Mathematics for Big Data
Statistics - Assignment 1

Aina Brichs Ralló

Balin Lin

Universitat Autònoma
de Barcelona

Exercise 1

Consider a simple regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, ..., n$, where the errors ϵ_i are iid $N(0, \sigma^2)$,in other words $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Find a three-dimensional sufficient statistics for the parameters.

Answer:

$y = (y_1, y_2, ..., y_n$ and $(x_1, x_2, ..., x_n)$ are known constants, whereas β_0 and β_1 are unknown.

The density function for y_i in a linear regression is given by the following expression:

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left[y_i - (\beta_0 + \beta_1 x_i)\right]^2\right\}$$

Therefore, the likelihood can be expressed as:

$$\mathrm{L}(\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left[y_i - (\beta_0 + \beta_1 x_i)\right]^2\right\}$$

If instead of multiplication, we use summation, we see that the expression is transformed to:

$$\mathrm{L}(\mathbf{y}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

And by separating the different terms inside the summation, we get that:

$$\mathrm{L}(\mathbf{y}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) + \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)^2\right)\right\}$$

Therefore,

$$\mathrm{L}(\mathbf{y}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\beta_0\sum_{i=1}^{n} y_i - 2\beta_1\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)^2\right)\right\}$$

From this final expression of the likelihood of a linear regression model, we can conclude that a three-dimensional sufficient statistics is given by:

$$\boxed{\mathrm{T} = \left(\sum_{i=1}^{n} y_i^2, \quad \sum_{i=1}^{n} y_i, \quad \sum_{i=1}^{n} y_i x_i\right)}$$