SF2930 Regression Analysis
Exam Generator
Version for the re-exam 2023-06-07

Timo Koski

Innehåll

1 Introduction

1.1 Instructions

This is the final version. Errors and typos and bad formulations will sought for and corrected, whenever detected.

> In the Exam the students will be provided with two documents:
>
> 1. First: an A4 sheet pointing out by problem number (e.g., 4.11) those problems in this generator to be solved in the Exam hall.
>
> 2. Second: The complete Exam generator (incl. Appendices, that is section section 9) WILL BE AVAILABLE IN THE EXAM. When solving a problem in the exam, one may refer to formulas and statement in problems in the generator NOT included in the exam without deriving or proving them.

The formulas in section 9 can be used in the solutions of any exam problems without proof or derivation.

1.2 Linear Algebra for Warm-Up

The problems in section are not very likely to appear as exam questions, but can be useful in more serious candidates for exam assignments.

Problem 1.1.

The ON-basis of $\mathbb{R}^n$ is

$$\mathfrak{E}_1 = \begin{pmatrix} 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \dots \quad \mathfrak{E}_j = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad \dots \quad \mathfrak{E}_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \end{pmatrix}. \tag{1}$$

Let A be any $n \times n$ matrix. Then

$$\mathfrak{E}_i^T A \tag{2}$$

is the ith row of A.

- Express in words the meaning of the following expressions:

$$A\mathfrak{E}_i, \quad \mathfrak{E}_i^T A \mathfrak{E}_i \quad \mathfrak{E}_i^T A \mathfrak{E}_j \quad \mathfrak{E}_j^T A \mathfrak{E}_i. \tag{3}$$

Problem 1.2.

Let $A^T = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ where $\mathbf{a}_i$ is the ith column vector of the $k \times n$ matrix A^T. Thus $\mathbf{a}_i^T$ is the ith row vector of A (an $n \times k$ matrix), i.e.,

$$A = \begin{pmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_n^T \end{pmatrix}.$$

Let B be a $k \times k$ matrix. Check that

$$ABA^T = \begin{pmatrix} \mathbf{a}_1^\intercal B\mathbf{a}_1 & \mathbf{a}_1^\intercal B\mathbf{a}_2 & \cdots & \mathbf{a}_1^\intercal B\mathbf{a}_n \\ \mathbf{a}_2^\intercal B\mathbf{a}_1 & \mathbf{a}_2^\intercal B\mathbf{a}_2 & \cdots & \mathbf{a}_2^\intercal B\mathbf{a}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_n^\intercal B\mathbf{a}_1 & \mathbf{a}_n^\intercal B\mathbf{a}_2 & \cdots & \mathbf{a}_n^\intercal B\mathbf{a}_n \end{pmatrix}. \tag{4}$$

2 Matrix Tricks and Linear Regression

Problem 2.1.

Consider the $k \times k$ matrix

$$S := \frac{1}{n-1} X_c^T X_c. \tag{5}$$

where X_c is given in (64).

a) Is S positive definite? Justify your answer. *Aid:* Slide 22/81 in Lecture 3.

b) Check that

$$s_{jk} = \mathfrak{E}_j^T S \mathfrak{E}_k = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \tag{6}$$

(c.f. (3)), where $\mathfrak{E}_j$ and $\mathfrak{E}_k$ are now column vectors in the ON-basis of $\mathbb{R}^k$, see (1).

s_{jk} in (6) is an estimate of the covariance between the regressor j and regressor number k in the population underlying the rows in the matrix $X_R = \left(\mathbf{x}_1^T \ldots \mathbf{x}_n^T\right)^T$, see also (63). Hence the matrix S is called the **sample covariance matrix**.

Problem 2.2.

The centering matrix is

$$C_{ce} = \mathbb{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T.$$

See section 9.1 and (97) for the detailed definition.

a) Check that C_{ce} is a projection matrix, c.f. section 9.3.

b) Show that

$$\mathbf{y}^T C_{ce} C_{ce} \mathbf{y} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{7}$$

c) Is the centering matrix invertible? Your answer should contain an explanation, not merely one of Yes or No.

Problem 2.3.

A training set of observed responses $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and the corresponding $n \times (k+1)$ data matrix X are available. We fit a multiple least squares model to the training set and obtain the LSE $\hat{\boldsymbol{\beta}}$. Then

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = H\mathbf{y}$$

is the predictor vector, where H is the hat matrix and $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T$.

a) $\hat{e}_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$ are the LS residuals. Show that

$$\sum_{i=1}^{n} \hat{e}_i = 0$$

by studying

$$\hat{\mathbf{e}}^T \mathbf{1}_n,$$

where $\hat{\mathbf{e}} = (\hat{e}_1, \hat{e}_2 \ldots \hat{e}_n)^T$ is an $n \times 1$ vector.

b) $\hat{y}_i$ are the LS-predictors of y_i, respectively, for $i = 1, \ldots, n$. Show that

$$\frac{1}{n} \sum_{1=}^{n} \hat{y}_i = \bar{y}$$

by studying

$$\hat{\mathbf{y}}^T \mathbf{1}_n,$$

where $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2 \ldots \hat{y}_n)^T$ is an $n \times 1$ vector.

c) Show that

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \min_{\beta} \|\mathbf{y} - \mathbf{1}_n \beta\|^2$$

where $\mathbf{y} = (y_1, y_2 \ldots y_n)^T$ is an $n \times 1$ vector and β is a scalar.

Problem 2.4.

$\mathbf{x}$ and $\mathbf{y}$ are two column vectors in $\mathbb{R}^n$. Then we define the (sample) correlation coefficient $\operatorname{cor}(\mathbf{x}, \mathbf{y})$ by

$$\operatorname{cor}(\mathbf{x}, \mathbf{y}) := \frac{\mathbf{x}^T C_{ce} \mathbf{y}}{\sqrt{\mathbf{x}^T C_{ce} \mathbf{x}} \sqrt{\mathbf{y}^T C_{ce} \mathbf{y}}} \tag{8}$$

A training set of observed responses $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and the corresponding $n \times (k+1)$ data matrix X are available. We fit a multiple least squares model to the training set and obtain the LSE $\hat{\boldsymbol{\beta}}$. Then

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = H\mathbf{y}$$

is the predictor vector, where H is the hat matrix. Show by means of (8) that

$$\text{cor}\,(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{\text{SS}_R}{\text{SS}_T}}.$$

Aid: Use the hat matrix formula for $\hat{\mathbf{y}}$.

Problem 2.5.

a) Set, see (96) in section 9.1 for details,

$$P_{\mathbf{1}_n} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T.$$

Check that $P_{\mathbf{1}_n}$ is a projection matrix, see section 9.3.

b) Show that the range space, see (98), of $P_{\mathbf{1}_n}$ is

$$\mathcal{R}\,(P_{\mathbf{1}_n}) = \{\mathbf{v} \in \mathbb{R}^n | \mathbf{v} = \hat{v}\mathbf{1}_n\}\,,$$

where

$$\hat{v} = \frac{\mathbf{v}^T\mathbf{1}_n}{\parallel \mathbf{1}_n \parallel^2}.$$

In other words, $\mathcal{R}\,(P_{\mathbf{1}_n})$ is the subspace of vectors proportional to the vector $\mathbf{1}_n$.

c) We have that the centering matrix, see (97) in section 9.1 is also written as

$$C_{ce} = \mathbb{I}_n - P_{\mathbf{1}_n}.$$

We know that C_{ce} is a projection matrix. What is the range space of C_{ce}? c) How are the range spaces of $P_{\mathbf{1}_n}$ and C_{ce} related to each other?

3 Covariance Matrices, Random Vectors, Minimum Mean Square Estimation

Problem 3.1.

An $n \times 1$ random vector $\mathbf{X}$ has expectation $\boldsymbol{\mu}_{\mathbf{X}} = E\,[\mathbf{X}]$, and An $n \times 1$ random vector $\mathbf{Y}$ has expectation $\boldsymbol{\mu}_{\mathbf{Y}} = E\,[\mathbf{Y}]$. Define their cross-covariance matrix as

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) := E\left[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})\,(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})^T\right].$$

It is easily seen that (you need not prove this)

$$\mathrm{Cov}(\mathbf{Y},\mathbf{X}) = \mathrm{Cov}(\mathbf{X},\mathbf{Y})^T. \tag{9}$$

Clearly

$$C_{\mathbf{X}} = \mathrm{Cov}(\mathbf{X},\mathbf{X}). \tag{10}$$

a) Show that

$$\mathrm{Cov}(\mathbf{X},\mathbf{Y}) = E\left[\mathbf{X}\mathbf{Y}^T\right] - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{Y}}^T.$$

b) A and B are matrices of suitable dimensions. Show that

$$\mathrm{Cov}(A\mathbf{X}, B\mathbf{Y}) = A\mathrm{Cov}(\mathbf{X},\mathbf{Y})B^T. \tag{11}$$

c) A,B and C are matrices of suitable dimensions. $\mathbf{Z}$ is an $n \times 1$ random vector. Show that

$$\mathrm{Cov}(A\mathbf{X}, B\mathbf{Y} + C\mathbf{Z}) = \mathrm{Cov}(A\mathbf{X}, B\mathbf{Y}) + \mathrm{Cov}(A\mathbf{X}, C\mathbf{Z}). \tag{12}$$

Problem 3.2.

$\mathbf{X}$ is an $n \times 1$ random vector with $\boldsymbol{\mu}$ as mean vector and Σ as a positive definite covariance matrix. The **Mahalanobis distance** between a vector $\mathbf{x} \in \mathbb{R}^n$ and the mean $\boldsymbol{\mu}$ is defined as

$$d_M(\mathbf{x}, \boldsymbol{\mu}; \Sigma) := \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}. \tag{13}$$

a) Find

$$E\left[d_M^2(\mathbf{X}, \boldsymbol{\mu}; \Sigma)\right].$$

Aid: Use $E\left[(\mathbf{X}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X}-\boldsymbol{\mu})\right] = \mathrm{Tr}\, E\left[(\mathbf{X}-\boldsymbol{\mu}) \Sigma^{-1} (\mathbf{X}-\boldsymbol{\mu})^T\right]$ and and a linear change of variable a factorization $\Sigma = AA^T$, where A is invertible. See the statements in section 9.12.

b) If $\mathbf{v} \in \mathbb{R}^n$ and A is a real, symmetric, positive-definite $n \times n$ matrix, then the set

$$\mathbb{D}(\mathbf{v}, h) = \{\mathbf{x} \in \mathbb{R}^n | (\mathbf{x}-\mathbf{v})^{\mathsf{T}} A (\mathbf{x}-\mathbf{v}) \leq h\}$$

is an ellipsoid with radius h centered at $\mathbf{v}$. The eigenvectors of A are the principal axes of $\mathbb{D}$. Consider the Mahalanobis ellipsoid

$$\mathbb{D}_M(\boldsymbol{\mu}, h) = \{\mathbf{x} \in \mathbb{R}^n | (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \leq h\} \tag{14}$$

Assume that $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$. Find $h_{0.05}$ such that

$$P(\mathbf{X} \in \mathbb{D}_M(\boldsymbol{\mu}, h_{0.05})) = 0.95.$$

Aid: See section 9.12.

Problem 3.3.

X is $n \times k$ and $\operatorname{rank} X = r \leq n < k$. The BMP - inverse of X is

$$X^{-} = \left(X^T X\right)^{-} X^T. \tag{15}$$

Consider the OLS

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\epsilon},$$

where $\operatorname{rank} X = r \leq n \leq k$. $\widehat{\boldsymbol{\beta}}^{\dagger} = \left(X^T X\right)^{-} X^T \mathbf{y}$. We have $\parallel \mathbf{x} \parallel = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$. Show that

$$\frac{1}{n} \operatorname{E} \parallel X\widehat{\boldsymbol{\beta}}^{\dagger} - X\boldsymbol{\beta}_* \parallel^2 = \frac{\sigma^2}{n} r.$$

You can use the facts that $\left(\left(X^T X\right)^{-}\right)^T = \left(\left(X^T X\right)^T\right)^{-}$ and $\operatorname{Tr} XX^{-} = r$ without proving these.

Problem 3.4.

