Regularization Methods for Linear Regression
Simple linear regression

M1 Math et Interactions – UEVE/ENSIIE

Autumn semester 2016

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

école nationale supérieure d'informatique pour l'industrie et l'entreprise

Outline

Model

Estimation

Residuals and Prediction

Analysis of Variance

Diagnostic

Simple Regression
General purpose

Idea

Explain the variations of a **quantitative** variable Y based on observations of a **quantitative** variable x

Examples

- Blood pressure = f(age)
- Wheat yield = f(quantity of fertilizer)
- Ozone concentration = f(temperature)
- Treatment effect = f(dose)
- pesticide rate = f(age of the fish) (example pursued during practicals)

Simple Regression

Being specific about the Variables at play

Vocabulary

The roles of Y and x are **not symetric**:

- Y is the **response** variable, or **output**
- x is the **explicative**, **input**, **covariate**, or **predictor**

Remarks

- Y is a random variable
- the covariate may be random (X) or controlled (x)
  - we consider it as fixed here (hence the notation x)
- **careful** note the difference between upper and lower case

Simple linear regression
Model

We asumme that the true relationship between Y and x is linear:

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

- β_0 is the **intercept** (**constant term**)
- β_1 is the **slope** (**pente**)
- ε is the error term or **noise**
  - describe a measurement uncertainty,
  - individual variability,
  - some factors unexplained by the model

$\rightsquigarrow$ In practice , β_0, β_1 and ε are unknown

Simple linear regression
Statistical Hypotheses

$\rightsquigarrow$ Mandatory for performing statistical inference (tests, ...) !

Hypotheses on the error term

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

Collecting data / sampling

Let $\{(Y_i, x_i)\}_{i=1}^n$ be a n-sample. We have

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

with $\{\varepsilon_i\}_{i=1}^n$ independent, identically distributed.

Simple linear regression
What Linearity?

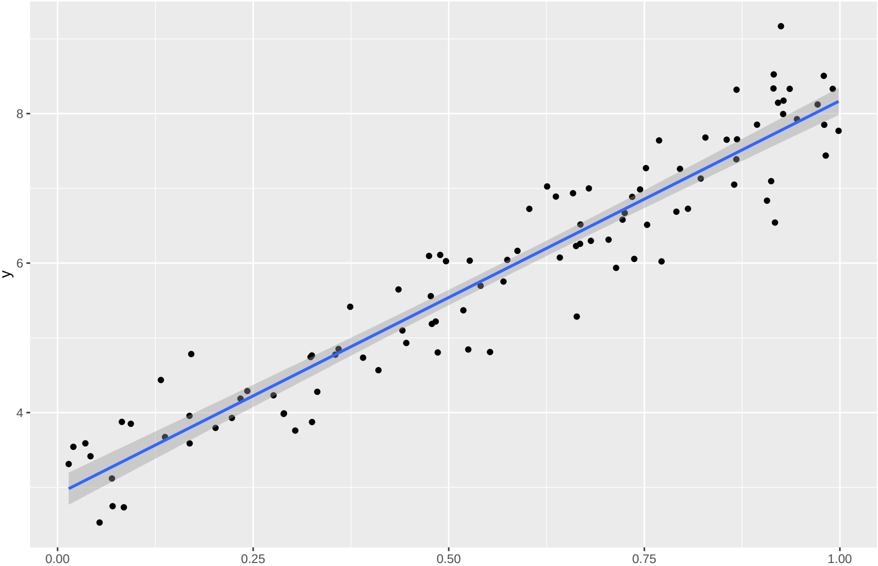
The model is **linear regarding the parameters** (not in x)

```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)
```

Simple linear regression

Linearity (model 1)

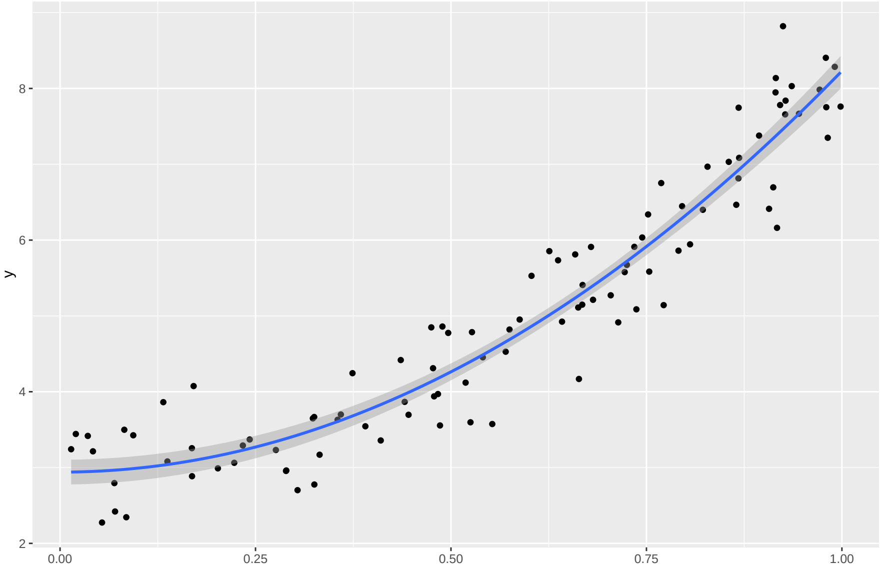
```
ggplot(d1,aes(x,y)) + geom_point() + stat_smooth(method="lm", formula=y~x)
```



Simple linear regression

Linearity (model 2)

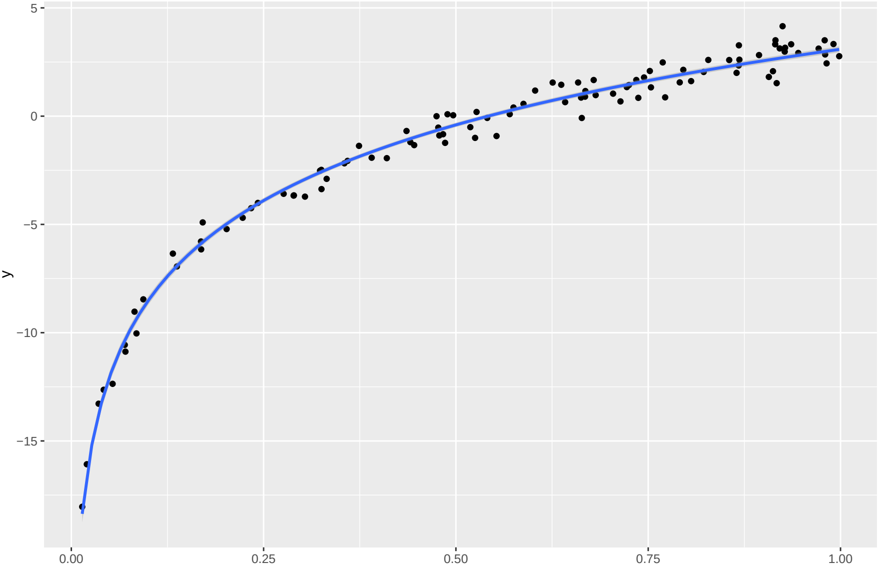
```
ggplot(d2,aes(x,y)) + geom_point() + stat_smooth(method="lm", formula=y~I(x^2))
```



