Linear Regression models & Regularization part1

Mathilde Mougeot

ensIIE & ENS Paris-Saclay, France

2025

Agenda

Lessons

- 3 plenatory lessons : CM1 CM2 CM3
- 3 Practical work sessions using R : TP1, TP2, TP3
- Agenda : CM1, TP1, CM2, TP2, CM3, TP3.

Before next Pratical session, install on your computer

1. R software, `https://www.r-project.org/`
2. Rstudio, `https://www.rstudio.com/`
3. Package Swift
   install.packages(swirl) ; library(swirl) ; swirl()
   → **ToDo before next TP** section 1 : R programming.
   1 : (basic) → 8 : (logic)

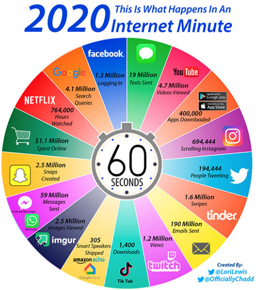
Documents are available (passwd : HMV2025

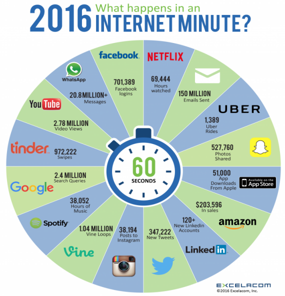
- `https://sites.google.com/site/MougeotMathilde/teaching`

A word on data and predictive models

Due to digitalization, data are available everywhere

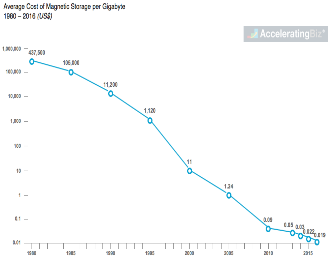
- Industry (sensors records. Ex. Temperature, Pressure ...)
- Finance : transactions... Marketing : consumer data.
  All Cell phone applications are recording data : (GPS, mail, musique ...)
- web data. Ex : social networks (GAFA data)

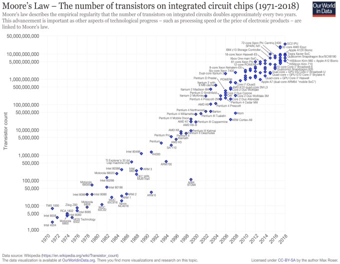




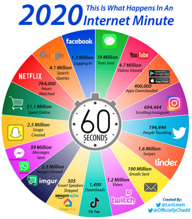
A word on data

Due to digitalization, data are stored and various applications propose decision making elements





A word on data and predictive models

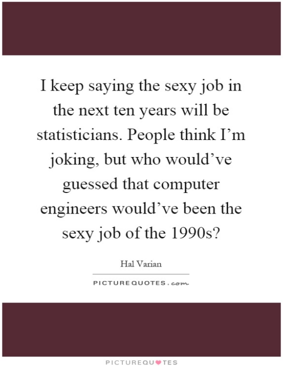


Nowadays, predictive models are crucial for monitoring & diagnosis

- Industry : health monitoring, Energy...
- Finance : forecast of the evolution of the market
- Marketing : ranking customers
- Health. Tele medicine

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

A word on data and Data Scientists



Hal Varian, Chief Economics at Google.

→ Statistical learning is a key ingredient in the training of a modern data scientist.

MRR...a first step !

- With the explosion of "Big data" problems, statistical learning has become a very hot field in many scientific areas.
- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.

Agenda for this first lesson

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 statistical model !
Gauss 1785 ; Legendre 1805.

