Applied Statistics Preliminary Examination
Theory of Linear Models
August 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).
  - SSE (sum of squared errors). Also called *residual sum of squares*.
  - 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.
  - $\boldsymbol{j}_n = (1, \dots, 1)^T$ 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.
  - For a generic linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}^T = (y_1, \dots, y_n)$, the design matrix $\boldsymbol{X}$ is of dimensions $n \times k$, and $\widehat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$, the R^2 is defined as:

$$R^2 = \left(\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y} - n\overline{y}^2\right) / (\boldsymbol{y}^T\boldsymbol{y} - n\overline{y}^2), \qquad \overline{y} = \sum_{i=1}^{n} y_i/n.$$

    We also establish the following notation: if x_{ij} is the (i,j)-th element of the design matrix $\boldsymbol{X}$, let $\overline{x}_j = \sum_{i=1}^{n} x_{ij}/n$ be the mean of the j-th column.

1. Let $\boldsymbol{x} \sim \mathrm{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a full rank diagonal covariance matrix whose i-th diagonal element is σ_i^2. Define the following matrices:

$$\boldsymbol{A} = (\boldsymbol{I}_n - \boldsymbol{P})\,\boldsymbol{\Sigma}^{-1}, \qquad \boldsymbol{P} = \frac{1}{s^2}\boldsymbol{\Sigma}^{-1}\boldsymbol{J}_n, \qquad s^2 = \boldsymbol{j}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{j}_n.$$

   (a) Find the distribution of $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$.

   (b) Find the distribution of $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$.

   (c) Show that $\boldsymbol{P}$ is a projection matrix.

   (d) Find the rank of $\boldsymbol{A}$.

   (e) Find the distribution of $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$.

2. Consider the linear model $\boldsymbol{y} \sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$ with elements $\boldsymbol{y}^T = (y_1, \ldots, y_n)$, $\boldsymbol{X}$ is a full-rank $n \times k$ matrix. Let $V = C(\boldsymbol{X})$ and $V_0 = C(\boldsymbol{j}_n)$ denote the column spaces of $\boldsymbol{X}$ and $\boldsymbol{j}_n$, respectively, and let $\widehat{\boldsymbol{y}}^T = (\widehat{y}_1, \ldots, \widehat{y}_n) = \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T$, where $\widehat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$. In addition, define the following sums of squares:

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2, \qquad SSR = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2, \qquad SSE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

   (a) Specify the form of $\boldsymbol{P}_V$ and $\boldsymbol{P}_{V_0}$, the projection matrices onto the spaces V and V_0, respectively.

   (b) Show that $(\boldsymbol{y} - \overline{y})^T\boldsymbol{j}_n = 0$.

   (c) Show that $\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})(y_i - \widehat{y}_i) = 0$.

   (d) Show that $SST = SSR + SSE$.

   (e) Starting from the definition of R^2 given in the“possibly useful results” of page 1, show that $R^2 = SSR/SST$.

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

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, 2, 3, \text{ and } j = 1, \ldots, n,$$

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

   (a) Show that $\mathrm{rk}(\boldsymbol{X}) = 3$, and determine the general form of all the *estimable* functions: $\eta = \boldsymbol{\lambda}^T\boldsymbol{\beta}$.

   (b) Determine if the null hypothesis $H_0 : \alpha_2 = 2(\alpha_1 + \alpha_3)$ is *testable*, and if so propose a test statistic and specify its distribution under both H_0 and H_1. (Only numerical values for simple quantities are needed, provide formulas for the remaining terms.)

   (c) Repeat part (b) for the null hypothesis $H_0 : \mu + \alpha_1 = 2(\mu + \alpha_2) = 3(\mu + \alpha_3)$.

   (d) Specify a matrix $\boldsymbol{C}$ of full (row) rank 3 such that the null hypothesis $H_0 : \boldsymbol{C}\boldsymbol{\beta}$ is testable, and construct a test statistic as in the previous two parts.

Design of Experiment: Prelim Problems

August 2026

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

Question 1

An agricultural scientist is conducting a study to determine how different irrigation methods and fertilizer types affect the yield of individual plants, measured in grams. The first factor (A), irrigation method, has a fixed effect consisting of three levels: drip, sprinkler, and flood. The second factor (B), fertilizer type, also has a fixed effect with two levels: organic and synthetic. In this experiment, the individual plants serve as the experimental units.

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 at **one field**. All 120 plants on the filed are randomly and equally assigned to the six irrigation–fertilizer combinations.

(b) Randomized Complete Block Design
The experiment is conducted at **four randomly selected fields**, each with 30 plants. Suppose each field serves as a block. The 30 plants within each field are randomly and equally assigned to the six irrigation–fertilizer combinations.

