Applied Statistics Preliminary Examination
Theory of Linear Models
August 2024

Instructions:

- This preliminary examination consists of two parts: Linear Models and Design of Experiments.
- For this Linear Models portion, work all 3 problems. Neither calculators nor electronic devices of any kind are allowed. Show all your work and clearly state any theorem or fact that you use. Each of the 14 parts carries an equal weight of 10 points.
- Abbreviations/Acronyms:
  - IID – Independent and Identically Distributed
  - LSE – Least Squares Estimator.
  - OLS – Ordinary Least Squares estimator; synonymous with LSE.
  - GLS – Generalized Least Squares estimator.
  - UMVU – Uniform Minimum Variance Unbiased (Estimator).
  - BLUE – Best Linear Unbiased Estimator.
- Notation:
  - $\boldsymbol{x}^T$ or $\boldsymbol{A}^T$: indicates the transpose of vector $\boldsymbol{x}$ or matrix $\boldsymbol{A}$.
  - $\mathrm{tr}(\boldsymbol{A})$ and $|\boldsymbol{A}|$: denotes the trace and determinant, respectively, of $\boldsymbol{A}$.
  - $\boldsymbol{I}_n$: the $n \times n$ identity matrix.
  - $\boldsymbol{j}_n = (1, \ldots, 1)^T$ is an n-vector of ones, and $\boldsymbol{J}_{m,n}$ is an $m \times n$ matrix of ones.
  - $\mathbb{E}(\boldsymbol{x})$ and $\mathbb{V}(\boldsymbol{x})$: expectation and variance(-covariance) of a random vector $\boldsymbol{x}$.
  - $x_i \sim IID(\mu, \sigma^2)$ indicates that the random variables $x_1, \ldots, x_n$ are IID with mean μ and variance σ^2 with *no specific distribution assumed*.
  - $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: the m-dimensional random vector $\boldsymbol{x}$ has a multivariate normal distribution (or univariate normal if $m = 1$) with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Also: MVN_m.
  - $X \sim t(n, \lambda)$: a t distribution with n degrees of freedom and noncentrality parameter λ. If $\lambda = 0$, we write simply $X \sim t(n)$.
  - $X \sim F(n_1, n_2, \lambda)$: an F distribution with n_1 numerator and n_2 denominator degrees of freedom, and noncentrality parameter λ. If $\lambda = 0$, we write simply $X \sim F(n_1, n_2)$.
- Possibly useful results:
  - Note the inverse for the patterned matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} 3 & 0 & -2 \\ 0 & 2 & 1 \\ -2 & 1 & 3 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{A}^{-1} = \frac{1}{7}\begin{bmatrix} 5 & -2 & 4 \\ -2 & 5 & -3 \\ 4 & -3 & 6 \end{bmatrix}$$

1. Answer the following questions for the matrix $\boldsymbol{A}$ and vector $\boldsymbol{c}$ defined as:

$$\boldsymbol{A} = \begin{bmatrix} 2/3 & 0 & \sqrt{2}/3 \\ 0 & 1 & 0 \\ \sqrt{2}/3 & 0 & 1/3 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} \sqrt{2} \\ 0 \\ 1 \end{bmatrix}$$

In addition, let $\boldsymbol{x} = (x_1, x_2, x_3)^T$ be an arbitrary vector in $\mathbb{R}^3$.

(a) Determine the rank of $\boldsymbol{A}$.

(b) Find the projection of $\boldsymbol{x}$ onto $\mathcal{C}(A)$ and $\mathcal{C}(A)^{\perp}$, the column space of $\boldsymbol{A}$ and its orthogonal complement, respectively.

(c) Find the eigenvalues of $\boldsymbol{A}$.

(d) Find either the inverse of $\boldsymbol{A}$ (if non-singular), or a generalized inverse of $\boldsymbol{A}$ (if singular).

(e) Is the system of linear equations $\boldsymbol{Ax} = \boldsymbol{c}$ *consistent*? Justify your answer and, if so, find one solution.

2. Consider the following vector of 4 observations $\boldsymbol{y} = (y_1, y_2, y_3, y_4)^T$ from the linear model:

$$\begin{aligned} y_1 &= \theta_1 + \theta_2 + \varepsilon_1 \\ y_2 &= \theta_1 - \theta_3 + \varepsilon_2 \\ y_3 &= \theta_1 - \theta_2 - \theta_3 + \varepsilon_3 \\ y_4 &= \theta_3 + \varepsilon_4 \end{aligned}$$

where the $\varepsilon_i \sim IID\ N(0, \sigma^2)$, and $\{\theta_1, \theta_2, \theta_3, \sigma^2\}$ are unknown parameters. We are particularly interested in making inference on the linear combination: $\eta = \theta_1 + \theta_2 + \theta_3$. (Hint: Note the "possibly useful results" in the Instructions.)

(a) Find the UMVU estimator of η and determine its distribution.

