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 $\mathrm{rank}(\mathbf{H}) = p$.

Solution.

1. Symmetry is trivial. For idempotency,
$$\mathbf{H}_\mathbf{X}\mathbf{H}_\mathbf{X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}\mathbf{I}_p(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}_\mathbf{X}.$$
2. If $\mathbf{v}$ is an eigenvector of $\mathbf{H}$ associated to the eigenvalue λ, then $\mathbf{H}\mathbf{v} = \lambda\mathbf{v}$ by definition. But $\mathbf{H}$ is idempotent, so $\mathbf{H}^2\mathbf{v} = \lambda\mathbf{H}\mathbf{v} = \lambda^2\mathbf{v}$ and the only solutions of $\lambda^2 = \lambda$ are $\{0, 1\}$.
3. The matrix $\mathbf{H}$ is symmetric and idempotent. It remains to show its image is $\mathscr{S}(\mathbf{X})$. For any $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{H}\mathbf{y} = \mathbf{X}\widehat{\beta}$ with $\widehat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \in \mathbb{R}^p$. Thus $\mathrm{im}(\mathbf{H}) \subseteq \mathscr{S}(\mathbf{X})$, while at the same time $\mathbf{H}\mathbf{X} = \mathbf{X}$, so $\mathrm{im}(\mathbf{H}) \supseteq \mathscr{S}(\mathbf{X})$.

   $\mathbf{H}$ is not well-defined if $\mathbf{X}$ does not have rank p since the inverse $\mathbf{X}^\top\mathbf{X}$ does not exist. (but the solution in this case is rather simple: replace the matrix inverse by the pseudoinverse)
4. The trace is invariant to cyclic permutations of its arguments, so
$$\mathrm{tr}(\mathbf{H}) = \mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\right) = \mathrm{tr}(\mathbf{I}_p) = p.$$
   The trace is also equal to the sum of the eigenvalues of $\mathbf{H}$, which are either 0 or 1. There must be p non-zero eigenvalues, so by the spectral theorem $\mathrm{rank}(\mathbf{H}) = p$.

   If the columns of $\mathbf{X}$ are linearly dependent, there exists a non-zero vector $\mathbf{v} \in \mathbb{R}^p$ such that $\mathbf{X}\mathbf{v} = 0_p$, so $\mathbf{X}^\top\mathbf{X}\mathbf{v} = 0_p$ and $\mathbf{X}^\top\mathbf{X}$ is not injective, thus not invertible.

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 $\mathbf{P} = \mathbf{Q}$.

Solution. There are many ways to prove this. First, the column vectors of $\mathbf{P}$ are elements of $\mathcal{V}$. Consider a basis $\mathbf{V}$ of p orthogonal vectors in $\mathcal{V}$ and a basis of $n - p$ vectors $\mathbf{W}$ for $\mathcal{V}^\perp$. We can express the ith column vector of $\mathbf{P}$ as $\mathbf{p}_i = \mathbf{V}\alpha + \mathbf{W}\gamma$ for some coefficients $\alpha \in \mathbb{R}^p, \gamma \in \mathbb{R}^{n-p}$. Because $\mathbf{P}$ is idempotent, $\mathbf{P}\mathbf{p}_1 = \mathbf{p}_1$ and so $\gamma = 0_{n-p}$. This shows columns of $\mathbf{P} \in \mathcal{V}$, so $\mathbf{Q}\mathbf{P} = \mathbf{P}$ since $\mathbf{Q}$ is a projector. Similarly, $\mathbf{P}\mathbf{Q} = \mathbf{Q}$. Using symmetry,
$$\mathbf{Q} = \mathbf{P}\mathbf{Q} = \mathbf{P}^\top\mathbf{Q}^\top = (\mathbf{Q}\mathbf{P})^\top = \mathbf{P}^\top = \mathbf{P}.$$

Alternatively: for any $\mathbf{v} \in \mathscr{V}$, $\mathbf{v} = \mathbf{P}\beta$ for some β. Pre-multiply both sides by $\mathbf{P}$ and use the idempotency of projection matrices to get $\mathbf{P}\mathbf{v} = \mathbf{P}\mathbf{P}\beta = \mathbf{P}\beta = \mathbf{v}$.

We thus have $\mathbf{P}\mathbf{v} = \mathbf{v} = \mathbf{Q}\mathbf{v}$ for any $\mathbf{v} \in \mathcal{V}$. Since any vector $\mathbf{x} \in \mathbb{R}^n$ can be uniquely decomposed into two orthogonal vectors $\mathbf{x} = \mathbf{v} + \mathbf{w}$, where $\mathbf{v} \in \mathcal{V}$ and $\mathbf{w} \in \mathcal{V}^\perp$, $\mathbf{Q}\mathbf{x} = \mathbf{v} = \mathbf{P}\mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^n$ and thus $\mathbf{P} = \mathbf{Q}$.

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. ($H_1 = X_1(X_1^\top X_1)^{-1}X_1^\top$)

Solution. The key is to note that $\mathbf{H}\mathbf{X}_1 = \mathbf{X}_1$ since the columns of $\mathbf{X}_1$ are in $\mathscr{S}(\mathbf{X})$. It follows that $\mathbf{H}\mathbf{H}_1 = \mathbf{H}_1$ and, by transposing, that $\mathbf{H}_1\mathbf{H} = \mathbf{H}_1$. The matrix $\mathbf{H} - \mathbf{H}_1$ is symmetric since both $\mathbf{H}$ and $\mathbf{H}_1$ are symmetric.

[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.

The idempotency follows from the observation that

$$\begin{aligned}(\mathbf{H}-\mathbf{H}_1)(\mathbf{H}-\mathbf{H}_1) &= \mathbf{H}\mathbf{H}-\mathbf{H}_1\mathbf{H}-\mathbf{H}\mathbf{H}_1+\mathbf{H}_1\mathbf{H}_1\\ &= \mathbf{H}-\mathbf{H}_1\pm\mathbf{H}_1\mathbf{H}_1\\ &= \mathbf{H}-\mathbf{H}_1.\end{aligned}$$

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. $\frac{\partial}{\partial X}\mathrm{tr}(X) = I_n$.

Solution. a) Denote $y = Ax$. Hence, $y_i = \sum_{j=1}^n A_{ij}x_j$ and thus $\frac{\partial}{\partial x_j}y_i = A_{ij}$. We obtain

$$\frac{\partial}{\partial x}y = \begin{bmatrix}\frac{\partial}{\partial x_1}y_1 & \frac{\partial}{\partial x_1}y_2 & \cdots & \frac{\partial}{\partial x_1}y_n\\ \frac{\partial}{\partial x_2}y_1 & \frac{\partial}{\partial x_2}y_2 & \cdots & \frac{\partial}{\partial x_2}y_n\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial}{\partial x_n}y_1 & \frac{\partial}{\partial x_n}y_2 & \cdots & \frac{\partial}{\partial x_n}y_n\end{bmatrix} = \begin{bmatrix}A_{11} & A_{21} & \cdots & A_{n1}\\ A_{12} & A_{22} & \cdots & A_{n2}\\ \vdots & \vdots & \ddots & \vdots\\ A_{1n} & A_{2n} & \cdots & A_{nn}\end{bmatrix} = A^\top.$$

b) Denote $y = x^\top Ax = \sum\limits_{i=1}^n\sum\limits_{j=1}^n A_{ij}x_ix_j$. We have

$$\begin{aligned}\frac{\partial}{\partial x_k}y &= \sum_{i\neq k}A_{ki}x_i + \sum_{i\neq k}A_{ik}x_i + 2A_{kk}x_k\\ &= \sum_{i=1}^n A_{ki}x_i + \sum_{i=1}^n A_{ik}x_i = (Ax)_k + (A^\top x)_k = (Ax+A^\top x)_k.\end{aligned}$$

Thus

$$\frac{\partial}{\partial x}y = Ax + A^\top x = (A+A^\top)x.$$

c) Denote $y = \mathrm{tr}(X) = \sum\limits_{i=1}^n X_{ii}$. Then $\frac{\partial}{\partial X_{ij}}y = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases}1, & \text{if } i=j,\\ 0, & \text{if } i\neq j\end{cases} \tag{1}$$

is the Kronecker delta. Thus $\frac{\partial}{\partial X}y = I_n$.

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

Solution. (a) Recall that $\mathbf{X}$ is full rank if and only if $\mathbf{X}$ is injective and if and only if $\ker(\mathbf{X}) = \{0_p\}$. If $\mathbf{v}\in\mathbb{R}^p\setminus\{0_p\}$,

$$\mathbf{v}^\top\mathbf{B}\mathbf{v} = \mathbf{v}^\top\mathbf{X}^\top\Omega\mathbf{X}\mathbf{v} = (\mathbf{X}\mathbf{v})^\top\Omega\mathbf{X}\mathbf{v} > 0$$

since $\mathbf{X}\mathbf{v}\neq 0_n$ and Ω is positive definite. It follows that $\mathbf{B}$ is also positive definite and thus invertible. The second part follows from the first upon taking $\Omega = \mathbf{I}_n$, which is positive definite.

(b) Counter-example. With $\mathbf{X} = (1,1)^\top$ and $\Omega = \left(\begin{smallmatrix}1 & 0\\ 0 & -1\end{smallmatrix}\right)$, we get $\mathbf{X}^\top\Omega\mathbf{X} = 0$. In general, if Ω has one positive eigenvalue a and one negative eigenvalue b, one can find a matrix $\mathbf{X}$ such that $\mathbf{X}^\top\Omega\mathbf{X} = 0$.

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

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

Solution. An easy calculation.

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 $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Solution. By writing explicitly the relationship $\mathbf{Q}\Sigma = \mathbf{I}_n$, we get

$$\begin{aligned} \mathbf{Q}_{11}\Sigma_{11} + \mathbf{Q}_{12}\Sigma_{21} &= \mathbf{I}_k \\ \mathbf{Q}_{21}\Sigma_{12} + \mathbf{Q}_{22}\Sigma_{22} &= \mathbf{I}_{p-k} \\ \mathbf{Q}_{21}\Sigma_{11} + \mathbf{Q}_{22}\Sigma_{21} &= \mathbf{O}_{p-k,k} \\ \mathbf{Q}_{11}\Sigma_{12} + \mathbf{Q}_{12}\Sigma_{22} &= \mathbf{O}_{k,p-k}. \end{aligned}$$

Recall that we can only invert matrices whose double indices are identical and that both $\mathbf{Q}$ and Σ are symmetric, so $\Sigma_{12} = \Sigma_{21}^\top$. One easily obtains

(a) $\Sigma_{12}\Sigma_{22}^{-1} = -\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}$ making use of the last equation.

(b) $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \mathbf{Q}_{11}^{-1}$ by substituting $\mathbf{Q}_{12}$ from the last equation into the first.

(c) One can cleverly choose $B \coloneqq \begin{pmatrix} \mathbf{I} & -\Sigma_{12}\Sigma_{22}^{-1} \\ \mathbf{O} & \mathbf{I} \end{pmatrix}$, noting that $\det(B) = \det\left(B^\top\right) = 1$. Computing the quadratic form $B\Sigma B^\top$, we get $\det(\Sigma) = \det(\Sigma_{22}) \det(\Sigma_{1|2})$ 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 $1 \leq k < n$. 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.

Solution. Without the loss of generality, assume means are 0. It is easy to generalize the solution below for the case of a non-zero mean. Following the hint, we write:

$$\begin{aligned} f(y_1|y_2) \propto f(y_1|y_2)f(y_2) = f(y_1,y_2) &\propto \exp\left(-\frac{1}{2}(y_1,y_2)\mathbf{Q}(y_1,y_2)^\top\right) \\ &\propto \exp\left(-\frac{1}{2}y_1^\top\mathbf{Q}_{11}y_1 - y_1^\top\mathbf{Q}_{12}y_2 - \frac{1}{2}y_2^\top\mathbf{Q}_{22}y_2\right) \\ &\propto \exp\left(-\frac{1}{2}(y_1+\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}y_2)^\top\mathbf{Q}_{11}(y_1+\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}y_2)\right) \end{aligned}$$

Firstly, we are using proportionality signs in our calculation, which is often convenient, and here it is almost necessary to keep the solution of this exercise simple. Note that every density has to integrate into one, so whatever factors that do not depend on the variable y_1 can be discarded. Secondly, on the first line of the calculation, we just wrote the density of multivariate normal using the precision matrix, discarding constants as described above. On the second line, we just developed the expression w.r.t. to the blocks of the precision matrix. On the final line, we *completed the square* in y_1 and separated the term depending on y_1 from the rest. And voila! Up to proportionality we have a Gaussian in y_1 considering y_2 as fixed. This means that we know the conditional distribution of Y_1 given Y_2.

Once we established that the conditional distribution is Gaussian, we can just read the mean and variance from the exponential. We see the precision matrix is $\mathbf{Q}_{11}$ which by Problem 7 (c) is exactly the Schur complement and Problem 7 (a) gives us an expression for the mean:

$$\begin{aligned} \mu_{1|2} &= -\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}y_2 \\ &= \Sigma_{12}\Sigma_{22}^{-1}y_2 \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

Finally, go through the argument again to see what changes when the means are not assumed to be zero.

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 $k = \operatorname{tr}(\mathbf{A}\Sigma)$.

Solution. Recall the affine transformation property of the normal distribution:

$$Y \sim \mathcal{N}(\mu, \Sigma) \implies \mathbf{B}Y + \theta \sim \mathcal{N}(\theta + \mathbf{B}\mu, \mathbf{B}\Sigma\mathbf{B}^\top). \tag{S1}$$

The Gaussian distribution is a location-scale family.

(a) Follows from (S1) and the fact that $\mathbf{A}$ is orthogonal, so $\mathbf{A}^\top\mathbf{A} = \mathbf{I}_n$.

(b) The matrix $\mathbf{C}$ is invertible because its diagonal elements are all strictly positive. Since $\mathbf{C}^{-1}(Y-\mu) = \mathbf{C}^{-1}Y - \mathbf{C}^{-1}\mu$, it follows from (S1) that $\mathbf{C}^{-1}(Y-\mu)$ is normal with mean $\mathbf{C}^{-1}\mu - \mathbf{C}^{-1}\mu = 0_n$ and covariance $\mathbf{C}^{-1}\Sigma\mathbf{C}^{-\top} = \mathbf{C}^{-1}\mathbf{C}\mathbf{C}^\top\mathbf{C}^{-\top} = \mathbf{I}_n$.

(c) By definition, the $\chi^2(k)$-distribution is the distribution of $X^\top X$ for $X \sim \mathcal{N}_k(0_k, \mathbf{I}_k)$. We rewrite $Z^\top\mathbf{H}Z$ in the form $X^\top X$ by using the spectral decomposition

$$\mathbf{H} = \mathbf{U}\Lambda\mathbf{U}^\top = \sum_{i=1}^{n}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top \overset{\lambda_i\in\{0,1\}}{\underset{\text{Prob. 1.2}}{=}} \sum_{i=1}^{k}\mathbf{u}_i\mathbf{u}_i^\top =: \underbrace{\widetilde{\mathbf{U}}}_{n\times k}\widetilde{\mathbf{U}}^\top$$

to set

$$Z^\top\mathbf{H}Z = Z^\top\widetilde{\mathbf{U}}\underbrace{\widetilde{\mathbf{U}}^\top Z}_{=:X} = X^\top X.$$

And indeed

$$\text{cov}(X) = \text{cov}(\widetilde{\mathbf{U}}^\top Z) = \widetilde{\mathbf{U}}^\top \underbrace{\text{cov}(Z)}_{=\mathbf{I}_n} \widetilde{\mathbf{U}} = \widetilde{\mathbf{U}}^\top \widetilde{\mathbf{U}} = \mathbf{I}_k$$

such that X has the desired distribution, which shows that $Z^\top \mathbf{H} Z = X^\top X \sim \chi^2(k)$.

(d) Since Σ is invertible it is positive definite. Write its Cholesky decomposition $\Sigma = \mathbf{C}\mathbf{C}^\top$, where $\mathbf{C}$ is invertible. From b), $Z \coloneqq \mathbf{C}^{-1}(Y-\mu) \sim \mathcal{N}_n(0_n, \mathbf{I}_n)$ and

$$(Y-\mu)^\top \Sigma^{-1}(Y-\mu) = (Y-\mu)^\top \mathbf{C}^{-\top}\mathbf{C}^{-1}(Y-\mu) = Z^\top Z = Z^\top \mathbf{I}_n Z.$$

The result now follows from c) since the identity matrix $\mathbf{I}_n$ is a projection matrix of rank n.

