)~sqrt(Population),Kyoto)
Kyoto <- cbind(Kyoto, residuals.sqrt=residuals(M1.sqrt))
summary(M1.sqrt)
```

```
##
## Call:
## lm(formula = sqrt(Emissions) ~ sqrt(Population), data = Kyoto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1188 -0.9248  0.3296  0.9923  2.4512
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.99048    0.69618   1.423    0.178
## sqrt(Population)  0.09905    0.00437  22.667 7.79e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.347 on 13 degrees of freedom
## Multiple R-squared:  0.9753,Adjusted R-squared:  0.9734
## F-statistic: 513.8 on 1 and 13 DF,  p-value: 7.786e-12
```

```
ggplot(Kyoto, aes(sqrt(Population),residuals.sqrt)) + geom_point()
```

Diagnostic: transformation racine carrée

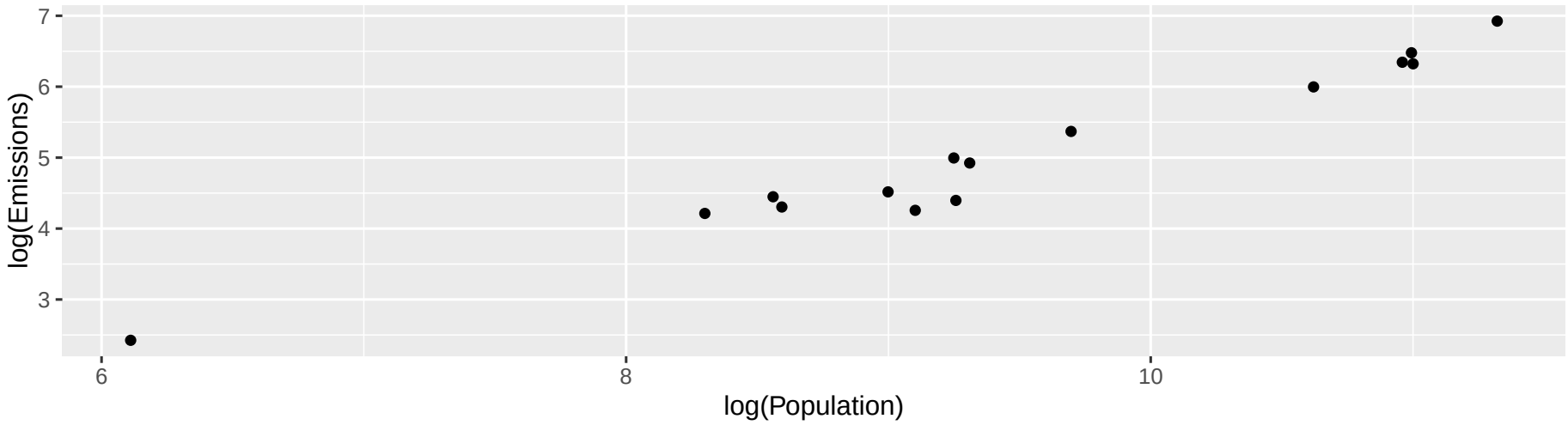
Application aux données Kyoto (V)



Diagnostic: transformation logarithmique

Application aux données Kyoto (VI)

