In this problem sheet, we consider the problem of linear regression with p predictors and one intercept,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon\,,$$

where $\mathbf{y}^t = (y_1, \ldots, y_n)$ is the column vector of target values, $\beta^t = (\beta_0, \ldots, \beta_p)$ is the column vector of coefficients, $\epsilon^t = (\epsilon_1, \ldots, \epsilon_n)$ is the vector of random errors, and $\mathbf{X}$ is the $n \times (p+1)$ matrix of observations given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}.$$

We assume that ϵ has a multivariate normal distribution with covariance matrix $\sigma^2 I$. The case $p = 1$ is referred to as simple linear regression.

Problem 0.

(i) Show that the least square solution is given by $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$. Argue that if $\mathbf{X}$ has rank $p+1$, then $\mathbf{X}^t\mathbf{X}$ is indeed invertible.

(ii) Let $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{y}$, where $\mathbf{H}$ is referred to as the hat matrix. Denote by $\mathbf{x}_j$ the j-th column of $\mathbf{X}$. Show that the inner product between $\hat{\mathbf{y}} - \mathbf{y}$ and $\mathbf{x}_j$ is 0. Deduce a geometrical interpretation of the least square solution.

(iii) Explain with words the interpretation of the regression coefficient $\hat{\beta}_j$ in a multivariate linear regression setting, with non-orthogonal inputs.

(iv) Let $\mathbf{X} = \mathbf{QR}$ be the QR decomposition of $\mathbf{X}$. Give a geometrical interpretation of the matrix $\mathbf{Q}$. Then show that $\hat{\mathbf{y}} = \mathbf{Q}\mathbf{Q}^t\mathbf{y}$.

(v) Show that $\hat{\beta}$ is an unbiased estimator of β, and derive the covariance matrix of $\hat{\beta}$.

(vi) Show that $\hat{\sigma}^2 = \sum_i (y_i - \hat{y}_i)^2/(n-p-1)$ is unbiased for σ^2.

(vii) How would you test if a particular predictor is associated with the response variable? Under normally distributed errors ϵ, which test would you use?

(viii) What is the difference between a confidence interval and a prediction interval? Which one is wider and why?

(ix) Give a hypothesis test to detect outliers.

(x) How would you check other model assumptions such as normality and constant variance?

Problem 1.
The least square estimator $\hat{\beta}$ of β is optimal in a certain sense. This is made precise with the result below, known as the Gauss-Markov Theorem.

Gauss-Markov Theorem. The least square estimator $\hat{\beta}$ has minimum variance amongst all unbiased linear estimators of β.

Two remarks:

(i) Linear should be understood as linear with respect to $\mathbf{y}$, that is of the form $\mathbf{B}\mathbf{y}$, where $\mathbf{B}$ is some $(p+1) \times n$ matrix. This way, the LS estimator $\hat{\beta}$ is indeed linear since $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$.

(ii) It is possible to define a partial ordering in the class of real symmetric matrices. We say that $\mathbf{B}_1 \preceq \mathbf{B}_2$ if $\mathbf{B}_2 - \mathbf{B}_1$ is a positive semi-definite matrix. In other words, we have that for any vector $\mathbf{x}$, $\mathbf{x}^t\mathbf{B}_1\mathbf{x} \leq \mathbf{x}^t\mathbf{B}_2\mathbf{x}$. Equivalently, the matrix $\mathbf{B}_2 - \mathbf{B}_1$ has non-negative eigenvalues.

The goal of this problem is to prove the Gauss-Markov Theorem.

(a) Let $\tilde{\beta} = \mathbf{B}\mathbf{y}$ be another unbiased linear estimator of β. Show that $\mathbf{B}\mathbf{X} = \mathbf{I}$.

(b) Show that $\mathrm{Cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}) = 0$.

