Problem 1. Define $\mathbf{H} \coloneqq \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, where $\mathbf{X}$ is a non-stochastic $n \times p$ full rank matrix with $p \leq n$. Show that

1. $\mathbf{H}$ is idempotent and symmetric, meaning that $\mathbf{H}^2 = \mathbf{H}$ and $\mathbf{H}^\top = \mathbf{H}$.
2. the eigenvalues of $\mathbf{H}$ are either 0 or 1.
3. $\mathbf{H}$ is a projection matrix onto the column space of $\mathbf{X}$, $\mathscr{S}(\mathbf{X})$. Is this still the case if the columns of $\mathbf{X}$ are not linearly independent?
4. the trace of $\mathbf{H}$, $\mathrm{tr}(\mathbf{H})$, is equal to p and thus $\mathrm{rank}(\mathbf{H}) = p$.

Problem 2. Show that orthogonal projection matrices[1] are unique: if $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal projection matrices onto a subspace $\mathscr{V}$ of $\mathbb{R}^n$, then $\mathbf{P} = \mathbf{Q}$.

Problem 3. Suppose the $n \times p$ full-rank design matrix $\mathbf{X}$ ($n \geq p$) can be written as $[\,\mathbf{X}_1\ \mathbf{X}_2\,]$ with blocks $\mathbf{X}_1$, an $n \times p_1$ matrix, and $\mathbf{X}_2$, an $n \times p_2$ matrix. Show that $\mathbf{H} - \mathbf{H}_1$ is an orthogonal projection matrix. ($H_1 = X_1(X_1^\top X_1)^{-1}X_1^\top$)

Problem 4. Suppose that $A, X \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$. Show that

1. $\frac{\partial}{\partial x}Ax = A^\top$;
2. $\frac{\partial}{\partial x}x^\top Ax = (A + A^\top)x$; $\quad$ [Note the special case $\frac{\partial}{\partial x}x^\top x = 2x$.]
3. $\frac{\partial}{\partial X}\mathrm{tr}(X) = I_n$.

Problem 5. Let $\mathbf{X}$ be an $n \times p$ full rank real matrix with $p \leq n$ and Ω an $n \times n$ positive definite matrix, meaning that $\mathbf{v}^\top\Omega\mathbf{v} > 0$ for all $\mathbf{v} \in \mathbb{R}^n \setminus \{0_n\}$.

1. Show that $\mathbf{B} = \mathbf{X}^\top\Omega\mathbf{X}$ is positive definite and thus invertible. Deduce from this fact that $\mathbf{X}^\top\mathbf{X}$ is invertible.
2. Show that $\mathbf{B}$ is not necessarily invertible if we only assume that Ω is real, symmetric and invertible.

Problem 6. Let $Y_1, \ldots, Y_n$ be i.i.d. from $\mathcal{N}(\mu, \sigma^2)$.

Show that the log-likelihood satisfies

$$\ell(\mu, \sigma^2) = -\frac{1}{2}\left\{ n\log\sigma^2 + \frac{1}{\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2 \right\} + \text{const}$$

and the maximum likelihood (ML) estimates of μ and σ^2 are

$$\hat{\mu} = \bar{y} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2.$$

Problem 7. Let Σ be an $p \times p$ positive definite covariance matrix. We define the precision matrix $\mathbf{Q} = \Sigma^{-1}$. Suppose the matrices are partitioned into blocks,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \text{ and } \Sigma^{-1} = \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix}$$

with $\dim(\Sigma_{11}) = k \times k$ and $\dim(\Sigma_{22}) = (p-k) \times (p-k)$. Prove the following relationships

(a) $\Sigma_{12}\Sigma_{22}^{-1} = -\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}$

(b) $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \mathbf{Q}_{11}^{-1}$

(c) $\det(\Sigma) = \det(\Sigma_{22})\det(\Sigma_{1|2})$ where $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

[1] Note: the projection is orthogonal, not the matrix — the latter is not invertible if $p < n$! The three defining properties of an orthogonal projection matrix onto $\mathcal{V}$ are (1) $\mathbf{P}\mathbf{v} = \mathbf{v}$ for any $\mathbf{v} \in \mathcal{V}$, (2) symmetry and (3) idempotency.

Problem 8. Let $Y \sim \mathcal{N}_n(\mu, \Sigma)$ and consider the partition

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Y_1 is a $k \times 1$ and Y_2 is a $(n-k) \times 1$ vector for some $1 \leq k < n$. Show that the conditional distribution of $Y_1 \mid Y_2 = y_2$ is $\mathcal{N}_k(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{1|2})$ and $\Sigma_{1|2}$ is the Schur complement of Σ_{22}.

Hint: write the joint density as $p(y_1, y_2) = p(y_1 \mid y_2)p(y_2)$ and express the joint density in terms of the precision matrix $\mathbf{Q}$. It suffices to consider terms in $p(y_1, y_2)$ that depend only on y_1 (why?). The conditional distribution can then be identified by its functional form directly.

Problem 9. Let $Z \sim \mathcal{N}_n(0_n, \mathbf{I}_n)$ and $Y \sim \mathcal{N}_n(\mu, \Sigma)$ with Σ positive definite.

(a) Let $\mathbf{A}$ be an orthogonal matrix. Show that $\mathbf{A}^\top Z \sim \mathcal{N}_n(0_n, \mathbf{I}_n)$.

(b) Show that $\mathbf{C}^{-1}(Y - \mu) \sim \mathcal{N}_n(0_n, \mathbf{I}_n)$ where $\mathbf{C}$ is the Cholesky root of Σ, the unique lower triangular matrix with positive diagonal elements such that $\Sigma = \mathbf{C}\mathbf{C}^\top$.

(c) Let $\mathbf{H}$ be a $n \times n$ projection matrix of rank $k \leq n$ with real entries. Show that $Z^\top \mathbf{H} Z \sim \chi^2(k)$.

(d) Show that $(Y - \mu)^\top \Sigma^{-1}(Y - \mu) \sim \chi^2(n)$.

(e) Let $\mathbf{A}$ be a non-negative definite matrix. If $\mathbf{A}\Sigma\mathbf{A} = \mathbf{A}$, then show that $(Y - \mu)^\top \mathbf{A}(Y - \mu) \sim \chi^2(k)$, where $k = \operatorname{tr}(\mathbf{A}\Sigma)$.

Problem 10. Consider a singular value decomposition (SVD) of the design matrix $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U}$ is an $n \times p$ orthonormal matrix (meaning $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$ and the columns of $\mathbf{U}$ are orthogonal vectors), $\mathbf{D}$ is an $p \times p$ diagonal matrix and $\mathbf{V}$ is an $p \times p$ orthogonal matrix. Show that the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ does not depend on $\mathbf{V}$.

Problem 11. (Non-linear $\leftrightarrow$ linear models). This exercise has the goal of showing that a non-linear model can (sometimes) be transformed into a linear one. For instance, the model $y = \beta_1(x + \beta_3)^{\beta_2}(\varepsilon^2 + 1)$ can be written as

$$\log(y) = \underbrace{\log(\beta_1)}_{\beta_1^*} + \underbrace{\beta_2}_{\beta_2^*} \log(x + \beta_3) + \underbrace{\log(\varepsilon^2 + 1)}_{\varepsilon^*},$$

with β_3 fixed, and $\begin{bmatrix} 1 & \log(x + \beta_3) \end{bmatrix}$ as design matrix. Moreover, we need $\beta_1 > 0, x + \beta_3 > 0$ in order to do the transformation.

