

Regularization Methods for Linear Regression

Mathilde Mougeot

ENSIIE

2017-2018

Agenda

Lessons

- 3 plenary lessons (MRR)
- 3 Practical work sessions using R (mrr)
- 10 Project sessions (ipR)

Before next session, install on your computer

- ① R software, <https://www.r-project.org/>
- ② Rstudio, <https://www.rstudio.com/>

Documents are available (at this stage)

- <https://sites.google.com/site/MougeotMathilde/teaching>

A word on data and predictive models

Data are everywhere

- Industry (Temperature, IR sensors...)
- Finance : transactions
- Marketing : consumer data.
- on your phone (GPS, mail, musique ...)

→ Data base are available everywhere : from small data set to Big Data

A word on data and predictive models

Data are everywhere

- Industry (Temperature, IR sensors...)
- Finance : transactions
- Marketing : consumer data.
- on your phone (GPS, mail, musique ...)

→ Data base are available everywhere : from small data set to Big Data

Nowadays, predictive models are crucial for monitoring, for diagnosis

- Industry : Health monitoring, Energy...
- Finance : forecast of the evolution of the market
- Marketing : scoring
- Health

→ Machine learning models are used to mine, to operate the data.

Regularization Methods for Linear Regression

-Linear regression and Regularized Linear Regression belongs to the **Predictive model family**.

-Linear regression is an old model but still very useful !

Gauss, 1785 ; Legendre 1805

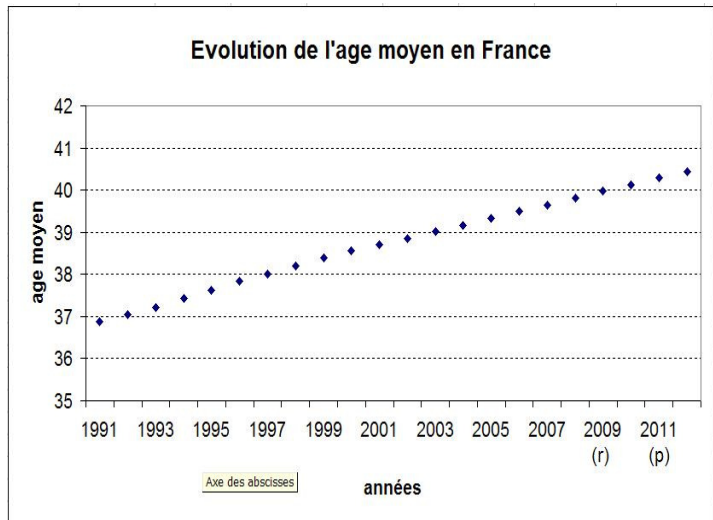
Outline of the lesson

- Motivations
- Ordinary Least Square -OLS- (geometrical approach)
- The linear Model (probabilistic approach)
- Using R software for modeling

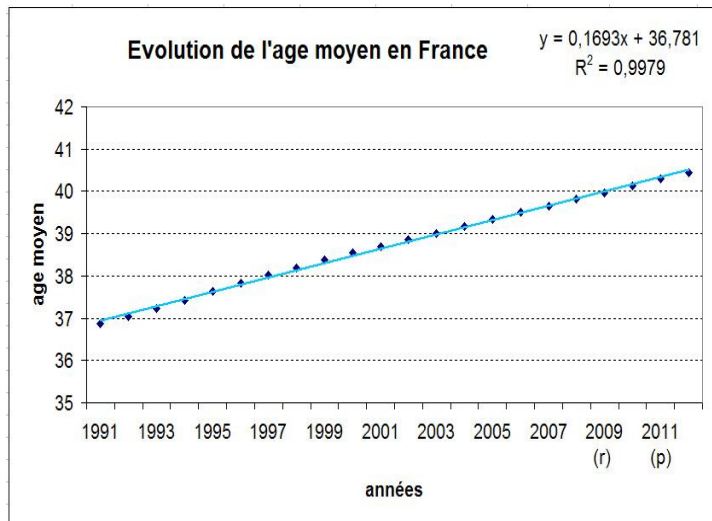
Evolution of the average age of the French population

	A	B	C	D	E	F	G
1	Évolution de l'âge moyen et de l'âge médian jusqu'en 2012						
2	Source : Insee, estimations de population.						
3		Âge moyen			Âge médian		
4		Ensemble	Hommes	Femmes	Ensemble	Hommes	Femmes
5	1991	36,9	35,3	38,4	33,7	32,4	35,0
6	1992	37,0	35,5	38,5	34,0	32,7	35,3
7	1993	37,2	35,7	38,7	34,3	32,9	35,6
8	1994	37,4	35,9	38,9	34,6	33,2	35,9
9	1995	37,6	36,1	39,1	34,9	33,6	36,2
10	1996	37,8	36,3	39,3	35,2	33,9	36,5
11	1997	38,0	36,5	39,5	35,5	34,1	36,8
12	1998	38,2	36,7	39,7	35,8	34,4	37,1
13	1999	38,4	36,9	39,8	36,1	34,7	37,4
14	2000	38,6	37,0	40,0	36,3	35,0	37,7
15	2001	38,7	37,2	40,1	36,6	35,3	38,0
16	2002	38,9	37,3	40,3	36,9	35,5	38,2
17	2003	39,0	37,5	40,4	37,1	35,8	38,5
18	2004	39,2	37,6	40,6	37,4	36,0	38,8
19	2005	39,3	37,8	40,8	37,7	36,2	39,1
20	2006	39,5	38,0	40,9	37,9	36,4	39,3
21	2007	39,7	38,1	41,1	38,1	36,7	39,6
22	2008 (r)	39,8	38,3	41,3	38,3	36,9	39,8
23	2009 (r)	40,0	38,5	41,4	38,6	37,1	40,0
24	2010 (p)	40,1	38,6	41,6	38,8	37,4	40,3
25	2011 (p)	40,3	38,8	41,7	39,0	37,6	40,5
26	2012 (p)	40,4	38,9	41,9	39,3	37,9	40,7
27	p : données provisoires, résultats arrêtés à fin 2011.						
28	r : données révisées.						
29	Champ : France.						

