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

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

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

Problem 3. Suppose the $n \times p$ full-rank design matrix X can be written as $[X_1\ X_2]$ with blocks X_1, an $n \times p_1$ matrix, and X_2, an $n \times p_2$ matrix. Show that $H - H_1$ 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. $\frac{\partial}{\partial X} \operatorname{tr}(X) = I_n$.

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

1. Show that $B = X^\top \Omega X$ is positive definite and thus invertible. Deduce from this fact that $X^\top X$ is invertible.
2. Show that 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 $\mathcal{N}(\mu, \sigma^2)$.

1. 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

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

2. Show that for μ fixed, the ML estimate of σ^2 is given by

$$\widehat{\sigma}^2_\mu = \frac{n-1}{n} s^2 \left\{ 1 + \frac{t(\mu)^2}{n-1} \right\},$$

where

$$s^2 = \frac{1}{n-1}\sum_{j=1}^n (y_j - \bar{y})^2 \qquad \text{and} \qquad t(\mu) = \sqrt{n}\,\frac{\bar{y} - \mu}{s}.$$

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

3. For a fixed μ, we take the ML estimate of σ^2 (depending on μ) and plug it in the likelihood to obtain the so-called profile likelihood for μ, which only depends on μ, not on σ. Show the profile likelihood is given by:
$$\ell_p(\mu) = -\frac{n}{2}\log[s^2\{1+t(\mu)^2/(n-1)\}] + \text{const.}$$

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

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

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

(a) $\Sigma_{12}\Sigma_{22}^{-1} = -Q_{11}^{-1}Q_{12}$

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

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

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 $1 \le k < n$. 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 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 $Z \sim \mathcal{N}_n(0_n, I_n)$ and $Y \sim \mathcal{N}_n(\mu, \Sigma)$ with Σ positive definite.

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

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

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

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

Problem 10. Consider a singular value decomposition (SVD) of the design matrix $X = UDV^\top$, where U is an $n \times p$ orthonormal matrix (meaning $U^\top U = I_p$ and the columns of U are orthogonal vectors), D is an $p \times p$ diagonal matrix and V is an $p \times p$ orthogonal matrix.

1. Show that H does not depend on V.
2. Give a formula for the ordinary least square estimate $\widehat{\beta}$, showing that the only inverse it involves is the inverse of a diagonal matrix.

Problem 11. (Non-linear ↔ 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 X. Calculate the elements of $X^\top X$, $X^\top Y$ and $(X^\top X)^{-1}$.

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. (Models in R)

In R, a model formula has the following general form

```
reponse~expression
```

where the left-hand side of the formula, `reponse`, can sometimes be absent, and the right-hand side, `expression`, is a collection of terms linked by operators, usually as an arithmetical expression. Let us suppose, for example, that

$$y = \begin{pmatrix} 217 \\ 143 \\ 186 \\ 121 \\ 157 \\ 143 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & 3 \\ 0 & 2 & 1 \\ 1 & 2 & 2 \\ 0 & 2 & 3 \end{pmatrix},$$

and let `x`, `a`, `b` be the columns of $X = [x, a, b]$.

a) A *factor* is a variable that represents a categoric/qualitative variable (command `as.factor()` in R). For example, if `a` is a factor, then `y~a` represents the model

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

Formally, it is written with indicators:

$$y_j = \beta_0 + \alpha_1 I_{(a_j = \text{“}1\text{”})} + \alpha_2 I_{(a_j = \text{“}2\text{”})} + \varepsilon_j, \tag{1}$$

where $I_E = 1$ if the expression E is true, and 0 otherwise. NoOtice that the “1” and “2” of “vector” `a` do *not* represent the number 1 and 2, but some categories, groups, classes or levels. For example, we can have “1” = “normally fed”, and “2” = “fed with growth inibitor”.

Let us suppose that `a` and `b` are factors:

I. Give the design matrix corresponding to model (1), as well as the vector of variables.

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

III. Suppress the column corresponding to α_1 of this matrix in order to have an injective 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~a`, (ii) `y~a+b`, (iii) `y~x+a-1` (iv) `y~b+x-1`.

b) Let us suppose that `a` and `b` are (still) factors: `a`*interaction* is represented in the form `a:x` or `a:b`. For example, `y~a:x` represents

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

that also writes

$$y_j = \beta_0 + \alpha_1 I_{(a_j=\text{“}1''')} x_j + \alpha_2 I_{(a_j=\text{“}2'')} x_j + \varepsilon_j$$

with the indicators. I.e., a model with different slopes for groups “1” and “2”, but with the same intercept.

