# MULTILINEAR REGRESSION

—

**Ariane FERREIRA / Philippe CASTAGLIOLA**

Ecole des Mines de Nantes & IRCCyN, Nantes, France

aprosa@emn.fr

# Introduction

- The *response* $y$ of a system is supposed to depend on $k$ controlable input variables $\xi_1, \ldots, \xi_k$ called *natural variables*

$$y = f(\xi_1, \ldots, \xi_k) + \varepsilon$$

- $f()$ is the true unknown response function.
- $\varepsilon$ (error) is a rv verifying $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$.
- $\xi_1, \ldots, \xi_k$ are expressed in natural units : pressure in Pascal, temperature in Celcius, etc.
- In practice, it is impossible to know what the true response function is.

# Introduction

- Idea: approximate the unknown true response function $f()$ with a “simple” known function

$$y = a_0 + a_1 x_1(\xi_1, \ldots, \xi_k) + \cdots + a_{p-1} x_{p-1}(\xi_1, \ldots, \xi_k) + \varepsilon$$

- $x_1(\xi_1, \ldots, \xi_k), \ldots, x_{p-1}(\xi_1, \ldots, \xi_k)$ are $p-1$ functions of the natural variables $\xi_1, \ldots, \xi_k$ called *regression variables*.
- We simply denote $x_1, \ldots, x_{p-1}$.
- $a_0, \ldots, a_{p-1}$ are the $p$ *regression coefficients*.

# Examples with $k = 2$

- Linear model ($p = 1 + k$ regression coefficients)

$$y = a_0 + a_1 \underbrace{\xi_1}_{x_1} + a_2 \underbrace{\xi_2}_{x_2} + \epsilon$$

- Linear model plus interactions ($p = 1 + k(k+1)/2$ regression coefficients)

$$y = a_0 + a_1 \underbrace{\xi_1}_{x_1} + a_2 \underbrace{\xi_2}_{x_2} + a_3 \underbrace{\xi_1 \xi_2}_{x_3} + \epsilon$$

- Quadratical model ($p = 1 + k(k+3)/2$ regression coefficients)

$$y = a_0 + a_1 \underbrace{\xi_1}_{x_1} + a_2 \underbrace{\xi_2}_{x_2} + a_3 \underbrace{\xi_1^2}_{x_3} + a_4 \underbrace{\xi_2^2}_{x_4} + a_5 \underbrace{\xi_1 \xi_2}_{x_5} + \epsilon$$

# Coded variables

- The natural variables $\xi_1, \ldots, \xi_k$ are often defined on very different scales.
- $\longrightarrow$ we can use dimensionless *coded variables* $u_1, \ldots, u_k$

$$u_j = \frac{\xi_j - \xi_j^{\min}}{\xi_j^{\max} - \xi_j^{\min}} \Rightarrow u_j \in (0, 1)$$

or

$$u_j = \frac{\xi_j - (\xi_j^{\max} + \xi_j^{\min})/2}{(\xi_j^{\max} - \xi_j^{\min})/2} \Rightarrow u_j \in (-1, 1)$$

- $\xi_j^{\min}$ et $\xi_j^{\max}$ are the minimum and maximum values $\xi_j$ can take.

# Estimation of the regression coefficients

- We have to perform $n > p$ experiments.
- For the $i$th experiment
  - we set up the $\xi_{i,1}, \ldots, \xi_{i,k}$.
  - we observe the value for $y_i$
  - we compute the $x_{i,1}, \ldots, x_{i,p-1}$.

| Experiments | Natural variables | | | Regression variables | | | Responses |
|---|---|---|---|---|---|---|---|
| $1$ | $\xi_{1,1}$ | $\ldots$ | $\xi_{1,k}$ | $x_{1,1}$ | $\ldots$ | $x_{1,p-1}$ | $y_1$ |
| $2$ | $\xi_{2,1}$ | $\ldots$ | $\xi_{2,k}$ | $x_{2,1}$ | $\ldots$ | $x_{2,p-1}$ | $y_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $\xi_{n,1}$ | $\ldots$ | $\xi_{n,k}$ | $x_{n,1}$ | $\ldots$ | $x_{n,p-1}$ | $y_n$ |

