EXERCISE CLASS : Linear regression

For \( i = 1, \ldots, n \), we consider \( y_i \in \mathbb{R} \) and \( x_i = (x_{i,0}, \ldots, x_{i,p})^T \in \mathbb{R}^{p+1} \) with \( x_{i,0} = 1 \). The OLS estimator is any coefficient vector \( \hat{\theta}_n = (\hat{\theta}_{n,0}, \ldots, \hat{\theta}_{n,p})^T \in \mathbb{R}^{p+1} \) such that

\[
\hat{\theta}_n \in \arg \min_{\theta \in \mathbb{R}^{p+1}} \sum_{i=1}^n (y_i - x_i^T \theta)^2.
\]

With the notations

\[
X = \begin{pmatrix}
X_1^T \\
\vdots \\
X_n^T
\end{pmatrix} = \begin{pmatrix}
x_{1,0} & \cdots & x_{1,p} \\
\vdots & \ddots & \vdots \\
x_{n,0} & \cdots & x_{n,p}
\end{pmatrix} \in \mathbb{R}^{n \times (p+1)}, \quad Y = \begin{pmatrix}
y_1 \\
\vdots \\
y_n
\end{pmatrix}.
\]

We have

\[
\hat{\theta}_n \in \arg \min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|. \tag{1}
\]

Let \( X = (1_n, \tilde{X}) \) and introduce \( \hat{\mu}_X = (\tilde{X}^T 1_n)/n \) and \( \hat{\mu}_Y = (1_n^T Y)/n \). Define the centred version of \( Y \) and \( \tilde{X} \), given by \( Y_c = Y - 1_n \hat{\mu}_Y \) and \( \tilde{X}_c = \tilde{X} - 1_n \hat{\mu}_X \), respectively. Consider the following alternative procedure:

\[
\hat{\theta}_{n,c} = \arg \min_{\theta \in \mathbb{R}^p} \|Y_c - \tilde{X}_c \theta\|, \tag{2}
\]

for which, the predictor at \( \tilde{x} \in \mathbb{R}^p \) is given by \( \hat{\mu}_X + (\tilde{x} - \hat{\mu}_X)^T \hat{\theta}_{n,c} \).

**Exercise 1.** Aim is to show that

\[
\min_{\theta \in \mathbb{R}^p} \|Y_c - \tilde{X}_c \theta\| = \min_{\theta \in \mathbb{R}^{p+1}} \|Y - X\theta\|.
\]

and, assuming that \( X \) has full rank, we have the following relationship between the traditional OLS and the OLS based on centred data,

\[
\hat{\theta}_{n,0} = \hat{\mu}_Y - \hat{\mu}_X \hat{\theta}_{n,c}, \quad (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,p}) = \hat{\theta}_{n,c}^T. \tag{3}
\]

Consequently, the 2 methods give the same predictor.

1. Start by obtaining that the inequality \( \geq \) holds true.
2. Then show that for any sequence \( (z_i) \), and for all \( z \in \mathbb{R} \), it holds that \( \|Z - z 1_n\| \geq \|Z - \bar{z} 1_n\| \), where \( Z = (z_1, \ldots, z_n) \) and \( \bar{z} = n^{-1} \sum_{i=1}^n z_i \).
3. Find \( \hat{\mu}_X \) such that, for any \( \theta_0 \in \mathbb{R} \) and \( \theta \in \mathbb{R}^p \), \( \|Y - \theta_0 1_n - \tilde{X} \theta\| \geq \|Y - \hat{\mu}_X 1_n - \tilde{X} \theta\| \).
4. Conclude that \( \min_{\theta \in \mathbb{R}^p} \|Y_c - \tilde{X}_c \theta\| = \min_{\theta \in \mathbb{R}^p, \theta_0 \in \mathbb{R}} \|Y - X(\theta_0, \theta^T)^T\| \).
5. Use the Lebesgue projection theorem to conclude that whenever \( \text{ker}(X) = \{0\} \), we have (3).