Simple linear regression

Linearity (model 3)

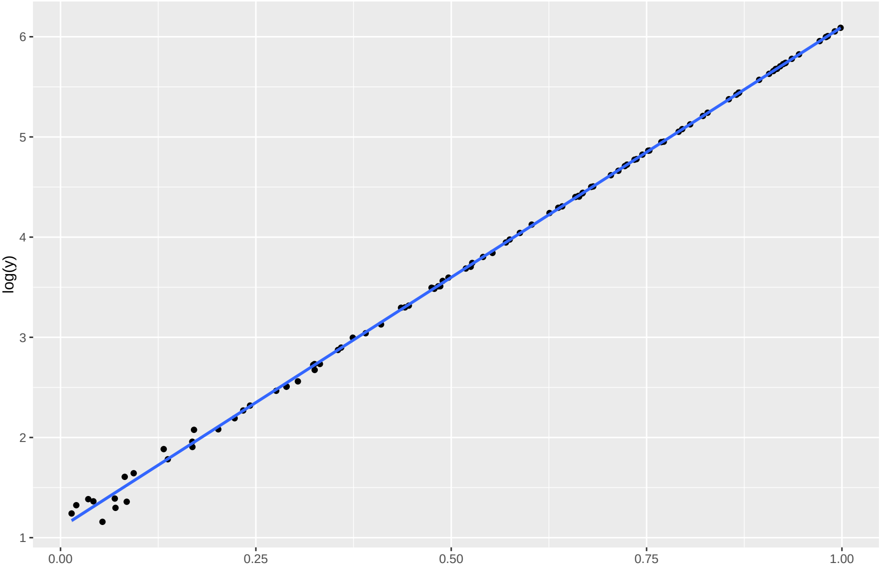
```
ggplot(d3,aes(x,y)) + geom_point() + stat_smooth(method="lm", formula=y~I(log(x)))
```



Simple linear regression

Linearity (model 4)

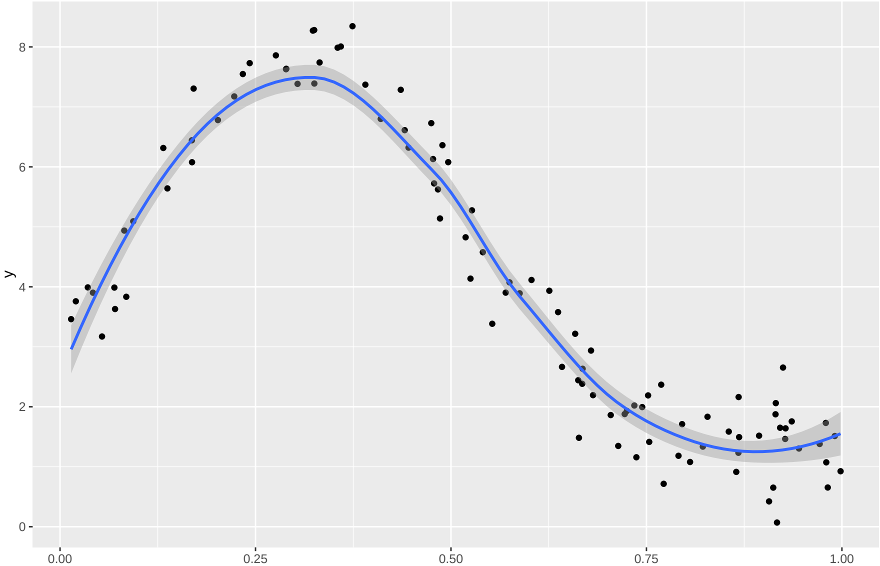
```
ggplot(d4,aes(x,log(y))) + geom_point() + stat_smooth(method="lm", formula=y ~ x)
```



Simple linear regression

Linearity (model 5)

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



Simple linear regression
Summary

Statistical goals

1. Estimating the parameters β_0, β_1 et σ^2

2. Testing the nullity of β_0, β_1, i.e. the effect of the covariate

3. Predicting Y for a new observation x_0

4. Testing the relevance of the model

Recurrent Example

Kyoto data set (I)

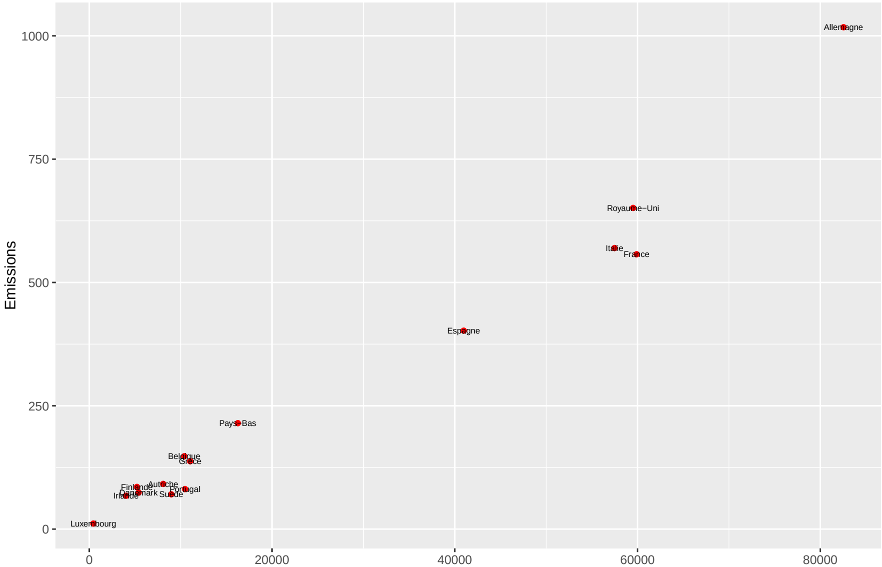
```
#### Infos
# European contries
# Population: Thousands
# Emissions:  Mil. tons CO2
# US population for prediction: 291049
Kyoto <- read.table(file='Emissions.txt',header=F)
colnames(Kyoto) <- c("Country","Population","Emissions")

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
```

Recurrent Example

Kyoto data set (II)

