Linear Regression (part 1)

Predictive Modeling & Statistical Learning

Gaston Sanchez

CC BY-SA 4.0

Linear Regression

Advertising Data from ISL

```r
# file in folder data/ of github repo
Advertising <- read.csv("data/Advertising.csv", row.names = 1)
```

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75.0	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2

(first 8 rows)

Advertising Data from ISL

Advertising consists of:

- the `Sales` of a product in 200 different markets
- the advertising budgets for three different media:
  - `TV`
  - `Radio`
  - `Newspaper`
- It is not possible to directly increase the sales of the product
- On the other hand, it is possible to control the advertising expenditure in each of the 3 media

Introduction

- Suppose we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$
- We assume there is some relationship between Y and $[X_1, \ldots, X_p]$. that can be written in a general form as

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

- f represents the systematic information that the predictors provide about Y
- ϵ represents an *error* term that is a catch-all for what we miss with the model

Data set `Advertising`

Response:

- Y: `Sales`

Predictors:

- X_1: `TV`
- X_2: `Radio`
- X_3: `Newspaper`

Relationship:

$$\texttt{Sales} = f(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper}) + \epsilon$$

Introduction

One possibility for $f()$ is a linear relationship of the form:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Introduction

One possibility for $f()$ is a linear relationship of the form:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

- It assumes a linear dependence of Y on the predictors
- $\beta_0, \beta_1, \ldots, \beta_p$ are unknown constants also known as the model *coefficients* or *parameters*
- The linearity is in the parameters (i.e. coefficients)

Linear relationship

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

$$\texttt{Sales} = \beta_0 + \beta_1 \ \texttt{TV} + \beta_2 \ \texttt{Radio} + \beta_3 \ \texttt{Newspaper} + \epsilon$$

Examples of linear models

A couple of examples of other possible linear models

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 log(X_2) + \beta_3 X_1 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 (X_1^{X_2}) + \varepsilon$$

Non-linear models

Some models are not linear in the parameters:

$$Y = \beta_0 + \beta_1 X_1^{\beta_2} + \varepsilon$$

Some relationships can be transformed to linearity, for example:

$$Y = \beta_0 X_1^{\beta_1} \varepsilon$$

can be linearized by taking logs (and reexpressing some of the parameters)

$$log(Y) = log(\beta_0) + \beta_1 log(X_1) + log(\varepsilon)$$

Introduction

The challenge involves finding parameter estimates denoted by

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$$

that provide the "best" approximation for Y:

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

or more commonly

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

Introduction

- Linearity is a BIG assumption.
- True regression functions are rarely linear.
- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Simple Linear Regression

Simple Linear Regression with one predictor

- Simple Linear Regression = Univariate regression
- One predictor variable X and one response variable Y

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Simple Linear Regression with one predictor

We assume a linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where:

- β_0 and β_1 are two unknown constants also known as *coefficients* or *parameters*
- β_0 represents the *intercept*
- β_1 represents the *slope*
- ε is a vector of error terms

Simple Linear Regression with one predictor

In vector notation:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}$$

where:

- $\mathbf{y}$ is the vector representing the response variable
- $\mathbf{x}$ is the vector representing the predictor variable
- $\boldsymbol{\varepsilon}$ is the vector representing the error term

Some vector-matrix notation

In matrix notation:

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times 2}{\mathbf{X}} \times \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

which can be represented by:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Simple Linear Regression with one predictor

Note that if the data is center (mean $= 0$)

$$\mathbf{y} = \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}$$

then there is no intercept term β_0

Some vector-matrix notation

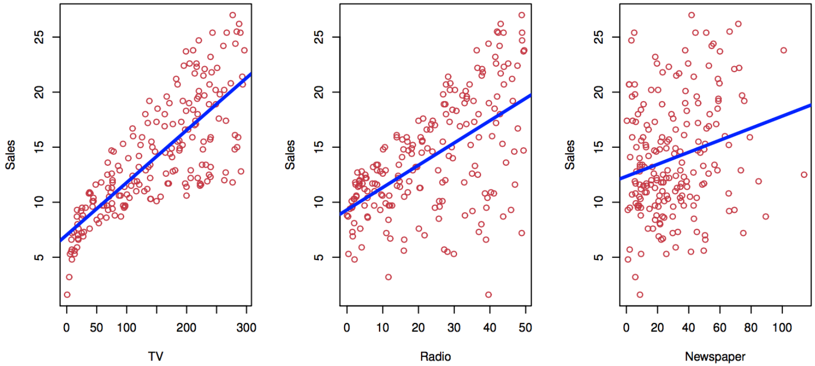
With centered data we have:

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times 1}{\mathbf{x}} \times \beta_1 + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

which can be represented by:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Various simple regressions



Simple Linear Regression with one predictor

Assuming the model

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\epsilon}$$

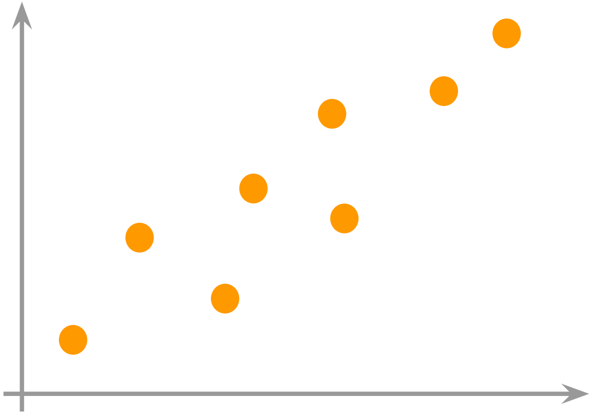
and given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}$$

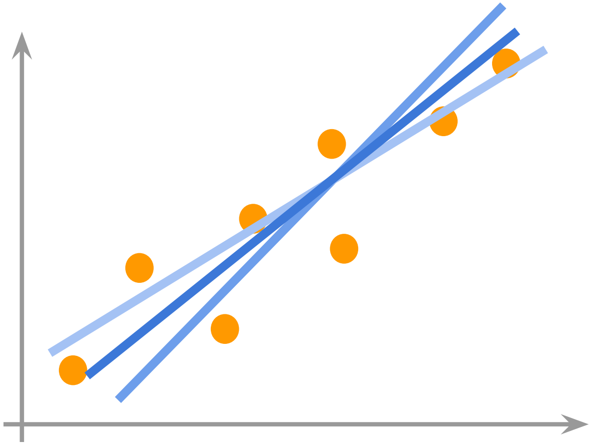
where $\hat{\mathbf{y}}$ indicates the predicted $\mathbf{y}$

Fitting a Line

Fitting a Line

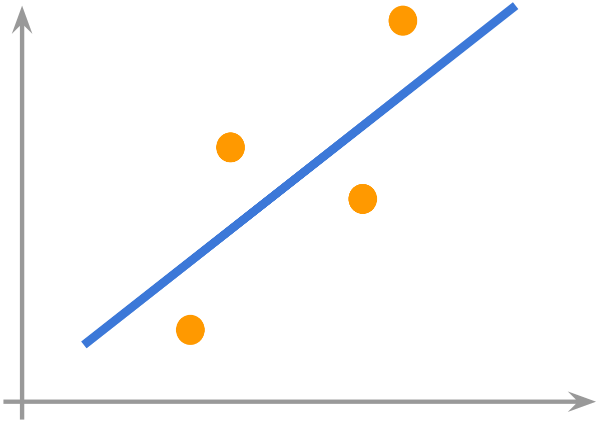


Fitting a Line

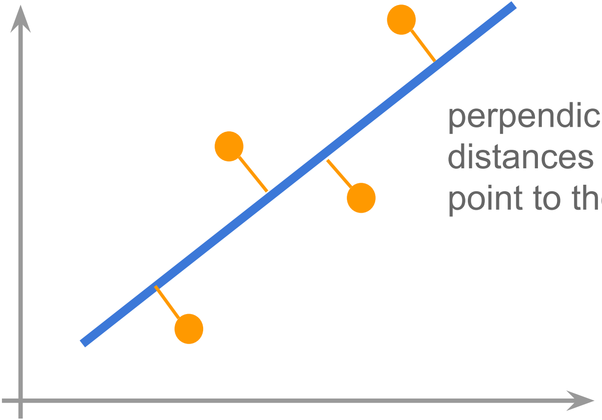


How to find the "best" fitting line?

