

1. [20 points] For the general linear model $Y = X\mathbf{b} + \mathbf{e}$, with uncorrelated errors having mean zero and variance \mathbf{s}^2 , suppose that the design matrix X has full rank p . Let $\hat{Y} = X\hat{\mathbf{b}}$ denote the projection of Y onto the space spanned by the columns of X . Show that $\sum_{i=1}^n \text{Var}(\hat{Y}_i) = \mathbf{s}^2 p$.

For the GLM $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, $E[\mathbf{e}] = 0$, $\text{Cov}[\mathbf{e}] = ((\text{cov}(\mathbf{e}_i, \mathbf{e}_j))) = \mathbf{s}^2 \mathbf{I}$, the estimator $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{P}\mathbf{Y}$, where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the orthogonal projection matrix onto the space spanned by the columns of \mathbf{X} , is uniquely defined, even if the column rank(\mathbf{X}) is not full. Now,

$$\text{Cov}(\hat{\mathbf{Y}}) = \text{Cov}(\mathbf{P}\mathbf{Y}) = \mathbf{P}\text{Cov}(\mathbf{Y})\mathbf{P}' = \mathbf{s}^2 \mathbf{P}\mathbf{P}'.$$

However, since \mathbf{P} is a symmetric idempotent matrix,

$$\text{Cov}(\hat{\mathbf{Y}}) = \mathbf{s}^2 \mathbf{P}\mathbf{P}' = \mathbf{s}^2 \mathbf{P}^2 = \mathbf{s}^2 \mathbf{P}.$$

Furthermore, the eigen-values $I_i, i=1,2,\dots,p$, of \mathbf{P} are either 0 or 1 [Eigen-values of \mathbf{P}^2 are equal to the square of the eigen-values of \mathbf{P} , and $\mathbf{P}^2 = \mathbf{P}$ implies that these values must be either 0 or 1]. Now,

$$\sum_{i=1}^n \text{Var}(\hat{Y}_i) = \text{trace}[\text{Cov}(\hat{\mathbf{Y}})] = \text{trace}[\mathbf{s}^2 \mathbf{P}] = \mathbf{s}^2 \text{trace}[\mathbf{P}] = \mathbf{s}^2 \sum_{i=1}^p I_i = \mathbf{s}^2 \text{rank}(\mathbf{X}).$$

The problem specified that \mathbf{X} has full column rank, so one can follow an easier approach, since $\text{trace}[\mathbf{P}] = \text{trace}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \text{trace}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = \text{trace}[\mathbf{I}_p] = p$.

2. [60 points] Let $Y_t = \mathbf{h}_t + \mathbf{e}_t$, where $\mathbf{h}_t, t \geq 0$ is a piece-wise linear function given by

$$\mathbf{h}_t = \begin{cases} \mathbf{b}_0 + \mathbf{b}_1 t, & 0 \leq t \leq t^* \\ \mathbf{b}_0^* + \mathbf{b}_2 t, & t^* \leq t \leq 10 \end{cases}$$

The parameters $\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_0^*$, and \mathbf{b}_2 are unknown. However, since the function is known to be continuous, they satisfy a restriction that $\mathbf{b}_0^* + \mathbf{b}_2 t^* = \mathbf{b}_0 + \mathbf{b}_1 t^*$, where $t^* = 5.5$ is known.

- a. [10 points] Suppose that the errors $\{\mathbf{e}_t\}$ in the observations $Y_t, t = 1, 2, \dots, 10$ have mean zero, variance \mathbf{s}^2 , and are uncorrelated. Express these observations into a general linear model form $Y = X\mathbf{b} + \mathbf{e}$, where $\mathbf{b}' = (\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2)$, i.e., define the vector \mathbf{Y} and the matrix \mathbf{X} , and the covariance matrix of the error vector.

