Session 1 : Exercises

1 Unidimensional model

Assumptions and notations The data is generated according to this model.

$$y_i = \theta_0^\star + \theta_1^\star x_i + \varepsilon_i,$$
$$\varepsilon_i \overset{i.i.d}{\sim} \varepsilon, \text{ for } i = 1, \dots, n$$
$$\mathbb{E}(\varepsilon) = 0$$

The statistical problem is to estimate the parameters $\widehat{\theta}_0$ and $\widehat{\theta}_1$ from the observations of y and x. with the *ordinary least squares* (OLS) method.

Exercice 1. Give the estimator for the coefficient for the basic model.
Solution. The OLS method consists in minimizing the sum of squared residuals, i.e., the differences between observed values and predicted values :

$$\min \sum_{i=1}^n \widehat{\varepsilon}_i^2 = \min \sum_{i=1}^n (y_i - \widehat{y}_i)^2 = \min \sum_{i=1}^n (y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i)^2 \equiv \min f(\widehat{\theta}_0, \widehat{\theta}_1),$$

Check the first order condition

$$\frac{\partial f}{\partial \widehat{\theta}_0} = 0, \quad \frac{\partial f}{\partial \widehat{\theta}_1} = 0.$$

From the first equation :

$$\frac{\partial f}{\partial \widehat{\theta}_0} = 0 \quad \Rightarrow \quad -2\sum_{i=1}^n (y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i) = 0 \quad \Rightarrow \quad \boxed{\widehat{\theta}_0 = \bar{y} - \bar{x}\widehat{\theta}_1}.$$

From the second equation,

$$\begin{aligned}
\frac{\partial f}{\partial \widehat{\theta}_1} = 0 &\iff -2\sum_{i=1}^n \left(y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i\right) x_i = 0 \\
&\iff \sum_{i=1}^n \left(y_i x_i - \bar{y} x_i - \widehat{\theta}_1 x_i^2 + \widehat{\theta}_1 x_i^2\right) = 0 \\
&\iff \boxed{\widehat{\theta}_1 = \frac{\sum_{i=1}^n y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}}
\end{aligned}$$

The Hessian of f is positive definite iff $\mathbb{V}\mathrm{ar}(x) > 0$, so this critical point is indeed a minimum iff $\mathbb{V}\mathrm{ar}(x) > 0$.

$$f(\widehat{\theta}_0, \widehat{\theta}_1) = \sum_{i=1}^{n}(y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i)^2,$$

$$\text{Hessian } H = \begin{bmatrix} \frac{\partial^2 f}{\partial \widehat{\theta}_0^2} & \frac{\partial^2 f}{\partial \widehat{\theta}_0 \partial \widehat{\theta}_1} \\ \frac{\partial^2 f}{\partial \widehat{\theta}_1 \partial \widehat{\theta}_0} & \frac{\partial^2 f}{\partial \widehat{\theta}_1^2} \end{bmatrix} = \begin{bmatrix} 2n & 2\sum_{i=1}^n x_i \\ 2\sum_{i=1}^n x_i & 2\sum_{i=1}^n x_i^2 \end{bmatrix},$$

$$\det(H) = 4\left(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2\right).$$

which is grater than 0 iff $\mathbb{V}\mathrm{ar}(x) > 0$.

Exercice 2. In the slides we give the estimator for the coefficient for the basic model when the data is centered.

Solution. Defining centered variables :

$$x_i^c = x_i - \bar{x}, \qquad y_i^c = y_i - \bar{y}, \qquad \widehat{y}_i^c = \widehat{y}_i - \bar{y}, \qquad \widehat{\varepsilon}_i = y_i^c - \widehat{y}_i^c,$$

the OLS method can be rewritten as :

$$y_i^c = \widehat{\theta}_0 + \widehat{\theta}_1 x_i^c + \widehat{\varepsilon}_i.$$

We want to minimize the sum of squared residuals using centered data : $\sum_{i=1}^n \widehat{\varepsilon}_i^2$.