This assignment deals with **Linear Minimal Mean Square Estimation**. We have two $n \times 1$ random vectors $\mathbf{X}$ and $\mathbf{Y}$ with zero expectation vectors. We consider a situation, where $\mathbf{Y}$ is a hidden and $\mathbf{X}$ is an observed measurement. We know the cross covariance matrix $\operatorname{Cov}(\mathbf{Y}, \mathbf{X}) = E\left[\mathbf{Y}\mathbf{X}^T\right]$ and the covariance matrix $\mathrm{C}_{\mathbf{X}}$, where we assume that $\mathrm{C}_{\mathbf{X}}^{-1}$ exists. The goal is to find a linear estimator $A\mathbf{X}$ of $\mathbf{Y}$ such that the mean square error

$$\mathrm{MSE} = \operatorname{E} \parallel \mathbf{Y} - A\mathbf{X} \parallel^2$$

is minimized. We have

$$\operatorname{E} \parallel \mathbf{Y} - A\mathbf{X} \parallel^2 = \operatorname{E}\left[(\mathbf{Y} - A\mathbf{X})^T (\mathbf{Y} - A\mathbf{X})\right] = \operatorname{Tr} \operatorname{E}\left[(\mathbf{Y} - A\mathbf{X})(\mathbf{Y} - A\mathbf{X})^T\right]$$

We find the linear minimal mean square estimator by the following steps.

a) Check that if

$$A = \operatorname{Cov}(\mathbf{Y}, \mathbf{X}) \mathrm{C}_{\mathbf{X}}^{-1}, \tag{16}$$

then we have the orthogonality condition

$$\operatorname{Cov}(\mathbf{Y} - A\mathbf{X}, \mathbf{X}) = \operatorname{E}\left[(\mathbf{Y} - A\mathbf{X})\mathbf{X}^T\right] = \mathbf{0}_n. \tag{17}$$

b) Let B be any $n \times n$ matrix. Write

$$(\mathbf{Y} - B\mathbf{X})(\mathbf{Y} - B\mathbf{X})^T = ((\mathbf{Y} - A\mathbf{X}) + (A - B)\mathbf{X})(\mathbf{Y} - A\mathbf{X}) + (A - B)\mathbf{X})^T$$

where A is given by (16). Then we get by a direct expansion that

$$(\mathbf{Y} - B\mathbf{X})(\mathbf{Y} - B\mathbf{X})^T = (\mathbf{Y} - A\mathbf{X})(\mathbf{Y} - A\mathbf{X})^T$$

$$-(A-B)\mathbf{X}(\mathbf{Y}-A\mathbf{X})^T-(\mathbf{Y}-A\mathbf{X})((A-B)\mathbf{X})^T+(A-B)\mathbf{X}\left((A-B)\mathbf{X}\right)^T.$$

(you need not re-check this, in case you think this is correct). Then show that

$$\begin{aligned}\mathrm{E}\left[(\mathbf{Y}-B\mathbf{X})\,(\mathbf{Y}-B\mathbf{X})^T\right] &= \mathrm{E}\left[(\mathbf{Y}-A\mathbf{X})(\mathbf{Y}-A\mathbf{X})^T\right]\\ &\quad+\mathrm{E}\left[(A-B)\mathbf{X}\left((A-B)\mathbf{X}\right)^T\right].\end{aligned}\tag{18}$$

Aid: You will need (9) in the form

$$\mathrm{Cov}((A-B)\mathbf{X},(\mathbf{Y}-A\mathbf{X}))=\mathrm{Cov}(\mathbf{Y}-A\mathbf{X}),(A-B)\mathbf{X}))^T.$$

and certain rules of covariance matrices of linear transformations.

c) Now show that

$$\mathrm{E}\parallel\mathbf{Y}-B\mathbf{X}\parallel^2=\mathrm{E}\parallel\mathbf{Y}-A\mathbf{X}\parallel^2+\mathrm{E}\parallel(A-B)\mathbf{X}\parallel^2$$

and draw the conclusion that A as in (16) gives the linear minimal mean square estimator $A\mathbf{X}$.

Problem 3.5.

We continue with Minimal Linear Mean Square Estimation. $\mathbf{X}$ and $\mathbf{Y}$ are $n\times 1$ random vectors with zero expectation vectors. We have the cross covariance matrix $\mathrm{Cov}(\mathbf{Y},\mathbf{X})=E\left[\mathbf{Y}\mathbf{X}^T\right]$ and the covariance matrix $\mathrm{C}_{\mathbf{X}}$, where we assume that $\mathrm{C}_{\mathbf{X}}^{-1}$ exists. By the preceding assignment above we have that

$$A=\mathrm{Cov}(\mathbf{Y},\mathbf{X})\mathrm{C}_{\mathbf{X}}^{-1},\tag{19}$$

minimizes the mean square error of estimating $\mathbf{Y}$ by a linear map of $\mathbf{X}$.

Find that the **Linear Minimal Mean Square Error** (LMME) is

$$\mathrm{LMME}=\mathrm{E}\parallel\mathbf{Y}-A\mathbf{X}\parallel^2=\mathrm{Tr}\left[C_{\mathbf{Y}}-\mathrm{Cov}(\mathbf{Y},\mathbf{X})\mathrm{C}_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X},\mathbf{Y})\right].\tag{20}$$

Aid: Start with

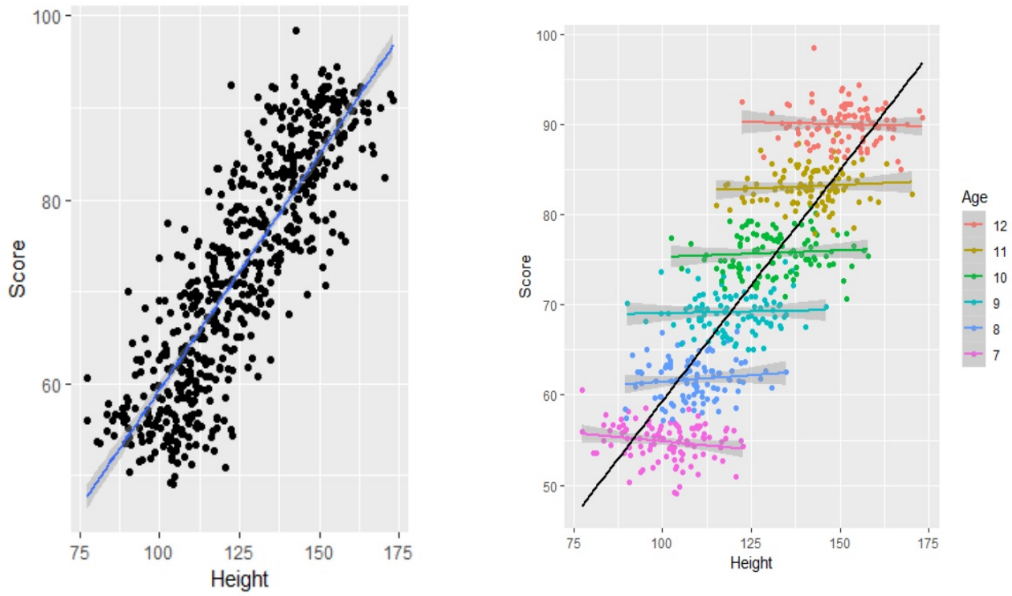
$$\mathrm{E}\parallel\mathbf{Y}-A\mathbf{X}\parallel^2=\mathrm{E}\left[(\mathbf{Y}-A\mathbf{X})^T(\mathbf{Y}-A\mathbf{X})\right]=\mathrm{Tr}\,\mathrm{E}\left[(\mathbf{Y}-A\mathbf{X})\,(\mathbf{Y}-A\mathbf{X})^T\right],$$

expand, compute the expectations and use the rules for traces in section 9.4.
A Solution to this Problem is found in Section 9.16.

4 Linear Regression

4.1 Simple Linear Regression

Problem 4.1.



In the left hand field you see a plot of a training set with predictor x =body height (in feet and inches) and y = score in mathematics exams in an US elementary school and a least squares regression line fitted to it. One might conclude that taller students perform better in mathematics. In the right hand field we see the same data and regression line after an additional analysis. Explain what is found now! What is the phenomenon known as?

Problem 4.2.

In (109) we have the LSE of the regression coefficient as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})\, x_i}{\sum_{i=1}^{n} (x_i - \overline{x})\, x_i}.$$

Rewrite this as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})\, (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

by means of the rules in section 9.8.

Problem 4.3.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the LSE of β_0 and β_1 in simple linear regression. β_0^*, β_1^* denote the true poulation values of the intercept and the regression coefficient.

a) Show that

$$\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{-\bar{x}\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{21}$$

What does this mean in terms of the estimated regression line?
Aid (1): the definition of Cov is

$$\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = E\left[\left(\hat{\beta}_0 - \beta_0^*\right)\left(\hat{\beta}_1 - \beta_1^*\right)\right]$$

Aid (2): You can use the means and variances in (110) even if we are not assuming normal regression.

b) Let us assume normal linear regression. What is the joint distribution of bivariate random variable

$$\left(\hat{\beta}_0, \hat{\beta}_1\right).$$

Recapitulate explicitly the mean vector and the covariance matrix.

Problem 4.4.

We consider simple linear regression in the equivalent centered form:

$$Y_i = \alpha + \beta_1 (x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, n \tag{22}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and

$$\alpha = \beta_0 + \beta_1 \bar{x}. \tag{23}$$

In the matrix form this involves

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X_c = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \\ 1 & x_n - \bar{x} \end{pmatrix}, \boldsymbol{\beta}_c = \begin{pmatrix} \alpha \\ \beta_1 \end{pmatrix}.$$

The least square criterion (cost function) to be minimized is

$$Q(\boldsymbol{\beta}_c) = \|\mathbf{y} - X_c \boldsymbol{\beta}_c\|^2.$$

Hence we know that

$$\hat{\boldsymbol{\beta}}_c = (X_c^T X_c)^{-1} X_c^T \mathbf{y}.$$

a) Show by a direct computation that

$$X_c^T X_c = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{pmatrix}.$$

b) It is now clear how to find $(X_c^T X_c)^{-1}$, if we assume that not all x_i are equal. Show under this assumption that the LSE of the parameters of the centered model are

$$\hat{\boldsymbol{\beta}}_c = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \frac{\sum_{i=1}^{n}(x_i - \bar{x}) y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{pmatrix}. \tag{24}$$

c) Derive the predictor equation for $\widehat{y}_c$ in the centered simple linear regression. Is this equivalent to the predictor $\widehat{y}$ in the uncentered model?

Problem 4.5.

Let $\hat{\alpha}$ and $\hat{\beta}_1$ be given as in (24).

a) Show that

$$\operatorname{Cov}\left(\hat{\alpha}, \hat{\beta}_1\right) = 0 \tag{25}$$

b) Suppose that $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ in (22) is a normal random vector, i.e.,

$$\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2 \mathbb{I}_n\right). \tag{26}$$

What does (25) imply in this case?

Problem 4.6.

Sometimes, with regards to content, it can be reasonable to assume that the regression line passes through the origin ($\beta_0 = 0$). The corresponding regression model is

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

with the following assumptions:

$$\begin{aligned} E[\varepsilon_i] &= 0, \quad i = 1, \ldots, n \\ \operatorname{Var}(\varepsilon_i) &= \sigma^2, \quad i = 1, \ldots, n \\ \{\varepsilon_i \mid i = 1, \ldots, n\} &\quad \text{stochastically independent} \\ \varepsilon_i &\sim N(0, \sigma^2), \quad i = 1, \ldots, n \end{aligned}$$

a) Derive the LS-estimator $\widehat{\beta}_1$ for β_1. Is $\widehat{\beta}_1$ also the ML-estimator?

b) Show that the residuals in general do not sum up to zero. However, why does this hold in the linear regression model including the intercept β_0? Explain the difference.

Problem 4.7.

Suppose that the components of the random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ are independent with means $= 0$, but that

$$C_{\boldsymbol{\varepsilon}} = \sigma^2 \begin{pmatrix} \frac{1}{w_1} & 0 & 0 & \ldots & 0 \\ 0 & \frac{1}{w_2} & 0 & \ldots & 0 \\ 0 & \ddots & \vdots & \ldots & 0 \\ 0 & 0 & 0 & \ldots & \frac{1}{w_n} \end{pmatrix}. \tag{27}$$

where the $w_i = w(x_i)$ are the values of some known positive weight function evaluated at the regressor values.

We consider simple theoretic linear regression in a centered form:

$$E[Y_i] = \beta_0 + \beta_1 (x_i - \bar{x}(w)), \quad i = 1, 2, \ldots, n,$$

where

$$\bar{x}(w) = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

is the weighted mean. We minimize the weighted LS-criterion

$$Q_W(\beta_0, \beta_1) = \sum_{1=1}^{n} w_i \left(y_i - (\beta_0 + \beta_1 (x_i - \bar{x}(w))\right)^2.$$

a) Check first that

$$\frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}(w))}{\sum_{i=1}^{n} w_i} = 0. \tag{28}$$

This can be useful in b), c) and/or d) below. You are allowed to use (28) there, even if you have failed to establish (28).

b) Check that the weighted LSE is

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \\ \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}(w)) y_i}{\sum_{i=1}^{n} w_i (x_i - \bar{x}(w))^2} \end{pmatrix}. \tag{29}$$

c) Show that

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}(w)) Y_i}{\sum_{i=1}^{n} w_i (x_i - \bar{x}(w))^2}$$

are unbiased.

d) Show that

$$\mathrm{Var}[\hat{\beta}_0] = \frac{\sigma^2}{\sum_{i=1}^{n} w_i}$$

and determine in similar manner $\mathrm{Var}[\hat{\beta}_1]$. Remember to justify your calculations.