Evolution of the average age of the French population



Modeling the average age of the French population

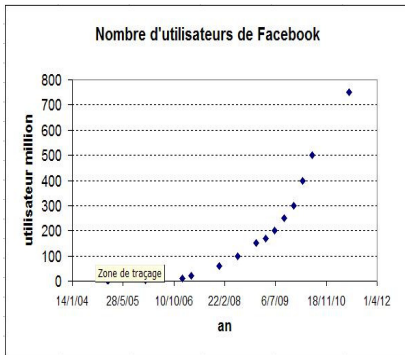


Application : Social Networks

Facebook users :

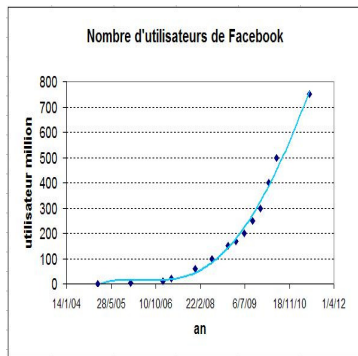
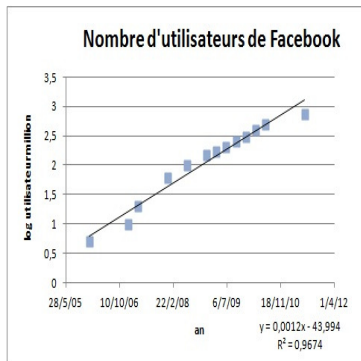
an	user(million)
31/12/04	0
31/12/05	5
31/12/06	10
31/3/07	20
30/12/07	60
30/6/08	100
30/12/08	150
30/3/09	170
30/6/09	200
30/9/09	250
30/12/09	300
30/3/10	400
30/6/10	500
30/6/11	750

Evolution of the number of Facebook users



Motivations : investment, forecast

Modeling the evolution of the number of Facebook users



Introduction : Regression model

- (Y, X) : couple of variables
 Y : Target quantitative variable
 $X = (X_1, X_2, \dots, X_p)$: Co-variates, quantitative variables
- The goal is to propose a Regression model to explain Y given X .
The parameters of the model are computed using a set of data

$$Y = \mathcal{F}_{data}(X) = \mathcal{F}_{data}(X_1, \dots, X_p)$$

here, \mathcal{F} is a linear function.

- Questions :
 - What are the performances of this model ?
 - What are the main explicative variables ?
 - Is-it possible to use the model to predict new values ? to forecast ?
 - Can we improve the model ?

Boston Housing Data

The original data are $n = 506$ observations on $p = 14$ variables,

medv	median value, being the target variable
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river ; 0 otherwise)
nox	nitric oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per USD 10,000
ptratio	pupil-teacher ratio by town
b	$1000(B - 0.63)^2$ where B is the proportion of blacks by town
lstat	percentage of lower status of the population
medv	median value of owner-occupied homes in USD 1000's

Boston Housing Data

The data :

nř	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
1	0.006	18	2.3	0	0.53	6.57	65.2	4.09	1	296	15.3	396.9	4.9	24.0
2	0.027	0	7.0	0	0.46	6.42	78.9	4.96	2	242	17.8	396.9	9.1	21.6
3	0.027	0	7.0	0	0.46	7.18	61.1	4.96	2	242	17.8	392.8	4.0	34.7
4	0.032	0	2.1	0	0.45	6.99	45.8	6.06	3	222	18.7	394.6	2.9	33.4
5	0.069	0	2.1	0	0.45	7.14	54.2	6.06	3	222	18.7	396.9	5.3	36.2
...

Boston Housing Data

- Model $Y = \mathcal{F}_{data}(X)$
- Evaluate the performances of the model
- What are the most important variables? (variable selection)
 - sparse models, less complex, best performances
- Inference and simulation
 - Ponctual estimation for new values of the co-variables
 - Confidence interval computation.

Outline

- Applications
- Ordinary Least Square (OLS) / Moindre Carrés Ordinaires (MCO)
- Linear Model
- Regularization methods : ridge, lasso

Ordinary Least Square (OLS)

Ordinary Least Square (OLS)

- Values/Variables :
 - $Y, Y \in \mathbb{R}$ **value/ Target variable**
 - $X = (X^1, ..., X^p), X \in \mathbb{R}^p$ **values/ covariates**

Ordinary Least Square (OLS)

- Values/Variables :
 - $Y, Y \in \mathbb{R}$ **value/ Target variable**
 - $X = (X^1, \dots, X^p), X \in \mathbb{R}^p$ **values/ covariates**
- Data : $S = \{(x_i, y_i) \mid i = 1, \dots, n, y_i \in \mathbb{R}, x_i \in \mathbb{R}^p\}$

Ordinary Least Square (OLS)

- Values/Variables :
 - $Y, Y \in \mathbb{R}$ **value/ Target variable**
 - $X = (X^1, \dots, X^p), X \in \mathbb{R}^p$ **values/ covariates**
- Data : $S = \{(x_i, y_i) \mid i = 1, \dots, n, y_i \in \mathbb{R}, x_i \in \mathbb{R}^p\}$
- Goal : Modeling Y **linearly** with X , with a "small" ϵ term

$$Y = \beta_1 + \beta_2 X_2 + \dots + \beta_p X_p + \boxed{\epsilon}$$

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \boxed{\epsilon}$$

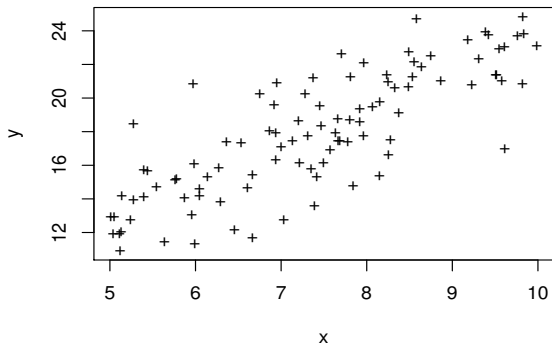
$$Y = \sum_j^p \beta_j X_j + \boxed{\epsilon}$$

Ordinary Least Square (OLS) Simple Linear Regression model

Simple Linear Regression : example

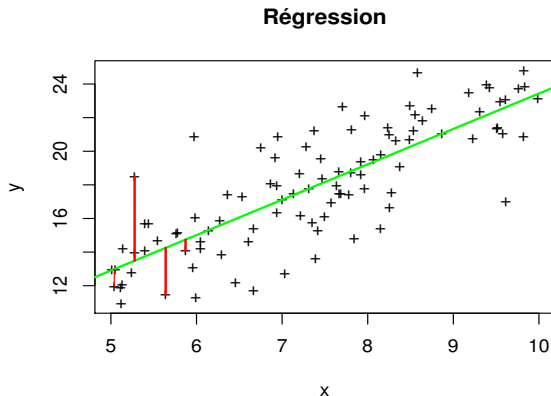
We only have one co-variable (X) to explain the target variable (Y).

