# Probability and Statistics 2 (GCED)
# Linear Model

Marta Pérez-Casany and Jordi Valero Bayà

Department of Statistics and Operations Research
Technicat University of Catalonia

Facultat d'Informàtica de Barcelona, First Semester 2019

# 0. Some real situations

We are interested in determining the **insulin production** of pancreatic tissue at different **glucose concentrations**

| Contentration | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.53 | 1.61 | 3.75 | 2.89 | 3.26 | 2.83 | 2.86 | 2.59 |
| 2 | 3.15 | 3.96 | 3.59 | 1.89 | 1.45 | 3.49 | 1.56 | 2.44 |
| 3 | 3.89 | 4.8 | 3.69 | 5.7 | 5.62 | 5.79 | 4.75 | 5.33 |
| 4 | 8.18 | 5.64 | 7.36 | 5.33 | 8.82 | 5.26 | 8.75 | 7.10 |
| 5 | 5.86 | 5.46 | 5.69 | 6.49 | 7.81 | 9.03 | 7.49 | 8.98 |

Question: Does the insulin production depend on the glucose concentration?

We want to study the Cadmium concentration in liver, kidney and pancreas of different fish.

Two groups of fishes depending of if the water where they live has: 1) normal concentrarion of cadmium, 2) high concentrarion of cadmium.

| Fish Id. | Group | Liver | Kidney | Pancreas |
|---|---|---|---|---|
| 1 | A | 0.38 | 0.09 | 0.65 |
| 2 | B | 0.14 | 0.36 | 0.9 |
| 3 | B | 0.18 | 0.29 | 0.34 |
| 4 | A | 0.24 | 0.19 | 0.43 |

Questions:

- Are there differences in cadmium concentration in the two groups?
- Is the cadmium absorved for the different organs in a similar way?

In a database information system that allows its users to search backward for several days, wanted to depvelop a formula to predict the time it would take to search as a function of the days.

| Observation Id. | Time | Days |
|---|---|---|
| 1 | 0.65 | 1 |
| 2 | 0.79 | 2 |
| 3 | 1.36 | 4 |
| 4 | 2.26 | 8 |
| 5 | 3.59 | 16 |
| 6 | 5.39 | 25 |

Questions: Is it possible to predict the search time when the number of days is known?

# General objective

All these situations have in common that one is interested in describing the behaviour of a r.v. $Y$ known as **dependent variable** as a funcion of some other variables known as: **independent, explicatives or covariates**, also **factors** if they are cathegorical.

The explicative variables will represent, in most of the cases, **the experimental conditions**.

# DEFINITION OF LINEAR MODEL

# 1. Definition of Linear Model

**Objetive:** to explain the behaviour of a r.v. $Y$ as a function of $X_1, X_2, \cdots, X_{p-1}$ .

Given $n \in \mathbb{Z}^+$, $\forall i \in \{1, 2, \cdots, n\}$ let $Y_i$ be the variable related to $Y$ when $X_1 = x_{i1}, X_2 = x_{i2}, \cdots, X_{p-1} = x_{ip-1}$, where $x_{ik} \in \mathbb{R}, \forall i, j$.

**Definition:**

$$\forall i \quad , Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{i(p-1)}\beta_{p-1} + e_i = \mu_i + e_i$$

**Hipothesis:**

- $\forall i \in \{1, 2, \cdots, n\}$, $e_i \sim \mathrm{N}(0, \sigma_i^2)$;
- $\forall i \in \{1, 2, \cdots, n\}$, $\sigma_i^2 = \sigma^2$ (homocedasticity);
- $\forall i, j \in \{1, 2, \cdots, n\}$ $i \neq j$, $e_i$ indep.of $e_j$.
- $X$ values are fixed or, if random, are independent of errors.

$\beta_0$ is known as **intercept**.

