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

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

be a linear model without the intercept term that satisfies the standard assumptions (i)-(v). Instead of the usual least squares criterion consider the following constrained least squares criterion,

$$g(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2, \quad \lambda > 0,$$

where $\|\boldsymbol{\beta}\|^2 = \boldsymbol{\beta}^\top\boldsymbol{\beta}$.

(a) Show that $\mathbf{b}_\lambda = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$ minimizes the criterion $g(\boldsymbol{\beta})$.

(b) Compute $\mathbb{E}[\mathbf{b}_\lambda]$. Is the estimator $\mathbf{b}_\lambda$ unbiased?

(c) Compute $\mathrm{Cov}(\mathbf{b}_\lambda)$.

(d) Show that $\mathrm{Cov}(\mathbf{b}) - \mathrm{Cov}(\mathbf{b}_\lambda)$ is positive definite, where $\mathbf{b}$ is the least squares estimator. Is this a violation of Gauss-Markov theorem?

Solution.

(a) Denote $f(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}$. Then

$$g(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}.$$

By Exercise 1.2a) we have

$$f'(\boldsymbol{\beta}) = -2\mathbf{y}^\top\mathbf{X} + 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}$$

and by generalizing the result of Exercise 1.1c) we have

$$\frac{\partial\boldsymbol{\beta}^\top\boldsymbol{\beta}}{\partial\boldsymbol{\beta}} = 2\boldsymbol{\beta}^\top.$$

Thus

$$g'(\beta) = -2\mathbf{y}^\top\mathbf{X} + 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X} + 2\lambda\boldsymbol{\beta}^\top$$

and by setting $g'(\boldsymbol{\beta}) = \mathbf{0}$ we obtain the following equation,

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y}.$$

Notice that the matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is invertible even if $\mathbf{X}$ includes linearly dependent columns. Now for the zero of the derivative $g'(\boldsymbol{\beta})$ we have

$$\mathbf{b}_\lambda = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}.$$

It remains to show that $\mathbf{b}_\lambda$ in fact minimizes $g(\boldsymbol{\beta})$ and is not, for example, a saddle point. By applying 1.1e) we get

$$g''(\boldsymbol{\beta}) = 2\mathbf{X}^\top\mathbf{X} + 2\lambda\mathbf{I}.$$

Matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is positive definite and thus $\mathbf{b}_\lambda$ minimizes $g(\boldsymbol{\beta})$.

(b) Remember that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{X}\boldsymbol{\beta}$. Then

$$\begin{aligned}\mathbb{E}[\mathbf{b}_\lambda] &= \mathbb{E}[(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}] = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbb{E}[\mathbf{y}] \\ &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta}.\end{aligned} \tag{1}$$

We can still expand the expression for $\mathbb{E}[\mathbf{b}_\lambda]$ so that it is easier to see if $\mathbf{b}_\lambda$ is biased. Notice that

$$\begin{aligned}&\mathbf{X}^\top\mathbf{X} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}) - \lambda\mathbf{I} \\ \Rightarrow&(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{I} - \lambda(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1} \\ \Rightarrow&(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} - \lambda(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}.\end{aligned} \tag{2}$$

By combining Equations (1) and (2) we get

$$\mathbb{E}[\mathbf{b}_\lambda] = \boldsymbol{\beta} - \lambda(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}.$$

Now $\mathbf{b}_\lambda$ is unbiased if and only if

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta} = \mathbf{0} \quad \forall\ \boldsymbol{\beta} \in \mathbb{R}^k.$$

However, the matrix $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}$ is invertible and thus it has trivial null space. Hence,

$$\mathbb{E}[\mathbf{b}_\lambda] \neq \boldsymbol{\beta}, \quad \forall\ \boldsymbol{\beta} \in \mathbb{R}^k \setminus \{\mathbf{0}\}.$$

That is, estimator $\mathbf{b}_\lambda$ is biased.

