Multiple Linear Regression: exercises

Exercice 1 (Properies of LS estimators)
Consider the classical regression model $Y = X\beta + \varepsilon$.

- *Recall the basic assumptions*
- *Recall the expression for* $\hat{\beta}$
- *Give the expectation and variance for* $\hat{\beta}$
- *Consider the vector* $\hat{\varepsilon} = Y - X\hat{\beta}$*, give its expecation and variance*
- *Prove that* $\mathrm{Cov}(\hat{\varepsilon}, \hat{Y}) = 0$.

Exercice 2 (†Gauss-Markov theorem)
The Gauss Markov Theorem states that the least square estimator in a regression model is the best unbiased linear estimator. Prove the Gauss-Markov theorem, hint consider an other unbiased linear estimator $\tilde{\beta} = AY$ *and prove that for any fixe vector* $\alpha \in \mathbb{R}^p$, $\mathrm{V}(\alpha'\tilde{\beta}) \geq \mathrm{V}(\alpha'\hat{\beta})$.

Exercice 3 (Plotting variables)
We have a dependent variable Y *and an explanatory variable* X*. We have taken* $n = 2$ *observations and found*

$$(x_1, y_1) = (4, 5) \quad \textit{and} \quad (x_2, y_2) = (1, 5).$$

Plot the variables, estimate β *using the model* $y_i = \beta x_i + \varepsilon_i$ *then plot* $\hat{Y}$.
We now have a dependent variable Y *and 2 explanatory variables* X *and*Z*, we have taken* $n = 3$ *observations*

$$(x_1, z_1, y_1) = (3, 2, 0), \quad (x_2, z_2, y_2) = (3, 3, 5) \quad \textit{and} \quad (x_3, z_3, y_3) = (0, 0, 3).$$

Plot the variables, provide an estimate of β *using the model* $y_i = \beta x_i + \gamma z_i + \varepsilon_i$ *and plot* $\hat{Y}$.

Exercice 4 (Nested models)
Given X *a matrix of size* $n \times p$ *with* p *linearly independent vectors of* $\mathbb{R}^n$*. We note* X_q *the matrix consisting in the first* q $(q < p)$ *vectors of* X*. We have the following two models :*

$$\begin{aligned} Y &= X\beta + \varepsilon \\ Y &= X_q\gamma + \epsilon. \end{aligned}$$

Compare R^2 *between the two models.*

Exercice 5 *We look at how Y varies as a function of two independent variables x and z. We have n observations of these variables. We write $X = (\mathbb{1}\ x\ z)$ where $\mathbb{1}$ is the vector of constants and x, z are the vectors of the explanatory variables.*

1. *We obtain the following results :*

$$X'X = \begin{pmatrix} 30 & 0 & 0 \\ ? & 10 & 7 \\ ? & ? & 15 \end{pmatrix}.$$

   (a) *Provide the missing values.*

   (b) *What is n equal to?*

   (c) *Calculate the empirical linear correlation between x and z.*

2. *The empirical linear regression of Y on $\mathbb{1}, x, z$ yields*

$$Y = -2\mathbb{1} + x + 2z + \hat{\varepsilon}, \qquad \text{SSR} = \|\hat{\varepsilon}\|^2 = 12.$$

   (a) *Calculate the empirical mean $\bar{y}$.*

   (b) *Calculate the sum of squares explained by the regression (SSE), the total sum of squares (SST) and the coefficient of determination.*

Exercice 6 (Orthogonal regression)
Consider the linear regression model

$$Y = X\beta + \varepsilon,$$

where $Y \in \mathbb{R}^n$, X is a matrix of size $n \times p$ with p orthogonal vectors, $\beta \in \mathbb{R}^p$ and $\varepsilon \in \mathbb{R}^n$. Consider U the matrix of the first q columns of X and V the matrix of the last $p-q$ columns of X. Using OLS, we obtain the following estimates :

$$\begin{aligned} \hat{Y}_X &= \hat{\beta}_1^X x_1 + \cdots + \hat{\beta}_p^X x_p \\ \hat{Y}_U &= \hat{\beta}_1^U x_1 + \cdots + \hat{\beta}_q^U x_q \\ \hat{Y}_V &= \hat{\beta}_{q+1}^V x_{q+1} + \cdots + \hat{\beta}_p^V x_p. \end{aligned}$$

Note also $\text{SSE}(A)$ the norm squared of $P_A Y$.

1. *Show that $\text{SSE}(X) = \text{SSE}(U) + \text{SSE}(V)$.*

2. *Choose a variable called x_I and show that the estimate β_I remains the same no matter which model is used.*