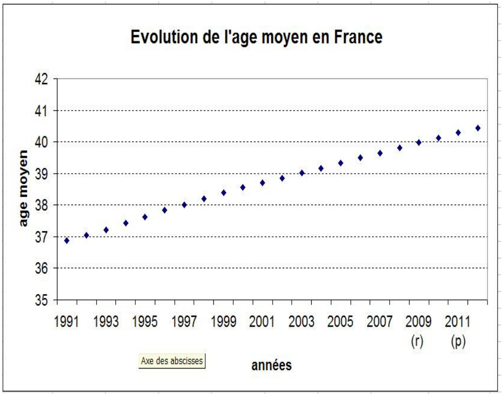
Outline of this first lesson

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

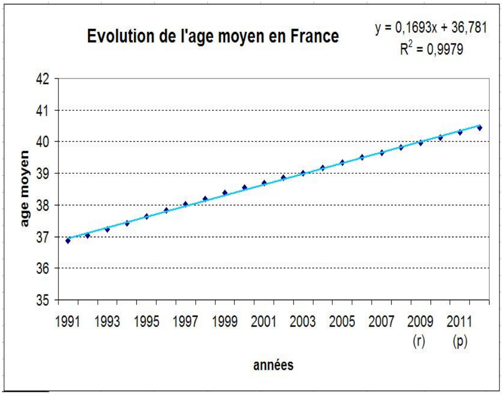
Evolution of the average age of the French population

	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

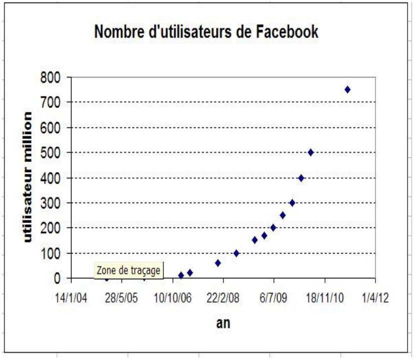


Application : Social Networks

Facebook users :

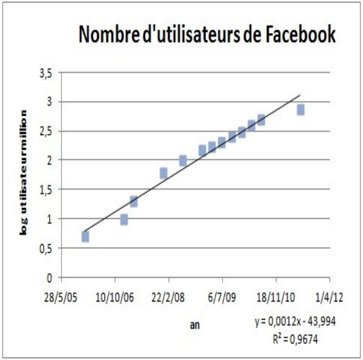
an	user(million)
31/12/04	0
31/12/05	5
31/12/06	10
31/3/07	20
30/12/07	60
30/6/08	100
30/12/08	150
30/3/09	170
30/6/09	200
30/9/09	250
30/12/09	300
30/3/10	400
30/6/10	500
30/6/11	750

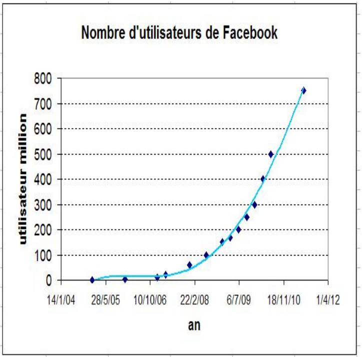
Evolution of the number of Facebook users



Motivations : investment, forecast

Modeling the evolution of the number of Facebook users





Introduction : Regression model

- (Y,X) : couple of variables
  Y : Target quantitative variable
  $X = (X_1, X_2, \ldots, X_p)$: Co-variates, quantitative variables

- The data set : $\mathcal{D}_n = \{(y_i, x_i), y_i \in \mathbb{R}, x_i \in R^p, 1 \leq i \leq n\}$

$\rightarrow$ The goal is to propose a Regression model to **explain Y given X**.

$\rightarrow$ The parameters of the model are computed using the set of data $\mathcal{D}_n$

$$Y \simeq \mathcal{F}_{\text{data set}}(X) \simeq \mathcal{F}_{\text{data set}}(X_1, \ldots, X_p)$$

In this case, $\mathcal{F}$ is a **linear function**.

- Potential questions concerning the modeling phenomena :
  - What are the performances of the regression model ?
  - What are the main explicative variables ?
  - What about predicting new values ? to forecast ?
  - Is-it possible to use an alternative model ? with less variables ?
  - Is-it possible to improve the model ?

Boston Housing Data

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

medv	median value. **Target variable (Y)**
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 :

n°	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

Different points of view may exist for studying the data with a model :

$$Y \simeq \mathcal{F}_{\textbf{data set}}(X)$$

- Mining approach
  - Evaluate the performances of the model
  - What are the most important variables ? (variable selection)
    $\rightarrow$ sparse models, less complex, best performances
- Predictive approach
  - Inference and simulation
    $\rightarrow$ Ponctual estimation for new values of the co-variables
    $\rightarrow$ 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**

