Preliminary Examination: LINEAR MODELS

Answer all questions and show all work.
Q1 is 30 points; Q2 is 35 points, and Q3 is 35 points.

1. Let $Y$ be an $n$-dimensional response vector. $X_1$ and $X_2$ are $n \times p$ and $n \times q$ matrices, respectively. Suppose that the correct model is:

$$Y = X_1 \beta_1 + X_2 \beta_2 + \epsilon,$$

where $E(\epsilon) = 0$ and var$(\epsilon) = \sigma^2 I$. Suppose that we fit the following incorrect model:

$$Y = X_1 \beta_1 + \epsilon$$

with the same assumptions for $\epsilon$. Assume that $X_1$ is full rank.

a. Consider the ordinary least squares (OLS) estimator $\hat{\beta}_1$ of $\beta_1$ under the correct model. Let $\tilde{\beta}_1$ denote the OLS estimator of $\beta_1$ when we fit the incorrect model. Assuming in part (a) that $\beta_2 = 0$, compare $\hat{\beta}_1$ and $\tilde{\beta}_1$ in terms of bias and variance.

Hint:

$$\left( \begin{array}{cc} A & B \\ B' & C \end{array} \right)^{-1} = \left( \begin{array}{cc} A^{-1} & 0 \\ 0' & 0 \end{array} \right) + \left( \begin{array}{c} -A^{-1}B \\ 0' \end{array} \right) \left( C - B'A^{-1}B \right)^{-1} \left( \begin{array}{c} C - B'A^{-1}B \\ 0 \end{array} \right)$$

b. Assume in part (b) that $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. When we fit the incorrect model and would like to test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, suppose we use the usual F test statistics assuming the incorrect model:

$$F = \frac{SSM/p}{SSE/(n - p)}.$$ 

Give the expressions of the model sum of squares ($SSM$) and residual sum of squares ($SSE$). What are the actual distributions of $SSM$ and $SSE$ under $H_0$ (and the correct model), respectively? Comment on the validity of this F test for $H_0 : \beta_1 = 0$. 
2. Consider the normal linear model \( Y = X\beta + \varepsilon \) where \( X \) is an \( n \times p \) design matrix with full rank and \( \varepsilon \sim \mathcal{N}_p(0, \sigma^2 I_n) \). Let \( \hat{\beta} \) be the ordinary least squares (OLS) estimator of \( \beta \). The ridge regression estimator \( \hat{\beta}_R(\lambda) \) of \( \beta \) is the vector-value of \( \beta \) that minimizes
\[
Q(\beta) = (Y - X\beta)'(Y - X\beta) + \lambda \beta'\beta,
\]
where \( \lambda > 0 \) is a fixed real number.

a. Find an expression for \( \hat{\beta}_R(\lambda) \). Show that you can write \( \hat{\beta}_R(\lambda) = W_\lambda \hat{\beta} \) for some matrix \( W_\lambda \) that depends on \( \lambda \) and give the explicit expression for \( W_\lambda \).

b. Find the mean, bias and variance of \( \hat{\beta}_R(\lambda) \).

c. Show that \( \text{Var}(\hat{\beta}_R(\lambda)) < \text{Var}(\hat{\beta}) \) in the sense that \( \text{Var}(\hat{\beta}_R(\lambda)) - \text{Var}(\hat{\beta}_R(\lambda) - \beta) \) is a positive definite matrix. Hint: You may use \( X'X = PDP' \) where \( D \) is a diagonal matrix and \( P \) is an orthogonal matrix.

d. Find the mean squared error of \( \hat{\beta}_R(\lambda) \), \( \text{MSE}(\lambda) = E\{ (\hat{\beta}_R(\lambda) - \beta)'(\hat{\beta}_R(\lambda) - \beta) \} \).

e. Assuming now that \( X'X = dI_p \) where \( d \) is a fixed constant, find the optimum value of \( \lambda \) that minimizes the \( \text{MSE}(\lambda) \).

3. This problem concerns the situation where doubts are casted on the stability assumption of the regression coefficient and the independence assumption of the errors. Consider the following change-point model:
\[
Y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n,
\]
where
\[
\mu_i = \begin{cases} 
\beta_1, & \text{if } i \leq n/2; \\
\beta_2, & \text{otherwise}.
\end{cases}
\]
Suppose that we only observe \( Y_1, \ldots, Y_n \). For simplicity, assume that the sample size is even, namely \( n = 2m \) for some integer \( m > 0 \). Let \( (\hat{\beta}_1, \hat{\beta}_2) \) be the ordinary least squares (OLS) estimator of \( (\beta_1, \beta_2) \).

a. Find the OLS estimator \( (\hat{\beta}_1, \hat{\beta}_2) \).

Assume that the errors satisfy
\[
\varepsilon_i = e_i - a e_{i-1}, \quad i = 1, \ldots, n,
\]
where \( a \in \mathcal{R} \) is a parameter controlling the dependence strength, and \( e_k, k = 0, 1, \ldots, n, \) are independent normal random variables with mean zero and variance \( \sigma^2 > 0 \).

b. Assume that \( a = 0 \). Find the joint distribution of \( (\hat{\beta}_1, \hat{\beta}_2) \). Are \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) independent in this case? Howe does the variance of \( \hat{\beta}_1 \) change when \( n \to \infty \) (for example, whether it decreases to zero linearly in \( n \), quadratically or at some other rate)?

c. Assume that \( a = 0 \). Find an unbiased estimator of \( \sigma^2 \) and derive a test for:
\[
H_0 : \beta_1 - \beta_2 = 0 \quad \text{vs} \quad H_a : \beta_1 - \beta_2 \neq 0
\]
You need to specify the test statistic and its distribution under the null hypothesis.

d. Now assume that $a = 1$. Find the joint distribution of $(\hat{\beta}_1, \hat{\beta}_2)$. Are $\hat{\beta}_1$ and $\hat{\beta}_2$ independent in this case? How does the variance of $\hat{\beta}_1$ change when $n \to \infty$ in this case (for example whether it decreases to zero linearly in $n$, quadratically or at some other rate)? Compare your result with the one in part (b) and comment on the effect of dependence among the errors. Is dependence always a “bad” thing?

e. Now suppose that $0 < a < 1$. Find the distribution of $\hat{\beta}_1$. How does the variance of $\hat{\beta}_1$ change when $n \to \infty$ (for example whether it decreases to zero linearly in $n$, quadratically or at some other rate)? Compare your result with the ones in parts (b) and (d) and comment.