(e) Using the solution in part d), we can write

$$(Y-\mu)^\top \mathbf{A}(Y-\mu) = (\mathbf{C}Z)^\top \mathbf{A}\mathbf{C}Z = Z^\top \mathbf{C}^\top \mathbf{A}\mathbf{C}Z = Z^\top \mathbf{H} Z.$$

Note that $\mathbf{H}$ is a symmetric matrix. Also,

$$\mathbf{H}^2 = \mathbf{C}^\top \mathbf{A}\mathbf{C}\mathbf{C}^\top \mathbf{A}\mathbf{C} = \mathbf{C}^\top \mathbf{A}\Sigma\mathbf{A}\mathbf{C} = \mathbf{C}^\top \mathbf{A}\mathbf{C} = \mathbf{H}.$$

So, $\mathbf{H}$ is idempotent. This shows that $\mathbf{H}$ is a projection matrix. So, by part c), $Z^\top \mathbf{H} Z \sim \chi^2(k)$, where k is the rank of $\mathbf{H}$. Now,

$$\text{rank}(\mathbf{H}) = \text{tr}(\mathbf{H}) = \text{tr}(\mathbf{C}^\top \mathbf{A}\mathbf{C}) = \text{tr}(\mathbf{A}\mathbf{C}\mathbf{C}^\top) = \text{tr}(\mathbf{A}\Sigma).$$

Note: Part d) can be seen as a special case of part e).

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 $\mathbf{V}$.

Solution. The fact that both $\mathbf{U}$ and $\mathbf{V}$ are orthonormal means that $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_p$. The hat matrix is

$$\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top = \mathbf{U}\Omega\mathbf{V}^\top\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\top\mathbf{V}\mathbf{D}\mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top,$$

since $\mathbf{D} = \mathbf{D}^\top$ and $\mathbf{V}^{-1} = \mathbf{V}^\top$, thus $(\mathbf{V}\mathbf{D}^2\mathbf{V}^\top)^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\top$.

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

Solution. Here is an example of solution (there could be others). The fixed parameters are underlined (e.g. $\underline{\beta_0}$).

a) $y = \begin{pmatrix}1 & \frac{1}{x} & \frac{1}{x^2}\end{pmatrix}(\beta_0 \quad \beta_1 \quad \beta_2)^\top + \varepsilon$ with $x \neq 0$

b) $y = (\frac{1}{1+\underline{\beta_1} x})(\beta_0) + \varepsilon$ with $x \neq 0$

c) $y = (1/x)(\gamma) + \varepsilon$ with $\gamma = \beta_0/\beta_1$ or $y = (\frac{1}{x\underline{\beta_1}})(\beta_0) + \varepsilon$ with $x \neq 0$

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

e) $y = (1 \quad x^{\underline{\beta_2}})(\beta_0 \quad \beta_1)^\top + \varepsilon$

f) $y = (1 \quad x_1^{\beta_2} \quad x_2^{\beta_4})(\beta_0 \quad \beta_1 \quad \beta_3)^\top + \varepsilon$

g) $\log(y) = (1 \quad \log(x_1) \quad \log[\cos(x_2)])(\log(\beta_1) \quad \beta_2 \quad \beta_3)^\top + \log(\varepsilon)$ with $x_1, \varepsilon > 0$ and $\cos(x_2) > 0$

h) $\log(y - \underline{\beta_1}) = (\log(x_1) \quad \log[2 + \cos(x_2)])(\beta_2 \quad \beta_3)^\top + \log(\varepsilon^2 + 1)$ with $x_1 > 0$.

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?

Solution. a) The design matrix is

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix}$$

One can straightforwardly calculate

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \quad \mathbf{X}^\top Y = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}$$

and use the 2×2 matrix inversion formula to get

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

b) Formula for $\widehat{\beta}$ follows easily by multiplying $\widehat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y$, though we are only interested in the second element of the resulting vector.

Assume now that the data are standardized, i.e. both x and Y have (empirical) mean zero and (empirical) variance one. Then $\widehat{\beta}_1$ reduces to the empirical correlation coefficient between x and Y, and it is the slope of the regression line when data are plotted. When we alleviate the assumption that our data are standardized, the interpretation of $\widehat{\beta}_1$ as the slope of the regression line is retained.

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

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

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:

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

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

I. Give the design matrix corresponding to model (2).

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) $\mathtt{z} = (0, 1, 5, 2, 1, 1)^\top$, 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`?

Solution. a) I. From the model equation (2), the regression function is

$$\mathbb{E}y_j = \beta_0 + \alpha_1 \mathbb{1}_{a_j=1} + \alpha_2 \mathbb{1}_{a_j=2}$$

from which we can easily read the regression matrix. The first vector of ones corresponds to the intercept. The second vector is vector of ones followed by zeros, because our dataset is ordered that way: group 1 precedes group 2 in vector `a`. The third vector is then the complement of the second one.

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \beta = (\beta_0, \alpha_1, \alpha_2)^\top.$$

II. Superposition of the second and third columns is exactly the first column, hence the matrix is not full rank. The consequence is that we cannot invert $\mathbf{X}^\top\mathbf{X}$, so our usual (unique) estimator

$$\widehat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \mathrm{y}$$

is not defined.

Let us provide a brief explanation. We model the mean stress value and this model only allows it to be different for the two groups (female/male). So there are only two quantities the mean stress value can attain, say μ_1 for females and μ_2 for males. But the model has 3 parameters, so there is too much freedom and multiple values of the parameters lead to the same fit (more precisely, to the same fitted values), so the model has no means to distinguish between these values of the parameters. This behavior is, of course, undesirable. One way to remedy is to get rid of one of the parameters.

III. $\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$

Suppressing the column corresponding to α_1 corresponds to leaving the parameter from the regression function, which now becomes:

$$\mathbb{E}y_j = \beta_0 + \alpha_2 \mathbb{1}_{a_j=2}$$

Now, β_0 is the mean of each observation in the group $a_j = 1$ and α_2 is the difference between the average of group $a_j = 2$ and the average of group $a_j = 1$.

Any other parameter could have been suppressed instead of α_1. However, suppressing the first level of a factor to obtain the interpretation in the previous paragraph is the default in `R`, and we will always take this approach.

IV. (i) Regressing `y` on `b` is similar. We again suppress the first level of `b` to have a full-rank design matrix, so this time, $\mathbf{X}$ will have 3 columns: 1 for intercept and 3-1=2 for factor `b`. The regression function and the design matrix are:

$$\mathbb{E}y_j = \beta_0 + \gamma_2 \mathbb{1}_{(b_j=2)} + \gamma_3 \mathbb{1}_{(b_j=3)} \quad , \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

(ii) Model `y~x+a` is almost like the one we had before (`y~a`), with the difference that there will be now one extra parameter and thus one extra column corresponding to the linear term in `x`:

$$\mathbb{E}y_j = \beta_0 + \beta_1 x_j + \alpha_2 \mathbb{1}_{(a_j=2)} \quad , \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

(iii) The final model `y~a+b` contains two factors. The first level will be suppressed for both.

$$\mathbb{E}y_j = \beta_0 + \alpha_2 \mathbb{1}_{(a_j=2)} + \gamma_2 \mathbb{1}_{(b_j=2)} + \gamma_3 \mathbb{1}_{(b_j=3)} \quad , \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

Notice that the design matrix of `y~a+b` is the "union" of the design matrices for models `y~a` and `y~b`.

b) Here we just provide the regression functions. You can verify the answers about the design matrices in `R` using the code below.

(i) $\mathbb{E}y_j = \beta_0 + \beta_1 x_j + \gamma_2 \mathbb{1}_{(b_j=2)} + \gamma_3 \mathbb{1}_{(b_j=3)} + \underbrace{\beta_2 x_j \mathbb{1}_{(b_j=2)} + \beta_3 x_j \mathbb{1}_{(b_j=3)}}_{=:(\star)}$, where $(\star)$ denotes the extra terms added by the interaction compared to model without the interaction: `y ~ x + b`.

(ii) $\mathbb{E}y_j = \beta_0 + \beta_1 x_j + \beta_2 z_j + \beta_3 x_j z_j$

(iii) $\mathbb{E}y_j = \beta_0 + \alpha_2 \mathbb{1}_{(a_j=2)} + \gamma_2 \mathbb{1}_{(b_j=2)} + \gamma_3 \mathbb{1}_{(b_j=3)} + \delta_5 \mathbb{1}_{(a_j=2)} \mathbb{1}_{(b_j=2)} + \delta_6 \mathbb{1}_{(a_j=2)} \mathbb{1}_{(b_j=3)}$

(iv) $\mathbb{E}y_j = \beta_0 + \beta_1 z_j + \beta_2 x_j^2$

You can verify the answers about the design matrices in `R`. First, import the data set manually (just copy-paste the code below):

```
y <- c(217,143,186,121,157,143)
X <- matrix(c(1,0,2,0,1,0,1,1,1,2,2,2,1,2,3,1,2,3,0,1,5,2,1,1), 6, 4)
dfX <- data.frame(y = y, x = X[,1], a = as.factor(X[,2]), b = as.factor(X[,3]), z = X[,4])
```

Now you can use the command

```
model.matrix(y~expression, data = dfX)
```

where you specify properly the expression to check your answers:

```
model.matrix(y~x*b, data = dfX) # (i)
model.matrix(y~x*z, data = dfX) # (ii)
model.matrix(y~a*b, data = dfX) # (iii)
model.matrix(y~z+I(x^2), data = dfX) # (iv)
```

Notice that model (iv) does not contain the main effect of `x`. Such a model can be rarely useful.

c) The triple interaction term may be too hard to just write down the formula from the top of the head. Note that the model is equivalent to `y ~ x + a + b + x:a + x:b + a:b + x:a:b`. I recommend reading this developed expression left to right and writing down the regression function in steps, writing it first only for `y ~ x`, then for `y ~ x + a`, etc. It leads to the following:

$$\mathbb{E}y_j = \beta_0 + \underbrace{\beta_1 x_j}_{\texttt{x}} + \underbrace{\alpha_2 \mathbb{1}_{(a_j=2)}}_{\texttt{a}} + \underbrace{\gamma_2 \mathbb{1}_{(b_j=2)} + \gamma_3 \mathbb{1}_{(b_j=3)}}_{\texttt{b}} + \underbrace{\beta_2 x_j \mathbb{1}_{(a_j=2)}}_{\texttt{x:a}} + \underbrace{\beta_3 x_j \mathbb{1}_{(b_j=2)} + \beta_4 x_j \mathbb{1}_{(b_j=3)}}_{\texttt{x:b}}$$
$$+ \underbrace{\delta_5 \mathbb{1}_{(a_j=2)} \mathbb{1}_{(b_j=2)} + \delta_6 \mathbb{1}_{(a_j=2)} \mathbb{1}_{(b_j=3)}}_{\texttt{a:b}} + \underbrace{\beta_5 x_j \mathbb{1}_{(a_j=2)} \mathbb{1}_{(b_j=2)} + \beta_6 x_j \mathbb{1}_{(a_j=2)} \mathbb{1}_{(b_j=3)}}_{\texttt{x:a:b}}$$

Several notes are in order here. Firstly, we can use whatever symbols and subscripts to denote the parameters. Try to find your own system. Secondly, the triple interaction term `x:a:b` corresponds to multiplying the simpler interaction `a:b` with x_j. The applied meaning of this will hopefully become clear later.

d) Model (i) has 3 parameters while model (ii) only has 2 parameters. One can in fact show that model (i) is more general. Model (i) has the regression function

$$\mathbb{E}y_j = \beta_0 + \alpha_2 \mathbb{1}_{(b_j=2)} + \alpha_3 \mathbb{1}_{(b_j=2)}$$

from which we can deduce

$$\begin{aligned} b_j = 1 \quad &\Rightarrow \quad \mathbb{E}y_j = \beta_0 \\ b_j = 2 \quad &\Rightarrow \quad \mathbb{E}y_j = \beta_0 + \alpha_2 \\ b_j = 3 \quad &\Rightarrow \quad \mathbb{E}y_j = \beta_0 + \alpha_3 \end{aligned}$$

so we can see that the difference between $b_j = 1$ and $b_j = 2$ is given by α_2, while the difference between $b_j = 1$ and $b_j = 3$ is given by α_3, which has no relationship with α_2. Hence the sample is split into 3 groups by variable `b`, and every group is allowed to have a different mean.

On the other hand, with the model (ii) we have

$$\mathbb{E}y_j = \beta_0 + \beta_1 b_j$$

from which we can deduct

$$\begin{aligned} b_j = 1 \quad &\Rightarrow \quad \mathbb{E}y_j = \beta_0 + \beta_1 \\ b_j = 2 \quad &\Rightarrow \quad \mathbb{E}y_j = \beta_0 + 2\beta_1 \\ b_j = 3 \quad &\Rightarrow \quad \mathbb{E}y_j = \beta_0 + 3\beta_1 \end{aligned}$$

so we can see that the difference between $b_j = 1$ and $b_j = 2$ is given by β_1, while the difference between $b_j = 1$ and $b_j = 3$ is given by $2\beta_1$. This model thus linearly constrains the differences between the three groups: the

difference between the mean of the third group and the first group is exactly double the difference between the second group and the first group.

Note: When building a linear model and an ordinal variable such as `b` is available, one has to decide whether to include that variable as a factor or as a numerical variable based on the consideration in the previous paragraph. In our toy example, can we assume that the effect of being single vs. being married is exactly the same as the effect of being divorced vs being married? In this case, for sure not. So we should start with `b` as a factor first, and maybe later simplify the model to `b` being numeric, based on what the data actually suggest.

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)?

Solution. The plots for every subquestion are given in the figure below.

a) The data are stored as an `.RData` file, hence it can be simply loaded as `load("percentile.RData")`. Then one can form the scatterplot using `plot(DATA$percentile ~ DATA$GPA)`.

b)
```
m1 <- lm(percentile ~ GPA, data=DATA)
summary(m1)
abline(m1$coefficients[1],m1$coefficients[2])
```

The previous commands tell us that the correlation between *percentile* and *GPA* is negative (estimated regression coefficient is -3.773). Improving a student's *GPA* by 1 leads to a decrease in his percentile by 3.773. This seems somewhat counterintuitive. One would expect that better *GPA* should be associated with better *percentile*, provided that the education system is working.

c)
```
m2 <- lm(percentile ~ GPA+as.factor(grade), data=DATA)
summary(m2)
plot(DATA$percentile[DATA$grade==8] ~ DATA$GPA[DATA$grade==8],
     col="blue",xlim=c(1,4), ylim=c(10,100), main="c)")
points(DATA$percentile[DATA$grade==12] ~ DATA$GPA[DATA$grade==12],
       col="red",pch=0)
abline(m2$coefficients[1],m2$coefficients[2],col="blue")
abline(m2$coefficients[1]+m2$coefficients[3],m2$coefficients[2],col="red")
```

Once the variable *grade* is accounted for by the model, not only *GPA* becomes significant but the negative dependence from part b) suddenly becomes positive, as one would expect. Since *grade* has only 2 levels (students are either from the 8th grade or 12th) it makes sense to treat different classes like two different groups (hence the coloring in the plot). Quantitatively, the model says that if student A has *GPA* larger by 1 than the *GPA* of student B, student A's percentile is expected to be higher than that of student B by 16.884.

d) The code here is only a slight modification of the previous one. Note that if one naturally wants both the interaction and the main terms, the $*$ operator can be used as

```
m3 <- lm(percentile ~ GPA*as.factor(grade), data=DATA)
```

While the model from part c) only allowed for the intercept to be different for the two groups of students and the slope was fixed to be the same, model `m3` allows for both the intercept and the slope to be different for the two groups of students. Qualitative conclusions remain roughly the same, but one can notice that *GPA* has a slightly stronger effect among the 8th grade students. To put it in numbers, if student A has *GPA* larger by 1 than *GPA* of student B, student A's percentile is expected to be higher by 20.626 (respectively by

20.626-7.862=12.764) than that of student B in the case of both students being in the 8th grade (respectively the 12th grade).

The variable *grade* is the so-called confounder of the relationship between *percentile* and *GPA*. If *grade* is not accounted for, the model produces completely wrong results. In this case, including *grade* changes the negative relationship to a positive one, which is called Simpson's paradox. Paradox, because even though higher values of *GPA* are naturally associated with higher values of *percentile* in both of the two classes appearing in our data set, it seems at the first glance that the overall correlation between *GPA* and *percentile* is negative. A sensible explanation of this could be the following: younger students usually have better GPA's because they put more effort into their studies, but they are not yet educated enough to be able to score higher than their older colleagues on a standardized test.

Often, a confounder is not taken into account in a study, which then leads to insensible conclusions and subsequent tabloid headings such as "Want to go to Harvard? Fail high school first!".

