Advanced Statistics
Part one: Parametrical Statictics

M. Pérez-Casany

Dept. of Statistics and Operations Research and DAMA-UPC
Technical University of Catalunya

First Semester: Second Session

Session 2: Parameter Estimation, Hypothesis Testing and Linear Model

1. Point Estimation: mm, mls, mle

Let X be a r.v. and $x = (x_1, x_2, \cdots, x_n)$ a sample from X.

Assume that X has dpf (pmf is de discrete case)

$$f(x;\theta), \quad \theta \in \Omega$$

where Ω is known as **parameter space**.

We have a set of prob. distrib.

$$\{f(x;\theta)/\theta \in \Omega\}$$

Objective of the point estimation: to find the value $\theta_0 \in \Omega$ that is more consistent with the sample x.

The different methodologies consist on finding the θ value that **maximizes** of **minimizes** a given function.

1. Point Estimation: mm, mls, mle

The **Moment Estimation Method (mm)**

In the one dimensional case, the parameter estimation, $\tilde{\theta}$, is the solution of the equation:

$$\mu_\theta = E_\theta(X) = \overline{x}$$

In the two-dimensional case, it is the solution of the system of equations:

$$\begin{cases} \mu_\theta = E_\theta(X) = \overline{x} \\ E_\theta(X^2) = \frac{1}{n}\sum_i x_i^2 \end{cases}$$

Similarly it is generalized for higer dimensional prob. distrib.

1. Point Estimation: mm, mls, mle

- For the Binomial(m, p) distrib. with $n = 1$

$$\theta = p, \qquad \mu_\theta = n\,p = x \longrightarrow \tilde{p} = x/n$$

- For the Poisson distrib.

$$\theta = \lambda, \qquad \mu_\theta = \lambda = \overline{x} \longrightarrow \tilde{\lambda} = \overline{x}$$

- For the exponential distrib.

$$\theta = \lambda, \qquad \mu_\theta = E_\theta(X) = \frac{1}{\lambda} = \overline{x} \longrightarrow \tilde{\lambda} = (\overline{x})^{-1}$$

- For the Normal (μ, σ^2) distrib.

$$\theta = (\mu, \sigma^2), \qquad \left\{ \begin{array}{l} \mu_\theta = E_\theta(X) = \overline{x} \\ \sigma^2_\theta = Var_\theta(X) = S^2 \end{array} \right.$$

1. Point Estimation: mm,mls, mle

Observation: Given that $\overline{x}$ is the statistic that minimizes

$$f(a) = \sum (x_i - a)^2,$$

and the moment estimation method makes $\mu_{\tilde{\theta}} = \overline{x}$, it may seen as a **minimum least squares method** (mls) since the parameter estimator $\tilde{\theta}$ is the one that minimizes

$$\sum_i (x_i - \mu_\theta)^2$$

1. Point Estimation: mm,mle, mls

The **Maximum likelihood Method (mle)**

Takes as parameter estimation, $\hat{\theta}$, the value such that maximizes the **likelihood function** defined by:

$$L(\theta; x) = f(x_1, \theta) \cdot f(x_2, \theta) \cdots f(x_n; \theta)$$

which is equivalent to maximize its logarithm:

$$l(\theta; x) = \sum_i \log f(x_i; \theta)$$

This maximization responds to the question: given that we have observed x, **which is the value that maximizes the prob. to observe x under the assumed model?**

1. Point Estimation: mm,mle, mls

Quite often the two methods bring to the same estimator. This is the case of the binomial, Poisson or Normal distributions.

An estatistic $t(x)$ is said to be an **unbiased** estimator for θ when

$$E(t(x)) = \theta$$

Observation: the unbiasedness is not a general property of the mle, but usually mle are unbiased.

Observe that:

$$t_1(x) = \frac{1}{n}\sum(x_i - \overline{x})^2$$

is a biased estimator for σ^2 and

$$t_2(x) = \frac{1}{n-1}\sum(x_i - \overline{x})^2$$

Is not.

2. Measures of quality of an estimator

We want an estimator be

1) **unbiased** and
2) with **minimum variance**

The Crámer Rao inequality says that an unbiased estimator $t(x)$ for θ has minimum variance if and only if:

$$Var(t(x)) = \left(E\left(\sum_i \frac{\partial^2 \log L(x_i,\theta)}{\partial\theta^2}\right)\right)^{-1}$$

3. Confidence Interval estimation

Let $x_1, x_2, \cdots, x_n$ be a sample from a r.v. X.

Let $t_1(x)$ and $t_2(x)$ be two statistics such that $t_1(x) < t_2(x)$.