(c) Notice that for covariance we have the following property.

$$\mathrm{Cov}(\mathbf{A}\mathbf{b}_\lambda) = \mathbf{A}\mathrm{Cov}(\mathbf{b}_\lambda)\mathbf{A}^\top,$$

where $\mathbf{A} \in \mathbb{R}^{k \times k}$ is a nonrandom invertible matrix. This property of scatter estimators is often called affine equivariance. Proving the affine equivariance of the covariance is out of the scope of this course. For more details, see the course Multivariate Statistical Analysis (MS-E2112).

Notice that we can express $\mathbf{b}_\lambda$ as

$$\mathbf{b}_\lambda = \mathbf{A}_\lambda \mathbf{b},$$

where $\mathbf{b}$ is the least squares estimator and $\mathbf{A}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X}$. Remember that $\mathrm{Cov}(\mathbf{b}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$. Then by affine equivariance of covariance we have

$$\begin{aligned} \mathrm{Cov}(\mathbf{b}_\lambda) &= \mathrm{Cov}(\mathbf{A}_\lambda \mathbf{b}) = \mathbf{A}_\lambda \mathrm{Cov}(\mathbf{b}) \mathbf{A}_\lambda^\top \\ &= \sigma^2 \mathbf{A}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{A}_\lambda^\top. \end{aligned}$$

(d) Expand the expression for $\mathrm{Cov}(\mathbf{b}) - \mathrm{Cov}(\mathbf{b}_\lambda)$,

$$\begin{aligned} &\mathrm{Cov}(\mathbf{b}) - \mathrm{Cov}(\mathbf{b}_\lambda) \\ &= \sigma^2 \left((\mathbf{X}^\top \mathbf{X})^{-1} - \mathbf{A}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{A}_\lambda^\top \right) \\ &= \sigma^2 \mathbf{A}_\lambda \left(\mathbf{A}_\lambda^{-1} (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{A}_\lambda^{-1})^\top + (\mathbf{X}^\top \mathbf{X})^{-1} \right) \mathbf{A}_\lambda^\top \\ &= \sigma^2 \mathbf{A}_\lambda \left(2\lambda (\mathbf{X}^\top \mathbf{X})^{-2} + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-3} \right) \mathbf{A}_\lambda^\top \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \left(2\lambda \mathbf{I} + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-1} \right) (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}. \end{aligned}$$

Denote $\mathbf{B} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$ and $\mathbf{C} = 2\lambda \mathbf{I} + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-1}$. Notice that both $\mathbf{B}$ and $\mathbf{C}$ are positive definite. Let $\mathbf{x} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$. Then $\mathbf{y} = \mathbf{B}\mathbf{x}$ is nonzero and

$$\mathbf{x}^\top \left(\mathrm{Cov}(\mathbf{b}) - \mathrm{Cov}(\mathbf{b}_\lambda) \right) \mathbf{x} = \sigma^2 \mathbf{x}^\top \mathbf{B}^\top \mathbf{C} \mathbf{B} \mathbf{x} = \sigma^2 \mathbf{y}^\top \mathbf{C} \mathbf{y} > 0.$$

Now we have proven that $\mathrm{Cov}(\mathbf{b}) - \mathrm{Cov}(\mathbf{b}_\lambda)$ is positive definite, from which it follows that $\mathrm{Cov}(\mathbf{b}_\lambda) - \mathrm{Cov}(\mathbf{b})$ is not positive semidefinite. However, this is not a violation of Gauss-Markov theorem since the estimator $\mathbf{b}_\lambda$ is biased.

Notes about this exercise:

- The variant of linear regression considered in this exercise is called ridge regression. For more detail about ridge regression, see [HTF09].
- Ridge regression is one way to deal with multicollinearity. Notice that ridge regression solution exists and is unique even if there are linearly dependent variables.

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,

$$y = \alpha_0 + \alpha_1 x + \varepsilon, \tag{3}$$
$$y = \beta_0 + \beta_1 x + \beta_2 z + \nu, \tag{4}$$

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 (3) is larger than the coefficient of determination for model (4).

References

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.