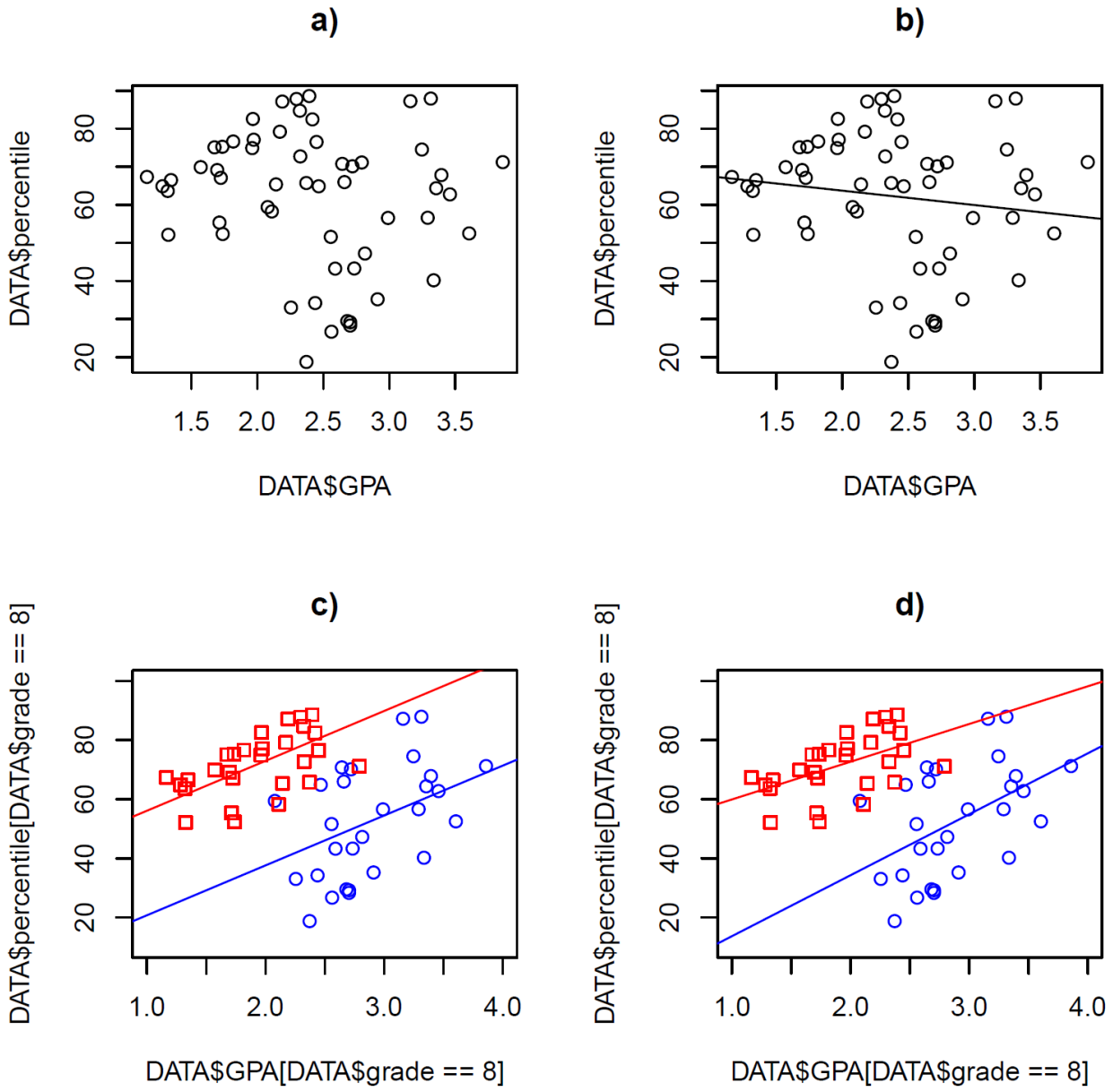

Figure 1: Standardised residuals as a function of values adjusted for four Gaussian models.

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?

Solution. The solutions are not unique, they depend on the ordering of variables.

a) A possible regression function can be

$$\mathbb{E}Y = \beta_0 + \beta_1 u + \beta_2 w + \beta_3 g + \beta_4 ug + \beta_5 wg.$$

Note that since g attains only two values, it does not matter in this case whether it is considered as a factor or as a continuous variable. But it would be more natural to consider it as a factor:

$$\mathbb{E}Y = \beta_0 + \beta_1 u + \beta_2 w + \beta_3 \delta_{[g=1]} + \beta_4 u\delta_{[g=1]} + \beta_5 w\delta_{[g=1]},$$

where δ is the identifier operator.

b) The following matrix is the design matrix corresponding to the regression function from part a):

$$X = \begin{pmatrix} 1 & 10 & 80 & 0 & 0 & 0 \\ 1 & 8 & 60 & 1 & 8 & 60 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

c) One would like to test $H_0 : \beta_5 = 0$ against $H_1 : \beta_5 \neq 0$. One possibility is to form the confidence interval for β_5 based on the t-distribution (see slide 87) and check whether 0 is contained in this confidence interval.

Note: Generally, the F-test is preferable to the t-test described above, but we will only learn about the F-test later.

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 $\mathbf{M}_{\mathbf{X}_1} \coloneqq \mathbf{I}_n - \mathbf{H}_{\mathbf{X}_1}$. 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 $\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2$. (Draw a 3D picture to visualize what this result actually says.)

Solution. We need to show that $\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$ is an orthogonal projection matrix, i.e., it is idempotent, symmetric and it spans $\mathscr{S}(\mathbf{X})$. Note that $\mathbf{X}_1^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2 = \mathbf{O}$, so $\mathbf{H}_{\mathbf{X}_1}\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2} = \mathbf{O}$ also. Since both $\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$ and $\mathbf{H}_{\mathbf{X}_1}$ are orthogonal projection matrices, the first two statements are obvious.

It remains to show that any vector $\mathbf{z} \in \mathscr{S}(\mathbf{X})$ is invariant under the action of $\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$ and that any vector orthogonal to this span is annihilated by $\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$. Since $\mathbf{X}$ is full rank, we can write $\mathbf{z} = \mathbf{X}\gamma = \mathbf{X}_1\gamma_1 + \mathbf{X}_2\gamma_2$ for some vector γ and subvectors γ_1 and γ_2. Then

$$\begin{aligned}(\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2})\mathbf{z} &= (\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2})(\mathbf{X}_1\gamma_1 + \mathbf{X}_2\gamma_2) \\ &= \mathbf{H}_{\mathbf{X}_1}(\mathbf{X}_1\gamma_1 + \mathbf{X}_2\gamma_2) + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}(\mathbf{X}_1\gamma_1 + \mathbf{X}_2\gamma_2) \\ &= \mathbf{X}_1\gamma_1 + \mathbf{H}_{\mathbf{X}_1}\mathbf{X}_2\gamma_2 + \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2\gamma_2 \\ &= \mathbf{X}_1\gamma_1 + \mathbf{X}_2\gamma_2\end{aligned}$$

upon noting that

$$\begin{aligned}\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}\mathbf{X}_1 &= \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2(\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_1 = \mathbf{O}, \\ \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}\mathbf{X}_2 &= \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2(\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2 = \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2.\end{aligned}$$

Take now $\mathbf{w} \in \mathscr{S}^\perp(\mathbf{X})$. We have

$$\begin{aligned}(\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2})\mathbf{w} &= \mathbf{H}_{\mathbf{X}_1}\mathbf{w} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}\mathbf{w} \\ &= 0 + \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2(\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{w} \\ &= \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2(\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top (\mathbf{I} - \mathbf{H}_{\mathbf{X}_1})\mathbf{w} = \mathbf{0}.\end{aligned}$$

Indeed, $\mathbf{H}_{\mathbf{X}_1}\mathbf{w} = \mathbf{0}$ because $\mathbf{w}$ is orthogonal to $\mathbf{X}$, thus also orthogonal to $\mathbf{X}_1$. At the same time, $\mathbf{X}_2^\top \mathbf{w} = \mathbf{0}$ by orthogonality. By the uniqueness of projection matrices, the result follows.

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 $\beta_2 - \beta_3$.

Solution. a) The covariance of $\hat{\beta}$ is given by $\mathsf{Var}\hat{\beta} = \sigma^2 (X^\top X)^{-1}$. Since we do not know σ^2, we estimate the covariance with $\widehat{\mathrm{var}}(\hat{\beta}) = S^2 (X^\top X)^{-1}$. Denoting $v_{ij} = ((X^\top X)^{-1})_{ij}$, $i = 0, 1, 2, 3$, $j = 0, 1, 2, 3$ (note that we start by the 0 indices). Hence, the i-th standardised error is estimated by $\widehat{\mathrm{SE}}(\hat{\beta}_i) = \sqrt{\widehat{\mathrm{var}}(\hat{\beta})_{ii}} = \sqrt{S^2 v_{ii}}$. For the correlation, we have

$$\widehat{\mathrm{corr}}(\hat{\beta}_i, \hat{\beta}_j) = \frac{\widehat{\mathrm{var}}(\hat{\beta})_{ij}}{\sqrt{\widehat{\mathrm{var}}(\hat{\beta})_{ii}}\sqrt{\widehat{\mathrm{var}}(\hat{\beta})_{jj}}} = \frac{S^2 v_{ij}}{\sqrt{S^2 v_{ii}}\sqrt{S^2 v_{jj}}} = \frac{v_{ij}}{\sqrt{v_{ii} v_{jj}}}.$$

b) We recall that the forecast is given by

$$\hat{y}_+ = x_+^\top \hat{\beta}.$$

Here, we have

$$\hat{y}_+ = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3.$$

For $x_1 = x_2 = x_3 = 1$, the expectation would increase of $4\hat{\beta}_1 = 4 \times 1.70 = 6.80$ if $x_1 = 5$, and of $4\hat{\beta}_2 = 4 \times 0.66 = 2.64$ if $x_2 = 5$. Explicitly,

$$\begin{aligned}
x_1 = x_2 = x_3 = 1 &\implies \hat{y}_+ = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 48.19 + 1.70 + 0.66 + 0.25 = 50.80 \\
x_1 = 5, x_2 = x_3 = 1 &\implies \hat{y}_+ = 48.19 + 1.70 \times 5 + 0.66 + 0.25 = 57.60 \\
x_1 = x_2 = 5, x_3 = 1 &\implies \hat{y}_+ = 48.19 + 1.70 \times 5 + 0.66 \times 5 + 0.25 = 60.24
\end{aligned}$$

c) Let us denote here $(X^\top X)^{-1} = (v_{ij})_{i,j=0}^3$. The entries v_{ij} can be read out of the R output provided in the assignment.

Recall that for the i-th coordinate of β, the confidence interval is

$$\hat{\beta}_i \pm \sqrt{S^2 v_{ii}}\, t_{n-p}(\alpha/2) = \hat{\beta}_i \pm \widehat{\mathrm{SE}}(\hat{\beta}_i)\, t_{n-p}(\alpha/2), \quad i = 0, 1, 2, 3.$$

Here, $n = 13$, $p = 4$, $\alpha = 0.05$, $t_9(0.975) = 2.262$, so we have the four intervals:

$$[39.34, 57.04], \quad [1.236, 2.164], \quad [0.5605, 0.7595], \quad [-0.1685, 0.6685].$$

For the test $\beta_3 = 0$ against $\beta_3 \neq 0$ we do not reject the null hypothesis because $0 \in [-0.1685, 0.6685]$.

More generally, if $c \in \mathbb{R}^p$, the confidence interval for $c^\top \beta$ is given by

$$c^\top \hat{\beta} \pm t_{n-p}(\alpha/2) \sqrt{S^2 c^\top (X^\top X)^{-1} c}.$$

Here we want a confidence interval for $c^\top \beta$ with $c = (0, 0, 1, -1)^\top$. We find

$$\begin{aligned} S^2 c^\top (X^\top X)^{-1} c &= S^2 v_{22} + S^2 v_{33} - 2\frac{v_{23}}{\sqrt{v_{22} v_{33}}} \sqrt{S^2 v_{22}} \sqrt{S^2 v_{33}} \\ &= \left(\widehat{\mathrm{SE}}(\hat\beta_2)\right)^2 + \left(\widehat{\mathrm{SE}}(\hat\beta_3)\right)^2 - 2\,\widehat{\mathrm{corr}}(\hat\beta_2, \hat\beta_3)\,\widehat{\mathrm{SE}}(\hat\beta_2)\,\widehat{\mathrm{SE}}(\hat\beta_3) \\ &= 0.044^2 + 0.185^2 - 2 \cdot (-0.089) \cdot 0.044 \cdot 0.185 \end{aligned}$$

Thus we have

$$[0.66 - 0.25 \pm \left\{0.044^2 + 0.185^2 - 2 \cdot 0.044 \cdot 0.185 \cdot (-0.089)\right\}^{1/2} t_9(0.95)] = [0.055, 0.765]$$

as 0.90 confindence interval for $\beta_2 - \beta_3$.

In `R`,

```
library(MASS)
fit<-lm(y~1+x1+x2+x3, data=cement)
confint(fit)
```

donne

```
2.5 %     97.5 %
(Intercept) 39.3411244 57.0461442
x1           1.2330935  2.1586869
x2           0.5568501  0.7569797
x3          -0.1678276  0.6678628
```

for the confidence intervals of each coordinate of β.

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.

Solution (a))**.** Let us compute $\tilde H$ for each case:

I. The singular values decomposition of $X_{n\times p}$ is $U\Lambda V^\top$, with $\Lambda_{n\times p}$ diagonal, i.e.,

$$\Lambda = \begin{pmatrix} \Lambda_1 \\ 0 \end{pmatrix},$$

where Λ_1 is a $p \times p$ diagonal matrix. Since $\tilde X = AX = \Lambda V^\top$,

$$\tilde X^\top \tilde X = V \Lambda_1^2 V^\top,$$

and its inverse is given by $V\Lambda_1^{-2}V^\top$ (Λ_1 is invertible since X is injective) and

$$\tilde H = \tilde X (\tilde X^\top \tilde X)^{-1} \tilde X^\top = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}.$$

II. Since $\tilde{X} = R = (R_1, 0)^\top$, we have

$$\tilde{H} = \tilde{X}(\tilde{X}^\top \tilde{X})^{-1}\tilde{X}^\top = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence, in the two cases

$$\hat{\tilde{y}} = (\tilde{y}_1, \dots, \tilde{y}_p, 0, \dots, 0)^\top$$

and

$$\tilde{e} = (0, \dots, 0, \tilde{y}_{p+1}, \dots, \tilde{y}_n)^\top.$$

The first p coordinates of $\hat{\tilde{y}}$ are equal to those of $\tilde{y}$, the last $n-p$ are zeros. The first p coordinates of $\tilde{e}$ are zeros, and its last $n-p$ coordinates are $\tilde{y}_i$, $i = n-p, \dots, n$. De plus,

$$\tilde{e} = (I - \tilde{H})\tilde{y} \sim \mathcal{N}\left((I-\tilde{H})\tilde{X}\beta, (I-\tilde{H})\sigma^2 I(I-\tilde{H})\right) = \mathcal{N}\left(0, \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2 I_{n-p} \end{pmatrix}\right),$$

and thus the residuals are independent in this case (indeed, the first p are all 0 and the last $n-p$ are all i.i.d. Gaussians). Notice that, usually, the residuals are not independent!

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 $\operatorname{var}(\varepsilon) = \sigma^2 I_n$ (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 $\hat{\beta}_1$?

Solution. a) The model can be written as

$$y = [X]\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \varepsilon, \quad \text{with } X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

The necessary and sufficient condition for the matrix X to be full rank is that the x_i's are not all the same.

b) From the model assumption we know that $\operatorname{var}(y) = \sigma^2 I_n$. We recall that $\hat{\beta}$ is a linear transformation of y, i.e. $\hat{\beta} = Ay$ with $A = (X^\top X)^{-1}X^\top$. Thus (recall that $X^\top X$ is symmetric, so $(X^\top X)^\top = X^\top X$)

$$\begin{aligned} \operatorname{var}(\hat{\beta}) &= A \operatorname{var}(y) A^\top \\ &= \sigma^2 (X^\top X)^{-1} X^\top \left[(X^\top X)^{-1} X^\top\right]^\top \\ &= \sigma^2 (X^\top X)^{-1} X^\top X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} \end{aligned}$$

c) The variance of $\hat{\beta}_1$ is the second diagonal element of the variance matrix of $\hat{\beta}$ which is

$$\operatorname{var}(\hat{\beta}_1) = [\operatorname{var}(\hat{\beta}_1)]_{22} = \sigma^2 [(X^\top X)^{-1}]_{22} = \frac{\sigma^2}{\det(X^\top X)} [X^\top X]_{11}.$$

From the form of X we have

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \implies X^\top X = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}.$$

So, the determinant of $X^\top X$ as a function of $x = (x_1, \ldots, x_n)^\top$ is

$$f(x) = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2.$$

Summarizing, we have the dependence of $\mathrm{var}(\hat{\beta}_1)$ as a function of x is

$$\mathrm{var}(\hat{\beta}_1)(x) = \frac{\sigma^2 n}{f(x)},$$

and finding the $\arg\min \mathrm{var}(\hat{\beta}_1)(x)$ is equivalent to finding the $\arg\max f(x)$. Being convex, $f(x)$ attains its minimum if $x_i = c$ for any $i = 1, \ldots, n$ and its maximum is attained on the boundary of the domain $[-1, 1]^n$, more specifically for $x_i \in \{-1, 1\}$. (Double differentiation shows that the Hessian of f is $\mathbf{H}_n = 2(n\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\top)$ for all x, with $\mathbf{I}_n$ the n by n identity matrix and $\mathbf{1}_n = (1, \ldots, 1)^\top$. The null space of $\mathbf{H}_n$ is $\mathrm{span}\{\mathbf{1}_n\}$, its first $(n-1)$ eigenvalues are $2n$ and the last is 0. Thus, $\mathbf{H}_n$ is always positive semi-definite.) As a consequence, $\sum_{i=1}^n x_i^2 = n$ and $\sum_{i=1}^n x_i = n_+ - n_-$, where n_+ is the number of x_i's attaining the value $+1$ and n_- is the number of x_i's attaining the value -1.
When n is even the optimal value can be attained for $n_+ = n_- = n/2$, so $f(x) = n^2$ and $\mathrm{var}(\hat{\beta}_1) = \sigma^2/n$.
When n is odd we have a sub-optimal case and the maximum value is attained for $n_+ - 1 = n_- = (n-1)/2$ (or alternatively $n_+ = n_- - 1 = (n-1)/2$), so $f(x) = n^2 - 1$ and $\mathrm{var}(\hat{\beta}_1) = \sigma^2 n/(n^2-1)$.

