Preliminary Examination: Linear Models

Q1 counts for 60 points and Q2, 40 points.

Answer questions by showing all of your work.

1. (60 points) For a linear model $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, consider the centered model $y_i = \alpha + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_k(x_{ik} - \bar{x}_k) + \epsilon_i$ where $\bar{x}_j = \sum_{i=1}^n x_{ij}/n$ for j=1, …, k, which is also written by $\boldsymbol{y} = [\mathbf{1}\ \ \boldsymbol{X}_c]\begin{pmatrix}\alpha \\ \boldsymbol{\beta}_1\end{pmatrix} + \boldsymbol{\varepsilon}$ where $\boldsymbol{X}_c = \begin{pmatrix} x_{11} - \bar{x}_1 & \dots & x_{1k} - \bar{x}_k \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \dots & x_{nk} - \bar{x}_k \end{pmatrix}$.

   (a) Prove that the (ordinary) least squares estimators for α and $\boldsymbol{\beta}_1$ are $\bar{y}$ and $(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\boldsymbol{y}$, respectively.

   (b) Find the distribution of SSR_c/σ^2 where $\text{SSR}_c = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$. Check all the conditions if you apply a theorem you memorize.

   (c) Prove that $\hat{\boldsymbol{\varepsilon}} \sim \text{N}(\mathbf{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{P}))$ where $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1\ \ \dots \hat{\varepsilon}_n)^\top$ and $\hat{\varepsilon}_i = y_i - \hat{y}_i$ after defining $\boldsymbol{P}$ appropriately.

   (d) Find the distribution of SSE/σ^2 where $\text{SSE} = \sum_{i=1}^n \hat{\varepsilon}_i^2$. Check all the conditions if you apply a theorem you memorize.

   (e) The response variables can be written as $\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}$ from the original model where $\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y}$ and $\boldsymbol{y} = \hat{\alpha}\mathbf{1} + \hat{\boldsymbol{y}}_c + \hat{\boldsymbol{\varepsilon}}$ from the centered form where $\hat{\boldsymbol{y}}_c = \boldsymbol{P}_c\boldsymbol{y}$.

      i. Show that $\boldsymbol{P}_c\boldsymbol{y} = \left(\boldsymbol{P} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}$.

      ii. For SSE defined in part (d), show that $\text{SSE} = \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{P}_c - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}$.

      iii. Show that SSR_c defined in part (b) and SSE defined in part (d) are independent.

2. (40 points) We consider the linear regression model $\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\boldsymbol{X} = (\boldsymbol{X}_1,\ \boldsymbol{X}_2)$$

for some $\boldsymbol{X}_1 \in \mathbb{R}^{n\times p_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n\times p_2}$. To accommodate for this block matrix form, we write $\boldsymbol{\beta} \in \mathbb{R}^{p_1+p_1}$ as,

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix},$$

where $\boldsymbol{\beta}_1 \in \mathbb{R}^{p_1}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{p_2}$. Throughout this question, assume that the design matrix $\boldsymbol{X}$ is deterministic with full rank and $n > p_1 + p_2$. The error vector $\boldsymbol{\varepsilon}$ has multivariate normal distribution with zero mean and a diagonal covariance matrix with common diagonal element σ^2.

Let

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p_1+p_2}}{\arg\min} \|\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta}\|,\ \hat{\boldsymbol{\beta}}_1 = \underset{\boldsymbol{\beta}_1\in\mathbb{R}^{p_1}}{\arg\min} \|\mathbf{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1\|,\ \hat{\boldsymbol{\beta}}_2 = \underset{\boldsymbol{\beta}_2\in\mathbb{R}^{p_2}}{\arg\min} \|\mathbf{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2\|,$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector.

(a) Construct an example of $(\mathbf{Y},\ \boldsymbol{X})$ where $\hat{\boldsymbol{\beta}}^T \neq (\hat{\boldsymbol{\beta}}_1^T,\ \hat{\boldsymbol{\beta}}_2^T)$.

(b) Prove that $\hat{\boldsymbol{\beta}}^T = (\hat{\boldsymbol{\beta}}_1^T,\ \hat{\boldsymbol{\beta}}_2^T)$ if $\boldsymbol{X}_1^T\boldsymbol{X}_2$ is a zero matrix.

(c) Prove that $\hat{\boldsymbol{\beta}}_1^T$ and $\hat{\boldsymbol{\beta}}_2^T$ are independent if $\boldsymbol{X}_1^T\boldsymbol{X}_2$ is a zero matrix.

(d) Prove that $\|\mathbf{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1\|^2+\|\mathbf{Y}-\boldsymbol{X}_2\boldsymbol{\beta}_2\|^2 \geq \|\mathbf{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2$ if $\boldsymbol{X}_1^T\boldsymbol{X}_2$ is a zero matrix.

(e) Construct an example of $(\mathbf{Y},\ \boldsymbol{X})$ where

$$\|\mathbf{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1\|^2 + \|\mathbf{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2\|^2 = \|\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2.$$