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. Consider the following linear regression model, called Model 1:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ with } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$$

where $\mathbf{Y}$ is the n-dimensional response vector, $\mathbf{X}$ is the $n \times (p+1)$ design matrix, and $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. The ith row of $\mathbf{X}$ is $(1, x_{i1}, \ldots, x_{ip})'$, for $i = 1, \ldots, n$.

We consider the following transformation: Define an $n \times (p+1)$ matrix $\mathbf{Z}$ whose ith row is given below:

$$(1, z_{i1}, \ldots, z_{ip})' = (1, c_1 x_{i1}, \ldots, c_p x_{ip})'; i = 1, \ldots, n$$

where $c_1, \ldots, c_p$ are known and non-zero constants. We consider the following model, called Model 2:

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\eta}, \text{ with } \boldsymbol{\eta} \sim N(\mathbf{0}, \sigma_\eta^2 \mathbf{I}_n).$$

   a. Derive the least squares estimator for $\boldsymbol{\alpha}$ in terms of $\mathbf{X}$.

   b. Show that the fitted values under Model 1 and Model 2 are the same.

   c. Derive the mean square errors (MSE) from Model 1 and Model 2, respectively. Are they the same?

2. Consider a general linear model (GLM), $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $k_1 + k_2$ independent variables and $n \times 1$ vector of observations $\mathbf{Y}$ of a response variable. Call it the <u>full model</u>.

Let $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$ be a $p = 1 + k_1 + k_2$ dimensional vector parameter, with $\boldsymbol{\beta}_1 = (\beta_0, \beta_1, ..., \beta_{k_1})'$ and $\boldsymbol{\beta}_2 = (\beta_{k_1+1}, ..., \beta_p)'$. Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_1$ is an $n \times (1 + k_1)$ matrix whose first column has all entries 1's.

Consider also another GLM, $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ for the same $\mathbf{Y}$. Call it the <u>reduced model</u>.

Assume that $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$ under each model, where $\mathbf{I}_n$ is the identity matrix of order n. Let $SSR(F)$ and $SSE(F)$, respectively, be the sum of squares of regression, and the sum of squares of errors for the full model, $SSR(R)$ be the sum of squares of regression for the reduced model, and let $SSR(2|1) = SSR(F) - SSR(R)$.

   a. Write $SSE(F)$ as a quadratic form in $\mathbf{Y}$, $\mathbf{Y}'\mathbf{A}\mathbf{Y}$. Clearly define the matrix $\mathbf{A}$. Find the distribution of $SSE(F)$ under the *full* model. You may give the distribution of a constant multiple of $SSE(F)$.

b. Find the distribution of $SSE(F)$ under the *reduced* model. You may give the distribution of a constant multiple of $SSE(F)$.

c. Write $SSR(2|1)$ as a quadratic form in $\mathbf{Y}$, $\mathbf{Y}'\mathbf{B}\mathbf{Y}$. Clearly define the matrix $\mathbf{B}$. Find the distribution of $SSR(2|1)/\sigma^2$ under the *reduced* model.

d. Find the distribution of $SSR(2|1)/\sigma^2$ under the full model.

e. Let $MSR(2|1) = SSR(2|1)/k_2$, and $F = MSR(2|1)/MSE(F)$. Find the distribution of F under the *reduced* model. Show how it may be used to test the hypothesis $H_0 : \boldsymbol{\beta}_2 = 0$ vs. $H_1 : \boldsymbol{\beta}_2 \neq 0$ in the full model.

3. Consider the following linear model:

$$Y_i = (i/n)\beta + \epsilon_i,\ \ i = 1, 2, \ldots, n, \tag{1}$$

where the errors ϵ_i follow the time series model:

$$\epsilon_i = \rho\epsilon_{i-1} + \sqrt{1-\rho^2}e_i,\ \ \rho \in (-1, 1), \tag{2}$$

for $1 < i \leq \infty$. And $\epsilon_1 = \sqrt{1-\rho^2}e_1$. Assume that $\{e_i\}$ in (2) are iid random variables with $E(e_i) = 0$ and $E(e_i^2) = \sigma^2$, and e_i independent of ϵ_k for $k < i$. In (1), we can interpret the observations as a combination of a linear time trend (without intercept for simplicity) and time series noise. To answer the questions, you may use the following algebraic facts:

- $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$
- $\sum_{i=0}^{n} \rho^i = \frac{1-\rho^{n+1}}{1-\rho}$, where $|\rho| \neq 1$.
- $\sum_{i=1}^{\infty} i\rho^i = \frac{\rho}{(1-\rho)^2}$
- Consider two sequences $\{a_i\}$ and $\{b_i\}$. Assume that a_i is non-decreasing and non-negative with
$$\lim_{n\to\infty} \frac{a_n^2}{\sum_{i=1}^{n} a_i^2} = 0.$$
Also assume that $\sum_{i=1}^{\infty} i|b_i| < \infty$ and $\sum_{i=1}^{\infty} b_i \neq 0$. Then as $n \to \infty$,
$$\sum_{1\leq i<j\leq n} a_i a_j b_{j-i} \approx \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{j=1}^{\infty} b_j\right)$$

a. Find $cov(\epsilon_i, \epsilon_{i+k})$, $k \geq 0$. Hint: You can use mathematical induction, i.e., considering the case of $i = 1$ and $k = 1$, $i = 1$ and $k = 2$, and generalize. Or, you may first write ϵ_i in terms of $\{e_j\}$ and then derive the covariance based on the independent random variables $\{e_j\}$.

b. Denote by $\hat{\beta}_1$ the ordinary least squares estimator of β. Derive $\hat{\beta}_1$ and find $\lim_{n\to\infty} n\text{var}(\hat{\beta}_1)$.

c. Note that from (2) we have $\epsilon_i - \rho\epsilon_{i-1} = \sqrt{1 - \rho^2}e_i$. Thus, we define Z_i as follows:

(3) $$Z_i = Y_i - \rho\epsilon_{i-1} = (i/n)\beta + \sqrt{1-\rho^2}e_i$$

If Z_i's were observed, we can use the model in (3) to estimate β. Denote by $\hat{\beta}_2$ this ordinary least squares estimator of β from this. Derive $\hat{\beta}_2$ and find $\lim_{n\to\infty} n\text{var}(\hat{\beta}_2)$. Compare the latter with that in part (b).

d. Another method to estimate β is maximum likelihood estimation. Assume that $\{e_i\}$ are iid normally distributed. Derive the likelihood for parameters β, σ^2, ρ based on the model in (1) (i.e., we observe Y_i's)? Hint: to avoid inverting the covariance matrix, you may consider the following:

$$f(\epsilon_1, \ldots, \epsilon_n) = f(\epsilon_1)f(\epsilon_2|\epsilon_1)f(\epsilon_3|\epsilon_1, \epsilon_2)\cdots f(\epsilon_n|\epsilon_1, \ldots, \epsilon_{n-1}).$$