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

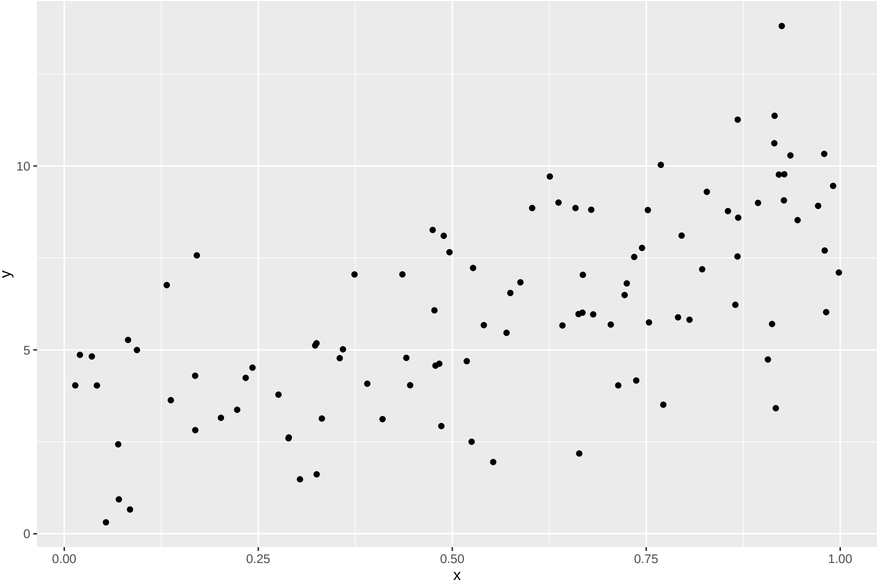


Outline

Outline

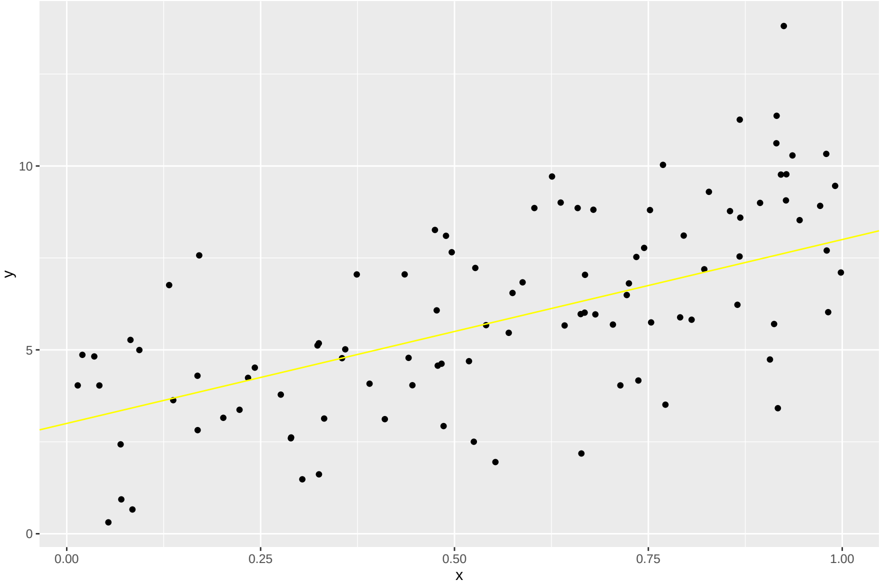
Ordinary Least Squares
Intuition

Suppose we draw some points in the **sample**.



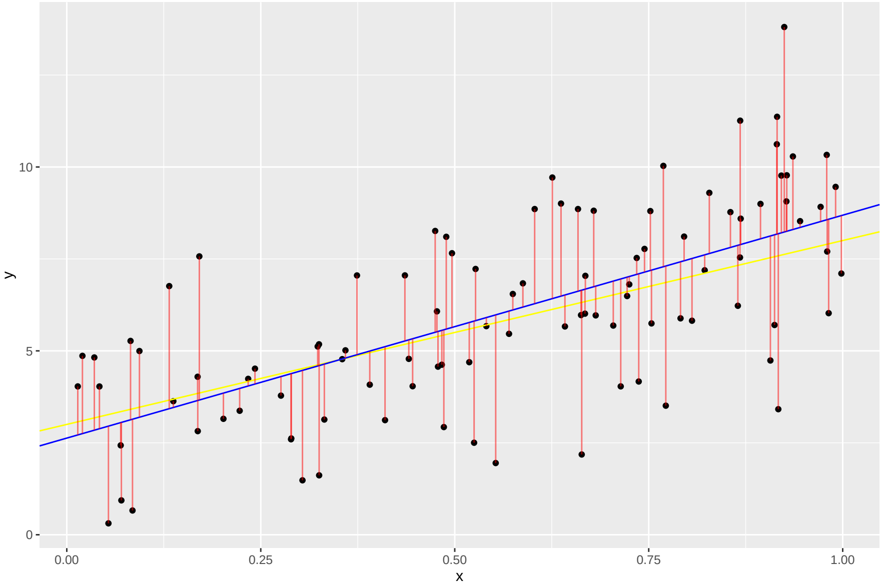
Ordinary Least Squares
Idea

The **“true”** line is the closest to the points of the whole **population**.



Ordinary Least Squares
Idea

We look for the **closest** line to the points of the **sample**



Ordinary Least Squares
Criterion

Formalism

- distance to a single point : $(y_i - x_i\beta_1 - \beta_0)^2$
- distance to the whole sample: $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$

$\rightsquigarrow$ Best line: intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ such that $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$ is minimum, among all possible values of β_0, β_1.

OLS estimator

The values estimated by OLS (the estimates) for β_0 et β_1 verify

Ordinary Least Squares
Criterion

Formalism

- distance to a single point : $(y_i - x_i\beta_1 - \beta_0)^2$
- distance to the whole sample: $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$

$\rightsquigarrow$ Best line: intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ such that $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$ is minimum, among all possible values of β_0, β_1.

OLS estimator

The values estimated by OLS (the estimates) for β_0 et β_1 verify

Formalism

- distance to a single point : $(y_i - x_i\beta_1 - \beta_0)^2$
- distance to the whole sample: $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$

$\rightsquigarrow$ Best line: intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ such that $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$ is minimum, among all possible values of β_0, β_1.

OLS estimator

The values estimated by OLS (the estimates) for β_0 et β_1 verify

Ordinary Least Squares

Criterion

Formalism

- distance to a single point : $(y_i - x_i\beta_1 - \beta_0)^2$
- distance to the whole sample: $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$

$\rightsquigarrow$ Best line: intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ such that $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$ is minimum, among all possible values of β_0, β_1.

OLS estimator

The values estimated by OLS (the estimates) for β_0 et β_1 verify

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

Formalism

- distance to a single point : $(y_i - x_i\beta_1 - \beta_0)^2$
- distance to the whole sample: $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$

$\rightsquigarrow$ Best line: intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ such that $\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$ is minimum, among all possible values of β_0, β_1.

OLS estimator

The values estimated by OLS (the estimates) for β_0 et β_1 verify

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

Ordinary Least Squares
Estimators

Theorem

The OLS estimators have the following expressions

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

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

Proof: by zeroing the derivative of the objective function, which is convex.