We can interpret the result in the following way: $\hat{\beta}_1$ is the slope of the line that best fits the data according to the linear regression. If all values of x_i are close to a single value (say 0) there will be "many" acceptably good linear fitting of the data and the slope can take values in a large set of values. Alternatively, small changes in the values of the y_i's can lead to large changes in the slope of the fitting line. On the contrary, if the x_i's are as spread as possible then even large changes in the values of the y_i's will have little effect on the value of the slope of the fitting line.

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

Solution. We have

$$\begin{aligned}
\mathrm{MSE}(c^\top\tilde{\beta}) &= \underbrace{\mathrm{bias}(c^\top\tilde{\beta})^2}_{=0,\text{ as }\tilde{\beta}\text{ unbiased}} + \mathrm{var}(c^\top\tilde{\beta}) = c^\top\mathrm{var}(\tilde{\beta})c, \\
\mathrm{MSE}(c^\top\hat{\beta}) &= \underbrace{\mathrm{bias}(c^\top\hat{\beta})^2}_{=0,\text{ as }\hat{\beta}\text{ unbiased}} + \mathrm{var}(c^\top\hat{\beta}) = c^\top\mathrm{var}(\hat{\beta})c.
\end{aligned}$$

So

$$\mathrm{MSE}(c^\top\tilde{\beta}) - \mathrm{MSE}(c^\top\hat{\beta}) = c^\top\mathrm{var}(\tilde{\beta})c - c^\top\mathrm{var}(\hat{\beta})c = c^\top(\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta}))c.$$

Hence, we have

$$\begin{aligned}
& \mathrm{MSE}(c^\top\tilde{\beta}) \geq \mathrm{MSE}(c^\top\hat{\beta}), \quad \forall c \in \mathbb{R}^p \\
\Leftrightarrow\quad & \mathrm{MSE}(c^\top\tilde{\beta}) - \mathrm{MSE}(c^\top\hat{\beta}) \geq 0, \quad \forall c \in \mathbb{R}^p \\
\Leftrightarrow\quad & c^\top(\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta}))c \geq 0, \quad \forall c \in \mathbb{R}^p \\
\Leftrightarrow\quad & \mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta}) \succeq 0.
\end{aligned}$$

Problem 21. (Diagnostic graphics)

a) Figure 2 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 3 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 3.

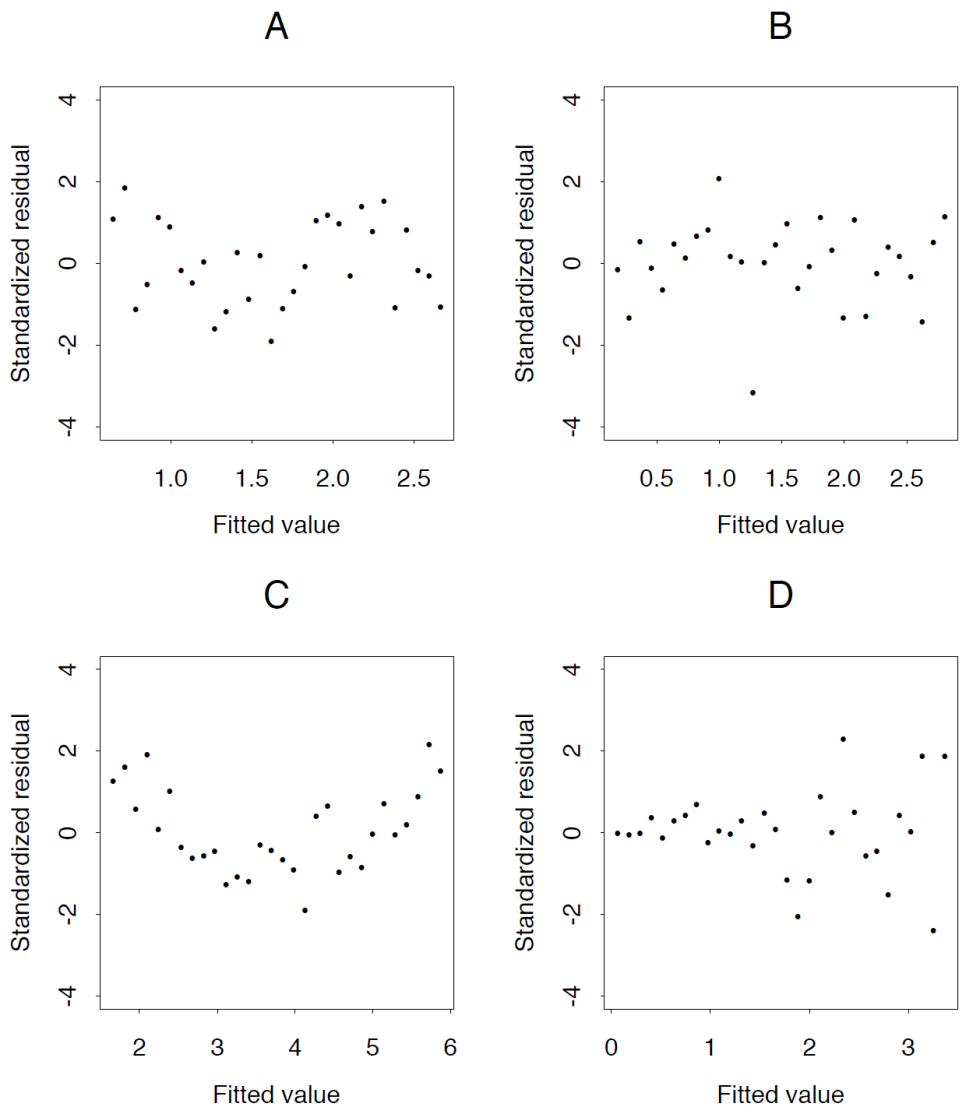

Figure 2: Standardised residuals as a function of values adjusted for four Gaussian models.

Solution. a) We know that $\text{cov}(e, \hat{y}) = 0$ and that the standardised residuals are standard Gaussian random variables (i.e. around 95% of the residuals must take values between -2 and 2 independently from the values of $\hat{y}_j$) if model assumptions are fulfilled ($\epsilon \sim N(0, \sigma^2 I)$, ...).

- Plot A : OK.
- Plot B : Problem = An outlier.
- Plot C : Problem = dependence between the fitted values and the standardised residuals. (see Example 8.24, page 390, Statistical Models, Davison).

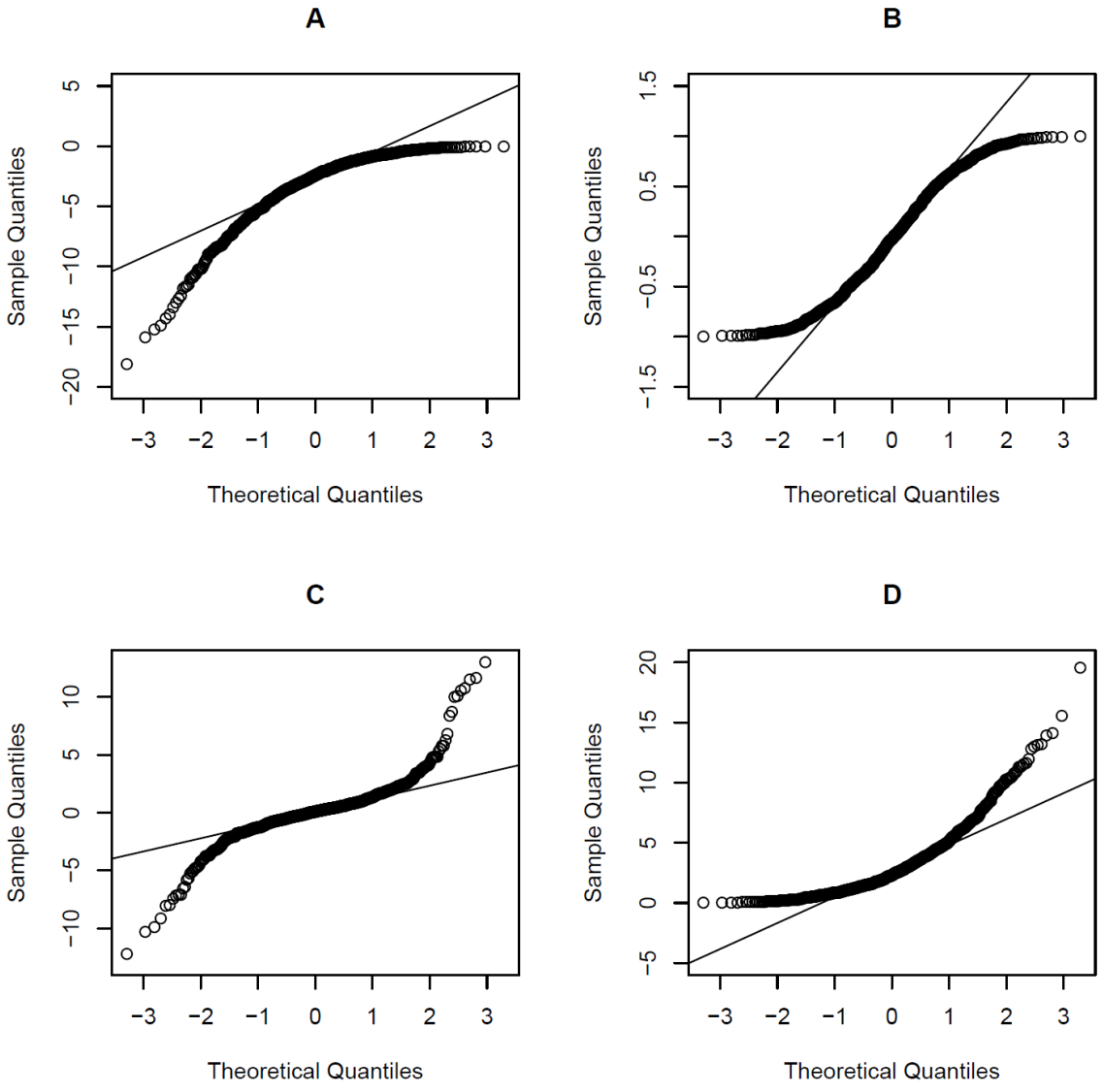

Figure 3: Four Q-Q Gaussian plots where the data do not follow a Gaussian law.

- Plot D : Problem = The variance of the residuals is not constant, heteroscedasticity.

b)
- Plot A: negative skewness coefficient.
- Plot B: tails lighter than Gaussian tails.
- Plot C: tails heavier than Gaussian tails.
- Plot D: positive skewness coefficient.

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 $[0,1]$.
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 $(v_1, \dots, v_n) \in A$, use the symmetry of the problem to write
   $$\mathbb{P}\left(V_{(1)} \le v_1, \dots, V_{(n)} \le v_n\right)$$
   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
   $$\frac{\partial^n}{\partial v_1 \dots \partial v_n} \mathbb{P}\left(V_{(1)} \le v_1, \dots, V_{(n)} \le v_n\right).$$
   Then, the marginal density of $V_{(k)}$ is found by integrating the joint density over all other variables.
3. Let $V_1, \dots, V_n \sim U([0,1])$ be i.i.d., and let
   $$V_{(1)} \le V_{(2)} \le \cdots \le V_{(n)}$$
   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.

Solution. 1. $x \mapsto \Phi(x)$ is strictly increasing, and $\Phi(\mathbb{R}) = (0,1)$. So, $\Phi(X_i) \in (0,1)$. If $x \in (0,1)$, then
$$\Pr[\Phi(X_i) \le x] = \Pr[X_i \le \Phi^{-1}(x)] = \Phi(\Phi^{-1}(x)) = x,$$
hence $\Phi(X_i) \sim U([0,1])$.

2. We start by computing $\mathbb{P}\left(V_{(1)} \le v_1, \dots, V_{(n)} \le v_n\right)$ under the assumption that $(v_1, \dots, v_n) \in A$. We first observe that
$$\{V_{(1)} \le v_1, \dots, V_{(n)} \le v_n\} \iff \bigcup_{\pi \in \Pi} \{V_{\pi(1)} \le v_1, \dots, V_{\pi(n)} \le v_n\},$$
where Π is the set of all possible permutations of $\{1, \dots, n\}$ (if this does not convince you, think about the simplest case with $n = 2$). Since the events in the union are disjoint and there are $n!$ possible permutations, we conclude that
$$\mathbb{P}\left(V_{(1)} \le v_1, \dots, V_{(n)} \le v_n\right) = \sum_{\pi \in \Pi} \mathbb{P}\left(V_{\pi(1)} \le v_1, \dots, V_{\pi(n)} \le v_n\right) \overset{i.i.d.}{=} n! \prod_{i=1}^{n} v_i.$$

By formula

$$f_{V_{(1)},\dots,V_{(n)}}(v_1,\dots,v_n) = \frac{\partial^n}{\partial v_1 \dots \partial v_n}\mathbb{P}\left(V_{(1)} \le v_1, \dots, V_{(n)} \le v_n\right) = n! I_A,$$

we derive that the joint probability density function is constant on A ($|A| = 1/(n!)$) and vanishes outside A. Finally, in order to calculate the density function for the k-th order statistics we integrate over all other variables:

$$\begin{aligned} f_{V_{(k)}}(v_k) &= \int_{[0,1]^n} n! I_A dv_1 \dots dv_n \\ &= n! \int_0^{v_k}\int_{v_1}^{v_k}\cdots\int_{v_{k-2}}^{v_k} dv_{k-1}\dots dv_2\, dv_1 \int_{v_k}^{v_{k+1}}\cdots\int_{v_k}^{v_n}\int_{v_k}^{1} dv_n \dots dv_{k+2}\, dv_{k+1} \\ &= \frac{n!}{(k-1)!(n-k)!} v_k^{k-1}(1-v_k)^{n-k} = n\binom{n-1}{k-1} v_k^{k-1}(1-v_k)^{n-k} \end{aligned} \quad (3)$$

3. Here $f_k(x) = \frac{n!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k}$, for $x \in [0,1]$, and zero otherwise. Since f_k is a density, $1 = \int f_k(x)dx$, (here and then, $\int = \int_0^1$) and thus

$$\int x^{k-1}(1-x)^{n-k}dx = \frac{(k-1)!(n-k)!}{n!},$$

or, more explicitly:

$$\int x^a(1-x)^b dx = \frac{a!b!}{(a+b+1)!}, \quad a, b \in \mathbb{N}.$$

Hence

$$\mathbb{E}[V_{(k)}] = \int x f_k(x)dx = \frac{n!}{(k-1)!(n-k)!}\int x^k(1-x)^{n-k}dx = \dots = k/(n+1).$$

4. $\mathbb{E}[X_{(k)}] = \mathbb{E}[\Phi^{-1}(\Phi(X_{(k)}))] \approx \Phi^{-1}(\mathbb{E}[\Phi(X_{(k)})]) = \Phi^{-1}(k/(n+1)) = z_{k/(n+1)}$. Thus, when $X_1, \dots, X_n$ are $N(0,1)$ i.i.d, we expect that their QQ normal plot is, "on average", on the line $y = x$.

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

$$\begin{aligned} \mathbb{E}[y_j] &= \beta_0, \quad j = 1, \dots, n-2, \\ \mathbb{E}[y_j] &= \beta_0 + \beta_1, \quad j = n-1, n. \end{aligned}$$

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.