Expression `y~a:b` represents the model

$$y_j = \beta_0 + \alpha_j + \varepsilon_j, \quad j = 1,\dots,6;$$

that also writes

$$y_j = \beta_0 + \sum_{i=1}^{2}\sum_{l=1}^{3} \gamma_{i,l} I_{(a_j=\text{“}i'')} I_{(b_j=\text{“}l'')} + \varepsilon_j.$$

I.e., a model with different *intercepts* for different combinations of levels for `a` and `b`. Find the design matrices corresponding to models

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

Among these matrices, which ones have linearly independent columns?

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 makes sense?

c) Add the variable *grade* to the model as a factor. How this changes 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 changed compared to part c)?

Problem 15. Assume a linear model was developed for 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 the (actual value of) weight of the patient.

a) Write down the regression function of the model.

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 give 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 X can be partitioned into two blocks as $[\,X_1\ X_2\,]$ and let $M_{X_1} := I_n - H_{X_1}$. Show that $H_X = H_{X_1} + H_{M_{X_1}X_2}$, where $H_{M_{X_1}X_2}$ is the projection on to the span of $M_{X_1}X_2$. (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 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 $t_9(0.975) = 2.262$ 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 $\widetilde{Y} = AY$ follows a linear Gaussian model as well,

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

with $\widetilde{X} = AX$. We will consider some particular case the 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 $\widehat{\tilde{y}}$ 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 $\widetilde{Y}$. 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* $\widetilde{H}$ 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 $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)^\top$.

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 $\widehat{\beta}_1$?

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

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

Show that

$$\text{MSE}(c^\top \widetilde{\beta}) \geq \text{MSE}(c^\top \widehat{\beta}), \quad \forall c \in \mathbb{R}^p,$$

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

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

Problem 21. (Diagnostic's graphics)

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

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 heavier 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 QQ plot to "see" whether a sample $x_1, \ldots, x_n$ comes from the normal distribution. Let $X_1, \ldots, X_n \sim N(0,1)$ be i.i.d, and let Φ be the cumulative distribution function of the normal law $N(0,1)$.

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

2. 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 $V_{(k)}$.

   Hint: Use the fact f_k is a density function.

3. 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.