In matrix form,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{n(p-1)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$\begin{array}{ccccc} \downarrow & \downarrow & \downarrow & \cdots & \downarrow \\ X_1 & X_2 & X_3 & \cdots & X_{p-1} \end{array}$$

Defining $Y_{n\times 1} = (Y_1, Y_2, \cdots, Y_n)^t$, $X_{n\times p} = (x_{ij})$, $\beta_{p\times 1} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_{p-1})^t$, $e_{n\times 1} = (e_1, e_2, \cdots, e_n)^t$, the model is written as:

$$Y = X\ \beta + e \Longleftrightarrow \mu = E(Y|X) = X\ \beta$$

$$Y|X \sim N(X\beta, \sigma^2 \cdot Id_n)$$

# Examples of linear models

The models used in the **analisis of variance** are LM with cathegorical coveriates.

**Example:** One wants to compare the *blood preasure* $(Y)$ in two types of individuals, those that have taken an special medication and those that have not.

$$Y_{ij} = \mu_i + e_{ij}, \ \forall i \in \{1,2\}, \ \forall j \in \{1,2,\cdots,n_i\};$$

in matrix form,

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{1n_1} \\ \vdots \\ e_{2n_2} \end{pmatrix}$$

The models known as **regresion models** are also a particular case of LM. In this case the covariates are continuous or discrete not cathegorical.

**Example:** One wants to study the level of a chemical agent in a plant ($Y$) as a function of the presence of this chemical on the floor ($X$).

$$Y_i = \beta_0 + x_i\beta_1 + e_i,\ \ i = 1, \cdots, n;$$

in matrix form,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

The models known as **Analysis of Covariance** are linear models in which the regression coefficients change by changing the levels of a cathegorical variables.

**Example:** One whants to study the levels of a given drug ($Y$) as a function of the dose ($X_1$). Moreover one has also consider the gender, since it is though that the efect may change depending on the gender ($X_2$).

$$Y_{ij} = \beta_{0i} + x_{ij}\beta_{1i} + + e_{ij}, \quad i \in \{1, 2\}, j \in \{1, 2, \cdots, n_i\}$$

in matrix form,

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ 1 & x_{12} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 \\ 0 & 0 & 1 & x_{21} \\ 0 & 0 & 1 & x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{02} \\ \beta_{12} \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{1n_1} \\ e_{21} \\ \vdots \\ e_{2n_2} \end{pmatrix}$$

# 2. The Null model

The null model is the one with just one parameter (the most simple model):

$$Y_i = \beta_0 + e_i, \cdots i = 1, \cdots, n$$

in matrix form:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} \beta_0 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

It is equivalent to study one sample from a r.v.

$\beta_0$ is denoted as *intercept* and we usually will consider models with intercept, in order that they contain the Null model as a submodel.

EXERCICE: Write in matrix form the models corresponding to the data of the following situations:

- Insulin production
- Cadmium concentration
- Time of searching

# PARAMETER ESTIMATION

# 3. Parameter vector estimation: Least Squares

Let $y = (y_1, y_2, \cdots y_n)^t$ be a realization of $Y$ and $\hat{\beta}$ a $\beta$ estimation.

1) **Minimum least square** estimation minimizes:

$$S(\beta) = ||y - \hat{y}||_2^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left( y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j \right)^2;$$

where $\hat{y} = \hat{\mu} = X\hat{\beta}$. Solution: $\hat{\beta} = (X^tX)^{-1}X^ty$, if $X^tX$ is not a singular matrix

2) **Weighted least squares** minimizes:

$$S(\beta) = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} w_i \left( y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2;$$

where $w_i^{-1} = Var(Y_i)$.

Solution: $\hat{\beta} = (X^tV^{-1}X)^{-1}X^tV^{-1}y$, if $X^tV^{-1}X$ is not a singular matrix, being $V = diag(w_i)$

Observation: No probability distribution for vector $Y$ is required

# 3. Parameter vector estimation: Maximum Likelihood

Esstimates maximize:

$$L(\beta;y)=(\sqrt{2\Pi}\sigma)^{-n}exp\Big(-\sum_{i=1}^{n}\frac{(y_i-\sum_{j=0}^{p-1}x_{ij}\beta_j)^2}{2\sigma^2}\Big);$$

which is equivalent to

$$l(\beta;y)=-n\log(\sqrt{2\Pi}\sigma)-\sum_{i=1}^{n}\frac{(y_i-\sum_{j=0}^{p-1}x_{ij}\beta_j)^2}{2\sigma^2}.$$

Let us define

$$U_j=\frac{\partial l}{\partial\beta_j}=\frac{1}{\sigma^2}\big(X^t(Y-X\beta)\big)_j\ \forall j.$$

The vector $U=(U_1,U_2,\cdots U_{p-1})^t$ is called **score vector**.

$$U_j=0\ \forall j\iff X^tY=X^tX\beta\iff\hat{\beta}=(X^tX)^{-1}X^tY;$$

if the rank of $(X^tX)$ is equal to $=p$. $\hat{\beta}$ es U.M.V.U.E.