Remarques

- does not depend on the Gaussian assumption of the noise
- do not misunderstand estimator/estimate (r.v/observation)
- we do no say a thing about $\sigma^2 \dots$

Ordinary Least Squares
Estimators

Theorem

The OLS estimators have the following expressions

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

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

Proof: by zeroing the derivative of the objective function, which is convex.

Remarques

- does not depend on the Gaussian assumption of the noise
- do not misunderstand estimator/estimate (r.v/observation)
- we do no say a thing about σ^2...

Ordinary Least Squares

Application to the Kyoto data set

```
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
```

Outline

Maximum likelihood

criterion

Formalism

- likelihood of a single point: $L(y_i) = f(y_i)$
- likelihood of the whole sample: $L(y_1, \dots, y_n) = \prod_{i=1}^{n} f(y_i)$
- log-likelihood : $\log L(y_1, \dots, y_n) = \sum_{i=1}^{n} \log f(y_i)$

ML Estimators

The values estimated by ML for β_0, β_1 et σ verify

Maximum likelihood
criterion

Formalism

- likelihood of a single point: $L(y_i) = f(y_i)$
- likelihood of the whole sample: $L(y_1, \dots, y_n) = \prod_{i=1}^n f(y_i)$
- log-likelihood : $\log L(y_1, \dots, y_n) = \sum_{i=1}^n \log f(y_i)$

ML Estimators

The values estimated by ML for β_0, β_1 et σ verify

Maximum likelihood
criterion

Formalism

- likelihood of a single point: $L(y_i) = f(y_i)$
- likelihood of the whole sample: $L(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i)$
- log-likelihood : $\log L(y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i)$

$\rightsquigarrow$ Best estimators: $(\beta_0, \beta_1, \sigma)$ maximizing L or $\log L$, measuring **how likely** are the current values of the parameters regarding the data (fixed)

ML Estimators

The values estimated by ML for β_0, β_1 et σ verify

Maximum likelihood
criterion

Formalism

- likelihood of a single point: $L(y_i) = f(y_i)$
- likelihood of the whole sample: $L(y_1, \dots, y_n) = \prod_{i=1}^{n} f(y_i)$
- log-likelihood : $\log L(y_1, \dots, y_n) = \sum_{i=1}^{n} \log f(y_i)$

ML Estimators

The values estimated by ML for β_0, β_1 et σ verify

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

Maximum likelihood

criterion

Formalism

- likelihood of a single point: $L(y_i) = f(y_i)$
- likelihood of the whole sample: $L(y_1, \dots, y_n) = \prod_{i=1}^n f(y_i)$
- log-likelihood : $\log L(y_1, \dots, y_n) = \sum_{i=1}^n \log f(y_i)$

ML Estimators

The values estimated by ML for β_0, β_1 et σ verify

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

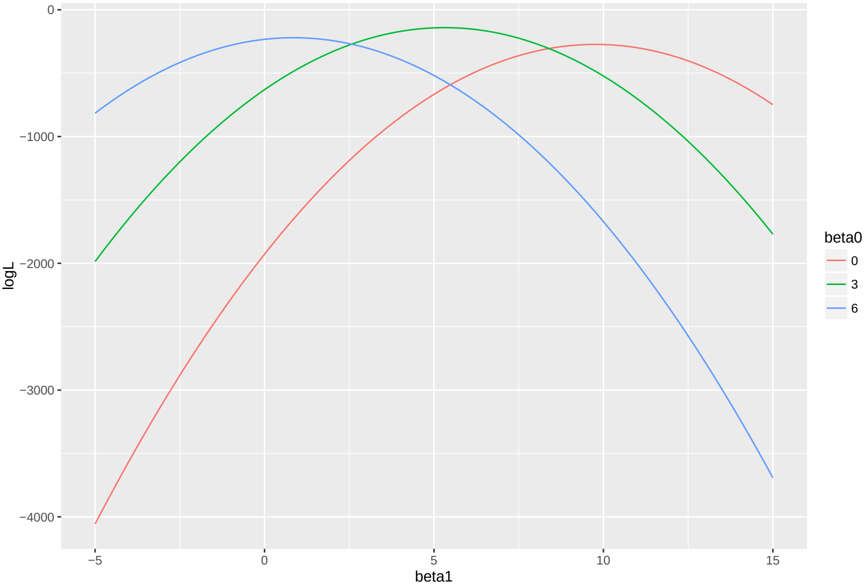
Maximul likelihood I

Intuition

```r
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)
```

Maximul likelihood II

Intuition



Maximum likelihood
Estimation

Theorem

The MLE have the following expression:

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

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

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

Proof:

By zeroing the derivatives of the objective function, which is concave.

Maximum likelihood

Practical estimation of the residual variance

We do not know β_1 nor β_0 ! If we replace them by their estimators,

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

we get an estimator which is *biased*. In practive, we use

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

Remark

The "-2" came from the 2 degrees of freedom lost by estimating β_0, β_1.