$$\frac{\partial f}{\partial \widehat{\theta}_0^c} = -2\sum_{i=1}^{n}(y_i^c - \widehat{\theta}_0^c - \widehat{\theta}_1 x_i^c) = 0 \quad \Rightarrow \quad \widehat{\theta}_0^c = \bar{y^c} - \widehat{\theta}_1 \bar{x^c} = 0,$$

$$\frac{\partial f}{\partial \widehat{\theta}_1} = -2\sum_{i=1}^{n}(y_i^c - \widehat{\theta}_0^c - \widehat{\theta}_1^c x_i^c)x_i^c = 0 \quad \Rightarrow \quad \widehat{\theta}_1^c = \frac{\sum_{i=1}^n x_i^c y_i^c}{\sum_{i=1}^n (x_i^c)^2} = \frac{\mathrm{Cov}(x,y)}{\mathbb{V}\mathrm{ar}(x)} = \widehat{\theta}_1.$$

Since $\widehat{\theta}_0^c = 0$, the regression model with centered data simplifies to :

$$\widehat{y}_i^c = \widehat{\theta}_1 x_i^c$$

Exercice 3. In the slides we derive a state the relationship between the slope $\widehat{\theta}_1$ and the correlation $\mathrm{corr}_n(\mathbf{y}, \mathbf{x})$

Solution. The linear correlation coefficient $\mathrm{corr}_n(\mathbf{y}, \mathbf{x})$ measures the degree of linear covariation between y and x and is defined as :

$$\mathrm{corr}_n(\mathbf{y}, \mathbf{x}) = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathbb{V}\mathrm{ar}(x)\mathbb{V}\mathrm{ar}(y)}} = \frac{\frac{1}{n}\sum_{i=1}^n y_i x_i - \bar{y}\bar{x}}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^n y_i^2 - \bar{y}^2\right)\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2\right)}} = \widehat{\theta}_1 \frac{\sqrt{\mathbb{V}\mathrm{ar}(x)}}{\sqrt{\mathbb{V}\mathrm{ar}(y)}}$$

where $\widehat{\theta}_1 =$.

We can check these properties

1) $\mathrm{corr}_n(\mathbf{y}, \mathbf{x}) \in [-1; 1]$:
2) $\mathrm{corr}_n(\mathbf{y}, \mathbf{x})$ is dimensionless.
3) $\mathrm{corr}_n(\mathbf{y}, \mathbf{x})$ is symmetric : $\mathrm{corr}_n(\mathbf{y}, \mathbf{x}) = \mathrm{corr}_n(\mathbf{x}, \mathbf{y})$.
4) $\mathrm{corr}_n(\mathbf{y}, \mathbf{x})$ is not affected by a change of variable :

Exercice 4. Show that the variance decomposes as

$$\sum_{i=1}^{n}(y_i^c)^2 = \sum_{i=1}^{n}(\widehat{y_i^c})^2 + \sum_{i=1}^{n}\widehat{\varepsilon}_i{}^2.$$

Solution. Analysis of variance allows decomposing the total variance into explained variance and residual variance in order to measure the quality of the regression model. By definition, we have $\widehat{\varepsilon}_i = y_i - \widehat{y}_i \iff y_i^c = \widehat{\varepsilon}_i + \widehat{y_i^c}$ with $\widehat{y_i^c} = \widehat{\theta}_1 x_i^c$, hence :

$$\sum_{i=1}^{n}(y_i^c)^2 = \sum_{i=1}^{n}(\widehat{y_i^c})^2 + \sum_{i=1}^{n}\widehat{\varepsilon}_i^2 + 2\sum_{i=1}^{n}\widehat{y_i^c}\widehat{\varepsilon}_i.$$

Let's see what happens with $\sum_{i=1}^{n}\widehat{y_i^c}\widehat{\varepsilon}_i$

$$\begin{aligned}
\sum_{i=1}^{n}\widehat{\varepsilon}_i\widehat{y}_i^c &= \sum_{i=1}^{n}\left(y_i^c - \widehat{y}_i^c\right)\widehat{y}_i^c = \sum_{i=1}^{n}\left(y_i^c - \widehat{y}_i^c\right)\widehat{\theta}_1 x_i^c \\
&= \widehat{\theta}_1\left(\sum_{i=1}^{n}\left(y_i^c - \widehat{\theta}_1 x_i^c\right)x_i^c\right) = \widehat{\theta}_1\left(\sum_{i=1}^{n}y_i^c x_i^c - \widehat{\theta}_1\sum_{i=1}^{n}(x_i^c)^2\right).
\end{aligned}$$

