4. Suppose that data \( \{(x_{ij}, y_{ij}): i = 1, \ldots, k, j = 1, \ldots, J\} \) can be modeled as having common slopes and possibly different intercepts using the linear model,

\[
Y_{ij} = \beta_i + \gamma x_{ij} + \epsilon_{ij},
\]

where \( \{\epsilon_{ij}\} \) are independently and identically distributed \( N(0, \sigma^2) \) random variables. Assume that no vector \( (x_{i1}, \ldots, x_{ij}) \), for \( i = 1, \ldots, k \), is proportional to the vector of 1s.

(a) [8 pts] Determine the ordinary-least-squares estimator of \( (\beta_1, \ldots, \beta_k, \gamma)^T \).

(b) [6 pts] Give an explicit expression for the size \( \alpha \) likelihood-ratio test of the hypothesis,

\[
H_O: \beta_1 = \cdots = \beta_k = 0 \quad \text{versus} \quad H_A: \text{not } H_O.
\]

(c) [9 pts] Compute the power of the test that you derived in part (b). (There are several ways of defining the non-centrality parameter for the test that you derived, for example, in terms of their scale factor. Pick any one of these, and use it consistently in this part and part (d).) Show that the power is independent of \( \gamma \).

(d) [2 pts] State the power of the test when \( \beta_1 = \cdots = \beta_k = 0 \) and \( \gamma = 2 \).
4. (a) [8 pts] Setting \( Y^T \equiv (Y_{11}, \ldots, Y_{1J}, \ldots, Y_{kJ}) \), we have
\[
Y = X\beta + \epsilon
\]
with
\[
X = \begin{pmatrix} I_k \otimes 1_J | x \end{pmatrix}
\]
an \( kJ \times (k + 1) \) matrix, \( I_k \) is a matrix order \( k \), \( \otimes \) denotes Kronecker product, \( 1_J \) is the unit column vector of length \( J \), \( x = (x_{11}, \ldots, x_{kJ})^T \), \( \beta \equiv (\beta_1, \ldots, \beta_k, \gamma)^T \), and \( \epsilon = (\epsilon_{11}, \ldots, \epsilon_{kJ})^T \). Solving the normal equations, or otherwise, yields
\[
\hat{\beta}_i = Y_i - \hat{\gamma} x_i, \quad i = 1, \ldots, k,
\]
and
\[
\hat{\gamma} = \frac{\sum_{i=1}^k \left( \sum_{j=1}^J Y_{ij}(x_{ij} - \bar{x}_i) \right)}{\sum_{i=1}^k s_{i,x}^2},
\]
where \( s_{i,x}^2 = \sum_{j=1}^J (x_{ij} - \bar{x}_i)^2 \).

(b) [6 pts] The size \( \alpha \) likelihood-ratio test rejects \( H_O \) if and only if
\[
F = \frac{\| Y_O - Y_A \|^2 / (k - 1)}{\| Y - Y_A \|^2 / (kJ - k - 1)} \geq F_{k,kJ-k-1}^\alpha,
\]
where \( \| \cdot \|^2 \) is the squared difference of two vectors; \( Y_O \) is the vector of estimated \( Y \)-means under \( H_O \), so that the estimated \( E \{ Y_{ij} \} \) value is equal to \( \hat{\gamma}_O x_{ij} \), where
\[
\hat{\gamma}_O = \frac{\sum_{i=1}^k \sum_{j=1}^J Y_{ij}x_{ij}}{\sum_{i=1}^k \sum_{j=1}^J x_{ij}^2};
\]
\( Y_A \) is the vector of estimated \( Y \)-means under \( H_A \), so that the estimated \( E \{ Y_{ij} \} \) is equal to \( \hat{\beta}_i + \hat{\gamma} x_{ij} \) given in the solution to part (a); and \( F_{s,r}^\alpha \) is the upper-\( \alpha \) critical point of the \( F \) distribution with \( s \) and \( r \) degrees of freedom.

(c) [9 pts] The power of the F test when the true intercept vector is \((\beta_i, \ldots, \beta_k)\) and the slope is \( \gamma \) is
\[
\pi(\beta_i, \ldots, \beta_k, \gamma) = P \left\{ F \geq F_{k,kJ-k-1}^\alpha(\delta^2) \right\},
\]
where the non-centrality parameter, \( \delta \), is specified by
\[
2\sigma^2 \delta^2 = \| Y_O - Y_A \|^2,
\]
where \( Y_{ij} \) in the expression above is replaced with \( \beta_i + \gamma x_{ij} \).
Expressing \( \| \mathbf{Y}_O - \mathbf{Y}_A \|^2 \) in terms of the \( \{ Y_{ij} \} \), we have

\[
\| \mathbf{Y}_O - \mathbf{Y}_A \|^2 = \sum_{i=1}^{k} \sum_{j=1}^{J} \left( \sum_{r=1}^{k} \sum_{t=1}^{J} \frac{Y_{rt} x_{rt}}{c_1} x_{ij} - \bar{Y}_i - \gamma (x_{ij} - \bar{x}_i) \right)^2
\]

\[
= \sum_{i=1}^{k} \sum_{j=1}^{J} \left( \sum_{r=1}^{k} \sum_{t=1}^{J} \frac{Y_{rt} x_{rt}}{c_1} x_{ij} - \frac{1}{J} \sum_{t=1}^{J} Y_{it} - (x_{ij} - \bar{x}_i) \right)^2
\]

where \( c_1 = \sum_{r=1}^{k} \sum_{t=1}^{J} x_{rt}^2 \) and \( c_2 = \sum_{r=1}^{k} \sum_{t=1}^{J} (x_{rt} - \bar{x}_r)^2 \). Then replacing \( Y_{ij} \) with \( \beta_i + \gamma x_{ij} \), gives

\[
2\sigma^2 \delta^2 = \sum_{i=1}^{k} \sum_{j=1}^{J} \left( \sum_{r=1}^{k} \sum_{t=1}^{J} \frac{\beta_r + \gamma x_{rt}}{c_1} x_{ij} - \frac{1}{J} \sum_{t=1}^{J} \beta_i + \gamma x_{it} \right)
\]

\[
- \frac{(x_{ij} - \bar{x}_i)}{c_2} \sum_{r=1}^{k} \sum_{t=1}^{J} (\beta_r + \gamma x_{rt})(x_{rt} - \bar{x}_r)
\]

\[
= \sum_{i=1}^{k} \sum_{j=1}^{J} \left( \frac{x_{ij}}{c_1} \sum_{r=1}^{k} \beta_r \sum_{t=1}^{J} x_{rt} - \beta_i - \frac{(x_{ij} - \bar{x}_i)}{c_2} \sum_{r=1}^{k} \beta_r (x_{rt} - \bar{x}_r)
\]

\[
+ \frac{x_{ij} \gamma}{c_1} \sum_{r=1}^{k} \sum_{t=1}^{J} x_{rt}^2 - \gamma (x_{ij} - \bar{x}_i) \sum_{r=1}^{k} \sum_{t=1}^{J} (x_{rt} - \bar{x}_r)
\]

\[
= \sum_{i=1}^{k} \sum_{j=1}^{J} \left( \frac{x_{ij}}{c_1} \sum_{r=1}^{k} \beta_r \sum_{t=1}^{J} x_{rt} - \beta_i - \frac{(x_{ij} - \bar{x}_i)}{c_2} \sum_{r=1}^{k} \beta_r (x_{rt} - \bar{x}_r)
\]

\[
+ \gamma (x_{ij} - \bar{x}_i) - \gamma (x_{ij} - \bar{x}_i)
\]

\[
= \sum_{i=1}^{k} \sum_{j=1}^{J} \left( \frac{x_{ij}}{c_1} \sum_{r=1}^{k} \beta_r \sum_{t=1}^{J} x_{rt} - \beta_i - \frac{(x_{ij} - \bar{x}_i)}{c_2} \sum_{r=1}^{k} \beta_r (x_{rt} - \bar{x}_r)
\]

\[
+ \gamma (x_{ij} - \bar{x}_i) - \gamma (x_{ij} - \bar{x}_i)
\]
which is independent of $\gamma$. (Other definitions of the non-centrality parameter are monotone functions of $\delta^2$.)

A more elegant method of solving this problem is to use the fact that $Y_O = P_O Y$ and $Y_A = P_A Y$, where $P_O$ and $P_A$ are the projection matrices onto the column spaces of the $kJ \times 1$ and $kJ \times (k + 1)$ matrices.

$$X_O = \begin{pmatrix} x \end{pmatrix} \quad \text{and} \quad X_A = \begin{pmatrix} I_k \otimes 1_J \mid x \end{pmatrix},$$

respectively. Both of these projections can be easily computed, the second by forming unit vectors that are pairwise orthogonal vectors using the Gram-Schmidt algorithm. Hence,

$$2\sigma^2 \delta^2 = \| P_O E\{Y\} - P_A E\{Y\} \|.$$

(d) [2 pts] Setting $\beta_1 = \cdots = \beta_k = 0$ in the expression in part (c) above for $2\sigma^2 \delta^2$, we immediately obtain zero for the non-centrality parameter, no matter what value $\gamma$ takes. This result shows that the test has size $\alpha$. 


6. An investigator proposes an unusual design to study conditions under which bacteria flourish. The experiment consists of two phases. Each phase of the experiment takes one week. The two phases are run over consecutive weeks. Only 10 petrie dishes can be experimented on in a given week.

Phase 1: In a temperature-controlled environment, the experimenter will grow 10 petrie dishes of bacteria under the null condition, which is called condition 0. It takes one week to grow these bacteria. At the end of the week, she will measure the bacteria count in each dish, say \( W_{0t} \), \( t = 1, \ldots, 10 \).

Phase 2: In the same temperature-controlled environment, the experimenter will grow 10 petrie dishes of bacteria. Each dish will be grown under its own unique condition (conditions 1, \ldots, 10), leading to the data \( W_{i1} \), \( i = 1, \ldots, 10 \).

The experimenter believes that the data are most appropriately analyzed on a transformed scale and wishes to work with \( Y_{it} = \log(W_{it}) \), where \( \log \) represents the natural logarithm. The model for the data is,

\[
Y_{it} = \mu_i + \epsilon_{it},
\]

for appropriate \( i \) and \( t \), where \( \{\mu_i\} \) are fixed but unknown constants, and \( \{\epsilon_{it}\} \) are independently and identically distributed \( N(0, \sigma^2) \) random variables.

Later in the question, \( \eta_i \) will denote the median of the distribution of \( Y_{it} \). Use \( \mu_i W \) to denote the mean of \( W_{it} \).

(a) [3 pts] Under the experimenter’s assumptions, provide a formula for a 95% confidence interval for \( \mu_1 - \mu_2 \).

(b) [6 pts] Suppose that the 10 treatments examined in Phase 2 of the experiment have a \( 2 \times 5 \) factorial structure. The level of Factor A is 1 for conditions 1 through 5 and it is 2 for conditions 6 through 10. The level of Factor B is \( 1 + i (\text{mod} 5) \) for condition \( i; i = 1, \ldots, 10 \). Under a main-effects model, but relying only on the estimate of \( \sigma^2 \) from Phase 1, provide a formula for a 95% confidence interval for \( \mu_1 - \mu_6 \). (Points will be deducted for a very wide interval!)

(c) [4 pts] Describe a graphical summary that would help you decide whether the main-effects model is sufficient, or whether there is a need to expand the model. Sketch pictures illustrating a situation where there is no need to expand the model and others illustrating a situation where there is a need to expand the model. Your artwork need not be perfect, but it should convey your point.

(d) [3 pts] Describe a hypothesis-testing method that would help you decide whether the main-effects model is adequate, or whether there is a need to expand the model.

(e) [4 pts] Explicitly describing appropriate assumptions, transform the interval in part (a) into a confidence interval for the ratio of the mean of \( W_{1t} \) to the mean of \( W_{2t} \). Can this interval be interpreted as an approximate 95% confidence interval for the ratio of the mean of \( W_{1t} \) to the mean of \( W_{2t} \)? Explain briefly.

(f) [5 pts] This experiment can (and should) be redesigned. Given the constraint of 10 petrie dishes per week and two weeks for the experiment, propose a design that would allow the investigator some possibility of examining the assumptions that
underlie the analysis, along with a means of estimating and testing for main effects of factors A and B. Along with your design, provide a skeleton of an ANOVA table, listing sources of variation and degrees of freedom for your design.
6. (a) [3 pts] The first week’s data provide an estimate of $\sigma^2$. The remainder of the part is the usual normal-theory model. A 95% CI for $\mu_1 - \mu_2$ is

$$Y_{11} - Y_{21} \pm t_{0.025}S\sqrt{2},$$

where

$$S^2 = \frac{1}{9} \sum_{i=1}^{10} (Y_{0i} - \bar{Y}_0)^2.$$

(b) [6 pts] Under the main-effects model, $\mu_1 - \mu_6$ is the effect of Factor A, as is $\mu_2 - \mu_7$, etc. Consequently, a 95% CI for this effect is

$$\frac{1}{5} \sum_{i=1}^{5} Y_{i1} - \frac{1}{5} \sum_{i=6}^{10} Y_{i1} \pm t_{0.025}S\sqrt{\frac{1}{5} + \frac{1}{5}}.$$

(c) [4 pts] An interaction plot provides a graphical means of assessing the need for an interaction term in the model. To use the plot, one assesses the parallelism of the curves. The pair of plots on the left do not show evidence of interaction. The pair of plots on the right show evidence of interaction.
(d) [3 pts] The two main methods are (i) the F-test for the presence of interactions and (ii) Tukey’s 1 df test for nonadditivity.

(e) [4 pts] Assume that the normal-theory model from the initial problem description holds. Define

\[ L = Y_{11} - Y_{21} - t_{0.025}S\sqrt{2}, \]

and

\[ U = Y_{11} - Y_{21} + t_{0.025}S\sqrt{2}. \]

Transform the endpoints of the interval (which is an interval for both \(\mu_1 - \mu_2\) and \(\eta_1 - \eta_2\)). The interval for \(e^{\bar{m}_1 - \bar{m}_2} = \frac{e^{\bar{m}_1}}{e^{\bar{m}_2}}\) is:

\((e^L, e^U)\).