Write, when possible, the following models as linear regressions, either by transforming and/or by fixing some parameters. Specify the new parameter (β^*), the new error (ε^*), restrictions (e.g. $\beta_1 > 0$) and give the design matrix, as in the example above:

a) $y = \beta_0 + \beta_1/x + \beta_2/x^2 + \varepsilon$

b) $y = \beta_0/(1 + \beta_1 x) + \varepsilon$

c) $y = \beta_0/(\beta_1 x) + \varepsilon$

d) $y = 1/(\beta_0 + \beta_1 x + \varepsilon)$

e) $y = \beta_0 + \beta_1 x^{\beta_2} + \varepsilon$

f) $y = \beta_0 + \beta_1 x_1^{\beta_2} + \beta_3 x_2^{\beta_4} + \varepsilon$

g) $y = \beta_1 x_1^{\beta_2} \cos(x_2)^{\beta_3} \varepsilon$

h) $y = \beta_1 + x_1^{\beta_2}(2 + \cos(x_2))^{\beta_3}(\varepsilon^2 + 1)$

Problem 12. Let $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1 \ldots, n$.

a) Write down the design matrix $\mathbf{X}$. Calculate the elements of $\mathbf{X}^\top\mathbf{X}$, $\mathbf{X}^\top Y$ and $(\mathbf{X}^\top\mathbf{X})^{-1}$.

b) Show that $\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. How do you interpret the estimate?

Problem 13. (Factors and Interactions – Linear Models in `R`)

In `R`, a model formula has the following general form `response~expression`. The right-hand side expression follows certain rules. For example, intercept is present unless removed by `-1` and powers have to be designated

with `I(x^2)`. For example, `y ~ x+I(x^2)-1` defines a model where `y` depends on `x` quadratically and the intercept is set to zero.

For this exercise, suppose that

$$\texttt{y} = \begin{pmatrix} 217 \\ 143 \\ 186 \\ 121 \\ 157 \\ 143 \end{pmatrix}, \quad \texttt{x} = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \texttt{a} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{pmatrix}, \quad \texttt{b} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 3 \end{pmatrix}.$$

We can assign a toy meaning to this toy data set for illustration purposes: let y_j be the stress level of the j-th measured individual. We would like to model the mean stress level based on the number of children the individual has (denoted x_j), the sex of the individual (denote a_j and labeled 1 for female and 2 for male), and the marital status of the individual (denoted b_j and labeled 1 for single, 2 for married and 3 for divorced). Notice that the values in vectors `a` and `b` are only labels here (denoting groups, classes, or levels).

a) A *factor* is a categorical/qualitative variable, which may not have a numerical meaning (e.g. a group-allocating variable such as `a` and `b`). For example, consider the following model of stress value based on sex only:

$$y_j = \beta_0 + \alpha_1 + \varepsilon_j, \quad j = 1, 2, 3; \qquad y_j = \beta_0 + \alpha_2 + \varepsilon_j, \quad j = 4, 5, 6;$$

i.e. the mean stress value is allowed to be different for males and females. We can write the model in a single equation using indicators:

$$y_j = \beta_0 + \alpha_1 \mathbb{1}_{(a_j=1)} + \alpha_2 \mathbb{1}_{(a_j=2)} + \varepsilon_j, \tag{1}$$

where $\mathbb{1}_E = 1$ if the expression E is true, and 0 otherwise.

I. Give the design matrix corresponding to model (1).

II. Notice that this matrix is *not* full-rank. What is the consequence on the parameters estimation?

III. Suppress the column corresponding to α_1 of this matrix in order to have a full-rank matrix. What is now the interpretation of the parameters β_0 and α_2?

IV. When the model includes the constant β_0, R automatically suppresses the first level of each factor. Give the design matrix corresponding to the following models:

(i) `y~b`, (ii) `y~x+a`, (iii) `y~a+b`

b) An *interaction* of two variables (say `a` and `x`) is written in `R` as `a:x` or `a*x`. Adding the interaction term `a:x` to the model `y~a+x`, i.e. forming the model `y~a+x+a:x` adds *product effect(s)* between the two variables into the model, e.g.

$$y_j = \beta_0 + \alpha_2 \mathbb{1}_{(a_j=2)} + \beta_1 x_j + \beta_2 x_j \mathbb{1}_{(a_j=2)} + \epsilon_j \,,$$

where the term $\beta_2 x_j \mathbb{1}_{(a_j=2)}$ was added by the interaction. Note that `a*x` is a shorthand for `y~a+x+a:x`, i.e. the operator '`*`' adds both the main terms and the interaction term to the model. This is convenient, because one is very rarely interested in having the interaction term without the main terms.

Assuming existence of a new continuous regressor (a new continuous variable) $\texttt{z} = (0, 1, 5, 2, 1, 1)^\top$, write down the regression function (a mathematical expression for $\mathbb{E}y_j$) of the following models and find the design matrices corresponding to those models.

(i) `y~x*b`, (ii) `y~x*z` (iii) `y~a*b`, (iv) `y~z+I(x^2)`.

c) Assuming further that we have many more observations than those $n = 6$ given above, write down the regression function of `y~x*a*b`.

d) Explain the difference between considering an ordinal variable (such as `b`) as a factor and considering it as a numerical variable:

(i) `y~as.factor(b)`, (ii) `y~as.numeric(b)`.

What happens when we use variable `a` instead of `b`?

Problem 14. (Confounders and Simpson's paradox) In this exercise, we are interested in the dependence of a standardized test *percentile* on the grade point average (*GPA*) of students of a certain high school in the US. The data file `percentile.RData` also contains the variable *grade*, which determines the study age of the students.

a) Load the data and create a scatterplot of *percentile* on *GPA*.

b) Fit the linear model `percentile~GPA` and add the regression line to your scatterplot from part a). What would be your conclusion about the relationship of *percentile* on *GPA* based on this model? How does the model quantify this relationship? Does this make sense?

c) Add the variable *grade* to the model as a factor. How does this change your qualitative conclusions? How does the new model quantify the dependency? Are the conclusions sensible now?

d) Add the interaction term between *GPA* and *grade* to your model. What is now different compared to part c)?

Problem 15. Assume a linear model was developed for the blood glucose concentration (Y) of a patient after giving u units of a medicament to the patient with weight w and sex g (0=male, 1=female). In this model, the effect of weight w and the medicament dose u on the glucose concentration Y is different for males and females. Contrarily, the increase of the medicament dose u by 1 has (for two patients of the same sex and weight) the same effect on Y regardless of the (actual value of the) weight of the patient.

a) Write down the regression function of the model, such that the model has the interpretation above.

b) Assume the first observation is based on a male, 80 kg, who was given 10 units of the medicament. The second observation is based on a female, 60 kg, who was given 8 units of the medicament. Write down the first two rows of the design matrix.

c) How would you test whether weight w has different effect on Y based on the sex g?