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

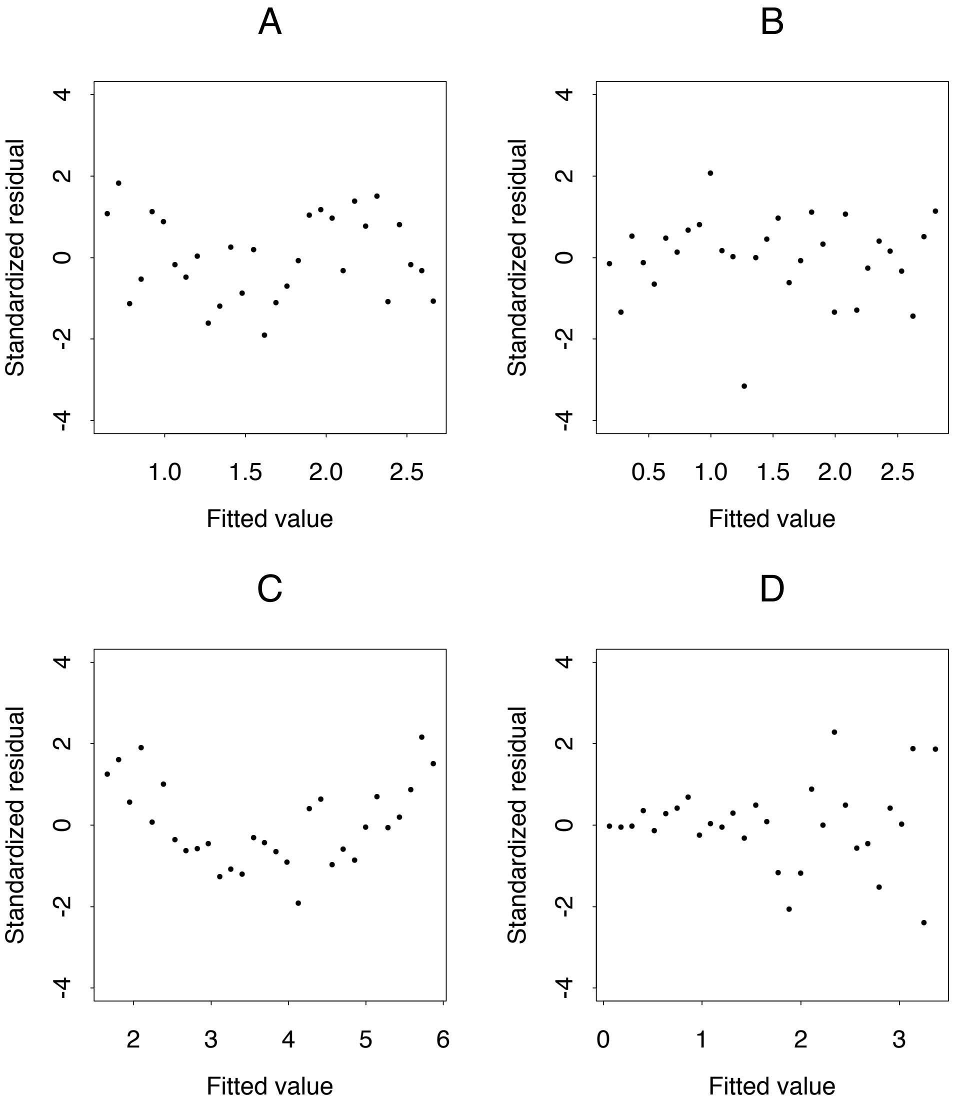

Figure 1: Standardised residuals as function of values adjusted for four Gaussian models.