Under these assumptions, we recognize the \(W_{it}\) as lognormals – with means \(\mu_i^W = e^{\mu_i + \frac{1}{2} \sigma^2}\). Then

\[ \frac{e^{\bar{m}_1}}{e^{\bar{m}_2}} = \frac{e^{\mu_1 + \frac{1}{2} \sigma^2}}{e^{\mu_2 + \frac{1}{2} \sigma^2}} = \frac{\mu_1^W}{\mu_2^W}. \]

(f) [5 pts] There are many possible designs. The simplest is a complete block design which dispenses with condition 0.

Week 1: Treatments 1 – 10
Week 2: Treatments 1 – 10

ANOVA table:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week</td>
<td>1</td>
</tr>
<tr>
<td>Tmts</td>
<td>9</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
</tr>
<tr>
<td>AB</td>
<td>4</td>
</tr>
<tr>
<td>Error</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
</tr>
</tbody>
</table>
2. 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 the weights can be modeled by the linear model

\[
Y_j = \sum_{i=1}^{N} w_i x_{ij} + \epsilon_j, \quad 1 \leq i \leq N, \quad 1 \leq j \leq b.
\]

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

\[
x_{ij} = \begin{cases} 
1, & \text{if object } i \text{ is used in weighing } j \\
0, & \text{otherwise}
\end{cases}
\]

The \( \epsilon_j \) are assumed to be independent and identically distributed normal random variables with mean 0 and unknown variance \( \sigma^2 \).

(a) (5 points) Assuming \( b = N \), find the least-squares estimates of the \( w_j \) and the variances of these estimates if only object \( j \) is weighed in weighing \( j \).

(b) (10 points) Assuming \( b = N \), find the least-squares estimates of the \( w_j \) and the variances of these estimates if all objects except object \( j \) are weighed in weighing \( j \).

(c) (3 points) Of the methods described in (a) and (b), which would you recommend? Why?

(d) (7 points) A sheikh summoned ten goldsmiths and asked each of them to mint 1000 gold coins of 64 grams each. The sheikh told them he understood that it would be impossible to mint coins with exactly the same weight, and that they were allowed to have a small amount of variability in the weights. To confirm that the vendors have followed his order, he plans to sample only a few coins from each goldsmith, with a total sample size of at most 75, and weigh them together in just one weighing, to check that specifications (only small variation from the target value of 64 grams) are met. Any goldsmith failing to meet specifications would be severely punished.

Assume that the weights of the coins are independently normally distributed with “small variability” of \( \sigma = 0.01 \) gram. All goldsmiths, except one, were honest and produced coins with correct specifications. The greedy goldsmith thought that, because the sheikh would only check the weights of coins from all the vendors together, it would be impossible to identify anyone who failed to meet specifications. Thus, he would produce coins weighing 63 grams each while controlling the variability as per specifications, to make a handsome profit and with no chance of getting caught for cheating.
Through an informant, the sheikh learned that one of the goldsmiths planned to produce coins weighing only 63 grams, but otherwise meeting specifications. Alas, the informant did not know the identity of the greedy goldsmith. But an advisor to the sheikh told him to proceed as he announced and proposed a special weighing design to identify the greedy goldsmith. By proceeding as planned they would be able to identify the greedy goldsmith without his knowing it.

On the day of the weighing, each vendor was asked to put his products on a table labeled with his name. The sheikh selected a random sample of $n_i$ coins from the $i$-th vendor. As announced, he used one weighing for the $n = \sum_{i=1}^{10} n_i$ sampled coins and observed their total weight $Y$. The weight was not close to the expected weight of $64n$ grams, but he was almost sure of the identity of the cheater. Express the model for the observation $Y$, given that the $i$-th vendor was the dishonest one, $\{i = 1, 2, \ldots, 10\}$. Find an experimental design that could be used by the sheikh to determine the guilty party with a very high probability.
Problem 2.

In matrix form our model is \((b = N)\):

\[
\begin{pmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_b \\
\end{pmatrix}
= \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1b} \\
X_{21} & X_{22} & \cdots & X_{2b} \\
\vdots & \vdots & \ddots & \vdots \\
X_{b1} & X_{b2} & \cdots & X_{bb} \\
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_b \\
\end{pmatrix}
+ \begin{pmatrix}
e_1 \\
e_2 \\
\vdots \\
e_b \\
\end{pmatrix}
\]

\(Y \rightarrow \vec{y}_{b \times 1}\)

\(X \rightarrow \vec{X}_{b \times b}\)

\(\beta \rightarrow \vec{\beta}_{b \times 1}\)

\(\epsilon \rightarrow \vec{\epsilon}_{b \times 1}\)

(a) In this case, \(X = \vec{I}_b\), hence \(X\) is of full rank and the L.S. ests are:

\[
\vec{\beta} = (X'X)^{-1}X'Y = (\vec{I}_b'\vec{I}_b)^{-1}\vec{I}_b'\vec{y} = \vec{\gamma}
\]

\[
\text{cov} \hat{\beta} = \sigma^2(X'X)^{-1} = \sigma^2(I_b'\vec{I}_b)^{-1} = \sigma^2 I_b
\]

\(\therefore \hat{\omega}_x = \text{L.S. est. of } \omega_x = \gamma_x \quad \text{var} \hat{\omega}_x = \sigma^2\)

(b) In this case \(X = \vec{J}_b - \vec{I}_b\).

We see

\[
X'X = (\vec{J}_b - \vec{I}_b)'(\vec{J}_b - \vec{I}_b) = b\vec{J}_b - 2\vec{J}_b + \vec{I}_b
\]

\[
= \vec{I}_b + (b-2)\vec{J}_b
\]
\[
(X'X)^{-1} = (I_b + (b-2)J_b)^{-1} = I_b - \frac{\frac{b-2}{1+ (b-2)b}}{\frac{b-2}{b^2 - 2b + 1}} J_b
\]
\[
= \frac{b-2}{(b-1)^2} J_b
\]

Thus the L.S. ests are
\[
\hat{\beta} = (X'X)^{-1} X'\hat{y} = (I_b - \frac{\frac{b-2}{(b-1)^2} J_b}{(b-1)^2}) (J_b - I_b) \hat{y}
\]
\[
= (-I_b + \frac{\frac{b-2}{(b-1)^2} J_b}{(b-1)^2}) \hat{y}
\]
\[
= (-I_b + \frac{(b-1)^2 - b(b-2) + (b-2)}{(b-1)^2} J_b) \hat{y}
\]
\[
= (-I_b + \frac{1}{b-1} J_b) \hat{y}
\]
\[
= \left( \frac{1}{b-1} y_0 - y_1, \ldots, \frac{1}{b-1} y_b - y_b \right)
\]
\[
\text{cov} \hat{\beta} = \sigma^2 (X'X)^{-1} = \sigma^2 (I_b - \frac{\frac{b-2}{(b-1)^2} J_b}{(b-1)^2})
\]
\[
\hat{\omega}_X = \text{L.S. est} \omega_x = \left( \frac{1}{b-1} \sum_{j=1}^{b} \frac{b}{J_j} \right) - \hat{\omega}_x
\]
\[
\text{var} \hat{\omega}_x = \text{diag. entries of cov} \hat{\beta} = \sigma^2 \left( 1 - \frac{\frac{b-2}{(b-1)^2}}{\sigma^2} \right)
\]

(c) Smaller variance is better, so I would recommend the method in (b).
(d) Inspired by the model in the statement of the problem, we might use

\[ y = \sum_{k=1}^{10} n_k w_k + \varepsilon \]

where \( y \) = weight of all the coins
\( n_k \) = number of coins from vendor \( k \)
\( w_k \) = mean weight of a coin from vendor \( k \)

\[ \varepsilon \sim N(0, \sigma^2) \]

\[ \sigma^2 = \sum_{k=1}^{10} \frac{n_k^2 (0.01)^2}{n_k} = (0.01)^2 \sum_{k=1}^{10} n_k \]

We know that if vendor 1 is the guilty vendor:
\[ w_1 = 63 \]
\[ w_k = 64 \quad \forall k \neq 1 \]

Thus

\[ y = 63 n_1 + \sum_{k=1, k \neq 1}^{10} 64 n_k + \varepsilon \]

\[ = 64 n - n_1 + \varepsilon \]

We can identify the guilty vendor with high probability if all the \( n_k \) are different, because the number of grams \( y \) is below 64n grams should be "close" (within \( \pm 2 \sigma^2 \)) with very high probability to the number of coins from vendor \( k = \) the guilty vendor.
In particular, an "estimate" of $\eta_a$ is

$$\hat{\eta}_a = 64\eta - y$$

$$\text{Var} \hat{\eta}_a = \text{Var} y = c^2 = (0.01)^2 \sum_{l=1}^{16} n_l^2$$

and we would like $\text{Var} \hat{\eta}_a$ to be small.

"A good design will make all the $\eta_a$ different while keeping $\sum n_a^2$ small."

Since $n$ must be $\leq 75$, a good choice is

$$n_l = 0, 1, 2, 3, \ldots, 9$$

$$\Rightarrow c^2 = (0.01)^2 (285)$$

$$\Rightarrow c = 0.01 \sqrt{285} \approx 0.17$$

$$\pm 2c = \pm 0.34$$
4. Let the random variable $\eta$ have density $\pi(\eta)$ supported on $(-\infty, \infty)$. Consider two real valued random variables $Y$ and $Z$. Conditional on $\eta$, $Y$ has density $f(y)$, which is free of $\eta$, and $Z$ has density $f(z|\eta)$ (all densities are with respect to Lebesgue measure). Suppose that $Y$ and $Z$ are independent conditional on $\eta$.

(a) (3 points) Prove that $Y$ and $Z$ are independent.

Conditional on $\theta$ and $\sigma^2$, let $X_1, \ldots, X_n$ be iid random variables with a $N(\theta, \sigma^2)$ distribution. Suppose that the following assumptions are made concerning $\theta$ and $\sigma^2$: $\theta$ follows a $N(\theta_0, \tau^2)$ distribution with both $\theta_0$ and $\tau^2$ known and $\sigma^2 = \sigma_0^2 > 0$.

(b) (6 points) Derive the (joint) marginal distribution of $(X_1, \ldots, X_n)$.

(c) (10 points) Find the distribution of the random variable

$$T(X_1, \ldots, X_n) = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 + \frac{n(\bar{X} - \theta_0)^2}{n\tau^2 + \sigma_0^2}$$

under the marginal distribution of $(X_1, \ldots, X_n)$.

(d) (6 points) Let $(x_1, \ldots, x_n)$ denote the observed values of $(X_1, \ldots, X_n)$. The Bayesian prior predictive p-value to validate the specified model is defined as

$$p = P(T(X_1, \ldots, X_n) \geq T(x_1, \ldots, x_n))$$

under the marginal distribution of $(X_1, \ldots, X_n)$. Compare this Bayesian prior predictive p-value to the classical p-values for testing $\sigma^2 = \sigma_0^2$ against $\sigma^2 \neq \sigma_0^2$ in this problem.
4. (a) 

\[ f(y, z) = \int_{-\infty}^{\infty} f(y)f(z \mid \eta)\pi(\eta)\,d\eta \]
\[ = f(y)\int_{-\infty}^{\infty} f(z \mid \eta)\pi(\eta)\,d\eta \]
\[ = f(y)f(z). \]

(b) \( \mathbf{X} = (X_1, \ldots, X_n)^T \) has a multivariate normal distribution.

\[
E(X_i) = E(E(X_i \mid \theta)) = E(\theta) = \theta_0,
\]
\[
Var(X_i) = E(Var(X_i \mid \theta)) + Var(E(X_i \mid \theta))
\]
\[ = E(\sigma_0^2) + Var(\theta) \]
\[ = \sigma_0^2 + \tau^2, \]
\[
E(X_i, X_j) = E(E(X_i, X_j \mid \theta))
\]
\[ = E(E(X_i \mid \theta)E(X_j \mid \theta)) \]
\[ = E(\theta^2) \]
\[ = \tau^2 + \theta_0^2, \]
\[
Cov(X_i, X_j) = E(X_i, X_j) - E(X_i)E(X_j)
\]
\[ = \tau^2 + \theta_0^2 - \theta_0^2 \]
\[ = \tau^2. \]

So,

\[ \mathbf{X} \sim MVN(\theta_0 \mathbf{1}, \sigma_0^2 \mathbf{I} + \tau^2 \mathbf{J}). \]

where, \( \mathbf{1} \) is a vector with all entries equal to one, \( \mathbf{I} \) is the identity matrix and \( \mathbf{J} \) is a matrix with all entries equal to 1.

(c) Given \( \theta \), well known results for independent normal samples imply that \( Y = (1/\sigma_0^2) \sum_{i=1}^{n} (X_i - \bar{X})^2 \) has a \( \chi^2_{n-1} \) distribution, free of \( \theta \), and \( Z = [n(\bar{X} - \theta_0)^2]/[n\tau^2 + \sigma_0^2] \) is independent of \( Y \) (but its distribution depends on \( \theta \)). From part (b) we have that \( Var(n\bar{X}) = n(\sigma_0^2 + \tau^2) + n(n - 1)\tau^2 \), so that the marginal distribution of \( \bar{X} \) is \( N(\theta_0, (n\sigma_0^2 + n^2\tau^2)/n) \) and the distribution of \( Z \) is \( \chi^2_1 \). By part (a), \( Y \) and \( Z \) are independent and \( T \sim \chi^2_n \).

(d) Let \( y \) be the observed value of \( Y = (1/\sigma_0^2) \sum_{i=1}^{n} (X_i - \bar{X})^2 \). The classical p-value is defined as \( P(Y \geq y) \) and is computed based on the \( \chi^2_{n-1} \) distribution. The Bayesian p-value includes, in addition, a measure of discrepancy between \( \bar{X} \) and \( \theta_0 \) and is computed based on a \( \chi^2_n \) distribution.
6. Consider the general linear model,

$$Y = X\beta + \epsilon,$$

where $Y$ is the $n \times 1$ vector of observations, $X$ is the $n \times p$ design matrix, $\beta$ is the $p \times 1$ vector of unknown parameters, and $\epsilon$ is an $n \times 1$ multivariate normal random vector with mean $0_n$ and covariance matrix $\sigma^2 I_n$. Here, $0_n$ denotes the $n \times 1$ vector all of whose entries are 0 and $I_n$ denotes the $n \times n$ identity matrix. Assume that $X$ is of full rank $p$. Let $\Phi$ be a real-valued nonnegative function on the space of $p \times p$ nonnegative definite matrices. A design matrix $X$ is said to be $\Phi$-optimal if it minimizes $\Phi(X'X)$ over the set of all possible design matrices.