# Estimation of the regression coefficients

- We get the following system of $n$ equations and $p$ unknowns

$$
\begin{array}{rcl}
y_1 & = & a_0 + a_1 x_{1,1} + a_2 x_{1,2} + \ldots + a_{p-1} x_{1,p-1} + \epsilon_1 \\
\vdots & \vdots & \vdots \\
y_n & = & a_0 + a_1 x_{n,1} + a_2 x_{n,2} + \ldots + a_{p-1} x_{n,p-1} + \epsilon_n
\end{array}
$$

- Matrix equation

$$
\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,p-1} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} a_0 \\ \vdots \\ a_{p-1} \end{pmatrix}}_{\mathbf{a}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}
$$

## Estimation of the regression coefficients

- The coefficients $a_0, \ldots, a_{p-1}$ can be obtained by minimising

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\mathbf{a})^T (\mathbf{y} - \mathbf{X}\mathbf{a})$$

- The solution is

$$\hat{\mathbf{a}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{C} \mathbf{X}^T \mathbf{y}$$

- $\mathbf{C} = (\mathbf{X}^T \mathbf{X})^{-1}$ is a symmetrical $(p, p)$ matrix.

## Estimation of the regression coefficients

- The *predicted* model is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{a}} = \mathbf{X}\mathbf{C}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$$

- $\mathbf{H} = \mathbf{X}\mathbf{C}\mathbf{X}^T$ is a $(n, n)$ symmetrical matrix verifying $\mathbf{H}^2 = \mathbf{H}$.

- Statistical properties of $\hat{\mathbf{a}}$

$$\begin{aligned} E(\hat{\mathbf{a}}) &= \mathbf{a} \\ V(\hat{\mathbf{a}}) &= \sigma^2\mathbf{C} \end{aligned}$$

## Vector of residuals

- The *vector of residuals* is

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

- We can show that

$$\mathbf{e} = (I - \mathbf{H})\mathbf{y}$$

- Orthogonal properies of $\mathbf{e}$

$$\begin{aligned}\mathbf{1}^T\mathbf{e} &= 0 \\ \hat{\mathbf{y}}^T\mathbf{e} &= 0 \\ \mathbf{X}^T\mathbf{e} &= \mathbf{0}\end{aligned}$$

# Definition of $SST$, $SSR$ and $SSE$

- Mean value of the response variable

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{\mathbf{1}^T\mathbf{y}}{n}$$

- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{y}^T\mathbf{y} - \frac{(\mathbf{1}^T\mathbf{y})^2}{n}$
- $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\mathbf{a}}^T\mathbf{X}^T\mathbf{y} - \frac{(\mathbf{1}^T\mathbf{y})^2}{n}$
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}^T\mathbf{y} - \hat{\mathbf{a}}^T\mathbf{X}^T\mathbf{y}$

## Definition of $MSR$ and $MSE$

- We notice that $SST = SSR + SSE$.
- From $SSR$ and $SSE$ we can define

$$\begin{aligned} MSR &= \frac{SSR}{p-1} \\ MSE &= \frac{SSE}{n-p} \end{aligned}$$

- We can proove that, **if the regression model is valid**, then $MSE$ is an unbiased estimator of $\sigma^2$

## Hypothesis tests for the regression coefficients

- Question: is there a linear relationship between the response $y$ and *at least* one regression variable?

$$H_0: \quad \forall j,\, a_j = 0$$
$$H_1: \quad \exists j,\, a_j \neq 0$$

- The hypothesis $H_0 : \forall j,\, a_j = 0$ is rejected if

$$1 - F_F\left(\frac{MSR}{MSE}, p-1, n-p\right) \leq \alpha$$

## Hypothesis tests for the regression coefficients

- Question: does the regression variable $x_j$ contribute to regression model in a significant way?

$$H_0: \quad a_j = 0$$

$$H_1: \quad a_j \neq 0$$

- The hypothesis $H_0 : a_j = 0$ is rejected if

$$2\left\{1 - F_T\left(\left|\frac{\hat{a}_j}{\sqrt{c_{jj} MSE}}\right|, n-p\right)\right\} \leq \alpha$$

- $c_{jj}$ is the diagonal entry of the matrix $\mathbf{C}$ corresponding to $\hat{a}_j$.

