

Applied Statistics Preliminary Examination

Theory of Linear Models

Aug 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.
 - $\mathbf{x}^T$ or $\mathbf{A}^T$: indicates transpose of vector $\mathbf{x}$ or matrix $\mathbf{A}$.
 - $\text{tr}(\mathbf{A})$ and $|\mathbf{A}|$: denotes the trace and determinant, respectively, of matrix $\mathbf{A}$.
 - $\mathbf{I}_n$: the $n \times n$ identity matrix.
 - $\mathbf{j}_n = (1, \dots, 1)^T$ is an n -vector of ones, and $\mathbf{J}_{m,n}$ is an $m \times n$ matrix of ones.
 - $\mathbb{E}(\mathbf{x})$ and $\mathbb{V}(\mathbf{x})$: expectation and variance of random vector $\mathbf{x}$.
 - $\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: the m -dimensional random vector $\mathbf{x}$ 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:

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

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

Problems:

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

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

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

- Find the distribution of $\sum_{i=1}^n y_i$ in simplest terms, as a function of θ and σ^2 .
- Find the distribution of Q_1/σ^2 .
- Find the distribution of Q_2/σ^2 .
- Find the distribution of:

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

- 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 $\mathbf{y}^T = (y_1, \dots, y_n)$ and corresponding errors $\boldsymbol{\epsilon}^T = (\epsilon_1, \dots, \epsilon_n)$:

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

where $r > 0$, $-\pi \leq \theta < \pi$, 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

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

- 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, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}^T = (\beta_1, \beta_2)$, and identify the form of the model matrix $\mathbf{X}$.
- Find expressions for $\hat{\boldsymbol{\beta}}$, the LSE of $\boldsymbol{\beta}$, and s^2 , the usual unbiased estimate of σ^2 . Also specify the distributions of $\hat{\boldsymbol{\beta}}$ and s^2 . (The expression for s^2 can be stated in terms of $\hat{\boldsymbol{\beta}}$.)
- Construct $(1 - \alpha)100\%$ confidence intervals for $\hat{\beta}_1$ and the mean response at the value $\mathbf{x}_0^T = (\cos(t_1), \sin(t_n))$.
- 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 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:

$$y_1 = \beta_1 + \beta_2 + \beta_2 + \epsilon_1$$

$$y_2 = \beta_1 + \beta_3 + \epsilon_2$$

$$y_3 = \beta_2 + \epsilon_3$$

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

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

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

(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).

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.
 - $\mathbf{x}^T$ or $\mathbf{A}^T$: indicates transpose of vector $\mathbf{x}$ or matrix $\mathbf{A}$.
 - $\text{tr}(\mathbf{A})$ and $|\mathbf{A}|$: denotes the trace and determinant, respectively, of matrix $\mathbf{A}$.
 - $\mathbf{I}_n$: the $n \times n$ identity matrix.
 - $\mathbf{j}_n = (1, \dots, 1)^T$ is an n -vector of ones, and $\mathbf{J}_{m,n}$ is an $m \times n$ matrix of ones.
 - $\mathbb{E}(\mathbf{x})$ and $\mathbb{V}(\mathbf{x})$: expectation and variance of random vector $\mathbf{x}$.
 - $\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: the m -dimensional random vector $\mathbf{x}$ 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 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}^T = (y_1, \dots, y_n)$, the design matrix $\mathbf{X}$ is of dimensions $n \times k$, and $\widehat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$, the R^2 is defined as:

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

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

Problems:

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

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

- Find the distribution of $\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$.
 - Find the distribution of $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$.
 - Show that $\mathbf{P}$ is a projection matrix.
 - Find the rank of $\mathbf{A}$.
 - Find the distribution of $\mathbf{x}^T \mathbf{A} \mathbf{x}$.
2. Consider the linear model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ with elements $\mathbf{y}^T = (y_1, \dots, y_n)$, $\mathbf{X}$ is a full-rank $n \times k$ matrix. Let $V = C(\mathbf{X})$ and $V_0 = C(\mathbf{j}_n)$ denote the column spaces of $\mathbf{X}$ and $\mathbf{j}_n$, respectively, and let $\hat{\mathbf{y}}^T = (\hat{y}_1, \dots, \hat{y}_n) = \hat{\boldsymbol{\beta}}^T \mathbf{X}^T$, where $\hat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$. In addition, define the following sums of squares:

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2, \quad SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- Specify the form of $\mathbf{P}_V$ and $\mathbf{P}_{V_0}$, the projection matrices onto the spaces V and V_0 , respectively.
 - Show that $(\mathbf{y} - \bar{y})^T \mathbf{j}_n = 0$.
 - Show that $\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$.
 - Show that $SST = SSR + SSE$.
 - 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 $\mathbf{y} = (y_{11}, \dots, y_{3n})^T$ from the linear model:

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

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

- Show that $\text{rk}(\mathbf{X}) = 3$, and determine the general form of all the *estimable* functions: $\boldsymbol{\eta} = \boldsymbol{\lambda}^T \boldsymbol{\beta}$.
- 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.)
- Repeat part (b) for the null hypothesis $H_0 : \mu + \alpha_1 = 2(\mu + \alpha_2) = 3(\mu + \alpha_3)$.
- Specify a matrix $\mathbf{C}$ of full (row) rank 3 such that the null hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta}$ is testable, and construct a test statistic as in the previous two parts.