Since the mean function is continuous, the two regression lines are equal at $t^* = 5.5$, i.e.,

$$\mathbf{b}_0^* = \mathbf{b}_0 - 5.5(\mathbf{b}_2 - \mathbf{b}_1) \text{ and the regression function reduces to}$$

$$\mathbf{h}_t = \begin{cases} \mathbf{b}_0 + \mathbf{b}_1 t, & 0 \leq t \leq 5.5 \\ \mathbf{b}_0 + \mathbf{b}_1 t + (\mathbf{b}_2 - \mathbf{b}_1)(t - 5.5), & 5.5 \leq t \leq 10 \end{cases}, \text{ or}$$

Once we are interested in inference about $(\mathbf{b}_2 - \mathbf{b}_1)$, we can use a one-to-one reparametrization from $(\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2)$ to $\mathbf{b}_0, \mathbf{b}_1$, and $(\mathbf{b}_2 - \mathbf{b}_1) = \mathbf{b}_d$ (say). Now, the model can be expressed as

$$\mathbf{h}_t = \mathbf{b}_0 + \mathbf{b}_1 t + \mathbf{b}_d (t - 5.5)^+, \quad t \in \{1, 2, \dots, 10\}, \text{ where } u^+ = \max(0, u).$$

Let the response vector be denoted by $\mathbf{Y}' = (Y_1, Y_2, \dots, Y_{10})$, the error vector by $\mathbf{e}' = (e_1, e_2, \dots, e_{10})$, with $\text{Cov}(\mathbf{e}) = \mathbf{S}^2 I_{10}$, then for the parameter vector $\mathbf{b}' = \{\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_d\}$, the design matrix \mathbf{X} is as follows:

$$\mathbf{X}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 1.5 & 2.5 & 3.5 & 4.5 \end{pmatrix}.$$

- b. [10 points] Obtain the normal equations for this model.

The normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$ for this problem are given by

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 10 & 55 & 12.5 \\ 55 & 385 & 110 \\ 12.5 & 110 & 41.25 \end{pmatrix}, \text{ and } \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^{10} Y_i \\ \sum_{i=1}^{10} iY_i \\ \sum_{i=6}^{10} (i-5.5)Y_i \end{pmatrix}.$$

- c. [5 points] Explain why $\mathbf{b}_1 - \mathbf{b}_2$ is estimable?

Since the matrix \mathbf{X} has full column rank, every component of $\mathbf{b}' = \{\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_d\}$ is estimable, hence $\mathbf{b}_1 - \mathbf{b}_2 = -\mathbf{b}_d$ is estimable.

- d. [10 points] Find the Gauss-Markov estimator $\hat{\mathbf{b}}_1 - \hat{\mathbf{b}}_2$ by solving the normal equations for $\mathbf{b}_1 - \mathbf{b}_2$ without necessarily finding the inverse of $\mathbf{X}'\mathbf{X}$.

Observe that for this system of equations, $(3 \quad -1 \quad 2)\mathbf{X}'\mathbf{X} = (0 \quad 0 \quad 10)$

Therefore,

$$\begin{aligned} \hat{\mathbf{b}}_d &= \frac{1}{10} \left[3 \sum_{i=1}^{10} Y_i - \sum_{i=1}^{10} iY_i + 2 \sum_{i=6}^{10} (i-5.5)Y_i \right] \\ &= \frac{1}{10} \left[\sum_{i=1}^5 (3-i)Y_i + \sum_{i=6}^{10} (i-8)Y_i \right]. \end{aligned}$$

Now, $\hat{\mathbf{b}}_1 - \hat{\mathbf{b}}_2 = -\hat{\mathbf{b}}_d$

- e. [10 points] Find the variance of $\hat{\mathbf{b}}_1 - \hat{\mathbf{b}}_2$.

Given the $\hat{\mathbf{b}}_d$ above and the uncorrelated Y's,

$$\text{Var}(-\hat{\mathbf{b}}_d) = \frac{\mathbf{s}^2}{100} \left[\sum_{i=1}^5 (3-i)^2 + \sum_{i=6}^{10} (i-8)^2 \right] = \frac{\mathbf{s}^2}{100} [10+10] = \mathbf{s}^2 / 5.$$

- f. [5 points] If the slope change point t^* was not known, would this model be still linear? Explain.
- No, in that case t^* would also be an unknown parameter. The model would not be linear in t^* .
- g. [10 points] For the reduced model for this problem, under the restriction $\mathbf{b}_1 - \mathbf{b}_2 = 1$, write down the normal equations for estimating the parameters $\mathbf{b}_0, \mathbf{b}_1$.