Problem 16. Suppose the $n \times p$ full-rank design matrix $\mathbf{X}$ can be partitioned into two blocks as $[\,\mathbf{X}_1\ \mathbf{X}_2\,]$ and let $\mathbf{M}_{\mathbf{X}_1} := \mathbf{I}_n - \mathbf{H}_{\mathbf{X}_1}$. Show that $\mathbf{H}_{\mathbf{X}} = \mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$, where $\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$ is the projection on to the span of $\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2$. (Draw a 3D picture to visualize what this result actually says.)

Problem 17. (Forecast and confidence intervals).

The following table gives the estimations, the standardised errors and the correlations for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ adjusted for $n = 13$ cement data of the example given at the course.

```
            Estimate         SE            Correlations of Estimates
(Intercept)    48.19      3.913              (Intercept)    x1    x2
        x1      1.70      0.205        x1       -0.736
        x2      0.66      0.044        x2       -0.416   -0.203
        x3      0.25      0.185        x3       -0.828    0.822  -0.089
```

a) Explain how we can compute the standardised errors and correlations in the table above.

b) For this model, what is the forecast of y for $x_1 = x_2 = x_3 = 1$? How much would the prediction increase if $x_1 = 5$? And if $x_1 = x_2 = 5$?

c) For this model, compute, using only the information above and the fact that the quantiles are $t_9(0.975) = 2.262$ and $t_9(0.95) = 1.833$, the 0.95 confidence intervals for β_0, β_1, β_2 and β_3. Compute also a 0.90 confidence interval for $\beta_2 - \beta_3$.

Problem 18. (Linear Gaussian models and space rotations) Let

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

be a Gaussian linear model, where X is injective, and $\varepsilon \sim N(0, \sigma^2 I)$. We know that if A is an orthogonal matrix, then $\tilde{Y} = AY$ follows a linear Gaussian model as well,

$$\tilde{Y} \sim \mathcal{N}(\tilde{X}\beta, \sigma^2 I),$$

with $\tilde{X} = AX$. We will consider some particular cases of the orthogonal matrix A:

I. $A = U^\top$, where $X = U\Lambda V^\top$ is the singular values decomposition of X.

II. $A = Q^\top$, where $X = QR$ is the QR decomposition of X

For each of these cases,

a) Compute the adjusted values $\hat{\tilde{y}}$ as functions of $\tilde{y}$. What can we say about their first p coordinates? And about their last $n - p$ coordinates?

b) Compute the residuals of model $\tilde{Y}$. What can we say about their first p residuals? And about their last $n - p$ residuals?

c) Recall that residuals are usually dependent. What do we notice here?

Hint: Start by computing the *hat matrix* $\tilde{H}$ for both cases I. and II.

Problem 19. (The best design)

Let us consider the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\beta_0, \beta_1 \in \mathbb{R}$, $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2 I_n$ (and $n \geq 2$).

a) Find the design matrix corresponding to this model and give a necessary and sufficient condition for it to be full rank.

b) Find the covariance matrix of the least squares estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^\top$.

c) Let us suppose that we can design the experiment by choosing $x_i \in [-1, 1]$ arbitrarily. Which is the best choice of x_i that minimises the variance of $\hat{\beta}_1$?

Problem 20. (Reformulation of the Gauss-Markov theorem)

Let $Y = X\beta + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0, \mathrm{var}(\varepsilon) = \sigma^2 I$. Let $\hat{\beta}$ be the least squares estimator of β, and $\tilde{\beta}$ another *linear* and *unbiased* estimator of β.

Show that

$$\mathrm{MSE}(c^\top \tilde{\beta}) \geq \mathrm{MSE}(c^\top \hat{\beta}), \quad \forall c \in \mathbb{R}^p,$$

is equivalent to the conclusion of the Gauss-Markov theorem. Here, $\mathrm{MSE}(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2)$ is the mean square error of $\hat{\theta}$.

Recall: $\mathrm{MSE}(\hat{\theta}) = \mathrm{bias}(\hat{\theta})^2 + \mathrm{var}(\hat{\theta})$.

Problem 21. (Diagnostic graphics)

a) Figure 1 represents the standardised residuals as a function of values adjusted for the linear model derived from four different datasets. For each case, discuss the adjusting and explain briefly how you would try to remedy the possible insufficiency.

b) Figure 2 shows four Q-Q Gaussian plots. In all the cases, the data do not follow the Gaussian distribution. In fact, the data are generated from a distribution with

   i) tails haevier than Gaussian tails;

   ii) tails lighter than Gaussian tails;

   iii) a positive skewness coefficient;

   iv) a negative skewness coefficient.

   Associate each case i)–iv) with a Q-Q plot of Figure 2.

Problem 22. (QQ plots)

The goal of this exercise is to justify the use of the QQ plot to "see" whether a sample $x_1, \dots, x_n$ comes from the normal distribution. Let $X_1, \dots, X_n \sim N(0, 1)$ be i.i.d, and let Φ be the cumulative distribution function of the normal law $N(0, 1)$.

1. Show that $\Phi(X_1), \dots, \Phi(X_n) \sim U([0, 1])$ are i.i.d., where $U([0, 1])$ denotes the uniform law on $[0, 1]$.

