Applied Statistics Preliminary Examination
Theory of Linear Models
May 2025

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 12 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 $\boldsymbol{A}$.
  - $\text{tr}(\boldsymbol{A})$ 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.
  - For a generic linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where 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:

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

  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.

Problems:

1. Let $\boldsymbol{X}_k$ be a full-rank $n \times k$ matrix whose column space $C(\boldsymbol{X}_k)$ consists of the linear span of the vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$, i.e., $C(\boldsymbol{X}_k) = \mathcal{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$, and let $\boldsymbol{X}_{k+1}$ be a full-rank $n \times (k+1)$ matrix whose column space consists of the linear span of the vectors that constitute $C(\boldsymbol{X}_k)$ and the extra vector $\boldsymbol{x}_{k+1}$, so that $C(\boldsymbol{X}_{k+1}) = \mathcal{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \boldsymbol{x}_{k+1})$. In addition, let $\boldsymbol{G}_k = (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1}$, $\boldsymbol{G}_{k+1} = (\boldsymbol{X}_{k+1}^T \boldsymbol{X}_{k+1})^{-1}$, $\boldsymbol{P}_k = \boldsymbol{X}_k \boldsymbol{G}_k \boldsymbol{X}_k^T$, and $\boldsymbol{P}_{k+1} = \boldsymbol{X}_{k+1} \boldsymbol{G}_{k+1} \boldsymbol{X}_{k+1}^T$.

   (a) Show that $\boldsymbol{P}_k$ is the projection matrix onto $C(\boldsymbol{P}_k)$.

   (b) Show that $C(\boldsymbol{P}_k) = C(\boldsymbol{X}_k)$.

   (c) Show that $\boldsymbol{P}_{k+1} - \boldsymbol{P}_k$ is a projection matrix.

   (d) Is $C(\boldsymbol{P}_k)$ orthogonal to $C(\boldsymbol{P}_{k+1} - \boldsymbol{P}_k)$? Justify.

2. Consider the two full-rank linear regression models for the same response vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ with sample mean $\overline{y}$ and sums of squares about the mean S_{yy} (see these definitions in the "Possibly useful results" of page 1):

$$\boldsymbol{y} = \boldsymbol{X}_k \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

   and

$$\boldsymbol{y} = \boldsymbol{X}_{k+1} \boldsymbol{b} + \boldsymbol{e}, \tag{2}$$

   where the design matrices are as defined in Problem 1, so that $C(\boldsymbol{X}_k) \subset C(\boldsymbol{X}_{k+1})$. As the notation suggests, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ and $\boldsymbol{b} = (b_1, \ldots, b_{k+1})^T$ are the corresponding coefficient vectors of appropriate dimension, with associated LSEs $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}$. The ultimate goal of this problem is to prove that, if R_k^2 and R_{k+1}^2 are the R^2 values for models (1) and (2) respectively, then R_{k+1}^2 cannot be smaller than R_k^2. (To simplify the problem we also assume that the first column of each design matrix consists of a vector of 1's, i.e., $\boldsymbol{x}_1 = \boldsymbol{j}_n$.) In what follows, x_{ij} is the (i,j)-th element of the design matrix $\boldsymbol{X}_k$, and $\overline{x}_j = \sum_{i=1}^n x_{ij}/n$ is the mean of the j-th column (see "Possibly useful results" of page 1).

   (a) Show that $\overline{y} = \sum_{j=1}^k \widehat{\beta}_j \overline{x}_j$.

   (b) If x_{ij} is the (i,j)-th element of $\boldsymbol{X}_k$, and $\overline{x}_j = \sum_{i=1}^n x_{ij}/n$ is the mean of the j-th column, show that $S_{yy} R_k^2 = \sum_{j=1}^k \sum_{i=1}^n \widehat{\beta}_j (x_{ij} - \overline{x}_j) y_i$.

   (c) By considering the difference $S_{yy}(R_{k+1}^2 - R_k^2)$ and noting the results of Problem 1, or otherwise, show that $R_k^2 \leq R_{k+1}^2$.

3. Consider the observations $\{y_{ij}\}$ from the following linear model on 6 distinct subjects:

$$y_{ij} = \alpha_i + (\mu + \gamma_i) x_{ij} + \epsilon_{ij}, \qquad i = 1, 2, \quad j = 1, 2, 3,$$

   where x_{ij} is the age of subject j whose gender is indicated by the index i. Let $\boldsymbol{\beta}^T = (\alpha_1, \alpha_2, \mu, \gamma_1, \gamma_2)$, assume that all subjects have different ages, and that $\{\epsilon_{ij}\} \sim$ IID $\mathrm{N}(0, \sigma^2)$.

   (a) Express the model in the usual vector-matrix form, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, write down the design matrix $\boldsymbol{X}$, and determine its rank.

   (b) Is the hypothesis $H_0 : \alpha_1 = \alpha_2 = \mu = 0$ testable? Carefully justify your answer.

(c) It is desired to test if the model can be reduced to common intercepts ($\alpha_1 = \alpha_2$) and slopes ($\gamma_1 = \gamma_2$) for the two genders. Carefully express this as an appropriate null hypothesis, and show that it is testable.