Fitting a Line

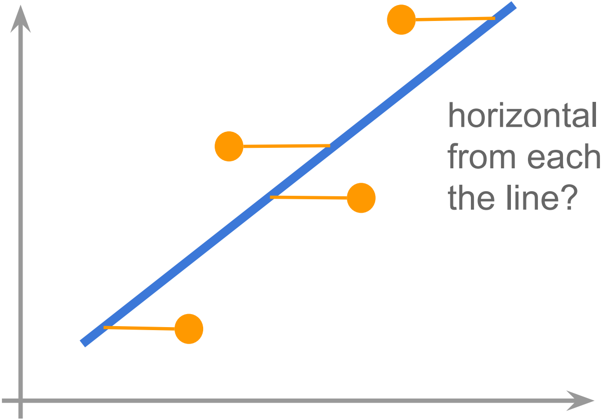


Fitting a Line



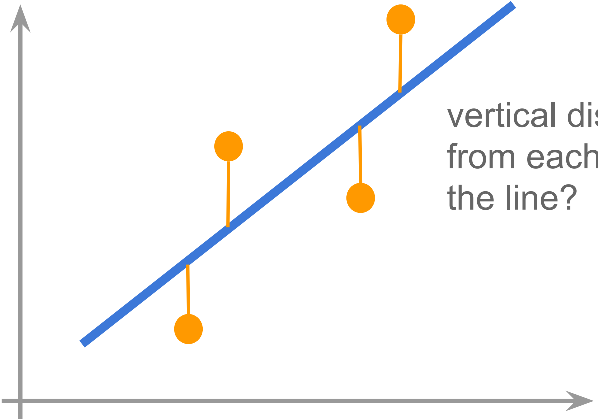
perpendicular distances from each point to the line?

Fitting a Line



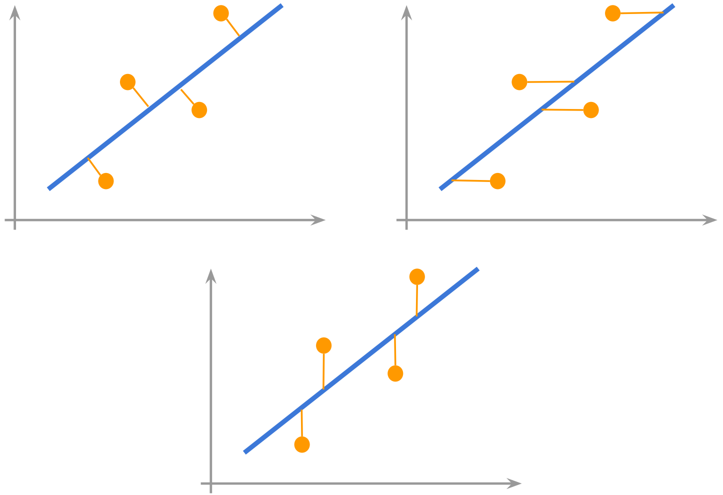
horizontal distances from each point to the line?

Fitting a Line

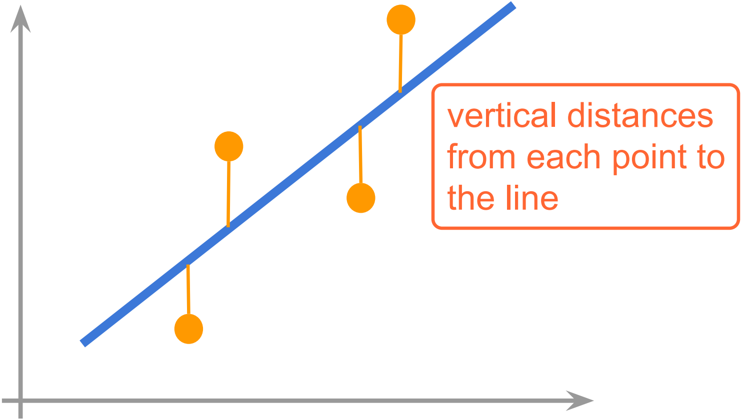


vertical distances from each point to the line?