Solution. a) From the fact that $\mathbb{E}[y] = X\beta$ we conclude that

$$X = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \implies X^\top X = \begin{pmatrix} n & 2 \\ 2 & 2 \end{pmatrix} \implies (X^\top X)^{-1} = \frac{1}{2n-4}\begin{pmatrix} 2 & -2 \\ -2 & n \end{pmatrix}.$$

The least squares estimator $\hat{\beta}$ is computed as

$$\hat{\beta} = (X^\top X)^{-1}X^\top y = \frac{1}{2n-4}\begin{pmatrix} 2 & -2 \\ -2 & n \end{pmatrix}\begin{pmatrix} (n-2)\tilde{y_1} + 2\tilde{y_2} \\ 2\tilde{y_2} \end{pmatrix} = \begin{pmatrix} \tilde{y_1} \\ -\tilde{y_1} + \tilde{y_2} \end{pmatrix}.$$

b) Note that the column space $\mathscr{S}(X)$ of the design matrix X can be spanned by two orthogonal vectors

$$v_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

Therefore, the hat matrix $H = X(X^\top X)^{-1}X^\top$ can be decomposed to the sum of two projection matrices $H = H_1 + H_2$, where

$$H_1 = v_1(v_1^\top v_1)^{-1}v_1^\top = \frac{1}{n-2}\begin{pmatrix} 1 & \dots & 1 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & \dots & 1 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 \end{pmatrix},$$

$$H_2 = v_2(v_2^\top v_2)^{-1}v_2^\top = \frac{1}{2}\begin{pmatrix} 0 & \dots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 0 \\ 0 & \dots & 0 & 1 & 1 \\ 0 & \dots & 0 & 1 & 1 \end{pmatrix},$$

and the trace of H is $\operatorname{tr}(H) = \operatorname{tr}(H_1) + \operatorname{tr}(H_2) = 2 = p$. The fitted values are calculated as

$$\hat{y} = Hy = H_1y + H_2y = \begin{pmatrix} \tilde{y_1} \\ \vdots \\ \tilde{y_1} \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{y_2} \\ \tilde{y_2} \end{pmatrix} = \begin{pmatrix} \tilde{y_1} \\ \vdots \\ \tilde{y_1} \\ \tilde{y_2} \\ \tilde{y_2} \end{pmatrix}.$$

c) The leverages are the diagonal values of the hat matrix. So,

$$h_{jj} = \begin{cases} 1/(n-2) & \text{for } j = 1, \dots, n-2, \\ 1/2 & \text{for } j = n-1, n. \end{cases}$$

Using the rule of thumb on slide 138 ($h_{jj} > 2p/n$), x_{n-1} and x_n are leverage points. Next, we proceed to the standardised residuals. Since $y_{n-1} = y_n = \tilde{y_2}$, i.e. $e_{n-1} = e_n = 0$, the estimator for the variance σ^2 is

$$s^2 = \frac{1}{n-2}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 = \frac{1}{n-2}\sum_{j=1}^{n-2}(y_j - \tilde{y_1})^2,$$

the standardised residuals are given by

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} = \begin{cases} \sqrt{\frac{n-2}{n-3}}(y_i - \tilde{y_1})/s & \text{for } i = 1, \dots, n-2, \\ 0 & \text{for } i = n-1, n. \end{cases}$$

Cook's statistics are

$$C_i = \frac{r_i^2 h_{ii}}{p(1-h_{ii})} = \frac{e_i^2}{ps^2}\frac{h_{ii}}{(1-h_{ii})^2} = \begin{cases} (n-2)(y_i - \tilde{y_1})^2/(2(n-3)^2 s^2) & \text{for } i = 1, \dots, n-2, \\ 0 & \text{for } i = n-1, n. \end{cases}$$

Therefore, even though $(x_{n-1}, y_{n-1}), (x_n, y_n)$ are leverage points, they have no influence on any other data as their Cook's statistics are zero. Why? The fitted values $\hat{y}_1 = \dots = \hat{y}_{n-2} = \tilde{y_1}$ are totally independent on the values of y_{n-1} and y_n.

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

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

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

Solution. Firstly, recall that t_ν-distribution arises as

$$\frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_\nu}}\sqrt{\nu},$$

where the two random variables in the previous symbolic expression are independent.

Secondly, by the lemma on slide 105

$$\operatorname{var}(\widehat{\beta}_j) = \sigma^2 v_{jj}\,,$$

where $v_{jj} = (X^\top X)^{-1}_{jj}$, and by the theorem on slide 81 we have $\frac{S^2}{\sigma^2}(n-p) \sim \chi^2_{n-p}$.

Combining the two facts, it is natural to estimate $\widehat{\sigma}^2 = S^2$ to get the estimator of the standard error. Note that by the theorem on slide 81 we also have the independence needed. Hence

$$t = \underbrace{\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{v_{jj}}}}_{\sim\mathcal{N}(0,1)} \quad \underbrace{\frac{\sigma}{S\sqrt{n-p}}}_{\frac{1}{\sqrt{(n-p)\frac{S^2}{\sigma^2}}} \sim \frac{1}{\sqrt{\chi^2_{n-p}}}} \qquad \sqrt{n-p} = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-p}}}\sqrt{n-p} \sim t_{n-p}\,.$$

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 $\widehat{\operatorname{corr}}(\hat{\beta}_2, \hat{\beta}_3) = -0.08911$, 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?

Solution. a) The column "`t value`" gives the statistics t for the hypothesis $\beta_i = 0$ defined by

$$T_i = \frac{\hat{\beta}_i}{\sqrt{S^2 v_{ii}}} = \frac{\hat{\beta}_i}{\widehat{\operatorname{SE}}(\hat{\beta}_i)},$$

where v_{ii} is the i-th diagonal element of the matrix $V = (X^\top X)^{-1}$. When the hypothesis $\beta_i = 0$ is true, we have that T_i follows Student's t-distribution with $n-p$ degrees of freedom. We will reject the null hypothesis $\beta_i = 0$ when the value of $|T_i|$ is large.

The column "`Pr(>|t|)`" gives the p-values for the bilateral tests t above. When we denote the observed value of T_i by τ_i, the p-value for the$i$-th test is given by

$$p_i = P(|T_i| > |\tau_i|) = 2(1 - t_{n-p}(|\tau_i|)) = 2t_{n-p}(-|\tau_i|)).$$

If $p_i < 0.05$, we reject the i-th hypothesis with a significance threshold of 5%.

For this example, with a significance threshold of 5%, we can reject the hypothesis $\beta_i = 0$ for $i = 0, 1, 2$, but not for $i = 3$.

b) In this case, the statistics t is given by

$$T = \frac{c^\top \hat{\beta}}{\sqrt{S^2 c^\top (X^\top X)^{-1} c}}$$

for $c = [0, 0, 1, -1]^\top$. We know that

$$\begin{aligned} S^2 c^\top (X^\top X)^{-1} c &= \left(\widehat{\operatorname{SE}}(\hat{\beta}_2)\right)^2 + \left(\widehat{\operatorname{SE}}(\hat{\beta}_3)\right)^2 - 2\,\widehat{\operatorname{corr}}(\hat{\beta}_2, \hat{\beta}_3)\,\widehat{\operatorname{SE}}(\hat{\beta}_2)\,\widehat{\operatorname{SE}}(\hat{\beta}_3) \\ &= 0.04423^2 + 0.18471^2 - 2 \cdot (-0.08911) \cdot 0.04423 \cdot 0.18471 = 0.03753. \end{aligned}$$

Hence

$$\tau = \frac{0.65691 - 0.25002}{\sqrt{0.03753}} = 2.10033$$

and we find the p-value

$$p = 2 \cdot t_9(-2.10033) = 0.06508.$$

Thus, we do not reject the null hypothesis with a significance threshold of 5%.

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 $\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2$.

Solution. We need to show that $\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$ is an orthogonal projection matrix, i.e., it is idempotent, symmetric and it spans $\mathscr{S}(\mathbf{X})$. Note that $\mathbf{X}_1^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2 = \mathbf{O}$, so $\mathbf{H}_{\mathbf{X}_1}\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2} = \mathbf{O}$ also. Since both $\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$ and $\mathbf{H}_{\mathbf{X}_1}$ are orthogonal projection matrices, the first two statements are obvious.

It remains to show that any vector $\mathbf{z} \in \mathscr{S}(\mathbf{X})$ is invariant under the action of $\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$ and that any vector orthogonal to this span is annihilated by $\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}$. Since $\mathbf{X}$ is full rank, we can write $\mathbf{z} = \mathbf{X}\gamma = \mathbf{X}_1\gamma_1 + \mathbf{X}_2\gamma_2$ for some vector γ and subvectors γ_1 and γ_2. Then

$$\begin{aligned}
(\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2})\mathbf{z} &= (\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2})(\mathbf{X}_1\gamma_1 + \mathbf{X}_2\gamma_2) \\
&= \mathbf{H}_{\mathbf{X}_1}(\mathbf{X}_1\gamma_1 + \mathbf{X}_2\gamma_2) + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}(\mathbf{X}_1\gamma_1 + \mathbf{X}_2\gamma_2) \\
&= \mathbf{X}_1\gamma_1 + \mathbf{H}_{\mathbf{X}_1}\mathbf{X}_2\gamma_2 + \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2\gamma_2 \\
&= \mathbf{X}_1\gamma_1 + \mathbf{X}_2\gamma_2
\end{aligned}$$

upon noting that

$$\begin{aligned}
\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}\mathbf{X}_1 &= \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2(\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_1 = \mathbf{O}, \\
\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}\mathbf{X}_2 &= \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2(\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2 = \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2.
\end{aligned}$$

Take now $\mathbf{w} \in \mathscr{S}^\perp(\mathbf{X})$. We have

$$\begin{aligned}
(\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2})\mathbf{w} &= \mathbf{H}_{\mathbf{X}_1}\mathbf{w} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}\mathbf{w} \\
&= 0 + \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2(\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{w} \\
&= \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2(\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top (\mathbf{I} - \mathbf{H}_{\mathbf{X}_1})\mathbf{w} = \mathbf{0}.
\end{aligned}$$

Indeed, $\mathbf{H}_{\mathbf{X}_1}\mathbf{w} = \mathbf{0}$ because $\mathbf{w}$ is orthogonal to $\mathbf{X}$, thus also orthogonal to $\mathbf{X}_1$. At the same time, $\mathbf{X}_2^\top \mathbf{w} = \mathbf{0}$ by orthogonality. By uniqueness of projection matrices (Exercise 1.2), the result follows.

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).

Solution. The coefficient estimates of (E2) is

$$\widetilde{\beta}_2 = (\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}y. \tag{S1}$$

Let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ denote the OLS estimates from running regression (E1). The orthogonal decomposition of y gives

$$y = \mathbf{X}_1\widehat{\beta}_1 + \mathbf{X}_2\widehat{\beta}_2 + \mathbf{M}_{\mathbf{X}}y. \tag{S2}$$

Premultiplying both sides of (S2) by $\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1}$ yields

$$\mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1} y = \mathbf{X}_2^\top \mathbf{M}_{\mathbf{X}_1} \mathbf{X}_2 \widehat{\beta}_2 \tag{S3}$$

since $\mathbf{M}_{\mathbf{X}} \mathbf{M}_{\mathbf{X}_1} \mathbf{X}_2 = \mathbf{M}_{\mathbf{X}} \mathbf{X}_2 = \mathbf{O}$. Solving (S3) gives back (S1), showing that $\widehat{\beta}_2 = \widetilde{\beta}_2$.

By premultiplying (S2) by $\mathbf{M}_{\mathbf{X}_1}$, we obtain instead

$$\mathbf{M}_{\mathbf{X}_1} y = \mathbf{M}_{\mathbf{X}_1} \mathbf{X}_2 \widehat{\beta}_2 + \mathbf{M}_{\mathbf{X}} y \tag{S4}$$

since $\mathbf{M}_{\mathbf{X}_1} \mathbf{M}_{\mathbf{X}} = \mathbf{M}_{\mathbf{X}}$. The regressand in (S4) is the same as that of regression (E2). The first term, $\mathbf{M}_{\mathbf{X}_1} \mathbf{X}_2 \widehat{\beta}_2$, must be the fitted value since $\widehat{\beta}_2$ is the OLS estimate of β_2. Thus, $\mathbf{M}_{\mathbf{X}} y$ must be the vector of residuals of (E2).

Deriving the expression for $\widehat{\beta}_2$ in the presence of multiple regressors involves tedious calculations with partitioned matrices. Use Frisch–Waugh-Lovell theorem when you have multiple regressors, but are only interested in a sub-vector of coefficient estimates such as $\widehat{\beta}_2$.

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 $\mathcal{F}(1, n-p)$.

Solution. By the FWL theorem, we can write the arguments of W as

$$\hat{\beta}_2 = (\mathbf{x}_2^\top \mathbf{M}_{\mathbf{X}_1} \mathbf{x}_2)^{-1} \mathbf{x}_2^\top \mathbf{M}_{\mathbf{X}_1} y, \qquad \mathrm{se}(\hat{\beta}_2) = \left[s^2 (\mathbf{x}_2^\top \mathbf{M}_{\mathbf{X}_1} \mathbf{x}_2)^{-1} \right]^{1/2}.$$

Clearly, $\mathrm{RSS}/(n-p) = s^2$ and thus it remains only to show that the numerator of F is

$$\mathrm{RSS}_0 - \mathrm{RSS} = \left[(\mathbf{x}_2^\top \mathbf{M}_{\mathbf{X}_1} \mathbf{x}_2)^{-1/2} \mathbf{x}_2^\top \mathbf{M}_{\mathbf{X}_1} y \right]^2 = y^\top \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1} \mathbf{x}_2} y.$$

First, we have

$$\mathrm{RSS}_0 - \mathrm{RSS} = \|\mathbf{M}_{\mathbf{X}_1} y\|^2 - \|\mathbf{M}_{\mathbf{X}} y\|^2 = \|(\mathbf{M}_{\mathbf{X}_1} - \mathbf{M}_{\mathbf{X}}) y\|^2.$$

Using an orthogonal decomposition, this expression can be further simplified to

$$\mathrm{RSS}_0 - \mathrm{RSS} = \|\mathbf{M}_{\mathbf{X}_1} \mathbf{H}_{\mathbf{X}} y\|^2 = \|\mathbf{M}_{\mathbf{X}_1} (\mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1} \mathbf{x}_2}) y\|^2 = \|\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1} \mathbf{x}_2} y\|^2$$

because $\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1} \mathbf{x}_2} \in \mathscr{M}^\perp(\mathbf{X}_1)$. Noting that $\|\mathbf{H}_{\mathbf{M}_{\mathbf{X}_1} \mathbf{x}_2} y\|^2 = y^\top \mathbf{H}_{\mathbf{M}_{\mathbf{X}_1} \mathbf{x}_2} y$, completes the proof.

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.

Solution. Since the ordering of the variable is different, the number in the table will be different from the ones in slide 168. Namely:

	Df	Red Sum Sq	F value	p-value
x_4	1	2715.8-883.9 = 1831.9	306.3	10^-7
x_3	1	883.9 -175.7 = 708.2	118.4	10^-6
x_2	1	175.7 -73.8 = 101.9	17.04	0.003
x_1	1	73.8 -47.9 = 26	4.3	0.07
Residual	8	47.9		

For calculation of the F-values one calculates the numerator as the correspondent reduction in the sum of squares (third column of the table) divided by the degrees of freedom added (in this case only 1) and the denominator as the residual sum of squares divided by the residual degrees of freedom ($47.9/8 = 5.98$). Notice that the denominator stays always the same (which is quite irrelevant now in the computer era :). To calculate the p-values, use R.

Adding variables in this (reverse) order would lead to a model with x_2, x_3 and x_4, while in the slide 164 variables x_1 and x_2 would be in the model instead.

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

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

Solution. 1. This means that all columns of X_1 are orthogonal to all columns of X_2. I.e., $\mathcal{M}(X_1) \perp \mathcal{M}(X_2)$.

2. We notice first that

   $$X^\top X = \begin{pmatrix} X_1^\top X_1 & 0 \\ 0 & X_2^\top X_2 \end{pmatrix},$$

   so

   $$\begin{aligned} H &= (X_1, X_2)\begin{pmatrix} (X_1^\top X_1)^{-1} & 0 \\ 0 & (X_2^\top X_2)^{-1} \end{pmatrix}(X_1, X_2)^\top \\ &= X_1(X_1^\top X_1)^{-1}X_1^\top + X_2(X_2^\top X_2)^{-1}X_2^\top = H_1 + H_2. \end{aligned}$$

   Then. since $X_1^\top X_2 = 0$, we have $H_1H_2 = 0$. Thus, $H_2H_1 = H_2^\top H_1^\top = (H_1H_2)^\top = 0$,

   $$HH_1 = (H_1 + H_2)H_1 = H_1^2 = H_1$$

et $H_1H = H_1^\top H^\top = (HH_1)^\top = H_1^\top = H_1$.

Interpretation: $H_1H_2 = 0$ comes from the fact that the space of columns of X_1 and X_2 are orthogonal, thus if we project them on $\mathcal{M}(X_2)$ and then on $\mathcal{M}(X_1)$, we obtain the null vector. The interpretation for $H_2H_1 = 0$ is similar. $HH_1 = H_1$ comes from the fact that to project on $\mathcal{M}(X_1)$ and then on $\mathcal{M}(X)$ is equivalent to projecting only on $\mathcal{M}(X_1)$, because $\mathcal{M}(X_1)$ is a subspace of $\mathcal{M}(X)$. For the same reason, $H_1H = H_1$ because we project on $\mathcal{M}(X)$ and then on $\mathcal{M}(X_1)$, so it is the same as projecting on $\mathcal{M}(X_1)$. Intuitively, We notice that if $X_1^\top X_2 \neq 0$, We have $HH_1 = H_1 = H_1H$, but $H_1H_2 \neq 0$ and $H_2H_1 \neq 0$.

3. Using the fact that $Hy = (H_1 + H_2)y$,
   (a) trivial;
   (b) comes from $H_2H_1 = 0$;
   (c) comes from $H(I - H_1) = H - H_1 = H_2$.
4. The fitted values under (M) (with (y, X) as data) are equal to
   (a) the sum of the fitted values under (M_1) and (M_2). (the model data (M_i) are (y, X_i))
   (b) the sum of the fitted values under (M_1) (given (y, X_1)) and of residuals of (M_1) fitted under (M_2) (the data are (e_1, X_2)).
   (c) the sum of the fitted values under (M_1) (given (y, X_1)) and of residuals of (M_1) fitted under (M) (the data are (e_1, X)).

