

TD n°2

stoehr@ceremade.dauphine.fr

Exercise 1. We consider the homoscedastic linear model without an intercept, *i.e.*, for $i \in \llbracket 1, n \rrbracket$,

$$Y_i = \beta x_i + \varepsilon_i,$$

where $x_1, \dots, x_n$ are fixed, and the errors $\varepsilon_1, \dots, \varepsilon_n$ are independent, have mean zero and common variance σ^2 . We assume that

$$\sum_{i=1}^n x_i^2 > 0, \quad \sum_{i=1}^n x_i \neq 0.$$

We define

$$\hat{\beta}_n = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}, \quad \tilde{\beta}_n = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}.$$

1. Interpret $\hat{\beta}_n$ and $\tilde{\beta}_n$ as Z -estimators by specifying their estimating functions. Which of these estimators coincides with the least squares estimator?
2. Compare the properties imposed on the residuals by these two methods.
3. Show that both estimators are unbiased for β .
4. Compare the variances of these estimators. When are they equal?
5. Interpret the result in terms of the weights assigned to the observations.

Exercise 2. We consider simple linear regression, *i.e.*, for $i \in \llbracket 1, n \rrbracket$,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $x_1, \dots, x_n$ are fixed real numbers that are not all equal, and the errors $\varepsilon_1, \dots, \varepsilon_n$ are independent, have mean zero and common variance σ^2 . We want to determine the line that best fits the observations in the least squares sense, that is, solve

$$(\hat{\beta}_{0,n}, \hat{\beta}_{1,n}) \in \arg \min_{(a,b) \in \mathbb{R}^2} \sum_{i=1}^n (Y_i - a - bx_i)^2.$$

We write

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{X})^2, \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{X})(Y_i - \bar{Y}).$$

1. Write down the design matrix $\mathbf{X}$ of this model. Show that it has rank 2 and deduce that the least squares estimator exists and is unique.
2. Give the two equations satisfied by $\hat{\beta}_{0,n}$ and $\hat{\beta}_{1,n}$ for the minimization problem. Use one of them to express $\hat{\beta}_{0,n}$ in terms of $\hat{\beta}_{1,n}$, $\bar{X}$ and $\bar{Y}$.

3. Establish the identities

$$S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{X}^2, \quad S_{xy} = \sum_{i=1}^n x_i Y_i - n\bar{X}\bar{Y}.$$

Deduce the expressions for $\hat{\beta}_{0,n}$ and $\hat{\beta}_{1,n}$.

4. Show that the fitted line passes through the centroid $(\bar{X}, \bar{Y})$ of the scatter plot and that the residuals sum to zero.

Exercise 3. We consider the homoscedastic linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}_\theta[\boldsymbol{\varepsilon}] = \mathbf{0}_n, \quad \text{Cov}_\theta[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n,$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a deterministic matrix of rank p , with $n > p$. Let $\hat{\boldsymbol{\beta}}_n$ denote the least squares estimator and

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top, \quad \mathbf{M} = \mathbf{I}_n - \mathbf{H}.$$

Recall that $\mathbf{H}$ and $\mathbf{M}$ are the matrices of the orthogonal projections onto $\text{Im}(\mathbf{X})$ and $\text{Im}(\mathbf{X})^\perp$, respectively.

1. Show that

$$\text{SCR} = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n\|^2 = \boldsymbol{\varepsilon}^\top \mathbf{M} \boldsymbol{\varepsilon}.$$

2. Using the properties of the trace, show that

$$\mathbb{E}_\theta[\text{SCR}] = \sigma^2 \text{tr}(\mathbf{M}).$$

3. Show that the trace of an orthogonal projection matrix equals its rank. Deduce the value of $\text{tr}(\mathbf{M})$.

4. Construct an unbiased estimator of σ^2 from SCR. Compute the bias of SCR/n and explain why we divide by $n - p$ rather than by n . Does this result require Gaussian errors?

Exercise 4. Let $p \geq 1$ be fixed. We consider a sequence of fixed-design regression models, *i.e.*, for every $n \geq p$, we consider the model

$$\mathbf{Y}_n = \mathbf{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \quad \mathbb{E}_\theta[\boldsymbol{\varepsilon}] = \mathbf{0}_n, \quad \text{Cov}_\theta[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n,$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathbf{X}_n \in \mathbb{R}^{n \times p}$ is a deterministic matrix whose i th row is denoted by $\mathbf{x}_{n,i}^\top$, and $\sigma^2 \in \mathbb{R}_+^*$. We assume that $\mathbf{X}_n$ has rank p for every $n \geq p$ and that

$$\frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n \xrightarrow{n \rightarrow \infty} \mathbf{Q},$$

where $\mathbf{Q}$ is a positive definite matrix. Let $\hat{\boldsymbol{\beta}}_n$ denote the least squares estimator.