The scatter plot is represented by :



Simple Linear Regression : example

For all observation couples i , $1 \leq i \leq n$ (Y_i, X_i),
the goal is here to minimize $(Y_i - (\beta_1 + \beta_2 X_i))^2$



Ordinary Least Square : simple linear model

Formalism :

- Distance to a single point : $(y_i - x_i\beta_2 - \beta_1)^2$
- Distance to the whole sample : $\sum_{i=1}^n (y_i - x_i\beta_2 - \beta_1)^2$
 → Best line : intercept $\hat{\beta}_1$ and slope $\hat{\beta}_2$ such that $\sum_{i=1}^n (y_i - x_i\beta_2 - \beta_1)^2$ is minimum, among all possible values of β_1 and β_2 .

OLS estimator :

The values estimated by OLS (the estimates) for β_1 and β_2 verify :

$$(\hat{\beta}_1^{ols}, \hat{\beta}_2^{ols}) = \arg \min_{\beta_1, \beta_2 \in \mathbb{R}^2} \left\{ \sum_{i=1}^n (y_i - x_i\beta_2 - \beta_1)^2 \right\}$$

Ordinary Least Square : simple linear model

OLS estimator (observations) :

The values estimated by OLS (the estimates) for β_1 and β_2 verify :

$$(\hat{\beta}_1^{ols}, \hat{\beta}_2^{ols}) = \arg \min_{\beta_1, \beta_2 \in \mathbb{R}^2} \left\{ \sum_{i=1}^n (y_i - x_i \beta_2 - \beta_1)^2 \right\}$$

OLS estimator (vector notations) : $y, x, 1_n$

The values estimated by OLS (the estimates) for β_1 and β_2 verify :

$$(\hat{\beta}_1^{ols}, \hat{\beta}_2^{ols}) = \arg \min_{\beta_1, \beta_2 \in \mathbb{R}^2} \|y - x\beta_2 - 1_n\beta_1\|_2^2$$

OLS estimator (matrix notations Y (n,1); X (n,2)) :

The values estimated by OLS (the estimates) for β_1 and β_2 verify :

$$(\hat{\beta}_1^{ols}, \hat{\beta}_2^{ols}) = \arg \min_{\beta_1, \beta_2 \in \mathbb{R}^2} \|Y - X\beta\|_2^2$$

Ordinary Least Square : simple linear model

Theorem :

The OLS estimators have the following expressions :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

Proof :

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

Ordinary Least Square : simple linear model

For the simple linear model, the correlation coefficient may be very useful :

- $r(x, y)$: correlation coefficient/ coefficient de corrélation linéaire

$$r(x, y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}$$

- $r(x, y) = 1$ if and only if $Y = aX + b$, linear relation between Y et X

Ordinary Least Square : simple linear model

For the simple linear model, the correlation coefficient may be very useful :

- $r(x, y)$: correlation coefficient/ coefficient de corrélation linéaire

$$r(x, y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}$$

- $r(x, y) = 1$ if and only if $Y = aX + b$, linear relation between Y et X

R-square used in multiple regression

- $R^2 = \frac{\text{Var}\hat{Y}}{\text{Var}(Y)}$
- $R^2 \in [0, 1]$
- **Simple regression** $R^2 = r^2$:

Ordinary Least Square : simple linear model

For the simple linear model, the correlation coefficient may be very useful :

- $r(x, y)$: correlation coefficient/ coefficient de corrélation linéaire

$$r(x, y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}$$

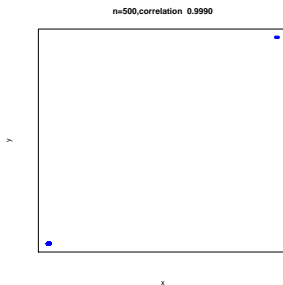
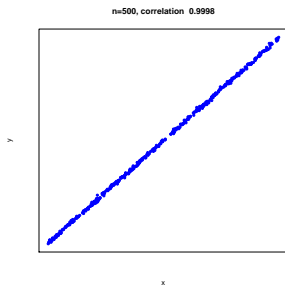
- $r(x, y) = 1$ if and only if $Y = aX + b$, linear relation between Y et X

R-square used in multiple regression

- $R^2 = \frac{\text{Var}\hat{Y}}{\text{Var}(Y)}$
- $R^2 \in [0, 1]$
- **Simple regression** $R^2 = r^2$:

Best Practices

The correlation coefficient equals 1 for these two cases :



Always looking at the data !!

Ordinary Least Square (OLS) Multiple Linear Regression model

Ordinary Least Square : multiple linear model

- **We suppose** $Y = \sum_j^p \beta_j X^j + \epsilon$ and $S = \{(x_i, y_i) \mid i = 1 \dots n, y_i \in \mathbb{R} \ x_i \in \mathbb{R}^p\}$
- The Quadratic error is defined by :

$$E(\beta) = \sum_i^n \epsilon_i^2 = \sum_i^n (y_i - \sum_j x_i^j \beta_j)^2$$

Ordinary Least Square : multiple linear model

- **We suppose** $Y = \sum_j^p \beta_j X^j + \epsilon$ and $S = \{(x_i, y_i) \mid i = 1 \dots n, y_i \in \mathbb{R} \ x_i \in \mathbb{R}^p\}$
- The Quadratic error is defined by :

$$E(\beta) = \sum_i^n \epsilon_i^2 = \sum_i^n (y_i - \sum_j x_i^j \beta_j)^2$$

with matrix notation :

$$E(\beta) = (Y - X\beta)^T (Y - X\beta)$$

Ordinary Least Square : multiple linear model

- **We suppose** $Y = \sum_j^p \beta_j X^j + \epsilon$ and $S = \{(x_i, y_i) \mid i = 1 \dots n, y_i \in \mathbb{R} \ x_i \in \mathbb{R}^p\}$
- The Quadratic error is defined by :

$$E(\beta) = \sum_i^n \epsilon_i^2 = \sum_i^n (y_i - \sum_j x_i^j \beta_j)^2$$

with matrix notation :

$$E(\beta) = (Y - X\beta)^T (Y - X\beta)$$

- **Goal :** To minimize the error $E(\beta)$ on the data set S .
To compute $\hat{\beta} \in \mathbb{R}^p$:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} E(\beta)$$

Ordinary Least Square. multiple regression model

- We aim to compute β which minimize :

$$\begin{aligned} E(\beta) &= ||Y - X\beta||_2^2 \\ &= (Y - X\beta)^T (Y - X\beta) \end{aligned}$$

- Assumption : $X^T X$ **invertible**. ($n \geq p$)

Theorem :

$$\hat{\beta}_{MCO} = (X^T X)^{-1} X^T Y$$

MCO

- Estimation

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

MCO

- Estimation

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

- Prediction Knowing $\hat{\beta}$ and given X_1, \dots, X_p ,
the prediction of the target can be computed : $\hat{Y} = \sum_j \hat{\beta}_j X_j$

$$\begin{aligned} \hat{Y} &= X \hat{\beta} \\ &= X (X^T X)^{-1} X^T Y \end{aligned}$$