Problem 4.8.

Suppose that the components of the random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ are independent with means $= 0$, but that

$$C_{\boldsymbol{\varepsilon}} = \sigma^2 \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}, \tag{30}$$

where $\lambda_i > 0$. We observe outcomes $y_1, \ldots, y_n$ of the true model (a constant with noise)

$$\mathbf{Y} = \mathbf{1}_n \beta_0 + \boldsymbol{\varepsilon}.$$

Now we minimize the weighted least squares criterion

$$Q_{\boldsymbol{\lambda}}(\beta_0) := \sum_{i=1}^{n} \frac{1}{\lambda_i} (y_i - \beta_0)^2 \tag{31}$$

as a function of β_0.

a) Check that the weighted LSE is

$$\hat{\beta}_0 = \frac{1}{\sum_{i=1}^{n} \frac{1}{\lambda_i}} \sum_{i=1}^{n} \frac{1}{\lambda_i} y_i. \tag{32}$$

b) Show that

$$\hat{\beta}_0 = \frac{1}{\sum_{i=1}^{n} \frac{1}{\lambda_i}} \sum_{i=1}^{n} \frac{1}{\lambda_i} Y_i$$

is unbiased.

c) Show that

$$\operatorname{Var}\left[\hat{\beta}_0\right] = \frac{\sigma^2}{\sum_{i=1}^{n} \frac{1}{\lambda_i}}.$$

Remember to justify your calculations.

Problem 4.9.

The phenomenon of fatigue was observed by engineers in metal materials in the mid-19th century, due to the shortcomings of the railway wagons after a short period of work. In the 1850s, August Wöhler, a railway engineer, conducted the first experimental fatigue program, testing wagon shafts for failure under alternating stress. The developed loads were recorded together with the number of rotations until failure, so it was possible to formulate the first S-N diagram (S = Stress, N = Number of Cycles) or Wöhler-curve. This assignment deals with what is called the statistical recalibration of the Wöhler curve. Let

$$Y = \text{Number of life cycles to failure}, \quad X = \text{stress}.$$

The linear part of the curve is stated in log-log terms as follows. We assume that Y, given $X = x$, has the lognormal distribution

$$(Y|X = x) \sim \text{LN}\left(\beta_0 + \beta_1 \ln(x), \sigma^2\right)$$

where x is greater than some $x_0 > 0$. There is a recorded set of data, n pairs of stress and life-cycles, $D_{tr} = \{x_i, y_i\}_{i=1}^{n}$.

a) Find simple linear regression equations for this training set, where β_0 is the intercept and β_1 is the slope.

b) Write down the LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$.

c) What are the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$?

e) What is now $\widehat{E}\,[Y|X = x]$, your estimate of the expected number of life cycles given the stress x?

References:

1. Barbosa, Joelton Fonseca and Correia, José AFO and Freire Junior, RCS and Zhu, Shun-Peng and De Jesus, Abílio MP: Probabilistic SN fields based on statistical distributions applied to metallic and composite materials: State of the art. *Advances in Mechanical Engineering*, 11, 8, 2019.

2. A positive random variable X is log-normally distributed, written as

$$X \sim \mathrm{LN}(\mu, \sigma^2),$$

if the natural logarithm of X is normally distributed with mean μ and variance σ^2, i.e.,

$$\ln(X) \sim N(\mu, \sigma^2)$$

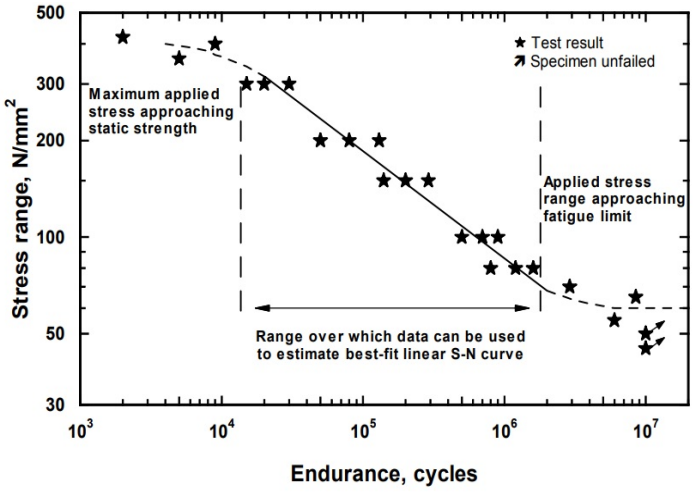


Fig.1 Typical fatigue endurance test data illustrating deviations from linear S-N curve.

4.2 Multiple Linear Regression

In every problem in this section and elsewhere, unless explicitly stated otherwise, it is assumed that the design matrix X has full column rank $= k + 1$ $(< n)$.

Problem 4.10.

In the OLS model,

$$\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}} = H\mathbf{Y}$$

is the predictor vector, where H is the hat matrix. Show that

$$\mathrm{Cov}\left(\mathbf{Y}, \mathbf{Y} - \hat{\mathbf{Y}}\right) = \sigma^2 \left(\mathbb{I}_n - H\right).$$

Aid: Recall (10) and (12).

Problem 4.11.

In the OLS model,

$$\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}} = H\mathbf{Y}$$

is the predictor vector, where H is the hat matrix. Show that

$$\mathrm{Cov}\left(\hat{\mathbf{Y}}, \mathbf{Y} - \hat{\mathbf{Y}}\right) = \mathbf{0}_n.$$

Aid: Recall (10) and (12).

Problem 4.12.

We have the OLS

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}},$$

and

$$\mathbf{e}_{LSE} = \mathbf{y} - X\widehat{\boldsymbol{\beta}} = \mathbf{y} - H\mathbf{y} = \mathbf{y} - \hat{\mathbf{y}}.$$

Explain the differences of concept in the OLS model between $\boldsymbol{\varepsilon}$, $\widehat{\boldsymbol{\varepsilon}}$ and $\mathbf{e}_{LSE}$.

Problem 4.13.

a) The hat matrix H is defined by

$$H = X\left(X^T X\right)^{-1} X^T.$$

Check that H is a projection matrix, see section 9.3.

b) We have

$$\mathbf{e}_{LSE} = \mathbf{y} - H\mathbf{y}$$

Show now that

$$\mathbf{e}_{LSE}^T H\mathbf{y} = 0.$$

Aid: You may need the idempotency of H.

c) Why is H called the hat matrix ?
Answer: It puts a hat on $\mathbf{y}$, i.e.,

$$\hat{\mathbf{y}} = H\mathbf{y}.$$

d) Show that

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{e}_{LSE}\|^2. \tag{33}$$

How can this be interpreted geometrically?

Problem 4.14.

A training set of observed responses $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and the corresponding $n \times (k+1)$ data matrix X are available from some source with the true model

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}.$$

Here $\boldsymbol{\varepsilon}$ has he mean vector $= \mathbf{0}_n$ and covariance matrix $\sigma^2\mathbb{I}_n$. We fit a multiple least squares model to the training set and obtain the LSE $\hat{\boldsymbol{\beta}}$.

The $m \times 1$ random vector $\mathbf{Y}_o$ is the response vector given by

$$\mathbf{Y}_o = X_o\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}_o$$

where $\boldsymbol{\varepsilon}_0$ has mean vector $= \mathbf{0}_m$ and covariance matrix $\sigma^2\mathbb{I}_m$. $\boldsymbol{\varepsilon}_0$ is independent of $\boldsymbol{\varepsilon}$. X_o is a $m \times (k+1)$ matrix of full column rank.

Before observing $\mathbf{Y}_o$, we want to predict $\mathbf{Y}_o$ by the predictor

$$\widehat{\mathbf{Y}}_o = X_o\hat{\boldsymbol{\beta}}.$$

a) Show that $E\left[\mathbf{Y}_o - \widehat{\mathbf{Y}}_o\right] = \mathbf{0}_m$.

b) Find the covariance matrix $C_{\boldsymbol{\varepsilon}_{pr}}$ of the prediction error $\boldsymbol{\varepsilon}_{pr} := \mathbf{Y}_o - \widehat{\mathbf{Y}}_o$.
Answer : $C_{\boldsymbol{\varepsilon}_{pr}} = \sigma^2\left(X_o(X^TX)^{-1}X_o^T + \mathbb{I}_m\right)$.
Aid: You can have use of (54) even if no normal distribution is involved here.

Problem 4.15.

We consider the OLS model with the following generalizations:

1) Correlated noise: The covariance matrix $C_{\boldsymbol{\varepsilon}}$ of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is

$$C_{\boldsymbol{\varepsilon}} = \sigma^2\mathbb{C}, \tag{34}$$

where $\mathbb{C}$ is a known positive definite matrix.

2) Neither σ^2 nor $\mathbb{C}$ depend on X.

Questions:

a) Introduce a new random vector

$$\mathbf{Z} := \mathbb{C}^{-1/2}\mathbf{Y}$$

Why is this possible? Quote briefly some relevant facts from linear algebra.

b) Show that we have the multiple regression

$$\mathbf{Z} = \left(\mathbb{C}^{-1/2}X\right)\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{\dagger}, \tag{35}$$

where $\boldsymbol{\varepsilon}^{\dagger}$ is a random vector satisfying $E\left[\boldsymbol{\varepsilon}^{\dagger}\right] = \mathbf{0}_n$, $C_{\boldsymbol{\varepsilon}^{\dagger}} = \sigma^2\mathbb{I}_n$.

c) The response data in your training set are now $z_i = C^{-1/2}y_i$, $i = 1, \ldots, n$,

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}.$$

Show (justify your solution carefully) now that LSE using (35) is

$$\hat{\boldsymbol{\beta}} = \left(X^T\mathbb{C}^{-1}X\right)^{-1}X^T\mathbb{C}^{-1}\mathbf{y}. \tag{36}$$

d) The covariance matrix in (30) is a special case of (34). Check that you get (32) from (36).

Problem 4.16.

In ridge regression, the ordinary LSE is replaced by $\widehat{\boldsymbol{\beta}}_{\lambda}$ defined by

$$\widehat{\boldsymbol{\beta}}_{\lambda} := \text{argmin}_{\boldsymbol{\beta}}\left(\|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2\right).$$

Using the rules of gradients for vector and matrix expressions found in Appendix 9.15, show that ($p = k + 1$)

$$\widehat{\boldsymbol{\beta}}_{\lambda} = \left(X^TX + \lambda\mathbb{I}_p\right)^{-1}X^T\mathbf{y}. \tag{37}$$

Whenever relevant, refer to the number of the rule you are evoking.

Problem 4.17.

The k regressors of an ordinary LS model as defined in section 9.10

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{38}$$

are **rescaled** as follows. The ith row in X is $\mathbf{x}_i^{\mathsf{T}} = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$ and the rescaled row is $\mathbf{z}_i^{\mathsf{T}}$ $= (1, c_1x_{i1}, c_2x_{i2}, \ldots, c_kx_{ik})$, where $c_i \neq 0$ for $i = 1, 2, \ldots, k.$. Let us define the $(k+1) \times (k+1)$ diagonal matrix

$$D = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & c_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_k \end{pmatrix}.$$

Then it follows that

$$Z = \begin{pmatrix} \mathbf{z}_1^\mathsf{T} \\ \mathbf{z}_2^\mathsf{T} \\ \vdots \\ \mathbf{z}_n^\mathsf{T} \end{pmatrix} = XD. \tag{39}$$

Next you regress the response data $\mathbf{y} = (y_1, \ldots, y_n)^T$ on Z and obtain

$$\widehat{\boldsymbol{\beta}}_z = \left(Z^T Z\right)^{-1} Z^T \mathbf{y} \tag{40}$$

a) Show that

$$\widehat{\boldsymbol{\beta}}_z = D^{-1}\widehat{\boldsymbol{\beta}},$$

where $\widehat{\boldsymbol{\beta}}$ is the LSE for regressing the response data $\mathbf{y} = (y_1, \ldots, y_n)^T$ on X.
Aid: Use (39) in (40) and simplify.

b) Show that the hat predictor is invariant w.r.t. the rescaling above in the sense that

$$\widehat{\mathbf{y}}_z = \widehat{\mathbf{y}}.$$

Here $\widehat{\mathbf{y}}_z$ is the predictor of $\mathbf{y}$ based on $\widehat{\boldsymbol{\beta}}_z$ and $\widehat{\mathbf{y}}$ is the predictor of $\mathbf{y}$ based on $\widehat{\boldsymbol{\beta}}$.