# PREDICTED AND RESIDUALS

# 4. Predicted values and raw residuals

The vector or predicted values is

$$\hat{Y} = X\hat{\beta} = (\hat{Y}_1, \hat{Y}_2, \cdots \hat{Y}_n)^t$$

If $(y_1, y_2, \cdots y_n)^t$ is a realization of vector $Y$, the raw residual of observation $y_i$ is equal to:

$$\text{raw residual} = y_i - \hat{y}_i = \hat{e}_i$$

# Geometrical interpretation of the Residual vector

$\hat{e} = Y - X\hat{\beta}$ is orthogonal to the columns of matrix $X$.

$$\begin{aligned} X^t\hat{e} &= X^t(Y - X\hat{\beta}) = X^t(Y - X(X^tX)^{-1}X^tY) \\ &= X^tY - X^tX(X^tX)^{-1}X^tY = X^tY - X^tY = 0 \end{aligned}$$

dibuix.pdf

# VARIANCE ESTIMATION

# 5. Residual variance estimation: Moment method

**Moment method** Assuming that $p = rang(X^t X)$, i.e that $X^t X$ has rang maximum, it is verified that:

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \sim \chi^2_{n-p}$$

Thus, equating the r.v. to its mean value, one has that:

$$E(\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2) = n - p \;\; \Leftrightarrow \;\; E(S^2) = \sigma^2$$

where

$$S^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \;\; .$$

Thus, $\mathbf{S^2}$ is an Unbiased estimator of $\sigma^2$ and it is also U.M.V.U.E. This estimator is known as **mean square error**.

# 5. Residual variance estimation: Maximum Likelihood

The log-likelihood function as a function og $\sigma$ is equal to:

$$l(\sigma^2;\mu) = -n\log(\sqrt{2\Pi\sigma^2}) - \sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{2\,\sigma^2};$$

differentiating and equation to zero one has that:

$$\frac{\partial l}{\partial\sigma^2} = \frac{-n}{\sigma^2} + \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(y_i-\mu_i)^2 = 0 \Longleftrightarrow \hat{\sigma}^2 = \left(\mathbf{1}-\frac{\mathbf{p}}{\mathbf{n}}\right)\mathbf{S}^2$$

Observation: If $n$ is large, both estimators are similar. For $p$ and $n$ relatively small they may differ quite a lot.

# Exercices

Exer. 1) Null model: $y_i = \beta_0 + e_i$, $i = 1, \cdots n$. Prove that:

$$\hat{\beta}_0 = \overline{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

Exer. 2) Simple regression model $y_i = \beta_0 + x_i\,\beta_1 + e_2$, $i = 1, \cdots, n$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2} = \frac{Cov(X,Y)}{Var(X)} = r_{XY}\frac{S_Y}{S_X}$$

Observation 1: $(\overline{x}, \overline{y})$ belongs to the regression line.

Observation 2: The correlation coefficient $(r_{XY})$ is a measure of the linear relation between $X$ and $Y$.