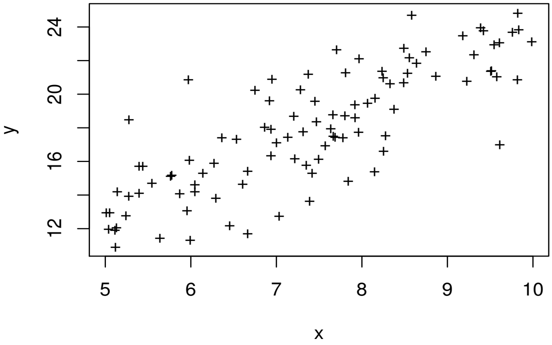
- Data set : $\mathcal{D}_n = \{(x_i, y_i)\ i = 1, ..., n,\ y_i \in \mathbb{R},\ x_i \in \mathbb{R}^p\}$

- Goal :
  - Modeling Y linearly using X with a "small" remaining $\boxed{\epsilon}$ term
  - $Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \boxed{\epsilon}$
  - Considering $X_1 = 1$, we write $Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \boxed{\epsilon}$
  - $Y = \sum_j^p \beta_j X_j + \boxed{\epsilon}$
    (p β_j parameters with "one constant/intercept" inside)

Ordinary Least Square (OLS)
Simple Linear Regression model

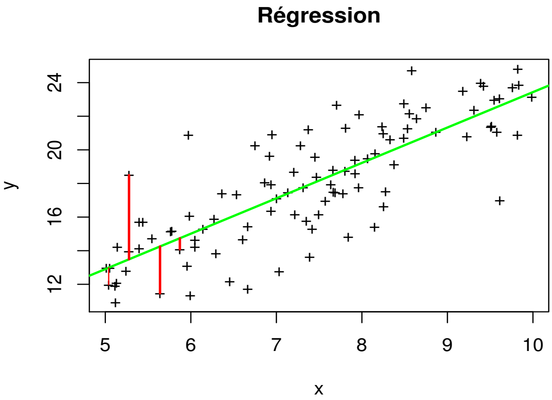
Simple Linear Regression : example

We only have one co-variate (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 $\sum_i^n (y_i - (\beta_1 + \beta_2 x_i))^2$



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 : $\sum_{i=1}^{n}(y_i - x_i\beta_2 - \beta_1)^2$
  $\rightarrow$ 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 :

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

OLS. Simple linear model. Several notations :

• OLS estimator (observations) :

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

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

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

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

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

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

The values estimated by OLS (the estimates) for $\beta = (\beta_1, \beta_2)^T$ verify :

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

Ordinary Least Square. Simple linear model

Theorem :

The OLS estimators $(\hat{\beta}_1, \hat{\beta}_2)$ for the simple linear model equals :

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

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

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 useful :

- $r(x, y)$: correlation coefficient/ coefficient de corrélation linéaire
$$r(x, y) = \frac{cov(X,Y)}{\sqrt{var(x)}\sqrt{var(y)}}$$
- $r(x, y) = 1$ if and only if $Y = aX + b$, linear relation between Y et X

R-square used in multiple regression

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

The Target and Variable point of view

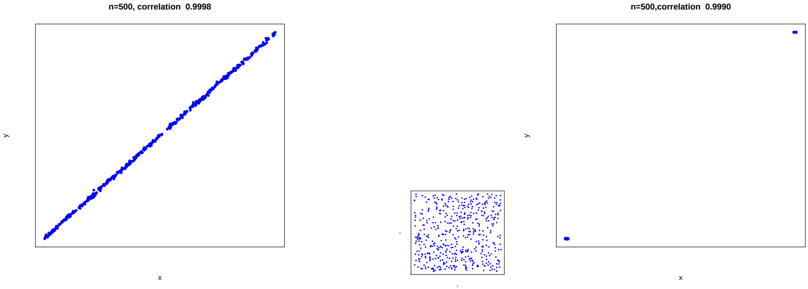
Target and co-variable play their own role !

- Model : $Y = \beta_1 + \beta_2 X + \epsilon$,
  $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$, $\hat{\beta}_2 = \frac{Cov(x,y)}{Var(x)} = r_{x,y} \frac{\sqrt{Var(y)}}{\sqrt{Var(x)}}$

- Model : $X = \alpha_1 + \alpha_2 Y + \epsilon$,
  $\hat{\alpha}_1 = \bar{x} - \hat{\alpha}_2 \bar{y}$, $\hat{\beta}_2 = \frac{Cov(x,y)}{Var(y)} = r_{x,y} \frac{\sqrt{Var(x)}}{\sqrt{Var(y)}}$

Correlation. A fake friend

The value of r may be sometimes non informative. Illustration with a joint distribution (y, x) with a correlation coefficient equaled to 0.999



$\rightarrow$ A scatter plot shows two very different distributions.
Always look at the data to detect suspicious points ! !