Let $P_D$ be the set of all positive definite $p \times p$ matrices. Let $T : P_D \rightarrow P_D$ be a function. We say $\Phi$ is invariant to the function $T$ if for all $M \in P_D$, $\Phi(T(M)) = \Phi(M)$.

Suppose $T_1, \ldots, T_t$ is a collection of $t$ functions from $P_D \rightarrow P_D$. Suppose also that $\Phi$ is convex and invariant to $T_1, \ldots, T_t$.

(a) (7 points) Show that the $p \times p$ positive definite matrix $X'X = \frac{1}{t} \sum_{i=1}^{t} T_i(X'X)$ satisfies

$$\Phi(X'X) \leq \Phi(X'X)$$

with equality if $X'X = T_i(X'X)$ for all $i$.

(b) (8 points) Let

$$AVG = \{X'X \mid X'X = \frac{1}{t} \sum_{i=1}^{t} T_i(X'X) \text{ for some design matrix } X \text{ of rank } p\} \subset P_D.$$

Prove that if $X'X_{opt}$ minimizes $\Phi$ over $AVG$ and if there exists a design $X^*$ such that $X'X^* = X'X_{opt}$ then $X^*$ is $\Phi$-optimal.

(c) (10 points) For the special case with $p = 2$, $n$ even, $\Phi(X'X) = \det(X'X)^{-1}$ and the linear model

$$Y_j = \beta_0 + \beta_1 x_j + \epsilon_j, 1 \leq j \leq n,$$

where $-1 \leq x_1, \ldots, x_n \leq 1$, show that the design that takes half its observations at $x_i = -1$ and half at $x_i = 1$ is $\Phi$-optimal. You can take it as known that $\Phi$ is convex and that it is invariant to the two transformations $T_1(M) = M$ and

$$T_2(M) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} M \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$
Problem 6

(a) \[ \Phi (x'x) = \Phi \left( \frac{1}{\epsilon} \sum_{k=1}^{t} T_k (x'x) \right) \]
\[ \leq \frac{1}{\epsilon} \sum_{k=1}^{t} \Phi (T_k (x'x)) \]

\[ \Phi (x'x) \quad \text{convexity} \]
\[ \geq \frac{1}{\epsilon} \sum_{k=1}^{t} \Phi (x'x) = \Phi (x'x) \]

Also, \[ \Phi (x'x) = \Phi \left( \frac{1}{\epsilon} \sum_{k=1}^{t} T_k (x'x) \right) = \Phi \left( \frac{1}{\epsilon} \frac{t}{2} x'x \right) = \Phi (x'x) \]

(b) Let \( x^* \) be such that \( x^* x^* = x'x \).
Then for all design matrices \( x \)
\[ \Phi (x'x) \geq \Phi (x'x) \geq \Phi (x'x^* + x'x^* \Phi (x'x) \]
by (a)

So \( x^* \) is \( \Phi \)-optimal.

(c) For the linear model given
\[ x'x = \left( \begin{array}{ccc}
\frac{1}{\epsilon} & \frac{2}{\epsilon^2} & \frac{3}{\epsilon^3} \\
\frac{2}{\epsilon^2} & \frac{4}{\epsilon^4} & \frac{6}{\epsilon^6} \\
\frac{3}{\epsilon^3} & \frac{6}{\epsilon^6} & \frac{9}{\epsilon^9}
\end{array} \right) \]

\[ \Rightarrow x'x = \frac{1}{2} T_1 (x'x) + \frac{1}{2} T_2 (x'x) \]
\[ = \frac{1}{2} \left( \begin{array}{ccc}
\frac{1}{\epsilon} & \frac{2}{\epsilon^2} & \frac{3}{\epsilon^3} \\
\frac{2}{\epsilon^2} & \frac{4}{\epsilon^4} & \frac{6}{\epsilon^6} \\
\frac{3}{\epsilon^3} & \frac{6}{\epsilon^6} & \frac{9}{\epsilon^9}
\end{array} \right) + \frac{1}{2} \left( \begin{array}{ccc}
\frac{1}{\epsilon} & \frac{2}{\epsilon^2} & \frac{3}{\epsilon^3} \\
\frac{2}{\epsilon^2} & \frac{4}{\epsilon^4} & \frac{6}{\epsilon^6} \\
\frac{3}{\epsilon^3} & \frac{6}{\epsilon^6} & \frac{9}{\epsilon^9}
\end{array} \right) \]
\[ = \left( \begin{array}{ccc}
\frac{1}{\epsilon} & 0 & 0 \\
0 & \frac{1}{\epsilon^2} & \frac{1}{\epsilon^3} \\
0 & \frac{1}{\epsilon^3} & \frac{1}{\epsilon^6}
\end{array} \right) \]
\[
\text{Now}
\begin{align*}
\det (X'X^{-1}) &= \left( \frac{\sum X_i^2}{n} \right)^{-1} \\
\text{For } -1 \leq X_i \leq +1 \text{ it is easy to see that}
\sum_i X_i^2 \text{ has max value } = n, \text{ so the min possible value for } \det (X'X)^{-1} = \frac{1}{n^2}.
\end{align*}
\]
Thus,
\[
X'X = \begin{pmatrix} n & 0 & 0 \\ 0 & n & 0 \\ 0 & 0 & n \end{pmatrix}
\]
minimizes \( \bar{\beta} \) ou AVG.

It is easy to see that the design matrix \( X^* \) corresponding to taking half the observations at +1 and half at -1 satisfies
\[
X^*X^* = \begin{pmatrix} n & 0 & 0 \\ 0 & n & 0 \\ 0 & 0 & n \end{pmatrix} = X'X_{opt}
\]
so \( X^* \) is \( \bar{\beta} \)-optimal.
3. Consider an unknown, vector-valued parameter $\theta = (\theta_1, \ldots, \theta_p)'$ with parameter space $\Theta = \mathcal{R}^p$. Our problem is to estimate $\theta$ with action space $\mathcal{A} = \Theta$. Let $\pi(\theta)$ denote our probability density function for $\theta$. (Typically, $\pi$ is a posterior density for $\theta$ based on some observations.) Let $\mu$ and $\Sigma$ denote the mean and covariance matrix of $\theta$ under $\pi$.

(a) (7 points) Consider the loss function

$$L_1(\theta, a) = |\theta - a|^2 = (\theta - a)'(\theta - a) = \sum_{i=1}^{p}(\theta_i - a_i)^2.$$ 

Find the Bayes action $a^{(1)}$ and its expected loss $\rho_1$.

(b) (7 points) Let $B$ be a specified $p \times p$ matrix. Find the Bayes action $a^{(B)}$ and its expected loss $\rho_B$ under the loss function

$$L_B(\theta, a) = |B \theta - a|^2.$$ 

(c) (7 points) Let $c_1$ and $c_B$ be specified positive constants. Find the Bayes action $a^{(1B)}$ and its expected loss $\rho_{1B}$ under the loss function

$$L_{1B}(\theta, a) = c_1 L_1(\theta, a) + c_B L_B(\theta, a).$$ 

(d) (4 points) Write $a^{(1B)}$ as a convex combination of $a^{(1)}$ and $a^{(B)}$. Also, suppose that $c_1 + c_B = 1$, implying that $L_{1B}$ is a convex combination of $L_1$ and $L_B$. Is $\rho_{1B}$ a convex combination of $\rho_1$ and $\rho_B$? Interpret the terms in $\rho_{1B}$. 
(a) Let $\rho(a) = E^\pi(|\theta - a|^2) = E^\pi(|\theta|^2) - 2a'E^\pi(\theta) + |a|^2$. This object is finite (since $\theta$ has a covariance) and is an upward opening "parabola" as a function of $a$. Hence, the minimizer is the solution to

$$\nabla[-2a'E^\pi(\theta) + |a|^2] = 0$$

which yields $a = E^\pi(\theta)$. Hence, $a^{(1)} = \mu$ and the optimized value $\rho_1 = E^\pi(|\theta - \mu|^2) = \text{tr}(\Sigma)$, the trace of $\Sigma$.

Notes: If you differentiate under the expectation, you should say something about the interchange of order. Also, if not adept at multivariate calculus, you can easily obtain the result based on the coordinate-based form $\rho(a) = E^\pi(\sum_{i=1}^p (\theta_i - a_i)^2)$.

(b) Similar to (a), the Bayes action is a solution to

$$\nabla[-2a'B E^\pi(\theta) + |a|^2] = 0$$

which yields $a = BE^\pi(\theta)$. Hence, $a^{(B)} = B\mu$ and the optimized value $\rho_B = E^\pi(|B\theta - B\mu|^2) = \text{tr}(B\Sigma B')$. This follows by noting that under $\pi$, $Y = B\theta$ has mean $B\mu$ and covariance matrix $B\Sigma B'$.

(c) Similar to (a) and (b), the Bayes action is a solution to

$$\nabla[-2c_1 a'E^\pi(\theta) - 2c_B a'B E^\pi(\theta) + (c_1 + c_B)|a|^2] = 0$$

which yields

$$(c_1 + c_B)a = c_1 E^\pi(\theta) + c_B B E^\pi(\theta)$$

Hence,

$$a^{(1B)} = \alpha \mu + (1 - \alpha) B\mu$$

$$= \alpha a^{(1)} + (1 - \alpha) a^{(B)}$$

where

$$\alpha = \frac{c_1}{c_1 + c_B}.$$}

This is a convex combination.
Next
\[
\rho_{1B} = c_1 E^y [\theta - (\alpha \mu + (1 - \alpha)B\mu)]^2 + c_B E^y |B\theta - (\alpha \mu + (1 - \alpha)B\mu)|^2 \\
= c_1 E^y [\theta - \mu - ((\alpha - 1)\mu + (1 - \alpha)B\mu)]^2 \\
+ c_B E^y |B\theta - B\mu - (\alpha \mu + (-\alpha)B\mu)|^2 \\
= c_1 [\text{tr}(\Sigma) + (1 - \alpha)^2 |\mu - B\mu|^2] + c_B [\text{tr}(B\Sigma B') + \alpha^2 |\mu - B\mu|^2] \\
= c_1 \text{tr}(\Sigma) + c_B \text{tr}(B\Sigma B') + [1 - \alpha^2 + \alpha^2] |\mu - B\mu|^2 \\
= c_1 \rho_1 + c_B \rho_B + [c_1 (1 - \alpha)^2 + c_B \alpha^2] |\mu - B\mu|^2
\]

(d) Note that \(\alpha = c_1\). We see that \(\rho_{1B}\) is equal to a convex combination of \(\rho_1\) and \(\rho_B\) plus an additional positive quantity. This quantity arises as “Bayesian biases”. That is, \(a^{(1)}\) is the mean of \(\theta\) and \(a^{(B)}\) is the mean of \(B\theta\), but \(a^{(1B)}\) is biased for both quantities. Also, note that the formula for \(\rho_{1B}\) can also be obtained from the conditional variance formula.
6. In the following, \( I_m \) is an \( m \times m \) identity matrix, \( 0_m \) is an \( m \times 1 \) vector of zero elements, and \( J_m = 1_m \cdot 1_m' \), where \( 1_m \) is an \( m \times 1 \) vector of 1’s. You may use, without proof, the fact that

\[
[I_m + \phi J_m]^{-1} = \left[ I_m - \frac{\phi}{1 + m\phi} J_m \right].
\]

(a) (3 points) Consider the general linear model

\[
Y = W\theta + \epsilon, \quad \epsilon \sim N(0_n, V\sigma^2), \tag{1}
\]

where \( \theta \) is a vector of \( p \) unknown parameters, \( \sigma^2 \) is a constant, and \( W \) and \( V \) are known matrices of sizes \( n \times p \) and \( n \times n \), respectively. Give a formula for a joint confidence region for a set of contrasts \( C'\theta \) using Scheffé’s method. Here, \( C' \) is \( q \times p \) of rank \( q \) (and \( q \geq 1 \)).

(b) (6 points) Now consider the following linear model:

\[
Y_{ijt} = \gamma_i + \tau_j + \epsilon_{ijt}, \tag{2}
\]

\[
\epsilon_{ijt} \sim N(0, \sigma^2_E), \quad \gamma_i \sim N(0, \sigma^2_\gamma),
\]

\[i = 1, 2; \quad j = 1, 2; \quad t = 1, 2;\]

where all random variables on the right hand side of the model are mutually independent. Write the model as \( Y = Z\gamma + X\tau + \epsilon \), where

\[
Y = [Y_{111}, Y_{112}, Y_{121}, Y_{122}, Y_{211}, Y_{212}, Y_{221}, Y_{222}]',
\]

\[\gamma = [\gamma_1, \gamma_2] \quad \text{and} \quad \tau = [\tau_1, \tau_2]\]

and find \( Z, X \). Show that the variance-covariance matrix of \( Z\gamma + \epsilon \) is of the form

\[
V\sigma^2_E = \begin{bmatrix}
aI_m + bJ_m & 0_m \\
0_m & aI_m + bJ_m
\end{bmatrix}.
\]

(c) (4 points) State the distribution of \( Y \) and find the best linear unbiased estimator of \( \tau \). Give a condition for \( C'\tau \) to be estimable under model (2), where \( C' \) is defined in part (a). Justify your answer.

(d) (6 points) For given constant vector \( d \) and estimable set of functions \( C'\tau \), state a test statistic for testing

\[
H_0 : C'\tau = d \quad \text{versus} \quad H_1 : C'\tau \neq d,
\]

where \( C' \) is defined in part (a). Find the expected value of the numerator of the test statistic.

