Preliminary Examination:
LINEAR MODELS

Answer all questions and show all work.

1. The spring balance weighing model is the following. N objects are available to be weighed on a scale. Their weights are to be determined by a series of b weighings. A weighing consists of placing any collection of the objects on the scale and then reading the weight Y of this collection of objects. Assume that the weights can be modeled by the linear model

$$Y_j = \sum_{i=1}^{N} w_i x_{ij} + \epsilon_j,\ 1 \le i \le N,\ 1 \le j \le b,$$

where Y_j is the observed weight on weighing j, w_i is the unknown weight of object i, and

$$x_{ij} = \begin{cases} 1, & \text{if object } i \text{ is used in weighting } j \\ 0, & \text{otherwise.} \end{cases}$$

The $\{\epsilon_j\}$ are assumed to be independent and identically distributed normal random variables with mean 0 and unknown variance σ^2.

a. Assuming $b = N$. Find the least-squares estimates of $\{w_j\}$ and the variances of these estimates *if only object j is weighed in weighing j.*

b. Assuming $b = N$. Find the least-squares estimates of $\{w_j\}$ and the variances of these estimates *if ALL objects except object j are weighed in weighing j.*

c. Of the methods described in parts (a) and (b), which would you recommend? Why?

2. Suppose that the following linear regression model is postulated

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

where $\mathbf{Y}$ is an $n \times 1$ vector of random variables, $\mathbf{X}_1$ is an $n \times p$ full rank matrix of known constants, $\boldsymbol{\beta}_1$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ with σ^2 unknown. Let $\mathbf{b}_1$ be the least squares estimator of $\boldsymbol{\beta}_1$ under the postulated model, and let $\hat{\mathbf{Y}} = \mathbf{X}_1\mathbf{b}_1$.

Suppose that the true model is actually

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2,$$

where $\mathbf{X}_2$ is an $n \times q$ matrix of known constants, $\boldsymbol{\beta}_2 \neq \mathbf{0}$ is a $q \times 1$ vector of unknown parameters, and $\boldsymbol{\varepsilon}_2$ has the same distribution as $\boldsymbol{\varepsilon}$.

a. Show that $\mathbf{b}_1$ is generally a biased estimator of $\boldsymbol{\beta}_1$. State any conditions under which $\mathbf{b}_1$ is unbiased.

b. Find the covariance matrix of $\mathbf{b}_1$.

c. Consider the decomposition of the total sum of squares of $\mathbf{Y}$ into regression sum of squares (SSR) and residual sum of squares (SSE) for the postulated model. In other words, $SSR = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$, and $SSE = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}})$. Find the expected value of SSE. Simplify the expression as much as possible. Is $MSE \equiv SSE/(n-p)$, the usual estimator of σ^2 unbiased?

d. To test the hypothesis $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ against $H_a : \boldsymbol{\beta} \neq \mathbf{0}$, suppose that we use the usual F-test based on the test statistic $F = \frac{SSR/p}{SSE/(n-p)}$, which assumes incorrectly that the postulated model is true. What are the actual distributions of SSR and SSE under H_0? Comment on the validity of the F-test.

e. Consider the least squares estimator of $\mathbf{b}_1^*$ of $\boldsymbol{\beta}_1$ under the *true model* assuming that $\mathbf{X} = [\mathbf{X}_1 \;\; \mathbf{X}_2]$ is of full rank. When $\boldsymbol{\beta}_2 = \mathbf{0}$, compare $\mathbf{b}_1$ and $\mathbf{b}_1^*$ in terms of bias and variance.

Hint: The following matrix identity may be useful:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{I} \end{bmatrix} (\mathbf{C} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1} \left[-\mathbf{B}'\mathbf{A}^{-1} \quad \mathbf{I}\right]$$

3. Let $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $j = 1, \ldots, n$, $i = 1, 2, 3$, and $\epsilon_{ij} \sim N(0, \sigma^2)$. Derive a test for $H : \tau_2 = (\tau_1 + \tau_3)/2$.

4. An experiment was conducted to study mean plant height of two genotypes exposed to three watering levels. The experiment was conducted in 4 greenhouses. Each greenhouse contained three tables. On each table, were 2 pots with 1 plant in each pot. The 2 plants on any given table consisted of 1 plant of one genotype and 1 plant of the other genotype, with genotypes randomly assigned to the pots. Within each greenhouse, the three watering levels were randomly assigned to the three tables, with one table per watering level. Thus, the two plants on any given table received the same amount of water throughout the experiment. At the conclusion of the experiment, the height of each plant was recorded.

For $i = 1, 2, 3, 4$; $j = 1, 2, 3$; and $k = 1, 2$; let y_{ijk} denote the height recorded for the plant associated with greenhouse i, watering level j, and genotype k. Consider the following model that will be referred to henceforth as MODEL 1.

$$y_{ijk} = \mu + g_i + w_j + t_{ij} + \gamma_k + \phi_{jk} + \epsilon_{ijk}, \;\; i = 1, 2, 3, 4; j = 1, 2, 3; k = 1, 2;$$

where the g_i terms are $N(0, \sigma_g^2)$, the t_{ij} terms are $N(0, \sigma_t^2)$, the ϵ_{ijk} terms are $N(0, \sigma^2)$. All these random terms are mutually independent, and the remaining terms in the model are unknown fixed parameters.

a. According to MODEL 1, what is the *correlation* between the heights of two plants growing together on the same table?

b. In terms of parameters in MODEL 1, write down the null hypothesis of no watering level main effects.

c. Now suppose that data have been collected and analyzed using MODEL 1. Let `GH`, `WL`, and `GENO` represent factors greenhouse, watering level, and genotype respectively. For the following ANOVA table, complete the column of degrees of freedom.

```
Analysis of Variance Table
Response: y
               Df        Sum Sq
GH                       113.3
WL                       321.8
GENO                       2.5
GH:WL                    116.4
GH:GENO                   11.7
WL:GENO                   75.1
Error                     14.5
```

d. Suppose our goal is to analyze the data under the assumption that MODEL 1 is correct. Use information in the output to compute the two F-statistics that can be used to test for watering level main effects and genotype main effects, respectively.