Statistical Methods - Homework #3

1. Consider the multiple linear regression model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim (0, \sigma^2) \quad \text{independently for } i = 1, \dots, n, \tag{1}$$

where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n, \boldsymbol{\beta} \in \mathbb{R}^p$ and $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times p}$ is full rank. $(\operatorname{rank}(X) = p)$ Letting $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$, (1) can equivalently be written as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \text{ and } \operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n.$$

(A) Find the ridge estimator of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}_\lambda$, that minimizes

$$\ell(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} = \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}.$$

Solution: Differentiate $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and set to zero:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = -2X^\top (\boldsymbol{y} - X\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = \mathbf{0}.$$

Rearranging gives the normal equations for the ridge estimator:

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

Since $X^\top X + \lambda I_p$ is positive definite (hence invertible) for any $\lambda > 0$,

$$\boxed{\hat{\boldsymbol{\beta}}_\lambda = (X^\top X + \lambda I_p)^{-1} X^\top \boldsymbol{y}.} \tag{2}$$

(B) Find $\mathrm{E}(\hat{\boldsymbol{\beta}}_\lambda)$. What is the value of λ that makes $\mathrm{E}(\hat{\boldsymbol{\beta}}_\lambda) = \boldsymbol{\beta}$?

Solution: Substituting $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ into (2) and taking expectation:

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\beta}}_\lambda) &= (X^\top X + \lambda I_p)^{-1} X^\top \mathrm{E}(\boldsymbol{y}) \\ &= (X^\top X + \lambda I_p)^{-1} X^\top X \boldsymbol{\beta}. \end{aligned} \tag{3}$$

In general, $\mathrm{E}(\hat{\boldsymbol{\beta}}_\lambda) \neq \boldsymbol{\beta}$ unless $\lambda = 0$, in which case

$$(X^\top X + 0 \cdot I_p)^{-1} X^\top X = (X^\top X)^{-1} X^\top X = I_p,$$

so $\mathrm{E}(\hat{\boldsymbol{\beta}}_0) = \boldsymbol{\beta}$. Hence, $\hat{\boldsymbol{\beta}}_\lambda$ is biased for $\lambda > 0$, and the value $\boxed{\lambda = 0}$ makes the estimator unbiased.

(C) Find $\mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda)$. Also, show that

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda) \preceq \mathrm{Var}(\hat{\boldsymbol{\beta}}_0).$$

(Hint: You may use eigen-value decomposition on $X^\top X$: There exists a matrix $P \in \mathbb{R}^{p\times p}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{p\times p}$ such that $P^\top P = PP^\top = I$ and $X^\top X = P\Lambda P^\top$).

Solution: Since $\hat{\boldsymbol{\beta}}_\lambda = (X^\top X + \lambda I_p)^{-1}X^\top \boldsymbol{y}$ is a linear function of $\boldsymbol{y}$ with $\mathrm{Var}(\boldsymbol{y}) = \sigma^2 I_n$,

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda) &= (X^\top X + \lambda I_p)^{-1}X^\top \cdot \sigma^2 I_n \cdot X(X^\top X + \lambda I_p)^{-1} \\ &\boxed{= \sigma^2 (X^\top X + \lambda I_p)^{-1}X^\top X(X^\top X + \lambda I_p)^{-1}.} \end{aligned} \tag{4}$$

For the OLS estimator ($\lambda = 0$): $\mathrm{Var}(\hat{\boldsymbol{\beta}}_0) = \sigma^2(X^\top X)^{-1}$.

To prove $\mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda) \preceq \mathrm{Var}(\hat{\boldsymbol{\beta}}_0)$, we show $D := \mathrm{Var}(\hat{\boldsymbol{\beta}}_0) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda) \succeq 0$. Let $X^\top X = P\Lambda P^\top$ be the eigendecomposition, where $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ with $\lambda_j > 0$. Then $X^\top X + \lambda I_p = P(\Lambda + \lambda I_p)P^\top$, so

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}_0) &= \sigma^2 P\Lambda^{-1}P^\top, \\ \mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda) &= \sigma^2 P(\Lambda + \lambda I_p)^{-1}\Lambda(\Lambda + \lambda I_p)^{-1}P^\top.\end{aligned}$$

Hence

$$D = \sigma^2 P\underbrace{\left[\Lambda^{-1} - (\Lambda + \lambda I_p)^{-1}\Lambda(\Lambda + \lambda I_p)^{-1}\right]}_{=:\,\Delta}P^\top.$$

Since Δ is diagonal, its j-th entry is

$$\Delta_{jj} = \frac{1}{\lambda_j} - \frac{\lambda_j}{(\lambda_j + \lambda)^2} = \frac{(\lambda_j + \lambda)^2 - \lambda_j^2}{\lambda_j(\lambda_j + \lambda)^2} = \frac{\lambda(\lambda + 2\lambda_j)}{\lambda_j(\lambda_j + \lambda)^2} \geq 0$$

for all $\lambda \geq 0$ and $\lambda_j > 0$. Thus $\Delta \succeq 0$, and therefore $D = \sigma^2 P\Delta P^\top \succeq 0$, i.e.,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda) \preceq \mathrm{Var}(\hat{\boldsymbol{\beta}}_0). \qquad \square$$

(D) Show the bias-variance decomposition of mean-squared error(MSE):

$$\mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^\top\right] = \mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda) + (\boldsymbol{\beta} - E[\hat{\boldsymbol{\beta}}_\lambda])(\boldsymbol{\beta} - E[\hat{\boldsymbol{\beta}}_\lambda])^\top.$$

Solution: Let $\boldsymbol{\mu} := \mathrm{E}[\hat{\boldsymbol{\beta}}_\lambda]$ and write $\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta} = (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{\beta})$. Then

$$\begin{aligned}&\mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^\top\right] \\ &= \mathrm{E}\left[\Big((\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{\beta})\Big)\Big((\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{\beta})\Big)^\top\right] \\ &= \mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu})^\top\right] + \mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu})\right](\boldsymbol{\mu} - \boldsymbol{\beta})^\top + (\boldsymbol{\mu} - \boldsymbol{\beta})\mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu})^\top\right] + (\boldsymbol{\mu} - \boldsymbol{\beta})(\boldsymbol{\mu} - \boldsymbol{\beta})^\top.\end{aligned}$$

Since $\mathrm{E}[\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu}] = \mathbf{0}$, the two cross terms vanish. Therefore,

$$\mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^\top\right] = \underbrace{\mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu})^\top\right]}_{=\,\mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda)} + \underbrace{(\boldsymbol{\mu} - \boldsymbol{\beta})(\boldsymbol{\mu} - \boldsymbol{\beta})^\top}_{=\,(\boldsymbol{\beta} - \mathrm{E}[\hat{\boldsymbol{\beta}}_\lambda])(\boldsymbol{\beta} - \mathrm{E}[\hat{\boldsymbol{\beta}}_\lambda])^\top},$$

which gives the desired bias-variance decomposition. $\square$