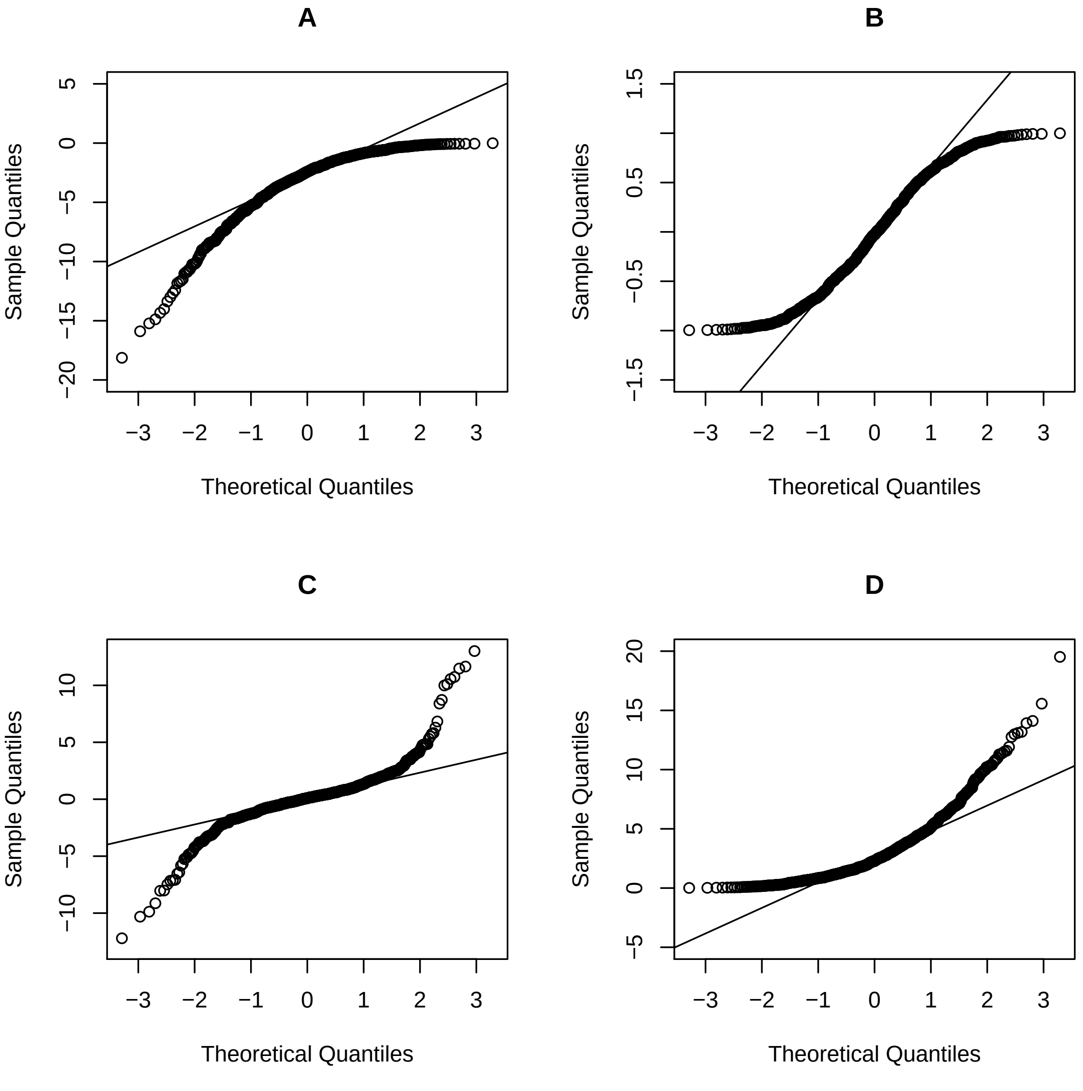

Figure 2: Four Q-Q Gaussian plots where the data do not follow a Gaussian law.

4. ***Bonus:*** Prove that $V_{(k)} \sim \text{Beta}(k, n+1-k)$ with probability density function:

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

Attention: Even if there are not many calculations, it is not an easy exercise.

Tip: 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

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

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

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

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

Problem 23. We consider the linear model with $n \ge p = 2$, where

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

a) Writing the model in the form $y = X\beta + \varepsilon$, find the least squares estimator of β as function of y_n and of $\bar{y}_0 = (n-1)^{-1}\sum_{j1}^{n-1} y_j$. Comment the form of this estimator.

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

c) 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 (slide 90), 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|)`”. What is the meaning of these values? Comment the observed values.

typo: corr(beta_2, beta_3) = -0.08911

b) Knowing that $\widehat{\text{corr}}(\widehat{\beta_1}, \widehat{\beta_2}) = -0.08911$, what 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. Suppose the $n \times p$ full-rank design matrix X can be partitioned into two blocks as $[\,X_1\ X_2\,]$ and let $M_{X_1} := I_n - H_{X_1}$. Show that $H_X = H_{X_1} + H_{M_{X_1}X_2}$, where $H_{M_{X_1}X_2}$ is the projection on to the span of $M_{X_1}X_2$.

Problem 27. (Frisch–Waugh–Lovell theorem) Consider the linear regression $\boldsymbol{Y} = X_1\beta_1 + X_2\beta_2 + \varepsilon$ with $\mathsf{E}(\varepsilon) = 0_n$. Let y be the observed response and suppose the $n \times p$ full-rank design matrix X can be written as the partitioned matrix $[X_1\ X_2]$ with blocks X_1, an $n \times p_1$ matrix, and 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

$$\boldsymbol{y} = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + e, \tag{E1}$$

Define the orthogonal projection matrix H_X as usual and $H_{X_i} = X_i(X_i^\top X_i)^{-1}X_i^\top$ for $i = 1, 2$. Similarly, define the complementary projection matrices $M_{X_1} = I_n - H_{X_1}$ and $M_{X_2} = I_n - H_{X_2}$.

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

$$M_{X_1}\boldsymbol{y} = M_{X_1}X_2\beta_2 + \text{residuals}. \tag{E2}$$

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 $y = X_1\beta_1 + x_2\beta_2 + \boldsymbol{\varepsilon}$ under the assumption that $X = (X_1^\top, x_2^\top)^\top$ is an $n \times p$ full-rank non-stochastic design matrix with x_2 an $n \times 1$ column vector and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$. 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,

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

where $\mathrm{se}(\widehat{\beta}_2) = \left[s^2 \mathsf{Var}\left(\widehat{\beta}_2\right)/\sigma^2\right]^{1/2}$. Under the null hypothesis $\mathscr{H}_0 : \beta_2 = 0$, $W \sim \mathcal{T}(n-p)$ 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 146) when x_4, x_3, x_2 and x_1 are added 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 146.

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

$$H_1 H_2, H_2 H_1, H H_1, H_1 H.$$

   What do you notice, which is the geometrical interpretation?