```
ggplot(Kyoto, aes(log(Population),log(Emissions))) + geom_point()
```



Diagnostic: transformation logarithmique

Application aux données Kyoto (VII)

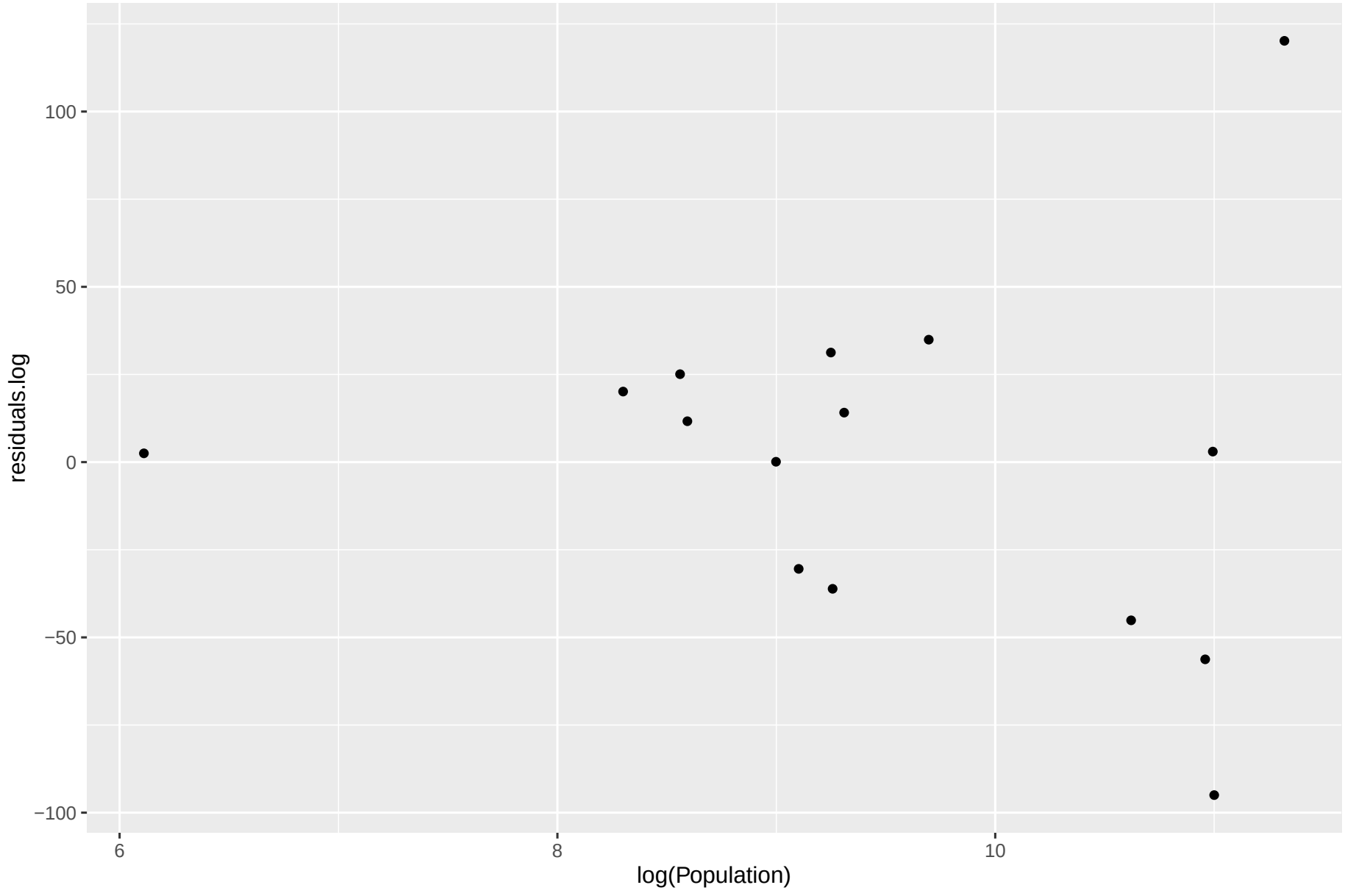
```
M1.log <- lm(log(Emissions)~log(Population),Kyoto)
Kyoto <- cbind(Kyoto, residuals.log=residuals(M1))
summary(M1.log)
```

```
##
## Call:
## lm(formula = log(Emissions) ~ log(Population), data = Kyoto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49102 -0.03698  0.02216  0.13590  0.29505
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.97210    0.43865  -6.776 1.31e-05 ***
## log(Population)  0.84816    0.04586  18.493 1.02e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2357 on 13 degrees of freedom
## Multiple R-squared:  0.9634,Adjusted R-squared:  0.9606
## F-statistic:   342 on 1 and 13 DF,  p-value: 1.018e-10
```