MCO

- Estimation

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

- Prediction Knowing $\hat{\beta}$ and given X_1, \dots, X_p ,
the prediction of the target can be computed : $\hat{Y} = \sum_j \hat{\beta}_j X_j$

$$\begin{aligned} \hat{Y} &= X \hat{\beta} \\ &= X (X^T X)^{-1} X^T Y \end{aligned}$$

- P Projection matrix on the Hyperplan (hat matrix)

$$\begin{aligned} P &= X (X^T X)^{-1} X^T \\ P^2 &= P \end{aligned}$$

MCO

- Estimation

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

- Prediction Knowing $\hat{\beta}$ and given X_1, \dots, X_p ,
the prediction of the target can be computed : $\hat{Y} = \sum_j \hat{\beta}_j X_j$

$$\begin{aligned}\hat{Y} &= X \hat{\beta} \\ &= X (X^T X)^{-1} X^T Y\end{aligned}$$

- P Projection matrix on the Hyperplan (hat matrix)

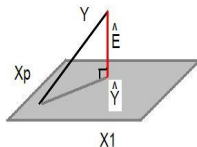
$$\begin{aligned}P &= X (X^T X)^{-1} X^T \\ P^2 &= P\end{aligned}$$

- Residuals

$$- \hat{\epsilon} = Y - \hat{Y}$$

- **Remarque : no assumption on the law or distribution of ϵ**

Ordinary Least Square. Geometrical interpretation



$$\begin{array}{l} Y \\ \in \mathcal{R}^n \end{array} = \begin{array}{l} \sum_j^p X^j \beta_j \\ \in \mathcal{R}^p \end{array} + \begin{array}{l} \epsilon \\ \in \mathcal{R}^{(n-p)} \end{array}$$

Ordinary Least Square. Properties

- Orthogonality :
 - $\hat{Y} \perp \hat{\epsilon}$
 - $X_j \perp \hat{\epsilon} \quad \forall j \in [1 \dots p] \quad \langle X^j, \hat{\epsilon} \rangle = 0$

Ordinary Least Square. Properties

- Orthogonality :
 - $\hat{Y} \perp \hat{\epsilon}$
 - $X_j \perp \hat{\epsilon} \quad \forall j \in [1 \dots p] \quad \langle X^j, \hat{\epsilon} \rangle = 0$
- Residual average :
 - $\sum_i \hat{\epsilon}_i = 0$ if there is an intercept in the model $X^1 = (1, 1, \dots, 1)$
 - the average point belongs to the hyperplan
 - $\bar{\hat{Y}} = \bar{Y}$

Ordinary Least Square. Properties

- Orthogonality :
 - $\hat{Y} \perp \hat{\epsilon}$
 - $X_j \perp \hat{\epsilon} \quad \forall j \in [1 \dots p] \quad \langle X^j, \hat{\epsilon} \rangle = 0$
- Residual average :
 - $\sum_i \hat{\epsilon}_i = 0$ if there is an intercept in the model $X^1 = (1, 1, \dots, 1)$
 - the average point belongs to the hyperplan
 - $\bar{\hat{Y}} = \bar{Y}$
- Analysis of Variance -ANAVAR- (Pythagore)
 - $var(Y) = var(\hat{Y}) + var(\hat{E})$

Multiple Linear model : example with R

```
head(mydata,3);  
y x1 x2 x3  
1 -2.20 0.38 0.98 0.46  
2 -1.75 0.11 0.62 0.37  
3 -0.24 0.80 0.59 0.87  
...  
> modlm=lm(y ~ x1+x2+x3,data=mydata);  
Call :  
lm(formula = y ~ x1+x2+x3, data = mydata)  
Coefficients :  
(Intercept) x1 x2 x3  
0.02754 1.98163 -3.03612 0.01903
```

Multiple Linear model : example with R

```
> modlm=lm(y ~ x1+x2+x3,data=mydata);
> summary(modlm)
lm(formula = y ~ x1+x2+x3, data = mydata)
```

Residuals :

```
Min 1Q Median 3Q Max
-0.29 -0.075 -0.0035 0.073 0.281
```

Coefficients :