# 6. Parameter interpretation

Given that we are assuming that:

$$\forall i \quad , Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip-1}\beta_{p-1} + e_i = \mu_i + e_i,$$

Being $Y_i$ the response under conditions $X_i = (x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots, x_{ip-1})$ and by $Y_i^*$ the response under conditions $X_i^* = (x_{i1}, x_{i2}, \cdots, x_{ij}+1, \cdots, x_{ip-1})$, one has that

$$Y_i^* - Y_i = \hat{\beta}_j,$$

Thus,

- $\hat{\beta}_j$ is the mean change obtained by increasing one unit the covariate $x_j$ while leaving the other covariates equal.
- If $\hat{\beta}_0$ designs the intercept estimation, it is interpreted as the mean response at the origin (all covariates equal to zero)

The residual standard deviation $\hat{\sigma}$ is the error associated to our predictions, 95% of our predictions will have an error in

$$(-t_{n-p,\alpha/2}\,\hat{\sigma}, t_{n-p,\alpha/2}\,\hat{\sigma}) \simeq (-1.95\,\hat{\sigma}, 1.95\,\hat{\sigma})$$

# INFERENCE ON THE MODEL PARAMETERS

# 7. $\hat{\beta}$ distribution

If $\beta^*$ is the true parameter vector (please do not cofuse the intercept, real number, with the mean vector of $Y$, even if we denote them with the same letter),

$$\hat{\beta}|X \sim N(\beta^*, \sigma^2(X^tX)^{-1});$$

because it is a linear combination of Normal r.v.'s

$$E(\hat{\beta}|X) = (X^tX)^{-1}X^tE(Y|X) = (X^tX)^{-1}X^tX\beta^* = \beta^+;$$

and given that $\hat{\beta} - \beta_0 = (X^tX)^{-1}X^t(Y - X\beta_0)$,

$$\begin{aligned} E((\hat{\beta}|X - \beta^+)(\hat{\beta}|X - \beta^*)^t) &= (X^tX)^{-1}X^tE((Y - X\beta^*)(Y - X\beta^*)^t)X(X^tX \\ &= \sigma^2(X^tX)^{-1}. \end{aligned}$$

Observation 1: $\sigma^2(X^tX)^{-1}$ is the inverse of what is known as the Fisher information matrix.

Observation 2: The components of $\hat{\beta}$ are not indep. r.v.'s

# 8. Inference for the model parameters

Given that, $\hat{\beta}|X \sim N(\beta^*, \sigma^2(X^tX)^{-1})$;

each component verifies: $\hat{\beta}_i|X \sim N(\beta_i^*, \sigma^2[(X^tX)^{-1}]_{ii})$

Thus, for a given $a \in \mathbb{R}$, we can test:

$$H_0 : \beta_i^* = a$$

$$H_1 : \beta_i^* \neq a$$

at level $\alpha$ by computing

$$\frac{\hat{\beta}_i - a}{S \cdot \sqrt{[(X^tX)^{-1}]_{ii}}}$$

and rejecting the null hypothesis if

$$|\frac{\hat{\beta}_i - a}{S \cdot \sqrt{[(X^tX)^{-1}]_{ii}}}| \geq t_{n-p,1-\alpha/2}$$

In particular, it is specially interesting to test

$$H_0 : \beta_i^* = 0$$

$$H_1 : \beta_i^* \neq 0$$

since do not reject $H_0$ will imply that covariate $X_i$ has not a significative infuence on $Y$.

By default this test is given by default in any software output.

**Confidence interval** for the parameters:

$$\hat{\beta}_i \pm t_{n-p,\alpha/2}\, S \cdot \sqrt{[(X^tX)^{-1}]_{ii}}$$

Observation: If the interval contains the **zero value**, the corresponding covariate is **not statistically significative**.

# 9. Anova table and Omnibus test

Given that $Y_i - \overline{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$ and that $\sum_{i=1}^{n}(Y_i - \hat{Y}_i) \cdot (\hat{Y}_i - \overline{Y}) = 0$,

one has that:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

Let us denote each sum prespectively by:

- $TSS$ (total sum of squares)
- $RSS$ (residual sum of squares)
- $RegSS$ (regression sum of squares)

Observation: $\hat{\sigma^2} = S^2 = RSS/(n-p)$

The ANOVA table is equal to:

| Source | SS | d.f. | $MSS$ | F |
|---|---|---|---|---|
| Regression | $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ | $p - 1$ | $RegSS/p$ | $F_0 = \frac{RegSS/p}{RSS/(n-p)}$ |
| Residuals | $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ | $n - p$ | $RSS/(n-p)$ | |
| Total | $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ | $n - 1$ | | |