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?

Solution. 1. since

$$(X^\top X)^{-1} = \begin{pmatrix}(X_1^\top X_1)^{-1} & 0 & \dots & 0\\ 0 & (X_2^\top X_2)^{-1} & 0 & \vdots\\ \vdots & 0 & \ddots & 0\\ 0 & \dots & 0 & (X_k^\top X_k)^{-1}\end{pmatrix}$$

and

$$(X_L^\top X_L)^{-1} = \operatorname{diag}((X_i^\top X_i)^{-1} : i \in L).$$

Son

$$H = X_1(X_1^\top X_1)^{-1}X_1^\top + \cdots + X_k(X_k^\top X_k)^{-1}X_k^\top = H_1 + \cdots + H_k$$

and

$$H_L = \sum_{i\in L} X_i(X_i^\top X_i)^{-1}X_i^\top = \sum_{i\in L} H_i.$$

2. If $i = j$, $H_iH_j = H_i^2 = H_i$ and if $i \neq j$, $H_iH_j = X_i(X_i^\top X_i)^{-1}X_i^\top X_j(X_j^\top X_j)^{-1}X_j^\top = 0$ because $X_i^\top X_j = 0$.

3.

$$\hat\beta = (X^\top X)^{-1}X^\top y = \begin{pmatrix} (X_1^\top X_1)^{-1} & 0 & \dots & 0 \\ 0 & (X_2^\top X_2)^{-1} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & (X_k^\top X_k)^{-1} \end{pmatrix} \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_k^\top \end{pmatrix} y = \begin{pmatrix} (X_1^\top X_1)^{-1}X_1^\top y \\ (X_2^\top X_2)^{-1}X_2^\top y \\ \vdots \\ (X_k^\top X_k)^{-1}X_k^\top y \end{pmatrix}.$$

4. First of all, we notice that

$$e_L := y - H_L y = y - \sum_{i\in L} H_i y$$

and

$$e_{L\cup\{j\}} := y - H_{L\cup\{j\}}y = y - \sum_{i\in L\cup\{j\}} H_i y.$$

Moreover,

$$\begin{aligned} (I - H_{L\cup\{j\}})e_L &= (I - H_{L\cup\{j\}})(I - H_L)y && (4)\\ &= (I - H_L - H_{L\cup\{j\}} + H_{L\cup\{j\}}H_L)y && (5)\\ &= (I - H_{L\cup\{j\}})y && (6)\\ &= e_{L\cup\{j\}}. && (7) \end{aligned}$$

Then $e_{L\cup\{j\}}$ is a orthogonal projection of e_L, so $e_L - e_{L\cup\{j\}} \perp e_{L\cup\{j\}}$ and

$$\|e_{L\cup\{j\}}\|^2 + \|e_L - e_{L\cup\{j\}}\|^2 = \|e_L\|^2.$$

So

$$RSS_L - RSS_{L\cup\{j\}} = \|e_L\|^2 - \|e_{L\cup\{j\}}\|^2 = \left\|e_L - e_{L\cup\{j\}}\right\|^2 = \|H_j y\|^2$$

is independent of L.

5. The interpretation with respect to ANOVA is that in this case, the addition of a variable X_j does not depend on the variables that we already have in the model (this is not the general case).

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?

Solution. 1. Here we will use the following test to add or not the j-th variable to the model $y = \beta_0 + \sum_{i\in L}\beta_i x_i$:

$$F = \frac{RSS(\hat\beta_L) - RSS(\hat\beta_{L\cup\{j\}})}{RSS(\hat\beta_{\text{full}})/(13-5)},$$

where $\hat\beta_{\text{full}}$ represents the estimator of β for the complete model. Since $RSS(\hat\beta_L) - RSS(\hat\beta_{L\cup\{j\}}) \sim \sigma^2\chi^2_1$ under the hypothesis $H_0 : \beta_j = 0$, and that $RSS(\hat\beta_{\text{full}}) \sim \sigma^2\chi^2_{n-p}$ and it is independent of $RSS(\hat\beta_L) - RSS(\hat\beta_{L\cup\{j\}})$, $F \sim F_{1,8}$ under H_0. In particular, the distribution of F does not depend on the size of L. The critical value of this test at level 5% is 5.32.

Forward selection

- Initial model : $y = \beta_0 + \epsilon$
- Stage 1 : $y = \beta_0 + \beta_4 x_4 + \epsilon$, $F = \frac{2715.8-883.9}{47.9/(13-5)} = 305.95 > 5.32$.
- Stage 2 : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$, $F = 135.13 > 5.32$.
- Stage 3: $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $F = 4.47 < 5.32$.

Final model : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$.

Backward selection

- Initial model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$
- Stage 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$, $F = \frac{48-47.9}{47.9/(13-5)} = 0.0167 < 5.32$.
- Stage 2 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $F = 1.65 < 5.32$.
- Stage 3 : $y = \beta_0 + \beta_2 x_2 + \epsilon$, $F = 141.70 > 5.32$.

Final model : $y = \beta_0 + \beta_2 x_2 + \beta_1 x_1 + \epsilon$.

2. (a) Mallow's C_p work as AIC: we choose the model with the minimal C_p. Here's the table with all the C_p:

Model	RSS	C_p	Model	RSS	C_p	Model	RSS	C_p
- - - -	2715.8	442.58	1 2 - -	57.9	2.67	1 2 3 -	48.1	3.03
1 - - -	1265.7	202.39	1 - 3 -	1227.1	197.94	1 2 - 4	48.0	3.02
- 2 - -	906.3	142.37	1 - - 4	74.8	5.49	1 - 3 4	50.8	3.48
- - 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

(b) With forward selection, we choose $y = \beta_0 + \sum_{i\in\{1,2,4\}}\beta_i x_i$, while the backward selection gives the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. This last model is the one with the smallest C_p among all others.

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

Solution. For the Gaussian linear models $y \sim N(X\beta, \sigma^2 I_n)$, the likelihood of (β, σ^2) is given by

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta)\right).$$

Then, the log-likelihood is

$$l(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta).$$

We have that the maximum likelihood estimator of β and σ^2 are

$$\hat{\beta} = (X^\top X)^{-1} X^\top y, \quad \hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^\top (y - X\hat{\beta}).$$

So, the maximum of log-likelihood is

$$l(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\underbrace{(y - X\hat{\beta})^\top (y - X\hat{\beta})}_{=n\hat{\sigma}^2} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2}.$$

From the AIC definition, we obtain that

$$\text{AIC} = -2l(\hat{\beta}, \hat{\sigma}^2) + 2p = n\log(2\pi) + n\log\hat{\sigma}^2 + n + 2p = n\log\hat{\sigma}^2 + 2p + \text{const.}$$

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

$$\text{CV} = \sum_{k=1}^{n}(y_k - x_k^\top \hat{\beta}_{-k})^2. \tag{8}$$

can be written as

$$\text{CV} = \sum_{k=1}^{n}\frac{(y_k - x_k^\top \hat{\beta})^2}{(1 - h_{kk})^2}. \tag{9}$$

What is the advantage of using (9) instead of (8)?

Solution. a) First note that $X_{-k}^\top X_{-k} = X^\top X - x_k x_k^\top$ and $x_k^\top (X^\top X)^{-1} x_k = h_{kk}$. By the Sherman-Morrison formula, we have

$$\begin{aligned}(X_{-k}^\top X_{-k}^\top)^{-1} = (X^\top X - x_k x_k^\top)^{-1} &= (X^\top X)^{-1} + \frac{(X^\top X)^{-1}x_k x_k^\top (X^\top X)^{-1}}{1 - x_k^\top (X^\top X)^{-1}x_k} \\ &= \left(I + \frac{(X^\top X)^{-1}x_k x_k^\top}{1 - h_{kk}}\right)(X^\top X)^{-1}, \\ &= \left(I + \frac{(X^\top X)^{-1}x_k x_k^\top}{1 - h_{kk}}\right)(X^\top X)^{-1},\end{aligned}$$

b) First, we can calculate

$$X^\top y = (x_1, \dots, x_n)y = \sum_{i=1}^{n} y_i x_i = X_{-k}^\top y_{-k} + y_k x_k, \tag{10}$$

hence

$$\begin{aligned} x_k^\top (X^\top X)^{-1} X_{-k}^\top y_{-k} &= x_k^\top (X^\top X)^{-1}(X^\top y - x_k y_k) = \hat{y}_k - h_{kk} y_k = y_k - e_k - h_{kk} y_k \\ &= (1 - h_{kk}) y_k - e_k. \end{aligned} \tag{11}$$

Using (a) together with two equations (10) and (11) above, we get

$$\begin{aligned} \hat{\beta}_{-k} &= (X_{-k}^\top X_{-k})^{-1} X_{-k}^\top y_{-k} \\ &\overset{(a)}{=} \left(I + \frac{(X^\top X)^{-1} x_k x_k^\top}{1 - h_{kk}} \right) (X^\top X)^{-1} X_{-k}^\top y_{-k} \\ &\overset{(10)}{=} (X^\top X)^{-1}(X^\top y - y_k x_k) + (1 - h_{kk})^{-1} (X^\top X)^{-1} x_k x_k^\top (X^\top X)^{-1} X_{-k}^\top y_{-k} \\ &\overset{(11)}{=} \hat{\beta} - (X^\top X)^{-1} x_k y_k + (1 - h_{kk})^{-1} (X^\top X)^{-1} x_k [(1 - h_{kk}) y_k - e_k] \\ &= \hat{\beta} - (1 - h_{kk})^{-1} e_k (X^\top X)^{-1} x_k. \end{aligned}$$

c) From (b), we know

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

If we use formula (8) we have to conduct n regressions to estimate all the β_{-j}, $j = 1, \dots, n$, and then proceed to n adjustments. On the other hand, if we use formula (9) only the adjustment of the model with the complete data is required. This makes it feasible to actually perform "leave-one-out" cross-validation for a linear model.

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 $1 - h_{11}, \dots, 1 - h_{nn}$. 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 $\mathbb{E}[CV] \approx (n + p)\sigma^2$.

Solution. From the previous exercise, we know that the "leave-one-out" cross-validation is given by

$$CV = \sum_{k=1}^{n} \frac{e_k^2}{(1 - h_{kk})^2}.$$

In matrix notation, this is equivalent to

$$\begin{aligned} CV &= e^\top D^{-2} e \\ &= y^\top (I - H) D^{-2} (I - H) y \end{aligned}$$

To calculate its expectation, we make use of the well-known formula $\mathbb{E}[y^\top A y] = \mathbb{E}[y]^\top A \mathbb{E}[y] + \mathrm{tr}(A \cdot \mathrm{cov}[y])$ with $\mathbb{E}[y] = \mu$ and $\mathrm{cov}[y] = \mathrm{cov}[\varepsilon] = \sigma^2 I_n$. To see why the formula holds,

$$\begin{aligned} \mathbb{E}[y^\top A y] &= \mathbb{E}[\mathrm{tr}(y^\top A y)] = \mathbb{E}[\mathrm{tr}(A \cdot y y^\top)] = \mathrm{tr}(A \cdot \mathbb{E}[y y^\top]) \\ &= \mathrm{tr}(A(\mathbb{E}[y]^\top \mathbb{E}[y] + \mathrm{cov}[y])) = \mathbb{E}[y]^\top A \mathbb{E}[y] + \mathrm{tr}(A \cdot \mathrm{cov}[y]). \end{aligned}$$

Applying the formula above, we now have

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

since $I - H$ is a projection matrix (symmetric and idempotent). As the final step of the calculation, we prove

$$\begin{aligned}\operatorname{tr}(D^{-2}(I-H)) &= \sum_{k=1}^{n}(1-h_{kk})^{-2}\cdot(1-h_{kk}) \\ &= \sum_{k=1}^{n}(1-h_{kk})^{-1} = \operatorname{tr}(D^{-1}),\end{aligned}$$

hence the result follows. Note that $\mu \in \mathcal{M}(X)$ only if the **model is correct**, so that $(I-H)\mu = 0$! In the case where the model is correct,

$$\mathbb{E}[CV] = \sigma^2 tr(D^{-1}) = \sigma^2 \sum_{k=1}^{n}(1-h_{kk})^{-1} \approx \sigma^2 \sum_{k=1}^{n}(1+h_{kk}) = \sigma^2(n+\operatorname{tr}(H)) = \sigma^2(n+p),$$

since $0 \leq h_{kk} \leq 1$ are usually small (influential points having high leverages pose issues for the model) and $(1-x)^{-1} \approx 1+x$ for small $x \approx 0$ as the Taylor expansion of $f(x) = (1-x)^{-1}$ at $x = 0$ is $1+x$.

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.

Solution. a)
```
library(RSADBE)
data(SPD)

m1 <- lm(Y ~ ., data = SPD)
m.backward <- step(m1, direction = "backward")

m0 <- lm(Y ~ 1, data = SPD)
my.scope <- formula(SPD)
m.forward <- step(m0, scope = my.scope, direction = "forward", data = SPD)
```

The forward/backward stepwise give the model `Y ~ X1 + X3` with AIC of 118.00.

b)
```
install.packages("leaps")
library(leaps)
library(car)

leaps <- regsubsets(formula(SPD), data = SPD)
plot(leaps)
subsets(leaps)
```

The model with the best BIC value is `Y ~ X1` with BIC of -27.50.

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.

Solution. a) Using the SVD decomposition $\mathbf{Z} = \mathbf{U}\mathbf{\Omega}\mathbf{V}^\top$, we have

$$\mathbf{Z}\hat{\gamma}_\lambda = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I}_{p-1})^{-1}\mathbf{Z}^\top y = \mathbf{U}\mathbf{\Omega}(\mathbf{\Omega}^2 + \lambda\mathbf{I}_{p-1})^{-1}\mathbf{\Omega}\mathbf{U}^\top y = \sum_{j=1}^{p-1} \frac{\omega_j^2}{\omega_j^2 + \lambda}\left(\mathbf{u}_j^\top y\right)\mathbf{u}_j.$$

Note that $\mathbf{Z}\mathbf{Z}^\top = \mathbf{U}\mathbf{\Omega}^2\mathbf{U}^\top$ has the eigenvectors u_j with corresponding eigenvalues ω_j^2. The coefficients associated to the basis vectors $\mathbf{u}_j$ with the smallest eigenvalue ω_j^2 get shrunk the most towards zero.

b) First, note that the OLS corresponds to the case where $\lambda = 0$, and $\hat{\gamma}_{\text{OLS}} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top y = \mathbf{Z}^\top y$. Also, we know that the fitted values with the OLS method is invariant to the scale of the variables: a rescaling of a column, say $z_j \mapsto kz_j$ leads to the solution $\hat{\gamma}_j \mapsto \hat{\gamma}_j/k$, so that the fitted values $\hat{y} = \mathbf{1}_n\overline{y} + \mathbf{Z}\hat{\gamma}_{\text{OLS}}$ remains unchanged (To rephrase it, we have not changed the column space of the design matrix). However, this is **not** the case for penalized methods: rescaling amounts to the different amount of shrinkage to the covariates, which can be seen from (a), and the fitted values $\hat{y} = \mathbf{1}_n\overline{y} + \mathbf{Z}\hat{\gamma}_\lambda$ change. Therefore, imposing a uniform criteria $\mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_{p-1}$ ensures that the penalty is consistent, and in this case,

$$\hat{\gamma}_\lambda = (\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I}_{p-1})^{-1}\mathbf{Z}^\top y = \frac{1}{1+\lambda}\mathbf{Z}^\top y = \frac{1}{1+\lambda}\hat{\gamma}_{\text{OLS}},$$

the shrinkage effect is uniform over all variables. This standardizing procedure is automatically done by any good R package.

c) We provide two solutions, one using calculations of the norm, and one using the definition of ridge regression.

- Let $\mathbf{Z} = \mathbf{U}\mathbf{\Omega}\mathbf{V}^\top$ be the singular value decomposition of $\mathbf{Z}$ with $\mathbf{D} = \text{diag}(\omega_1, \ldots, \omega_{p-1})$. This gives us the eigendecomposition of $\mathbf{Z}^\top\mathbf{Z} = \mathbf{V}\mathbf{\Omega}^2\mathbf{V}^\top$, thus

$$\hat{\gamma}_\lambda = \mathbf{V}(\mathbf{\Omega}^2 + \lambda\mathbf{I}_{p-1})^{-1}\mathbf{\Omega}\mathbf{U}^\top y = \sum_{j=1}^{p-1} \frac{\omega_j}{\omega_j^2 + \lambda}(\mathbf{u}_j^\top y)\mathbf{v}_j.$$

Since a $(p-1)\times(p-1)$ matrix $\mathbf{V}$ is orthogonal,

$$\left\|\hat{\gamma}_\lambda\right\|_2^2 = y^\top\mathbf{U}\mathbf{\Omega}(\mathbf{\Omega}^2 + \lambda\mathbf{I}_{p-1})^{-2}\mathbf{\Omega}\mathbf{U}^\top y = \sum_{j=1}^{p-1}\left(\frac{\omega_j}{\omega_j^2+\lambda}\right)^2(\mathbf{u}_j^\top y)^2,$$

Since $\lambda \mapsto (w/(w^2+\lambda))^2$ is a decreasing function for any $w > 0$, the expression above is decreasing in λ.