Part I – Consistency.

1. Recall the expression for $\hat{\beta}_n$. Show that

$$\hat{\beta}_n - \beta = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \sum_{i=1}^n \mathbf{x}_{n,i} \varepsilon_i.$$

Deduce that $\hat{\beta}_n$ is unbiased and determine its covariance matrix.

2. Show that

$$\mathbb{E} \left[\left\| \hat{\beta}_n - \beta \right\|^2 \right] = \frac{\sigma^2}{n} \text{tr} \left\{ \left(\frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n \right)^{-1} \right\} \xrightarrow{n \rightarrow \infty} 0.$$

Deduce the consistency of $\hat{\beta}_n$.

Part II — Asymptotic normality. We now assume that the variables $(\varepsilon_i)_{i \geq 1}$ are *i.i.d.*, in addition to having mean zero and variance σ^2 , and that the design also satisfies

$$\max_{1 \leq i \leq n} \frac{\|\mathbf{x}_{n,i}\|}{\sqrt{n}} \xrightarrow{n \rightarrow \infty} 0.$$

3. Show that

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_{n,i} \varepsilon_i.$$

Why can the usual central limit theorem for *i.i.d.* variables not be applied directly to the sum in this expression?

4. Let $\mathbf{u} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$. Define

$$Z_{n,i} = \frac{\mathbf{u}^\top \mathbf{x}_{n,i} \varepsilon_i}{\sqrt{n}}.$$

Check that $Z_{n,1}, \dots, Z_{n,n}$ are independent, have mean zero and are square-integrable. Show that

$$\sum_{i=1}^n \text{Var}[Z_{n,i}] \xrightarrow{n \rightarrow \infty} \sigma^2 \mathbf{u}^\top \mathbf{Q} \mathbf{u} > 0.$$

5. Let $\delta > 0$. Show that

$$\sum_{i=1}^n \mathbb{E} \left[Z_{n,i}^2 \mathbf{1}_{\{|Z_{n,i}| > \delta\}} \right] \leq \left\{ \frac{1}{n} \sum_{i=1}^n (\mathbf{u}^\top \mathbf{x}_{n,i})^2 \right\} \mathbb{E} \left[\varepsilon_1^2 \mathbf{1}_{\left\{ |\varepsilon_1| > \frac{\delta \sqrt{n}}{\|\mathbf{u}\| \max_{1 \leq j \leq n} \|\mathbf{x}_{n,j}\|} \right\}} \right].$$

6. Deduce that the Lindeberg condition holds.

7. Apply the Lindeberg–Feller theorem and then the Cramér–Wold criterion to establish

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_{n,i} \varepsilon_i \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}_p(\mathbf{0}_p, \sigma^2 \mathbf{Q}).$$

Conclude that

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}_p(\mathbf{0}_p, \sigma^2 \mathbf{Q}^{-1}).$$

Part III — Why control the weights of the observations?

8. Consider the case $p = 1$ and the design

$$x_{n,1} = \sqrt{n}, \quad x_{n,i} = 0, \quad i \in \llbracket 2, n \rrbracket.$$

Check that the first assumption on the design holds, but not the second. Compute $\hat{\beta}_n - \beta$ and $\sqrt{n}(\hat{\beta}_n - \beta)$. If the errors take the values -1 and 1 with probability $1/2$ each, is the estimator consistent? Is it asymptotically normal at rate $\sqrt{n}$? Interpret.

Appendix

Lindeberg–Feller theorem. For each n , let $Z_{n,1}, \dots, Z_{n,n}$ be independent, mean-zero, square-integrable real random variables. Suppose that

$$\sum_{i=1}^n \text{Var}[Z_{n,i}] \xrightarrow{n \rightarrow \infty} s^2 \in]0, +\infty[$$

and that, for every $\delta > 0$,

$$\sum_{i=1}^n \mathbb{E} \left[Z_{n,i}^2 \mathbb{1}_{\{|Z_{n,i}| > \delta\}} \right] \xrightarrow{n \rightarrow \infty} 0.$$

Then,

$$\sum_{i=1}^n Z_{n,i} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, s^2).$$

Cramér–Wold theorem. Let $(\mathbf{Z}_n)_{n \geq 1}$ be a sequence of random vectors in $\mathbb{R}^p$ and let $\mathbf{Z}$ be a random vector in $\mathbb{R}^p$. Then,

$$\mathbf{Z}_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathbf{Z} \quad \Longleftrightarrow \quad \forall \mathbf{u} \in \mathbb{R}^p, \quad \mathbf{u}^\top \mathbf{Z}_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathbf{u}^\top \mathbf{Z}.$$