The **omnibus** test is defined as:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p - 1 = 0$$

$$H_1 : \text{it exists } i, \text{such that } \beta_i \neq 0$$

$$H_0 \text{ is rejected when } F_0 \geq F_{1-\alpha,p,n-p}$$

Observation 1: This test compares our model with the null model

Observation 2: The Omnibus test is not equivalent to test if each parameter equals to zero

# INFERENCE ON THE PREDICTED VALUES

# 11. Predicted values distribution

One defines **vector of predicted values** as $\hat{Y} = X\hat{\beta}$.

$$\hat{Y}|X \sim N(X\beta^*, \sigma^2 X(X^tX)^{-1}X^t);$$

because it is a linear combination of Normal r.v.'s

$$E(\hat{Y}|X) = XE(\hat{\beta}|X) = X\beta^*;$$

and

$$\begin{aligned} E((\hat{Y}|X - X\beta^*)(\hat{Y}|X - X\beta^*)^t) &= XE((\hat{\beta}|X - \beta^*)(\hat{\beta}|X - \beta^*)^t)X^t \\ &= X\sigma^2(X^tX)^{-1}X^t \\ &= \sigma^2 X(X^tX)^{-1}X^t. \end{aligned}$$

The matrix $X(X^tX)^{-1}X^t$ is called **hat matrix**. Since

$$\hat{Y} = X(X^tX)^{-1}X^tY$$

**$\sigma^2$ times the hat matrix is the matrix of variances and covariances of the predictions.**

# 12. PI for predictions

Let $X^* = (x_1^*, x_2^*, \cdots, x_{p-1}^*)^t$ be some particular experimental conditions.

The predicted value at $X^*$ is $\hat{y^*} = (X^*)^t \, \hat{\beta}$.

Remind that under normality and the null model, the CI for $\mu$ is computed as $\overline{y} \pm t_{1-\alpha/2,n-p} S/\sqrt{n}$.

In this case, a $100(1-\alpha)\%$ prediction interval (PI) for this future observation is:

$$\hat{y}^* \pm t_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \, (1 + (X^*)^t (X^t X)^{-1} X^*)}$$

For the particular case of simple linear regression the PI is

$$\hat{y}^* \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(X^* - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}$$

# 13. CI for mean responses

The fitted value at $X^*$ is $\hat{y}^* = (X^*)^t \, \hat{\beta}$.

Let $\mu^* = E(Y|X^*)$, one has that $E(\hat{y}^*) = (X^*)^t \, \hat{\beta} = E(Y|X^*)$.

In consequence, $\hat{y}^*$ **is an unbiased estimator of** $E(Y|X^*)$.

Moreover,

$$Var(\hat{y}^*) = \sigma^2 (X^*)^t (X^t X)^{-1} X^*,$$

and thus, a CI for $\mu_0$ is:

$$\hat{y}^* \pm t_{1-\alpha/2, n-p} \sqrt{\hat{\sigma}^2 \, (X^*)^t (X^t X)^{-1} X^*}$$

For the particular case of simple linear regression the PI is

$$\hat{y}^* \pm t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(X^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right)}$$

piciplot.pdf

# MORE ABOUT RESIDUALS

# 14. Residual vector distribution

The i-thm residual is defined as $\hat{e}_i = y_i - \hat{y}_i$.

The vector $\hat{e} = (Y_1 - \hat{Y}_1, Y_2 - \hat{Y}_2, \cdots, Y_n - \hat{Y}_n)^t$ is known as **residual vector** and verifies:

$$\hat{e}|X \sim N(0, \sigma^2(Id - X(X^tX)^{-1}X^t));$$

because it is a linear combination of Normal distributed r.v.'s

$$E(\hat{e}) = X\beta^* - X\beta^* = 0;$$

and given that

$$E(YY^t) = E(\hat{Y}\hat{Y}^t) + E((Y - \hat{Y})(Y - \hat{Y})^t),$$

one has

$$E((Y - \hat{Y})(Y - \hat{Y})^t) = \sigma^2 Id - \sigma^2 X(X^tX)^{-1}X^t = \sigma^2(Id - X(X^tX)^{-1}X^t).$$