(e) (6 points) Let \( \phi = \sigma^2_\gamma / \sigma^2_E \) and let \( C' = [1 \quad -1] \). Assuming that the distribution of your test statistic in part (d) is non-central \( F \), does the power of the test depend on the value of \( \sigma^2_\gamma \)? If so, in which way?
Solution to Question 6

Q II Spring 2011

a) A 100(1-\alpha)\% joint confidence region for \( \theta \) is given by

\[
\begin{bmatrix} C'\widehat{\theta} - C\theta \end{bmatrix} \left( C'(W'W)^{-1}W' \right) C' \theta - C\theta \right) \leq \frac{\hat{\sigma}^2}{\hat{\sigma}^2} F_{n-r-r, \alpha}
\]

where \( \hat{\sigma}^2 \) is an unbiased estimator of \( \sigma^2 \) and \( r = \text{rank}(X) \),
and \( \hat{\theta} = (W'W)^{-1}W'y \)

b) \( y = Z\theta + X\varepsilon + e \)

\[
\begin{array}{ccc|ccc}
Y_{111} & Z = & 1 & 0 & X = & 1 & 0 \\
Y_{112} & & 1 & 0 & & 1 & 0 \\
Y_{121} & & 1 & 0 & & 0 & 1 \\
Y_{122} & & 1 & 0 & & 0 & 1 \\
Y_{211} & & 0 & 1 & & 1 & 0 \\
Y_{212} & & 0 & 1 & & 1 & 0 \\
Y_{221} & & 0 & 1 & & 0 & 1 \\
Y_{222} & & 0 & 1 & & 0 & 1 \\
\end{array}
\]

\[
\text{Var}(Z\theta + \varepsilon) = \text{Var}(Z\theta + \varepsilon) = \text{Var}(Z\theta) + \text{Var}(\varepsilon) \text{ due to uncorrelatedness}
\]

\[
= Z I\sigma^2_{\theta} Z' + I\sigma^2_{\varepsilon}
\]

\[
= \begin{bmatrix} \sigma^2_{\theta} + I\sigma^2_{\varepsilon} = I\sigma^2_{\varepsilon} + I\sigma^2_{\varepsilon} \\
I\sigma^2_{\varepsilon} + I\sigma^2_{\varepsilon} \\
\end{bmatrix}
\]

which is the required form with \( a = 1 \), \( b = \frac{\sigma^2_{\theta}}{\sigma^2_{\varepsilon}} \).
c) \( Y \sim N(X \beta, \sigma^2 \mathbf{I}_e) \)

so the BLUE of \( \beta \) is the generalized L.S. estimator.

\[
\hat{\beta} = (X' \hat{V}^{-1} X)^{-1} X' \hat{V}^{-1} y
\]

(OLS estimator)

\[
E[\hat{\beta}] = \beta
\]

\[
= \beta
\]

\[
= \beta
\]

\[
= \beta
\]

\[
\Rightarrow \quad C'(X' \hat{V}^{-1} X)^{-1} X' \hat{V}^{-1} X = C' \quad \text{for estimability to hold}
\]

d) \( H_0: \quad \beta = \beta_0 \quad \text{vs.} \quad H_a: \quad \beta \neq \beta_0 \)

\[
\hat{\beta} \sim F_{k, m-k, \sigma^2}
\]

So \( \hat{\beta} \) is a test statistic for testing \( H_0 \).

\[
E \left[ \left( \hat{\beta} - \beta_0 \right)' A^{-1} \left( \hat{\beta} - \beta_0 \right) \right] = \sigma^2 \text{trace} \left( A^{-1} \right)
\]

where \( A = (C' (X' \hat{V}^{-1} X)^{-1} C)^{-1} \)

\[
\text{Var}(\hat{\beta}) = \text{Var}(\hat{\beta}) = C' (X' \hat{V}^{-1} X)^{-1} X' \hat{V}^{-1} \Sigma \hat{V}^{-1} X (X' \hat{V}^{-1} X)^{-1} C = \sigma^2 \text{ trace} \left( A^{-1} \right)
\]

\[
\quad = C' (X' \hat{V}^{-1} X)^{-1} C \sigma^2 \quad \text{since } \beta \text{ is estimable}
\]

\[
E(\hat{\beta}' A \hat{\beta}) = \frac{1}{2} \text{trace} \left( A^{-1} \right) + \frac{1}{2} (C' \hat{\beta} - \beta_0)' (C' (X' \hat{V}^{-1} X)^{-1} C)' (C' \hat{\beta} - \beta_0)
\]

\[
= \sigma^2 + \frac{1}{2} (C' \hat{\beta} - \beta_0)' (C' (X' \hat{V}^{-1} X)^{-1} C)' (C' \hat{\beta} - \beta_0)
\]
2) Power is an increasing function of the non-centrality parameter. The non-centrality parameter is
\[ \frac{1}{2} \text{exp} \left( C' C - d \right) \left( C' C - C \right) \frac{1}{2} \text{exp} \left( C' C - d \right) \]

\[ \frac{2 \phi}{\gamma} \]

\[ V = V \text{ is a function of } \sigma_0^2 \text{ through } \phi. \text{ So the question is whether } C' C' \text{ depends on } \sigma_0^2. \]

\[ V^{-1} = \left( I_4 + J_4 \phi \right) = \left( I_4 + J_4 \phi \right)^{-1} \]

\[ \left( I_4 + J_4 \phi \right)^{-1} \]

\[ V^{-1} X = X - \frac{\phi}{1+4\phi} J_8 x_2 \]

\[ X' V^{-1} X = X' X - \frac{2 \phi}{1+4\phi} J_{8x2} \]

\[ = 4 I_2 - \frac{8 \phi}{1+4\phi} J_2 = 4 \left[ I_2 + \left( \frac{-2\phi}{1+4\phi} \right) J_2 \right] \]

\[ \text{This is full rank. So using the given formula,} \]

\[ (X' V^{-1} X)^{-1} = \left[ I_2 - \frac{-2\phi}{1+4\phi} \right] \frac{1}{1+2 \phi \frac{2\phi}{1+4\phi}} \left[ I_2 + \frac{2\phi}{1+4\phi} J_2 \right] \]

Then with \[ c = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \]

\[ c' (X' V^{-1} X)^{-1} c = \frac{3}{4} \text{ which does not depend on } \phi \text{.} \]
5. Suppose that the following linear regression model is postulated

\[ Y = X_1 \beta_1 + \epsilon, \]

where \( Y \) is an \( n \times 1 \) vector of random variables, \( X_1 \) is an \( n \times p \) full rank matrix of known constants, \( \beta_1 \) is a \( p \times 1 \) vector of unknown parameters, and \( \epsilon \sim N(0, \sigma^2 I_n) \) with \( \sigma^2 \) unknown. Let \( b_1 \) be the least squares estimator of \( \beta_1 \) under the postulated model, and let \( \hat{Y} = X_1 b_1 \).

Suppose that the true model is actually

\[ Y = X_1 \beta_1 + X_2 \beta_2 + \epsilon_2, \]

where \( X_2 \) is an \( n \times q \) matrix of known constants, \( \beta_2 \neq 0 \) is a \( q \times 1 \) vector of unknown parameters, and \( \epsilon_2 \) has the same distribution as \( \epsilon \).

(a) [4 pts] Show that \( b_1 \) is generally a biased estimator of \( \beta_1 \). State any conditions under which \( b_1 \) is unbiased.

(b) [3 pts] Find the covariance matrix of \( b_1 \).

(c) [5 pts] Consider the decomposition of the total sum of squares of \( Y \) into regression sum of squares (SSR) and residual sum of squares (SSE) for the postulated model. In other words, \( \text{SSR} = \hat{Y}'\hat{Y} \) and \( \text{SSE} = (Y - \hat{Y})'(Y - \hat{Y}) \).

Find the expected value of SSE. Simplify the expression as much as possible. Is \( \text{MSE} = \text{SSE}/(n-p) \), the usual estimator of \( \sigma^2 \) unbiased?

(d) To test the hypothesis \( H_0 : \beta_1 = 0 \), suppose that we use the usual \( F \)-test based on the test statistic \( F = \frac{\text{SSR}/p}{\text{SSE}/(n-p)} \), which assumes incorrectly that the postulated model is true.

   i. [6 pts] What are the actual distributions of SSR and SSE under \( H_0 \)?

   ii. [2 pts] Comment on the validity of the \( F \)-test.

(e) [5 pts] Consider the least squares estimator \( b_1^* \) of \( \beta_1 \) under the true model assuming that \( X = [X_1 \ X_2] \) is of full rank. When \( \beta_2 = 0 \), compare \( b_1 \) and \( b_1^* \) in terms of bias and variance.

**Hint:** The following matrix identity may be useful:

\[
\begin{bmatrix}
A & B \\
B' & C
\end{bmatrix}^{-1} = \begin{bmatrix}
A^{-1} & 0 \\
0 & 0
\end{bmatrix} + \begin{bmatrix}
-A^{-1}B \\
0
\end{bmatrix} (C - B'A^{-1}B)^{-1} \begin{bmatrix}
-A'B^{-1} \\
-I
\end{bmatrix}.
\]
Part II

5. (a) From \( b_1 = (X'_1X_1)^{-1}X'_1Y \), we have

\[
E(b_1) = (X'_1X_1)^{-1}X'_1E(Y) = (X'_1X_1)^{-1}X'_1(X_1\beta_1 + X_2\beta_2) = \beta_1 + (X'_1X_1)^{-1}X'_1X_2\beta_2.
\]

Unless \((X'_1X_1)^{-1}X'_1X_2\beta_2 = 0\), \(b_1\) is biased. If \(X'_1X_2 = 0\), that is, the columns of \(X_1\) and \(X_2\) are orthogonal, then \(b_1\) is unbiased.

(b) \n\[
\text{var}(b_1) = \text{var}((X'_1X_1)^{-1}X'_1Y) = (X'_1X_1)^{-1}X'_1(\sigma^2I_n)X_1(X'_1X_1)^{-1} = (X'_1X_1)^{-1}\sigma^2.
\]

(c) First note that \(\text{SSE} = (Y - \hat{Y})'(Y - \hat{Y}) = Y'(I_n - X_1(X'_1X_1)^{-1}X'_1)Y\). Letting \(P_1 \equiv X_1(X'_1X_1)^{-1}X'_1\) (the projection onto the column space of \(X_1\)), we have

\[
E(\text{SSE}) = E(Y'(I_n - P_1)Y) = tr([(I_n - P_1)E(YY')]
\]

\[
= tr([(I_n - P_1)(\sigma^2I_n + (X_1\beta_1 + X_2\beta_2)(X_1\beta_1 + X_2\beta_2)')]
\]

\[
= \sigma^2 tr[(I_n - P_1)] + (X_1\beta_1 + X_2\beta_2)'(I_n - P_1)(X_1\beta_1 + X_2\beta_2)
\]

\[
= (n - p)\sigma^2 + \beta'_2X'_2(I_n - P_1)X_2\beta_2.
\]

As \(E(\text{MSE}) = \sigma^2 + \beta'_2X'_2(I_n - P_1)X_2\beta_2/(n - p) \neq \sigma^2\) in general, MSE is not unbiased in estimating \(\sigma^2\).

(d) i. Under \(H_0\), \(Y \sim N(X_2\beta_2, \sigma^2I_n)\). Note that \(\text{SSR} = Y'P_1Y\). Since \(P_1\) is idempotent and its rank is \(p\), \(\text{SSR}/\sigma^2\) follows chi-square distribution with \(df = p\) and non-centrality parameter \(\frac{1}{2\sigma^2}\beta'_2X'_2P_1X_2\beta_2\). Similarly, we can see that \(\text{SSE}/\sigma^2\) follows chi-square distribution with \(df = n - p\) and non-centrality parameter \(\frac{1}{2\sigma^2}\beta'_2X'_2(I_n - P_1)X_2\beta_2\).

In addition, \(P_1(I_n - P_1) = 0\) implies that \(\text{SSR}\) and \(\text{SSE}\) are independent.

ii. With the non-zero non-centrality parameters in general, the \(F\)-test statistic under \(H_0\) no longer follows the \(F\) distribution with \(df_1 = p\) and \(df_2 = n - p\) (the usual null distribution), and thus the \(F\)-test is not valid.

(e) As the LSE of \(\beta_1\) under the true model, \(b_1^*\) is unbiased. \(b_1\) is also unbiased when \(\beta_2 = 0\). The covariance matrix of \(b_1^*\) corresponds to the leading \(p \times p\) submatrix of \((X'X)^{-1}\sigma^2\), where

\[
X'X = \begin{bmatrix}
X'_1X_1 & X'_1X_2 \\
X'_2X_1 & X'_2X_2
\end{bmatrix}.
\]
Using the hint, we obtain the leading $p \times p$ submatrix of $(X'X)^{-1}$ as

$$(X'_1X_1)^{-1} + (X'_1X_1)^{-1}X'_1X_2(X'_2X_2 - X'_2X_1(X'_1X_1)X'_1X_2)^{-1}X'_1X_2(X'_1X_1)^{-1}.$$ 

Since $X'_2X_2 - X'_2X_1(X'_1X_1)X'_1X_2 = X'_2(I_n - P_1)X_2$ and it is positive definite, its inverse is also positive definite. From this, we can verify that $\text{var}(b^*_1) - \text{var}(b_1)$ is non-negative definite, and thus $b_1$ has smaller variance than $b^*_1$ in general.

Note also that by the Gauss-Markov theorem, when $\beta_2 = 0$, $b_1$ is the BLUE of $\beta_1$. So, for all $\ell \in R^p$, $\text{var}(\ell'b^*_1) \geq \text{var}(\ell'b_1)$. 

10
4. 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,

\[ y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \varepsilon_i; \quad i = 1, \ldots, n, \]

where \( \varepsilon_1, \ldots, \varepsilon_n \) are independent and identically distributed observations from a normal distribution with mean 0 and variance \( \sigma^2 \). Suppose that the true regression coefficient vector \( \beta \equiv (\beta_1, \beta_2, \beta_3)' \) is \( (\beta_1', \beta_2', 0)' \) with \( \beta_1' \neq 0 \) and \( \beta_2' \neq 0 \).