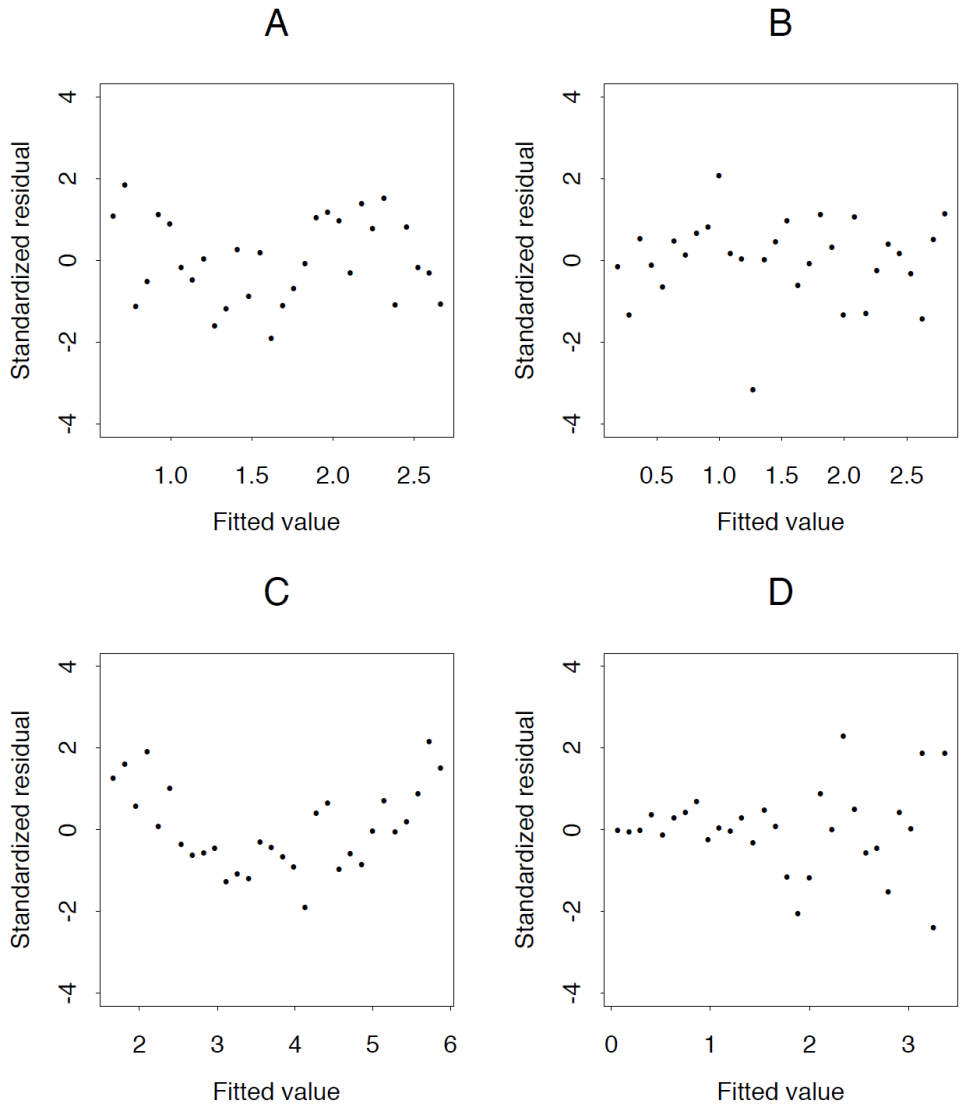

Figure 1: Standardised residuals as a function of values adjusted for four Gaussian models.

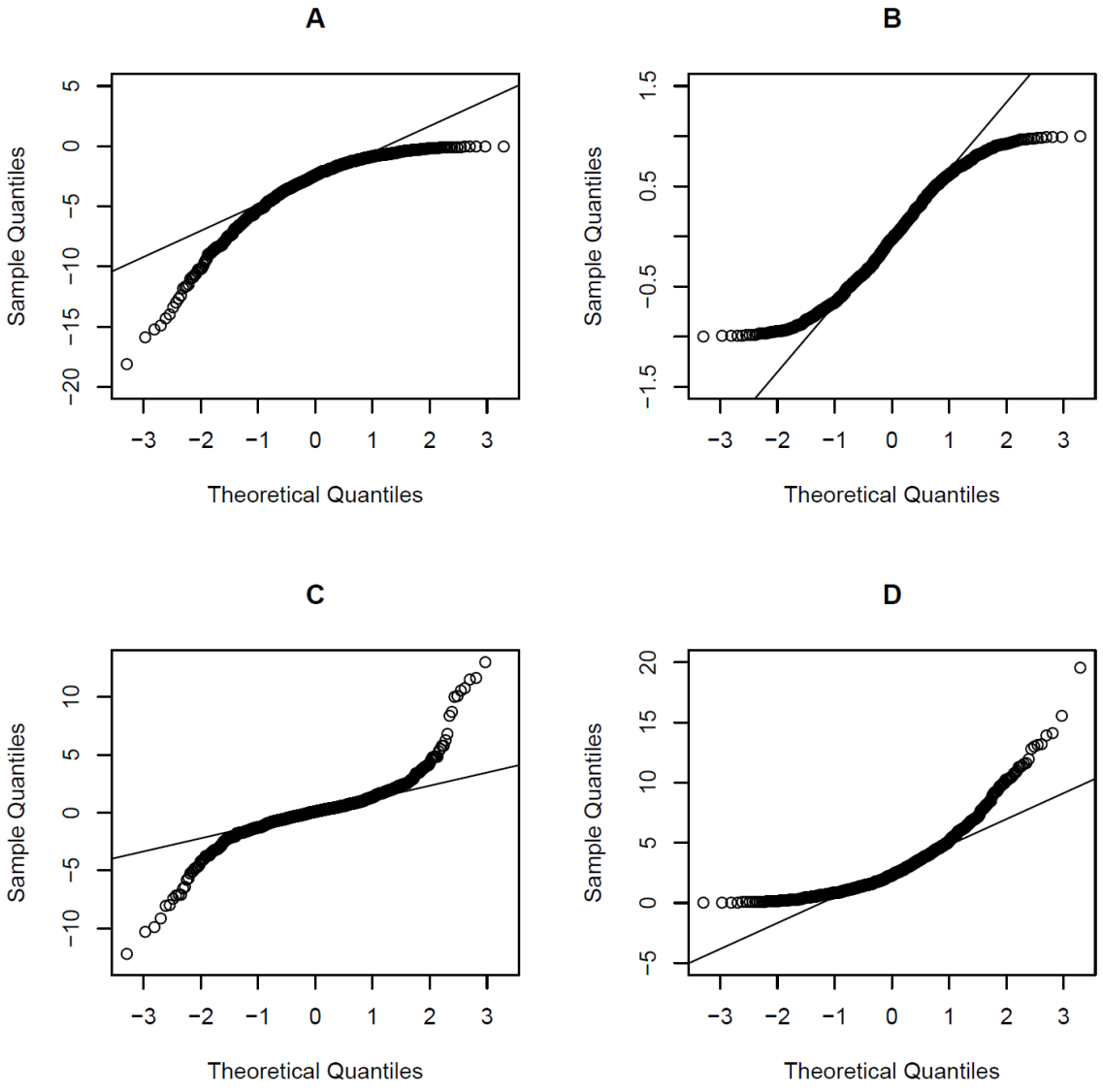

Figure 2: Four Q-Q Gaussian plots where the data do not follow a Gaussian law.

2. (***Bonus***, i.e. this part can be skipped, we just need the form of the density below.)
For the kth order statistic $V_{(k)}$ of a sample of n uniform variables on $[0,1]$, as given in subproblem 3 below, prove that $V_{(k)} \sim \text{Beta}(k, n+1-k)$ with probability density function:

$$f_k(x) = n\binom{n-1}{k-1} x^{k-1}(1-x)^{n-k}, \quad x \in [0,1].$$

Hint: Even though there are not many calculations, it is not an easy exercise. Let $A = \{0 < v_1 < \cdots < v_n < 1\} \subset [0,1]^n$. For $(v_1, \ldots, v_n) \in A$, use the symmetry of the problem to write

$$\mathbb{P}\left(V_{(1)} \leq v_1, \ldots, V_{(n)} \leq v_n\right)$$

as a n variables multiple integral. It is not advisable to compute explicitly this integral, but we can find a (very!) easy explicit formula for the joint distribution

$$\frac{\partial^n}{\partial v_1 \ldots \partial v_n} \mathbb{P}\left(V_{(1)} \leq v_1, \ldots, V_{(n)} \leq v_n\right).$$

Then, the marginal density of $V_{(k)}$ is found by integrating the joint density over all other variables.

3. Let $V_1, \ldots, V_n \sim U([0,1])$ be i.i.d., and let

$$V_{(1)} \leq V_{(2)} \leq \cdots \leq V_{(n)}$$

be the associated order statistics. Compute the expectation of $V_{(k)}$.

4. Let z_α be the quantile α of the normal law $N(0,1)$, defined by

$$\Phi(z_\alpha) = \alpha.$$

Explain why $\mathbb{E}[X_{(k)}] \approx z_{k/(n+1)}$. A rigorous justification is *not* necessary. Link it with the QQ plot.

Hint: It is necessary to approximate $\mathbb{E}[f(X)] \approx f(\mathbb{E}[X])$ for a function f slightly non linear.