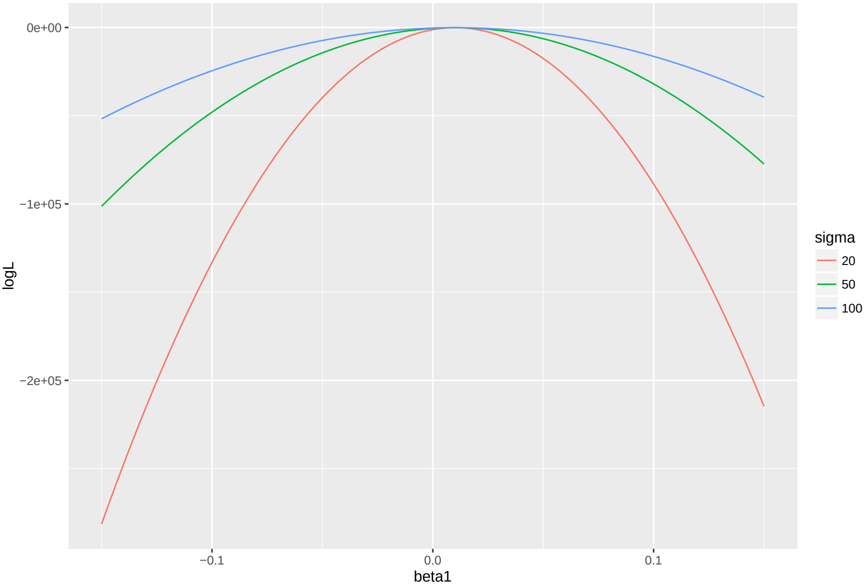
Maximum likelihood I

Application to the Kyoto data set

```
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 likelihood II

Application to the Kyoto data set



Outline

Parameters Estimation

Properties of β_0 et β_1 (I)

General Case

$\hat{B}_0$ and $\hat{B}_1$ are unbiased estimators of β_0 and β_1, with variances given by

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

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

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

Gaussian Case

If the noise is Gaussian, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, then

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

Parameters Estimation

Properties of β_0 et β_1 (I)

General Case

$\hat{B}_0$ and $\hat{B}_1$ are unbiased estimators of β_0 and β_1, with variances given by

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

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

and covariance $\operatorname{cov}(\hat{B}_0, \hat{B}_1) = -\frac{\sigma^2 \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}$

Gaussian Case

If the noise is Gaussian, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, then

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

Parameters Estimation

Properties of B_0 et B_1 (II)

Gauss-Markov Theorem

- Gaussian case $\hat{B}_0$ and $\hat{B}_1$ are the best unbiased estimators (i.e. with minimal variance).
- General case $\hat{B}_0$ and $\hat{B}_1$ are the best **linear** unbiased estimators.

Theorem

- The residual variance σ^2 is estimated with no bias by:

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

- With Gaussian noise, we moreover have

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

Parameters Estimation

Properties of B_0 et B_1 (II)

Gauss-Markov Theorem

- Gaussian case $\hat{B}_0$ and $\hat{B}_1$ are the best unbiased estimators (i.e. with minimal variance).
- General case $\hat{B}_0$ and $\hat{B}_1$ are the best **linear** unbiased estimators.

Theorem

- The residual variance σ^2 is estimated with no bias by:

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

- With Gaussian noise, we moreover have

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

Outline

Testing the parameters: the slope
With the Gaussian assumption

Testing the nullity of β_1 (the slope)

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

Test Statistic and decision rule

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

$p-$value (degree of significance)

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

Testing the parameters: the slope
With the Gaussian assumption

Testing the nullity of β_1 (the slope)

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

Test Statistic and decision rule

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

$p-$value (degree of significance)

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

Testing the parameters: the slope
With the Gaussian assumption

Testing the nullity of β_1 (the slope)

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

Test Statistic and decision rule

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

$p-$value (degree of significance)

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

Testing the parameters: the intercept
With the Gaussian assumption

Testing the nullity of β_0 (the intercept)

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

Test Statistic and decision rule

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

$p-$value (degree of significance)

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

Testing the parameters: the intercept

With the Gaussian assumption

Testing the nullity of β_0 (the intercept)

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

Test Statistic and decision rule

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

$p-$value (degree of significance)

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

Testing the parameters: the intercept

With the Gaussian assumption

Testing the nullity of β_0 (the intercept)

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

Test Statistic and decision rule

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

$p-$value (degree of significance)

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

Testing the parameters

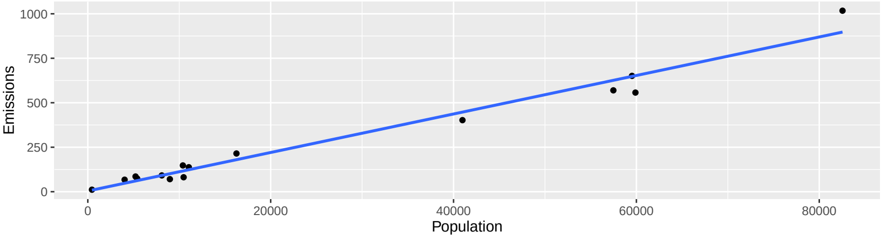
Application to the Kyoto data set (I)

```
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
```

Testing the parameters

Application to the Kyoto data set (II

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



Outline

Prediction, predictor

Problem

The value predicted by the model for the ith individual is

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

but β_0, β_1 and ε_i are unknown.

Idea

The estimators the estimates of β_0 and β_1 let us define

- a **predictor**: $\hat{Y}_i = \hat{B}_0 + \hat{B}_1 x_i$
- a **prediction**: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Estimating the noise: the residuals

Proposition

Let $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ the residual for point i. We have:

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

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

Remarks

- We have $\sum \hat{\varepsilon}_i = 0$
- Contrary to ε_i, the residuals $\hat{\varepsilon}_i$ are not independent
- The more far away x_i from the mean $\bar{\mathbf{x}}$ (the barycenter), the higher the variance of the prediction error

Predicting a new observation

Predicted value

Let x_0 be a new observation. The value predicted by the model is $Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0$. This value can be approximated by

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

Remarks

They is two sources of error with such a prediction:

- Uncertainties due to the estimation of β_0 et β_1
- We do not know the noise ε_0 at point x_0

Prediction: confidence interval

Let x_0 be a new observation and $\hat{Y}_0$ the corresponding prediction.

Proposition (Distribution of $\hat{Y}_0$)

Under the Gaussian assumption and from the joint distribution of (B_0, B_1), *we derive*

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

Remarks

- $\mathbb{V}(\hat{Y}_0) = \mathbb{V}(B_0 + B_1 x) \neq \mathbb{V}(B_0) + \mathbb{V}(B_1 x)$ because $\mathrm{cov}(B_0, B_1) \neq 0$.
- $\mathbb{V}(\hat{Y}_0)$ take into account the error committed estimating $\beta_0 + \beta_1 x$.
- *The more we estimate* $\mathbb{E}(Y_0)$ *from a point* x_0 *which is far away (resp. close to)* $\bar{\mathbf{x}}$, *the higher the variance (resp. smaller).*

Prediction: confidence interval

Let x_0 be a new observation and $\hat{Y}_0$ the corresponding prediction.

Proposition (Distribution of $\hat{Y}_0$)

Under the Gaussian assumption and from the joint distribution of (B_0, B_1), *we derive*

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

Remarks

- $\mathbb{V}(\hat{Y}_0) = \mathbb{V}(B_0 + B_1 x) \neq \mathbb{V}(B_0) + \mathbb{V}(B_1 x)$ because $\operatorname{cov}(B_0, B_1) \neq 0$.
- $\mathbb{V}(\hat{Y}_0)$ take into account the error committed estimating $\beta_0 + \beta_1 x$.
- *The more we estimate* $\mathbb{E}(Y_0)$ *from a point* x_0 *which is far away (resp. close to)* $\bar{\mathbf{x}}$, *the higher the variance (resp. smaller).*

Prediction: prediction interval

For the **prediction** interval, one has to take into account the estimation of incompressible noise, i.e., $\hat{\sigma}^2$.

Prediction Interval

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

$\rightsquigarrow$ for a new observed value, we add the error due to the random draw.

Prediction, residuals

Application to the Kyoto data set (I)

```r
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
```

```r
sum(residuals(model))
```

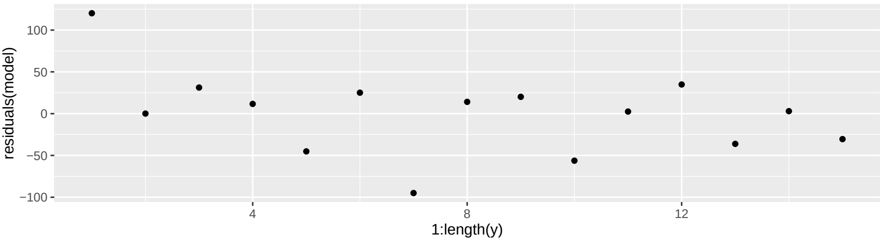
```
## [1] -7.81597e-14
```