# 15. Standarized and Studentized residuals

Thus, an estimation of the variance of the $(y_i - \hat{y}_i)$ is obtained as:

$$S^2 \cdot (1 - [X(X^tX)^{-1}X^t]_{ii}).$$

Denoting by $h_{ii} = [X(X^tX)X^t]_{ii}$, the **standarized residuals** are defined as:

$$\textbf{Stand.Res} = \frac{y_i - \hat{y}_i}{s\sqrt{1 - h_{ii}}}$$

And the **studentized residuals** as:

$$\textbf{Student.Res} = \frac{y_i - \hat{y}_i}{s_{(-i)}\sqrt{1 - h_{ii}}}$$

where $S^2_{(-i)}$ is the variance estimation when the $i$-thm observation is suppress from the model.

# IMPORTANT TO TAKE INTO ACCOUNT

# 16. Determination coefficient

The **determination coefficient** is a goodness-of-fit measure defined as:

$$R^2 = \frac{RegSS}{TSS} \cdot 100\%,$$

and interpreted as the amount of variability in the response captured by the explanatory variables.

As closer to one better is the model fit, if the model assumptions are fulfilled.

# 17. Multicollinearity

Definition: A term used in regression analysis to indicate situations where the explanatory variables are related by a linear function.... (the Cambridge Dictionary of Statistics, B.S. Everitt)

When two or more predictor variables are linearly dependent, they are called collinear.

**Multicollinearity** implies that $\mathbf{det}(\mathbf{X}^{\mathbf{t}} \cdot \mathbf{X})$ is very small or zero in the extrem case.

If collinearity exists

- The model interpretation is very difficult.
- The variance of the $\hat{\beta}_i$ parameters is large.
- Matrix $X^t \cdot X$ may be singular which makes impossible to compute the parameter estimations.

Comment: Even if there is multicollinearity, the predictions are correct if the model is correct.

# 17. Multicollinearity detection

To detect multicollinearity it is appropiate:

1) To perform the multiple scattered plot of all the pairs of explanatory variables, and to compute the **Correlations among pairs of predictors**:

$$r_{x_1 x_2} = \frac{\sum_i x_{1i}x_{2i} - n\overline{x}_1\,\overline{x}_2}{\left[\sum_i x_{1i}^2 - n\overline{x}_1^2\right]^{1/2}\left[\sum_i x_{2i}^2 - n\overline{x}_2^2\right]^{1/2}}$$

Observation: If a correlation term is relatively high, one of the $x$'s has to be eleminated and redo the analysis.

# 17. Multicollinearity detection: VIF

2) It is convinient to compute the **Variance Inflation Factor (VIF)** of each variable.

The VIF is defined as:

$$VIF(X_j) = \frac{1}{1 - R^2(X_j)},$$

being $R^2(X_j)$ the $R^2$ obtained by regressing the variable $X_i$ with respect to the rest of explanatory variables.

| VIF | Conclusion |
|---|---|
| VIF=1 | not correlated |
| $1 < VIF < 5$ | moderately correlated |
| $VIF > 5$ | highly correlated |

Observation: $1 - R^2(x_j)$ is known as **tolerance**.

It may be proved that:

$$Var(\hat{\beta}_j) = VIF(X_j) \cdot \frac{\sigma^2}{\sum_{i^1}^{n}(x_{ij} - \overline{x_j})^2},$$

thus as larger is the VIF more variance will have the estimated coefficient, and consequently we can rely less on it.

The procedure is like this: if one or more variables have a large VIF, then the one with the larger VIF is supresed and the model is fitted again with one variable less, if there are still variables with a larger VIF we keep doing like this, otherwise we stop removing variables.

**Model fitting is an iterative process**.

# 18. Leverage

Given an observation $Y_i$, its leverage is defined as the element ii-thm of the hat matrix, that is:

$$h_{ii} = [X(X^t X)X^t]_{ii}$$

it is a measure of the distance between $(x_{i1}, x_{i2}, \cdots, x_{ip-1})$ and the centroid $(\overline{x_1}, \overline{x_2}, \cdots, \overline{x_{p-1}})$.