Problem 23. We consider the linear model with $n > 8$ and $p = 2$, where

$$\mathbb{E}[y_j] = \beta_0, \quad j = 1, \ldots, n-2,$$
$$\mathbb{E}[y_j] = \beta_0 + \beta_1, \quad j = n-1, n.$$

a) Writing the model in the form $y = X\beta + \varepsilon$, find the least squares estimator $\hat{\beta}$ of β as a function of $\tilde{y}_1 = (n-2)^{-1}\sum_{j=1}^{n-2} y_j$ and $\tilde{y}_2 = (y_{n-1} + y_n)/2$.

b) Calculate the hat matrix for this model, verify that its trace is equal to p and find the fitted values $\hat{y}$.

c) Suppose $y_{n-1} = y_n = \tilde{y}_2$. Find the leverages h_{jj}, the standardised residuals, and Cook's statistics. Comment on this.

Problem 24. (t-test)

Let $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $X \in \mathbb{R}^{n \times p}$ of full column rank. Let us denote the t-statistic for the j-th parameter as

$$t = \frac{\widehat{\beta}_j - \beta_j}{\widehat{\text{se}}(\widehat{\beta}_j)},$$

where $\text{se}(\widehat{\beta}_j) = (\text{var}(\widehat{\beta}_j))^{1/2}$ is the standard deviation of the estimator $\widehat{\beta}_j$ and $\widehat{\text{se}}(\widehat{\beta}_j)$ is a suitable estimator of thereof. Show that $t \sim t_{n-p}$.