Consider the following three candidate models:

Model 1: \( y_i = x_{i1}\beta_1 + \varepsilon_i \)
Model 2: \( y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i \)
Model 3: \( y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \varepsilon_i \)

Let \( \hat{\beta}_k \) be the least squares estimator of \( \beta \) 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 \( x = x_i \equiv (x_{i1}, x_{i2}, x_{i3})' \) is available and denoted by \( z_i, \ i = 1, \ldots, n \). Then the average squared prediction error of the fitted Model \( k \) is defined as

\[ PE_k = \frac{1}{n} \sum_{i=1}^{n} (z_i - \hat{x}_i'\hat{\beta}_k)^2. \]

(a) (3 points) Show that the expected average squared prediction error conditional on the data \( \{y_i, x_i\}, i = 1, \ldots, n \), is

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

(b) Consider the overall unconditional expected prediction error of Model \( k \) (yet conditional on \( x_i, i = 1, \ldots, n \)) that is defined as

\[ \Gamma_{k,n} = E_{Y_1, \ldots, Y_n}(E_{Z_1, \ldots, Z_n}(PE_k)). \]

i. (5 points) Verify that

\[ \Gamma_{k,n} = \sigma^2 + k\sigma^2/n + \beta'X'(I - P_k)X\beta/n \]

where \( X \) is the full design matrix, and \( P_k \equiv X_k(X_k'X_k)^{-1}X_k' \) with \( X_k \) being the \( n \) by \( k \) design matrix for Model \( k \).

ii. (5 points) Assume that the first two covariates are sufficiently different in the sense that \( \|(I - P_1)X_2\|_2^2 = O(n) \), where \( X_2 \) is the second column of \( X \). For large \( n \), which model has the smallest unconditional expected prediction error?
(c) In practice, the future values $z_i$ are not available, and we need to estimate the expected prediction error $\Gamma_{k,n}$ using the data. For example, the leave-one-out cross validation method estimates $\Gamma_{k,n}$ by holding out each data point $(y_i, x_i)$ and predicting $y_i$ based on the model fitted to the remaining $(n-1)$ observations. The leave-one-out cross validation error of Model $k$ is defined as

$$CV_{k,n} \equiv \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i' \hat{\beta}_k^{[i]})^2,$$

where $\hat{\beta}_k^{[i]}$ is the least squares estimator of $\beta$ under Model $k$ with the $i$th observation deleted. Under some assumptions on $X$, it can be shown (but you don’t need to show) that

$$CV_{1,n} = \Gamma_{1,n} + o_p(1),$$

and

$$CV_{k,n} = \epsilon' \epsilon / n + 2k\sigma^2 / n - \epsilon' P_k \epsilon / n + o_p(1/n)$$

for $k = 2, 3$.

i. (5 points) Is the leave-one-out cross validation error consistent for the unconditional expected prediction error $\Gamma_{k,n}$ when $k = 2, 3$ in the sense that $CV_{k,n} - \Gamma_{k,n}$ converges to zero in probability as $n$ increases?

ii. (7 points) If we choose the model with the smallest cross validation error for each $n$, what is the limiting probability that the correct model (Model 2) is chosen by leave-one-out cross validation?
5. Solution for question 4.

(a)

\[ PE_k = \frac{1}{n} \sum_{i=1}^{n} (z_i - x_i' \hat{\beta}_k)^2 \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \{ (x_i' \beta + \epsilon_i) - x_i' \hat{\beta}_k \}^2 \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \{ \epsilon_i^2 + (x_i' \beta - x_i' \hat{\beta}_k)^2 + 2 \epsilon_i (x_i' \beta - x_i' \hat{\beta}_k) \}, \]

where \( \epsilon_i, i = 1, \ldots, n, \) are iid with \( N(0, \sigma^2) \) and independent of the data. Thus

\[ E_{z_1, \ldots, z_n}(PE_k) = \frac{1}{n} \sum_{i=1}^{n} \{ \sigma^2 + (x_i' \beta - x_i' \hat{\beta}_k)^2 \} = \sigma^2 + \frac{1}{n} \sum_{i=1}^{n} (x_i' \beta - x_i' \hat{\beta}_k)^2. \]

(b) i.

\[ \sum_{i=1}^{n} (x_i' \beta - x_i' \hat{\beta}_k)^2 \]

\[ = (X \beta - X \hat{\beta}_k)' (X \beta - X \hat{\beta}_k) \]

\[ = (X \beta - X_k(X_k'X_k)^{-1}X_k'Y)' (X \beta - X_k(X_k'X_k)^{-1}X_k'Y) \]

\[ = (X \beta - P_k(X \beta + \epsilon))'(X \beta - P_k(X \beta + \epsilon)) \]

\[ = ((I - P_k)X \beta - P_k \epsilon)'((I - P_k)X \beta - P_k \epsilon) \]

\[ = \beta'X'(I - P_k)X \beta + \epsilon'P_k \epsilon. \]

Hence,

\[ E_{Y_1, \ldots, Y_n} \sum_{i=1}^{n} (x_i' \beta - x_i' \hat{\beta}_k)^2 = \beta'X'(I - P_k)X \beta + E\epsilon'P_k \epsilon \]

\[ = \beta'X'(I - P_k)X \beta + k\sigma^2, \]

which yields the expression for \( \Gamma_{k,n}. \)

ii. Since the true model depends on the first two covariates only, \( (I - P_k)X \beta = 0 \) for \( k = 2, 3. \) Therefore \( \Gamma_{k,n} = \sigma^2 + k\sigma^2/n \) for \( k = 2, 3 \) while \( \Gamma_{1,n} = \sigma^2 + \sigma^2/n + \beta'X'(I - P_1)X \beta/n. \) From the assumption, note that

\[ \beta'X'(I - P_1)X \beta/n = \|(I - P_1)X[2] \beta_2^*\|^2/n = O(1), \]

and as a result, \( \Gamma_{1,n} \) is larger than \( \Gamma_{2,n} \) for sufficiently large \( n. \) Model 2 has the smallest expected prediction error asymptotically.
(c) i. Under the assumption that $\varepsilon \sim N(0, \sigma^2 I)$, $\varepsilon' \varepsilon$ follows $\sigma^2 \chi^2(n)$ and hence $\varepsilon' \varepsilon / n$ converges to $\sigma^2$ in probability. Similarly, $\varepsilon' P_k \varepsilon / n - k\sigma^2 / n$ converges to zero in probability. Combining these results, we have

$$CV_{k,n} - \Gamma_{k,n} = \varepsilon' \varepsilon / n + 2k\sigma^2 / n - \varepsilon' P_k \varepsilon / n + o_p(1/n) - (\sigma^2 + k\sigma^2 / n)$$

$$= (\varepsilon' \varepsilon / n - \sigma^2) + (k\sigma^2 / n - \varepsilon' P_k \varepsilon / n) + o_p(1/n)$$

$$\rightarrow 0 \text{ in probability as } n \rightarrow \infty.$$  

ii. Let the index of the chosen model be $k^*_n \equiv \arg \min_{k=1, 2, 3} CV_{k,n}$. First, the answer to part ii in (b) implies that $\lim_{n \rightarrow \infty} P(k^*_n = 1) = 0$ as $CV_{1,n} = \sigma^2 + O(1) + o_p(1)$ and $CV_{k,n} = \sigma^2 + o_p(1)$ for $k = 2, 3$. So, $\lim_{n \rightarrow \infty} P(k^*_n = 2) = \lim_{n \rightarrow \infty} P(CV_{2,n} < CV_{3,n})$. From

$$P(CV_{2,n} < CV_{3,n})$$

$$= P(\varepsilon' P_3 \varepsilon / n - \varepsilon' P_2 \varepsilon / n - 2\sigma^2 / n + o_p(1/n) < 0)$$

$$= P(\varepsilon'(P_3 - P_2)\varepsilon - 2\sigma^2 + o_p(1) < 0),$$

and the fact that $\varepsilon'(P_3 - P_2)\varepsilon \sim \sigma^2 \chi(1)$, we have

$$\lim_{n \rightarrow \infty} P(k^*_n = 2) = P(\chi(1) < 2) \approx 0.8427 \text{ and } \lim_{n \rightarrow \infty} P(k^*_n = 3) = P(\chi(1) > 2).$$

There is a positive probability that Model 3 is chosen incorrectly by the leave-one-out cross validation method!
4 (a) [3 points] Show that a least squares estimator of $\theta$ in the general linear model

$$Y = W\theta + \epsilon, \quad \epsilon \sim (0, I\sigma^2)$$

is given by $\hat{\theta} = (W'W)^{-1}W'y$, where $(W'W)^{-1}$ is a generalized inverse of $W'W$. Here, $Y$ and $\epsilon$ are $n \times 1$ random vectors of response and error variables, respectively, $\theta$ is a $p \times 1$ parameter vector, $W$ is an $n \times p$ matrix and $y$ is a vector of observed values of $Y$.

(b) Now partition $W$ and $\theta$ as $W = [X \ Z]$ and $\theta = [\tau' \ \gamma']'$ to obtain the following general linear model:

$$Y = X\tau + Z\gamma + \epsilon \quad \epsilon \sim (0, I\sigma^2). \quad (1)$$

This model could be used for a block design where $\tau$ is a parameter vector of the effects of $v$ treatments, $\gamma$ is a parameter vector of the effects of $b$ blocks, and $X$ is the $n \times v$ design matrix of rank $v$ which has element 1 in row $u$ column $i$ if the $u$th observation is on treatment $i$; and $Z$ is the $n \times b$ design matrix of rank $b$ which has element 1 in row $u$ column $j$ if the $u$th observation is on block $j$.

(i) [3 points] Obtain an expression for the least squares estimator, $\hat{\tau}$, for $\tau$ in the form $C\hat{\tau} = Q$, where the expressions for $C$ and $Q$ are in terms of $X$ and $Z$.

(ii) [3 points] Prove that $L_v$ is an eigenvector of $C$ and that the rank of $C$ must be less than $v$.

(c) Let $C^-$ be the Moore-Penrose generalized inverse of $C$. Let $h_i'\tau, i = 1, \ldots, g$, be a set of $g$ contrasts in the treatment effects $\tau$, and let $H = [h_1, \ldots, h_g]'$.

(i) [2 points] Give a formula for the average variance of the contrast estimators $h_1'\hat{\tau}, h_2'\hat{\tau}, \ldots, h_g'\hat{\tau}$ in terms of $H$ and $C^-$.

(ii) [2 points] Suppose that $C$ has rank $v - 1$. Let $x_1, x_2, \ldots, x_v$ be a set of orthonormal eigenvectors for $C$ corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{v-1}, 0$. Write $C^-$ in terms of its spectral decomposition and, hence, give a formula for the average variance of the least squares estimators of the contrasts $h_1'\hat{\tau}, h_2'\hat{\tau}, \ldots, h_g'\hat{\tau}$ in terms of $H$ and the eigenvalues and eigenvectors of $C$. 

14
(iii) [6 points] The objective in designing a blocked experiment is often to minimize the average variance of the least squares estimators of a set of contrasts \( H \hat{\tau} \). Thus, it is useful to have a lower bound for their average variance.

For \( i = 1, \ldots, v - 1 \), set \( u_i^2 = \lambda_i^{-1} x_i' H' H x_i \) and \( v_i^2 = \lambda_i \). Use the Cauchy-Schwarz inequality to obtain a lower bound for the average variance of the above contrast estimators, in the class of designs for which \( \text{trace}(C) = t \) where \( t \) is some positive constant. [This bound is a function of some or all of the eigenvalues and eigenvectors of \( C \), \( t \) and \( H \)]. In terms of \( H \) and \( x_1, \ldots, x_v \), give values of \( \lambda_1, \lambda_2, \ldots, \lambda_v \) for which the bound is achieved.

(iv) [3 points] Suppose that \( C = aI_y - bv^{-1} J_y \). [The \( C \) matrices for a complete block design and for a balanced incomplete block design have this form]. Show that, if the rows of \( H \) are a set of \( g = v - 1 \) orthonormal contrasts, then these form a set of \( v - 1 \) orthonormal eigenvectors of \( C \). Obtain the corresponding eigenvalues.

(v) [3 points] Using any of the information that you have gained so far, and \( C \) and \( H \) as defined in part (iv), prove that the average variance of the contrast estimators \( H \hat{\tau} \) is a minimum over all block designs of the same size and with fixed \( \text{trace}(C) = t \).
4. Solution

(a) The least squares estimator minimizes the sum of squares of the errors, which is equivalent to minimizing the length of \( y - W\theta \). Since \( W\theta \) lies in the column space of \( W \), \( y - W\theta \) is minimized if it is orthogonal to the column space of \( W \). Thus we need \( \hat{\theta} \) such that

\[ W'(y - W\theta) = 0; \]

that is, \( W'W\hat{\theta} = W'y \), or \( \hat{\theta} = (W'W)^{-1}W'y \), where \( (W'W)^{-1} \) is a generalized inverse of \( W'W \).

(b) (i) Obtain an expression for \( \hat{\tau} \):

\[ W'W\hat{\theta} = W'y \]

\[
\begin{bmatrix}
X'X & X'Z \\
Z'X & Z'Z
\end{bmatrix}
\begin{bmatrix}
\hat{\tau} \\
\hat{\gamma}
\end{bmatrix}
= \begin{bmatrix}
X'y \\
Z'y
\end{bmatrix}
\]

This gives the following equations

\[ X'X\hat{\tau} + X'Z\hat{\gamma} = X'y \quad (2) \]
\[ Z'X\hat{\tau} + Z'Z\hat{\gamma} = Z'y \quad (3) \]

Note that \( Z'Z \) is \( b \times b \) of rank \( b \), so its inverse exists. Pre-multiply (3) by \( X'Z(Z'Z)^{-1} \) and subtract from (2) to give

\[ (X'X - X'Z(Z'Z)^{-1}Z'X)\hat{\tau} = X'y - X'Z(Z'Z)^{-1}Z'y, \quad (4) \]

where \( C = X'X - X'Z(Z'Z)^{-1}Z'X \) and \( Q = X'y - X'Z(Z'Z)^{-1}Z'y \). We can write equation (4) as

\[ X'(I - P)X\hat{\tau} = X'(I - P)y \]

with

\[ P = Z(Z'Z)^{-1}Z'. \]

(ii)

\[ C1_v = X'(I - P)X1_v = X'(I - P)1_n. \]
But

\[ P1_n = Z(Z'Z)^{-1}Z'1_n = Z \begin{bmatrix} \frac{1}{k_1} \\ \frac{1}{k_2} \\ \vdots \\ \frac{1}{k_b} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_b \end{bmatrix} = Z1_b = 1_n, \]

where \( k_j \) is the number of observations in block \( j \). So \( C1_v = X'(I - P)1_n = 0_v = 01_v \) and \( 1_v \) is an eigenvector of \( C \) with eigenvalue zero. Since the rank of a matrix is equal to the number of non-zero eigenvalues, the rank of \( C \) must be at most \( v - 1 \).