(b) Find the UMVU estimator of σ^2 and determine its distribution. (You don't need to compute the exact algebraic expression for the UMVU; it suffices to compute the fitted values of $\boldsymbol{y}$ and explain how it would be obtained from these.)

(c) Carry out the following inferences: (i) a 95% confidence interval for η, and (ii) a test of $H_0: \theta_3 = 0$.

(d) If the model has heteroscedastic errors, $\varepsilon_i \sim IID\ N(0, \sigma^2/i^2)$, for $i = 1, \ldots, 4$, find the BLUE of $\boldsymbol{\beta} = (\theta_1, \theta_2, \theta_3)^T$, and determine its distribution. (You don't need to compute the exact algebraic expression for the BLUE but give sufficient detail to enable its computation.)

3. Consider the observations $\boldsymbol{y} = (y_{11}, \ldots, y_{32})^T$ from the linear model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}; \quad i = 1,2,3, \text{and } j = 1,2.$$

where the $\varepsilon_{ij} \sim IID\ N(0, \sigma^2)$ and $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2)^T$ are unknown parameters to be estimated. Note that the model can be written in the usual vector-matrix form: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$.

(a) Find the rank of the design matrix $\boldsymbol{X}$. (For future reference, let $\boldsymbol{G}$ denote a generalized inverse of $\boldsymbol{X}$, but you don't need to find this.)

(b) Determine which of the following functions are *estimable*:

$$\eta_1 = \mu, \quad \eta_2 = \mu + \alpha_1, \quad \eta_3 = \mu + \beta_1, \quad \eta_4 = \alpha_1 - \alpha_2, \quad \eta_5 = \beta_1 - \beta_2$$

(c) For the estimable functions in (b), find the BLUE for each and give its distribution.

(d) Show (with justification) that the following null hypothesis is *testable*:

$$H_0: \alpha_1 = \alpha_2 = \alpha_3$$

and specify how to carry out the test.

(e) Propose a set of *side-conditions* that will reparameterize the model to full rank. Specify what the new full rank matrix $\widetilde{\boldsymbol{X}}$ is, as well as the reduced parameter set.

Applied Statistics Preliminary Examination
Design of Experiments
May 2024

Instructions:

- For this Design of Experiments portion, work all 3 problems. Each of the 12 parts carries an equal weight of 10 points.

1. An experiment was conducted to determine the effects of various factors on the breaking strength of chromium alloys. Three factors were initially considered in the experiment:
   - The Alloy Type: Chromium hydride (Level 1), Nichrome (Level 2), and Ferrochrome (Level 3).
   - The Temperature at which the breaking point was measured: High (H), Medium (M), and Low (L).
   - The Technician who performed the operation of breaking the alloy. Three technicians were randomly chosen from those working at the company conducting the study and are labeled generically as Tech A, Tech B, and Tech C.

   For each of the following variations of this experiment, write down the appropriate statistical model and assumptions. Include any and all interactions that can be tested.

   (a) For each combination of the three factor levels, one alloy specimen was pressured until it broke, and the breaking strength of the specimen was measured in pounds per square inch (psi).
   (b) For each combination of the three factor levels, four alloy specimens were pressured until they broke, and the breaking strength of each specimen was measured in psi.
   (c) All tests were performed on the same strength testing apparatus, so the tests had to be spread out over time during the same week. For each combination of the three factor levels, four alloy specimens were pressured until they broke, and the breaking strength of each specimen was measured in psi. However, all of Tech A's tests were performed on Monday of that week, all of Tech B's tests were performed on Wednesday, and all of Tech C's tests were performed on Friday. This introduced another factor, Day with levels Monday (M), Wednesday (W), and Friday (F).
   (d) Due to time constraints, only 9 specimens could be tested, one at each possible combination of the Alloy Type and Temperature. Each technician was randomly assigned three of the combinations subject to the constraint that each technician had each Alloy Type and each Temperature exactly one time. The 9 observations are described in the following table:

Alloy Type	Temp	Tech	Breaking Strength
1	H	B	y_{112}
1	M	C	y_{123}
1	L	A	y_{131}
2	H	C	y_{213}
2	M	A	y_{221}
2	L	B	y_{232}
3	H	A	y_{311}
3	M	B	y_{322}
3	L	C	y_{333}

2. A factory is located near the confluence of two rivers (see sketch below). A consumer protection group believes that the factory is releasing a carcinogenic chemical into the water. In order to determine whether or not the factory is releasing the chemical into the water, the group wishes to collect samples upstream of the factory (at Locations 1 and 2) and compare to samples downstream of the factory (at Location 3). (The reason that they cannot collect samples *at* the location of the factory is because the area marked with dashed lines is fenced-in land owned by the factory and they cannot trespass.)