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.02754 0.01503 1.833 0.0674 .
x1          1.98163 0.01577 125.652 <2e-16 ***
x2          -3.03612 0.01621 -187.286 <2e-16 ***
x3           0.01903 0.01576  1.208 0.2277
— Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error : 0.1009 on 496 degrees of freedom
Multiple R-squared : 0.9904, Adjusted R-squared : 0.9904
F-statistic : 1.707e+04 on 3 and 496 DF, p-value : < 2.2e-16
```

Ordinary Least Square, quality of the adjustment

- R^2 , R-square (French : coefficient de détermination)
 - $R^2 = \frac{\text{var}(\hat{Y})}{\text{var}(Y)}$, remark : no unit

Ordinary Least Square, quality of the adjustment

- R^2 , R-square (French : coefficient de détermination)
 - $R^2 = \frac{\text{var}(\hat{Y})}{\text{var}(Y)}$, remark : no unit
 - $\cos^2 w = R^2 = \frac{\|\hat{Y} - \bar{\hat{Y}}_{1,n}\|^2}{\|Y - \bar{Y}_{1,n}\|^2}$
 w : angle between the centered vector $(Y - \bar{Y}_{1,n})$ and its centered prediction $(\hat{Y} - \bar{\hat{Y}}_{1,n})$

Ordinary Least Square, quality of the adjustment

- R^2 , R-square (French : coefficient de détermination)
 - $R^2 = \frac{\text{var}(\hat{Y})}{\text{var}(Y)}$, remark : no unit
 - $\cos^2 w = R^2 = \frac{\|\hat{Y} - \bar{\hat{Y}}_{1,n}\|^2}{\|Y - \bar{Y}_{1,n}\|^2}$
 w : angle between the centered vector $(Y - \bar{Y}_{1,n})$ and its centered prediction $(\hat{Y} - \bar{\hat{Y}}_{1,n})$
- $\text{var}(\hat{E}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
 $= (1 - R^2)\text{var}(Y)$, unit of Y^2

Ordinary Least Square : Quality of the Ajustement

- R-square (for **few** variables)
 - $R^2 = \cos^2 \omega = \frac{\text{Var} \hat{Y}}{\text{Var}(Y)}$
 - $R^2 \in [0, 1]$
 - $R^2 =$ **increases mechanically with the number of variables**

Ordinary Least Square : Quality of the Adjustment

- R-square (for **few** variables)
 - $R^2 = \cos^2 \omega = \frac{\text{Var} \hat{Y}}{\text{Var}(Y)}$
 - $R^2 \in [0, 1]$
 - $R^2 =$ **increases mechanically with the number of variables**
- Adjusted R-squared is sometimes preferred (penalization with the number of variables)
 - $R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-p}$
 - R_{adj}^2 may be negative

Ordinary Least Square : Quality of the Adjustment

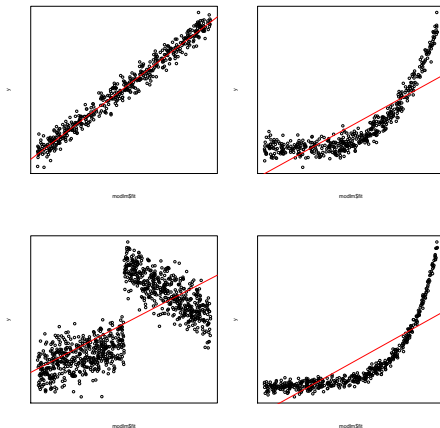
- R-square (for **few** variables)
 - $R^2 = \cos^2 \omega = \frac{\text{Var} \hat{Y}}{\text{Var}(Y)}$
 - $R^2 \in [0, 1]$
 - $R^2 =$ **increases mechanically with the number of variables**
- Adjusted R-squared is sometimes preferred (penalization with the number of variables)
 - $R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-p}$
 - R_{adj}^2 may be negative
- Residual study :
 - $\hat{\epsilon}_i = y_i - \hat{y}_i \quad \forall i \in 1..n$
 - Visualization of
 - $(\hat{\epsilon}_i, i) \quad \forall i \in 1..n$
 - $(\hat{\epsilon}_i, y_i)$ homoscedastic vs heteroscedastic bissectrice model

Ordinary Least Square : Quality of the Adjustment

- R-square (for **few** variables)
 - $R^2 = \cos^2 \omega = \frac{\text{Var} \hat{Y}}{\text{Var}(Y)}$
 - $R^2 \in [0, 1]$
 - $R^2 =$ **increases mechanically with the number of variables**
- Adjusted R-squared is sometimes preferred (penalization with the number of variables)
 - $R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-p}$
 - R_{adj}^2 may be negative
- Residual study :
 - $\hat{\epsilon}_i = y_i - \hat{y}_i \quad \forall i \in 1..n$
 - Visualization of
 - $(\hat{\epsilon}_i, i) \quad \forall i \in 1..n$
 - $(\hat{\epsilon}_i, y_i)$ homoscedastic vs heteroscedastic bissectrice model
- Prediction : Visualization of
 - $(\hat{y}_i, y_i) \quad \forall i \in 1..n$
 - comparison with the first bisector.

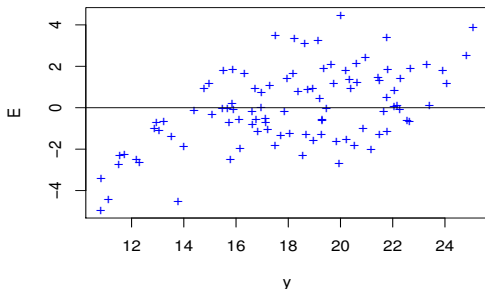
Ordinary Least Square : Quality of the Adjustment

Graphics (y_i, \hat{y}_i) $1 \leq i \leq j$ **VERY USEFUL**



Ordinary Least Square : Student Residual graph

$$\frac{\hat{\epsilon}_i}{S_E} = \frac{y_i - \hat{y}_i}{S_E} \text{ (no unit term)}$$

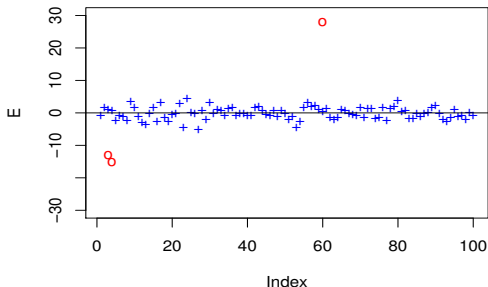


Residual graph

→ Random distribution. There is no information to be capture

Ordinary Least Square : Student Residual graph

$$\frac{\hat{\epsilon}_i}{S_E} = \frac{y_i - \hat{y}_i}{S_E} \text{ (with non unit)}$$

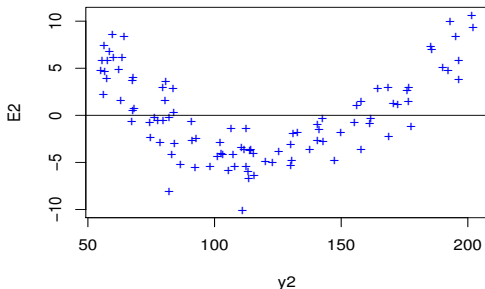


Residual graph function of Y

→ Large values for some points ? Outliers detection ?

Ordinary Least Square : Student Residual graph

$$\frac{\hat{\epsilon}_i}{S_E} = \frac{y_i - \hat{y}_i}{S_E} \text{ (with non unit)}$$

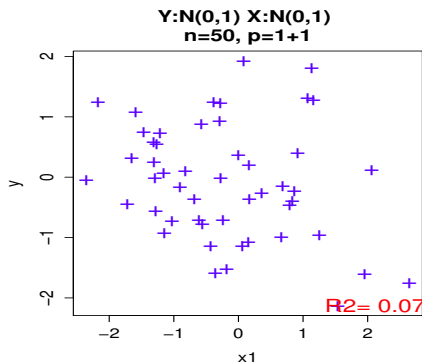


Grphe des résidus en fonction de Y

- there is still some information in the residuals.
- The model needs to be changed.

Ordinary Least Square : curse of dimension

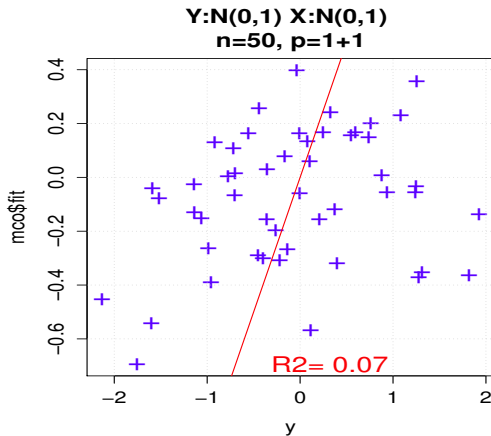
Data set : $\{(y_i, x_i) | 1 \leq i \leq n\}$. One target variable, one covariable



$$\rightarrow R^2 =, R^2_{adj} = -0.02$$

Ordinary Least Square : illustration of the impact of the number of covariables on the model

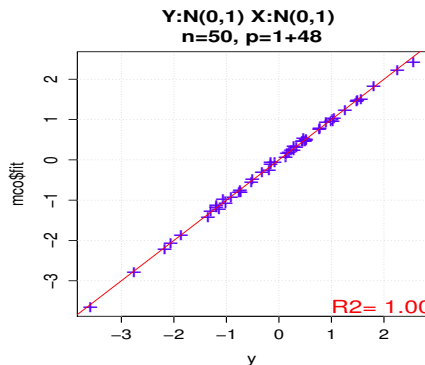
initial data and OLS line



$$\rightarrow R^2 =, R^2_{adj} = -0.02$$

Ordinary Least Square : illustration of the impact of the number of covariables on the model

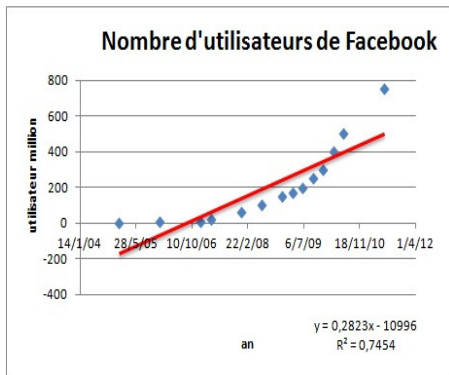
initial data and **48 more covariables** $\mathcal{N}(0, 1)$ are added to the initial data set.



$$\rightarrow R^2 = 0.99, R^2_{adj} = 0.93$$

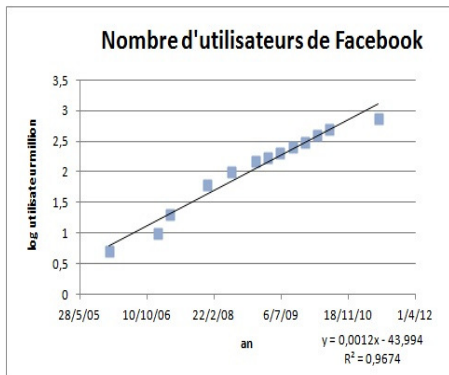
Ordinary Least Square : need to change the model (1/2)

initial data set :



Ordinary Least Square : need to change the model (1/2)

logarithmic transformation



MCO Regression. Some limits :

If $X^T X$ is non invertible

- $n \gg p$, collinearity between some X_j .
 - Pseudo-inverse, the solution is not unique
 - Variable selection
- $p \gg n$, when the number of variables is larger than the number of observations
 - Regularization method
 - Ridge -L2-, Lasso -L1-.
 - Variable selection

OLS model

Ponctual estimation.

OLS, $X^T X$ non invertible \rightarrow

Pseudo inverse computation

Solution ($n > p$), $X^T X$ is non invertible with the rank k , $k < p$:

$$\begin{aligned} X^T X &= U \Sigma^2 U^T \\ &= U \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \vdots & 0 & 0 \\ 0 & 0 & \sigma_k^2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} U^T \\ &= U_k \Sigma_k^2 U_k^T \end{aligned}$$

$$(X^T X)^{*-1} = U_k \Sigma_k^{2^{-1}} U_k^T \quad \text{avec} \quad \Sigma_k^2 = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \vdots & 0 \\ 0 & 0 & \sigma_k^2 \end{pmatrix}$$

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

The solution non unique

Outline

- Motivations
- Ordinary Least Square
- **Linear Model**
- Penalized regression, ridge, lasso

Linear Model

Probabilistic assumption on the residuals

Linear model

- We write : $Y = X\beta + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$

- We have

- $\epsilon_i = Y_i - \sum X_i^j \beta_j$ avec $f(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\epsilon^2}{2\sigma^2}}$ i.i.d.

Linear model

- We write : $Y = X\beta + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$

- We have

$$\epsilon_i = Y_i - \sum X_i^j \beta_j \quad \text{avec } f(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\epsilon^2}{2\sigma^2}} \quad \text{i.i.d.}$$

- Residual density & Maximum Likelihood Estimation

$$f(\epsilon_1, \dots, \epsilon_n) = \prod_i f(\epsilon_i)$$

Linear model

- We write : $Y = X\beta + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$

- We have

$$\epsilon_i = Y_i - \sum X_i^j \beta_j \quad \text{avec } f(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\epsilon^2}{2\sigma^2}} \quad \text{i.i.d.}$$

- Residual density & Maximum Likelihood Estimation

$$\begin{aligned} f(\epsilon_1, \dots, \epsilon_n) &= \prod_i f(\epsilon_i) \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{\sum \epsilon_i^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2} \sigma^{2n/2}} e^{-\frac{\|Y - X\beta\|^2}{2\sigma^2}} \end{aligned}$$

Linear model

- We write : $Y = X\beta + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$

- We have

$$- \epsilon_i = Y_i - \sum X_i^j \beta_j \quad \text{avec } f(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\epsilon^2}{2\sigma^2}} \quad \text{i.i.d.}$$

- Residual density & Maximum Likelihood Estimation

$$\begin{aligned} f(\epsilon_1, \dots, \epsilon_n) &= \prod_i f(\epsilon_i) \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{\sum \epsilon_i^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2} \sigma^{2n/2}} e^{-\frac{\|Y - X\beta\|^2}{2\sigma^2}} \end{aligned}$$

- the goal is to compute $\hat{\beta}$, σ^2 solutions of the maximum likelihood Estimation (MLE)

Same solution for the MLE and the OLS :

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

$$- \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{i=n} \hat{\epsilon}_i^2$$

Linear model : What are the laws of the estimators ?

$$Y = X\beta + \epsilon \text{ avec } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Law of the estimators :

- Law of $\hat{\beta}$?
- Law of \hat{Y} ?
- Law of $\hat{\sigma}^2$?

Benefits

- let to compute confidence intervals for β and Y .
- let to test the parameters.

Law of $\hat{\beta}$:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^T X)^{-1})$$

Law of $\hat{\beta}$:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^T X)^{-1})$$

Expectation and Variance of $\hat{\beta}$? :

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{et } Y = X\beta + \epsilon$$

Law of $\hat{\beta}$:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^T X)^{-1})$$

Expectation and Variance of $\hat{\beta}$? :

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{et } Y = X\beta + \epsilon$$

- $\mathbb{E}(\hat{\beta}) = \beta$ Non biased estimator $\mathbb{E}(\hat{\beta}) - \beta = 0$

Law of $\hat{\beta}$:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^T X)^{-1})$$

Expectation and Variance of $\hat{\beta}$? :

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{et } Y = X\beta + \epsilon$$

- $\mathbb{E}(\hat{\beta}) = \beta$ Non biased estimator $\mathbb{E}(\hat{\beta}) - \beta = 0$

- $\text{Var}(\hat{\beta}) = \sigma^2(X^T X)^{-1}$

$$\begin{aligned} \text{Var}(\hat{\beta}) &= (X^T X)^{-1} X^T ([\text{var}(Y)] X (X^T X)^{-1}) \\ &= (X^T X)^{-1} X^T ([\text{var}(\epsilon)] X (X^T X)^{-1}) \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

- $\mathbb{E}[(\hat{\beta} - \beta)^2] = \text{Var}(\hat{\beta}) + 0$

Recall $\text{Var}(aY) = a\text{Var}(Y)a^T$

Law of \hat{Y}

$$\hat{Y} \sim \mathcal{N}(X\beta, \sigma^2 X(X^T X)^{-1} X^T)$$

Expectation and Variance of \hat{Y} ?, $\hat{Y} = X\hat{\beta}$

Law of \hat{Y}

$$\hat{Y} \sim \mathcal{N}(X\beta, \sigma^2 X(X^T X)^{-1} X^T)$$

Expectation and Variance of \hat{Y} ?, $\hat{Y} = X\hat{\beta}$

- $\mathbb{E}(\hat{Y}) = X\beta$
 $\mathbb{E}(\hat{Y}) = \mathbb{E}(X\hat{\beta}) = X\mathbb{E}(\hat{\beta}) = X\beta = \mathbb{E}(Y)$

Law of \hat{Y}

$$\hat{Y} \sim \mathcal{N}(X\beta, \sigma^2 X(X^T X)^{-1} X^T)$$

Expectation and Variance of \hat{Y} ?, $\hat{Y} = X\hat{\beta}$

- $\mathbb{E}(\hat{Y}) = X\beta$
 $\mathbb{E}(\hat{Y}) = \mathbb{E}(X\hat{\beta}) = X\mathbb{E}(\hat{\beta}) = X\beta = \mathbb{E}(Y)$

- $\text{Var}(\hat{Y}) = \sigma^2 X(X^T X)^{-1} X^T$

$$\begin{aligned}\text{Var}(\hat{Y}) &= \text{Var}(X\hat{\beta}) \\ &= X \text{Var}(\hat{\beta}) X^T \\ &= \sigma^2 X(X^T X)^{-1} X^T\end{aligned}$$

Law of $\hat{\epsilon}$

$$\hat{\epsilon} \sim \mathcal{N}(0, \sigma^2(I_n - X(X^T X)^{-1}X^T))$$

Expectation and Variance of $\hat{\epsilon} = Y - \hat{Y}$? :

Law of $\hat{\epsilon}$

$$\hat{\epsilon} \sim \mathcal{N}(0, \sigma^2(I_n - X(X^T X)^{-1}X^T))$$

Expectation and Variance of $\hat{\epsilon} = Y - \hat{Y}$? :

- $\mathbb{E}(\hat{\epsilon}) = 0$

Law of $\hat{\epsilon}$

$$\hat{\epsilon} \sim \mathcal{N}(0, \sigma^2(I_n - X(X^T X)^{-1}X^T))$$

Expectation and Variance of $\hat{\epsilon} = Y - \hat{Y}$? :

- $\mathbb{E}(\hat{\epsilon}) = 0$
- $Var(\hat{\epsilon}) = \sigma^2(I_n - X(X^T X)^{-1}X^T)$

$$\begin{aligned} Var(\hat{\epsilon}) &= Var(Y - \hat{Y}) \\ &= Var(Y - X\hat{\beta}) \\ &= \sigma^2(I_n) - XVar(\hat{\beta})X^T \\ &= \sigma^2(I_n - X(X^T X)^{-1}X^T) \end{aligned}$$

Recal : $Var(aY) = aVar(Y)a^T$

Linear model : law of the estimators

Under the assumption that ϵ_i are i.i.d. with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Theorem

