Validation of the model

Model:

$$\forall i \in \{1, \dots, n\}, \quad y_i = \beta_0 + \beta_1 X_i^{(1)} + \dots + \beta_q X_i^{(q)} + \varepsilon_i$$

$$\underline{Y} = \mathbb{X}\beta + U$$

assumptions :

- $\mathbb{X}$: a full rank matrix
- $U \sim \mathcal{N}(0, \sigma^2 I_n)$

Two scopes:

- Estimation
- Validation

Main difficulty: to verify that $U \sim \mathcal{N}(0, \sigma^2 I_n)$ because no observation!

Because no observation of the ε_i,

we are considering the $\hat{\varepsilon}_i$ (residuals)

<u>Rk</u>: $\hat{\varepsilon}_i$ are estimations of the ε_i

P_X : matrix associated to the orthogonal projection on $\mathcal{C}_X$

$$P_X = X \cdot \left({}^tX X\right)^{-1} {}^tX$$

ε_i	$\hat{\varepsilon}_i$
$E[\varepsilon_i] = 0$	$E[\hat{\varepsilon}_i] = 0$
$V[U] = \sigma^2 I_n$	$V[\hat{U}] = \sigma^2 (I_n - P_X)$

We deduce that $V(\hat{\varepsilon}_i)$ depends on i.

$\Rightarrow$ no homoscedasticity for the variance of the residuals

To suppress this non homogeneity, we make some normalization

notation: h_{ij} the term at position (i,j) into the T_x matrix

let introduce:

$$r_i := \frac{\hat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}}$$

$$\Rightarrow V[r_i] = 1 \quad \forall i \in \{1, \dots, n\}$$

Since σ^2 is unknown, we can not compute the r_i in practice

$\rightarrow$ we prefer the standardized residuals whose definition is:

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_r \sqrt{1 - h_{ii}}}$$

Rk:

$$\hat{\sigma}_n^2 = \frac{1}{n - \mathrm{rk}\, X} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

$$\longrightarrow \quad (n - \mathrm{rk}\, X) \frac{\hat{\sigma}_n^2}{\sigma^2} \sim \chi^2(n - \mathrm{rk}\, X)$$

$$t_i = \frac{r_i \sim \mathcal{N}(0;1)}{\sqrt{(n-rk\,X)\,\frac{\hat{\sigma}_n^2}{\sigma^2} \Big/ (n-rk\,X)}}$$

$$\hat{\cdot} \sim \chi^2(n-rk\,X)$$

$$r_i = \frac{\hat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}} \quad \text{with } \hat{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2(1-h_{ii}))$$

We do not know the distribution of the t_i because we do not have the independance between $\hat{\xi}_i$ and $\hat{\sigma}_n^2$

This explains why we introduce the studentized residuals:

$$r_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

where $\hat{\sigma}_{(i)}$ is an estimation of σ^2 without using the individual i

In practice, how to construct $\hat{\sigma}_{(i)}$?

- $\underline{Y}_{(i)}$: the vector $\underline{Y}$ where we delete the row i.
- $\mathbb{X}_{(i)}$: the matrix $\mathbb{X}$ where we delete the row i
- $U_{(i)}$: vector U where we delete the row i

$$\Rightarrow \underline{Y}_{(i)} = \mathbb{X}_{(i)} \beta + U_{(i)}$$

We are able to estimate β and σ^2 with those $(n-1)$ observations

$$\Rightarrow \hat{\beta}_{(i)} = \left(\mathbb{X}_{(i)}^{\prime} \mathbb{X}_{(i)} \right)^{-1} \mathbb{X}_{(i)}^{\prime} \underline{Y}_{(i)}$$

$$\hat{Y}_{(i)} = X_{(i)} \hat{\beta}_{(i)}$$

$$\hat{\sigma}^2_{(i)} = \frac{1}{n - \operatorname{rk} X_{(i)}} \times {}^t\hat{Y}_{(i)} \hat{Y}_{(i)}$$

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

with $\hat{y}_i = \begin{pmatrix} 1 & x_i^{(1)} & \dots & x_i^{(p)} \end{pmatrix} \hat{\beta}_{(i)}$

Rk: Thus, $\hat{\varepsilon}_i$ can be viewed as the prediction error

$$t_i^* = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(i)} \sqrt{1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i}}$$

$$\sim T(n-k-1)$$

$$x_i^T = (1 \; x_i^{(1)} \; \cdots \; x_i^{(r)})$$

<u>Rk</u>: The computation of $\hat{\sigma}_{(i)}$

is in some sense a cross-validation

method.

What is cross-validation?

$\mathcal{L}$: learning sample

We split the learning sample into ℓ folds

integer between 2 and n

→ we create V folds denoted

$\mathcal{L}_1, \dots, \mathcal{L}_V$

1° step:

we take $\mathcal{L}_2 \cup \mathcal{L}_3 \cup \dots \cup \mathcal{L}_V$ as the new learning sample

→ we build our model thanks to the new learning sample

for example: for linear model

$\hat{\beta}$.

We take $\mathcal{L}_1$ which is for this step a test sample.

$\rightarrow$ $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ for the individuals i that are into $\mathcal{L}_1$

<u>Step 2</u>: we take $\mathcal{L}_1 \cup \mathcal{L}_3 \cup \ldots \cup \mathcal{L}_V$

as the new learning sample

$\longrightarrow$ we build a new model

for example in linear model: $\hat{\beta}_2$

$\rightarrow$ $\mathcal{L}_2$ is a test sample for this model

$\rightarrow$ $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ for all the individuals into $\mathcal{L}_2$

We do the same for step 3 until step V.

$\rightarrow$ (10) - validation error:

$$:= \frac{1}{V} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

This error is a way to evaluate the performance of the model constructed thanks to the whole learning sample!

Work:

1) Simulated observations like this:

$n = 1000$

$p = 4$

$$X^{(1)} \sim \mathcal{E}(0.2)$$

$$X^{(2)} \sim \mathcal{U}([-2,2])$$

$$X^{(3)} \sim T(7)$$

$$X^{(4)} \sim \mathcal{B}(10; 0.5)$$

$$Y = 3 - 2 \times X^{(1)} + 3 \times X^{(2)} + X^{(3)} - 5 \times X^{(4)} + U$$

with $U \sim \mathcal{N}(0; 1)$

c) Construct the estimation of the linear model

3) Evaluate the performance of your model with cross-validation