Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #2 (régression sur une variable continue - 1)

Préambule

If

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

then

$$X_1 \mid X_2 = x_2 \ \sim\ \mathcal{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2),\ (1-\rho^2)\sigma_1^2\right).$$

i.e.

$$\mathbb{E}(X_1 \mid X_2 = x_2) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2) = \underbrace{\mu_1 - \rho\frac{\sigma_1}{\sigma_2}\mu_2}_{\beta_0} + \underbrace{\rho\frac{\sigma_1}{\sigma_2}}_{\beta_1} x_2$$

and

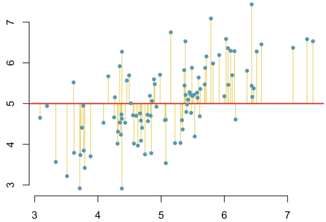
$$\text{Var}(X_1 \mid X_2 = x_2) = \underbrace{(1-\rho^2)\sigma_1^2}_{\text{constant}} < \sigma_1^2 = \text{Var}(X_1)$$

Consider the following model

$$y_i = \beta_0 + \varepsilon_i$$

where

- β_0 is an unknown parameter
- ε_i is the unobservable random error term (or residual)



Consider n observations y_i. The residual sum of squares is

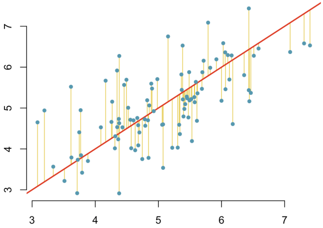
$$RSS(\beta_0) = \sum_{i=1}^{n} \big(y_i - \beta_0\big)^2$$

Set

$$\widehat{\beta}_0 = \text{argmin}\{RSS(\beta_0)\} = \overline{y}$$

Droite de régression (x continue)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



where

- β_0 and β_1 are unknown regression parameters
- ε_i is the unobservable random error term (or residual)

Consider n pairs of observations (x_i, y_i). The residual sum of squares is

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big(y_i - [\beta_0 + \beta_1 x_i]\Big)^2$$

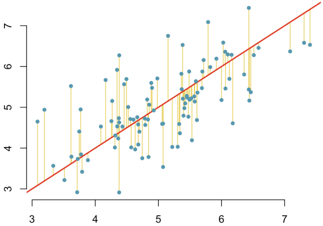
Consider

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \text{argmin}\,\{RSS(\beta_0, \beta_1)\}$$

Droite de régression (x continue)

First order conditions are here

$$\left.\frac{\partial RSS(\beta_0,\beta_1)}{\partial\beta_0}\right|_{(\widehat{\beta}_0,\widehat{\beta}_1)} = 0$$



while

$$\left.\frac{\partial RSS(\beta_0,\beta_1)}{\partial\beta_1}\right|_{(\widehat{\beta}_0,\widehat{\beta}_1)} = 0$$

Then $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$, while

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2} = \text{corr}(\mathbf{x},\mathbf{y}) \cdot \frac{s_y}{s_x}$$

or

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

Droite de régression (x continue)

The fitted values (or predictions) are

$$\widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and (fitted) residuals are

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i$$

Observe (from the first first order condition) that

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0 \text{ , since } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(provided that there is an intercept term - β_0 - in the model) and (from the second first order condition)

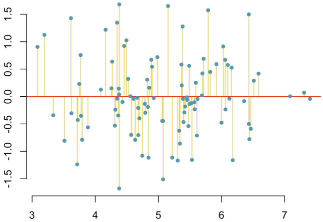
$$\sum_{i=1}^{n} x_i \widehat{\varepsilon}_i = 0$$

Droite de régression (x continue)

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0 \text{ means } \hat{\bar{y}} = \beta_0 + \hat{\beta}_1 \bar{x}$$

i.e. the regression line passes through the means $(\overline{x}, \overline{y})$

$$\sum_{i=1}^{n} x_i \widehat{\varepsilon}_i = 0 \text{ means } \text{corr}(\mathbf{x}, \widehat{\boldsymbol{\varepsilon}}) = 0$$



if we consider model

$$\widehat{\varepsilon}_i = \alpha_0 + \alpha_1 x_i + \eta_i,$$

then $\hat{\alpha}_0 = \hat{\alpha}_1 = 0$ (least squares).

Droite de régression (x continue)

Observe that

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\substack{\text{total sum of squares}\\ TSS}} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\substack{\text{residual sum of squares}\\ RSS}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\substack{\text{explained sum of squares}\\ ESS}}$$

The determination coefficient R^2 is

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

One can write

$$R^2 = \widehat{\beta}_1^2 \frac{s_x}{s_y} = \frac{s_{xy}^2}{s_x s_y} = \text{corr}(\mathbf{x},\mathbf{y})^2$$

Finally, the estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{RSS}{n-2}$$

Droite de régression (x continue)

Parmi les autres interprétations des estimateurs, notons que les paramètres sont des fonctions linéaires des y_i:

$$\widehat{\beta}_1 = \sum_{i=1}^{n} \omega_{1,i} \cdot y_i \text{ where } \omega_{1,i} = \frac{x_i - \overline{x}}{s_x^2}$$

and

$$\widehat{\beta}_0 = \sum_{i=1}^{n} \omega_{0,i} \cdot y_i \text{ where } \omega_{0,i} = \frac{1}{n} - \overline{x}\omega_{1,i}$$

La notation ω ne signifie pas vraiment que l'on ait ici des poids : les ω_i peuvent être négatifs. Par exemple, on notera que

$$\sum_{i=1}^{n} \omega_{1,i} = 0, \quad \sum_{i=1}^{n} \omega_{1,i} \cdot x_i = 1, \quad \sum_{i=1}^{n} {\omega_{1,i}}^2 = \frac{1}{s_x^2}$$

Droite de régression (x continue)

Under technical assumptions (discussed later), $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators of β_0 and β_1 respectively,

$$\mathbb{E}[\widehat{\beta}_0] = \beta_0 \text{ and } \mathbb{E}[\widehat{\beta}_1] = \beta_1$$

Variances are respectively

$$\text{Var}[\widehat{\beta}_0] = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{s_x^2}\right) \quad , \quad \text{Var}[\widehat{\beta}_1] = \frac{\sigma^2}{ns_x^2}$$

but since σ is unknown, those variances are estimated by

$$\widehat{\text{Var}}[\widehat{\beta}_0] = \widehat{\sigma}^2\left(\frac{1}{n} + \frac{\overline{x}^2}{s_x^2}\right) \text{ and } \widehat{\text{Var}}[\widehat{\beta}_1] = \frac{\widehat{\sigma}^2}{ns_x^2}$$

which gives a standard error

$$s_{\hat{\beta}_1} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^{,2}}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Droite de régression (x continue)

and

$$s_{\hat{\beta}_0} = s_{\hat{\beta}_1}\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2} = \sqrt{\frac{1}{n(n-2)}\left(\sum_{i=1}^n \hat{\varepsilon}_j^{,2}\right)\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}$$

Recall that

$$\underbrace{\sum_{i=1}^n (y_i-\bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^n (y_i-\hat{y}_i)^2}_{\text{RSS}} + \underbrace{\sum_{i=1}^n (\hat{y}_i-\bar{y})^2}_{\text{ESS}}$$

and the variance of y is estimated by

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i-\bar{y})^2 = \frac{\text{TSS}}{n-1}$$

Droite de régression (x continue)

while the variance of ε is estimated by

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 = \frac{\text{RSS}}{n-2}$$

Finally, if we assume that ε are normally distributed, we can prove that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are Gaussian estimators.