if $p \leq n$ and $X^T X$ invertible,

The vector $\begin{pmatrix} \hat{\beta} \\ \hat{\epsilon} \end{pmatrix}$ of dimension $(p + n)$ is a gaussian vector

with mean $\begin{pmatrix} \beta \\ 0 \end{pmatrix}$, and

and variance $\sigma^2 \begin{pmatrix} (X^T X)^{-1} & 0 \\ 0 & I_n - X(X^T X)^{-1}X^T \end{pmatrix}$

Loi $\hat{\sigma}^2$

$$\frac{n-p}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-p}^2$$

We note : $\hat{\sigma}^2 = \frac{||\hat{\epsilon}||^2}{n-p}$

Loi $\hat{\sigma}^2$

$$\frac{n-p}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-p}^2$$

We note : $\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|^2}{n-p}$

$\|\hat{\epsilon}\|^2 = \sum_i^n \hat{\epsilon}_i^2$ $\|\hat{\epsilon}\|^2$ suit une loi $\sigma^2 \chi^2(n-p)$ (Cochran theorem)

Loi $\hat{\sigma}^2$

$$\frac{n-p}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-p}^2$$

We note : $\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|^2}{n-p}$

$\|\hat{\epsilon}\|^2 = \sum_i^n \hat{\epsilon}_i^2$ $\|\hat{\epsilon}\|^2$ suit une loi $\sigma^2 \chi^2(n-p)$ (Cochran theorem)

Then, the expectation of $\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|^2}{n-p}$ is σ^2 ,
($\mathbb{E}(\chi^2(n-p)) = n-p$)

We deduce the law of $\hat{\sigma}^2$:

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p} \chi^2(n-p)$$

Recall : Student theorem.

$U \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2(d)$, U and V are independant, then, we have
 $Z = \frac{U}{\sqrt{V/d}}$ follows a Student law of parameter d .

Significativity test of $\hat{\beta}_j$, σ^2 **inconnu**

- Student Statistics : T

Significativity test of $\hat{\beta}_j$, σ^2 **inconnu**

- Student Statistics : T
- Significativity test (bilateral)
 - $H_0 : \beta_j = 0$
 - $H_1 : \beta_j \neq 0$

Significativity test of $\hat{\beta}_j$, σ^2 **inconnu**

- Student Statistics : T
- Significativity test (bilateral)
 - $H_0 : \beta_j = 0$
 - $H_1 : \beta_j \neq 0$
- Decision with a risk α , **Reject of H_0 if**
 - $\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 S_{j,j}}} > t_{n-p}(1 - \alpha/2)$ with $S_{j,j}$ jème term of the diagonal of $(X^T X)^{-1}$
 - $\text{pvalue} < \alpha$

Significativity test of $\hat{\beta}_j$, σ^2 **inconnu**

- Student Statistics : T
- Significativity test (bilateral)
 - $H_0 : \beta_j = 0$
 - $H_1 : \beta_j \neq 0$
- Decision with a risk α , **Reject of H_0 if**
 - $\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 S_{j,j}}} > t_{n-p}(1 - \alpha/2)$ with $S_{j,j}$ jème term of the diagonal of $(X^T X)^{-1}$
 - $\text{pvalue} < \alpha$
- Conclusion :
 - β_j is significantly different of zero
 - X_j a une influence dans le modèle

Significativity test of $\hat{\beta}_j$, σ^2 **inconnu**

- Student Statistics : T
- Significativity test (bilateral)
 - $H_0 : \beta_j = 0$
 - $H_1 : \beta_j \neq 0$
- Decision with a risk α , **Reject of H_0 if**
 - $\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 S_{j,j}}} > t_{n-p}(1 - \alpha/2)$ with $S_{j,j}$ jème term of the diagonal of $(X^T X)^{-1}$
 - $\text{pvalue} < \alpha$
- Conclusion :
 - β_j is significantly different of zero
 - X_j a une influence dans le modèle

Not true if there exists colinearity between the variables

Global significativity of the model

Test of the model with a risk α

$$H_0 : \beta_2 = \beta_3 = \dots = \beta_p = 0$$

$$H_1 : \exists j = 2, \dots, p, \beta_j \neq 0$$

Global significance of the model

Test of the model with a risk α

$$H_0 : \beta_2 = \beta_3 = \dots = \beta_p = 0$$

$$H_1 : \exists j = 2, \dots, p, \beta_j \neq 0$$

Statistics

$$F = \frac{n-p}{p-1} \frac{\|\hat{Y} - \bar{\hat{Y}}\|^2}{\|Y - \hat{Y}\|^2} \sim Fisher(p-1, n-p)$$

Global significativity of the model

Test of the model with a risk α

$$H_0 : \beta_2 = \beta_3 = \dots = \beta_p = 0$$

$$H_1 : \exists j = 2, \dots, p, \beta_j \neq 0$$

Statistics

$$F = \frac{n-p}{p-1} \frac{\|\hat{Y} - \bar{\hat{Y}}\|^2}{\|Y - \hat{Y}\|^2} \sim \text{Fisher}(p-1, n-p)$$

Remarque : $\frac{n-p}{p-1} \frac{\|\hat{Y} - \bar{\hat{Y}}\|^2}{\|Y - \hat{Y}\|^2} = \frac{SSE/(p-1)}{SSR/(n-p)}$ (E : Estimated ; R : Residuals)

Decision rule

- si $F_{obs} > q_{\alpha}^F$, H_0 is rejected, and there exist a coefficient which is not zero.

The regression is "useful"

- si $F_{obs} \leq q_{\alpha}^F$, H_0 is acceted, all the coefficients are supposed to be null

The regression is not "useful"

Global significance of the model

- Fisher Statistic

Global significance of the model

- Fisher Statistic
- Significativity test (bilateral)
 - $H_0 : \beta_2 = \dots = \beta_p = 0$
 - $H_1 : \exists \beta_j \neq 0$

Global significance of the model

- Fisher Statistic
- Significativity test (bilateral)
 - $H_0 : \beta_2 = \dots = \beta_p = 0$
 - $H_1 : \exists \beta_j \neq 0$
- Decision with a risk α , **Reject H_0 if**
 - si $\frac{n-p}{p-1} \frac{R^2}{1-R^2} > f_{p-1, n-p}(1 - \alpha)$
 - si $\text{pvalue} < \alpha$

→ The linear model has an added value

Global significance of the model

Remarque1 : sur la statistique de Fisher :

$$F = \frac{n-p}{p-1} \frac{R^2}{1-R^2}$$

The R^2 coefficient increase mechanically with the number of variables

Global significance of the model

Remarque1 : sur la statistique de Fisher :

$$F = \frac{n-p}{p-1} \frac{R^2}{1-R^2}$$

The R^2 coefficient increase mechanically with the number of variables

Remarque : the adjusted R^2 may be used

$$R_{adj}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p}$$

The R_{adj}^2 does not increase with the number of variables.

Boston Housing Data

The original data are $n = 506$ observations on $p = 14$ variables,

medv	being the target variable
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitric oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per USD 10,000
ptratio	pupil-teacher ratio by town
b	$1000(B - 0.63)^2$ where B is the proportion of blacks by town
lstat	percentage of lower status of the population
medv	median value of owner-occupied homes in USD 1000's

Boston Housing Data

Les données :

nř	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
...

Boston Housing Data

MCO sous R