3. Show that each of the following quantities are equal to Hy:
   (a) $H_1 y + H_2 y$;
   (b) $H_1 y + H_2 e_1$, with $e_1 = (I - H_1) y$;
   (c) $H_1 y + H e_1$.
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

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

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 $\widehat{\beta} = (X^\top X)^{-1} X^\top y = (\widehat{\beta}_1^\top, \ldots, \widehat{\beta}_k^\top)^\top$. We denote by δ_{ij} Kronecker's delta: $\delta_{ij} = 1$ if $i = j$, 0 otherwise. For a ordered set $L \subset \{1, \ldots, k\}$ we define $X_L = (X_i : i \in L)$ and $\widehat{\beta}_L = (\widehat{\beta}_i^\top : i \in L)^\top$. For example, if $L = \{1, 2, 4\}$, $X_L = (X_1, X_2, X_4)$ and

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

We define $RSS_L = \left\| y - H_L y \right\|^2$, 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 $H_i H_j = \delta_{ij} H_i$.

3. Show that $\widehat{\beta}_j = (X_j^\top X_j)^{-1} X_j^\top y$.

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 intercept* 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 significative variables at level 5%. Utilise the F-test
$$F = \frac{RSS(\widehat{\beta}_L) - RSS(\widehat{\beta}_{L\cup\{j\}})}{RSS(\widehat{\beta}_{\text{full}})/(13-5)}$$
to decide if the addition of the j-th variable is significative.

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, based on a response vector of size n, with p covariables and σ^2 unknown, can be written as :
$$\text{AIC} = n \log \widehat{\sigma}^2 + 2p + \text{const},$$
where $\widehat{\sigma}^2 = SS_p/n$ is the miximum likelihood estimator of σ^2

Problem 34. (Cross validation and number of regressions)

Let $y = X\beta + \epsilon$, $\widehat{\beta}$ denote the OLS estimator of β, X_{-j} denote the design matrix, which arises from X by dropping the k-th row x_k (which is understood for mathematical purposes as a column vector), and $\widehat{\beta}_{-j}$ be the estimator based on the data set without the k-th observation (symbolically, $y_{-k} = X_{-k}\beta_{-k} + \epsilon_{-k}$, again y_{-k} and ϵ_{-k} denote the vectors that arise from y and ϵ by dropping the k-th entry).

a) Using the Sherman-Morrison formula
$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u},$$

show that

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

b) Show that

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

to conclude that

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

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

$$\mathrm{CV} = \sum_{j=1}^{n} (y_j - x_j^\top \widehat{\beta}_{-j})^2. \tag{2}$$

can be written as

$$\mathrm{CV} = \sum_{j=1}^{n} \frac{(y_j - x_j^\top \widehat{\beta})^2}{(1-h_{jj})^2}. \tag{3}$$

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 $X = [1_n \ Z]$ be an $n \times p$ design matrix with centered inputs Z, meaning that $Z^\top 1_n = 0_{p-1}$. Consider the model $y = 1_n\alpha + Z\gamma + \boldsymbol{\varepsilon}$, where $\mathsf{E}(\boldsymbol{\varepsilon}) = 0_n$ and $\mathsf{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$.

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

$$\widehat{y}_{\text{ridge}} = \overline{y}1_n + \sum_{j=1}^{p-1} \frac{d_j^2}{d_j^2 + \lambda} u_j u_j^\top y,$$

where u_j are the left singular column vectors of Z and d_j the elements of a diagonal matrix to be determined. Discuss what happens to $\widehat{y}_{\text{ridge}}$ when some of the $\{d_j^2\}_{j=1}^{p-1}$ are close to zero.

b) What happens to the ridge estimates if the columns of Z are orthogonal? Explain why it is preferable to standardize the columns of Z so they have approximately unit variance.

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

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

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

Hint: Use the convexity for the first part.

Problem 39. Unlike the ridge regression, lasso solution is not always unique. However, the adjusted values are unique: let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ be two lasso solutions (for the same smoothing parameters λ).

a) Show that $X\widehat{\beta}_1 = X\widehat{\beta}_2$, using convexity.

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

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

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

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

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

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". *(Note that X denotes here a generic vector of regressors just as Y denotes a generic random variable from which Y_i's are sampled.)*
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. (CV and ridge regression)

1. Explain how to use cross-validation to choose the tuning parameter λ for ridge regression.
2. In the spirit of Problem 34, show that leave-one-out CV is computationally tractable in case of ridge regression.

Problem 42. (Generalized least squares)
Consider the linear model $Y = X\beta + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y}$ is an $n\times 1$ vector of responses, 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 $\Sigma \neq \sigma^2 I_n$ 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

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

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

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

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

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

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 $(x_1, z_1), (x_2, z_2), \ldots, (x_n, z_n)$, it 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 $\widehat{f}$, then$\widehat{f} \in N(x_1, x_2, \ldots, x_n)$.