```
ggplot(Kyoto, aes(log(Population),residuals.log)) + geom_point()
```

Diagnostic: transformation logarithmique

Application aux données Kyoto (VII)



Diagnostic: test de normalité

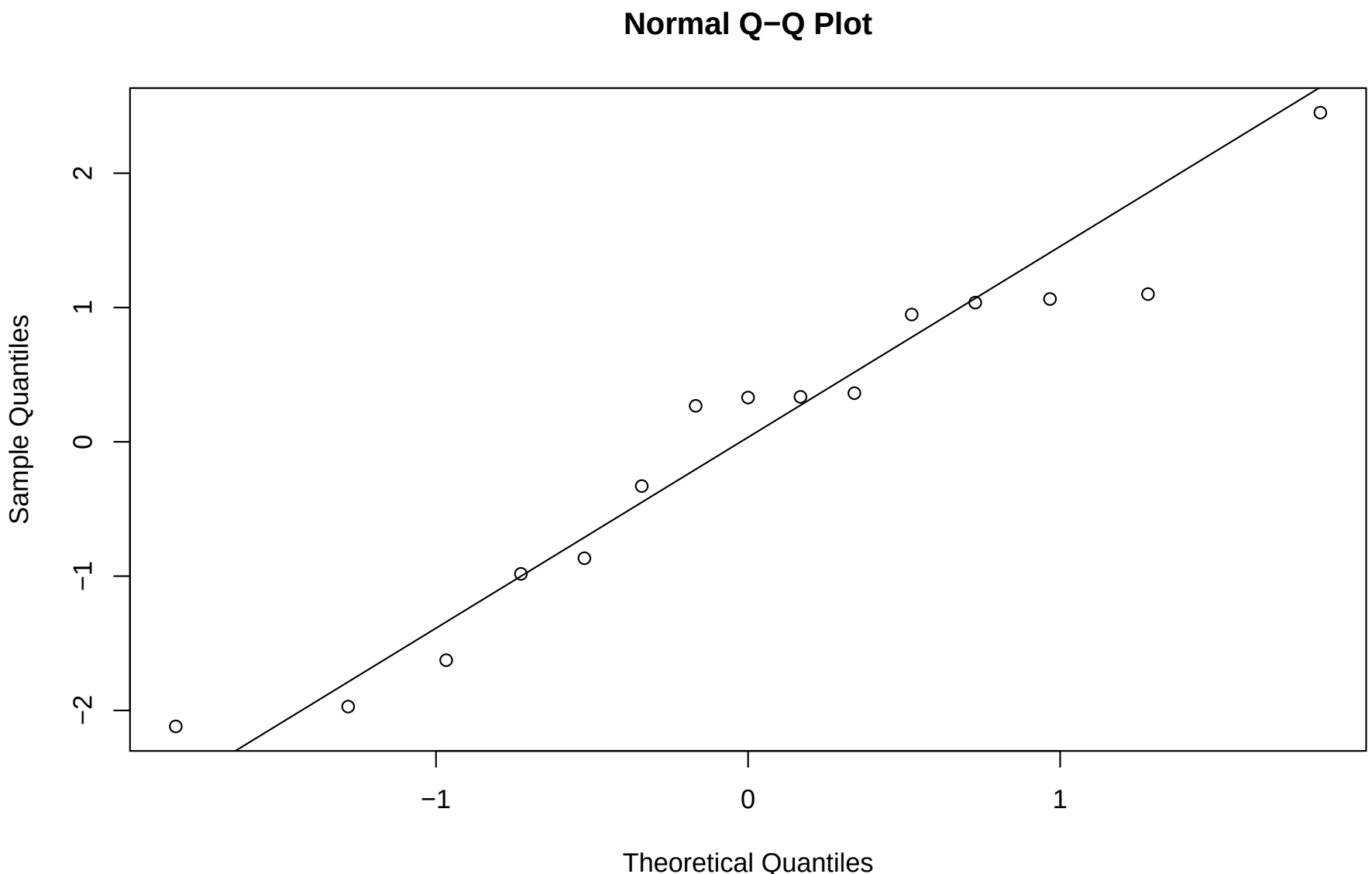
Application aux données Kyoto (VIII)

```
shapiro.test(residuals(M1.sqrt))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(M1.sqrt)
## W = 0.94777, p-value = 0.4901
```

Diagnostic: test de normalité

Application aux données Kyoto (IX)

```
qqnorm(residuals(M1.sqrt)); qqline(residuals(M1.sqrt))
```



Diagnostic: test d'indépendance

Application aux données Kyoto (X)

Tests d'indépendance des résidus

```
library(car)
durbinWatsonTest(M1.sqrt)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.3678903       2.31636   0.474
##  Alternative hypothesis: rho != 0
```

Modèle final

Application aux données Kyoto (XI)