Applied Statistics Preliminary Examination
Theory of Linear Models
May 2015

Instructions:

- Do all 3 Problems. Neither calculators nor electronic devices of any kind are allowed. 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); BLUE (best linear unbiased 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.
  - $\mathbb{E}(X)$ and $\mathbb{V}(X)$: expectation and variance of random variable X.
  - $\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: the m-dimensional random vector $\mathbf{x}$ has a normal distribution with mean $\boldsymbol{\mu}$ 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.
  - If $\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given in partitioned form as

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

with $m_1 = \dim(\mathbf{x}_1)$, then the conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$ is

$$\mathbf{x}_1|\mathbf{x}_2 \sim N_{m_1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2),\ \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right).$$

1. Let $\mathbf{y} = (y_1, y_2, y_3)^T \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

   (a) Find the distribution of $y_1|y_2 = 2$ and $y_3 = 0$, which denotes the conditional distribution of y_1 given $y_2 = 2$ and $y_3 = 0$.

   (b) Find the distribution of $y_1 - y_2 + y_3|y_1 + y_2 - y_3 = 1$.

   (c) Prove or disprove: $y_1 + y_2 + y_3$ and $3y_1^2 + 9y_2^2 + 3y_3^2 - 12y_1y_2 + 10y_1y_3 - 12y_2y_3$ are independent.

   (d) Determine the distribution of $y_1/|y_3 - 1|$.

2. Consider the model $y_i = \beta_0 + i\beta_1 + \varepsilon_i$, $i = 1, \ldots, n$, with $\varepsilon_1, \ldots, \varepsilon_n \sim$ IID $\mathrm{N}(0, \sigma^2)$, $\mathbf{y} = (y_1, \ldots, y_n)^T$, and let $\hat{\boldsymbol{\beta}}$ denote the LSE of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. Define and note the following statistics and basic results:

$$t_1 = \sum_{i=1}^{n} y_i, \qquad t_2 = \sum_{i=1}^{n} iy_i, \qquad u_n = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \qquad v_n = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

   (a) Find, in simplest terms, an expression for $\hat{\boldsymbol{\beta}}$, and show that $\hat{\boldsymbol{\beta}}$ depends on $\mathbf{y}$ only through the (sufficient) statistics t_1 and t_2.

   (b) Compute the distribution of $\hat{\boldsymbol{\beta}}$. Does $\mathbb{V}(\hat{\beta}_0) \to 0$ and $\mathbb{V}(\hat{\beta}_1) \to 0$ as $n \to \infty$?

   (c) Compute a $(1 - \alpha)100\%$ confidence region for $\boldsymbol{\beta}$.

   (d) Construct a size α test of $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

   (e) Consider now the model $y_i = \beta/i + \varepsilon_i$, $i = 1, \ldots, n$, with $\varepsilon_1, \ldots, \varepsilon_n \sim \mathrm{IID}(0, \sigma^2)$, not necessarily normal. If $\hat{\beta}$ is the LSE of β, does it now follow that $\mathbb{V}(\hat{\beta}) \to 0$ as $n \to \infty$?

3. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, \sigma^2\mathbf{I}_4)$, with

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ -1 & 1 & 0 \\ -1 & -1 & -1 \end{pmatrix}, \qquad \text{and we define} \qquad \mathbf{G} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

   (a) What is the rank of $\mathbf{X}$?

   (b) Verify that $\mathbf{G}$ is a generalized inverse of $\mathbf{X}^T\mathbf{X}$.

   (c) Determine which of the following functions of $\boldsymbol{\beta}$ are *estimable*:

$$\eta_1 = \beta_1, \qquad \eta_2 = \beta_1 - \beta_2, \qquad \eta_3 = \beta_1 + \frac{1}{2}\beta_3.$$

   (d) Find BLUEs for the estimable functions in (c), and compute their corresponding distributions.