$\boxed{\text{Always study carefully the data}}$

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
  $S = \{(x_i, y_i)\ i = 1...n,\ y_i \in \mathbb{R}\ x_i \in \mathbb{R}^p\}$
- 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 : $X^T X$ **inversible**. ($n \geq p$)

Theorem :

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

MCO

- **Estimation.**

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

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

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

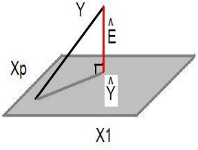
- P **Projection matrix** on the Hyperplan (hat matrix)

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

- **Residuals.**
  - $\hat{\epsilon} = Y - \hat{Y}$

No assumption on the law or distribution of the residuels ϵ

OLS. Geometrical interpretation



$$Y \quad = \quad \sum_j^p X^j \beta_j \quad + \quad \epsilon$$

Space	Space	Space
dim n	dim p	dim $(n-p)$

Ordinary Least Square. Properties

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

- Residual average :
  - $\sum_i \hat{\epsilon}_i = 0$ if there is an intercept in the model $X^1 = (1, 1, \dots, 1)$

  $\rightarrow$ 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)
    - $R^2 = \frac{var(\hat{Y})}{var(Y)}$, **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 $(\hat{Y} - \bar{\hat{Y}}1_{,n})$

- $var(\hat{E}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  $= (1 - R^2)var(Y)$, **unit of Y^2 !**

Ordinary Least Square. Study of the Adjustment

- R-square
  - $R^2 = \cos^2 \omega = \frac{Var \hat{Y}}{Var(Y)}$
  - $R^2 \in [0, 1]$
  - $R^2 =$ increases mechanically with the number of variables
  - reliable information in case of few variables, useless in high dimension.

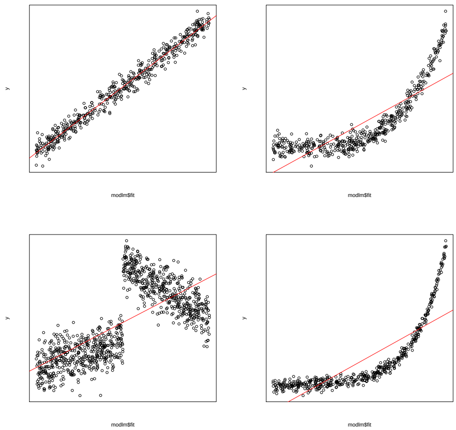
- Adjusted R-squared is sometimes preferred
  (see section 2 : penalization with the number of variables)
  - $R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p}$. R^2_{adj} may be negative.

- Residual study : **mandatory**
  - $\hat{\epsilon}_i = y_i - \hat{y}_i \qquad \forall i \in 1..n$
  - Vizualization of
    - $(\hat{\epsilon}_i, y_i)$ homoscedastic vs heteroscedastic model (or $(\hat{\epsilon}_i / S_E, y_i)$)
    - $(\hat{\epsilon}_i, i)$ $\forall i \in 1..n$

- Prediction study : **mandatory**
  - Vizualization of $(\hat{y}_i, y_i) \quad \forall i \in 1..n$. Comparison with the first bisector.

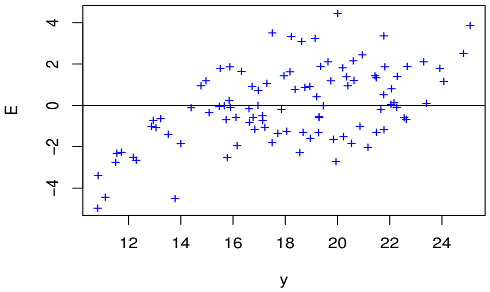
Ordinary Least Square : Quality of the Adjustement