The leverage verifies that for any $i$, $1/n \leq h_{ii} \leq 1$.

Observations such that:

$$h_{ii} > 3p/n \text{ or } h_{ii} > 0,99$$

have a large leverage and need to be observed carefully.

# 19. Influential observations: Cook's distance

Definition: An influential observation is an observation that has a disproportionate influence on the values of the regression coefficients (the Cambridge Dictionary of Statistics, B. S. Everitt)

The Cook's distance is computed as:

$$D_i = \frac{h_{ii}}{1 - h_{ii}} \cdot \frac{\hat{e}_i^2 / S^2}{p + 1}, \; i = 1 \cdots n$$

and it is large either if the observation has a large leverage or if it has a large standarized residual.

Observation 1: Outliers are influencial observations

Observation 2: If the leverage and the residuals are both big, clearly we have and influential observation.

Figure: Two Regression lines: with all the observations (blue), without influential point (black).

# 20. Influential observations: DFFITS

Another measure that is quite similar to the Cook's distance is:

$$DFITS_i = \sqrt{\frac{h_{ii}}{1-h_{ii}}} \cdot StudentRes_i, \quad i = 1, \cdots, n$$

It is verified that

$$D_i \simeq \frac{(DFFITS_i)^2}{(p+1)}$$

# 21. Influential observations: DFBETA, DFBETAS

To know if an observation is influent, it has sense to compute its impact on each one of the model parameters.

Given the $i$-thm observation, $i = 1, \cdot, n$, for $j = 0, \cdots p - 1$,

$$D_{ij} = \beta_j - \beta_{j(-i)}\,(DFBETA)\;\; and/or\;\; D^*_{ij} = \frac{\beta_j - \beta_{j(-i)}}{S_{(-i)}(\beta_j)}\,(DFBETAS)$$

This will give a total number of $n \times p$ values that are usually better examined graphically.

If for a given $i$ one or several DFBETA or DFBETAS are large, we have to look carefully to that observation.

# MODEL ADEQUACY CHECKING

# 22. Goodness of fit measures: $R^2$

More about the **Coefficient of determination**

It is defined as:

$$R^2 = \frac{RegSS}{TSS} \cdot 100\%$$

but it is also equal to:

$$R^2 = c^t R_{xx} c,$$

where $c = (r_{x_1 y}, r_{x_2 y}, \cdots r_{x_p y})$ and $R_{xx} = (r_{x_i x_j})_{i,j}$ being $r_{a,b}$ the linear correlation of vectors $a$ and $b$.

# 22. Goodness of fit measures: $R^2$

1) $R^2 \in (0,1)$
2) as larger it is, better is the fit
3) It is a measure of the correlation between the observed values and the one predicted by the model.
4) If $X$ columns have zero correlation, then $R_{xx} = Id_p$ and thus

$$R^2 = c^t \cdot c = ||c||^2$$

5) For simple linear regression, $R^2 = (r_{xy})^2$.

Observation: As large are the components of $c$, more correlated is the response variable with the explanatory variables.

Alternative expression for the $R^2$

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

And this it is interpreted as: **proportion of variability in the data explained by the covariates** or **one minus the proportion of variability in the data not explained by the covariates.**

Exer 5. Prove that in the case of Simple Linear Regression,

$$R^2 = (r_{xy})^2$$

# 22. Goodness of fit measures: adjusted $R^2$

It is obvious that as many covariates appear in the model, larger is going to be the $R^2$ value.

In order to penalize models with more covariates and to be able to compare models with different number of covariates, one computes the adjusted $R^2$ as follows:

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

Observation: If $p$ increases, the numerator of the fraction increases and thus $R^2_{adj}$ decreases with respect to $R^2$.

# 23. Goodness of fit measures: Residual analysis

**Checking the model assumptions**

- Scatter plot of $y$ vs $x$, linearity must be observed.
- Scatter plot of $\hat{e}_i$ vs $\hat{y}_i$, no trend must be observed.
- qq-plot for $\hat{e}_i$, linearity must be observed.
- Sometimes, plot $\hat{e}_i$ vs order in which the observations are conducted.