(c) Show that for two random vectors U and V, the covariance matrix Σ_{U+V} of the random vector $U + V$ satisfies

$$\Sigma_{U+V} = \Sigma_U + \Sigma_V + \mathrm{Cov}(U, V) + \mathrm{Cov}(V, U)\,.$$

(d) Deduce from (b) and (c) that

$$\Sigma_{\tilde{\beta}} = \Sigma_{\tilde{\beta}-\hat{\beta}} + \Sigma_{\hat{\beta}}\,,$$

and conclude.

Problem 2.
Using geometrical considerations and Pythagora's theorem, we saw during the lectures that for the general linear regression model with intercept, the Total Sum of Squares (TSS) can be decomposed as a sum of Explained Sum of Squares (ESS) and Residual Sum of Squares (RSS),

$$\begin{aligned} TSS = \sum (y_i - \bar{y})^2 &= \sum \left(\hat{y}_i - \bar{\hat{y}}\right)^2 + \sum (y_i - \hat{y}_i)^2 \\ &= ESS + RSS\,, \end{aligned}$$

where $\bar{y} = n^{-1}\sum y_i$ and $\bar{\hat{y}} = n^{-1}\sum \hat{y}_i$.

(a) Check that this decomposition holds using direct calculations.

Hint: You may use the fact that $\mathbf{X}^t(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$, and conclude from the first equation corresponding to the column of ones in $\mathbf{X}$ that $\bar{y} = \bar{\hat{y}}$.

(b) Show that in the case of simple linear regression, the R^2 coefficient defined as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

is equal to the square of the empirical correlation coefficient r, defined as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum(x_i - \bar{x})^2)(\sum(y_i - \bar{y})^2)}} .$$

Problem 3.
Show that for the simple linear regression model, the variance of the LS coefficient estimates are given by

$$\operatorname{Var} \hat{\beta}_0 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}}{\sum(x_i - \bar{x})^2} \right) \qquad \operatorname{Var} \hat{\beta}_1 = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} .$$

Note that from the expression of $\operatorname{Var} \hat{\beta}_1$, the more variability there is in the xs, and the less there is in the estimation of the slope: we get a better estimate as the input variable is more spread out. This makes sense, right?

Problem 4.
In linear regression, the k-th diagonal element h_{kk} of the projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, referred to as the leverage of x_k, quantifies how much observation k contributes to the LS estimate.

(a) Show that for simple linear regression, the leverage corresponding to the k-th observation can be written

$$h_{kk} = \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} ,$$

Conclude that necessarily, h_{kk} belongs to the interval $[1/n, 1]$. Which observation corresponds to the smallest leverage?

(b) Show that for a multivariate regression model, $0 \leq h_{kk} \leq 1$.

Hint: Use that fact that $\mathbf{H}$ is idempotent and symmetrical.

(c) If $h_{kk} = 0$ or 1, then $h_{kj} = 0$ for all $j \neq k$.

(d) For all $j \neq k$, $-1/2 \leq h_{kj} \leq 1/2$.

Problem 5.
Consider multivariate linear regression with independent errors $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

(a) Show that the least square estimator of β is equal to the maximum likelihood estimator.

(b) What is the maximum likelihood estimator of σ^2? Compare its expression with the unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \,.$$

Problem 6.
Prove the Sherman-Morisson-Woodbury Theorem: for any non-singular $p \times p$ matrix $\mathbf{A}$ and $p \times 1$ column vectors $\mathbf{u}$ and $\mathbf{v}$,

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^t)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^t\mathbf{A}^{-1}}{1 + \mathbf{v}^t\mathbf{A}^{-1}\mathbf{u}} \,.$$

Problem 7.
We consider simple linear regression, with one predictor and an intercept. The LS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed from a training sample $\{(x_i, y_i)\}$ of size n.

(i) Show that $\hat{\beta}_0 + \hat{\beta}_1 x \sim \mathcal{N}(\beta_0 + \beta_1 x, \gamma_n \sigma^2)$, where

$$\gamma_n = \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \,.$$

(ii) Deduce from *(i)* a confidence interval for the mean response at x.

(iii) How would you modify the confidence interval in *(ii)* to obtain a prediction interval for the response variable y at x?