Also, plug-in the slope in the previous equation

$$\widehat{\theta}_1 = \frac{\sum_{i=1}^{n}x_i^c y_i^c}{\sum_{i=1}^{n}(x_i^c)^2} \implies \sum_{i=1}^{n}x_i^c y_i^c - \widehat{\theta}_1 n\mathbb{V}\mathrm{ar}(x) = 0 \implies \sum_{i=1}^{n}\widehat{\varepsilon}_i\widehat{y}_i^c = 0.$$

Therefore

$$\sum_{i=1}^{n}(y_i^c)^2 = \sum_{i=1}^{n}(\widehat{y_i^c})^2 + \sum_{i=1}^{n}\widehat{\varepsilon}_i{}^2.$$

This is the analysis of variance equation that describes the decomposition of the total variability of the point cloud into explained variations and residual variations. Indeed :

$$\begin{aligned}
\mathrm{SST} &= \sum_{i=1}^{n}(y_i^c)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{(variance of } y\text{, sum of squares total)} \\
\mathrm{SSR} &= \sum_{i=1}^{n}(\widehat{y_i^c})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \quad \text{(variance of } \widehat{y}\text{, sum of squares regression)} \\
\mathrm{SSE} &= \sum_{i=1}^{n}\widehat{\varepsilon}_i{}^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \quad \text{(variance of } \varepsilon\text{, sum of squares error)},
\end{aligned}$$

with SST the total sum of squares, SSE the explained sum of squares (by the regression line), and SSR the residual sum of squares. We write the analysis of variance :

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}.$$

Exercice 5. Show that the R^2, defined as the ratio of the SSR and the SST, is equivalent to

$$1 - \frac{\sum_{i=1}^{n}\widehat{\varepsilon}_i{}^2}{\sum_{i=1}^{n}(y_i^c)^2}.$$

Solution. The ordinary least squares estimates for simple linear regression are

$$\widehat{\theta}_0 = \bar{y} - \widehat{\theta}_1\bar{x}, \quad \widehat{\theta}_1 = \frac{\mathrm{cov}(x,y)}{\mathbb{V}\mathrm{ar}(x)}.$$

The coefficient of determination R^2 for simple linear regression is defined as

$$R^2 = \frac{\sum_{i=1}^n (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\widehat{\theta}_0 + \widehat{\theta}_1 x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Substituting $\widehat{\theta}_0 = \bar{y} - \widehat{\theta}_1 \bar{x}$ yields

$$R^2 = \frac{\sum_{i=1}^n \left(\widehat{\theta}_1 (x_i - \bar{x})\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

This simplifies to

$$R^2 = \widehat{\theta}_1^2 \, \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2}{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2} = \widehat{\theta}_1^2 \, \frac{\mathbb{V}\mathrm{ar}(x)}{\mathbb{V}\mathrm{ar}(y)}.$$

Since

$$\widehat{\theta}_1 = \frac{\mathrm{cov}(x,y)}{\mathbb{V}\mathrm{ar}(x)},$$

then

$$R^2 = \left(\frac{\mathrm{cov}(x,y)}{\sqrt{\mathbb{V}\mathrm{ar}(x)\,\mathbb{V}\mathrm{ar}(y)}}\right)^2.$$

So we conclude

$$\boxed{R^2 = \mathrm{corr}(x,y)^2.}$$

2 Multidimensional case

Assumptions and notations We assume that the data is generated following the model : $\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}$ where :

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \theta_0^\star \\ \vdots \\ \theta_p^\star \end{pmatrix}}_{\boldsymbol{\theta}^\star} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

and so $y_i = \theta_0^\star + \sum_{j=1}^p \theta_j^\star x_{i,j} + \varepsilon_i$. Using the matrix notation, we can notate the following quantities as :

Total variance of the data (SST) :

$$\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|_2^2 = \sum_{i=1}^n (y_i - \bar{y})^2$$

The goal of OLS is

$$\widehat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^{p+1}} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^{p+1}} \sum_{i=1}^n \left[y_i - \left(\theta_0 + \sum_{j=1}^p \theta_j x_{i,j} \right) \right]^2.$$

The minimizer is unique iff

$$\mathrm{Ker}(X) = \{0\}.$$

Assuming this condition, the solution is denoted by $\widehat{\boldsymbol{\theta}}$, and the fitted values (predictions) are

$$\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\theta}}.$$

and the residual, the error in the prediction is (SSE) :

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - X\widehat{\boldsymbol{\theta}}$$

We can then decompose the variability as follows :