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



Ordinary Least Square : Student Residual graph

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

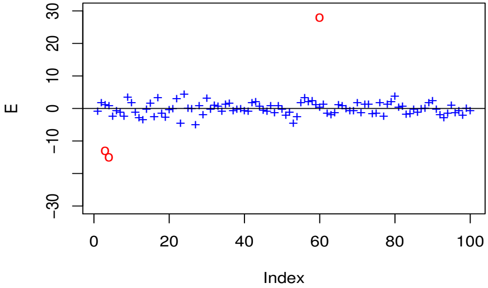


Residual graph

→ Random distribution. There is no information to be capture

Ordinary Least Square : Student Residual graph

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

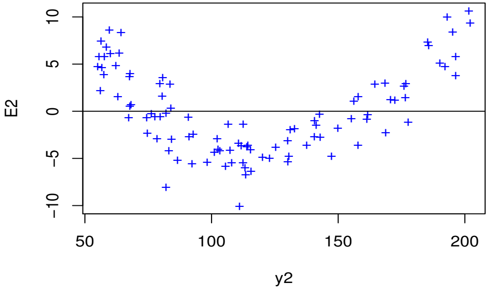


Residual graph function of Y

$\rightarrow$ **Large values for some points ? Outliers detection ?**

Ordinary Least Square : Student Residual graph

$\frac{\hat{\epsilon}_i}{\hat{S}_E} = \frac{y_i - \hat{y}_i}{\hat{S}_E}$ (with no unit)



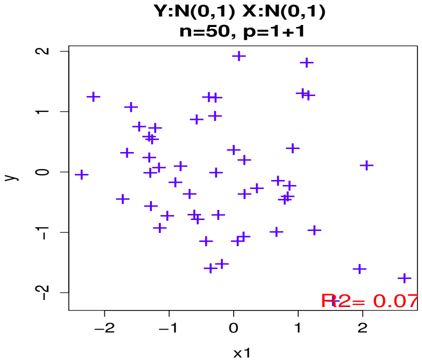
Graphe des résidus en fonction de Y

$\rightarrow$ there is still some information in the residuals.

$\rightarrow$ The model needs to be changed.

Ordinary Least Square : curse of dimension.

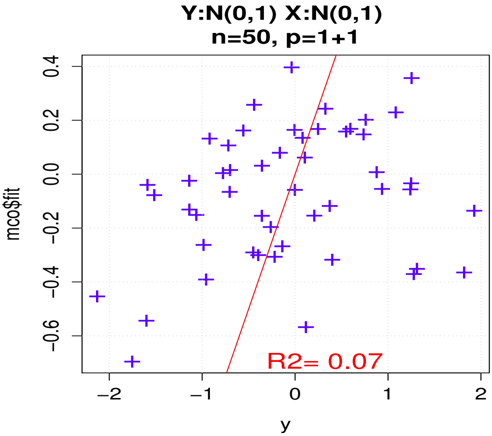
Data set : $\{(y_i, x_i) 1 \leq i \leq n\}$. One target variable, one covariable



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

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

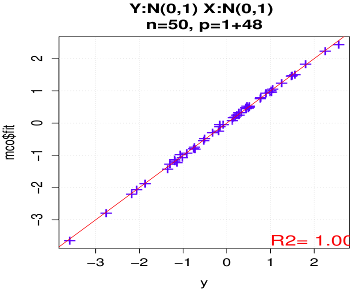
initial data and OLS line



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

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.

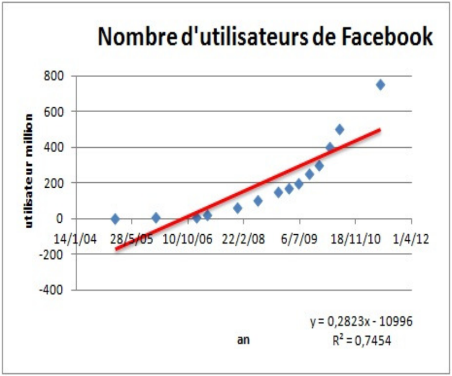


$\rightarrow R^2 = 0.99,\ R^2_{adj} = 0.93$

Despite the graph, the fit is completely arbitrary. No linear structure to catch !