Problem 4.18.

$$\mathbf{e}_{LSE} = \mathbf{y} - X\widehat{\boldsymbol{\beta}} = \mathbf{y} - H\mathbf{y} = \mathbf{y} - \widehat{\mathbf{y}}$$

is the vector of LSE residuals.

$$\mathbf{e}_{LSE} = (\widehat{e}_1, \widehat{e}_2, \ldots, \widehat{e}_n)^T.$$

a) Show that $X^T\mathbf{e}_{LSE} = \mathbf{0}_k$.
When you look at the scalar product of the first row in X^T and $\mathbf{e}_{LSE}$ this means

$$\sum_{i=1}^{n} \widehat{e}_i = 0. \tag{41}$$

b) Show now that

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{y}_i = \bar{y}. \tag{42}$$

c) At a certain moment in the lectures it was obtained that

$$\mathbf{y}^T C_{ce} C_{ce} \mathbf{y} = \widehat{\mathbf{y}}^T C_{ce} \widehat{\mathbf{y}} + \widehat{\mathbf{y}}^T \mathbf{e}_{LSE} + \mathbf{e}_{LSE}^T \widehat{\mathbf{y}} + \mathbf{e}_{LSE}^T \mathbf{e}_{LSE}.$$

We know that $\mathbf{y}^T C_{ce} C_{ce} \mathbf{y} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_{\mathrm{T}}$ by (7) and section 9.14. Show no that

$$SS_{\mathrm{R}} + SS_{\mathrm{Res}} = \widehat{\mathbf{y}}^T C_{ce} \widehat{\mathbf{y}} + \widehat{\mathbf{y}}^T \mathbf{e}_{LSE} + \mathbf{e}_{LSE}^T \widehat{\mathbf{y}} + \mathbf{e}_{LSE}^T \mathbf{e}_{LSE}.$$

Thus we have established the identity (117).

4.3 Estimability and the Gauss-Markov Theorem

The Gauss-Markov theorem states that LSE is best linear unbiased estimate of the regresssion coefficients. In this subsection the Gauss-Markov theorem is studied by an approach different from the one applied in Lecture 4 slides and in MVP. The material below is found in Chapter 2.8 of Julian R. Faraway: *Practical regression with R,* 2002, which is downloadable from canvas via the page SF2930 Course Plan. The proof is based on the notion of estimable linear combinations.

Problem 4.19.

An ordinary LS model as defined in section 9.10

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{43}$$

where $\boldsymbol{\beta}$ is $(k+1) \times 1$. We let $\boldsymbol{c}$ be an arbitrary $(k+1) \times 1$ vector in $\mathbb{R}^{k+1}$. Next we introduce the $1 \times n$ vector $\boldsymbol{a}^T$ by

$$\boldsymbol{a}^T := \boldsymbol{c}^T (X^T X)^{-1} X^T. \tag{44}$$

a) Show that

$$\mathrm{E}\left[\boldsymbol{a}^T \mathbf{Y}\right] = \boldsymbol{c}^T \boldsymbol{\beta}. \tag{45}$$

One says that $\boldsymbol{c}^T\boldsymbol{\beta}$ is **estimable**, since there is an unbiased estimator $\boldsymbol{a}^T\mathbf{Y}$ of it.

b) Check that we also have

$$\mathrm{E}\left[\boldsymbol{a}^T \mathbf{Y}\right] = \boldsymbol{a}^T X \boldsymbol{\beta}. \tag{46}$$

Problem 4.20.

Here you need the notion of the column space or range $\mathcal{R}(X)$ of a matrix X, as presented in the Appendix 9.2.

a) Verify that if (45) and (46) are valid for every $\boldsymbol{\beta}$, then

$$\boldsymbol{c} \in \mathcal{R}\left(X^T\right)$$

b) X is an $n \times (k+1)$ matrix with full column rank $= k+1$. Show now that

$$\boldsymbol{c} \in \mathcal{R}\left(X^T X\right). \tag{47}$$

Aid: In other words, you are through with this, once you have verified that

$$\mathcal{R}\left(X^T\right) \subset \mathcal{R}\left(X^T X\right).$$

This is the second part of the proof of the statement in Proposition 9.1. We give a stepwise guide for this.

1. Take any $\mathbf{w} \in \mathcal{R}\left(X^T\right)$. Thus there exists a $\mathbf{b} \in \mathbb{R}^n$ such that

$$\mathbf{w} = X^T\mathbf{b}.$$

Define by use of the hat matrix the projected vector $\widehat{\mathbf{b}} \in \mathcal{R}(X)$ by

$$\widehat{\mathbf{b}} = H\mathbf{b}.$$

Check now that

$$X^T\widehat{\mathbf{b}} = X^T\mathbf{b}. \tag{48}$$

2. Since $\widehat{\mathbf{b}} \in \mathcal{R}(X)$, there exists $\mathbf{v}$ such that $\widehat{\mathbf{b}} = X\mathbf{v}$. Now compute

$$X^TX\mathbf{v}$$

and draw the desired conclusion.

Problem 4.21.

Now we prove the first part of the Gauss-Markov theorem. We are relying on (47), which you can invoke without a successfull proof thereof. We have seen in the first Problem of this subsection that $\boldsymbol{c}^T\boldsymbol{\beta}$ is estimable, if X has full column rank. Let thus $\boldsymbol{a}^T\mathbf{Y}$ be any unbiased estimator of $\boldsymbol{c}^T\boldsymbol{\beta}$. By the preceding problems, (45) and (46) hold and imply (47). By definition of $\mathcal{R}\left(X^TX\right)$, (47) implies that there exists a $(k+1) \times 1$ vector $\boldsymbol{z}$ such that

$$\boldsymbol{c} = X^TX\boldsymbol{z}. \tag{49}$$

If $\widehat{\boldsymbol{\beta}}$ is the LSE of (43), then (49) implies

$$\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} = \boldsymbol{z}^TX^TX\widehat{\boldsymbol{\beta}} = \boldsymbol{z}^TX^T\mathbf{Y}. \tag{50}$$

Then we get

$$\begin{aligned}\mathrm{Var}\left[\boldsymbol{a}^T\mathbf{Y}\right] &= \mathrm{Var}\left[\boldsymbol{a}^T\mathbf{Y} - \boldsymbol{c}^T\widehat{\boldsymbol{\beta}} + \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right] = \mathrm{Var}\left[\left(\boldsymbol{a}^T\mathbf{Y} - \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right) + \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right] \\ &= \mathrm{Var}\left[\boldsymbol{a}^T\mathbf{Y} - \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right] + \mathrm{Var}\left[\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right] + 2\mathrm{Cov}\left(\boldsymbol{a}^T\mathbf{Y} - \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}, \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right).\end{aligned}$$

where we used a familiar formula for the variance of a sum of two random variables from the first course.

a) Use to (50) to write

$$\mathrm{Cov}\left(\boldsymbol{a}^T\mathbf{Y} - \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}, \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right) = \mathrm{Cov}\left(\boldsymbol{a}^T\mathbf{Y} - \boldsymbol{z}^TX^T\mathbf{Y}, \boldsymbol{z}^TX^T\mathbf{Y}\right)$$

Then show that

$$\mathrm{Cov}\left(\boldsymbol{a}^T\mathbf{Y} - \boldsymbol{z}^TX^T\mathbf{Y}, \boldsymbol{z}^TX^T\mathbf{Y}\right) = 0.$$

Aid: You may need the standard rule of covariance computation (11) and (47).

b) Hence we have

$$\operatorname{Var}\left[\boldsymbol{a}^T\mathbf{Y}\right] = \operatorname{Var}\left[\boldsymbol{a}^T\mathbf{Y} - \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right] + \operatorname{Var}\left[\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right]. \tag{51}$$

Now draw the conclusion in the Gauss-Markov theorem.

c) We need to establish the uniqueness of $\widehat{\boldsymbol{\beta}}$ as the minimum variance unbiased linear estimator. This means that if $\operatorname{Var}\left[\boldsymbol{a}^T\mathbf{Y}\right] = \operatorname{Var}\left[\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right]$, then $\boldsymbol{a}^T\mathbf{Y} = \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}$.

Aid: See what happens in (51), if $\operatorname{Var}\left[\boldsymbol{a}^T\mathbf{Y}\right] = \operatorname{Var}\left[\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}\right]$.

Problem 4.22.

If X has full column rank, $\operatorname{rank} X = k+1 < n$, check that

$$X^+ = (X^TX)^{-1}X^T$$

is the BMP inverse of X by verifying the conditions MP1-MP4 in section 9.7.

Problem 4.23.

In this problem X has full row rank $n < k+1$.

If X has full row rank, $\operatorname{rank} X = n < k+1$, then

$$X^+ = X^T(XX^T)^{-1}$$

is the BMP inverse of X.

a) We know the normal equations:

$$X^TX\boldsymbol{\beta} = X^T\mathbf{y}. \tag{52}$$

For X with full row rank we define

$$\boldsymbol{\beta}^+ = X^+\mathbf{y}.$$

Check now that $\boldsymbol{\beta}^+$ is a solution to (52).

b) We have the BMP predictor

$$\widehat{\mathbf{y}}^+ = X\boldsymbol{\beta}^+ = XX^+\mathbf{y}$$

Show that $\widehat{\mathbf{y}}^+$ is an interpolation of the training set.

Problem 4.24.

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}$$

and

$$\mathrm{E}\left[\boldsymbol{\varepsilon}\right] = \mathbf{0}_n, C_{\boldsymbol{\varepsilon}} = \sigma^2\mathbb{I}_n.$$

$$\widehat{\mathbf{Y}} = H\mathbf{Y}, \widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}}.$$

Let the variance σ^2 be estimated by

$$\widehat{\sigma^2} = \frac{1}{n-k-1}\widehat{\boldsymbol{\varepsilon}}^T\widehat{\boldsymbol{\varepsilon}}.$$

Show that $\widehat{\sigma^2}$ is unbiased.
Aid: Insert $\widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}}$ compute the required expectetion and use the rules for traces in section 9.4.

4.4 Normal Linear Regression

Problem 4.25.

Prove in formal detail:

$$\mathbf{Y} \sim N_n\left(X\boldsymbol{\beta}_*, \sigma^2\mathbb{I}_n\right).$$

Problem 4.26.

The LSE of the regressor parameters is

$$\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}.$$

a) Show that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_* + (X^TX)^{-1}X^T\boldsymbol{\varepsilon}. \tag{53}$$

b) Show that

$$\widehat{\boldsymbol{\beta}} \sim N_{k+1}\left(\boldsymbol{\beta}_*, \sigma^2(X^TX)^{-1}\right).$$

Remember to justify carefully.

Problem 4.27.

$$\widehat{\mathbf{Y}} = H\mathbf{Y},$$

where H is the hat matrix.

a) Show that

$$\widehat{\mathbf{Y}} \sim N_n\left(X\boldsymbol{\beta}_*, \sigma^2H\right). \tag{54}$$

b)

$$\widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}}.$$

Show (with appropriate justification) that

$$\widehat{\boldsymbol{\varepsilon}} \sim N_n\left(\mathbf{0}_n, \sigma^2\left(\mathbb{I}_n - H\right)\right) \tag{55}$$

Problem 4.28.

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right).$$

Thus

$$\mathbf{Y} \sim N_n\left(X\boldsymbol{\beta}, \sigma^2 \mathbb{I}_n\right)$$

and the p.d.f .is

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(\sigma^2 2\pi)^n}} e^{-\frac{1}{2\sigma^2}\|\mathbf{y} - X\boldsymbol{\beta}\|^2}.$$

We have the $-1\cdot$ loglikelihood function

$$l_{\mathbf{y}}\left(\boldsymbol{\beta}, \sigma^2\right) := -\ln f_{\mathbf{Y}}(\mathbf{y}) = \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2 \tag{56}$$

For $j = 0, 1, \ldots, k$ it holds that

$$\frac{\partial}{\partial \beta_j} l_{\mathbf{y}}\left(\boldsymbol{\beta}, \sigma^2\right) = \frac{1}{2\sigma^2}\frac{\partial}{\partial \beta_j} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2 = \frac{1}{2\sigma^2}\frac{\partial}{\partial \beta_j} Q(\boldsymbol{\beta})$$

Hence the maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\beta}}_{ML}$ can be found by solving first w.r.t. $\boldsymbol{\beta}$ the equations

$$\frac{1}{2}\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}_k$$

Since $\mathbf{y}^T X\boldsymbol{\beta} = \boldsymbol{\beta}^T X^T \mathbf{y}$ we get

$$Q(\boldsymbol{\beta}) = \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} - 2\boldsymbol{\beta}^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y}. \tag{57}$$

a) Show now that

$$\frac{1}{2}\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\widehat{\boldsymbol{\beta}}_{ML}} = \mathbf{0}_k \Leftrightarrow X^T X\widehat{\boldsymbol{\beta}}_{ML} = X^T \mathbf{y}$$

Aid: Use the rules in section 9.15.

b) Hence, if X has full column rank, $\widehat{\boldsymbol{\beta}}_{ML}$ equals the LSE $\widehat{\boldsymbol{\beta}}$ and the LSE residuals are

$$\mathbf{e}_{LSE} = \mathbf{y} - X\widehat{\boldsymbol{\beta}}_{ML}.$$

Therefore

$$Q\left(\widehat{\boldsymbol{\beta}}_{ML}\right) = \parallel \mathbf{y} - X\widehat{\boldsymbol{\beta}}_{ML} \parallel^2 = \mathbf{e}_{LSE}^T \mathbf{e}_{LSE}.$$

When we substitute this in (56) we get

$$l_{\mathbf{y}}\left(\widehat{\boldsymbol{\beta}}_{ML}, \sigma^2\right) = \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2}\mathbf{e}_{LSE}^T \mathbf{e}_{LSE}.$$

Find now the MLE $\widehat{\sigma}^2_{ML}$ of σ^2.

c) Test the correctness of your result in b) by verifying that

$$l_{\mathbf{y}}\left(\widehat{\boldsymbol{\beta}}_{ML}, \widehat{\sigma}^2_{ML}\right) = \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\widehat{\sigma}^2_{ML}) + \frac{n}{2} \tag{58}$$

Problem 4.29.