```r
## valeurs estimés
head(fitted(model))
```

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

Prediction, residuals
Application to the Kyoto data set (II)

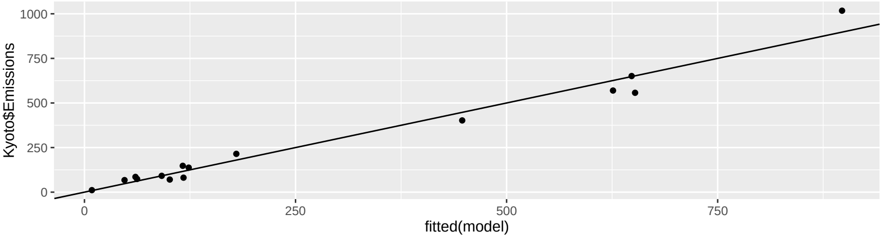
```
qplot(1:length(y),residuals(model), geom='point')
```



Prediction, residuals

Application to the Kyoto data set (III)

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



Prediction, residuals

Application to the Kyoto data set (IV)

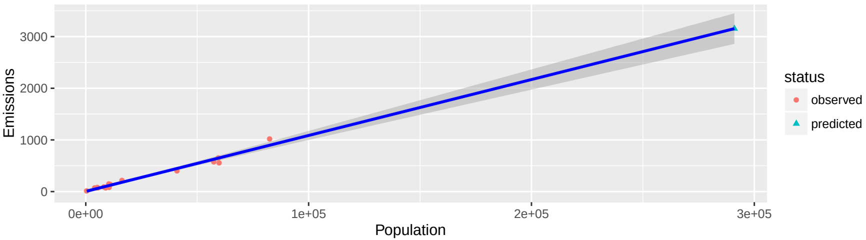
US Population for the prediction: 291049

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

##       fit      lwr      upr
## 1 3154.03 2858.296 3449.764

US <- predict(model, newdata=data.frame(Population=291049), interval="prediction")
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 , US[1]),
    status     = factor(c(rep("observed",nrow(Kyoto)),"predicted")))
```

Prediction, residuals

Application to the Kyoto data set (V)

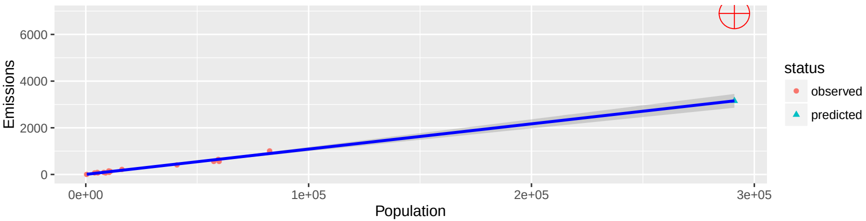
```
ggplot(Kyoto2,aes(x=Population,y=Emissions,colour=status,shape=status)) +
    geom_point() + stat_smooth(method=lm,colour="blue",fullrange=TRUE)
```



Prediction: careful!
Application to the Kyoto data set (VI)

If the new point does not follow the same model as the point from the training set. . .

```
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)
```



Outline

Decomposing the variance

Theorem of total variance

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

Vocabulary

- TSS = Total Sum of Squares
  $\rightsquigarrow$ Total variability to explain
- ESS = Explained Sum of Squares
  $\rightsquigarrow$ variability explained by the model
- RSS = Residual Sum of Squares
  $\rightsquigarrow$ Residual variability, not explained by the model

Decomposing the variance

Interpretation

Theorem of Total Variance (Pythagoras!)

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ and $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_n)^\top$, then

$$TTS = RSS + ESS$$
$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|_2^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2 + \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|_2^2$$

Hence,

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

- The variability explained by the model is **independent** from the residual variability.
- Geometrically, $\hat{\mathbf{Y}}$ is the **orthogonal projection** of $\mathbf{Y}$ on the subspace of $\mathbb{R}^n$ spawn by $\mathbf{x}$.

Decomposing the variance

Interpretation

Theorem of Total Variance (Pythagoras!)

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\top$, then

$$TTS = RSS + ESS$$
$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|_2^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2 + \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|_2^2$$

Hence,

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

- The variability explained by the model is **independent** from the residual variability.
- Geometrically, $\hat{\mathbf{Y}}$ is the **orthogonal projection** of $\mathbf{Y}$ on the subspace of $\mathbb{R}^n$ spawn by $\mathbf{x}$.

Coefficient of Determination
Definition

Coefficient of Determination

The of coefficient of determination is defined by

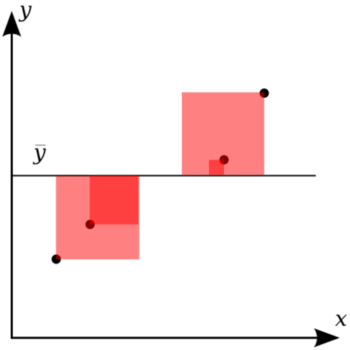
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Remark

The coefficient of determination can be interpreted as the percentage of variance explained by the model.

Coefficient of Determination

Interpretation for simple linear regression

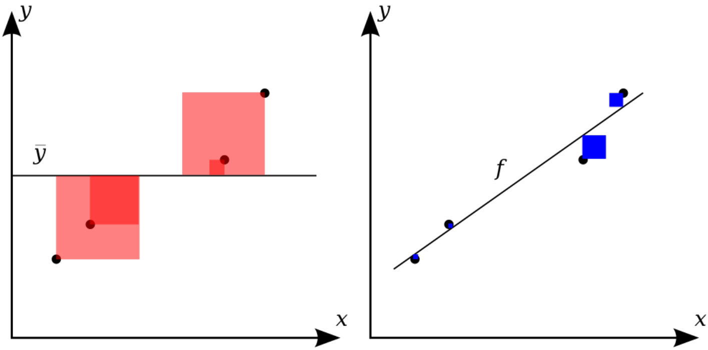


Model with the intercept

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

Coefficient of Determination

Interpretation for simple linear regression

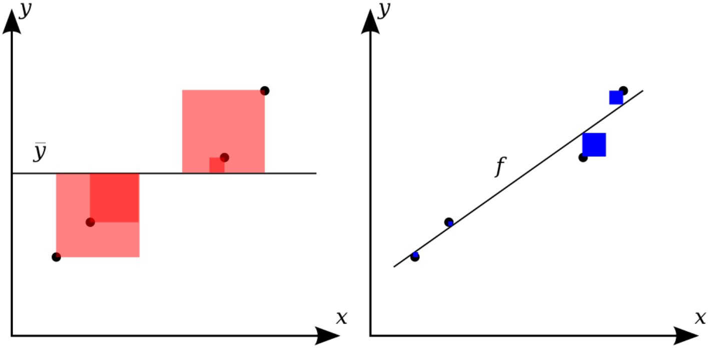


Model with intercept and slope

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

Coefficient of Determination

Interpretation for simple linear regression



Coefficient of determination

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

Testing the relevance of the model (I)