Which Criterion?



Fitting a Line



Estimation of Parameters

Estimation of the parameters

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for $\mathbf{y}$ based on the ith value of $\mathbf{x}$

Estimation of the parameters

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for $\mathbf{y}$ based on the ith value of $\mathbf{x}$
- Then $e_i = y_i - \hat{y}_i$ represents the ith residual

Estimation of the parameters

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for $\mathbf{y}$ based on the ith value of $\mathbf{x}$
- Then $e_i = y_i - \hat{y}_i$ represents the ith residual
- We define the Residual Sum of Squares (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

- The **Least Squares** approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

Estimation of the parameters

The starting point is to write the model as:

$$\mathbf{e} = \mathbf{y} - (\beta_0 + \beta_1 \mathbf{x})$$

For convenience we define a quadratic loss function

$$L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

To minimize L we take partial derivatives with respect to each of the two parameters

Estimation of the parameters

Taking partial derivatives w.r.t each of the two parameters:

$$\frac{\partial L}{\partial \beta_0} = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

and

$$\frac{\partial L}{\partial \beta_1} = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

Estimation of the parameters

The solutions for β_0 and β_1 would be ontained by solving the so-called *normal equations*

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

and

$$\sum_{i=1}^{n}(x_i y_i - x_i \beta_0 - \beta_1 x_i^2) = 0$$

Estimation of the parameters by OLS

The **Least Squares** coefficients are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Estimation of the parameters by OLS

Notice that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

is equivalent to:

$$\hat{\beta}_1 = \frac{cov(\mathbf{x}, \mathbf{y})}{var(\mathbf{x})}$$

Example: Advertising Data

```r
# number of observations
n <- nrow(Advertising)

# model matrix
x <- Advertising$TV

# reponse variable
y <- Advertising$Sales
```

Example: Advertising Data

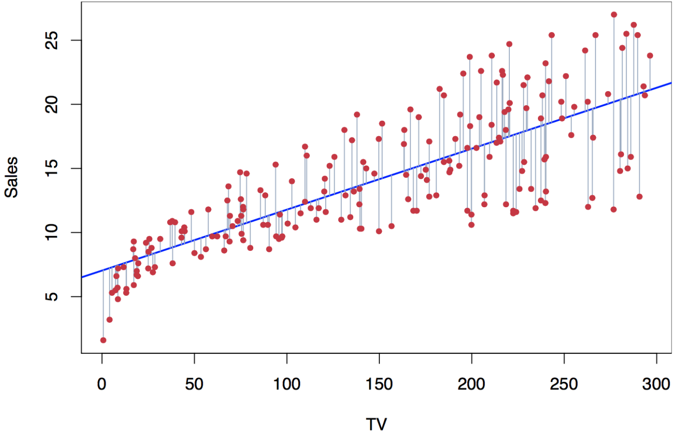
```
# slope
b1 <- cov(x, y) / var(x)
b1

## [1] 0.04753664

# intercept
b0 <- mean(y) - b1 * mean(x)
b0

## [1] 7.032594
```

Example: Advertising Data



The least squares fit for the regression of `Sales` on `TV`.

In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Another perspective

Projection

Notice that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Can be expressed in vector notation as:

$$\hat{\beta}_1 = \frac{\mathbf{y}^\intercal \mathbf{x}}{\mathbf{x}^\intercal \mathbf{x}}$$

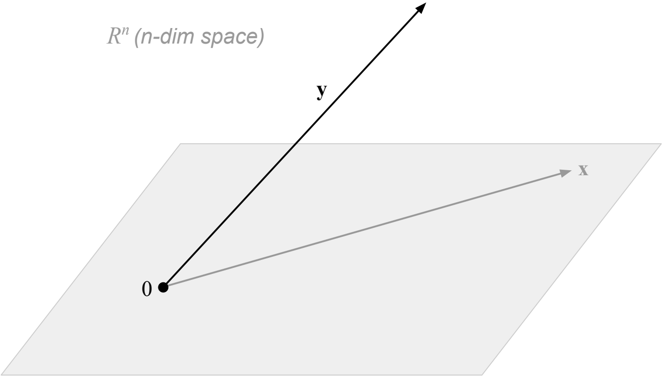
with $\mathbf{x}$ and $\mathbf{y}$ mean-centered.

