Applied Statistics Preliminary Examination
Theory of Linear Models
May 2026

Instructions:

- Do all 3 Problems. Neither calculators nor electronic devices of any kind are allowed. Show all your work, clearly stating 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); BLUE (best linear unbiased estimator). Sometimes the LSE may be designated OLS (ordinary least squares) estimator, in order to differentiate it from the GLS (generalized least squares) estimator.
- Notation.
  - $\boldsymbol{x}^T$ or $\boldsymbol{A}^T$: indicates transpose of vector $\boldsymbol{x}$ or matrix $\boldsymbol{A}$.
  - $\mathrm{tr}(\boldsymbol{A})$ and $|\boldsymbol{A}|$: denotes the trace and determinant, respectively, of matrix $\boldsymbol{A}$.
  - $\boldsymbol{I}_n$: 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 of random vector $\boldsymbol{x}$.
  - $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: the m-dimensional random vector $\boldsymbol{x}$ has a normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
  - $X \sim t(n, \lambda)$: 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 and n_2 numerator and denominator degrees of freedom respectively, and noncentrality parameter λ. If $\lambda = 0$ we write simply: $X \sim F(n_1, n_2)$.
- Possibly useful results.
  - Trig Identities:

$$\sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$$

$$\sin(\alpha - \beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta$$

1. Let $\boldsymbol{y} \sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$ with elements $\boldsymbol{y}^T = (y_1, \ldots, y_n)$, $\boldsymbol{X}$ is a full-rank $n \times k$ matrix with column space $C(\boldsymbol{X})$, $k < n$, $\boldsymbol{\beta}$ is a conformable vector, and denote by $\boldsymbol{P}$ the projection matrix onto $C(\boldsymbol{X})$. Introduce the notation $\theta = \boldsymbol{j}_n^T\boldsymbol{X}\boldsymbol{\beta}$, and define the following quadratic forms:

$$Q_1 = \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}, \qquad \text{and} \qquad Q_2 = \boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y},$$

and note that $Q_2 = SSE$, the usual sums of squares error from a least squares fit.

   (a) Find the distribution of $\sum_{i=1}^n y_i$ in simplest terms, as a function of θ and σ^2.

   (b) Find the distribution of Q_1/σ^2.

   (c) Find the distribution of Q_2/σ^2.

   (d) Find the distribution of:

$$\left(\frac{n-k}{n}\right)\frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}}{SSE}.$$

   (e) Determine whether or not the statistic

$$U = \sqrt{\frac{n-k}{n}}\left(\frac{\sum_{i=1}^n y_i}{\sqrt{SSE}}\right),$$

   has a familiar distribution based on the normal: t, χ^2, or F. Carefully justify your answer.

2. Consider the following time series regression model with regular oscillations for the vector of observations $\boldsymbol{y}^T = (y_1, \ldots, y_n)$ and corresponding errors $\boldsymbol{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_n)$:

$$y_i = r\sin(t_i + \theta) + \epsilon_i, \qquad i = 1, \ldots, n, \qquad \{\epsilon_i\} \sim \text{iid } \mathrm{N}(0, \sigma^2),$$

where $r > 0$, $-\pi \le \theta < \pi$, and $\sigma^2 > 0$ are unknown parameters. The known time points $\{t_i\}$ are evenly spaced in such a way that the following relations are satisfied:

$$\sum_{i=1}^n \sin^2(t_i) = \sum_{i=1}^n \cos^2(t_i) = \frac{n}{2},$$

and

$$\sum_{i=1}^n \sin(t_i) = \sum_{i=1}^n \cos(t_i) = \sum_{i=1}^n \sin(t_i)\cos(t_i) = 0.$$

   (a) Show that with the introduction of the new parameters $\beta_1 = r\sin\theta$ and $\beta_2 = r\cos\theta$, the time series regression can be written in the usual linear model form, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}^T = (\beta_1, \beta_2)$, and identify the form of the model matrix $\boldsymbol{X}$.

   (b) Find expressions for $\widehat{\boldsymbol{\beta}}$, the LSE of $\boldsymbol{\beta}$, and s^2, the usual unbiased estimate of σ^2. Also specify the distriutions of each. (The expression for s^2 can be stated in terms of $\widehat{\boldsymbol{\beta}}$.)