The interval $(t_1(x), t_2(x))$ is said to be a **confidence interval** for parameter θ with **condifence level** $(1-\alpha)$ if, and only if, $(1-\alpha)\%$ of the interval will contain the true value.

That is:

$$\theta \in (t_1(x), t_2(x))$$

will be true for $(1-\alpha)\%$ of the samples.

3. Confidence Interval estimation

Conf. Interval for μ under Normal(μ, σ^2) assumption.

Given that

$$X \sim N(\mu, \sigma^2) \Longrightarrow \overline{x} \sim N(\mu, \sigma^2/n),$$

after some steeps one has that the interval for μ is equal to:

1) $\overline{x} + -Z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$

   if σ^2 is known, being $Z_{1-\alpha}$ the quantile $(1-\alpha)$ in a N$(0,1)$ distribution.

2) $\overline{x} + -t_{1-\alpha,n-1} \cdot \frac{S}{\sqrt{n}}$

   if σ^2 is unknown, being $t_{1-\alpha,n-1}$ the quantile $(1-\alpha)$ in a t_{n-1} distribution.

4. Hypothesis Testing

Assume the $X \sim f(x;\theta)$, and that one wants to decide betweent he two hypothesis

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta \neq \theta_0$$

is such a way that:

- $P(\text{reject} H_0 | H_0 \text{is true}) \leq \alpha$
- $P(\text{accept} H_0 | H_0 \text{is false})$ small as possible

The test is defined by means of a **decision rule** of the form:

if the statistic $t(x)$ is larger than a given value, reject H_0

4. Hypothesis Testing

If $X \sim N(\mu, \sigma^2)$ and we want to test:

$$H_0 : \mu = \mu_0 \;\; vs \;\; \mu \neq \mu_0$$

We reject H_0 when:

- $|\overline{x} - \mu_0| \geq Z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$ if σ is known
- $|\overline{x} - \mu_0| \geq t_{1-\alpha,n-1} \cdot \frac{S}{\sqrt{n}}$ if σ is unknown

5.1. Definition of Linear Model

Objetive: to explain the vehaviour of a r.v. Y (dependent var.) as a function of $X_1, X_2, \cdots, X_p$ (independent var. or covariates).

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

Definition:

$$\forall i \;\; , Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + e_i$$

Hipothesis:

- $\forall i \in \{1, 2, \cdots, n\}$, $Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$;
- $\forall i \in \{1, 2, \cdots, n\}$, $\sigma_i^2 = \sigma^2$;
- $\forall i, j \in \{1, 2, \cdots, n\}\, i \neq j$, Y_i indep.of Y_j.

In matrix form,

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

$$\begin{array}{ccccc} \downarrow & \downarrow & \downarrow & \cdots & \downarrow \\ X_1 & X_2 & X_3 & \cdots & X_p \end{array}$$

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

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

Observation:

The variables $X_1, X_2, \cdots, X_p$ may be a function of another set of variables. It may exists $\{Z_1, Z_2, \cdots, Z_m\}$, $m \in \mathbb{N}$, such that

$$X_i = g_i(Z_1, Z_2, \cdots, Z_m), \ \ \forall i \in \{1, 2, \cdots, p\}.$$

Examples:

- $p = 3$, $m = 1$; $\quad X_1 = 1 = Z_1^0$, $X_2 = Z_1$, $X_3 = Z_1^2$

$$\forall i \qquad Y_i = \beta_1 + z_{i1}\beta_2 + z_{i1}^2\beta_3.$$

- $p = 2$, $m = 3$; $\quad X_1 = e^{Z_1}\, Z_2$, $X_2 = Z_3 - Z_2$

$$\forall i \qquad Y_i = e^{z_{i1}}\, z_{i2}\beta_1 + (z_{i3} - z_{i2})\beta_2.$$

5.2 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,

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

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).

$$Y_i = \beta_1 + x_i\beta_2 + e_i,\ i = 1, \cdots, n;$$

in matrix form,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

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).

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

in matrix form,

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

5.3 Examples of non-linear models

Example: On wants to study the milk productiuon in cows as a function of the days since the day they gave birth. If x_i is the number of days from the birth day and Y_i denotes the production of milk in liters, the suitable model is

$$Y_i = \exp\big(\beta_0 + \beta_1 x_i + \log(x_i)\big) + e_i \ \text{ donde } e_i \sim \mathrm{N}(0, \sigma^2)$$

Example: One wants to study the quality of matter provided by a different providers. To that end, from each provider one randomly selects a set of b shipment, and from each one one obtains n observations. the suitable model is:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + e_{(ij)k}$$

where
$i = 1, \cdots a, \ j = 1, \cdots, b, \ k = 1, \cdots, n$ y $e_{(ij)k} \sim \mathrm{N}(0, \sigma^2)$.