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2\mathbb{I}_n\right).$$

We have used a training data set

$$\mathcal{D}_{tr} = \{\left(y_i, \mathbf{x}_i^T\right)_{i=1}^n\}$$

to find the LSE $\widehat{\boldsymbol{\beta}}$, which determines the hat matrix so that

$$\widehat{\mathbf{Y}} = H\mathbf{Y}, \quad \widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}}.$$

a) Show that

$$\widehat{\boldsymbol{\varepsilon}} = (\mathbb{I}_n - H)\boldsymbol{\varepsilon}. \tag{59}$$

b) Set

$$\mathbf{Z} := \frac{\boldsymbol{\varepsilon}}{\sigma}. \tag{60}$$

Check that $\mathbf{Z} \sim N_n\left(\mathbf{0}_n, \mathbb{I}_n\right)$, i.e., $\mathbf{Z}$ is a standard normal vector. Check that now that

$$\frac{\widehat{\boldsymbol{\varepsilon}}^T\widehat{\boldsymbol{\varepsilon}}}{\sigma^2} = \mathbf{Z}^T(\mathbb{I}_n - H)\mathbf{Z}. \tag{61}$$

Justify in detail. *Aid:* The formula in (59) may be useful here.

c) The unbiased estimate of variance σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n-k-1}\widehat{\boldsymbol{\varepsilon}}^T\widehat{\boldsymbol{\varepsilon}}.$$

Show now that

$$(n-k-1)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2\left(n-k-1\right). \tag{62}$$

Aid: Check proposition 9.3 and the section 9.4 for rules on traces.

5 Properties of the Hat Matrix and Diagnostics of Regression Models

Let

$$X_R := \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} \tag{63}$$

so that the design matrix is partitioned as

$$X = (\mathbf{1}_n \quad X_R)\,.$$

Note that

$$C_{ce}X_R = \left(X_R - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T X_R\right) = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nk} - \bar{x}_k \end{pmatrix},$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$, $j = 1, \ldots, k$. We set

$$X_c := \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nk} - \bar{x}_k \end{pmatrix} \tag{64}$$

This is the matrix of centered regressor/covariate values.

> The exam questions in this subsection are based on the material in the slides of Lecture 6. We provide a summary for convenience of reference.

Multiple regression in component form is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n, \quad n > k + 1.$$

The centered model is

$$Y_i = \alpha + \beta_1 (x_{i1} - \bar{x}_1) + \cdots + \beta_k (x_{ik} - \bar{x}_k) + \varepsilon_i, \quad i = 1, \ldots, n \tag{65}$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$, $j = 1, \ldots, k$, and

$$\alpha = \beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_k \bar{x}_k. \tag{66}$$

Now we can write the equations in (65) in matrix form using (64) as

$$\mathbf{Y} = (\mathbf{1}_n, X_c) \begin{pmatrix} \alpha \\ \boldsymbol{\beta}_R \end{pmatrix} + \boldsymbol{\varepsilon}. \tag{67}$$

where

$$\boldsymbol{\beta}_R = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

and X_c is given in (64).

It is shown in Lecture 6 that the LSE of the centered model are

$$\begin{aligned} \hat{\alpha} &= \bar{y} \\ \hat{\boldsymbol{\beta}}_R &= \left(X_c^T X_c\right)^{-1} X_c^T \mathbf{y} \end{aligned} \tag{68}$$

Next, advance from Lecture 6 to the exam questions.

First write (67) as

$$\mathbf{Y} = \mathbf{1}_n \alpha + X_c \boldsymbol{\beta}_R + \boldsymbol{\varepsilon}.$$

Then the predictor (as a function of the training data) is

$$\hat{\mathbf{y}} = \mathbf{1}_n \hat{\alpha} + X_c \hat{\boldsymbol{\beta}}_R$$

Hence the corrresponding hat matrix is

$$H_c = X_c \left(X_c^T X_c\right)^{-1} X_c^T \tag{69}$$

Problem 5.1.

In view of (68) we have

$$\hat{\mathbf{y}} = \mathbf{1}_n \bar{y} + X_c \hat{\boldsymbol{\beta}}_R$$

a) Do some simple steps of matrix calculus to derive the identity

$$\hat{\mathbf{y}} = \left(\frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T + H_c\right)\mathbf{y}. \tag{70}$$

(*Aid*: $\bar{y} = \frac{1}{n}\mathbf{1}_n^T \mathbf{y}$).

b) Show by (70) that the hat matrix of the non-centered model is

$$H = \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T + X_c \left(X_c^T X_c\right)^{-1} X_c^T. \tag{71}$$

c) Show that the elements h_{ii} on the main diagonal of H are given as

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})^T \left(X_c^T X_c\right)^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}). \tag{72}$$

where

$$\mathbf{x}_i^T = (x_{i1}, x_{i2} \ldots, x_{ik})$$

is the ith row in X_R and

$$\bar{\mathbf{x}}^T = (\bar{x}_1, \bar{x}_2 \ldots, \bar{x}_k)$$

is the $1 \times k$ vector of means of columns in X_R.
Aid: The equation (72) follows from (71), when one takes $A = X_c$ and $B = \left(X_c^T X_c\right)^{-1}$ in (4).

d) Show that $h_{ii} \geq \frac{1}{n}$.

Problem 5.2.

The hat matrix H plays an important part in the diagnostics of linear regression. This is due to the fact that it determines, as has been shown, the variances and covariances of $\widehat{\mathbf{Y}}$ and the residuals $\widehat{\boldsymbol{\varepsilon}}$ (recall (54) and (55).

a)

$$\widehat{\mathbf{y}} = \begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{pmatrix} = H \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

We have

$$\widehat{y}_i = \mathfrak{E}_i^T \widehat{\mathbf{y}}.$$

Use (2) to write the right hand side of

$$\mathfrak{E}_i^T \widehat{\mathbf{y}} = \mathfrak{E}_i^T H \mathbf{y}$$

as a finite sum (display the sum) and compute by means of this sum the partial derivative

$$\frac{\partial \widehat{y}_i}{\partial y_i}$$

and interpret h_{ii} by this.

b) Show that (6.2)

$$h_{ii} = \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i$$

is the generic element in the main diagonal of the hat matrix H. *Aid:* The expression (4) may be useful.

c) In (72) we have shown that the elements h_{ii} on the main diagonal of H can also be written as

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})^T \left(X_c^T X_c\right)^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}). \tag{73}$$

In this expression the vector of means $\bar{\mathbf{x}}$ can be regarded as a centroid. Check that for the simple centered linear regression model the elements of the hat matrix H_c are

$$h_{ij}(c) = \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \quad \text{and} \quad h_{ii}(c) = \frac{(x_i - \bar{x})^2}{S_{xx}}.$$

Here you start with

$$X_c := \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}, \tag{74}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Check now using (73) that for the simple linear regression

$$\frac{1}{n} \leq h_{ii} < 1.$$

Problem 5.3.

This exam item deals with hidden extrapolation on pp. 107− 110 of MPV. On page 110 we read the following:

outside the RVH, extrapolation is required.

The diagonal elements h_{ii} of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are useful in detecting **hidden extrapolation**. The values of h_{ii} depend both on the Euclidean distance of the point $\mathbf{x}_i$ from the centroid and on the density of the points in the RVH. In general, the point that has the largest value of h_{ii}, say $h_{\max}$, will lie on the boundary of the RVH in a region of the x space where the density of the observations is relatively low. The set of points $\mathbf{x}$ (not necessarily data points used to fit the model) that satisfy

$$\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} \leq h_{\max}$$

is an ellipsoid enclosing all points inside the RVH (see Cook [1979] and Weisberg [1985]). Thus, if we are interested in prediction or estimation at the point $\mathbf{x}_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$, the location of that point relative to the RVH is reflected by

$$h_{00} = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$$

a) We have used a training set to find the LSE $\widehat{\boldsymbol{\beta}}$ regarded as n outcomes of the true model

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}$$

We are predicting the value of the response Y at $\mathbf{x}_o^T = (1, x_{o1} \ldots, x_{ok})$ by

$$\hat{Y} = \mathbf{x}_0^T \widehat{\boldsymbol{\beta}},$$

when

$$Y = \mathbf{x}_0^T \boldsymbol{\beta}_* + \varepsilon.$$

How does $h_{\max}$ in the quote of MPV above reflect the mean square error of the the random variable $Y - \hat{Y}$?

b) Explain the problem of hidden extrapolation in predicting new responses and estimating the mean response at given point $\mathbf{x}_0$ in the multiple linear regression. Justify your explanations by sketching the graph and explain how to detect this problem by using the properties of the hat matrix $H = X(X^\top X)^{-1}X^\top$.

Problem 5.4.

Leverage is a technical term for the measurement of how far away a $1 \times k$ regressor vector $\mathbf{x}_i^T$ is from the rest of the regressor vectors in X, i.e., far away in $\mathbb{R}^k$. High-leverage points, if any, are outliers with respect to the RVH. This makes the predictor likely to pass close to a high leverage observation. Hence high-leverage points have the potential to cause large changes in the predictors if they are deleted: if this happens, then they are said to be **influential** points. The h_{ii}, i.e., the elements of the main diagonal of H, are used as the measures of leverage of the correspoding vector $\mathbf{x}_i$.

a) The distance related to h_{ii} reflects basically on the squared and sample based Mahalanobis distance. By the squared and sample based Mahalanobis distance between a vector $\mathbf{x}_i$ in finite set of vectors and its centroid we mean the squared Mahalanobis distance in (13), when the sample covariance matrix S in (5) is used in place of Σ. We write

$$d_M^2\left(\mathbf{x}_i, \bar{\mathbf{x}}; S\right) = \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T S^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right).$$

Show that

$$(n-1)\left(h_{ii} - \frac{1}{n}\right) = d_M^2(\mathbf{x}_i, \bar{\mathbf{x}}; S).$$

b) The predicted residual error sum of squares (PRESS) is one of the techniques of cross-validation used in linear regression analysis for detecting influential observations.

The PRESS statistic is

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2.$$

How is this obtained?

c) Use the formula (derived by Sherman-Morrison-Woodbury theorem, which you need not to recapitulate or invoke here)

$$\hat{y}_i = h_{ii} y_i + (1 - h_{ii})\hat{y}_{(i)},$$

where h_{ii} is the element on the main diagonal of H, to rewrite PRESS so that you only need to use predictors from full data linear regression. What important property of h_{ii}s is crucial here?

Problem 5.5.

a) Consider

$$H_o\text{: } \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

and the alternative hypothesis

$$H_1\text{: } \beta_j \neq 0 \text{ for at least one } \beta_j$$

What null hypothesis about multiple linear regression are we testing here ?

b) We assume normal multiple linear regression for the hypothesis testing. To test this we need H_c in the centered model. It has been shown in another problem in this exam generator that, if H_o is true,

$$\frac{\widehat{\varepsilon}^T\widehat{\varepsilon}}{\sigma^2} \sim \chi^2\left(n-k-1\right). \tag{75}$$

This is valid even for the centered model. In the same manner we can show that for the model component of variance that

$$\frac{SS_{\mathrm{R}}}{\sigma^2} = \frac{\widehat{\mathbf{y}}^T C_{ce}\widehat{\mathbf{y}}}{\sigma^2} \sim \chi^2(k). \tag{76}$$

This is valid even for the centered model. It has been found that these two quadratic forms $\widehat{\sigma^2}$ and SS_{R} are independent. It has been shown in another problem that

$S_1 \sim \chi^2\left(d_1\right)$ and $S_2 \sim \chi^2\left(d_2\right)$ are independent. Set

$$V := \frac{S_1/d_1}{S_2/d_2}.$$

What is the distribution of V?

c) Explain how the following ANOVA table for multiple linear regression is used to test the hypotheses above at a significance level α. Remember to state explicitly the test statistic and its distribution, if H_o is true.

Source	df	Sum of Squares	MSS
Regression	k	SS_{R}	SS_{R}/k
Residual	$n-k-1$	SS_{Res}	SS_{Res}/(n-k-1)
Total	$n-1$	SS_{T}	

Source = source of variation, df= degrees of freedom, SS= sum of squares, MSS= mean sum of squares.