(c) (i) Using the fact that \((I - P)\) is symmetric and idempotent,

\[ \text{Var}(Q) = X'(I - P)I\sigma^2(I - P)X = X'(I - P)X\sigma^2 = C\sigma^2. \]

So,

\[ \text{Var}(H\hat{\cdot}) = HC^{-} \text{Var}(Q)C^{-}H' = HC^{-}CC^{-}H'\sigma^2 = HC^{-}H'\sigma^2, \]

since \( C^{-} \) satisfies \( C^{-}CC^{-} = C^{-} \) (Moore-Penrose inverse). So the average variance is \( g^{-1}\text{trace}(HC^{-}H')\sigma^2 \).

(ii)

\[ C^{-} = \sum_{i=1}^{v-1} \lambda_i^{-1}x_i'x_i'. \]

So, the average variance is

\[ \sigma^2 g^{-1}\text{trace}(HC^{-}H') = \sigma^2 g^{-1}\text{trace}\left( \sum_{i=1}^{v-1} \lambda_i^{-1}Hx_i'x_i'H' \right) = \sigma^2 g^{-1} \sum_{i=1}^{v-1} \lambda_i^{-1}x_i'H'Hx_i. \]

(iii) From (ii), the average variance is \( \sigma^2 g^{-1}\sum_{i=1}^{v-1} u_i^2 \). The Cauchy-Schwarz inequality says

\[ (\sum_{i=1}^{v-1} u_i v_i)^2 \leq (\sum_{i=1}^{v-1} u_i^2)(\sum_{i=1}^{v-1} v_i^2) \quad \text{with equality holding when } u_i = \alpha v_i \text{ for all } i. \]
and some constant $\alpha$, so

$$\left(\sum_{i=1}^{v-1} \sqrt{x_i'H'Hx_i}\right)^2 \leq \left(\sum_{i=1}^{v-1} \lambda_i^{-1}x_i'H'Hx_i\right) \left(\sum_{i=1}^{v-1} \lambda_i\right).$$

But $\sum_{i=1}^{v-1} \lambda_i = t$. So,

$$\text{Ave. Var}(H\hat{\tau}) \geq t^{-1} \left(\sum_{i=1}^{v-1} \sqrt{x_i'H'Hx_i}\right)^2 \frac{\sigma^2}{g}.$$

The equality is satisfied if $u_i = \alpha v_i$ for all $i$ and some constant $\alpha$; that is, if

$$\lambda_i^{-1/2} \sqrt{x_i'H'Hx_i} = \alpha \lambda_i^{1/2},$$

or if the eigenvalues of $C$ satisfy

$$\lambda_i = \alpha^{-1} \sqrt{x_i'H'Hx_i},$$

for some constant $\alpha$.

(iv) Since the rows of $H$ are contrasts, it follows that $h_j'1_v = 0$ so $(aI - bv^{-1}J_v)h_j = ah_j$ for $j = 1, \ldots, v - 1$. So a set of orthonormal eigenvectors for $C$ is $\{h_1, \ldots, h_{v-1}, 1_v\}$ corresponding to eigenvalues $a, \ldots, a, 0$.

(v) Using the orthonormal eigenvectors $\{h_1, \ldots, h_{v-1}, 1_v\}$, $h_j'H'Hh_j = 1$ so (5) is satisfied with $\alpha^{-1} = a$, and the average variance is a minimum.
1. Consider a simple linear regression model,

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \]

where \( E[\varepsilon_i] = 0 \), \( \text{var}[\varepsilon_i] = \sigma^2 \), and \( \{\varepsilon_i\} \) are mutually independent normal random variables.

Let \( Y' = (Y_1, Y_2, \cdots, Y_n) \), \( \beta' = (\beta_0, \beta_1) \), and \( X' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \). Furthermore, for known constants \( \{z_i : i = 1, 2, \ldots, n\} \),

let \( Z' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_n \end{bmatrix} \), and assume that \( Z'X \) is a non-singular matrix.

(a) [3 points] Show that the instrumental-variable estimator, \( \tilde{b} = (Z'X)^{-1}Z'Y \), is an unbiased estimator of the vector \( \beta \).

(b) [5 points] Find the sampling distribution of \( \tilde{b} = (\tilde{\beta}_0, \tilde{\beta}_1)' \), including its mean vector and variance-covariance matrix. Furthermore, let \( b = (b_0, b_1)' \) be the OLS estimator of \( \beta \). Prove that \( \text{var}(\tilde{\beta}_1) \geq \text{var}(b_1) \).

(c) Let \( P_Z \) denote the projection matrix on the column space \( C(Z) \).

i [4 points] Show that \( Y'(I - P_Z)Y/\sigma^2 \) has a chi-squared distribution. Give its degrees of freedom and the non-centrality parameter.

ii [3 points] What is the value of the non-centrality parameter when \( \beta_1 = 0 \)? Justify your answer.

(d) Let the residuals corresponding to the instrumental-variable estimator \( \tilde{b} \) be denoted by \( \tilde{\varepsilon} = Y - X\tilde{b} \).

i [4 points] Show that \( \tilde{b} \) and \( \tilde{\varepsilon} \) are independently distributed and give the sampling distribution of \( \tilde{\varepsilon} \).

ii [4 points] Now let \( Q \) denote the \( n \times n \) matrix so that \( \tilde{\varepsilon} = QY \). Show that the matrix \( Q \) is not symmetric. Is the matrix \( Q \) idempotent? Explain why or why not.

iii [2 points] Give a necessary and sufficient condition for the distribution of \( \tilde{\varepsilon}'\tilde{\varepsilon}/\sigma^2 \) to be a chi-squared distribution.
QII Solutions - Spring 2009

Disclaimer: These are not meant to be model solutions, but are intended to give an outline of how the problem might be solved.

1. (a) \( E[\tilde{b}] = (Z'X)^{-1}Z'E[Y] = (Z'X)^{-1}Z'X\beta = \beta \), since \( Z'X \) is non-singular.

(b) Now, \( \text{var}(\tilde{b}) = (Z'X)^{-1}Z'\text{var}(Y)Z(Z'X)^{-1} = \sigma^2(Z'X)^{-1}Z'Z(Z'X)^{-1}. \)

Since \( \tilde{b} \) is a linear form in \( Y \), it follows that \( \tilde{b} \sim N[\beta, \sigma^2(Z'X)^{-1}Z'Z(Z'X)^{-1}] \).
Furthermore, since \( \tilde{b} \) is the unique linear unbiased estimator of \( \beta \) with minimum variance, and \( \tilde{b} \) is a linear unbiased estimator, it follows that \( \text{var}(\tilde{b}_1) \geq \text{var}(\tilde{b}_1) \).

(c) Note that \( Y \sim N(X\beta, \sigma^2I) \) and \( I - P_Z \) is idempotent with rank \( n - 2 \).

i It follows that \( \frac{1}{\sigma^2}Y'(I - P_Z)Y \sim \chi^2_{n-2}(\varphi = \frac{1}{\sigma^2}\beta'X'(I - P_Z)X\beta). \)

ii If \( \beta_1 = 0 \), then \( X\beta = \beta_01_n \). Now, since \( 1_n \in C(Z) \), it follows that \( \beta'X'(I - P_Z) = 0 \). Hence, \( \varphi = 0 \).

(d) Now, \( E[\tilde{e}] = E[Y - X\tilde{b}] = X\beta - X\beta = 0. \)

i Since \( \tilde{b} = (Z'X)^{-1}Z'Y \) and \( \tilde{e} = (I - X(Z'X)^{-1}Z')Y = QY \), where, \( Q = (I - X(Z'X)^{-1}Z') \), clearly \( \tilde{e} \) and \( \tilde{b} \) are linear forms in \( Y \), which has a multivariate normal distribution. It follows that the joint distribution of \( \tilde{e} \) and \( \tilde{b} \) is multivariate normal. Furthermore, since \( (Z'X)^{-1}Z'Q = 0 \), it follows that \( \text{cov}(\tilde{e}, \tilde{b}) = 0 \). Therefore, \( \tilde{e} \) and \( \tilde{b} \) are independently distributed. Finally, \( \text{var}(\tilde{e}) = \text{var}(QY) = \sigma^2[QQ'] \). Thus, \( \tilde{e} \sim N(0, \sigma^2QQ') \).

ii Note that \( Q' = I - Z(Z'X)^{-1}X' \neq I - X(Z'X)^{-1}Z' \), therefore it is not symmetric. However, \( Q \) is idempotent, since

\[
QQ = I - 2X(Z'X)^{-1}Z + X(Z'X)^{-1}Z'X(Z'X)^{-1}Z' = I - 2X(Z'X)^{-1}Z' + X(Z'X)^{-1}Z' = Q.
\]

iii For the distribution of \( \tilde{e}'\tilde{e}/\sigma^2 \) to be chi-squared, it is necessary and sufficient that \( QQ' \) be symmetric and idempotent. [Clearly, it is symmetric. However, it can be verified that it is not idempotent. Thus, \( \frac{1}{\sigma^2}\tilde{e}'\tilde{e} \) does not have a chi-squared distribution. Note that one can also write the condition above as “\( QQ' \) is symmetric and idempotent”. In fact, it can be shown that the two are equivalent.]
6. [25 points] Consider the linear regression model,

\[ y_i = x_{i1} \beta_1 + \cdots + x_{iK} \beta_K + \varepsilon_i; \quad i = 1, \ldots, n, \]

where \( \varepsilon_1, \ldots, \varepsilon_n \) are independent and identically distributed \( N(0, \sigma^2) \) random variables. We assume that \( \sigma^2 \) is known and, without loss of generality, that \( \sigma^2 = 1 \). Also, assume that the true \( K \)-dimensional vector \( \beta \) is of the form \( (\beta_1, \ldots, \beta_{k_0}, 0, \ldots, 0)' \) for some \( k_0 \leq K \). Now, consider model-selection criteria of the form,

\[ C(k, \lambda) \equiv \text{RSS}(k) + \lambda k; \quad k = 0, \ldots, K, \]

where \( \text{RSS}(k) \) is the residual sum of squares of the least squares fit with the first \( k \) covariates only, and \( \lambda > 0 \) represents a penalty for overfitting. For example, \( \lambda = 2 \) corresponds to the Mallows \( C_p \) criterion and the Akaike information criterion (AIC), and \( \lambda = \log(n) \) corresponds to the Bayesian information criterion (BIC). Generally, the model dimension is chosen by minimizing the criterion over \( k = 0, \ldots, K \).

Given the true model dimension \( k_0 \), this question examines the effect of adding extra variables on the model-selection criterion \( C(k, \lambda) \) by restricting the selection range \( k \) to \( k_0, \ldots, K \). Define \( \hat{k}_\lambda = \arg \min_{k_0 \leq k \leq K} C(k, \lambda) \).

(a) [4 points] Fit a linear model with the first \( k \) covariates to the data and call the fit \( \hat{Y}_k \).

Let \( P_k = X'_k(X_kX_k)^{-1}X'_k \), where \( X_k \) is the design matrix of a linear model with the first \( k \) covariates. Show that for \( k = k_0, \ldots, K \), the residual random vector \( Y - \hat{Y}_k \) is \( (I - P_k) \varepsilon \) for the fitted model, where \( Y = (y_1, \ldots, y_n)' \), \( \hat{Y}_k = (\hat{y}_1, \ldots, \hat{y}_n)' \), and \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)' \).

(b) [3 points] Verify that \( C(k, \lambda) = \varepsilon' \varepsilon - \delta_k(\lambda) \), for \( k = k_0, \ldots, K \), where \( \delta_k(\lambda) = \varepsilon'P_k\varepsilon - \lambda k \).

(c) [10 points] Consider the sequence of random variables, \( Z_i \equiv \varepsilon'(P_i - P_{i-1})\varepsilon \), for \( i = 1, \ldots, K \). By convention, define \( P_0 = 0 \), the zero matrix. Derive their joint distribution.

(d) [3 points] For the sequence defined in (c), show that

\[ \delta_k(\lambda) = \sum_{i=1}^{k} (Z_i - \lambda). \]

(e) The results in (c) and (d) show that \( \delta_k(\lambda) \) is a sequence of partial sums.

i. [2 points] How is \( \hat{k}_\lambda \) related to the path defined by the sequence \( \{\delta_k(\lambda) : k = k_0, \ldots, K\} \)?

ii. [3 points] How would you simulate the distribution of \( \hat{k}_\lambda \) to study statistical properties of the model size selected by the criterion \( C(k, \lambda) \)?
6. (a) The least squares estimator for the linear model with the first \( k \) covariates is \( \hat{\beta}_k = (X_k'X_k)^{-1}X_k'Y \). So,

\[
Y - \hat{Y}_k = Y - X_k\hat{\beta}_k \\
= (I - X_k(X_k'X_k)^{-1}X_k')Y \\
= (I - X_k(X_k'X_k)^{-1}X_k')(X_{k_0}\beta_{k_0} + \varepsilon) \\
= (I - X_k(X_k'X_k)^{-1}X_k')\varepsilon.
\]

The last equality follows from the fact that \( X_{k_0}\beta_{k_0} \) is in the column space of \( X_k \) for \( k \geq k_0 \). Thus, \( Y - \hat{Y}_k = (I - P_k)\varepsilon \).