Projection

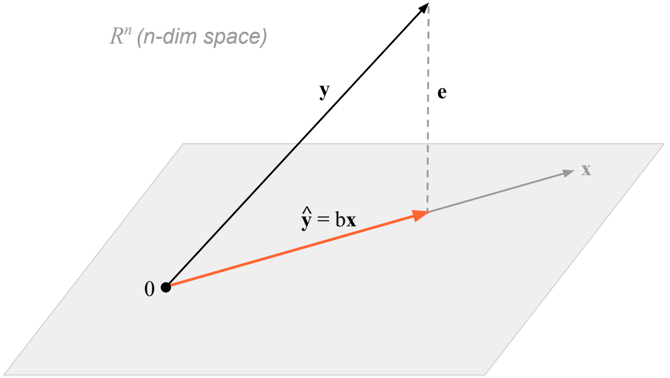
Thus, with centered variables $\mathbf{x}$ and $\mathbf{y}$, the fitted values $\hat{\mathbf{y}}$ are given by:

$$\begin{aligned}\hat{\mathbf{y}} &= \hat{\beta}_1 \mathbf{x} \\ &= \left(\frac{\mathbf{y}^\intercal \mathbf{x}}{\mathbf{x}^\intercal \mathbf{x}}\right) \mathbf{x} \\ &= \mathbf{x} \left(\frac{\mathbf{y}^\intercal \mathbf{x}}{\mathbf{x}^\intercal \mathbf{x}}\right) \\ &= \mathbf{x}(\mathbf{x}^\intercal \mathbf{x})^{-1} \mathbf{x}^\intercal \mathbf{y}\end{aligned}$$

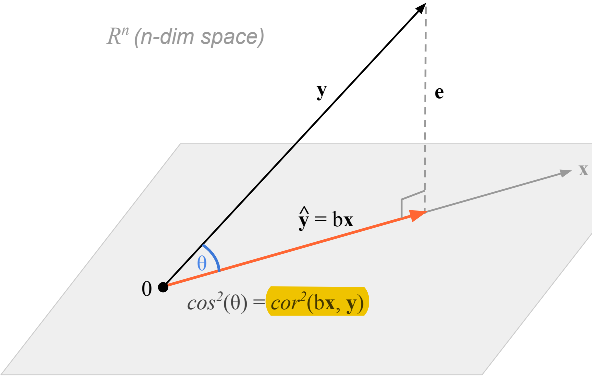
From variables perspective



From variables perspective



From variables perspective



Some Remarks

- There is nothing in the Least Squares method that requires statistical inference: formal tests of null hypotheses or confidence intervals.
- In its simplest form, regression analysis can be performed without statistical inference.
- The inferential part can sometimes be very useful but goes beyond the definition of a regression analysis.

Some Comments

- Linear Regression is a "simple" approach to supervised learning.
- Don't get fooled by the word "simple".
- "simple" $\neq$ easy / boring / uninteresting.
- I will use the terms *Regression Analysis* and *Regression Model* interchangeably.

References

- **Linear Models with R** by Julian Faraway (2015). CRC Press.
- **Modern Regression Methods** by Thomas Ryan (1997). Wiley.
- **Modern Multivariate Statistical Techniques** by Julian Izenman (2008). Springer.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 11: Classification and prediction methods*. Wiley.

References (French Literature)

- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 17: La regression multiple et le modele lineaire general*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 5: La Regression Multiple*. Dunod, Paris.
- **Regression avec R** by Cornillon and Matzner-Lober (2011). Springer.
- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 3, section 3.2: Regression multiple, modele lineaire*. Dunod, Paris.
- **Traitement des donnees statistiques** by Lebart et al. (1982). *Unit 3: Modele Lineaire*. Dunod, Paris.