2. Theoretical exercises

Demo exercises

2.1 Prove the Gauss-Markov theorem.

Solution. Let the standard assumptions (i)-(v) of the lecture slides be satisfied. Under the standard assumptions, Gauss-Markov theorem states that the least squares estimator,

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y},$$

is the best linear unbiased estimator (BLUE) for the regression coefficients $\boldsymbol{\beta}$. In this context, the best estimator is the estimator with the smallest variance. Let $\mathbf{b}^*$ be a linear unbiased estimator for the regression coefficients. In order to prove the Gauss-Markov theorem, we need to show that,

$$\mathrm{Cov}(\mathbf{b}^*) - \mathrm{Cov}(\mathbf{b})$$

is positive semidefinite for every $\mathbf{b}^*$. We proved that $\mathbf{b}$ is an unbiased estimator in the theoretical exercises of week 1. In addition, by the theoretical exercises of week 1, we have that,

$$\mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}.$$

Let,

$$\mathbf{b}^* = \mathbf{C}\mathbf{y} = (\mathbf{D} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y},$$

where $\mathbf{C} = \mathbf{D} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$ is a non-random matrix of size $(k+1) \times n$. Since $\mathbf{b}^*$ is assumed to be unbiased, we have that,

$$\begin{aligned}\mathbb{E}(\mathbf{b}^*) &= \mathbb{E}\left[\left(\mathbf{D} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\right)\mathbf{y}\right] = \left(\mathbf{D} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\right)\mathbf{X}\boldsymbol{\beta} \\ &= (\mathbf{D}\mathbf{X} + \mathbf{I})\boldsymbol{\beta},\end{aligned}$$

which gives $\mathbf{D}\mathbf{X} = 0$, since the equation above has to hold for every $\boldsymbol{\beta}$. Recall that, $\mathrm{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, where σ^2 is the variance of the residual terms. Hereby, the covariance matrix is,

$$\begin{aligned}\mathrm{Cov}(\mathbf{b}^*) &= \mathbb{E}\left[(\mathbf{b}^* - \mathbb{E}(\mathbf{b}^*))(\mathbf{b}^* - \mathbb{E}(\mathbf{b}^*))^\top\right] = \mathbb{E}\left[(\mathbf{C}\mathbf{y} - \mathbb{E}(\mathbf{C}\mathbf{y}))(\mathbf{C}\mathbf{y} - \mathbb{E}(\mathbf{C}\mathbf{y}))^\top\right] \\ &= \mathbb{E}\left[\mathbf{C}(\mathbf{y} - \mathbb{E}(\mathbf{y}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))^\top \mathbf{C}^\top\right) = \mathbf{C}(\mathrm{Cov}(\mathbf{y}))\mathbf{C}^\top = \sigma^2\mathbf{C}\mathbf{C}^\top \\ &= \sigma^2\left(\mathbf{D} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\right)\left(\mathbf{D} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\right)^\top \\ &= \sigma^2\left(\mathbf{D}\mathbf{D}^\top + \mathbf{D}\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}^\top + (\mathbf{X}^\top \mathbf{X})^{-1}\right) \\ &= \sigma^2\mathbf{D}\mathbf{D}^T + \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2\mathbf{D}\mathbf{D}^\top + \mathrm{Cov}(\mathbf{b}).\end{aligned}$$

Furthermore, the difference of the covariance matrices is

$$\mathrm{Cov}(\mathbf{b}^*) - \mathrm{Cov}(\mathbf{b}) = \sigma^2 \mathbf{D}\mathbf{D}^\top,$$

which is a positive semidefinite matrix, since $\mathbf{D}\mathbf{D}^\top$ is symmetric and

$$\mathbf{a}^\top(\mathbf{D}\mathbf{D}^\top)\mathbf{a} = \mathbf{c}^\top\mathbf{c} = ||\mathbf{c}||_2^2 \geq 0,$$

where $\mathbf{c} = \mathbf{D}^\top\mathbf{a}$ and $|| \cdot ||_2$ is the ordinary l^2-vector norm. Since the matrix is positive semidefinite, it follows that the variances of the least squares estimators are smaller (or at most equal) than the variances of the estimator $\mathbf{b}^*$. Note that, the equality is involved above, since the matrix $\mathbf{D}$ is not necessary of full-rank.

2.2 Let,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbf{X} \in \mathbb{R}^{n\times(k+1)},$$

be a linear model that satisfies the standard assumptions (i)-(v). Furthermore, let $\boldsymbol{\beta}$ satisfy the constraint,

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r},$$

where $\mathbf{R}$ is a full-rank $m \times (k+1)$-matrix with $m < k+1$. Derive the constrained least squares estimator for $\boldsymbol{\beta}$. Use the method of Lagrange multipliers, and recall that,

$k+1$	= number of variables
m	= number of constrains
n	= number of observations

Solution.