Problem 25. When we adjust the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to the cement data set (n=13, slide 55), `R` gives us the following table:

```
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  48.19363      3.91330    12.315   6.17e-07 ***
x1            1.69589      0.20458     8.290   1.66e-05 ***
x2            0.65691      0.04423    14.851   1.23e-07 ***
x3            0.25002      0.18471     1.354   0.209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a) Explain in details how we compute the values in the columns "`t value`" and "`Pr(>|t|)`". Which is the significance of these values? Comment the observed values.

b) Knowing that $\widehat{\mathrm{corr}}(\hat{\beta}_2, \hat{\beta}_3) = -0.08911$, which is the p value for the null hypothesis $\beta_2 - \beta_3 = 0$? Try to find the value of the test statistics without using `R`. For a test with a threshold of 5%, can we reject the null hypothesis?

Problem 26. [REDUNDANT] Suppose the $n \times p$ full-rank design matrix $\mathbf{X}$ can be partitioned into two blocks as $[\,\mathbf{X}_1\ \mathbf{X}_2\,]$ and let $\mathbf{M}_{\mathbf{X}_1} \coloneqq \mathbf{I}_n - \mathbf{H}_{\mathbf{X}_1}$. Show that $\mathbf{H}_{\mathbf{X}} = \mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$, where $\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$ is the projection on to the span of $\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2$.

Problem 27. (Frisch–Waugh–Lovell theorem) Consider the linear regression $y = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon$ with $\mathbb{E}\varepsilon = 0_n$. Let y be the observed response and suppose the $n \times p$ full-rank design matrix $\mathbf{X}$ can be written as the partitioned matrix $[\mathbf{X}_1\ \mathbf{X}_2]$ with blocks $\mathbf{X}_1$, an $n \times p_1$ matrix, and $\mathbf{X}_2$, an $n \times p_2$ matrix. Let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ be the ordinary least square (OLS) parameter estimates from running this regression. Suppose we run least squares on this model to obtain

$$y = \mathbf{X}_1\widehat{\beta}_1 + \mathbf{X}_2\widehat{\beta}_2 + e, \tag{E1}$$

Define the orthogonal projection matrix $\mathbf{H}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ as usual and $\mathbf{H}_{\mathbf{X}_i} = \mathbf{X}_i(\mathbf{X}_i^\top\mathbf{X}_i)^{-1}\mathbf{X}_i^\top$ for $i = 1, 2$. Similarly, define the complementary projection matrices $\mathbf{M}_{\mathbf{X}_1} = \mathbf{I}_n - \mathbf{H}_{\mathbf{X}_1}$ and $\mathbf{M}_{\mathbf{X}_2} = \mathbf{I}_n - \mathbf{H}_{\mathbf{X}_2}$.

Prove the Frisch–Waugh–Lovell (FWL) theorem, i.e., show that the ordinary least square estimates $\widehat{\beta}_2$ and the residuals e from (E1) are identical to those obtained by running ordinary least squares on the regression

$$\mathbf{M}_{\mathbf{X}_1}y = \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2\beta_2 + \text{residuals}. \tag{E2}$$

Hint: starting from (E1) *assuming* $\widehat{\beta}_2$ *has been computed, pre-multiply both sides so as to obtain an expression in terms of* $\widehat{\beta}_2$ *only on the right-hand side and show the latter coincides with the least square estimate from* (E2).

Problem 28. (t-test vs. F-test for model-submodel testing, requires the previous problem)

Consider the linear regression $y = \mathbf{X}_1\beta_1 + \mathbf{x}_2\beta_2 + \boldsymbol{\varepsilon}$ under the assumption that $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{x}_2^\top)^\top$ is an $n \times p$ full-rank non-stochastic design matrix with $\mathbf{x}_2$ an $n \times 1$ column vector and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0_n, \sigma^2\mathbf{I}_n)$. We are interested in testing whether the parameter $\beta_2 = 0$: the Wald test t-statistic W and the Fisher test statistic F for this hypothesis are, respectively,

$$W = \frac{\hat{\beta}_2}{\mathrm{se}(\hat{\beta}_2)}, \qquad F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}/(n-p)},$$

where $\mathrm{se}(\hat{\beta}_2) = \left[s^2\mathsf{Var}\left(\hat{\beta}_2\right)/\sigma^2\right]^{1/2}$. Under the null hypothesis $\mathcal{H}_0 : \beta_2 = 0$, $W \sim \mathcal{T}(n-p)$ and $F \sim \mathcal{F}(1, n-p)$. Show algebraically that $W^2 = F$.

Note that the two statistics lead to the same inference because the square of a $\mathcal{T}(n-p)$ distributed random variable has distribution $\mathcal{F}(1, n-p)$.

Problem 29. We consider the cement data with $n = 13$. The residuals sum of squares (RSS) for all the possible models (containing always the denoted variables and the intercept) are given below:

Model	RSS	Model	RSS	Model	RSS
- - - -	2715.8	1 2 - -	57.9	1 2 3 -	48.1
1 - - -	1265.7	1 - 3 -	1227.1	1 2 - 4	48.0
- 2 - -	906.3	1 - - 4	74.8	1 - 3 4	50.8
- - 3 -	1939.4	- 2 3 -	415.4	- 2 3 4	73.8
- - - 4	883.9	- 2 - 4	868.9		
		- - 3 4	175.7	1 2 3 4	47.9

Calculate the analysis of variance table (as in slide 163) adding x_4, x_3, x_2 and x_1 to the model in this order, and test which term should be included in the model for the threshold $\alpha = 0.05$. Compare with slide 164.

Problem 30. (Orthogonal variables) Let us consider the regression

$$y = X\beta + \varepsilon = (X_1,\, X_2)\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon,$$

where $X = (X_1, X_2)$, $\beta^\top = (\beta_1^\top, \beta_2^\top)$, X_1 is $n \times p_1$, X_2 is $n \times p_2$ (both injective) such that

$$X_1^\top X_2 = 0_{p_1 \times p_2}.$$

Let H_i be the hat matrix associated to X_i.

1. What is the geometrical interpretation of $X_1^\top X_2 = 0$?
2. Calculate H as a function of X_i and of H_i, then, calculate the products

   $$H_1H_2, H_2H_1, HH_1, H_1H.$$

   What do you notice, which is the geometrical interpretation?
3. Show that each of the following quantities are equal to Hy:
   (a) $H_1y + H_2y$;
   (b) $H_1y + H_2e_1$, with $e_1 = (I - H_1)y$;
   (c) $H_1y + He_1$.
4. Interpret these equalities in relation to the models

   $$y = X\beta + \varepsilon \qquad (M)$$

   and to its submodels

   $$y = X_1\beta_1 + \varepsilon, \qquad (M_1)$$
   $$y = X_2\beta_2 + \varepsilon. \qquad (M_2)$$

Problem 31. (Orthogonal variables and ANOVA)

Let us consider the regression

$$y = X\beta + \varepsilon = (X_1, \dots, X_k) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \varepsilon$$

where X_i is $n \times p_i$, all the X_i are injective, and

$$i \neq j \implies X_i^\top X_j = 0.$$

Let H be the hat matrix associated to X, H_i the hat matrix associated to X_i and $\hat{\beta} = (X^\top X)^{-1}X^\top y = (\hat{\beta}_1^\top, \dots, \hat{\beta}_k^\top)^\top$. We denote by δ_{ij} Kronecker's delta: $\delta_{ij} = 1$ if $i = j$, 0 otherwise. For an ordered set $L \subset \{1, \dots, k\}$ we define $X_L = (X_i : i \in L)$ and $\hat{\beta}_L = (\hat{\beta}_i^\top : i \in L)^\top$. For example, if $L = \{1, 2, 4\}$, $X_L = (X_1, X_2, X_4)$ and

$$\hat{\beta}_L = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_4 \end{pmatrix}.$$

We define $RSS_L = \|y - H_Ly\|^2$, where $H_L = X_L(X_L^\top X_L)^{-1}X_L^\top$.

1. Show that $H = H_1 + \cdots + H_k$ and that $H_L = \sum_{i \in L} H_i$.
2. Show that $H_iH_j = \delta_{ij}H_i$.
3. Show that $\hat{\beta}_j = (X_j^\top X_j)^{-1}X_j^\top y$.
4. For $j \notin L$, calculate

   $$RSS_L - RSS_{L \cup \{j\}},$$

   and show that this expression does not depend on L.
5. Which is the interpretation of point 4. with respect to ANOVA?

Problem 32. (Automatic model selection)

We consider the cement data. The residuals' sum of squares (RSS) and the Mallows' C_p for the model *containing the ordinate at the origin* are the following:

Model	RSS	C_p	Model	RSS	C_p	Model	RSS	C_p
- - - -	2715.8	442.58	1 2 - -	57.9		1 2 3 -	48.1	
1 - - -	1265.7	202.39	1 - 3 -	1227.1	197.94	1 2 - 4	48.0	
- 2 - -	906.3		1 - - 4	74.8	5.49	1 - 3 4	50.8	
- - 3 -	1939.4	314.90	- 2 3 -	415.4	62.38	- 2 3 4	73.8	7.325
- - - 4	883.9	138.62	- 2 - 4	868.9	138.12			
			- - 3 4	175.7	22.34	1 2 3 4	47.9	5