Hypothesis tested: nullity of β_1 (the slope)

$$\left\{\begin{array}{ll}\mathcal{M}_0: & \text{the simpler model}\\ \mathcal{M}_1: & \text{the more complex model}\end{array}\right. \Leftrightarrow \left\{\begin{array}{ll}\mathcal{M}_0: & Y_i = \beta_0 + \varepsilon_i\\ \mathcal{M}_1: & Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\end{array}\right.$$

Distribution of the SS under H_0

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

Moreover, $SCR \perp\!\!\!\perp SCM$

Testing the relevance of the model (I)

Hypothesis tested: nullity of β_1 (the slope)

$$\begin{cases} \mathcal{M}_0: & \text{the simpler model} \\ \mathcal{M}_1: & \text{the more complex model} \end{cases} \Leftrightarrow \begin{cases} \mathcal{M}_0: & Y_i = \beta_0 + \varepsilon_i \\ \mathcal{M}_1: & Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \end{cases}$$

Distribution of the SS under H_0

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

Moreover, $SCR \perp\!\!\!\perp SCM$

Testinf the relevance of the model (II)

Test Statistic: Fisher

Intuitively, we reject when the observed value of F is "large":

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

Proof...

Decision rule et $p-$value

We reject H_0 if $F \geq f_{1,n-2;1-\alpha}$ $\quad p-\text{val} = \mathbb{P}_{H_0}(\mathcal{F}_{1,n-2} \geq f(\text{obs}))$

Testinf the relevance of the model (II)

Test Statistic: Fisher

Intuitively, we reject when the observed value of F is "large":

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

Proof...

Decision rule et $p-$value

We reject H_0 if $F \geq f_{1,n-2;1-\alpha}$ $\quad p-\text{val} = \mathbb{P}_{H_0}\left(\mathcal{F}_{1,n-2} \geq f(\text{obs})\right)$

Analysis of variance

Summary Table

Source	Degrees de liberté	Sum og des Mean	Squares Squares	F
Model	1	ESS	EMS	$F = \frac{(n-2)EMS}{RSS}$
Residual	$n-2$	RSS	$\frac{RSS}{(n-2)}$	
Total	$n-1$	TSS		

Analysis of variance for simple linear regression I

Application to the Kyoto data set

```r
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
```

Analysis of variance for simple linear regression II

Application to the Kyoto data set

```
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
```

Analysis of variance for simple linear regression III

Application to the Kyoto data set

```
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
```

Outline

Recall the hypotheses of the regression model

Mostly related to the **noise**

1. Centered: $\mathbb{E}(Y) = \beta_0 + \beta_1 x$, then $\mathbb{E}(\varepsilon_i) = 0$
2. Homoscedastic: $\mathbb{V}(\varepsilon_i) = \sigma^2$ for all i,
3. Independent, $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$ for all $i \neq j$,
4. Gaussian: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Residual analysis

Diagnostic and solutions

We do not observe ε_i, so we the residual $\hat{\varepsilon}_i$ for the diagnostic

1. **Residuals graph**
   - looking for a tendency, heteroscedasticity, loss of centering
   - transformation the response Y_i and/or the x_i
2. Testing the independency (Durbin-Watson)
3. Testing the normality (Shapiro, Kolmogorov, χ^2)

Tolerance

- loss of Gaussianity: **few impact**, especially when the distribution remains symmetric
- independency: **important** for the inference (tests, estimation)

Residual analysis

Diagnostic and solutions

We do not observe ε_i, so we the residual $\hat{\varepsilon}_i$ for the diagnostic

1. **Residuals graph**
    - looking for a tendency, heteroscedasticity, loss of centering
    - transformation the response Y_i and/or the x_i
2. Testing the independency (Durbin-Watson)
3. Testing the normality (Shapiro, Kolmogorov, χ^2)

Tolerance

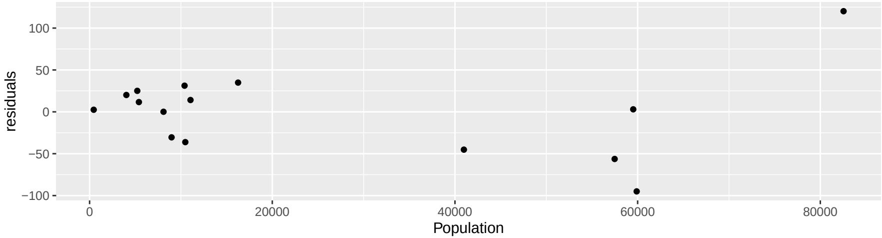
- loss of Gaussianity: **few impact**, especially when the distribution remains symmetric
- independency: **important** for the inference (tests, estimation)

Diagnostic

Application to the Kyoto data set (I)

Homoscedasticity ? Centering ? hum...

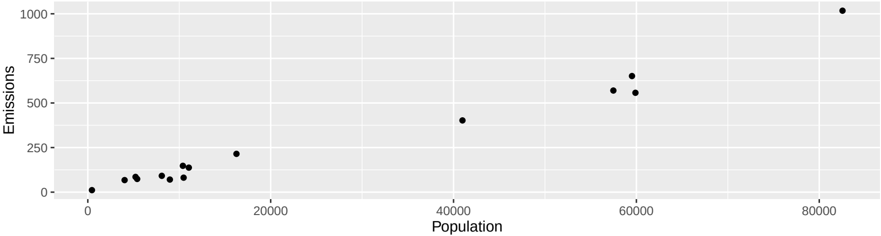
```
M1 <- lm(Emissions~Population,Kyoto)
Kyoto <- cbind(Kyoto, residuals=residuals(M1))
ggplot(Kyoto, aes(Population,residuals)) + geom_point()
```



Diagnostic: original data

Application to the Kyoto data set (II)

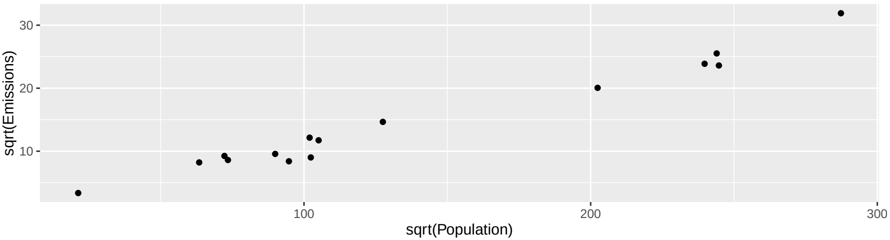
```
ggplot(Kyoto, aes(Population,Emissions)) + geom_point()
```