## Confidence intervals

- Confidence interval for the regression coefficient $a_j$

$$\begin{aligned}(a_j)_L &= \hat{a}_j - F_T^{-1}(1-\alpha/2, n-p)\sqrt{c_{jj}MSE} \\ (a_j)_U &= \hat{a}_j + F_T^{-1}(1-\alpha/2, n-p)\sqrt{c_{jj}MSE}\end{aligned}$$

- Confidence interval for the predicted model $\mathbf{x}^T\mathbf{a}$

$$\begin{aligned}(\mathbf{x}^T\mathbf{a})_L &= \hat{y} - F_T^{-1}(1-\alpha/2, n-p)\sqrt{MSE\mathbf{x}^T\mathbf{C}\mathbf{x}} \\ (\mathbf{x}^T\mathbf{a})_U &= \hat{y} + F_T^{-1}(1-\alpha/2, n-p)\sqrt{MSE\mathbf{x}^T\mathbf{C}\mathbf{x}}\end{aligned}$$

with $\mathbf{x}^T = (1, x_1, \ldots, x_{p-1})$.

# Multiple determination coefficient

- The *multiple determination coefficient* $R^2 \in (0, 1)$ is

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- The closer $R^2$ to 1, the better the regression model.
- But adding one extra regression variable always increases $R^2$.
- Definition of an *ajusted* coefficient $R_a^2$

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)$$

# Other vectors of residuals

- The *vector of standardized residuals* $\mathbf{d}$ is

$$\mathbf{d} = \frac{\mathbf{e}}{\sqrt{MSE}}$$

- The *vector of studentized residuals* $\mathbf{r}$ is

$$r_j = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

where $h_{ii}$ is the $i$th diagonal entry of the matrix $\mathbf{H}$.

- If $2\Phi(-|d_i|) \leq \alpha$ or $2\Phi(-|r_i|) \leq \alpha$ then
  - $y_i$ is an outlier.
  - or $y_i$ is in a region where the predicted model is not very good.

# Influent observations

- Goal: detect observations which strongly controls the model.
- We can decide that the $i$th observation is influent if

$$h_{ii} > \frac{2p}{n}$$

or if the *Cook's distance*

$$D_i = \frac{(\hat{\mathbf{a}}_{(i)} - \hat{\mathbf{a}})^T \mathbf{X}^T \mathbf{X} (\hat{\mathbf{a}}_{(i)} - \hat{\mathbf{a}})}{p\hat{\sigma}^2} = \frac{r_i^2 h_{ii}}{p(1 - h_{ii})} > 1$$

# Use of repeated observations

- We perform $m$ different experiments.
- For the $i$th experiment $(\xi_{i,1}, \ldots, \xi_{i,k})$ we have $n_i$ values $y_{i,1}, \ldots, y_{i,n_i}$ for the response.
- The total number of observed responses is $n = n_1 + \cdots + n_m$.
- Decomposition of $SSE$

$$SSE = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{i,j} - \hat{y}_i)^2 = \underbrace{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2}_{SSPE} + \underbrace{\sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2}_{SSLOF}$$

# Use of repeated observations

- We define

$$MSLOF = \frac{SSLOF}{m-p}$$
$$MSPE = \frac{SSPE}{n-m}$$

- We can proove that $MSPE$ is an unbiased estimator of $\sigma^2$, *independant* of the regression model.
- The regression model can be rejected if

$$1 - F_F\left(\frac{MSLOF}{MSPE}, m-p, n-m\right) \leq \alpha$$