- Variance explained by the regression (SSR) :

$$\|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}_n\|_2^2 = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2$$

- Variance of the error (SSE) :

$$\|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Finally, the decomposition holds (proof done bellow) :

$$\underbrace{\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|_2^2}_{\text{SST}} = \underbrace{\|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}_n\|_2^2}_{\text{SSR}} + \underbrace{\|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2}_{\text{SSE}}.$$

Exercice 6. The solution of the normal equations is unique when the columns of X are linearly independent, i.e., iff X is full rank, $rank(X) = p+1$ iff its kernel is trivial : $\mathrm{Ker}(X) = \{0\}$. We want to show here that

$$\mathrm{Ker}(X) = \mathrm{Ker}(X^\top X).$$

Solution.

1) First, show $\mathrm{Ker}(X) \subseteq \mathrm{Ker}(X^\top X)$: Let $\boldsymbol{\theta} \in \mathrm{Ker}(X)$, then $X\boldsymbol{\theta} = 0$. Multiplying both sides by $X^\top$, we get

$$X^\top X\boldsymbol{\theta} = X^\top 0 = 0,$$

so $\boldsymbol{\theta} \in \mathrm{Ker}(X^\top X)$.

2) Next, show $\mathrm{Ker}(X^\top X) \subseteq \mathrm{Ker}(X)$: Let $\boldsymbol{\theta} \in \mathrm{Ker}(X^\top X)$, then $X^\top X\boldsymbol{\theta} = 0$. Consider the quadratic form

$$\boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta} = \|X\boldsymbol{\theta}\|^2 = 0.$$

Hence, $X\boldsymbol{\theta} = 0$, so $\boldsymbol{\theta} \in \mathrm{Ker}(X)$.

Exercice 7. Show that the residuals are centered, $\sum \widehat{\epsilon}_i = 0$.

Solution.

Recall that the OLS estimator $\widehat{\boldsymbol{\theta}}$ satisfies the normal equations

$$X^\top X\widehat{\boldsymbol{\theta}} = X^\top \mathbf{y}.$$

Equivalently,

$$X^\top (X\widehat{\boldsymbol{\theta}} - \mathbf{y}) = \mathbf{0}_{p+1},$$

that is,

$$X^\top \hat{\boldsymbol{\varepsilon}} = \mathbf{0}_{p+1}.$$

In words : the residual vector $\hat{\boldsymbol{\varepsilon}}$ is orthogonal to all the columns of X. Since the first column of X is the vector $\mathbf{1}_n$ (corresponding to the intercept), we get

$$\hat{\boldsymbol{\varepsilon}}^\top \mathbf{1}_n = \sum_{i=1}^{n} \widehat{\epsilon}_i = 0.$$

Exercice 8. Show that the variance decomposes as

$$\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 + \|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}_n\|^2. \tag{1}$$

Solution.

From the normal equations we know that the residuals are orthogonal to the fitted values :

$$\langle \hat{\boldsymbol{\varepsilon}}, \widehat{\mathbf{y}} \rangle = 0,$$

and, since $\mathbf{1}_n$ (the intercept column) belongs to the column space of X, we also have

$$\langle \hat{\boldsymbol{\varepsilon}}, \bar{\mathbf{y}}\mathbf{1}_n \rangle = 0.$$

Thus,

$$\langle \hat{\boldsymbol{\varepsilon}}, \widehat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}_n \rangle = 0.$$

Now write

$$\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n = (\mathbf{y} - \widehat{\mathbf{y}}) + (\widehat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}_n).$$

Taking squared norms and using the orthogonality relation above gives

$$\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 + \|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}_n\|^2.$$

This is exactly the variance decomposition formula (1).

Exercice 9. Show that $0 \leqslant R^2 \leqslant 1$ and

$$R^2 = 1 - \frac{\|\mathbf{y} - \widehat{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|^2} \tag{2}$$

Solution.

Recall the definition of R^2

$$R^2 = \frac{\|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}_n\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|^2}.$$

Reordering the Equality (1) :

$$\|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}_n\|^2 = \|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|^2 - \|\mathbf{y} - \widehat{\mathbf{y}}\|^2.$$

Dividing both sides by $\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|^2$, we obtain

$$R^2 = \frac{\|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}_n\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|^2} = 1 - \frac{\|\mathbf{y} - \widehat{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|^2}.$$