   (c) Construct $(1-\alpha)100\%$ confidence intervals for $\hat{\beta}_1$ and the response y_0 at the value $\boldsymbol{x}_0^T = (\cos(t_1), \sin(t_n))$.

   (d) Show that for some $k > 0$, $k(\hat{\beta}_1^2 + \hat{\beta}_2^2) \sim \chi^2(m, \lambda)$, and identify the value of the scale factor k, the degrees of freedom m, and the noncentrality parameter λ.

3. Consider the observations $\{y_1, y_2, y_3\}$ from the following linear model:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 + \beta_2 + \epsilon_1 \\ y_2 &= \beta_1 + \beta_3 + \epsilon_2 \\ y_3 &= \beta_2 + \epsilon_3 \end{aligned}$$

where $\{\epsilon_1, \epsilon_2, \epsilon_3\} \sim$ IID $\mathrm{N}(0, \sigma^2)$.

(a) Characterize all the *estimable* functions of the form $\eta = \boldsymbol{\lambda}^T \boldsymbol{\beta}$, where $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \beta_3)$, and hence determine which of the following are estimable:

$$\eta_1 = \beta_1, \qquad \eta_2 = \beta_2, \qquad \eta_3 = \beta_3, \qquad \eta_4 = \beta_1 - 2\beta_2 + \beta_3.$$

(Note: verify in particular that η_4 is estimable.)

(b) Find the BLUE of η_4 and completely determine its distribution.

(c) Find another unbiased estimator of η_4, different from the BLUE, and verify that its variance is indeed larger.

(d) Propose a test statistic for testing $H_0 : \eta_4 = 0$, and specify its distribution under both H_0 and H_1. (You should provide numerical values for all the quantities involved except SSE.)

(e) Specify a valid set of parameter constraints/side conditions in order to reparametrize the model to one with a full rank model matrix. Specify the resulting new model matrix, and verify that the BLUE of η_4 is the same as in (a).

Design of Experiment: Prelim Problems

May 2026

Please Do All Problems. Each of the 13 parts carries an equal weight of 10 points.

Question 1

A manufacturing engineer is investigating the effects of various machine operating speeds and coolant types on the surface roughness of individual parts, which is measured in micrometers. The first factor (A) in the study is machine speed, which has a fixed effect with four levels: slow, moderate, fast, and very fast. The second factor (B) is coolant type, which also has a fixed effect consisting of Type I, Type II, and Type III. The experimental units for this study are the individual manufactured parts.

For each of the following scenarios, write down the appropriate statistical model and clearly state all assumptions. Assume there are no interaction effects.

(a) Completely Randomized Factorial Design

The experiment is conducted in **one manufacturing facility**. All **240 parts** produced during the study are randomly and equally assigned to the twelve machine speed–coolant treatment combinations.

(b) Randomized Complete Block Design

The experiment is conducted in **five manufacturing facilities**, each producing **48 parts**. Suppose each facility serves as a block. The parts within each facility are randomly and equally assigned to the twelve machine speed–coolant treatment combinations.

(c) Split-Plot Design

The experiment is conducted in **five manufacturing facilities**. In each facility, there are **four production runs**, one for each machine speed. Each production run produces **12 parts**, for a total of 240 parts in the study.

Because changing machine speed requires recalibration and downtime, machine speed is randomly assigned to entire production runs within each facility. Coolant type is then randomly assigned to parts within each production run, with **four parts receiving each coolant type**.

(d) Nested Design

The experiment is conducted in twenty-four manufacturing facilities. For each machine speed–coolant combination, **two different manufacturing facilities** are randomly selected and assigned exclusively to that combination. Thus, each facility operates at only one machine speed and uses only one coolant type during the study.

Within each facility, exactly **10 parts** are produced under the assigned machine speed and coolant combination, and surface roughness is measured for all parts.

Question 2

An environmental scientist conducts an experiment to investigate how sensor type and calibration method influence measurement error, which is recorded in parts per million (ppm). The experiment includes three factors. The first factor is sensor type, which is treated as a fixed effect and has four levels: Optical, Electrochemical, Infrared, and Photoionization.

Within each sensor type, the scientist considers sensor unit as a second factor. This factor is nested within sensor type and is treated as a random effect. For each sensor type, three independently manufactured sensor units are used, resulting in a total of twelve sensor units across all sensor types.