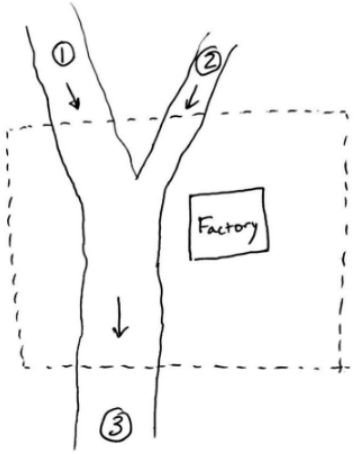
   Samples are taken at each of the marked locations over the course of 6 weeks (two samples per week). During weeks 1-3, Technician 1 collected the samples, while during weeks 4-6, Technician 2 collected the samples.

   The factors that are to be included in this model are Location (3 levels), Week (6 levels), and Technician (2 levels). Let y_{ijkl} denote the l^{th} measure of the concentration of the chemical ($l = 1,2$), in parts per million (ppm), at Location i ($i = 1,2,3$) during Week j ($j = 1,2,\dots,6$) by Technician k ($k = 1,2$). Assume that Location is a fixed factor, while Week and Technician are random factors. Also be sure to use the restricted (sum-to-zero) assumption for interactions between fixed and random factors.

   (a) Write down a model for this experiment. Clearly define all the terms and state all relevant assumptions. Include all interactions that can be included.
   (b) For each effect/term included in the model, give the degrees of freedom (df) and expected mean squares (EMS). Also specify the df and EMS for the error term as well as the total df.
   (c) For each effect/term included in the model, give the correct F ratio that would be used to test that effect/term. If no clear F ratio is present, then explain why.
   (d) The main goal of this study is to determine if the factory is releasing the chemical into the water. Let μ_i be the mean concentration of the chemical (in ppm) at Location i. The flow of water at Location 1 is twice that of Location 2, so the proper test is:

   $$H_0: \frac{2}{3}\mu_1 + \frac{1}{3}\mu_2 - \mu_3 = 0 \quad \text{vs.} \quad H_1: \frac{2}{3}\mu_1 + \frac{1}{3}\mu_2 - \mu_3 < 0$$

   Restate this hypothesis in terms of the model parameters given in (a). Is this hypothesis testable? Why or why not?



3. An experiment was conducted to determine the effectiveness of a treatment for a certain parasite that infects bobwhite quail. Bobwhite quail is a bird that is indigenous to the Eastern United States and parts of South America. A total of 16 infected quail were captured and half were randomly assigned to a control group which did not receive the treatment, while the other half were assigned to a treatment group which did receive the treatment. The birds were kept apart in individual cages, and the infection level of the parasite was measured by examining their droppings at the start of the experiment and then at 30 day intervals.

   Thus, the factors are as follows:
   - Treatment: Fixed factor with levels Control (C) and Administered (A).
   - Day: Fixed factor with levels 0 days (representing the first measurement time at the start of the experiment), 30 days, 60 days, 90 days, and 120 days.
   - Bird: Random factor that is nested within Treatment with levels 1C, 2C, ..., 8C (representing the 8 birds in the Control group) and 1A, 2A, ..., 8A (representing the 8 birds in the Administered group).

   The following model was applied by the researchers:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{k(i)} + \varepsilon_{ijk}$$

   Where:
   - y_{ijk} is the measured infection level for the i^{th} Treatment on the j^{th} Day for the k^{th} Bird (within the i^{th} Treatment). The infection level is always non-negative with larger values indicating higher infection levels.
   - α_i $(i = 1,2)$ is the fixed Treatment effect such that $\sum_i \alpha_i = 0$.
   - β_j $(j = 1,2,\dots,5)$ is the fixed Day effect such that $\sum_j \beta_j = 0$.
   - $\gamma_{k(i)}$ $(k = 1,2,\dots,8)$ is the random Bird effect such that $\gamma_{k(i)} \sim IID\ N(0, \sigma_\gamma^2)$.
   - $\varepsilon_{ijk} \sim IID\ N(0, \sigma^2)$.

   (a) Show that $Cov(y_{ijk}, y_{ij'k}) = \sigma_\gamma^2$ where $j \neq j'$, which is the covariance between infection levels for the same bird on days j and j'.

   (b) The fact shown in (a) means that the model implies that the covariance between the infection levels for the same bird is the same regardless of what days are being examined. Is this a reasonable assumption? Explain why or why not.

   (c) The expectation (and hope) of the researchers is that the treatment decreases the infection level for the birds. That is, they would expect that $\mathbb{E}[y_{2jk}]$ would decrease as j (the Day) increases. Explain what model assumptions would be violated if their hope that the treatment decreases the mean infection level is realized.

   (d) In the data that was collected for the experiment, all 8 of the birds in the control group were still infected at the conclusion of the experiment (i.e. $y_{1jk} > 0$), but 3 of the birds in the treated group were uninfected by the end of the study. In particular, Bird 2A had no infection on and after the 60th day ($y_{232} = y_{242} = y_{252} = 0$), and Bird 5A and 8A had no infection after the 90th day ($y_{245} = y_{255} = 0$ and $y_{248} = y_{258} = 0$). Explain what model assumptions are violated by this.