   (e) Justify your reasoning in determining if the following hypothesis is *testable*,

$$H_0 : \begin{Bmatrix} 4\beta_1 + 2\beta_3 = 0 \\ \beta_2 = \beta_1 \end{Bmatrix}.$$

   If it is testable then describe a suitable test of this hypothesis, and state the distribution of the test statistic under H_0 as well as under the alternative hypothesis,

$$H_1 : \begin{Bmatrix} 4\beta_1 + 2\beta_3 = 1 \\ \beta_2 = \beta_1 + 1 \end{Bmatrix}.$$

Design of Experiments: Prelim Problems
May 2015

Please Do All Problems

For each test, state the null and alternative hypotheses in terms of the model parameters

Note: Tukey's Studentized range distribution: If $\bar{Y}_1, \cdots, \bar{Y}_n$ are independent random variables with $N(\mu, \sigma^2)$ distribution then for $\hat{\sigma}$ being an unbiased estimator of σ, the statistic $(\max_i \bar{Y}_i - \min_i \bar{Y}_i)/(\hat{\sigma}/\sqrt{n})$ is said to have the Studentized range distribution.

1. Consider a balanced two-stage nested design with fixed factor A at a levels and random factor B (nested within factor A) at b levels. Assume there are n replications at each combination of the levels of A and B.

   (a) Write a statistical model for this design. Clearly define all the terms and state all relevant assumptions. (10 points)

   (b) Construct the ANOVA table for the model in part (a). Include the expressions for the sums of squares as well as the expected mean squares. (10 points)

   (c) Formulate Tukey's studentized range test for the levels of factor A. That is write the expression for the $100(1-\alpha)\%$ simultaneous confidence intervals for all pairwise differences of the means of the levels of A. (10 points)

   (d) Let Y_{ijk} denote the k-th replication from the i-th level of factor A and j-th level of factor B. Obtain an expression for the correlation coefficient between Y_{ijk} and $Y_{ijk'}$, where $k \neq k'$. (10 points)

   (e) Estimate the correlation coefficient in part (d). (10 points)

   (f) Explain how you would randomize this design. (10 points)

2. A management information system consultant conducted a small-scale study of five different daily summary reports (A: greatest amount of detail; B; C; D; E: least amount of detail). She used five sales executives in the study. Each was given one type of daily report for a month and then was asked to rate its helpfulness on a 25-point scale(0: no help; 25: extremely helpful). Over a five-month period, each executive received each type of report for one month according to the Latin square design shown below. The helpfulness ratings and summary statistics follow.

	Month					
Executive	March	April	May	June	July	Mean
Harrison	21 (D)	8 (A)	17 (C)	9 (B)	16 (E)	14.20
Smith	5 (A)	10 (E)	3 (B)	12 (C)	15 (D)	9.00
Carmichael	20 (C)	10 (B)	15 (E)	22 (D)	12 (A)	15.80
Loeb	4 (B)	17 (D)	3 (A)	9 (E)	10(C)	8.60
Munch	17 (E)	16 (C)	20 (D)	7 (A)	11 (B)	14.2
Mean	13.40	12.20	11.60	11.80	12.80	12.36
Report	A	B	C	D	E	
Mean	7.00	7.40	15.00	19.00	13.4	

$$\sum_i \sum_j \sum_k y_{ijk}^2 = 4596.00$$

(a) Write a linear model for this experiment. Clearly define all the terms and state all relevant assumptions. (10 points)

(b) Calculate the sum of squares. (20 points)

(c) Write the ANOVA table for this experiment and include the corresponding expected mean squares. (10 points)

(d) Test whether or not the five types of reports differ in mean helpfulness at the 0.01 level of significance. (10 points)

(e) Calculate the residuals for the month of March. (10 points)

(f) If you were to conduct pairwise comparisons based on Tukey's Studentized range statistic, what would the minimum significant difference value be for the 0.05 level of significance? (10 points)