The third factor is calibration method, which is also treated as a fixed effect and has three levels: Factory, Field, and Laboratory. Each individual sensor unit is tested under all three calibration methods. The order in which the calibration methods are applied is randomized for each sensor unit to reduce potential ordering effects. Altogether, the experimental design results in 36 observations, corresponding to the combination of four sensor types, three sensor units per type, and three calibration methods.

Let y_{ijk} denote the measurement error for the jth sensor unit of the ith sensor type using the kth calibration method, where $i = 1, \dots, 4$, $j = 1, 2, 3$, and $k = 1, 2, 3$.

The proposed mixed-effects model is

$$y_{ijk} = \mu + \alpha_i + B_{j(i)} + \gamma_k + (\alpha\gamma)_{ik} + \varepsilon_{ijk}.$$

The average measurement errors (averaged over sensor units and calibration methods) for each sensor type $\bar{y}_{i\cdot\cdot}$ are:

Sensor type	Average error (ppm)
Optical	8.2
Electrochemical	10.5
Infrared	7.6
Photoionization	9.1

(a) Write down the assumptions for the mixed-effects model and explain each term in the linear model.

(b) Complete the following ANOVA table, including sources of variation, degrees of freedom, and expected mean squares.

Source	df	MS	E(MS)
Sensor type A	_____	_____	_____
Sensor unit within type $B(A)$	_____	18	_____
Calibration method C	_____	42	_____
$A \times C$	_____	21	_____
Error	_____	6	_____
Total	_____		

(c) At 0.05 significance level, test whether calibration method has a statistically significant effect on measurement error. Clearly state the null and alternative hypotheses before conducting the test.

(d) Let $\sigma^2_{B(A)}$ denote the variance among sensor units within each sensor type. Construct a 95% confidence interval for $\sigma^2_{B(A)}$.

(e) Test whether there is a statistically significant interaction between sensor type and calibration method at 0.10 significance level. Clearly state the null and alternative hypotheses before conducting the test. Explain how the presence or absence of this interaction impacts the interpretation of the main effects.

(f) Construct 95% confidence intervals for all pairwise differences in mean measurement error among the sensor types. Write down the general formula for these confidence intervals, but perform the detailed calculation only for the comparison between the Optical sensor and the Infrared sensor.

Question 3

A cognitive psychologist studies how background music affects task completion time (measured in seconds). Fifteen participants are randomly selected and divided into three independent groups of five participants each. All participants perform tasks while listening to the same five background music pieces (labeled M1, M2, M3, M4, M5).

Task performance is known to vary across participants and can also be affected by session-specific conditions such as room temperature, lighting, or background noise. These session conditions differ across testing sessions and across groups. To simultaneously control for participant variability and session conditions, a 5×5 Latin square is implemented separately within each group. Thus, three replicated Latin squares are used.

Let y_{lijk} denote the task completion time for the lth replicate, ith session condition, jth participant, and kth music piece, where

$$l = 1, 2, 3, \quad i = 1, \ldots, 5, \quad j = 1, \ldots, 5, \quad k = 1, \ldots, 5.$$

Participants and session conditions are nested within replicate, while music pieces are common across replicates.

The proposed model is

$$y_{lijk} = \mu + \theta_l + A_{i(l)} + P_{j(l)} + \tau_k + \varepsilon_{lijk},$$

where:

- θ_l is the random effect of replicate l,
- $A_{i(l)}$ is the random effect of session condition i nested within replicate l,
- $P_{j(l)}$ is the random effect of participant j nested within replicate l,
- τ_k is the fixed effect of music piece k,
- ε_{lijk} is random error.

(a) Please complete the assumptions of the unrestricted model.

(b) Complete the following ANOVA table for this design by filling in the degrees of freedom and the expected mean squares.

Source of Variation	df	MS	E(MS)
Replicate	_____	110	_____
Session condition (Replicate)	_____	70	_____
Participant (Replicate)	_____	55	_____
Music piece	_____	130	_____
Error	_____	35	_____
Total	_____		

(c) Conduct a hypothesis test for whether background music affects task completion time at 0.05 significance level. State the null and alternative hypotheses before conducting the test.