1. Utilise the selection methods *forward selection* and *backward elimination* to chose some models for these data, including the significant variables at level 5%. Utilise the F-test

$$F = \frac{RSS(\hat{\beta}_L) - RSS(\hat{\beta}_{L\cup\{j\}})}{RSS(\hat{\beta}_{\text{full}})/(13-5)}$$

   to decide if the addition of the j-th variable is significant.

2. Another selection criterion is the Mallow's C_p:

$$C_p = \frac{SS_p}{s^2} + 2p - n.$$

   Notice that here s^2 is the variance estimator in the complete model.

   (a) How could we use this criterion? Calculate the missing C_p.

   (b) Which is the model selected by this criterion using the *forward selection*, and then *backward elimination*? Among all the models considered, which one is the best, according to this criterion?

Problem 33. (AIC and Gaussian linear models)

Show that the AIC criterion for a Gaussian linear model, base on a response vector of size n, with p covariates and σ^2 unknown, can be written as :

$$\text{AIC} = n\,\log\hat{\sigma}^2 + 2p + \text{const},$$

where $\hat{\sigma}^2 = SS_p/n$ is the maximum likelihood estimator of σ^2

Problem 34. (Cross validation and number of regressions)

Let $y = X\beta + \epsilon$, and $\widehat{\beta}$ denote the OLS estimator of β. The (leave-one-out) cross validation uses one observation (x_k, y_k) as the validation set and the remaining observations (X_{-k}, y_{-k}) as the training set and repeating the procedure for each $k = 1, \ldots, n$. With the k-th observations $x_k \in \mathbb{R}^p$ and $y_k \in \mathbb{R}$ deleted, let $X_{-k} \in \mathbb{R}^{(n-1)\times p}$, $y_{-k} \in \mathbb{R}^{n-1}$, and $\widehat{\beta}_{-k} \in \mathbb{R}^p$ denote the corresponding design matrix, the responses, and the OLS estimator, respectively (symbolically, $y_{-k} = X_{-k}\beta_{-k} + \epsilon_{-k}$).

a) Use the Sherman-Morrison formula

$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}$$

to show that

$$(X_{-k}^\top X_{-k})^{-1} = \left(I + \frac{(X^\top X)^{-1}x_k x_k^\top}{1 - h_{kk}}\right)\left(X^\top X\right)^{-1}.$$

b) Noting that $x_k^\top$ is the k-th row of the original design matrix X, show that

$$X_{-k}^\top y_{-k} = X^\top y - y_k x_k \quad \text{and} \quad x_k^\top (X^\top X)^{-1} X_{-k}^\top y_{-k} = (1 - h_{kk})y_k - e_k\,,$$

to conclude that

$$\hat{\beta}_{-k} = \hat{\beta} - \frac{e_k \left(X^\top X\right)^{-1} x_k}{1 - h_{kk}}.$$

c) Use the previous formula to deduce that the cross-validation criterion

$$\text{CV} = \sum_{k=1}^{n} (y_k - x_k^\top \hat{\beta}_{-k})^2. \tag{2}$$

can be written as

$$\mathrm{CV} = \sum_{k=1}^{n} \frac{(y_k - x_k^\top \hat{\beta})^2}{(1-h_{kk})^2}. \tag{3}$$

What is the advantage of using (3) instead of (2)?

Problem 35. Let us suppose that $y = \mu + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and that we adjusted to y a linear model with the full rank design matrix $X_{n\times p}$, $n \geq p$, and the corresponding hat matrix H. Let D be the diagonal matrix with elements $1-h_{11}, \ldots, 1-h_{nn}$. Using the previous exercise, show that

$$\mathbb{E}[CV] = \mu^\top (I-H) D^{-2} (I-H)\mu + \sigma^2 tr(D^{-1}),$$

and deduce that if μ belongs to the space generated by the columns of X, then $\mathbb{E}[CV] \approx (n+p)\sigma^2$.

Problem 36. (Model selection in `R`)

a) Use the criteria *backward stepwise* and *forward stepwise* to choose a model for the data"Supervisor Performance" (`SPD`) from `R` package `RSADBE`

   Which model has the best AIC value?

b) Using the package `leaps`, find the model with the best BIC value among all submodels.

Problem 37. (Ridge regression)

Let $\mathbf{X} = [\mathbf{1}_n\ \mathbf{Z}]$ be an $n \times p$ design matrix with centered inputs $\mathbf{Z}$, meaning that $\mathbf{Z}^\top \mathbf{1}_n = \mathbf{0}_{p-1}$. Consider the model $y = \mathbf{1}_n\beta_0 + \mathbf{Z}\gamma + \boldsymbol{\varepsilon}$, where $\mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0}_n$ and $\mathsf{Var}\,(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$. The ridge estimators are defined by

$$(\hat{\beta}_0, \hat{\gamma}_\lambda) = \underset{(\beta_0,\gamma)}{\arg\min} \|y - \mathbf{1}_n\beta_0 - \mathbf{Z}\gamma\|_2^2 + \lambda\|\gamma\|_2^2.$$

From slide 211, we know that the ridge estimators are given by

$$(\hat{\beta}_0, \hat{\gamma}_\lambda) = (\overline{y}, (\mathbf{Z}^\top\mathbf{Z} + \lambda \mathbf{I}_{p-1})^{-1}\mathbf{Z}^\top y)$$

a) Show that the fitted value of the ridge regression are

$$\hat{y}_\lambda = \overline{y}\mathbf{1}_n + \sum_{j=1}^{p-1} \frac{\omega_j^2}{\omega_j^2 + \lambda} \left(\mathbf{u}_j^\top y\right) \mathbf{u}_j,$$

   where $\mathbf{u}_j$ and ω_j are the left singular column vectors and the singular values of $\mathbf{Z}$, respectively. Discuss what happens to $\hat{y}_\lambda$ when some of the $\{\omega_j^2\}_{j=1}^{p-1}$ are close to zero.

b) What happens to the ridge estimates if the columns of $\mathbf{Z}$ are orthogonal, i.e. $\mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_{p-1}$? Explain why it is preferable to standardize the columns of $\mathbf{Z}$ so they have approximately unit variance.

c) Show that $\lambda \mapsto \left\|\hat{\gamma}_\lambda\right\|_2^2$ is a decreasing function.

Problem 38. Let $\lambda^* = 2\max_{1\leq j\leq q} |Z_j^\top y|$. Show that

$$\begin{cases} \lambda > \lambda^* \implies \hat{\gamma}_{\text{lasso}} = 0, \\ \lambda < \lambda^* \implies \hat{\gamma}_{\text{lasso}} \neq 0. \end{cases}$$

Hint: Use the convexity for the first part.

Problem 39. Let $\mathbf{X} = [\mathbf{1}_n\ \mathbf{Z}]$ be an $n \times p$ design matrix with centered inputs $\mathbf{Z}$, meaning that $\mathbf{Z}^\top \mathbf{1}_n = \mathbf{0}_{p-1}$. Consider the model $y = \mathbf{1}_n\beta_0 + \mathbf{Z}\gamma + \boldsymbol{\varepsilon}$, where $\mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0}_n$ and $\mathsf{Var}\,(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$. The ridge estimators are defined by

$$(\hat{\beta}_0, \hat{\gamma}_\lambda) = \underset{(\beta_0,\gamma)}{\arg\min} \|y - \mathbf{1}_n\beta_0 - \mathbf{Z}\gamma\|_2^2 + \lambda\|\gamma\|_1.$$

We know that $\hat{\beta}_0 = \overline{y}$ regardless of the smoothing parameter $\lambda \geq 0$, thus

$$\hat{\gamma}_\lambda = \underset{\gamma}{\arg\min} \|y - \mathbf{1}_n\overline{y} - \mathbf{Z}\gamma\|_2^2 + \lambda\|\gamma\|_1.$$

Unlike the ridge regression, lasso solution may not be unique. Nonetheless, the adjusted values are unique: let $\hat{\gamma}_1$ and $\hat{\gamma}_2$ be two lasso solutions (for the same smoothing parameters λ).

a) Show that $Z\hat{\gamma}_1 = Z\hat{\gamma}_2$, using convexity.

b) Show that, if $\lambda > 0$, then $\|\hat{\gamma}_1\|_1 = \|\hat{\gamma}_2\|_1$.

Problem 40. (Median regression)
Let $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$. Note that the median of a random variable Y is defined as

$$\text{med}(Y) = \arg\min_{c \in \mathbb{R}} \mathsf{E}|Y - c| \,.$$

Let $X_i = (1, x_i)^\top$ and

$$\widehat{\beta} = \arg\min_{\beta} \sum (Y_i - \beta^\top X_i)^2 \,, \qquad \widetilde{\beta} = \arg\min_{\beta} \sum |Y_i - \beta^\top X_i|$$

1. Show that $\mathsf{E}|Y - \beta^\top X|$ is minimized for $\beta^\top X = \text{med}(Y)$ and conclude why $\widetilde{\beta}$ is sometimes called the "median regression estimate".
2. Compare what are the estimators $\widehat{\beta}$ and $\widetilde{\beta}$ actually estimating in the cases of $\epsilon \sim N(0,1)$ and $\epsilon_i \sim Exp(1)$.

Problem 41. (Naive kernel density estimator)
Let $X_1, \ldots, X_n$ be a random sample from a distribution function F. Let $f = F'$ be the density. For every $x \in \mathbb{R}$, the estimator of f is given as

$$\widehat{f}(x) := \frac{F_n(x+h) - F_n(x-h)}{2h} \,,$$

where F_n is the empirical distribution function. Show that $\widehat{f}$ is a kernel density estimator (check out "kernel density estimation" on Wikipedia for definition), i.e. specify the weighting function, also known as the kernel.