Given $\mathbf{b}_d = -1$, the model reduces to

$$\mathbf{h}_t = \mathbf{b}_0 + \mathbf{b}_1 t - (t - 5.5)^+, \text{ or } \mathbf{h}_t + (t - 5.5)^+ = \mathbf{b}_0 + \mathbf{b}_1 t,$$

where $u^+ = \max(0, u)$, for $t \in \{1, 2, \dots, 10\}$.

Now for the response variable $Y_t^* = Y_t + (t - 5.5)^+$, we have the linear model

$$Y_t^* = \mathbf{b}_0 + \mathbf{b}_1 t, \quad t \in \{1, 2, \dots, 10\}.$$

The corresponding normal equations are given by

$$\mathbf{X}^{*\prime} \mathbf{X}^* = \begin{pmatrix} 10 & 55 \\ 55 & 385 \end{pmatrix} \text{ and } \mathbf{X}^{*\prime} \mathbf{Y}^* = \begin{pmatrix} \sum_{i=1}^{10} Y_i^* \\ \sum_{i=1}^{10} i Y_i^* \end{pmatrix}.$$

3. [5 points each] Explain why each of the following statement is True or False. If you provide incorrect explanation, even though your make correct choice, you will not get any points.

- a. True/False. In a general linear model, $\text{rank}(\mathbf{X})$ is full if and only if every component of the vector \mathbf{b} is estimable.

TRUE. If $\text{rank}(\mathbf{X})$ is full, every linear function of the parameters is estimable. In particular, every component of the parameter vector is estimable. Conversely, if every component of the parameter vector is estimable, it implies that every linear combination of the parameter vector is estimable. Hence the row space of \mathbf{X} has full rank.

- b. True/False. Let $\hat{\mathbf{y}} = \mathbf{c}' \hat{\mathbf{b}}$ be the B.L.U.E. estimator of an estimable function $\mathbf{y} = \mathbf{c}' \mathbf{b}$, and let $\hat{\mathbf{y}}$ be any other linear estimator of \mathbf{y} , then $\text{Var}(\hat{\mathbf{y}}) < \text{Var}(\hat{\mathbf{y}})$.

FALSE. Even though $\hat{\mathbf{y}} = \mathbf{c}' \hat{\mathbf{b}}$ is the best linear estimator among all unbiased estimators, $\hat{\mathbf{y}}$ could be a biased estimator.

- c. True/False. In the sample linear regression model $Y_i = \mathbf{m} + \mathbf{b}(x_i - \bar{x}) + \mathbf{e}_i, i = 1, 2, \dots, n$, where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, the best unbiased linear estimator of \mathbf{m} is the sample mean \bar{Y} , which is the projection of the vector Y onto the space generated by the vectors $\mathbf{1}$ and $\bar{\mathbf{x}}$, where $\mathbf{1}' = (1, 1, \dots, 1)$ and $\bar{\mathbf{x}} = \bar{x}\mathbf{1}$.

FALSE. Even though the B.L.U.E. of \mathbf{m} is the sample mean \bar{Y} , but $\bar{Y} \neq \bar{Y}\mathbf{1}$, which is the projection of the vector Y onto the space generated by the vectors $\mathbf{1}$ and $\bar{\mathbf{x}}$, where $\mathbf{1}' = (1, 1, \dots, 1)$ and $\bar{\mathbf{x}} = \bar{x}\mathbf{1}$. {Note that space is same as the space generated by the vector $\mathbf{1}$.}

- d. True/False. In the sample model $Y_i = \mathbf{b} x_i + \mathbf{e}_i, i = 1, 2, \dots, n$, with errors $\{\mathbf{e}_i\}$ having mean zero, variance \mathbf{s}^2 , and are uncorrelated, the B.L.U.E. of the parameter \mathbf{b} is given by the ratio estimator $\hat{\mathbf{b}} = \sum_{i=1}^n Y_i / \sum_{i=1}^n x_i$.

FALSE. In the no intercept model, it is easy to prove that $\hat{\mathbf{b}} = \sum_{i=1}^n x_i Y_i / \sum_{i=1}^n x_i^2$.