6 Choice of Regression Models

Problem 6.1.

Consider two models

$$\mathcal{M}_1 \colon E\left[\mathbf{Y}|X_1\right] = X_1\boldsymbol{\beta}_1 \quad \text{and} \quad \mathcal{M}_2 \colon E\left[\mathbf{Y}|X_2\right] = X_2\boldsymbol{\beta}_2,$$

where X_1 is $n \times (k_1+1)$, X_2 is $n \times (k_2+1)$, $\boldsymbol{\beta}_1$ is $(k_1+1)\times 1$, and $\boldsymbol{\beta}_2$ is $(k_2+1)\times 1$.

We suppose that the models are nested so that $\mathcal{M}_1 \subset \mathcal{M}_2$: more explicitly, $X_2 = [X_1\ \tilde{X}_2]$, where $\tilde{X}_2$ is $n \times (k_2 - k_1)$. In other words, the first (k_1+1) columns of X_2 are equal to those of X_1.

Denote by $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ the least squares estimates of $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Recall that the LSE errors are

$$Q_o^{(i)}(\widehat{\boldsymbol{\beta}}_i) = \|\mathbf{y} - X_i\widehat{\boldsymbol{\beta}}_i\|^2, \quad i = 1, 2.$$

The question is whether the larger, more flexible model $\mathcal{M}_2$ is significantly better than the more restricted model $\mathcal{M}_1$. It has been shown that an F-statistic to decide this question is

$$F_{\mathcal{M}} = \frac{\left(Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)\right)/(k_2-k_1)}{Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)/(n-k_2)}.$$

In this problem we study the relationship between the F-test and likelihood ratio test of the the null hypothesis that the true model lies in $\mathcal{M}_1$. We have the likelihood functions

$$L_{\mathbf{y}}\left(\boldsymbol{\beta}_i, \sigma^2\right) = \frac{1}{\sqrt{(\sigma^2 2\pi)^n}} e^{-\frac{1}{2\sigma^2}\|\mathbf{y} - X_i\boldsymbol{\beta_i}\|^2}, \quad i = 1, 2.$$

The likelihood ratio (LR) statistic is

$$\mathrm{LR} = \frac{\max_{\boldsymbol{\beta}_2, \sigma^2} L_{\mathbf{y}}\left(\boldsymbol{\beta}_2, \sigma^2\right)}{\max_{\boldsymbol{\beta}_1, \sigma^2} L_{\mathbf{y}}\left(\boldsymbol{\beta}_1, \sigma^2\right)}. \tag{77}$$

a) Why does it always hold that $\mathrm{LR} \geq 1$?

b) The MLEs of σ^2 are (you do not need to show this here)

$$\widehat{\sigma}_1^2 = \frac{\sum_{j=1}^n e_{1,j}^2}{n}, \widehat{\sigma}_2^2 = \frac{\sum_{j=1}^n e_{2,j}^2}{n}$$

in the models in $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Here $(e_{i,1}, \ldots, e_{i,n})$ are the least squares residuals w.r.t. model $\mathcal{M}_i$, $i = 1, 2$. Why does it always hold that

$$\widehat{\sigma}_1^2 \geq \widehat{\sigma}_2^2?$$

c) Show that

$$\mathrm{LR} = \left(\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2}\right)^{n/2}.$$

Aid: Use (77) and (58) in an approriate way.

d) Show that $F_{\mathcal{M}}$ is a monotone function of LR.

e) The LR test rejects the null hypothesis that the true model lies in $\mathcal{M}_1$ if $\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} > c$ for a chosen constant c. The F-test rejects the null hypothesis that the true model lies in $\mathcal{M}_1$, if $F_{\mathcal{M}} > F_\alpha(k_2 - k_1, n - k_2)$, the α-percentile. In view of the finding in d) above, what is the relationship between the F-test and the LR test for the choice of nested multiple regression models?

Problem 6.2.

This problem deals with Mallows' C_p criterion for the choice of model dimension in multiple regression. The main statements below are recapitulated without proof in MPV on pp. 334-335. The symbol p in MPV is $p = k + 1$ in this document.

We are faced with multiple regression models $\mathcal{M}_k$, where k is the number of predictors in the model. The intercept is included in every model; hence $\mathcal{M}_0$ is the model with only the intercept. M is the maximum number of possible predictors in a model. The true model for n observations of a response variable Y is OLS as in section 9.10 satisfying

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}_*$ is a $(k_*+1)\times 1$ vector and X is a $n\times(k_*+1)$ matrix. The true number of predictors k_* is unknown. The purpose of developing Mallows' C_p is to estimate $\hat{k}$ on the basis of n samples of Y so that overfitting is avoided.

Fix now any $k \in \{1, \ldots, M\}$. Then

$$\widehat{\mathbf{Y}} = \begin{pmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{pmatrix} = H_k \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \widehat{\boldsymbol{\varepsilon}} = \begin{pmatrix} \widehat{\varepsilon}_1 \\ \vdots \\ \widehat{\varepsilon}_n \end{pmatrix} = \mathbf{Y} - H_k\mathbf{Y}$$

where H_k is the hat matrix for the model $\mathcal{M}_k$. Thus H_k is

$$H_k = X_k \left(X_k^T X_k\right)^{-1} X_k^T,$$

where X_k^T is a $(k+1)\times n$ matrix. Hence $X_k^T X$ is a well-defined $(k+1)\times(k_*+1)$ matrix and $H_k X\boldsymbol{\beta}_*$ is a well-defined $n\times 1$ matrix. $\widehat{\boldsymbol{\varepsilon}}$ is the vector of LSE residuals computed in the model $\mathcal{M}_k$, i.e. $\widehat{\boldsymbol{\varepsilon}}(k)$, but we use a simpler notation for ease of writing.

a) Verify that

$$E\left[\widehat{\mathbf{Y}}\right] = H_k X\boldsymbol{\beta}_*$$

Hence

$$E\left[\widehat{Y}_i\right] \neq E\left[Y_i\right].$$

and the decomposition (80) has a useful content.

b) Show that

$$C_{\widehat{\mathbf{Y}}} = \sigma^2 H_k. \tag{78}$$

c) We have that

$$E\left[\widehat{\boldsymbol{\varepsilon}}\right] = \left(\mathbb{I}_n - H_k\right) X\boldsymbol{\beta}_*.$$

Show now that

$$C_{\widehat{\boldsymbol{\varepsilon}}} = \sigma^2 \left(\mathbb{I}_n - H_k\right). \tag{79}$$

d) Check that

$$E\left[\left(\widehat{Y}_i - E\left[Y_i\right]\right)^2\right] = \left(E\left[\widehat{Y}_i\right] - E\left[Y_i\right]\right)^2 + \mathrm{Var}\left[\widehat{Y}_i\right]. \tag{80}$$

Aid: Start with $E\left[\left(\widehat{Y}_i - E\left[Y_i\right]\right)^2\right] = E\left[\left(\widehat{Y}_i - E\left[\widehat{Y}_i\right] + E\left[\widehat{Y}_i\right] - E\left[Y_i\right]\right)^2\right]$. Next, expand the sum inside the expectation, and then compute the expectation.

e) Show that

$$\sum_{i=1}^{n} \mathrm{Var}\left[\widehat{Y}_i\right] = (k+1)\sigma^2.$$

Aid: Note that the variances $\mathrm{Var}\left[\widehat{Y}_i\right]$ are the diagonal elements of the covariance matrix of $\widehat{\mathbf{Y}}$, note (78), and recall Appendix C in the slides of Lecture 3.

f) Show that

$$E\left[\sum_{i=1}^{n} \widehat{\varepsilon}_i^2\right] = \sum_{i=1}^{n} \left(E\left[\widehat{Y}_i\right] - E\left[Y_i\right]\right)^2 + (n - (k+1))\sigma^2.$$

Aid: Use a well known formula from the first course for $E\left[\widehat{\varepsilon}_i^2\right]$. Then note (79) and use the results on slide 58/79 in Lecture 4 as in b) above.

g) Then show that

$$\begin{aligned}
\Gamma_k &:= \frac{1}{\sigma^2} E\left[\left(\widehat{Y}_i - E\left[Y_i\right]\right)^2\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(E\left[\widehat{Y}_i\right] - E\left[Y_i\right]\right)^2 + \sum_{i=1}^{n}\mathrm{Var}\left[\widehat{Y}_i\right]\right] \\
&= \frac{1}{\sigma^2} E\left[\sum_{i=1}^{n}\widehat{\varepsilon}_i^2\right] - n + 2(k+1).
\end{aligned}$$

Now Mallows replaces the mean with the sum of residual squares to define the (almost) final criterion of choice of model dimension as

$$\mathrm{C}_k := \frac{1}{\sigma^2}\sum_{i=1}^{n}\widehat{\varepsilon}_i(k)^2 - n + 2(k+1)$$

where $\widehat{\varepsilon}_i = \widehat{\varepsilon}_i(k)$ are the LSE residuals computed in the model $\mathcal{M}_k$, $k \in \{1, \ldots, M\}$. How should one estimate σ^2 ?

Problem 6.3.

a) Let X be $n \times (k+1)$ and $\boldsymbol{\beta}_*$ be $(k+1) \times 1$ and

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n, \right)$$
$$\mathbf{Z} = X\boldsymbol{\beta}_* + \boldsymbol{\nu}, \quad \boldsymbol{\nu} \sim N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right).$$

In this $\boldsymbol{\varepsilon}$ and $\boldsymbol{\nu}$ are independent.

We use a training data set on the responses in Y

$$\mathcal{D}_{tr} = \{\left(y_i, \mathbf{x}_i^T\right)_{i=1}^n\}$$

to find the LSE $\widehat{\boldsymbol{\beta}}$. Next we wish to estimate $\mathbf{Z}$ by means of

$$\widehat{\mathbf{Z}} = X\widehat{\boldsymbol{\beta}} = H\mathbf{Y}.$$

Compute the mean square error per sample of $\widehat{\mathbf{Z}}$ as

$$\frac{1}{n} \mathrm{E}\left[\left(\mathbf{Z} - \widehat{\mathbf{Z}}\right)^T \left(\mathbf{Z} - \widehat{\mathbf{Z}}\right)\right] = \sigma^2 \left(1 + \frac{k+1}{n}\right). \tag{81}$$

Aid: Expand, use the properties of H, the rule (11) may be useful, too. Compute the involved traces.

b) The AIC (=Akaike Information Criterion) for model choice in a model family defined by the p.d.f.s $f_{\mathbf{Y}}(y; \theta)$, where $p = \dim(\theta)$ is generally given by

$$\mathrm{AIC}_p = -2 \ln f_{\mathbf{Y}}\left(y; \widehat{\theta}_{\mathrm{MLE}}\right) + 2 \cdot p. \tag{82}$$

The best model in the sense of AIC is found by

$$p_{\mathrm{AIC}} = \mathrm{argmin}_p \left\{-2 \ln f_{\mathbf{Y}}\left(y; \widehat{\theta}_{\mathrm{MLE}}\right) + 2 \cdot p\right\}.$$

The assignment here is to apply AIC in multiple linear regression. Consider a family $\mathbb{M}$ of K normal multiple regression models

$$\mathcal{M}_k \colon E[\mathbf{Y}|X_k] = X_k\boldsymbol{\beta}_k, \quad k = 1, 2, \ldots, K,$$

where X_k is $n \times (k+1)$, $\boldsymbol{\beta}_i$ is $(k+1) \times 1$.

Show that AIC for model choice in $\mathbb{M}$ given training sets $\mathcal{D}_{tr} = \{(y_i)_{i=1}^n, X_k = \left(\mathbf{x}_{(1)}^T, \ldots, \mathbf{x}_{(k)}^T\right)^T\}$, $\mathbf{x}_{(l)}^T = \left(x_i^{(l)}\right)_{i=1}^n$ is

$$\mathrm{AIC}_k = C_n + n \ln\left(\widehat{\sigma}^2_{\mathrm{MLE}}\right) + 2 \cdot (k+1). \tag{83}$$

4.4 where $C_n = n \ln(2\pi) + n$. We do not include the variance σ^2 in the number of parameters, as it is shared by all models.

Aid: You allowed to use, without working out or presenting the details, the relevant results in Problem 4.28 in section 4.4.

c) Some statisticians/ data analysts take (81) as a starting point and define an approximate AIC by

$$\mathrm{AIC}_k^* = \widehat{\sigma}^2_{\mathrm{MLE}} \left(1 + \frac{k+1}{n}\right). \tag{84}$$

Describe concisely how AIC_k^* might incorporate the idea of parsimony, i.e., of optimal model fit and model complexity underlying the Akaike Information Criterion.

d) Give an approximate relation between $\frac{\mathrm{AIC}_k}{n}$ and a certain function of AIC_k^*. *Aid:* Approximation $\ln(1+x) \approx x$.

7 Generalized Linear Regression

In the problems of this section

$$\mathbf{x}^T\boldsymbol{\beta} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

Problem 7.1.