- It suffices to show that for any $\lambda_1 > \lambda_2$, $\|\hat{\gamma}_{\lambda_1}\|_2^2 \le \|\hat{\gamma}_{\lambda_2}\|_2^2$. Since

$$\hat{\gamma}_\lambda = \arg\min_\gamma \|y - \mathbf{1}_n\overline{y} - \mathbf{Z}\gamma\|_2^2 + \lambda\|\gamma\|_2^2,$$

we have

$$\begin{aligned}\|y - \mathbf{1}_n\overline{y} - \mathbf{Z}\hat{\gamma}_{\lambda_1}\|_2^2 + \lambda_1\|\hat{\gamma}_{\lambda_1}\|_2^2 &\le \|y - \mathbf{1}_n\overline{y} - \mathbf{Z}\hat{\gamma}_{\lambda_2}\|_2^2 + \lambda_1\|\hat{\gamma}_{\lambda_2}\|_2^2,\\ \|y - \mathbf{1}_n\overline{y} - \mathbf{Z}\hat{\gamma}_{\lambda_1}\|_2^2 + \lambda_2\|\hat{\gamma}_{\lambda_1}\|_2^2 &\ge \|y - \mathbf{1}_n\overline{y} - \mathbf{Z}\hat{\gamma}_{\lambda_2}\|_2^2 + \lambda_2\|\hat{\gamma}_{\lambda_2}\|_2^2.\end{aligned}$$

Then the difference of two inequalities becomes

$$(\lambda_1 - \lambda_2)\|\hat{\gamma}_{\lambda_1}\|_2^2 \le (\lambda_1 - \lambda_2)\|\hat{\gamma}_{\lambda_2}\|_2^2,$$

hence $\|\hat{\gamma}_{\lambda_1}\|_2^2 \le \|\hat{\gamma}_{\lambda_2}\|_2^2$.

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

Solution. Let $\widetilde{y} = y - \bar{y}\mathbb{1}$ be the centered response. Then $\widehat{\gamma} = \widehat{\gamma}_{\text{lasso}}$ minimizes function f defined as

$$f(\gamma) = g(\gamma) + \lambda\|\gamma\|_1 \quad \text{with} \quad g(\gamma) = \sum_{i=1}^{n}\Big(\widetilde{y}_i - \sum_{j=1}^{q} Z_{ij}\gamma_j\Big)^2.$$

We will study what happens with the two parts of f close to 0. For g, this will be done via derivative, while the non-differentiable term $\|\gamma\|_1$ we will be inspected directly (another approach would be to use the sub-gradient, an optimization-theory notion generalizing the concept of a derivative).

The partial derivatives of g at 0 are

$$\frac{\partial g}{\partial \gamma_j}(0) = -\sum_{i=1}^{n} 2\Big(\widetilde{y}_i - \sum_{j=1}^{q} Z_{ij}0\Big)Z_{ij} = -2Z_j^\top \widetilde{y} = -2Z_j^\top y, \quad j = 1, \dots, q,$$

where the last equality comes from the fact that $Z^\top \mathbb{1} = 0$.

For the case $\lambda < \lambda^*$, we will show that there exist v such that $f(v) < f(0)$. Let j be the coordinate for which $\lambda < 2|Z_j^\top y|$, and let e_j denote the j-th vector of the standard basis (i.e. zero but 1 in the j-th coordinate). For t small we have

$$f(te_j) = g(te_j) + \lambda\|te_j\| = g(te_j) + \lambda|t| = g(0) + t\big[-2Z_j^\top y + \lambda\text{sign}(t) + o(1)\big].$$

If $Z_j^\top y > 0$, $f(te_j) < g(0) = f(0)$ for $t > 0$ small enough. If $Z_j^\top y < 0$, the same is true for $t < 0$ small enough. Hence 0 is not the minimizer of f.

Now let $\lambda > \lambda^*$. We can estimate, using the Taylor expansion for g at 0, that for any v:

$$f(v) = g(0) + [\nabla g(0)]^\top v + o(v) + \lambda\|v\|_1 \geq g(0) + \big(\lambda - \underbrace{\|\nabla g(0)\|_\infty}_{\lambda^*}\big)\|v\|_1 + o(v).$$

Recall that $o(v)/\|v\|_1 \to 0$ for $v \to 0$. Hence, since $\lambda > \lambda^*$, 0 must be a strict local minimum of f. Since f is convex, 0 must be the only minimum.

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 $\|\hat{\gamma}_1\|_1 = \|\hat{\gamma}_2\|_1$.

Solution. Since both the estimators estimate the intercept the same (as the mean), so we can only focus on Z and γ estimates, denoted as $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$. Also, denote $y^* = y - \mathbf{1}_n\overline{y}$.

a) Assume that $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ both give an optimal objective value, henceforth denoted as α. Note first that $\|Y - Z\gamma\|_2^2$ is strictly convex in $Z\gamma$, hence for $t \in (0,1)$, we have

$$\|y^* - tZ\widehat{\gamma}_1 - (1-t)Z\widehat{\gamma}_2\|_2^2 \leq t\|y^* - Z\widehat{\gamma}_1\|_2^2 + (1-t)\|y^* - Z\widehat{\gamma}_1\|_2^2. \tag{12}$$

By the strict convexity, the equality holds if only if $Z\widehat{\gamma}_1 = Z\widehat{\gamma}_2$. Also, L^1-norm is convex, hence

$$\|t\widehat{\gamma}_1 + (1-t)\widehat{\gamma}_2\|_1 \leq t\|\widehat{\gamma}_1\|_1 + (1-t)\|\widehat{\gamma}_2\|_1.$$

Owing to the optimality of $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$, we obtain

$$\begin{aligned}
\alpha &\leq \|y^* - tZ\widehat{\gamma}_1 - (1-t)Z\widehat{\gamma}_2\|_2^2 + \lambda\|t\widehat{\gamma}_1 + (1-t)\widehat{\gamma}_2\|_1 \\
&\leq \big(t\|y^* - Z\widehat{\gamma}_1\|_2^2 + (1-t)\|y^* - Z\widehat{\gamma}_2\|_2^2\big) + \lambda\left(t\|\widehat{\gamma}_1\|_1 + (1-t)\|\widehat{\gamma}_2\|_1\right) \\
&= t\Big(\|y^* - Z\widehat{\gamma}_1\|_2^2 + \lambda\|\widehat{\gamma}_1\|_1\Big) + (1-t)\Big(\|y^* - Z\widehat{\gamma}_2\|_2^2 + \lambda\|\widehat{\gamma}_2\|_1\Big) \\
&= t\alpha + (1-t)\alpha = \alpha.
\end{aligned}$$

Hence equalities must be preserved in the previous chain, and the equality of (12) holds, i.e. $Z\widehat{\gamma}_1 = Z\widehat{\gamma}_2$.

b) This is evident from the previous part, because in the last inequality we used two upper estimates. If the inequality should be equality, both of the upper estimates must be sharp. From the sharpness of the first one we deduced part a), from the second one we can deduce part b), provided $\lambda > 0$.

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

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

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 $\beta^\top X = \operatorname{med}(Y)$ 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 $\epsilon_i \sim Exp(1)$.

Solution. 1. This is clear from how median is defined. $\widetilde{\beta}$ is modeling median of the response variable in the same way as $\widehat{\beta}$ is modeling the expectation.

2. Since both the median and expectation of a standard gaussian is zero, in the gaussian case the two estimators are estimating the same:

$$\mathbb{E}Y_i = \beta_0 + \beta_1 x_i = \operatorname{med}(Y_i) \,.$$

In the second case, $\operatorname{med}(Exp(1)) \neq \mathbb{E}Exp(1)$ and both of them are non-zero, hence $\widehat{\beta}_0$ is estimating a different constant than $\widetilde{\beta}_0$, and neither is really estimating β_0. β_1 is being estimated the same, since both the median and the expectation are linear. Hence the effect of the covariate on the response is the same in both cases.

Problem 41. (Naive kernel density estimator)
Let $X_1, \dots, X_n$ 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.

Solution. Write $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{[X_i, \infty)}(x)$. Then

$$\widehat{f}(x) = \frac{1}{2nh} \sum_{i=1}^n \mathbb{1}_{[x_i - h, X_i + h)}(x) = \frac{1}{2nh} \sum_{i=1}^n \mathbb{1}_{[-h,h)}(x - x_i) = \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} \mathbb{1}_{[-1,1)} \left(\frac{x - x_i}{h} \right) =: \frac{1}{nh} \sum_{i=1}^n K\Big(\frac{x - X_i}{h} \Big) \,,$$

where $K(y) = \frac{1}{2} \mathbb{1}_{[-1,1)}$. So the kernel corresponds to $U[-1,1)$ distribution.

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

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

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.

Solution. 1. The maximum likelihood estimator of β is

$$\arg\max_{\beta} L(y;\mathbf{X}) = \arg\max_{\beta} \frac{|\Sigma|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(y-\mathbf{X}\beta)^\top \Sigma^{-1}(y-\mathbf{X}\beta)\right\}.$$

and thus finding the MLE amounts to minimization of $(y-\mathbf{X}\beta)^\top\Sigma^{-1}(y-\mathbf{X}\beta)$.

2. Since $\Sigma = \mathbf{U}\Lambda\mathbf{U}^\top$ is positive definite, its inverse Σ^{-1} is well-defined and positive definite and by the spectral theorem admits a square root $\Sigma^{-1/2} = \mathbf{U}\Lambda^{-1/2}\mathbf{U}^\top$.

   One can rewrite the regression as the classical linear model setting by premultiplying by $\Sigma^{1/2}$. The normal equation can also be derived using vector calculus, by differentiating $(y-\mathbf{X}\beta)^\top\Sigma^{-1}(y-\mathbf{X}\beta)$ with respect to β and setting the derivative to zero. The normal equations are

$$\mathbf{X}^\top\Sigma^{-1}\mathbf{X}\beta = \mathbf{X}^\top\Sigma^{-1}y$$

   and since $\mathbf{X}^\top\Sigma^{-1}\mathbf{X}$ is a quadratic form and Σ is positive definite, the inverse is well-defined. Differentiating twice gives $2\mathbf{X}^\top\Sigma^{-1}\mathbf{X}$ and since the Hessian is positive, $\widehat{\beta}_{\text{GLS}}$ minimizes the distance and is therefore the maximum likelihood estimator of β.

3. By the transformation property, the estimator $\widehat{\beta}_{\text{GLS}}$ is Gaussian because $\boldsymbol{\varepsilon}$ is also Gaussian. Its mean and variance are

$$\mathbb{E}\widehat{\beta}_{\text{GLS}} = (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbb{E}Y = (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbf{X}\beta = \beta$$

   and

$$\begin{aligned}\mathsf{Var}\left(\widehat{\beta}_{\text{GLS}}\right) &= (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathsf{Var}\left(Y\right)\Sigma^{-1}\mathbf{X}(\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\\ &= (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\Sigma\Sigma^{-1}\mathbf{X}(\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\\ &= (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}.\end{aligned}$$

4. First, the ordinary least square (OLS) estimator is unbiased,

$$\mathbb{E}\widehat{\beta}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}Y = \beta.$$

   Let $Y^* = \Sigma^{-1/2}Y$. Then, the linear model with $Y^* = \Sigma^{-1/2}\mathbf{X}\beta + \boldsymbol{\varepsilon}^*$ satisfies the hypothesis of the Gauss–Markov theorem with $\boldsymbol{\varepsilon}^* \sim \mathcal{N}_n(0_n, \mathbf{I}_n)$ and the OLS estimator of this regression is BLUE. Since we premultiply by the matrix $\Sigma^{-1/2}$, the design matrix becomes $\Sigma^{-1/2}\mathbf{X}$ and so the BLUE estimator is $\widehat{\beta}_{\text{GLS}}$.

   Alternatively, proceed as in the proof of Gauss–Markov theorem to show that $\widehat{\beta}_{\text{GLS}}$ is BLUE.

   Let $\widetilde{\beta}$ be any linear unbiased estimator of β, necessarily of the form $\mathbf{A}Y$ with $\mathbf{A}\mathbf{X} = \mathbf{I}_n$. Write $\mathsf{Var}\left(\widetilde{\beta}\right) = \mathbf{A}\Sigma\mathbf{A}^\top$ and the difference between the variance of the estimators as

$$\begin{aligned}\mathsf{Var}\left(\widetilde{\beta}\right) - \mathsf{Var}\left(\widehat{\beta}_{\text{GLS}}\right) &= \mathbf{A}\Sigma\mathbf{A}^\top - (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\\ &= \mathbf{A}\left\{\Sigma - \mathbf{X}(\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\right\}\mathbf{A}^\top\\ &= \mathbf{A}\Sigma^{1/2}\left\{\mathbf{I}_n - \Sigma^{-1/2}\mathbf{X}(\mathbf{X}^\top\Sigma^{-1/2}\Sigma^{-1/2}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1/2}\right\}\Sigma^{1/2}\mathbf{A}^\top\\ &= \mathbf{A}\Sigma^{1/2}\mathbf{M}_{\Sigma^{-1/2}\mathbf{X}}\Sigma^{1/2}\mathbf{A}^\top.\end{aligned}$$

   Since $\mathbf{M}_{\Sigma^{-1/2}\mathbf{X}}$ is a projection matrix, it is idempotent and the difference $\mathsf{Var}\left(\widetilde{\beta}\right) - \mathsf{Var}\left(\widehat{\beta}_{\text{GLS}}\right)$ is a quadratic form and hence positive semi-definite. Since $\widehat{\beta}_{\text{OLS}}$ is a linear unbiased estimator (with $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$), it is not the BLUE in this particular example.

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?

Solution. Some preliminary remark:

a) In this exercise, we will denote j-th row of X by $x_j^\top$, thus $X^\top = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$.

b) Let us recall that under regularity assumptions, the MLE of θ is asymptotically Gaussian, with covariance matrix *the inverse of the Fisher information matrix*.

1. The model is $y = X\beta + \varepsilon$, with ε of zero mean and known variance σ^2. The variance of the LSE is thus $\sigma^2(X^\top X)^{-1}$.

   The density of y_j is $g(y_j - x_j^\top \beta)$, with g the density of ε_j. Therefore, we have

$$\ell(\beta) = \sum_{j=1}^n \log g(y_j - x_j^\top \beta), \quad \beta \in \mathbb{R}^p.$$

   Let us notice that $h_j(\beta) = \log g(y_j - x_j^\top \beta)$. From the "chain-rule", we have

$$\frac{\partial h_j}{\partial \beta} = -x_j \left. \frac{d \log g(u)}{du} \right|_{u = y_j - x_j^\top \beta}$$

   thus

$$\frac{\partial^2 h_j}{\partial \beta^2} = \frac{\partial}{\partial \beta} \frac{\partial h_j}{\partial \beta} = -\frac{\partial}{\partial \beta} \left(\left. \frac{d \log g(u)}{du} \right|_{u = y_j - x_j^\top \beta} \right) x_j^\top = x_j x_j^\top \cdot \left. \frac{d^2 \log g(u)}{du^2} \right|_{u = y_j - x_j^\top \beta}.$$

   We used the fact that if A is a matrix and $f(\beta)$ is a vectorial function such that $Af(\beta)$ is defined, then $\frac{\partial (Af(\beta))}{\partial \beta} = \frac{\partial (f(\beta))}{\partial \beta} A^\top$ (from the "chain-rule").

   Hence,

$$-\frac{\partial^2 \ell}{\partial \beta^2} \quad = \quad -\sum_j x_j x_j^\top \left. \frac{d^2 \log g(u)}{du^2} \right|_{u = y_j - x_j^\top \beta}$$

   and so,

$$I(\beta) = \mathbb{E}\left\{ -\frac{\partial^2 \ell}{\partial \beta^2} \right\} = \sum_{j=1}^n x_j x_j^\top \mathbb{E}\left\{ - \left. \frac{d^2 \log g(u)}{du^2} \right|_{u = y_j - x_j^\top \beta} \right\},$$

   where the expectation becomes

$$\int -\frac{d^2 \log g(u)}{du^2} g(u)du = i_g,$$

   with the change of variable $y_j - x_j^\top \beta = u$. That implies $I(\beta) = i_g X^\top X$, and so the MLE has as asymptotic variance $i_g^{-1}(X^\top X)^{-1}$.

   The asymptotic relative efficiency of least squares with respect to the MLE is thus

$$\left\{ \frac{|i_g^{-1}(X^\top X)^{-1}|}{|\sigma^2 (X^\top X)^{-1}|} \right\}^{1/p} = \frac{1}{i_g \sigma^2}.$$

2. We have $g(u) = (2\pi\sigma^2)^{-1/2} e^{-u^2/(2\sigma^2)}$ for $u \in \mathbb{R}$, thus $i_g = 1/\sigma^2$, which give an efficiency of 1. That is not surprising, since the LSE is exactly the MLE in this case.

