STAT5030 Assignment 3

Due: Mar 8, 2023

1. (Underfitted Model)

   True Model: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad E(\boldsymbol{\varepsilon}) = \boldsymbol{0},\ Var(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}.$

   Mis-specified Model: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon}, \quad E(\boldsymbol{\varepsilon}) = \boldsymbol{0},\ Var(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}.$

   Let the least squares estimate of $\boldsymbol{b}$ be $\hat{\boldsymbol{b}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$,

   (a) Prove that $\hat{\boldsymbol{b}}$ is a biased estimator of $\boldsymbol{b}$ and find the bias.

   (b) Find the variance of $\hat{\boldsymbol{b}}$.

   (c) Prove that S^2 is a biased estimator of σ^2 where

   $$S^2 = \frac{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{Y}}{n - r(\boldsymbol{X})}.$$

   (d) Show that S^2 overestimate σ^2.

   (e) Find $E(\hat{\varepsilon})$ and $Var(\hat{\varepsilon})$.

2. (Over-fitted model)

   True Model: $\boldsymbol{Y} = \boldsymbol{X_1}\boldsymbol{b_1} + \boldsymbol{\varepsilon}, \quad E(\boldsymbol{\varepsilon}) = \boldsymbol{0},\ Var(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}.$

   Mis-specified Model: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix}.$

   (a) What is $E(\hat{\boldsymbol{b}})$.

   (b) Evaluate $E(S^2)$.

   (c) Evaluate $Var(\hat{\boldsymbol{b}})$.

3. Suppose $\boldsymbol{Y} \sim N_3(\boldsymbol{0},\ \boldsymbol{I})$. Find the distribution of

   $$\frac{1}{3}[(Y_1 - Y_2)^2 + (Y_2 - Y_3)^2 + (Y_3 - Y_1)^2].$$

4. Consider the full rank model

   $$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i,p-1} + \varepsilon_i, \quad i = 1, 2, 3, \cdots, n$$

   where the ε_i are i.i.d. $N(0, \sigma^2)$ and the x_{ij} are standardized so that for $j = 1, 2, \cdots, p-1$, $\sum_i x_{ij} = 0$ and $\sum_i x_{ij}^2 = c$. In matrix notation, the model can be written as

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

   Show that

   $$\frac{1}{p}\sum_{j=0}^{p-1} var(\hat{\beta}_j)$$

   is minimized when the columns of $\boldsymbol{X}$ are mutually orthogonal.

5. For the regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon}$, $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $Var(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$. Let $\hat{\boldsymbol{b}}_k = (\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$.

   (a) Prove $E(\hat{\boldsymbol{b}}_k) = \boldsymbol{b} - k(\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{b}$.

   (b) Prove that $Var(\hat{\boldsymbol{b}}_k) = \sigma^2\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-2}$.

   (c) Show that for a fixed k, the estimator $\hat{\boldsymbol{b}}_k$ is unbiased for $\boldsymbol{b}$ if and only if $k = 0$.

6. Aerial observations Y_1, Y_2, Y_3 and Y_4 are made of angles $\theta_1, \theta_2, \theta_3$ and θ_4 respectively, of a quadrilateral on the ground. Let the observations be subject to independent normal errors with zero means and common variance σ^2,

   (a) Evaluate the least squares estimates of $\theta_1, \theta_2, \theta_3$ and θ_4.

   (b) Evaluate the unbiased estimator of σ^2.

   (c) Derive a test statistic for the hypothesis that the quadrilateral is a parallelogram with $\theta_1 = \theta_3$ and $\theta_2 = \theta_4$.

7. In order to estimate two parameters θ and ϕ, it is possible to make observations of tree types:

   (a) the first type have expectation θ,

   (b) the second type have expectation $\theta + \phi$, and

   (c) the third type have expectation $\theta - 2\phi$.

   All observations are subject to independent normal errors with zeros means and common variance σ^2. If m observations of type (a), m observations of type (b) and n observations of type (c) are made, find the least squares estimates $\hat{\theta}$ and $\hat{\phi}$. Prove that these estimates are uncorrelated if $m = 2n$.

8. Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}.$$

   where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Let $\hat{\boldsymbol{\beta}}$ be the least squares estimate of $\boldsymbol{\beta}$. Define $\tilde{\boldsymbol{\beta}} = c\hat{\boldsymbol{\beta}}$ where $c \leq 1$. The mean squared error(MSE) of $\tilde{\boldsymbol{\beta}}$ is

$$MSE(\tilde{\boldsymbol{\beta}}) = E(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

   (a) Prove that $MSE(\tilde{\boldsymbol{\beta}}) = c^2\sigma^2 tr(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + (c-1)^2\boldsymbol{\beta}^\top\boldsymbol{\beta}$.

   (b) Let c^* be the value of c such that $MSE(\tilde{\boldsymbol{\beta}})$ is a minimum. Find c^*.

   (c) Let $p = 5, \sigma^2 = 1, \boldsymbol{\beta}^\top = (1, 2, 3, 4, 5)$ and the eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}$ be $1, 2, 3, 4, 5$. Evaluate c^*.

9. True model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i, \quad i = 1, 2, 3, 4, 5 \quad \varepsilon_i \sim N(0, \sigma^2), iid.$

   Mis-specified model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, 3, 4, 5 \quad \varepsilon_i \sim N(0, \sigma^2), iid.$

   Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least squares estimates of β_0 and β_1 respectively. If $x = -2, -1, 0, 1, 2$. Find the bias of $\hat{\beta}_0$ and $\hat{\beta}_1$.

10. Let

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad E(\boldsymbol{\varepsilon}) = \boldsymbol{0},\ Var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}.$$

and the LS estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$. Now, assume additional information $(\boldsymbol{u}, \boldsymbol{H})$ are available where

$$\boldsymbol{u} = \boldsymbol{H}\boldsymbol{\beta} + \gamma, \quad E(\gamma) = \boldsymbol{0},\ Var(\gamma) = \boldsymbol{W}$$

and $\boldsymbol{W}$ is known.

(a) Find the generalized least squares estimate of $\boldsymbol{\beta}$ using all available information.

(b) Let $\hat{\boldsymbol{\beta}}_a = (\boldsymbol{H}^\top \boldsymbol{W}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^\top \boldsymbol{W}^{-1}\boldsymbol{u}$ be the generalized least squares estimate of $\boldsymbol{\beta}$ based ONLY on the additional information $(\boldsymbol{u}, \boldsymbol{H})$. Further, let the generalized least squares estimate of $\boldsymbol{\beta}$ using all available information obtained in part (a) be $\hat{\boldsymbol{\beta}}_k$. Show that

$$\hat{\boldsymbol{\beta}}_k = w_1 \hat{\boldsymbol{\beta}} + w_2 \hat{\boldsymbol{\beta}}_a.$$

What are w_1 and w_2? What is the value of $|w_1 + w_2|$, the determinant of $w_1 + w_2$?