$\epsilon \sim \mathrm{Logistic}(0,1)$, iff its probability density function (pdf) is

$$\sigma'(x) = \frac{d}{dx}\sigma(x) = \frac{e^{-x}}{(1+e^{-x})^2}.$$

a) Verify the primitive function

$$\int \sigma'(t)dt = \frac{1}{1+e^{-x}} + C \tag{85}$$

and determine C. Set

$$\sigma(x) := \frac{1}{1+e^{-x}}.$$

b) Show that

$$P\left(-\varepsilon \leq x\right) = P\left(\varepsilon \leq x\right). \tag{86}$$

You may use (85) even if you have failed to show it.

c) Define

$$Y^* = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon \tag{87}$$

and construct

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Verify that

$$P(Y = 1 \mid \mathbf{x}) = \sigma\left(\mathbf{x}^T\boldsymbol{\beta}\right).$$

You may use (86) even if you have failed to show it.

d) The r.v. Y in this assignment is a special case of generalized linear regression (GLM). The way to construct a GLM goes via a link function. Present and justify the link function in this assignment.

Problem 7.2.

Consider the normal multiple least squares model,

$$Y^* = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon, \tag{88}$$

where $\mathbf{x}^T = (1, x_1, x_2, \ldots, x_p)$, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$ and $\varepsilon \sim N(0,1)$, and is independent of $\mathbf{x}$.

a) Explain why we have

$$P\left(-\varepsilon \leq x\right) = P\left(\varepsilon \leq x\right). \tag{89}$$

Aid: Let $\Phi(x)$ be the cumulative distribution function of $N(0,1)$ and $\phi(x)$ be the corresponding pdf. Since $\phi(x) = \phi(-x)$, we have

$$\Phi(-x) = 1 - \Phi(x).$$

b) Construct

$$Y = \begin{cases} 1 & \text{if } Y^* > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Verify that

$$P(Y = 1 \mid \mathbf{x}) = \Phi\left(\mathbf{x}^T\boldsymbol{\beta}\right),$$

You may use (89) even if you have failed to show it.

c) The r.v. Y in this assignment is a special case of generalized linear regression (GLM). The way to construct a GLM goes via a link function. Present and justify the link function in this assignment.

Problem 7.3.

Y is an exponentially distributed random variable with parameter $\lambda > 0$, we write $Y \sim \text{Exp}(\lambda)$, if its probability density function is

$$f(y;\lambda) = \begin{cases} \lambda e^{-\lambda y} & y \geq 0, \\ 0 & y < 0. \end{cases}$$

λ is sometimes called the rate parameter. We want to construct a generalized linear model (exponential regression). We know that $\mathrm{E}\left[Y\right] = 1/\lambda$, this need not be derived here.

a) Find the canonical link function relating $\mathbf{x}^T\boldsymbol{\beta}$ to the mean. What does this require from $\mathbf{x}^T\boldsymbol{\beta}$?

b) Write down the probability density function

$$f(y; \mathbf{x}^T\boldsymbol{\beta}).$$

c) What is $\mathrm{E}\left[Y|\mathbf{x}\right]$ in exponential regression?

8 The Woodbury Matrix Identity & Ridge Regression

Problem 8.1.

Woodbury matrix identity is a generalization of the Sherman-Morrison-Woodbury theorem, and says that

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}, \tag{90}$$

where A, U, C and V are conformable matrices and the required inverses exist. You are not asked to prove this. By (37) we have the ridge regression estimate

$$\widehat{\boldsymbol{\beta}}_\lambda = \left(X^TX+\lambda\mathbb{I}_p\right)^{-1}X^T\mathbf{y}. \tag{91}$$

Now, apply the Woodbury matrix identity with $A = X^TX$, X has full column rank $p = k+1$, $\lambda\mathbb{I}_p = C$, $V = U = \mathbb{I}_p$ to express $\widehat{\boldsymbol{\beta}}_\lambda$ as in the right hand side of (90). Or, we find that

$$\widehat{\boldsymbol{\beta}}_\lambda = \left(X^TX\right)^{-1}X^T\mathbf{y} - \text{CORR} \tag{92}$$

a) What is the expression for CORR?

b) Let next $\lambda \to 0$. By (91), if X has full column rank,

$$\widehat{\boldsymbol{\beta}}_\lambda \to \left(X^TX\right)^{-1}X^T\mathbf{y}.$$

Find the limit for $\widehat{\boldsymbol{\beta}}_\lambda$ by means of (92), when $\lambda \to 0$.
Hint: Use (90) once more in CORR.

9 Collections of Formulas and Auxiliary Results

9.1 Matrices and Matrix Rules

- A and B conformal, $(AB)^T = B^T A^T$. A and B invertible, $(AB)^{-1} = B^{-1}A^{-1}$.$(A^T)^T = A$.
- A and B conformal, $(A+B)^T = A^T + B^T$
- A is $n \times n$ and invertible.

$$\left(A^T\right)^{-1} = \left(A^{-1}\right)^T. \tag{93}$$

Proof: $A^T \left(A^{-1}\right)^T = \left(A^{-1}A\right)^T = \mathbb{I}_n^T = \mathbb{I}_n$ and $\left(A^{-1}\right)^T A^T = \left(AA^{-1}\right)^T = \mathbb{I}_n^T = \mathbb{I}_n$. □
Hence, as $X^T X$ is $k \times k$, and symmetric

$$\left(\left(X^T X\right)^{-1}\right)^T = \left(X^T X\right)^{-1} \tag{94}$$

The $n \times n$ identity matrix is

$$\mathbb{I}_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & \ddots & \vdots & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \tag{95}$$

The $n \times 1$ vector of ones is denoted by

$$\mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n. \tag{96}$$

The centering matrix is

$$C_{ce} := \mathbb{I}_n - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T. \tag{97}$$

9.2 The Range Space (=Column Space) of a Matrix

Let A be any $n \times p$ matrix. Then the range space of A is defined by

$$\mathcal{R}(A) = \{\mathbf{x} \in \mathbb{R}^n | \text{there exists a } \mathbf{b} \in \mathbb{R}^p \text{ such that } \mathbf{x} = A\mathbf{b}\}. \tag{98}$$

Then we have the following statement.

Proposition 9.1. *X is an $n \times (k+1)$ matrix with full column rank $= k+1$. Then*

$$\mathcal{R}\left(X^T\right) = \mathcal{R}\left(X^T X\right). \tag{99}$$

Proof: We show only half of the assertion, namely that

$$\mathcal{R}\left(X^T X\right) \subset \mathcal{R}\left(X^T\right). \tag{100}$$

For that purpose take any $\mathbf{w} \in \mathcal{R}\left(X^T X\right)$. Thus there exists a $\mathbf{b} \in \mathbb{R}^{k+1}$ such that

$$\mathbf{w} = X^T X\mathbf{b} = X^T\left(X\mathbf{b}\right) = X^T\mathbf{c},$$

where $\mathbf{c} = X\mathbf{b} \in \mathbb{R}^n$. Therefore $\mathbf{w} \in \mathcal{R}\left(X^T\right)$ by definition of $\mathcal{R}\left(X^T\right)$. Hence we have established (100). □

9.3 Projection Matrix

Any $n \times n$ matrix P from $\mathbb{R}^n$ to a subspace (=range space of P), $\mathcal{R}(P) \subset \mathbb{R}^n$, is a called a *projection matrix*, if it is idempotent and symmetric, i.e.,

$$P^2 = P, \quad P^T = P.$$

9.4 Trace of a Square Matrix

Let A be a square matrix. The **trace** $\operatorname{Tr} A$ of A is the sum of the entries in main diagonal:

$$\operatorname{Tr}\begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix} = \textstyle\sum_1^k a_{jj}$$

- 1.If A is a $k \times n$-matrix, and B an $n \times k$-matrix, then $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$
- 2. In particular, if a is a column-vector, then $a^T a = \operatorname{Tr}\left(aa^T\right)$.
- 3. For any real numbers a and b, $\operatorname{Tr}(aC + bD) = a\operatorname{Tr} C + b\operatorname{Tr} D$

Exempel 9.2. The hat matrix H is $H = X(X^T X)^{-1}X^T$. Set $B = X^T$ and $A = X(X^T X)^{-1}$. Then

$$\operatorname{Tr} H = \operatorname{Tr} X(X^T X)^{-1}X^T = \operatorname{Tr} AB = \operatorname{Tr} BA,$$

where we used rule 2.. But $BA = X^T X(X^T X)^{-1} = \mathbb{I}_{k+1}$ (How is $k+1$ there?). Thus

$$\operatorname{Tr} H = \operatorname{Tr}\mathbb{I}_{k+1} = k + 1.$$

9.5 Factorization and Square Root of Covariance Matrices

If Σ is an $n \times n$ symmetric matrix, then Σ can be written as

$$\Sigma = ADA^T,$$

where A is an orthogonal matrix ($A^TA = AA^T = I_n$) and D is an $n \times n$ diagonal matrix, with the eigenvalues on the main diagonal. If Σ is a covariance matrix, its eigenvalues λ_i are non-negative. Then

$$\Sigma^{1/2} = AD^{1/2}A^T,$$

where $D^{1/2}$ is an $n \times n$ diagonal matrix, with $\sqrt{\lambda_i}$ on the main diagonal. One checks now that $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$.

When Σ is positive definite, its eigenvalues are positive and we define

$$\Sigma^{-1/2} = AD^{-1/2}A^T, \tag{101}$$

where $D^{-1/2}$ is an $n \times n$ diagonal matrix, with $1/\sqrt{\lambda_i}$ on the main diagonal. $\Sigma^{-1/2}$ is symmetric, since $D^{-1/2}$ is symmetric. Clearly, $D^{-1/2}D^{-1/2} = D^{-1}$. Then

$$\Sigma^{-1} = \Sigma^{-1/2}\Sigma^{-1/2}. \tag{102}$$

9.6 Linear Transformations of Covariance Matrices

$$E[\mathbf{X} + \mathbf{Y}] = \boldsymbol{\mu}_\mathbf{X} + \boldsymbol{\mu}_\mathbf{Y} \tag{103}$$

If $\mathbf{Z} = A\mathbf{X} + \mathbf{b}$, then

$$E[\mathbf{Z}] = A\boldsymbol{\mu}_\mathbf{X} + \mathbf{b}, \tag{104}$$

and

$$C_\mathbf{Z} = AC_\mathbf{X}A^T. \tag{105}$$

$$C_\mathbf{X} = E\left[\mathbf{X}\mathbf{X}^T\right] - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^T \tag{106}$$

$$\text{Var}\left[\mathbf{a}^T\mathbf{X}\right] = \mathbf{a}^TC_\mathbf{X}\mathbf{a} \tag{107}$$

9.7 Generalized Inverse

If a generalized inverse G of A satisfies the four conditions below, then G is called the Bjerhammar -Moore-Penrose (BMP) inverse.

MP1 $AGA = A$

MP2 $GAG = G$

MP3 $(AG)^T = AG$

MP4 $(GA)^T = GA$

9.8 Rules of Computation with Finite Sums

(1) $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n.$

(2) $\sum_{i=1}^{n} a \cdot x_i = a \sum_{i=1}^{n} x_i.$

(3) $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$

(4) $\sum_{i=1}^{n} (ax_i + by_i) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i$

(5) $\sum_{i=1}^{n} (x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.$
$\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i.$

(6) $\sum_{i=1}^{n} (x_i - \overline{x}) = 0.$

(7) $\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) = \sum_{i=1}^{n} (x_i - \overline{x}) y_i = \sum_{i=1}^{n} x_i (y_i - \overline{y}).$

(8) $\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}.$

(9) $\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2.$

9.9 Simple Linear Regression

9.9.1 The Model

Let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

be an $(n \times 1)$-vector, $(n \times 2)$-matrix, and (2×1)-vector, respectively. The simple linear regression is now given by

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an $n \times 1$ random vector. We assume the following:

1) $E[\boldsymbol{\varepsilon}] = \mathbf{0}_n$ (= the $n \times 1$ zero vector), i.e., $\mathbf{0}_n \in \mathbb{R}^n$.

2) The errors are uncorrelated: the covariance matrix $C_{\boldsymbol{\varepsilon}}$ of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is

$$C_{\boldsymbol{\varepsilon}} = \sigma^2 \mathbb{I}_n.$$

3) σ^2 does not depend on X.