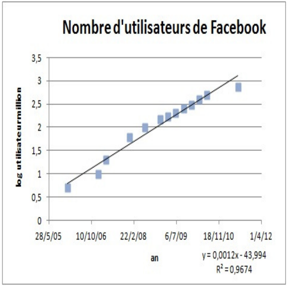
OLS. Transformation of the the initial model (1/2)

initial data set :

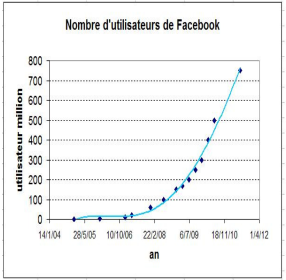


OLS. Transformation of the the initial model (2/2)

logarithmic transformation :
$Z = Log(Y)\ X = t,$
$Z = \alpha_1 + \alpha_2 t + \epsilon$



$\rightarrow \tilde{Y} = \exp\left(\hat{\alpha}_1 + \hat{\alpha}_2 t\right).$



MCO Regression. Some limits :

If $X^T X$ is non inversible

- $n >> p$, collinearity between some X_j.
  - Pseudo-inverse, the solution is not unique
  - Variable selection

- $p >> n$, 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^TX non inversible $\rightarrow$

Pseudo inverse computation.
Standard Value Decomposition (SVD) of X^TX

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

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

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

$\hat{\beta} = (X^TX)^{*-1}X^TY$

The solution non unique

Outline

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

Linear Model

Probabilistic assumption on the residuals

The probabilistic assumption on the residuals let to introduce :
- statistical tests on the value of the coefficients (to study if a variable is useful or not to explain the target),
-to test if the general liner model is useful (all coefficients equal zero).

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.
- Residual density & Maximum Likelihood Estimation (MLE)

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

- 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 ?

$Y = X\beta + \epsilon$ 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

$\rightarrow$ **let to compute confidence intervals for β and Y.**

$\rightarrow$ **let to test the parameters.**

Law of $\hat{\beta}$:

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

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

$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad and\ Y = X\beta + \epsilon$

- $\mathbb{E}(\hat{\beta}) = \beta$ Non biased estimator $\mathbb{E}(\hat{\beta}) - \beta = 0$
- $Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

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

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

Recall $Var(aY) = aVar(Y)a^T$

Law of $\hat{Y}$

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

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

- $\mathbb{E}(\hat{Y}) = X\beta$
  $\mathbb{E}(\hat{Y}) = \mathbb{E}(X\hat{\beta}) = X\mathbb{E}(\hat{\beta}) = X\beta = \mathbb{E}(Y)$
- $Var(\hat{Y}) = \sigma^2 \, X(X^T X)^{-1} \, X^T$

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

Law of $\hat{\epsilon}$

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

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 $X^T X$ inversible,

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 $\hat{\sigma}^2$

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

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

$||\hat{\epsilon}||^2 = \sum_i^n \hat{\epsilon}_i^2$ $\quad$ $||\hat{\epsilon}||^2$ follows a $\sigma^2\chi^2(n-p)$ law (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 $\hat{\sigma}^2$:
$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2(n-p) \qquad \rightarrow \qquad \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$

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 **unknown**

- 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 $(X^T X)^{-1}$
    - pvalue $< \alpha$
- Conclusion :
    - β_j is significatively different of zero
    - X_j is influent to explain the target variable

Not true if there exists colinearity between the variables

Global significativity of the model

Test of the model with a risk α

$$H_0 : \quad \beta_2 = \beta_3 = \ldots = \beta_p = 0$$
$$H_1 : \quad \exists j = 2, \ldots, p, \beta_j \neq 0$$

Statistics

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

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 $F_{obs} > q^F_\alpha$, H_0 is rejected, and there exist a coefficient which is not zero. **The regression is "useful"**
- si $F_{obs} \leq q^F_\alpha$, H_0 is acceted, all the coefficients are supposed to be null **The regression is not "useful"**

Global significativity of the model

- Fisher Statistics
- Significativity test (bilateral)
  - $H_0 : \beta_2 = \ldots = \beta_p = 0$
  - $H_1 : \exists \beta_j \neq 0$
- Decision with a rish 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 pvalue $< \alpha$

$\rightarrow$ The linear model has an added value

Global significativity of the model

Remarque1 : Fisher statistics :
$F = \frac{n-p}{p-1}\frac{R^2}{1-R^2}$
The R^2 coefficient increases mechanically with the number of variables

Remarque : the adjusted R^2 may be used

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

The R^2_{adj} does not increase ith 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 :

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

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
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

Linear models with R

code 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 / 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
```