Diagnostic: square-root transformation

Application aux données Kyoto (III)

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



Diagnostic: square-root transformation

Application to the Kyoto data set (IV)

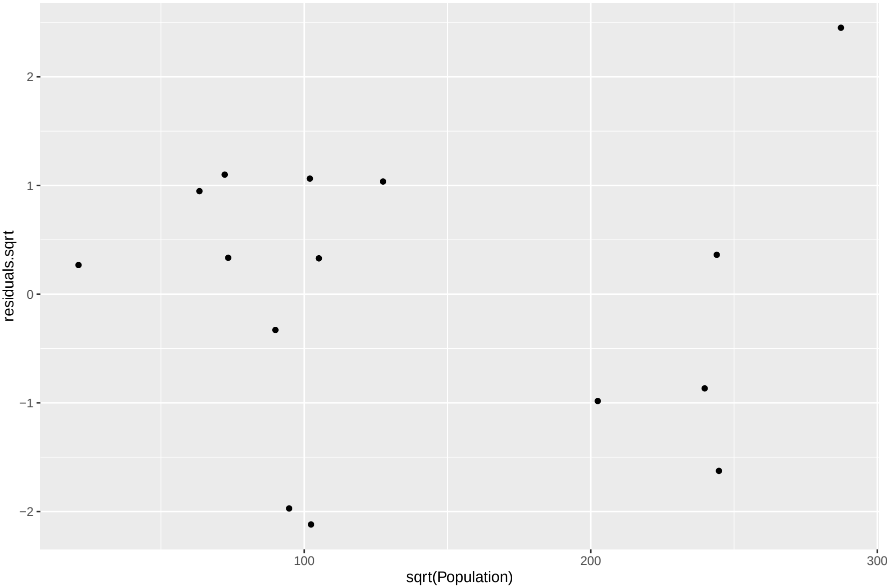
```
M1.sqrt <- lm(sqrt(Emissions)~sqrt(Population),Kyoto)
Kyoto <- cbind(Kyoto, residuals.sqrt=residuals(M1.sqrt))
summary(M1.sqrt)
```

```
##
## 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: square-root transformation

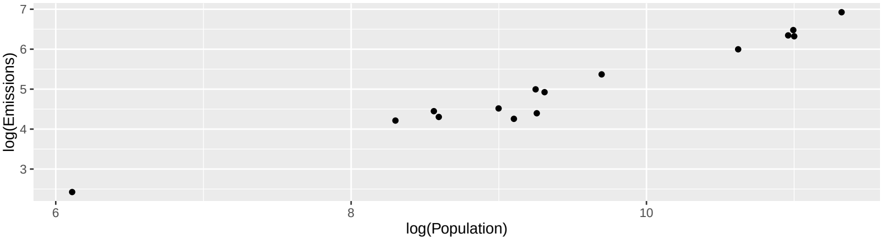
Application to the Kyoto data set (V)



Diagnostic: logarithmic transformation

Application to the Kyoto data set (VI)

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



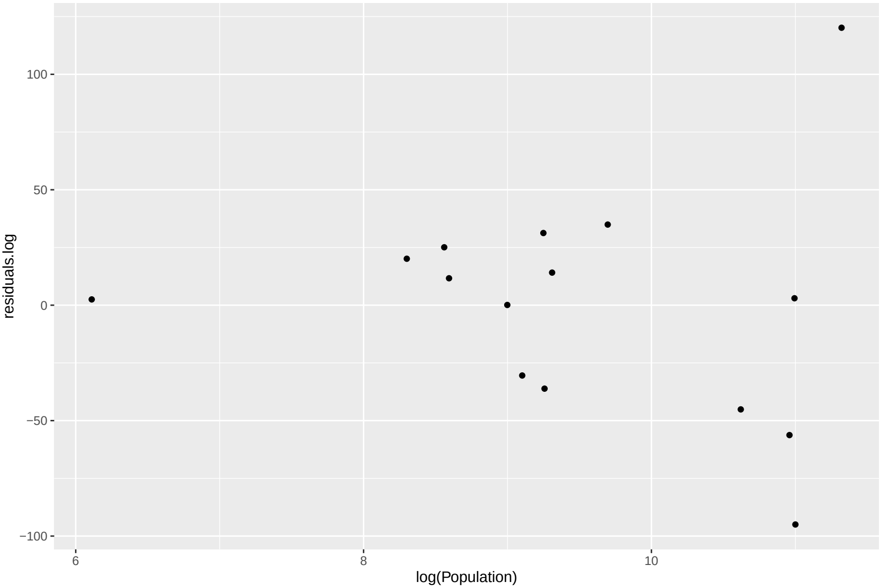
Diagnostic: logarithmic transformation

Application to the Kyoto data set (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: logarithmic transformation

Application to the Kyoto data set (VII)



Diagnostic: testing Gaussianity

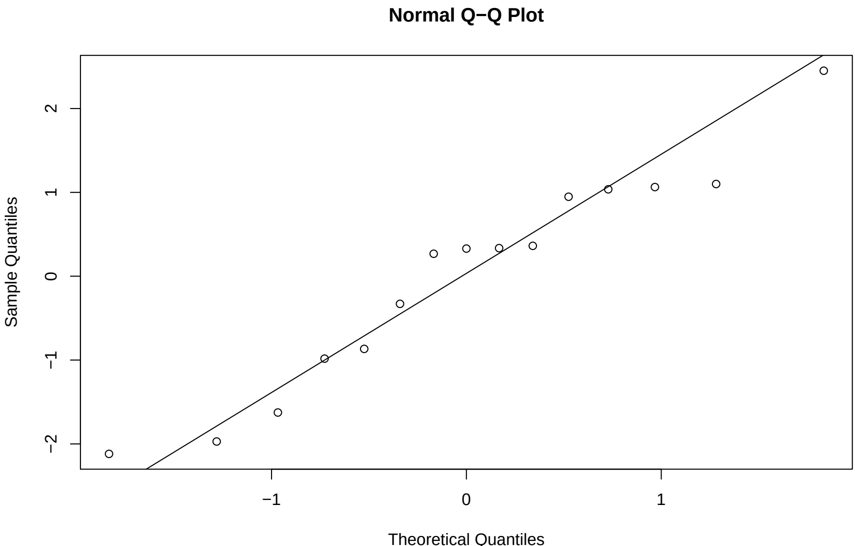
Application to the Kyoto data set (VIII)

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

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

Diagnostic: testing Gaussianity

Application to the Kyoto data set (IX)

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



Diagnostic: testing independency

Application to the Kyoto data set (X)

Testing the dependency of the residuals

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

##  lag Autocorrelation D-W Statistic p-value
##    1      -0.3678903       2.31636   0.466
##  Alternative hypothesis: rho != 0
```

Final model

Application to the Kyoto data set (XI)