(b) Since \( RSS(k) = (Y - \hat{Y}_k)'(Y - \hat{Y}_k) = \varepsilon'(I - P_k)\varepsilon \) from (a),

\[
C(k, \lambda) = RSS(k) + \lambda k = \varepsilon'(I - P_k)\varepsilon + \lambda k = \varepsilon'\varepsilon - (\varepsilon'P_k\varepsilon - \lambda k).
\]

(c) i. (6 points) First note that both \( P_k \) and \( P_{k-1} \) in the definition of \( Z_k \) are projections, and \( P_kP_{k-1} = P_{k-1} \) since each column of \( P_{k-1} \) is in the span of the columns of \( X_{k-1} \). So,

\[
(P_k - P_{k-1})^2 = P_k^2 - P_kP_{k-1} - P_{k-1}P_k + P_{k-1}^2 \\
= P_k - P_{k-1},
\]

which shows that \( P_k - P_{k-1} \) is idempotent. Furthermore \( \text{rank}(P_k - P_{k-1}) = \text{tr}(P_k - P_{k-1}) = k - (k - 1) = 1 \). From the assumption \( \varepsilon \sim N(0, I) \), the quadratic form \( \varepsilon'(P_k - P_{k-1})\varepsilon \) follows a \( \chi_1^2 \) distribution for \( k = 1, \ldots, K \).

ii. (4 points) Second, observe that for \( k > j \),

\[
(P_k - P_{k-1})(P_j - P_{j-1}) \\
= P_kP_j - P_kP_{j-1} - P_{k-1}P_j + P_{k-1}P_{j-1} \\
= P_j - P_{j-1} = 0.
\]

So, two quadratic forms, \( \varepsilon'(P_k - P_{k-1})\varepsilon \) and \( \varepsilon'(P_j - P_{j-1})\varepsilon \) are independent. By i and ii, \( Z_1, \ldots, Z_K \) are independent and identically distributed \( \chi_1^2 \).

Alternatively, the result can be proved by Cochran’s theorem. Observe that

\[
I = (I - P_K) + (P_K - P_{K-1}) + \cdots + (P_1 - P_0).
\]

Since \( \text{rank}(P_k - P_{k-1}) = 1 \) for \( k = 1, \ldots, K \), as shown before, and \( \text{rank}(I - P_K) = n - K \),

\[
\text{rank}(I) = \text{rank}(I - P_K) + \sum_{k=1}^{K} \text{rank}(P_k - P_{k-1}).
\]

Thus, for \( \varepsilon \sim N(0, I) \), the quadratic forms \( Z_k = \varepsilon'(P_k - P_{k-1})\varepsilon \); \( k = 1, \ldots, K \), are independent, and each \( Z_k \) is distributed \( \chi_1^2 \) by Cochran’s theorem.
(d) \[ \sum_{i=1}^{k} (Z_i - \lambda) = \sum_{i=1}^{k} (\epsilon'(P_i - P_{i-1})\epsilon - \lambda) = \epsilon'P_k \epsilon - \lambda k = \delta_k(\lambda). \]

(e) i. \( \hat{k}_\lambda = \arg \min_{k_0 \leq k \leq K} C(k, \lambda) = \arg \max_{k_0 \leq k \leq K} \delta_k(\lambda). \) So, \( \hat{k}_\lambda \) is alternatively defined as the index \( k \) with the maximum value of the random walk \( \delta_k(\lambda) \) among \( k \geq k_0. \)

ii. Since \( \hat{k}_\lambda \) is the index \( k \) with the maximum of the sequence of partial sums of iid random variables \( (\chi^2_1 - \lambda) \) from \( k_0 \) to \( K \), take the following steps repeatedly to simulate the distribution of \( \hat{k}_\lambda \).

A. Generate \( K \) values \( z_1, \ldots, z_K \) iid from \( \chi^2_1. \)

B. Compute the partial sums \( d_k = \sum_{i=1}^{k} (z_i - \lambda) \) for \( k = k_0, \ldots, K. \)

C. Find the index \( k \) with the maximum value of \( d_k \), which gives a realization of \( \hat{k}_\lambda. \)
4. This question involves a classical problem in randomized clinical trials. To make the problem simple, consider the comparison of one active drug and a placebo. Although in practice the patients have been stratified into homogeneous strata, in this question we consider an analysis of only one of the homogeneous strata of interest. The number of patients in the stratum is \( n \) (\( n \) even), and we randomize assignment of patients to one of the two treatments so that each treatment has \( n/2 \) patients. The randomization introduces extra variability but it does have benefits.

(a) [3 points] For \( i = 1, \ldots, n \), let

\[
\hat{j}_i = \begin{cases} 
1; & \text{patient } i \text{ is assigned the active drug} \\
2; & \text{patient } i \text{ is assigned the placebo}.
\end{cases}
\]

Define a probability space on the \( 2^n \) possible values of \( j \equiv (j_1, \ldots, j_n) \), where the probability measure \( P(\cdot) \) assigns equal probability to all \( j \) such that \( \sum_{i=1}^n I(j_i = 1) = \sum_{i=1}^n I(j_i = 2) = n/2 \); and \( P(j) = 0 \) for any other \( j \). Give a formula for that equal probability. [\( I(\cdot) \) denotes the indicator function.]

(b) Let \( \tau_j \) denote the effect of treatment \( j \) (\( j = 1, 2 \)), and suppose that the response from individual \( i \) (\( i = 1, \ldots, n \)) after receiving treatment \( j \) is

\[
Y_{ij} = \tau_j + e_i,
\]

where \( \{e_i : i = 1, \ldots, n\} \) are the individual effects, assumed fixed. Of course, each individual gets only one of the two treatments, but it is still possible to define \( Y_{ij} \) for all \( i \) and all \( j \).

For \( j = 1, 2 \), let \( i_j \equiv (i_{1j}, \ldots, i_{nj/2})' \) denote the set of individuals assigned to treatment \( j \); that is, \( i_{kj} \) refers to individual \( k \) being assigned treatment \( j \) and given the index \( k \) in the treatment-\( j \) group of patients. Then it is possible to show that

\[
P(i_{kj} = \ell) = n^{-1}
\]

\[
P(i_{kj} = \ell, i_{kj'} = \ell') = \frac{n(n-1)}{n} I((k, j) \neq (k', j'))
\]

for all \( j, j' = 1, 2, k, k' = 1, \ldots, n/2 \), and \( \ell \neq \ell' = 1, \ldots, n \). You do not have to show this, but you will need it for the rest of this question.

Define \( Y_{obs} \) as the vector of responses taken from individuals assigned treatment \( j \); \( j = 1, 2 \). Then from the model above, the \( k \)-th element of \( Y_{obs} \) is, \( \tau_j + e_{ikj} \). Further define

\[
Y_{obs} \equiv \begin{bmatrix} Y_{1obs} \\ Y_{2obs} \end{bmatrix}
\]

i. [7 points] Under the randomization distribution derived in (a), show that \( E(Y_{obs}) \) has the form,

\[
\mu 1_n + \begin{bmatrix} 1_{n/2} & 0 \\ 0 & 1_{n/2} \end{bmatrix} \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix},
\]
where $1_m$ is an $m$-dimensional vector of 1's, and give the expression for $\mu$. Further show that $\text{cov}(Y^{obs})$ has the form,

$$aI_n + b1_n1_n',$$

where $I_n$ is the $n \times n$ identity matrix, and give the expression for $a$ and $b$.

ii. [5 points] Describe how you would test the null hypothesis, $H_0: \tau_1 = \tau_2$, versus the alternative hypothesis, $H_1: \tau_1 \neq \tau_2$.

(c) Until now, $e \equiv (e_1, \ldots, e_n)'$ has been assumed fixed. These individual effects can be modeled statistically:

$$e \sim N(\nu, \Sigma),$$

and it is assumed that this source of random variation is independent of the randomization.

i. [5 points] Using (b) or otherwise, give an expression for

$$E(Y^{obs})$$

and

$$\text{cov}(Y^{obs}),$$

where both sources of random variation are accounted for.

ii. [2 points] Describe a test of $H_0: \tau_1 = \tau_2$, versus $H_1: \tau_1 \neq \tau_2$, based on your results.

(d) [3 points] Discuss any differences in inference on $\tau_1 - \tau_2$ between the situation in (b) and the situation in (c). What are the implications of this for randomized clinical trials?
4. (a) A probability space consists of a triple \((\Omega, \mathcal{F}, P)\). Here \(\Omega = \{j: j_i = 1, 2; \ i = 1, \ldots, n\}\), which is a finite set of \(2^n\) elements. The \(\sigma\)-algebra \(\mathcal{F}\) is the set of all subsets of \(\Omega\), and \(P(\cdot)\) is defined by a randomization distribution that chooses \(n/2\) patients from a total of \(n\). There are \(\binom{n}{n/2}\) ways of doing this, and hence

\[
P(j) = \begin{cases} \binom{n}{n/2}^{-1} & \text{if } \sum_i I(j_i = 1) = \sum_i I(j_i = 2) = n/2 \\ 0 & \text{otherwise} \end{cases}
\]

(b) i. The \(k\)-th element of \(Y_{j}^{\text{obs}}\) is \(Y_{k,j}^{\text{obs}} = \tau_j + e_{k,j}\). Then

\[
E(Y_{k,j}^{\text{obs}}) = \tau_j + \sum_{\ell=1}^{n} e_{\ell}/n = \mu + \tau_j,
\]

where \(\mu = \bar{e}\). Consequently,

\[
E(Y^{\text{obs}}) = \begin{bmatrix} \mu + \tau_1 \\ \vdots \\ \mu + \tau_n \end{bmatrix} = \mu 1_n + \begin{bmatrix} 1_{n/2} & 0 \\ 0 & 1_{n/2} \end{bmatrix} \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix}.
\]

Consider \(j = 1\) and calculate \(\text{cov}(Y_{i_{k1}}, Y_{i_{k'1}})\):

For \(k = k'\), it is:

\[
\sum_{\ell} e_{\ell}^2/n - (\bar{e})^2 = \sum_{\ell} (e_{\ell} - \bar{e})^2/n = S^2_e.
\]

For \(k \neq k'\), it is:

\[
\sum_{\ell \neq \ell'} e_{\ell}e_{\ell'}/n(n-1) - \bar{e}^2 = S^2_e - \sum_{\ell \neq \ell'} \sum_{\ell} (e_{\ell} - e_{\ell'})^2/(2n(n-1))
\]

\[
= S^2_e - \mathbf{e}'A\mathbf{e},
\]

where \(\mathbf{e}'A\mathbf{e} = \sum_{\ell \neq \ell'} (e_{\ell} - e_{\ell'})^2/(2n(n-1))\).

Likewise, the same algebra holds for all non-diagonal elements, and hence

\[
\text{cov}(Y^{\text{obs}}) = (\mathbf{e}'A\mathbf{e})I_n + (S^2_e - \mathbf{e}'A\mathbf{e})1_n1'_n = aI_n + b1_n1'_n
\]

ii. There are two approaches:

Define the test statistic

\[
D^{\text{obs}} \equiv (n/2)^{-1}(Y^{\text{obs}})'1_{n/2} - (n/2)^{-1}(Y^{\text{obs}})'1_{n/2}.
\]
or for that matter any other functions of $Y_{1}^{obs}$ and $Y_{2}^{obs}$ that reflects the difference between the two treatments. Under the null hypothesis, there are \( \binom{n}{n/2} \) values \{D_i\} that can be computed; order them. If $D^{obs}$ is below the $\alpha/2$ quantile or above the $(1 - \alpha/2)$ quantile, then reject $H_0$.

The other approach is to approximate the randomization distribution with a normal distribution (e.g., Kempthorne, 1952, Ch. 8) and use test statistic

$$T^{obs} \equiv \sqrt{n/4} \frac{D^{obs}}{S^{obs}},$$

where $(S^{obs})^2$ is the sample variance of entries of $Y^{obs}$. Under $H_0$, $T^{obs} \sim t_{n-1}$; then reject $H_0$ if $T^{obs} < t_{n-1}(\alpha/2)$ or $T^{obs} > t_{n-1}(1 - \alpha/2)$.

(c) i. The $k$-th element of $Y_{j}^{obs}$ is:

$$Y_{kj}^{obs} = \tau_j + e_{kj}.$$

Then, given $e$, from (b) we have that $E(Y_{kj}^{obs} | e) = \tau_j + \bar{e}$. Hence,

$$E(Y_{kj}^{obs}) = \tau_j + E(\bar{e}) = \tau_j + \bar{v}.$$

In what follows, write $S_{\epsilon}^2 \equiv e'Qe$. Then

$$\text{var}(Y_{kj}^{obs}) = E(\text{var}(Y_{kj}^{obs} | e)) + \text{var}(E(Y_{kj}^{obs} | e))$$

$$= E(S_{\epsilon}^2) + \text{var}(\bar{e})$$

$$= E(e'Qe) + \text{var}(\bar{e})$$

$$= tr(Q\Sigma) + 1'\Sigma 1/n^2,$$

where the result, $E(e'Qe) = tr(Q\Sigma)$ is not needed to get full credit. Consider, for example, $j = 1$ and calculate $\text{cov}(Y_{k1}^{obs}, Y_{k'1}^{obs})$. For $k \neq k'$:

$$\text{cov}(Y_{k1}^{obs}, Y_{k'1}^{obs}) = E(S_{\epsilon}^2 - e'\Lambda e) + \text{var}(\bar{v})$$

$$= tr((Q-A)\Sigma) + 1'\Sigma 1/n^2.$$

We obtain the same result for $\text{cov}(Y_{k1}^{obs}, Y_{k2}^{obs})$.

ii. Likewise, $E(Y^{obs})$ and $\text{cov}(Y^{obs})$ takes the same form as in (b), and the hypothesis tests are the same.

(d) There is no difference in the testing procedure between the situation in (b) and the situation in (c). This shows that randomization allows testing to proceed in the same way when patient effects are considered fixed or random. (The conclusions generalize to $J$ treatments.)
7. Let \( Y_0, Y_1, \ldots, Y_n \) follow a general linear model,
\[
Y_i = x_i^T \beta + \delta_i; \quad i = 0, 1, \ldots, n,
\]
where the regression vector \( \beta \) is \( p \)-dimensional, \( x_0, x_1, \ldots, x_n \) are known covariates, and \( \delta_0, \delta_1, \ldots, \delta_n \) are mean-zero error terms. Write \( Y \equiv (Y_1, \ldots, Y_n)' \), \( X \equiv (x_1, \ldots, x_n)' \), and \( \delta \equiv (\delta_1, \ldots, \delta_n)' \). Notice that \( Y_0 \) is not part of \( Y \) and plays a special role in this question. Define
\[
\Sigma_{YY} \equiv \text{cov}(Y), \quad \sigma_{Y0} \equiv \text{cov}(Y, Y_0), \quad \sigma_{00} = \text{var}(Y_0).
\]
The regression parameters \( \beta \) are unknown, but assume that \( \Sigma_{YY}, \sigma_{Y0}, \) and \( \sigma_{00} \) are known. Suppose that \( Y \) is observed with measurement error. The data are:
\[
Z = Y + \varepsilon,
\]
where \( \varepsilon \equiv (\varepsilon_1, \ldots, \varepsilon_n)' \) is independent of \( Y \), and \( \varepsilon_1, \ldots, \varepsilon_n \) are i.i.d. random variables with mean zero and known variance \( \sigma^2 \). Notice that there is no observation \( Z_0 \). We wish to make inference on \( Y_0 \) and \( \beta \) based on the data \( Z \) using only first- and second-moment assumptions. (In the case of \( Y_0 \) we call the inference prediction, and in the case of \( \beta \) we call the inference estimation.)