(d) Propose a test statistic for the test in (c) and specify its distribution under H_0. Give detailed explanations and provide numerical values for all the quantities that can actually be calculated.

Applied Statistics Preliminary Examination
Theory of Linear Models
August 2025

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 $\boldsymbol{A}$.
  - $\text{tr}(\boldsymbol{A})$ 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.
  - Note the inverse for the patterned matrix $\boldsymbol{A}$:

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

  - If $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors in a vector space with inner-product $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$, then the *Cauchy-Schwarz Inequality* states that:

$$|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| \leq \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle \langle \boldsymbol{b}, \boldsymbol{b} \rangle}.$$

Problems

1. Let $\boldsymbol{x} \sim \mathrm{N}_3(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_3)$, with $\boldsymbol{x}^T = (x_1, x_2, x_3)$, $\boldsymbol{\mu}^T = (3, -2, 1)$, and define the following derived random variables and vectors:

$$3y = 2x_1^2 + 2x_2^2 + 2x_3^2 - 2x_1x_2 - 2x_1x_3 - 2x_2x_3, \quad \boldsymbol{w} = \begin{bmatrix} x_1 + x_2 + x_3 \\ x_1 - x_3 \end{bmatrix}, \quad z = x_1 + x_2 + x_3.$$

   (a) Find the distribution of $\boldsymbol{w}$, and hence show that its two components are independent.

   (b) Show that y and z are independent.

   (c) Are y and $\boldsymbol{w}$ independent? Justify.

   (d) Find the distribution of y/σ^2.

2. Consider the linear model in the usual vector-matrix form: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2)$, and $\boldsymbol{X}$ is $n \times p$ of full-rank, i.e., $\mathrm{rk}(\boldsymbol{X}) = p$. Let $\widehat{\boldsymbol{\beta}}$ and s^2 denote respectively, the LSE of $\boldsymbol{\beta}$ and the usual unbiased estimate of σ^2. Letting $\boldsymbol{G} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$, recall the following basic results:

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2\boldsymbol{G}), \qquad \frac{(n-p)s^2}{\sigma^2} \sim \chi^2(n-p), \qquad \widehat{\boldsymbol{\beta}} \text{ is independent of } s^2.$$

   Consider now a single new model point $(\boldsymbol{x}_0, y_0)$, so that $y_0 = \boldsymbol{x}_0^T\boldsymbol{\beta} + \epsilon_0$. Given $\boldsymbol{x}_0$, we know that the best point prediction of $\mathbb{E}(y_0) = \boldsymbol{x}_0^T\boldsymbol{\beta} \equiv \mu_0$ is $\hat{y}_0 = \boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}$. Using the distribution of the standardized $\hat{y}_0$, one can then construct a (pointwise) confidence interval for the regression "line" μ_0. The purpose of this Problem is to construct a confidence band of coverage $(1-\alpha)100\%$ for the entire regression "line" that will hold simultaneously for all design points $\boldsymbol{x}_0$.

   (a) Show that:

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{G}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p\ s^2} \ \sim \ F(p, n-p).$$

   (b) By defining an inner-product for vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R}^p$ as $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{a}^T\boldsymbol{G}\boldsymbol{b}$, and taking $\boldsymbol{a} = \boldsymbol{x}_0$ and $\boldsymbol{b} = \boldsymbol{G}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, show that:

$$(\hat{y}_0 - \mu_0)^2 \leq \left(\boldsymbol{x}_0^T\boldsymbol{G}\boldsymbol{x}_0\right)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{G}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

   (c) Use the above results to derive the fact that a $(1-\alpha)100\%$ simultaneous confidence band for the entire regression "line" is given by

$$\hat{y}_0 \pm w,$$

   and determine the expression for w as a function of $\boldsymbol{x}_0$, $\boldsymbol{G}$, s^2, and an appropriate F distribution quantile.

3. Consider the observations $\boldsymbol{y} = (y_{11}, \ldots, y_{22})^T$ from the linear model:

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

where the $\{\epsilon_{ij}\} \sim$ IID $\mathrm{N}(0, \sigma^2)$, and $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \gamma_1, \gamma_2)^T$ 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) Find the rank of the design matrix $\boldsymbol{X}$, and show that it is rank-defficient (not of full-rank). (Note: if $\mathrm{rk}(\boldsymbol{X}) = k$, and the number of columns in $\boldsymbol{X}$ is p, the rank-defficiency is still defined to be $q = p - k$, even though $n = 4 < 5 = p$ in this overparametrized model.)

(b) Determine the general form of all the *estimable* functions: $\eta = \boldsymbol{\lambda}^T\boldsymbol{\beta}$.

(c) Show (with justification) that the set of *side conditions* $\alpha_1 + \alpha_2 = \gamma_1 - \gamma_2 = 0$ is not a valid and complete set that can be used to remove the overparametrization.

(d) In light of the previous part, propose a valid and complete set of side conditions that will successfully reparametrize the model to full rank.

(e) Write down the new (full-rank) model matrix $\widetilde{\boldsymbol{X}}$ and new parameter vector $\widetilde{\boldsymbol{\beta}}$ that would result from using the valid and complete set of side conditions proposed in the previous part (whence the model can be reformulated as $\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$).