The Lagrangian function is,

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= \boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} + 2\boldsymbol{\lambda}^\top(\mathbf{R}\boldsymbol{\beta} - \mathbf{r}) \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\boldsymbol{\lambda}^\top(\mathbf{R}\boldsymbol{\beta} - \mathbf{r}) \\ &= \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\lambda}^\top\mathbf{R}\boldsymbol{\beta} - 2\boldsymbol{\lambda}^\top\mathbf{r}, \end{aligned}$$

where $2\boldsymbol{\lambda}$ is the m-vector of Lagrange multipliers (the multiplier 2 is included for convenience). Note that $\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y}$ and $\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta}$ are scalars, and hence $\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y}$. Next, we differentiate the function $f(\boldsymbol{\beta}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, and set the derivatives equal to zero (recall matrix differentiation from the first theoretical exercises).

$$\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} = -2\mathbf{y}^\top\mathbf{X} + 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X} + 2\boldsymbol{\lambda}^\top\mathbf{R} = \mathbf{0}, \tag{1}$$

$$\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = 2\boldsymbol{\beta}^\top\mathbf{R}^\top - 2\mathbf{r}^\top = \mathbf{0}. \tag{2}$$

Equations (1) and (2) form a system of equations with unknown vectors $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. Note that, the corresponding equations are formulated as row vectors.

By right-multiplying Equation (1) with the matrix $-\frac{1}{2}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top$, we obtain

$$\mathbf{y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top - \boldsymbol{\beta}^\top\mathbf{R}^\top = \boldsymbol{\lambda}^\top\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top. \tag{3}$$

It can be shown that the matrix $\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top$ is invertible, the proof is omitted here. Next, we use Equation 2 to solve Equation 3 for the the vector $\boldsymbol{\lambda}$,

$$\begin{aligned}\boldsymbol{\lambda}^\top &= (\mathbf{y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top - \boldsymbol{\beta}^\top\mathbf{R}^\top)(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}\\ &= (\mathbf{b}^\top\mathbf{R}^\top - \mathbf{r}^\top)(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1},\end{aligned}$$

where,

$$\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

is the ordinary least squares estimator for the vector $\boldsymbol{\beta}$. Then, by substituting the obtained expression for $\boldsymbol{\lambda}^\top$ into Equation (1), we get that,

$$-\mathbf{y}^\top\mathbf{X} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X} + (\mathbf{b}^\top\mathbf{R}^\top - \mathbf{r}^\top)(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}\mathbf{R} = \mathbf{0}.$$

By solving for $\boldsymbol{\beta}$, we obtain the constrained least squares estimator $\mathbf{b_R}$:

$$\begin{aligned}\mathbf{b}_\mathbf{R}^\top &= \mathbf{y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} - (\mathbf{b}^\top\mathbf{R}^\top - \mathbf{r}^\top)(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\\ &= \mathbf{b}^\top - (\mathbf{b}^\top\mathbf{R}^\top - \mathbf{r}^\top)(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\\ \implies \mathbf{b}_\mathbf{R} &= \mathbf{b} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}).\end{aligned}$$

Homework

2.3 Consider the following data set containing three observations:

$$\begin{aligned}\mathbf{y}_1 &= (y_{11}, y_{12}) = (1, 2)\\ \mathbf{y}_2 &= (y_{21}, y_{22}) = (3, 4)\\ \mathbf{y}_3 &= (y_{31}, y_{32}) = (5, 6)\end{aligned}$$

a) Keep the first variable (coordinate) fixed and permute the second variable (coordinate). How many distinct permutations can be formed?

b) Keep the first variable (coordinate) fixed and permute the second variable (coordinate). Find every distinct permutation.

c) Form 5 bootstrap samples of the data.

d) Consider the following table with eight distinct scenarios. Which of the following are possible bootstrap samples?

1	2	3	4	5	6	7	8
(1,2)	(3,4)	(1,2)	(1,2)	(1,1)	(1,6)	(1,4)	(4,3)
(1,2)	(3,4)	(2,1)	(3,4)	(2,2)	(3,2)	(1,2)	(4,3)
(5,6)	(3,4)	(1,2)	(5,6)	(3,3)	(5,4)	(1,6)	(4,3)

2.4 Consider the following linear models,

$$y = \alpha_0 + \alpha_1 x + \varepsilon, \tag{4}$$

$$y = \beta_0 + \beta_1 x + \beta_2 z + \nu, \tag{5}$$

where we have n observations for the variables z, y and x. The estimates for the regression coefficients are given by the least squares method and are denoted with the hat symbol. When do the following claims hold true? (consider each part separately)

Note that some of the claims might not be true in any situation. Deduction with good reasoning is sufficient here.

a. $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \geq \sum_{i=1}^{n} \hat{\nu}_i^2$ ($\hat{\varepsilon}$ and $\hat{\nu}$ are the estimated residuals).

b. $\hat{\alpha}_1$ is statistically significant (5% significance level), but $\hat{\beta}_1$ is not.

c. $\hat{\alpha}_1$ is not statistically significant (5% significance level), but $\hat{\beta}_1$ is.

d. The coefficient of determination for model (4) is larger than the coefficient of determination for model (5).