(a) [1 point] Derive an expression for \( \Sigma_{ZZ} \equiv \text{cov}(Z) \) in terms of the parameters defined above.

(b) We first consider inference on \( \beta \) (i.e., estimation).
   i. [3 points] Write down the Generalized Least Squares (GLS) criterion for estimating \( \beta \). The resulting estimator is called \( \hat{\beta}_{GLS} \); derive a formula for it. [Assume that all necessary matrix inverses exist.]
   ii. [10 points] Consider any linear estimator,
   \[
   \hat{\beta}(A) \equiv AZ,
   \]
   where \( A \) is a \( p \times n \) matrix. Now define the \( p \times p \) matrix,
   \[
   Q(A) \equiv E\{[(\beta - AZ)(\beta - AZ)']\}.
   \]
   We wish to minimize \( \ell' Q(A) \ell \) with respect to \( A \) for every \( \ell \), where \( A \) is restricted to satisfy \( E(\hat{\beta}(A)) = \beta \).
   Show that \( \hat{\beta}_{GLS} \) is a solution to this optimization problem. [Hint: Minimize \( \ell' Q(A) \ell \) with respect to \( a \equiv A' \ell \).]

(c) We now consider inference on \( Y_0 \) (i.e., prediction)
   i. [3 points] Assume \( \beta \) is known in this part of the question (the assumption will be relaxed in the next part). Define the loss function,
   \[
   L(Y_0, \hat{Y}_0) \equiv (\hat{Y}_0 - Y_0)^2.
   \]
The optimal linear predictor is obtained by minimizing $E(L(Y_0, \hat{Y}_0))$; you may assume that it is given by,

$$\hat{Y}_0(\beta) = \beta'x_0 + \sigma_y^{-1}(Z'Z)^{-1}(Z'X\beta),$$

where we have chosen to emphasize $\beta$ (assumed known in this part of the question). Give an expression for the mean squared prediction error,

$$M_1(\beta) \equiv E(\hat{Y}_0(\beta) - Y_0)^2.$$  

How does $M_1(\beta)$ vary with $\beta$?

ii. [3 points] In reality, $\beta$ is unknown. One way to deal with this is to “plug in” an estimator, $\beta(A) \equiv A'Z$, to yield the estimated predictor, $\hat{Y}_0(\beta(A))$, and the estimated mean squared prediction error, $M(\beta(A))$. Show that $\hat{Y}_0(\beta(A))$ is of the form,

$$b(A)'Z,$$

and give a formula for $b(A)$. Also give a formula for $M(\beta(A))$.

(d) Now, $M_1(\beta(A))$ does not account for the variability in $\beta(A)$. From (c)ii, recall that $\hat{Y}_0(\beta(A)) = b(A)'Z$, where $\beta(A) = A'Z$. It is not hard to show (but you do not have to show it here) that any linear unbiased predictor can be written as $\hat{Y}_0(\beta(A))$ where $\beta(A)$ is an unbiased estimator of $\beta$. You may assume that

$$M_2(A) \equiv E(b(A)'Z - Y_0)^2 = \sigma_{Y0} - \sigma_{Y0}'\Sigma_{y0}^{-1}\sigma_{y0} + (x_0 - X'\Sigma_{x0}^{-1}\sigma_{y0})'Q(A)(x_0 - X'\Sigma_{x0}^{-1}\sigma_{y0})$$

i. [3 points] Use this result to find the Best Linear Unbiased Predictor (BLUP), $Y_0^*$, of $Y_0$.

ii. [2 points] Compare the two expressions, $M_1(\beta(A))$ and $M_2(A)$, and discuss their role in providing an expression for

$$M^* \equiv E(Y_0^* - Y_0)^2.$$  

Give an expression for $M^*$. 

12
\[ \Sigma_{ZZ} = \text{cov}(Z) = \text{cov}(Y) + \text{cov}(&) \\
= \Sigma_{YY} + \sigma^2 I \]

(b) i. The GLS criterion is to minimize

\[ (Z - X\beta)'\Sigma_{ZZ}^{-1}(Z - X\beta) \]

w.r.t. \( \beta \). Differentiate w.r.t. \( \beta \) and set the result equal to 0, to obtain

\[ X'(\Sigma_{ZZ}^{-1}Z - X\beta) = 0. \]

That is,

\[ (X'(\Sigma_{ZZ}^{-1}Z)\beta = X'(\Sigma_{ZZ}^{-1}Z), \]

and hence

\[ \hat{\beta}_{GLS} = (X'(\Sigma_{ZZ}^{-1}Z)^{-1}X'\Sigma_{ZZ}^{-1}Z. \]

ii. \( \ell'Q(A)\ell = E(\ell'\beta - \alpha'Z)^2 = a'\Sigma_{ZZ}a \), where \( a' = \ell'A \), and recall that \( E(AZ) = \beta \). Hence, minimize \( a'\Sigma_{ZZ}a \) subject to \( a'X\beta = \ell'\beta \), equivalently \( X'\alpha = \ell \). Using the method of Lagrange multipliers, minimize, with respect to \( a \) and \( m \):

\[ a'\Sigma_{ZZ}a - 2m'(X'\alpha - \ell) \]

Hence, solve

\[
\begin{align*}
a'\Sigma_{ZZ} &= m'X' \\
a'X &= \ell'.
\end{align*}
\]

That is,

\[ \ell' = a'X = m'(X'(\Sigma_{ZZ}^{-1}Z)), \]

or

\[ m' = \ell'(X'(\Sigma_{ZZ}^{-1}Z)^{-1}. \]

Hence,

\[ \ell' A \equiv a' = \ell'(X'(\Sigma_{ZZ}^{-1}Z)^{-1}X'\Sigma_{ZZ}^{-1}}, \]

for all \( \ell \),

which implies that

\[ A = (X'(\Sigma_{ZZ}^{-1}Z)^{-1}X'\Sigma_{ZZ}^{-1}) \]

and \( AZ = \hat{\beta}_{GLS}. \)

(c) i. \( \hat{Y}_0(\beta) = Y_0'\beta + \sigma_Y'\Sigma_{ZZ}^{-1}(Z - X\beta) \), and hence

\[ M_1(\beta) = \text{var}(\sigma_Y'\Sigma_{ZZ}^{-1}(\delta + \varepsilon) - \delta_0) = \sigma_{00} - \sigma_Y'\Sigma_{ZZ}^{-1}\sigma_Y, \]

which does not depend on \( \beta \).
ii. 
\[ \tilde{Y}_0(\beta) = \sigma_{Y_0}^{-1} \Sigma_{ZZ}^{-1} Z + (x_0 - X' \Sigma_{ZZ}^{-1} \sigma_{Y_0})' \beta. \]

Hence
\[ \begin{align*}
\tilde{Y}_0(\hat{\beta}(A)) & = \{ \Sigma_{ZZ}^{-1} \sigma_{Y_0} + \Lambda'(x_0 - X' \Sigma_{ZZ}^{-1} \sigma_{Y_0}) \}' Z \\
& \equiv b(A)' Z \\
M_1(\hat{\beta}(A)) & = \sigma_{y_0} - \sigma_{y_0} \Sigma_{ZZ}^{-1} \sigma_{y_0}.
\end{align*} \]

(d) i. The BLUP minimizes \( M_2(A) \) subject to \( E(\hat{\beta}(A)) = \beta \). From the expression given, this is equivalent to optimizing \( \ell' Q(A) \ell \) subject to \( E(\hat{\beta}(A)) = \beta \). From (b), the answer is \( \hat{\beta}_{GLS} \), and hence
\[ Y_0^* = x_0' \hat{\beta}_{GLS} + \sigma_{y_0} \Sigma_{ZZ}^{-1} (Z - X \hat{\beta}_{GLS}). \]

ii. Observe that \( M^* \) is \( M_2(A) \) with \( Q(A) = \text{var}(\hat{\beta}_{GLS}) = (X' \Sigma_{ZZ}^{-1} X)^{-1} \) substituted in.
Clearly, \( M_1(\hat{\beta}_{GLS}) \) is an underestimate, since
\[ M_1(\hat{\beta}_{GLS}) \leq M_2(A), \]
for all \( A \). Hence
\[ M_1(\hat{\beta}_{GLS}) \leq M^*. \]
3. Consider the general linear models,

\[ M_1: Y_1 = X_1 \beta + \varepsilon_1; \quad M_2: Y_2 = X_2 \beta + \varepsilon_2, \]

where \( Y_1 \) is an \( n_1 \times 1 \) vector, \( Y_2 \) is an \( n_2 \times 1 \) vector, \( X_1 \) and \( X_2 \) are not necessarily of full rank, and \( \varepsilon_1 \) and \( \varepsilon_2 \) are uncorrelated random vectors, with means \( 0 \) and variance-covariance matrices \( \sigma^2 I \).

(a) (3 points) Find the best linear unbiased estimators, \( T_1 \) and \( T_2 \), respectively, of functions \( \ell' \beta \) that are estimable under both models \( M_1 \) and \( M_2 \). In addition, find \( \text{Var}(T_i) \) for \( i = 1, 2 \).

(b) (3 points) Show that the minimum variance unbiased estimator of estimable \( \ell' \beta \) among all convex combinations \( T(\alpha) = \alpha T_1 + (1 - \alpha) T_2 \), where \( 0 \leq \alpha \leq 1 \), is given by \( T(\alpha^*) \) where \( \alpha^* = \omega_1/\omega_1 + \omega_2 \) where \( \omega_i^{-1} = (\text{Var}[T_i])^{-1}, i = 1, 2 \). Also find \( \text{Var}[T(\alpha^*)] \).

(c) (3 points) Consider the combined linear model,

\[ M_3: Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}. \]

Obviously, \( \ell' \beta \) in part (a) is estimable under model \( M_3 \). Find its best linear unbiased estimator \( T_3 \), and find \( \text{Var}(T_3) \).

(d) (5 points) Explain why the inequality \( \text{Var}(T_3) \leq \text{Var}([T(\alpha^*)]) \), holds in general.

(e) (6 points) Show that if either \( X_1 \) or \( X_2 \) is of rank one, then \( T(\alpha^*) = T_3 \).

(f) (5 points) Let \( A \) and \( B \) be nonnegative definite matrices. Prove that

\[ [a' A^{-1} a][a' B^{-1} a] \geq [a' (A + B)^{-1} a][a' A^{-1} a + a' B^{-1} a], \]

provided that \( a \in \mu(A) \cap \mu(B) \), where \( \mu(A) \) and \( \mu(B) \) denote the column spaces of \( A \) and \( B \) respectively, with equality when either \( A \) or \( B \) is of rank one.

[Hint: Use \( A = X_1' X_1 \) and \( B = X_2' X_2 \) in parts (d) and (e) above. You can use these parts here, even if you do not prove the earlier results.]
Solution: Problem 3

1. Using the notation, \( A = X_1'X_1 \) and \( B = X_2'X_2 \), the function \( \ell' \beta \) is estimable under each of the two models, provided \( \ell \in \mu(A) \cap \mu(B) \). Then \( T_1 = \ell' A^- X_1' Y_1 \), and \( T_2 = \ell' B^- X_2' Y_2 \). The variances are given by \( \text{Var}(T_1) = \sigma^2 \ell' A^- \ell \), and \( \text{Var}(T_2) = \sigma^2 \ell' B^- \ell \).

2. Since \( T_1 \) and \( T_2 \) are unbiased estimators of \( \ell' \beta \), so is \( T(\alpha) \) for all \( \alpha \). Furthermore, since, \((\varepsilon_1, \varepsilon_2)\) are uncorrelated random variables, \( T_1 \) and \( T_2 \) are also uncorrelated, and the variance of \( T(\alpha) \) is given by \( \text{Var}[T(\alpha)] = \alpha^2 \text{Var}(T_1) + (1 - \alpha)^2 \text{Var}(T_2) \). On setting its derivative with respect to \( \alpha \) equal to zero, it is easily seen that \( \alpha^* = \omega_1/(\omega_1 + \omega_2) \), with \( \text{Var}[T(\alpha^*)] = 1/(\omega_1 + \omega_2) \).

3. For the model M3, \( T_3 = \ell'(A + B)^- (X_1 \ X_2)' Y \), and \( \text{Var}(T_3) = \sigma^2 \ell'(A + B)^- \ell \).

4. Since \( T_3 \) is the best linear unbiased estimator based on \( Y_1 \) and \( Y_2 \), it has minimum variance among all linear unbiased estimators. Furthermore, \( T(\alpha^*) \) is also a linear unbiased estimator based on \( Y_1 \) and \( Y_2 \). Therefore, \( \text{Var}(T_3) \leq \text{Var}([T(\alpha^*)] \) holds in general.

5. Without loss of generality, assume that the rank of \( X_2 \) is one and the length of the vector \( \ell \) is equal to 1. Since \( \ell' \beta \) is an estimable function under both the models M1 and M2, each row of \( X_2 \) belongs to the space spanned by the vector \( \ell \), and \( \ell \in \mu(A) \). One can write the matrix \( A \) as a linear combination of the matrix \( \ell \ell' \) and other that are orthogonal to this matrix. Then it is easy to show that \( T_3 = \ell'(A + B)^- (X_1 \ X_2)' Y = T(\alpha^*) \).

6. This part follows from the part (e) after cancelling out \( \sigma^2 \), and rearranging terms.