**Exercise 2** (on-line ols and cross-validation). The goal of this exercise is to show that the OLS estimator \( \hat{\theta}_n \) associated with design matrix \( X_n(\in \mathbb{R}^{n \times (p+1)}) \) and output \( y_n(\in \mathbb{R}^n) \) can be easily updated when a new pair of observation \( (x_{n,1}, y_{n,1}) \in \mathbb{R}^{(p+1)} \times \mathbb{R} \) is given. We apply the result to cross validation procedure in the end.

To clarify the notation:

\[
X_{(n,1)} = \begin{pmatrix}
X_{(n)} \\
x_{(n,1)}^T
\end{pmatrix} \in \mathbb{R}^{(n+1) \times (p+1)}, \quad \text{and} \quad y_{(n,1)} = \begin{pmatrix}
y_{(n)} \\
y_{(n,1)}
\end{pmatrix} \in \mathbb{R}^{n+1}
\]

We assume from now on that \( X_n(\in \mathbb{R}^{n \times (p+1)}) \) are full column rank (i.e., the columns of each matrix are independent vectors).

NB : Some of the questions require some computation (in particular obtaining (4) and (6)). Even if you could not prove it, it can be use later.
1) Let \( A, B, C, D \) be matrices with respective sizes \((d, d), (d,k), (k,k), (k,d)\). Show that if \( A \) and \( C \) are invertible, then
\[
(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1}.
\] (4)

2) Obtain that
\[
(X^T_{(n+1)}X_{(n+1)})^{-1} = \left(X^T_{(n)}X_{(n)}\right)^{-1} - \frac{\zeta_{n+1}s^T_{n+1}}{1 + b_{n+1}}
\] (5)
where \( \zeta_{n+1} = (X^T_{(n)}X_{(n)})^{-1}x_{n+1} \) and \( b_{n+1} = x^T_{n+1}(X^T_{(n)}X_{(n)})^{-1}x_{n+1} \).

3) Express \( X^T_{(n+1)}y_{(n+1)} \) with respect to \( X^T_{(n)}y_{(n)} \) and \( y_{n+1}x_{n+1} \).

4) Show that the OLS estimator \( \hat{\theta}_{n+1} \) associated with design matrix \( X_{(n+1)} \) and output \( y_{(n+1)} \) can be obtained as follows:
\[
\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{u_{n+1}}{1 + b_{n+1}} \zeta_{n+1},
\] (6)
where \( u_{n+1} = y_{n+1} - x^T_{n+1}\hat{\theta}_n \).

5) Keeping in memory \( (X^T_{(n)}X_{(n)})^{-1} \) and \( \hat{\theta}_n \), explain how to update \( \hat{\theta}_{n+1} \) using a minimal number of operations of the kind : matrix \((p+1,p+1)\) times vector \((p+1,1)\). How many such operations are needed?

6) Using Equation (5) above, show that
\[
1 + b_{n+1} = \frac{1}{1 - h_{n+1}}
\]
where \( h_{n+1} = x^T_{n+1}(X^T_{(n+1)}X_{(n+1)})^{-1}x_{n+1} \).

7) The prediction of \( y_{n+1} \) given by the model is \( \hat{y}_{n+1} := x^T_{n+1}\hat{\theta}_{n+1} \). With the following formula
\[
\hat{y}_{n+1} = x^T_{n+1}\hat{\theta}_n + \frac{u_{n+1}b_{n+1}}{1 + b_{n+1}}.
\]
prove that
\[
y_{n+1} - \hat{y}_{n+1} = u_{n+1}(1 - h_{n+1}).
\]

8) Given some data \((y, X)\), leave-one-out cross-validation consists in computing the risk
\[
R_{cv} = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T\hat{\theta}_{(-i)})^2
\]
where \( \hat{\theta}_{(-i)} \) is the OLS estimator based on \((y_{(-i)}, X_{(-i)})\), i.e., the data \((y, X)\) without the \( i \)-th line. Applying what have been done so far, show that
\[
R_{cv} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(1 - \hat{h}_i)^2,
\]
with \( \hat{h}_i = x_i^T(X^TX)^{-1}x_i \) and \( \hat{y}_i = x_i^T\hat{\theta}_n \).