# SUMMARY

# 24. Multiple Regression: Summary

- Parameter estimation: $\hat{\beta} = (X^t \cdot X)^{-1} X^t y$
- Predicted values: $\hat{y}_i = (X\hat{\beta})_i$.
- Residual sum of squares: $RSS = \sum_i (y_i - \hat{y}_i)^2$.
- Variance estimation: $S^2 = \hat{\sigma}^2 = MSE = \frac{RSS}{n-p}$
- Standard deviation for $\beta$'s: $S_{\hat{\beta}_j} = S \cdot \sqrt{c_{jj}}$, being $c_{jj}$ the diagonal elements of $(X^t \cdot X)^{-1}$
- Conf. int for parameter: $\hat{\beta}_i \pm t_{1-\alpha/2,n-p} \cdot S_{\hat{\beta}_j}$
- Coef. of determination: $R^2 = \frac{RegSS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$

EXAMPLE

# 25. Example of multiple regression

**Example:** Seven programs were monitored to observe their resource demands. In particular, the number of disk I/O's, menory size (in Kilobytes), and CPU time (in milliseconds) were observed. One is interested in modelling CPU time ($Y$) as a function of the other two.

$$Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + e_i, \; i = 1, \cdots, n;$$

assuming the general linear model assumptions-

# Example of multiple Regression

| CPU Time ($y_i$) | Disk I/0's ($x_{1i}$) | Memory size ($X_{2i}$) |
|:-:|:-:|:-:|
| 2 | 14 | 70 |
| 5 | 16 | 75 |
| 7 | 27 | 144 |
| 9 | 42 | 190 |
| 10 | 39 | 210 |
| 13 | 50 | 235 |
| 20 | 83 | 400 |

# Example of multiple Regression

For the CPU time example,

$\hat{\beta} = (-0.1614, 0.1182, 0.02265)^t$.

The model is:

$$\text{CPU time} = -0.1614 + 0.1182 \cdot \text{Number of disk I/0's} + 0.0265 \cdot \text{Memory size}$$

Residual sum of squares: $RSS = \sum_i e_i^2 = 5.3$

$$S = \sqrt{SSE/(7-3)} = 1.2$$

Coefficient of determination: $R^2 = 1 - \frac{RSS}{SST} = 0.97$

# Example of multiple Regression

Estandard errors and conf. interv. for parameter estimations

| | $S_{\hat{\beta}_i}$ | Conf. interv. (90%) |
|---|---|---|
| $\beta_0$ | $1.2\sqrt{0.6297} = 0.9131$ | $(-2.11, 1.79)$ |
| $\beta_!$ | $1.2\sqrt{0.0280} = 0.1925$ | $(-0.29, 0.53)$ |
| $\beta_2$ | $1.2\sqrt{0.0012} = 0.0404$ | $(-0.06, 0.11)$ |

Observation: None of the parameters are significant at 90% confidence level.

# Example of multiple Regression

In the CPU example:

Computing the correlation coef. between I/O's and memory size we get:

$$R_{x_1 x_2} = 0.9947$$

This may be due to large programs (large memory size) doing more I/O's than small programs.

Peforming the two simple linear regressions one can see that both are useful to estimate CPU Time. So, one can consider each one of the regresors but not both.

# COMMON MISTAKES

# 26. Regression: Common mistakes

- Not verifying that the relationship is linear
- Relying on the automated results without visual verification
- Not taking into account that parameter estimations depend upon the units of the predicted and predictor variables.
- Confusing the Coeff. of Determination and the Coeff. of Correlation.
- Using highly correlated variables in the prediction
- Using regression to predict far beyond the measured range. The statistical confidence decreases as we move outside the measured range.

- Using too many predictor variables (overfitting).
- Measuring only a small subset of the complete range of the explanatory variables.
- Confusing correlation with causality: two variables may be hightly correlated but none of them controls the other one.

  Regression the I/O's variable with respect to CPU time, we can deduce that more CPU time may be used to predict the number of disk I/O's.

  Nevertheless, installing a faster CPU will not imply a reduction of the number of disk I/O's, inspite that the observed CPU time will be smaller.

Figure: Extrapolation in general is not possible