9.9.2 LSE

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\begin{cases} \frac{\partial}{\partial \beta_0} Q(\hat{\beta}_0, \hat{\beta}_1) & = 0 \\ \frac{\partial}{\partial \beta_1} Q(\hat{\beta}_0, \hat{\beta}_1) & = 0. \end{cases}$$

$$\underline{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.} \tag{108}$$

$\frac{\partial}{\partial \beta_1} Q(\hat{\beta}_0, \hat{\beta}_1) = 0 \Leftrightarrow \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0.$

$$\underline{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})\, x_i}{\sum_{i=1}^{n} (x_i - \overline{x})\, x_i}.} \tag{109}$$

Let β_0^*, β_1^* denote the true poulation values of the intercept and the regression coefficient. Assume that

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N_n \left(\mathbf{0}_n, \sigma^2 \mathbb{I}_n \right)$$

Then

$$\begin{aligned} \hat{\beta}_0 &\sim N \left(\beta_0^*, \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \right) \\ \hat{\beta}_1 &\sim N \left(\beta_1^*, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right). \end{aligned} \tag{110}$$

9.10 Multiple Linear Regression

The Ordinary Least Squares Model (OLS)

We have the model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{111}$$

where

$$X = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and we assume the following:

1) $E[\boldsymbol{\varepsilon}] = \mathbf{0}_n$ (= the $n \times 1$ zero vector), i.e., $\mathbf{0}_n \in \mathbb{R}^n$.

2) The covariance matrix $C_{\boldsymbol{\varepsilon}}$ of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is

$$C_{\boldsymbol{\varepsilon}} = \sigma^2 \mathbb{I}_n. \tag{112}$$

3) σ^2 does not depend on X.

4) $n > k+1$ and X has full rank.

9.11 Normal Linear Regression

OLS with

$$\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2 \mathbb{I}_n\right)$$
$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}.$$

Let $\mathbf{X} \sim N_n(\boldsymbol{\mu}, C)$ and $\mathbf{Y} = B\mathbf{X} + \mathbf{b}$, B is $m \times n$. Then

$$\mathbf{Y} \sim N_m\left(B\boldsymbol{\mu} + \mathbf{b}, BCB^T\right). \tag{113}$$

9.12 Distribution of Quadratic Forms of Normal Vectors

1. $\mathbf{X} \sim N_n(\mu, \Sigma)$ is an $n \times 1$ Gaussian vector, where Σ is positive definite. The quadratic form is

$$(\mathbf{X}-\mu)^T \Sigma^{-1} (\mathbf{X}-\mu).$$

We use the factorization of Σ^{-1} in (102) to get

$$(\mathbf{X}-\mu)^T \Sigma^{-1} (\mathbf{X}-\mu) = \left(\Sigma^{-1/2}(\mathbf{X}-\mu)\right)^T \Sigma^{-1/2}(\mathbf{X}-\mu).$$

Let $\mathbf{Z} := \Sigma^{-1/2}(\mathbf{X}-\mu)$. Then our rules of computation give that Z has the mean vector $E[Z] = \mathbf{0}_n$ and Z has the covariance matrix

$$C_{\mathbf{Z}} = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma^{1/2}\Sigma^{1/2}\Sigma^{-1/2} = \mathbf{I}_n.$$

$\mathbf{Z} \sim N_n(\mathbf{0}_n, \mathbb{I}_n)$. Hence

$$(\mathbf{X}-\mu)^T \Sigma^{-1} (\mathbf{X}-\mu) = \mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n).$$

This is Thm 9.1. in Gut, Allan: *An Intermediate Course in Probability. Second Edition.* Note that $z_i \sim N(0,1)$ are indepedent and that a finite sum of squares of independent standard normal r.v.'s has $\chi^2(n)$ (chi-squared distribution with n degrees of freedom).

2.

Proposition 9.3. *If* $\mathbf{Z} \sim N_n(\mathbf{0}_n, \mathbb{I}_n)$, *then*

$$\mathbf{Z}^T A \mathbf{Z} \sim \chi^2(r) \tag{114}$$

if and only if A *is an idempotent matrix with* $\operatorname{rank} A = r$.

9.13 Sherman-Morrison-Woodbury Theorem

Suppose A is is an invertible square $n \times n$ matrix and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are column vectors. Then $A + \mathbf{u}\mathbf{v}^\mathsf{T}$ is invertible iff $1 + \mathbf{v}^\mathsf{T} A^{-1}\mathbf{u} \neq 0$. In this case,

$$\left(A + \mathbf{u}\mathbf{v}^\mathsf{T}\right)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^\mathsf{T}A^{-1}}{1 + \mathbf{v}^\mathsf{T}A^{-1}\mathbf{u}}. \tag{115}$$

9.14 Fundamental Analysis of Variance Identity

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{116}$$

is known as the **Fundamental Analysis of Variance Identity**

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \leftrightarrow \ SS_{\mathrm{T}}$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \leftrightarrow \ SS_{\mathrm{R}} \textbf{ regression or model sum of squares}$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \leftrightarrow \quad SS_{\mathrm{Res}} \textbf{ Residual Sum of Squares}$$

$$SS_{\mathrm{T}} = SS_{\mathrm{R}} + SS_{\mathrm{Res}} \tag{117}$$

Coefficient of determination

$$R^2 = \frac{SS_{\mathrm{R}}}{SS_{\mathrm{T}}} = 1 - \frac{SS_{\mathrm{Res}}}{SS_{\mathrm{T}}}. \tag{118}$$

9.15 Matrix Derivatives

Let $\mathbf{A}$ be a $k \times k$ matrix of constants, $\mathbf{a}$ be a $k \times 1$ vector of constants, and $\mathbf{y}$ be a $k \times 1$ vector of varaibles.

1. If $\mathbf{z} = \mathbf{a}^\top \mathbf{y}$, then
$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \frac{\partial \mathbf{a}^\top \mathbf{y}}{\partial \mathbf{y}} = \mathbf{a}.$$
2. If $\mathbf{z} = \mathbf{y}^\top \mathbf{y}$, then
$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \frac{\partial \mathbf{y}^\top \mathbf{y}}{\partial \mathbf{y}} = 2\mathbf{y}.$$
3. If $\mathbf{z} = \mathbf{a}^\top \mathbf{A}\mathbf{y}$, then
$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \frac{\partial \mathbf{a}^\top \mathbf{A}\mathbf{y}}{\partial \mathbf{y}} = \mathbf{A}^\top \mathbf{a}.$$

4. If $\mathbf{z} = \mathbf{y}^\top \mathbf{A}\mathbf{y}$, then

$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \frac{\partial \mathbf{y}^\top \mathbf{A}\mathbf{y}}{\partial \mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{A}^\top \mathbf{y}.$$

If $\mathbf{A}$ is symmetric, then

$$\frac{\partial \mathbf{y}^\top \mathbf{A}\mathbf{y}}{\partial \mathbf{y}} = 2\mathbf{A}\mathbf{y}.$$

9.16 Solution of (20) (for all interested)

We expand using rules of transpose and multiplication

$$\begin{aligned}(\mathbf{Y} - A\mathbf{X})(\mathbf{Y} - A\mathbf{X})^T &= (\mathbf{Y} - A\mathbf{X})\left(\mathbf{Y}^T - (A\mathbf{X})^T\right) \\ &= \mathbf{Y}\mathbf{Y}^T - \mathbf{Y}(A\mathbf{X})^T - A\mathbf{X}\mathbf{Y}^T + A\mathbf{X}(A\mathbf{X})^T.\end{aligned}$$

Now we take the expectations

$$\begin{aligned}E\left[\mathbf{Y}\mathbf{Y}^T - \mathbf{Y}(A\mathbf{X})^T - A\mathbf{X}\mathbf{Y}^T + A\mathbf{X}(A\mathbf{X})^T\right] &= E\left[\mathbf{Y}\mathbf{Y}^T\right] - E\left[\mathbf{Y}(A\mathbf{X})^T\right] \\ &\quad -E\left[A\mathbf{X}\mathbf{Y}^T\right] + E\left[A\mathbf{X}(A\mathbf{X})^T\right].\end{aligned} \tag{119}$$

Since mean vectors are here zero vectors, we use next (11), i.e., $\mathrm{Cov}(A\mathbf{X}, B\mathbf{Y}) = A\mathrm{Cov}(\mathbf{X}, \mathbf{Y})B^T$ with

$$E\left[\mathbf{Y}(A\mathbf{X})^T\right] = \mathrm{Cov}(\mathbf{Y}, A\mathbf{X}) = \mathrm{Cov}(\mathbf{Y}, \mathbf{X})A^T = E\left[(\mathbf{Y}\mathbf{X}^T\right]A^T, \tag{120}$$

and

$$E\left[A\mathbf{X}\mathbf{Y}^T\right] = \mathrm{Cov}(A\mathbf{X}, \mathbf{Y}) = A\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = AE\left[\mathbf{X}\mathbf{Y}^T\right], \tag{121}$$

and

$$E\left[A\mathbf{X}(A\mathbf{X})^T\right] = \mathrm{Cov}(A\mathbf{X}, A\mathbf{X}) = A\mathrm{Cov}(\mathbf{X}, \mathbf{X})A^T = A\mathrm{C}_{\mathbf{X}}A^T. \tag{122}$$

By definition

$$E\left[\mathbf{Y}\mathbf{Y}^T\right] = \mathrm{C}_{\mathbf{Y}}. \tag{123}$$

When we insert from (120) -(123) in the right hand side of (119) we get

$$\begin{aligned}E\left[(\mathbf{Y} - A\mathbf{X})(\mathbf{Y} - A\mathbf{X})^T\right] &= \mathrm{C}_{\mathbf{Y}} \\ &\quad -E\left[(\mathbf{Y}\mathbf{X}^T\right]A^T - AE\left[\mathbf{X}\mathbf{Y}^T\right] \\ &\quad +A\mathrm{C}_{\mathbf{X}}A^T.\end{aligned} \tag{124}$$

Here the expression (19) of the optimal matrix A gives

$$\begin{aligned}E\left[\mathbf{Y}\mathbf{X}^T\right]A^T &= \mathrm{Cov}(\mathbf{Y}, \mathbf{X})A^T = \mathrm{Cov}(\mathbf{Y}, \mathbf{X})(\mathrm{C}_{\mathbf{X}}^{-1})^T\mathrm{Cov}(\mathbf{Y}, \mathbf{X})^T \\ &= \mathrm{Cov}(\mathbf{Y}, \mathbf{X})\mathrm{C}_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X}, \mathbf{Y}).\end{aligned} \tag{125}$$

Here we first invoked (9) by $\mathrm{Cov}(\mathbf{Y},\mathbf{X})^T = \left(\mathrm{Cov}(\mathbf{X},\mathbf{Y})^T\right)^T = \mathrm{Cov}(\mathbf{X},\mathbf{Y})$. Second, we used (93), as well as the fact that the covariance matrix $\mathrm{C}_{\mathbf{X}}$ is symmetric. Again by (19),

$$AE\left[\mathbf{X}\mathbf{Y}^T\right] = \mathrm{Cov}(\mathbf{Y},\mathbf{X})\mathrm{C}_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X},\mathbf{Y}). \tag{126}$$

Finally by (19)

$$\begin{aligned} A\mathrm{C}_{\mathbf{X}}A^T &= \mathrm{Cov}(\mathbf{Y},\mathbf{X})\underbrace{\mathrm{C}_{\mathbf{X}}^{-1}\mathrm{C}_{\mathbf{X}}}_{=\mathbb{I}_n}\left(\mathrm{Cov}(\mathbf{Y},\mathbf{X})\mathrm{C}_{\mathbf{X}}^{-1}\right)^T \\ &= \mathrm{Cov}(\mathbf{Y},\mathbf{X})\left(\mathrm{C}_{\mathbf{X}}^{-1}\right)^T\mathrm{Cov}(\mathbf{Y},\mathbf{X})^T \end{aligned}$$

By (9), $\mathrm{Cov}(\mathbf{Y},\mathbf{X})^T = \left(\mathrm{Cov}(\mathbf{X},\mathbf{Y})^T\right)^T = \mathrm{Cov}(\mathbf{X},\mathbf{Y})$. We note also again (93) and that the covariance matrix $\mathrm{C}_{\mathbf{X}}$ is symmetric. Hence we have found that

$$A\mathrm{C}_{\mathbf{X}}A^T = \mathrm{Cov}(\mathbf{Y},\mathbf{X})\mathrm{C}_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X},\mathbf{Y}). \tag{127}$$

Next we substitute (125), (126) and (127) in the right hand side of (124) we obtain

$$\begin{aligned} E\left[(\mathbf{Y}-A\mathbf{X})(\mathbf{Y}-A\mathbf{X})^T\right] &= \mathrm{C}_{\mathbf{Y}} \\ &\quad -\mathrm{Cov}(\mathbf{Y},\mathbf{X})\mathrm{C}_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X},\mathbf{Y}) - \mathrm{Cov}(\mathbf{Y},\mathbf{X})\mathrm{C}_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X},\mathbf{Y}) \\ &\quad +\mathrm{Cov}(\mathbf{Y},\mathbf{X})\mathrm{C}_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X},\mathbf{Y}). \\ &= \mathrm{C}_{\mathbf{Y}} - \mathrm{Cov}(\mathbf{Y},\mathbf{X})\mathrm{C}_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X},\mathbf{Y}). \end{aligned} \tag{128}$$

Clearly, the rightmost expression in (128) verifies (20). □