3. Let $\lambda = \sqrt{2}/\sigma$, where σ^2 is the variance. The Laplace density is $g(u) = (\lambda/2)\exp(-\lambda|u|)$, $u \in \mathbb{R}$. Let us remark that, since the MLE is regular, we have

$$i_g = \int -\frac{d^2 \log g(u)}{du^2} g(u) du = \int \left\{ \frac{d \log g(u)}{du} \right\}^2 g(u) du = \int \{-\lambda \text{sgn}(u)\}^2 (\lambda/2) \exp(-\lambda|u|) du = \lambda^2.$$

So, the asymptotic relative efficiency is $1/(\lambda^2\sigma^2) = 1/2$.

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

Solution. The adjusted values are

$$\hat{y}_i = \hat{g}(x_i) = \sum_{j=1}^n y_j \frac{K\left(\frac{x_j - x_i}{\lambda}\right)}{\sum_{k=1}^n K\left(\frac{x_k - x_i}{\lambda}\right)}.$$

By defining

$$S_{\lambda,ij} = \frac{K\left(\frac{x_j - x_i}{\lambda}\right)}{\sum_{k=1}^n K\left(\frac{x_k - x_i}{\lambda}\right)},$$

we have $\hat{y} = S_\lambda y$, where S_λ is called the *smoother matrix*. In non-parametric regression, this is the analogue of the hat matrix.

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

$$\min_{f \in C^2[a,b]} L(f), \text{ where } L(f) = \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_a^b \{f''(x)\}^2 dx, \quad \lambda > 0, \tag{13}$$

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 (13) has a solution $\hat{f}$, then$\hat{f} \in N(x_1, x_2, \ldots, x_n)$.

Solution. 1. Using integration by parts, we obtain that

$$\begin{aligned}
\int_a^b g''(x) h''(x) dx &= \underbrace{g''(x) h'(x)\Big|_a^b}_{=0, \text{ car } g''(a)=g''(b)=0} - \int_a^b g'''(x) h'(x) dx \\
&= -\sum_{i=1}^{n-1} g'''(x_i^+) \int_{x_i}^{x_{i+1}} h'(x) dx \\
&= -\sum_{i=1}^{n-1} g'''(x_i^+) \{h(x_{i+1}) - h(x_i)\} = 0.
\end{aligned}$$

Here, the second equality comes from the fact that $g'''(x) = 0$ inside the intervals (a, x_1) and (x_n, b) and that $g'''(x)$ equals to the constant $\lim_{x \to x_i^+} g'''(x) = g'''(x_i^+)$ inside the interval (x_i, x_{i+1}). To obtain the last equality finally, observe that $\tilde{g}(x_i) = g(x_i) = z_i$ hence $h(x_i) = 0$ for every i.

2. By direct computation we obtain that

$$\begin{aligned}\int_a^b \{\tilde{g}''(x)\}^2 dx &= \int_a^b \{g''(x) + h''(x)\}^2 dx \\ &= \int_a^b \{g''(x)\}^2 dx + 2\int_a^b g''(x)h''(x)dx + \int_a^b \{h''(x)\}^2 dx \\ &= \int_a^b \{g''(x)\}^2 dx + \int_a^b \{h''(x)\}^2 dx \geq \int_a^b \{g''(x)\}^2 dx.\end{aligned}$$

where we have equality if and only if $h''(x) \equiv 0$, so we must have $h(x) = kx + c$. But since $h(x_i) = 0$ for every i, it must be that$h(x) \equiv 0$. In particular we have equality if and only if $\tilde{g} = g$.

3. Let $\tilde{f} \in C^2[a, b] \backslash N(x_1, \ldots, x_n)$ and let $f \in N(x_1, \ldots, x_n)$ the spline which is interpolating the points $(x_i, \tilde{f}(x_i))$, $i = 1, \ldots, n$. By point (2)

$$\int_a^b \{\tilde{f}''(x)\}^2 dx > \int_a^b \{f''(x)\}^2.$$

Moreover

$$\sum_{i=1}^n (y_i - \tilde{f}(x_i))^2 = \sum_{i=1}^n (y_i - f(x_i))^2.$$

Hence, $L(\tilde{f}) > L(f)$ and we notice that if the minimum exists, it must belong to $N(x_1, \ldots, x_n)$.

Remark. Using the properties of splines, it it possible to show that a minimum always exists and is unique. Hence the problem $\min_{f \in C^2[a,b]} L(f)$ admits always a unique solution and this solution is a natural cubic spline.

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.

Solution. For the if part, let Ω be the covariance matrix of a random vector $Y \in \mathbb{R}^p$. Then, for any $a \in \mathbb{R}^p$, $a^\top \Omega a$ is the variance of the random variable $a^\top Y$. This shows that $a^\top \Omega a \geq 0$ for every $a \in \mathbb{R}^p$ and hence Ω is non-negative definite.

Conversely, let Ω be non-negative definite. So, we can write $\Omega = \mathbf{U}\Lambda\mathbf{U}^\top$, where $\mathbf{U} \in \mathbb{R}^{p \times p}$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_i \geq 0 \, \forall \, i$. Now, we can always find a random vector $X \in \mathbb{R}^p$ such that Λ is the covariance matrix of X (take independent random variables with variances given by the eigenvalues, and use them to form random vector X). Then, Ω is the covariance matrix of $Y = \mathbf{U}X$.

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.

Solution.

- not b) $\Rightarrow$ not a):
  Assume the jth eigenvalue of $\mathbf{\Omega}$ is $\lambda_j \leq 0$ (or $\lambda_j < 0$) with eigenvector u_j, then so is $u_j^\top \mathbf{\Omega} u_j = \lambda_j$, which contradicts positive definiteness (or semi-definiteness, respectively).

- b) $\Rightarrow$ a):
  Note that you can write singular value decomposition $\mathbf{\Omega} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ also as $\mathbf{\Omega} = \sum_{i=1}^p \lambda_i u_i u_i^\top$ where the λ_i are the diagonal entries of $\mathbf{\Lambda}$ and the u_i are the column vectors of $\mathbf{U}$ which form an orthonormal basis (ONB) of $\mathbb{R}^p$. Hence, $x^\top \mathbf{\Omega} x = \sum_{i=1}^p \lambda_i \underbrace{(u_i^\top x)^2}_{\geq 0} \overset{\lambda_1, \ldots, \lambda_p \geq 0}{\geq} \lambda_j (u_j^\top x)^2$ for any j. For $x \neq 0$, we can further choose j such that also $u_i^\top x \neq 0$ since the u_i form an ONB. Thus, we have that from $\lambda_1, \ldots, \lambda_p (>) \geq 0$ immediately implies $x^\top \Omega x (>) \geq 0$.

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.

Solution. For the manual calculation, we begin by calculating the eigenvalues of Σ as roots of the characteristic polynomial:

$$\det(\Sigma - \lambda \mathbf{I}) = (5/2 - \lambda)^2 - (3/2)^2 = 0 \qquad \rightsquigarrow \qquad \lambda_1 = 4 \quad \& \quad \lambda_2 = 1.$$

Then we can easily calculate the principal components, i.e. the eigenvectors:

$$(\Sigma - 4\mathbf{I})v = 0 \quad \rightsquigarrow \quad v_1 = (1,1)^\top \qquad \& \qquad (\Sigma - \mathbf{I})v = 0 \quad \rightsquigarrow \quad v_2 = (1,-1)^\top.$$

For points 2.-5. we give the code:

```
# 2.
Sigma <- matrix(c(5,3,3,5)/2, ncol=2)
EIG <- eigen(Sigma)
EIG$vectors[,1] # v_1
EIG$vectors[,2] # v_2
# 3.
sqrt_Sigma <- EIG$vectors %*% diag(sqrt(EIG$values)) %*% t(EIG$vectors)
n <- 100
X <- matrix(rnorm(2*n),ncol=2) %*% sqrt_Sigma # the data matrix
# 4.
SVD <- svd(X)
SVD$v[,1] # \hat{v}_1
SVD$v[,2] # \hat{v}_2
# 5.
plot(X[,1],X[,2],col="gray")
sign1 <- sign(sum(EIG$vectors[,1]*SVD$v[,1]))
sign2 <- sign(sum(EIG$vectors[,2]*SVD$v[,2]))
arrows(0,0,sqrt(EIG$values[1])*EIG$vectors[1,1],sqrt(EIG$values[1])*EIG$vectors[2,1],col="gray60")
arrows(0,0,sqrt(EIG$values[2])*EIG$vectors[1,2],sqrt(EIG$values[2])*EIG$vectors[2,2],col="gray60")
arrows(0,0,sign1*SVD$d[1]/sqrt(n)*SVD$v[1,1],sign1*SVD$d[1]/sqrt(n)*SVD$v[2,1])
arrows(0,0,sign2*SVD$d[2]/sqrt(n)*SVD$v[1,2],sign2*SVD$d[2]/sqrt(n)*SVD$v[2,2])
```

The code for point 5. looks quite complicated for the following reasons:

- Signs of eigenvectors and singular vectors are irrelevant. For example, the software can produce either $v_1 = (1,1)^\top$ or $v_1 = (-1,-1)^\top$. If we want to ensure that v_1 and its estimator $\widehat{v}_1$ are facing in a similar direction, we have to take care of the signs manually.
- The eigenvalue λ_1 captures variance in the direction of the first principal component. Note that variance is not useful for plotting, standard deviation is preferred with this respect. By multiplying v_1 by $\sqrt{\lambda_1}$, we are including the information on the data spread in the plot, making it look more natural.
- What is the relation between the eigenvalues of Σ and the singular values of the data matrix $\mathbf{X}$? Recall that the squared singular values of $\mathbf{X}$ are eigenvalues of $\mathbf{X}^\top\mathbf{X}$. However, $\mathbf{X}^\top\mathbf{X}$ is not the estimator of Σ, $\mathbf{X}^\top\mathbf{X}/n$ is! Putting everything together, the corresponding scale to $\sqrt{\lambda_1}$ is $\sigma_1/\sqrt{n}$.

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 $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ 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.

Solution. We can rearrange the objective as

$$\begin{aligned}\sum_{i=1}^{n} \|x_i - \mathbf{Q}\mathbf{Q}^\top x_i\|_2^2 = \|\mathbf{X} - \mathbf{X}\mathbf{Q}\mathbf{Q}^\top\|_F^2 &= \operatorname{tr}\Big[(\mathbf{X} - \mathbf{X}\mathbf{Q}\mathbf{Q}^\top)^\top(\mathbf{X} - \mathbf{X}\mathbf{Q}\mathbf{Q}^\top)\Big] \\ &= \operatorname{tr}(\mathbf{X}^\top\mathbf{X}) - 2\operatorname{tr}(\mathbf{X}^\top\mathbf{X}\mathbf{Q}\mathbf{Q}^\top) + \operatorname{tr}(\mathbf{Q}\mathbf{Q}^\top\mathbf{X}^\top\mathbf{X}\mathbf{Q}\mathbf{Q}^\top) \\ &= \operatorname{tr}(\mathbf{X}^\top\mathbf{X}) - \operatorname{tr}(\mathbf{Q}^\top\mathbf{X}^\top\mathbf{X}\mathbf{Q}) = \operatorname{tr}(\mathbf{X}^\top\mathbf{X}) - \operatorname{tr}(\mathbf{Q}^\top\mathbf{V}\mathbf{D}^2\mathbf{V}^\top\mathbf{Q})\end{aligned}$$

where we used the cyclic permutation property of the trace and the fact that $\mathbf{Q}$ is orthogonal. The first term of the final expression does not depend on $\mathbf{Q}$ and hence it can be dropped. The minus in front of the second term then changes the minimization problem to the maximization one, hence the problems are equivalent.

Now, if we choose $\mathbf{Q} = (v_1, \ldots, v_q)$, the objective value is $\sum_{i=1}^{q} d_i^2$. Given the order of the singular values, $d_1 \geq d_2 \geq \ldots$, and the the required orthogonality of $\mathbf{Q}$, this is the highest objective value one can obtain.

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:

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

Solution. We can create and plot the data using the following code:

```
x <- runif(100)*2
e <- rnorm(100)/sqrt(10) # notice the square-root
y <- 10 + 2*sin(pi*x) + e
plot(x,y)
```

Since we know how we generated the data, it is clear how to transform the x-axis to obtain a clear linear relationship:

```
plot(sin(pi*x),y)
```

So the transformation is $\widetilde{x} = \sin(\pi * x)$. Once we have the transformed plot, one can see that the constant 10 is the intercept, i.e. the value of the (imaginary) line at $\widetilde{x} = 0$, while 2 is the slope of that line.

Similarly for the second dependent variable, the transformation of the x-axis is $\widetilde{x} = \log(x)$, only this time, we also have to transform the y-axis as $\widetilde{y} = \log(y)$.

Note that the first dependent variable follows a linear model. The second one doesn't, it does only after a log-transformation. In the case of the first dependent variable, we could have probably guessed the transformation even without knowing how the data were generated. In the case of the second dependent variable, we probably would have been lost. But still, in the latter case, the variance seems to be increasing with increasing values of y, which points towards log-transformation of the response.

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.

Solution. (a) This is a nonlinear model with $\eta(\beta) = \beta_1 \cos(x - \beta_1)$. So, we can use Newton-Raphson method (slide 253 in the lecture notes). We start with an initial choice $\beta^{(0)} = (\beta_1^{(0)}, \beta_2^{(0)})^\top$. For $h = 0, 1, 2, \ldots$, we iteratively fit linear regression with design matrix $D^{(h)} = \nabla_\beta \eta(\beta)$ and response $y - \eta(\beta^{(h)})$, that is find $u^{(h)} = ({D^{(h)}}^\top D^{(h)})^{-1} {D^{(h)}}^\top (y - \eta(\beta^{(h)}))$, and update $\beta^{(h+1)} = \beta^{(h)} + u^{(h)}$. For this specific problem, $\frac{\partial \eta(\beta)}{\partial \beta_1} = \cos(x - \beta_2)$, $\frac{\partial \eta(\beta)}{\partial \beta_2} = \beta_1 \sin(x - \beta_2)$. So,

$$D^{(h)} = \begin{pmatrix} \cos(x_1 - \beta_2^{(h)}) & \beta_1^{(h)} \sin(x_1 - \beta_2^{(h)}) \\ \cos(x_2 - \beta_2^{(h)}) & \beta_1^{(h)} \sin(x_2 - \beta_2^{(h)}) \\ \vdots & \vdots \\ \cos(x_n - \beta_2^{(h)}) & \beta_1^{(h)} \sin(x_n - \beta_2^{(h)}) \end{pmatrix}, \qquad y - \eta(\beta^{(h)}) = \begin{pmatrix} y_1 - \beta_1^{(h)} \cos(x_1 - \beta_2^{(h)}) \\ y_2 - \beta_1^{(h)} \cos(x_2 - \beta_2^{(h)}) \\ \vdots \\ y_n - \beta_1^{(h)} \cos(x_n - \beta_2^{(h)}) \end{pmatrix}.$$

This is repeated until convergence.

(b) The problem can be reformulated as follows.

$$y = \beta_1 \cos(x - \beta_2) + \epsilon = \underbrace{\beta_1 \cos(\beta_2)}_{\beta_1^*} \underbrace{\cos(x)}_{x_1} - \underbrace{\beta_1 \sin(\beta_2)}_{\beta_2^*} \underbrace{\sin(x)}_{x_2} + \epsilon = \beta_1^* x_1 + \beta_2^* x_2 + \epsilon$$

This is a simple linear regression problem with $x_1 = \cos(x)$, $x_2 = \sin(x)$ and no intercept term.

(c) Here is the code:

```
### part (b):
m3 <- lm(y~I(cos(x)) + I(sin(x))-1,data=data1)
summary(m3)
plot(data1$x,data1$y)
points(x,fitted(m3),col="red",pch=0)

### part(a): Newton-Rhanpson for NLLS
# mean function
nu_cos <- function(x,beta) beta[1]*cos(x-beta[2])
# fitting algorithm
NNLS_cos <- function(beta, n_iter, data){
# data - data frame
# beta - starting value for the algorithm (in this case a vector)
# maxiter - no. of iterations, we do not check any stopping criterion for simplicity
  for(k in 1:n_iter){
    data_akt <- data.frame(x1=cos(data$x - beta[2]),
                           x2=beta[1]*sin(data$x - beta[2]),
                           y=data$y-nu_cos(data$x,beta) )
    m_akt <- lm(y~x1+x2-1, data=data_akt)
    u <- unname(coef(m_akt))
    beta <- beta + u
  }
```

```
  return(beta)
}

beta_0 <- c(0.5,-0.5) # starting point - try out different ones!
( beta_hat <- NNLS_cos(beta_0,100,data1) ) # let it run for 100 iterations, that should be enough
fitted_val <- beta_hat[1]*cos(data1$x - beta_hat[2])
# do we really get the same estimates?
plot(data1$x,data1$y)
points(x,fitted(m3),col="red",pch=0)
points(data1$x, fitted_val, col="blue", pch=4)
# yes! you can also verify it analytically
```