(c) Split-Plot Design
The experiment is conducted at **two randomly selected fields**. In each field, there are **three large plots**. Each large plot contains **20 plants**, for a total of 120 plants in the study. Because changing irrigation methods requires substantial equipment adjustments, irrigation is randomly applied to entire large plots within each field. Fertilizer type is then randomly and equally assigned to subplots within each irrigation plot, with 10 plants receiving each fertilizer.

(d) Nested Design
The experiment is conducted at **twelve randomly selected fields**. For each irrigation–fertilizer combination, **two different fields** are randomly selected and assigned exclusively to that combination. Thus, each field receives only one irrigation method and one fertilizer type. Within each field, exactly **10 plants** are grown, and individual plant yields are measured for all plants.

Question 2

A food scientist conducts an experiment to examine how food type and preservation method influence bacterial concentration, which is measured in colony-forming units per gram (CFU/g). The experiment involves three factors.

The first factor is food type, which is treated as a fixed effect and includes three specific categories: beef, chicken, and fish. Within each food type, the scientist samples multiple batches to account for natural variation among production lots. Batch is treated as a random effect nested within food type, with four independently sampled batches taken for each food type. This results in a total of twelve batches across the three food types. The third factor is preservation method, which is also treated as a fixed effect. Two preservation methods are studied: freezing and vacuum sealing. Each batch is divided into two equal portions, and one of the two preservation methods is randomly assigned to each portion to ensure an unbiased comparison.

Altogether, the experimental design yields 24 observations, corresponding to the combination of three food types, four batches per food type, and two preservation methods.

Let y_{ijk} denote the bacterial concentration for the jth batch of the ith food type using the kth preservation method, where $i = 1, 2, 3$, $j = 1, \ldots, 4$, and $k = 1, 2$.

The proposed mixed-effects model is

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

The average bacterial concentrations (averaged over food types and batches) for each preservation method ($\bar{y}_{..k}$) are:

Preservation method	Average response (CFU/g)
Freezing	45.8
Vacuum sealing	49.2

(a) Write down the assumptions for the unrestricted 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)
Food type A	______	200	____________
Batch within food $B(A)$	______	30	____________
Preservation C	______	______	____________
$A \times C$	______	24	____________
Error	______	10	____________
Total	______		

(c) At the $\alpha = 0.05$ significance level, test whether food type affects the mean bacterial concentration. Clearly state the null and alternative hypotheses before conducting the test.

(d) Let $\sigma^2_{B(A)}$ denote the variance among batches within food type.Test whether the batch-to-batch variability $\sigma^2_{B(A)}$ exceeds the random error variance σ^2 at 0.05 significance level. Clearly state the null and alternative hypotheses before conducting the test.

(e) Write a contrast that represents the interaction between food type and preservation method. Then write the formula for a 95% confidence interval for this interaction contrast, clearly identifying the estimate, standard error, and appropriate critical value.

(f) Derive the expected mean square for food type, $E(MS_A)$. Do not simply quote or apply the textbook EMS rules; instead, show the algebraic steps that lead to the final expression.

Question 3

A sensory scientist studies the effect of four beverage formulations (A, B, C, D) on taste rating scores. Ratings are recorded on a numerical scale, with larger values indicating better taste.

Twelve panelists are randomly selected for the study. It is known that taste sensitivity changes over the sequence of samples due to palate fatigue. To control for this effect, three replicated Latin squares are used.

Let y_{lijk} denote the taste rating for the lth replicate, ith tasting order, jth panelist, and kth beverage formulation, where $l = 1, 2, 3$, $i = 1, \dots, 4$, $j = 1, \dots, 4$, and $k = 1, \dots, 4$.

The proposed model is

$$y_{lijk} = \mu + \theta_l + \alpha_i + B_{j(l)} + \tau_k + \varepsilon_{lijk}.$$

Here:

- θ_l is the random effect of the lth replicate,
- α_i is the fixed effect of tasting order,
- $B_{j(l)}$ is the random effect of panelist j within replicate l,
- τ_k is the fixed effect of beverage formulation,
- ε_{lijk} is random error.

(a) State all assumptions of the unrestricted model.

(b) Complete the ANOVA table for this design.

Source of Variation	df	MS	E(MS)
Replicate	______	120	______________________
Order position	______	85	______________________
Panelist within replicate	______	60	______________________
Beverage formulation	______	140	______________________
Error	______	40	______________________
Total	______		

(c) Write down the formula for constructing 95% confidence intervals for all pairwise differences in mean taste ratings among the beverage formulations. Wherever possible, substitute the numerical values or estimated quantities (such as critical values and variance estimates) directly into the formula.