Preliminary Examination: LINEAR MODELS

Make sure to **show all your work, formulas and justifications and explain all answers clearly and fully in proper English.** You cannot get full credit unless you show your work, and partial credit will be granted based on work shown.

1. Answer the following two parts.

   a. If $\mathbf{Y} = \mathbf{1}\mu + \boldsymbol{\varepsilon}$, where $var(\boldsymbol{\varepsilon}) = \sigma_1^2\mathbf{I} + \sigma_2^2\mathbf{1}\mathbf{1}'$ for some $\sigma_1^2, \sigma_2^2 > 0$, prove that the BLUE of μ is $\bar{Y}$, the average of the entries in the vector $\mathbf{Y}$.

   b. Use the fact in part (a), along with whatever else you may know about best linear unbiased estimation, to compute the BLUE of μ in the following situation.

   Suppose

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \\ \mu \\ \mu \\ \mu \end{pmatrix}, \begin{pmatrix} 5 & 1 & 1 & 1 & 0 \\ 1 & 5 & 1 & 1 & 0 \\ 1 & 1 & 5 & 1 & 0 \\ 1 & 1 & 1 & 5 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} \right),$$

   and we observe

$$\mathbf{y} = \begin{pmatrix} 8 \\ 3 \\ 8 \\ 5 \\ 9 \end{pmatrix}.$$

   Note that finding the BLUE of μ in this situation does not require inverting any matrices. All the computations needed to obtain the BLUE can be carried out by hand easily.

2. The problem of selecting a model with the best predictive ability among a class of linear models is examined. Consider a linear regression model with three covariates,

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + \epsilon_i;\ i = 1, \ldots, n,$$

   where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed observations from a normal distribution with mean 0 and variance σ^2. Suppose that the true regression coefficient vector is $\boldsymbol{\beta} \equiv (\beta_1, \beta_2, \beta_3)'$.

Consider the following three candidate models:

$$\text{Model 1: } Y_i = X_{i1}\beta_1 + \epsilon_i$$
$$\text{Model 2: } Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i$$
$$\text{Model 3: } Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + \epsilon_i$$

Let $\hat{\boldsymbol{\beta}}_k$ be the least squares estimator of $\boldsymbol{\beta}$ under Model k for $k = 1, 2, 3$.

For theoretical comparison of the prediction accuracy of the three models with estimated coefficients, suppose that the future value of a response variable when $\mathbf{x} = \mathbf{x}_i \equiv (x_{i1}, x_{i2}, x_{i3})'$ is available and denoted by z_i, $i = 1, \ldots, n$. Then the average squared prediction error (PE) of the fitted Model k is defined as

$$PE_k \equiv \frac{1}{n}\sum_{i=1}^{n}(z_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_k)^2.$$

a. Show that the expected average squared prediction error conditional on the data $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, is

$$E_{Z_1,\ldots,Z_n}(PE_k) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_k)^2.$$

b. Now consider the overall unconditional expected prediction error of Model k (yet conditional on $\mathbf{x}_i$, $i = 1, \ldots, n$) that is defined as:

$$\Gamma_{k,n} \equiv E_{Y_1,\ldots,Y_n}(E_{Z_1,\ldots,Z_n}(PE_k)).$$

b-i Verify that

$$\Gamma_{k,n} = \sigma^2 + k\sigma^2/n + \boldsymbol{\beta}'\mathbf{X}(\mathbf{I} - \mathbf{P}_k)\mathbf{X}\boldsymbol{\beta}/n$$

where $\mathbf{X}$ is the full design matrix, and $\mathbf{P}_k \equiv \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'$ with $\mathbf{X}_k$ being the $n \times k$ design matrix for Model k.

b-ii Assume that the first two covariates are sufficiently different in the sense that $\|(\mathbf{I} - \mathbf{P}_1)\mathbf{X}_{[2]}\|^2 = O(n)$, where $\mathbf{X}_{[2]}$ is the second column of $\mathbf{X}$. Moreover, assume that the value of $\boldsymbol{\beta}$ is $(\beta_1^*, \beta_2^*, 0)'$ with $\beta_1^* \neq 0$ and $\beta_2^* \neq 0$. For large n, which model has the smallest unconditional expected prediction error and why?

3. Suppose that the true model is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon},$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. We choose to fit the *incorrect* model:

(1) $$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

where we also assume $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

Show that

$$E(SSE) = \sigma^2(n - r(\mathbf{X}))$$

and $\tilde{\sigma}^2$ is unbiased, where SSE is from fitting model (1) and $\tilde{\sigma}^2$ is the OLS estimator of σ^2 when fitting (1).

4. A state highway department studied the wear characteristics of five different paints at eight locations in the state. The standard, currently used paint (paint 1) and four experimental paints (paints 2, 3, 4, 5) were included in the study. *The eight locations were randomly selected, thus reflecting variations in traffic densities throughout the state*. At each location, a random ordering of the paints to the chosen road surface was employed. After a suitable period of exposure to weather and traffic, a combined measure of wear, considering both durability and visibility, was obtained. The data on wear $\{Y_{ij}\}$ are below:

Location	Paint (j)				
i	1	2	3	4	5
1	11	13	10	18	15
2	20	28	15	30	18
3	8	10	8	16	12
4	30	35	27	41	28

Location	Paint (j)				
i	1	2	3	4	5
5	14	16	13	22	16
6	25	27	26	33	25
7	43	46	41	55	42
8	13	14	12	20	13

a. State an appropriate statistical model including model assumptions.

b. Complete the ANOVA table below.

Source	df	SS	MS	$E(MS)$
Blocks	?	4826.375	?	?
Paint types	?	531.350	?	?
Error	?	122.250	?	?
Total	?	?		

c. We would like to test whether or not the mean wear differs for the five paints. State the null and alternative hypotheses, test statistic, its distribution under null hypothesis and associated degrees of freedom and decision rule.

d. From the data we obtain $\bar{Y}_{\cdot 1} = 20.5$, $\bar{Y}_{\cdot 2} = 23.625$, $\bar{Y}_{\cdot 3} = 19.0$, $\bar{Y}_{\cdot 4} = 29.375$, $\bar{Y}_{\cdot 5} = 21.125$. Now assume that Paints 1, 3, and 5 are white, whereas paints 2 and 4 are yellow. Estimate the difference in the mean wear for the two groups of paints as well as providing a 95% confidence interval. Interpret your findings.