5.4 Parameter vector estimation

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

Minimum least square estimation minimizes:

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

where $\hat{y} = \hat{\mu} = X\hat{\beta}$.

Máxima verosimilitud estimation maximizes:

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

which is equivalent to

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

Let us define

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

The vector $U = (U_1, U_2, \cdots U_p)^t$ is called **score vector.**

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

if the rank of (X^tX) is equal to $= p$.

$\hat{\beta}$ es U.M.V.U.E.

The matrix $\mathcal{J} = E(UU^t)$ is known as **Fisher information matrix.**

Under the Normality assumption,

$$\begin{aligned}
\mathcal{J} = E(UU^t) &= E\left(\frac{1}{\sigma^2} X^t (Y - X\beta)(Y - X\beta)^t X \frac{1}{\sigma^2}\right) \\
&= \frac{1}{\sigma^2} X^t E\big((Y - X\beta)(Y - X\beta)^t\big) X \frac{1}{\sigma^2} \\
&= \frac{1}{\sigma^2} X^t X.
\end{aligned}$$

$\mathcal{J}$ is important to perform inference about the model parameters.

Weighted least squares one wants to minimize:

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

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

Weighted least squares with correlated data one wants to minimize:

$$S(\beta) = (y - X\beta)^t V^{-1} (y - X\beta).$$

where $V = Var(YY^t)$.

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

5.5 $\hat{\beta}$ distribution

If β_0 is the true parameter value

$$\hat{\beta} \sim N(\beta_0, \sigma^2(X^tX)^{-1});$$

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

$$E(\hat{\beta}) = (X^tX)^{-1}X^tE(Y) = (X^tX)^{-1}X^tX\beta_0 = \beta_0;$$

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

$$\begin{aligned} E((\hat{\beta}-\beta_0)(\hat{\beta}-\beta_0)^t) &= (X^tX)^{-1}X^tE((Y-X\beta_0)(Y-X\beta_0)^t)X(X^t \\ &= \sigma^2(X^tX)^{-1}. \end{aligned}$$

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

5.6 Predicted values

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

$$\hat{Y} \sim N(X\beta_0, \sigma^2 X(X^tX)^{-1}X^t);$$

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

$$E(\hat{Y}) = XE(\hat{\beta}) = X\beta_0;$$

and

$$\begin{aligned} E((\hat{Y} - X\beta_0)(\hat{Y} - X\beta_0)^t) &= XE((\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)^t)X^t \\ &= X\sigma^2(X^tX)^{-1}X^t \\ &= \sigma^2 X(X^tX)^{-1}X^t. \end{aligned}$$

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

5.7 Residual vector

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:

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

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

$$E(\hat{e}) = X\beta_0 - X\beta_0 = 0;$$

and given that

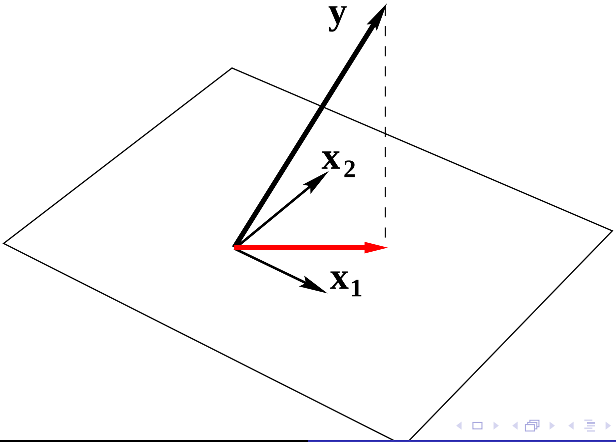
$$E(YY^t) = E(\hat{Y}\hat{Y}^t) + E((Y - \hat{Y})(Y - \hat{Y})^t),$$

one has

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

$\hat{e}$ is orthogonal to the columns of matrix X.

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



5.8 Residual variance estimation

Moment method

$$\hat{\sigma}^{\mathbf{2}} = \mathbf{S^2} = \frac{1}{n-r}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \text{ donde } r = rango(X^t X).$$

S^2 is an U.M.V.U. estimator

Maximum likelihood method.

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

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

5.10 Goodness of fit measures

Cofficient of multiple correlation

Is defined as

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

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

1) $R^2 \in (0, 1)$
2) as larger is its value as 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, $R^2 = c^2 \cdot c$.