Problem 42. (Generalized least squares)
Consider the linear model $Y = \mathbf{X}\beta + \boldsymbol{\varepsilon}$, where y is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ full-rank non-stochastic design matrix and the error vector $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0_n, \Sigma)$ for $\Sigma \neq \sigma^2 \mathbf{I}_n$ a *known* positive definite covariance matrix. Let y be the observed response vector.

1. Show that the maximum likelihood estimator (MLE) of β is the vector that minimizes

$$(y - \mathbf{X}\beta)^\top \Sigma^{-1} (y - \mathbf{X}\beta).$$

2. Show that the maximum likelihood estimator of β, known as generalized least squares estimator (GLS), is of the form

$$\widehat{\beta}_{\text{GLS}} = (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1} y.$$

3. Derive the distribution of $\widehat{\beta}_{\text{GLS}}$.
4. Show that the ordinary least squares (OLS) estimator $\widehat{\beta}$ is an unbiased estimator of β, but is not the best linear unbiased estimator (BLUE) of β. State carefully any result you use.

Problem 43. Consider the linear model $y = X\beta + \varepsilon$, with $\varepsilon_j \overset{iid}{\sim} g(\cdot)$; suppose that $\mathbb{E}(\varepsilon_j) = 0$ and $\text{var}(\varepsilon_j) = \sigma^2 < \infty$ is known. Suppose that the MLE of β is regular, with

$$i_g = \int -\frac{\partial^2 \log g(u)}{\partial u^2} g(u) du = \int \left\{ \frac{\partial \log g(u)}{\partial u} \right\}^2 g(u) du.$$

1. Show that the asymptotic relative efficiency (ARE) of the leas squares estimator of β relative to MLE of β is

$$\frac{1}{\sigma^2 i_g}.$$

2. What is it reduced to if g is the gaussian density?
3. What about if g is the density of the Laplace distribution?

Problem 44. Give the equivalent of the H matrix for non-parametric regression with kernel smoothing.

Problem 45. (Cubic spline)
Let $n \geq 2$ and $a < x_1 < x_2 < \cdots < x_n < b$. Denote by $N(x_1, x_2, \ldots, x_n)$ the space of natural cubic splines with knots $x_1, x_2, \ldots, x_n$. The goal of this exercise is to show that the solution to the problem

$$\min_{f \in C^2[a,b]} L(f), \text{ where } L(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b \{f''(x)\}^2 dx, \quad \lambda > 0, \tag{4}$$

must belong to $N(x_1, x_2, \ldots, x_n)$. In order to show this, we need the following theorem

> **Theorem.** For every set of points $(x_1, z_1), (x_2, z_2), \ldots, (x_n, z_n)$, there exists a natural cubic spline g interpolating those points. In other words, $g(x_i) = z_i$, $i = 1, \ldots, n$, for a unique natural cubic spline g. Moreover, the knots of g are $x_1, x_2, \ldots, x_n$.

1. Let g the natural cubic spline interpolating the points (x_i, z_i), $i = 1, \ldots, n$, and let $\tilde{g} \in C^2[a, b]$ another function interpolating the same points. Show that

$$\int_a^b g''(x)h''(x)dx = 0,$$

   where $h = \tilde{g} - g$.
   Hint: integration by parts

2. Using point (1) show that

$$\int_a^b \{\tilde{g}''(x)\}^2 dx \geq \int_a^b \{g''(x)\}^2 dx$$

   when the equality holds if and only if $\tilde{g} = g$.

3. Use point (2) to show that if the problem (4) has a solution $\hat{f}$, then$\hat{f} \in N(x_1, x_2, \ldots, x_n)$.

Problem 46. Prove the proposition on slide 29:

Let $\mathbf{\Omega} \in \mathbb{R}^{p \times p}$ be a real symmetric matrix. Then Ω is non-negative definite if and only if Ω is the covariance matrix of some random vector Y.

Problem 47. Show that the two definitions of a positive (semi-)definite matrix on lecture slide 26 are equivalent:

For a real symmetric $p \times p$ matrix $\mathbf{\Omega}$, show that the statements

a) for all $x \in \mathbb{R}^p \setminus \{0\}$, $x^\top \mathbf{\Omega} x > 0$ (or $x^\top \mathbf{\Omega} x \geq 0$), and

b) all eigenvalues of $\mathbf{\Omega}$ are positive (or non-negative)

are equivalent, defining $\mathbf{\Omega}$ as a positive definite (or semi-definite) matrix.

Problem 48. Let Y be a random variable with covariance $\Sigma = \begin{pmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{pmatrix}$.

1. Calculate the principal components v_1 and v_2.
2. Verify your calculation in `R`.
3. In `R`, simulate $n = 100$ data points from a distribution with mean zero and covariance Σ.
4. In `R`, find the principal components of the sample from the previous point, denoted by $\widehat{v}_1$ and $\widehat{v}_2$.
5. In `R`, plot the simulated data points together with the population and sample principal components.

Problem 48b. Let $\{x_1, \ldots, x_n\} \subset \mathbb{R}^p$, and $\mathbf{X}$ be a matrix with $x_i^\top$ in its i-th row. Let $\mathbf{X} = \mathbf{UDV}^\top$ be the SVD of $\mathbf{X}$. Show that for $q < p$ the optimization problem

$$\min_{\mathbf{Q} \in \mathbb{R}^{p \times q}, \mathbf{Q}^\top \mathbf{Q} = I} \sum_{i=1}^{n} \|x_i - \mathbf{Q}\mathbf{Q}^\top x_i\|_2^2$$

is equivalent to

$$\max_{\mathbf{Q}\in\mathbb{R}^{p\times q},\mathbf{Q}^\top\mathbf{Q}=I} \operatorname{tr}(\mathbf{Q}^\top\mathbf{V}\mathbf{D}^2\mathbf{V}^\top\mathbf{Q})$$

and conclude that $\mathbf{Q} = (v_1, \ldots, v_q)$ is a solution, where v_i is the i-th column of $\mathbf{V}$.

A note on the SVD: The full SVD of $\mathbf{X} \in \mathbb{R}^{n\times p}$ refers to the decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix with its columns forming a basis of $\mathbb{R}^n$, $\mathbf{V} \in \mathbb{R}^{p\times p}$ is an orthogonal matrix with its columns forming a basis of $\mathbb{R}^p$ and $\mathbf{D} \in \mathbb{R}^{n\times p}$ has non-zero entries only on the "diagonal". However, some authors (including us in this exercise) understand by SVD the *compact SVD*, which refers to the same decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, while (let $m = \min(n, p)$) $\mathbf{U} \in \mathbb{R}^{n\times m}$ and $\mathbf{V} \in \mathbb{R}^{m\times p}$ has orthogonal columns (but may not be full bases anymore), and $\mathbf{D} \in \mathbb{R}^{m\times m}$ is a diagonal matrix. Intuitively, going from the full SVD to the compact one, one just trims off an all-zero block of $\mathbf{D}$ to make it a square matrix and discards the corresponding parts of $\mathbf{U}$ or $\mathbf{V}$. The compact SVD is often the default in software packages, since one is seldom interested in the full SVD. It is often clear from the context, whether the full SVD or the compact SVD is considered. In the exercise above, the meaning of $\mathbf{D}^2$ would be unclear unless the compact SVD was considered. Recall that neither the full SVD nor the compact SVD are unique.

Problem 49. In `R`, generate a random vector (a regressor) $\mathtt{x} \in \mathbb{R}^{100}$ such that $x_j \subset [0, 2]$, and a random vector of errors $\mathtt{e} \in \mathbb{R}^{100}$ such that $e_j \sim N(0, 1/10)$. Then create the dependent random variable as

$$y_j = 10 + 2\sin(\pi * x_j) + e_j.$$

Plot the dependent random variable against the regressor. Secondly, find a transformation of the x-axis which reveals the approximate linear relationship between `x` and `y`. Can you see how the constants (10 and 2) affect the plots? Go through the same for the following dependent variable:

$$y_j = \exp(15 + 3\log(x) + e_j).$$

Problem 50. Let $y_i = \beta_1 \cos(x - \beta_2) + \epsilon$ for $i = 1, \ldots, 100$.

a) Can you obtain estimates for $\beta = (\beta_1, \beta_2)^\top$ directly by solving a sequence of least squares problems? How do the design matrices and responses for this sequence look like?

b) Can you obtain estimates for a suitable transformation of β by solving only a single least squares problem?

c) Simulate data in R using the following code:

```
x <- 1.5*pi*runif(100)
y <- 1*cos(x - (-1)) + rnorm(100)/2
data1 <- data.frame(x=x,y=y)
```

i.e. $\beta = (1, -1)^\top$ here. Treat β as unknown and estimate it using both (a) and (b). Find the fitted values using approach (a) and approach (b). Plot the raw data and both sets of fitted values to check if they are the same.