```
library(mlbench)
#Data data(BostonHousing) tab=BostonHousing;names(tab)
target="medv"; Y=tab[,target]; X=tab[,names(tab)!=target]; names(X)
#MCO resfit=lsfit(x=X,y=Y,intercept=T);
resfit$coef hist(resfit$res)
```

Cst	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
36.45	-0.10	0.046	0.020	2.68	-17.76	3.80	0.00	-1.47	0.30	-0.01	-0.95			

Boston Housing Data

Modèle Linéaire sous R

`reslm=lm(medv ~ .,data=tab); summary(reslm)` Résultats :

$n = 506$, $p = 14$

Residuals:

Min	1Q	Median	3Q	Max
-15.595	-2.730	-0.518	1.777	26.199

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.646e+01	5.103e+00	7.144	3.28e-12	***
crim	-1.080e-01	3.286e-02	-3.287	0.001087	**
zn	4.642e-02	1.373e-02	3.382	0.000778	***
indus	2.056e-02	6.150e-02	0.334	0.738288	
chas1	2.687e+00	8.616e-01	3.118	0.001925	**
nox	-1.777e+01	3.820e+00	-4.651	4.25e-06	***
rm	3.810e+00	4.179e-01	9.116	< 2e-16	***
age	6.922e-04	1.321e-02	0.052	0.958229	
dis	-1.476e+00	1.995e-01	-7.398	6.01e-13	***
rad	3.060e-01	6.635e-02	4.613	5.07e-06	***
tax	-1.233e-02	3.760e-03	-3.280	0.001112	**
ptratio	-9.527e-01	1.308e-01	-7.283	1.31e-12	***
b	9.312e-03	2.686e-03	3.467	0.000573	***
lstat	-5.248e-01	5.072e-02	-10.347	< 2e-16	***

Signif. codes: 0 '***' 0.001/ '**' 0.01/ '*' 0.05/ '.' 0.1/ ' '

Residual standard error: 4.745 on 492 degrees of freedom

Multiple R-squared: 0.7406, Adjusted R-squared: 0.7338

F-statistic: 108.1 on 13 and 492 DF, p-value: < 2.2e-16

Précautions

- Multicolinéarité

la solution des MCO nécessite de calculer $(X^T X)^{-1}$.

Lorsque le déterminant de cette matrice est très proche de zéro, le problème est mal conditionné.

- Choix des variables

Le coefficient de détermination R^2 augmente en fonction du nombre de variables.

si $p = n$ $R^2 = 1$, ce qui n'est pas forcément pertinent.

Démonstration sous R