Problem 8.
Consider two linear regression models

Model 1: $Y = \beta_0 + \beta_1 X + \epsilon$,
Model 0: $Y = \beta_0 + \epsilon$,

We want to test if the slope is needed in our model (simpler Model 0 preferred over Model 1). That is, we want to test for the null hypothesis $H_0: \ \beta_1 = 0$. For this we have two options: a t-test of an F-test. The goal of this problem is to show that in this simple setting, the two tests are equal.

(i) t-test. Let $\mathbf{X}$ be the $(n \times 2)$ the matrix of observations, the first column being a column of ones, and $\mathbf{y}^t = (y_1, \ldots, y_n)$ be the column vector of outputs in the training data set. Consider Model 1,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \,,$$

where $\beta^t = (\beta_0, \beta_1)$, and ϵ has a multivariate normal distribution with covariance matrix $\sigma^2 I$. We denote by $\hat{\beta}^t = (\hat{\beta}_0, \hat{\beta}_1)$ the least square estimate of β. Let v_{ij} be the entry in the i-th row and j-th column of the 2×2 matrix $(\mathbf{X}^t\mathbf{X})^{-1}$. Show that

$$z := \frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{v_{22}}} \sim t_{n-2}\,,$$

where $\hat{\sigma}^2$ is an unbiased estimator of σ^2

(ii) F-test. Denote by RSS_1 (respectively RSS_0) the residual sum of squares of the larger model (respectively, of the reduced model). We saw during the lectures that

$$F = \frac{(RSS_0 - RSS_1)}{RSS_1/(n-2)} \sim F_{1,n-2}\,.$$

Show that we can re-express F as

$$F = \frac{\hat{\beta}_1^2}{\hat{\sigma}^2}\sum(x_i - \bar{x})^2\,.$$

(iii) Conclude that the t-test and F-test are equal.

Problem 9.
One of the goals of linear regression is to predict the value of a target variable y for a new observation x. Let $\mathbf{x}_{n+1}^t = (1,\, x_{n+1,1}, \ldots,\, x_{n+1,p})$ be a new observation. We model the response variable by

$$y_{n+1} = \mathbf{x}_{n+1}^t\beta + \epsilon_{n+1}\,,$$

where $\mathbf{E}\,\epsilon_{n+1} = 0$, $\mathsf{Var}\,\epsilon_{n+1} = \sigma^2$, and $\mathsf{Cov}(\epsilon_{n+1}, \epsilon_i) = 0$, for $i = 1, \ldots, n$. The target value y_{n+1} is predicted using $\hat{y}_{n+1} = \mathbf{x}_{n+1}^t\hat{\beta}$, where $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$ is the least square estimate of β computed from the training data. The prediction error is defined as $\hat{\epsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1}$.

(a) Show that

$$\begin{aligned}\mathbf{E}\,\hat{\epsilon}_{n+1} &= 0\,,\\ \mathrm{Var}\,\hat{\epsilon}_{n+1} &= \sigma^2\left(1 + \mathbf{x}_{n+1}^t(\mathbf{X}^\mathbf{t}\mathbf{X})^{-1}\mathbf{x}_{n+1}\right)\,.\end{aligned}$$

(b) Show that the expression of the variance found in *(a)* in the case of simple linear regression $(p = 1)$ can be written as

$$\mathrm{Var}\,\hat{\epsilon}_{n+1} = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\,.$$

(c) For which value of x_{n+1} is the variance of the prediction error minimum? What is the value of the variance in that case?

Our goal is to generalise the result found in *(c)* in the multiple linear regression setting. We adopt the following notation

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{z}_1^t \\ \vdots & \vdots \\ 1 & \mathbf{z}_n^t \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{Z}_1 & \dots & \mathbf{Z}_p \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{Z} \end{pmatrix},$$

where the $\mathbf{Z}_j$ are column vectors and $\mathbf{Z}$ is a $n \times p$ matrix. The column means of $\mathbf{Z}$ are put into a vector $\bar{\mathbf{x}}^t = (\bar{x}_1, \dots, \bar{x}_p)$.

(d) Express $\mathbf{X}^t\mathbf{X}$ as a 2×2 block matrix, in terms of $\mathbf{Z}$, $\bar{\mathbf{x}}$ and n.

We recall the inversion formula for block matrices. Let $\mathbf{M}$ be an invertible matrix, such that

$$\mathbf{M} = \begin{pmatrix} \mathbf{T} & \mathbf{U} \\ \mathbf{V} & \mathbf{W} \end{pmatrix},$$

with $\mathbf{T}$ invertible. Then $\mathbf{Q} = \mathbf{W} - \mathbf{V}\mathbf{T}^{-1}\mathbf{U}$ is invertible and

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{T}^{-1} + \mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1} & -\mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1} \\ -\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1} & \mathbf{Q}^{-1} \end{pmatrix}.$$

(e) Express $(\mathbf{X}^t\mathbf{X})^{-1}$ as a 2×2 block matrix, in terms of n, $\bar{\mathbf{x}}$ and some matrix Γ^{-1}. Give an expression of Γ in terms of $\mathbf{Z}$, $\bar{\mathbf{x}}$ and n.

(f) Let $\mathbf{x}_{n+1}^t = (1 \;\; \mathbf{z}_{n+1}^t)$ be a new observation. Show that the variance of the prediction error is

$$\operatorname{Var} \hat{\epsilon}_{n+1} = \sigma^2 \left(1 + \frac{1}{n} + \frac{1}{n} (\mathbf{z}_{n+1} - \bar{\mathbf{x}})^t \Gamma^{-1} (\mathbf{z}_{n+1} - \bar{\mathbf{x}}) \right).$$

(g) Assume that Γ is symmetric positive definite. For which value of $\mathbf{x}_{n+1}$ is the variance of the prediction error minimal? What is the value of the variance in that case?

(h) Show that if $\mathbf{X}^t\mathbf{X}$ is invertible, then Γ is indeed symmetric positive definite.

Problem 10.

The *UK Building Research Station* collected data on weekly gaz consumption and average external mean temperature in a district in South-East England over a few months. A linear regression explaining gaz consumption as a function of the temperature is carried out in R:

(i) Write down the model and its assumptions.

(ii) Fill in the blanks in the R output.

(iii) Let $Y \sim t_{28}$. What is $\mathbf{P}(|Y| > 11.04)$?

(iv) Describe the test associated with the row `Temp` in the R output (the null hypothesis, the alternative, the law under the null, the decision rule).

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.11082  0.02672  0.25294  0.63803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72385    0.12974     ?     < 2e-16 ***
Temp        -0.27793       ?     -11.04 1.05e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-Squared: 0.8131,     Adjusted R-squared: 0.8064
F-statistic: 121.8 on 1 and 28 DF,  p-value: 1.046e-11
```

(v) Give an interpretation for `Multiple R-squared: 0.8131`.

(vi) Give an estimate of the variance of the error term in the simple linear regression model.

(vii) Explain and interpret the last line

`F-statistic: 121.8 on 1 and 28 DF, p-value: 1.046e-11`

(viii) Do you believe that the outside temperature has an effect of gaz consumption? Justify your answer.

Problem 11.
We are interested in the model $Y = X\beta + \epsilon$ under usual conditions. We obtained the following fit, on a learning sample of size $n = 21$:

$$\hat{y} = 6.683_{(2.67)} + 0.44_{(2.32)}x_1 + 0.425_{(2.47)}x_2 + 0.171_{(2.09)}x_3 + 0.009_{(2.24)}x_4\,,$$

and $R^2 = 0.54$. For each coefficient, the number between brackets represents the absolute value of the test statistic.

(i) What are the assumptions made on the model?

(ii) Test for $\beta_1 = 0$ at 5% level.

(iii) Can you test for $\beta_3 = 1$ against the two-sided alternative $\beta_3 \neq 1$?

(iv) Test for $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ at 5% level.