

Applied Statistics Preliminary Examination
Theory of Linear Models
May 2016

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 15 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.
 - $\mathbb{E}(X)$ 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 matrix \mathbf{A} is given in block form as

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{pmatrix},$$

then $|\mathbf{A}| = |\mathbf{B}| \cdot |\mathbf{C}|$.

1. Let $\mathbf{y} = (y_1, y_2, y_3, y_4)^T \sim N_4(\mu\mathbf{j}_4, \boldsymbol{\Sigma})$, where $\mathbf{j}_4 = (1, 1, 1, 1)^T$ and $\boldsymbol{\Sigma}$ is given in block form as:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Theta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Theta} \end{pmatrix}, \quad \text{with} \quad \boldsymbol{\Theta} = \begin{pmatrix} \theta_0 & \theta_1 \\ \theta_1 & \theta_0 \end{pmatrix}.$$

In addition, define $U = \sum_{i=1}^4 y_i$, $V = (y_1 - y_2)^2 + (y_3 - y_4)^2$, and $W = [(y_1 + y_2) - (y_3 + y_4)]^2$.

- For what values of θ_0 and θ_1 is $\boldsymbol{\Sigma}$ a valid (positive definite) covariance matrix?
 - Find the distributions of U , V , and W .
 - Are U , V , and W *pairwise* independent? Justify your arguments.
 - Compute $\mathbb{E}(U)$, $\mathbb{E}(V)$, $\mathbb{E}(W)$, and thus produce unbiased estimators for μ , θ_0 , and θ_1 that are functions of (U, V, W) .
 - Does there exist a constant c such that cW/V has an F distribution? Justify, and if so, find c .
2. Consider the linear model in centered form, $y_i = \alpha + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_k(x_{ik} - \bar{x}_k) + \epsilon_i$, for $i = 1, \dots, n$, where $n \geq 2$, and $\bar{x}_j = n^{-1} \sum_{i=1}^n x_{ij}$ denotes the sample mean of the j -th explanatory variable. Let $\mathbf{y} = (y_1, \dots, y_n)^T$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^T$, and suppose that the error vector $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ is correlated, so that the model can be written in matrix form as:

$$\mathbf{y} = (\mathbf{j}_n \quad \mathbf{X}_c) \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{V}), \quad \text{with} \quad \mathbf{V} = [(1 - \rho)\mathbf{I}_n + \rho\mathbf{J}_{n,n}],$$

where ρ is a known constant of appropriate value such that \mathbf{V} is positive definite. Note: $\mathbf{X}_c = (\mathbf{I}_n - n^{-1}\mathbf{J}_{n,n})\mathbf{X}_1$, where \mathbf{X}_1 is the matrix whose (i, j) -th entry is x_{ij} .

- By noting that one must have $\mathbb{V}(\sum_{i=1}^n y_i) > 0$, find a lower bound for the allowable value of ρ .
 - Provided $\rho \neq 1$ and $\rho \neq -1/(n-1)$, it can be shown that the inverse of \mathbf{V} exists and has the form, $\mathbf{V}^{-1} = a(\mathbf{I}_n - b\rho\mathbf{J}_{n,n})$, for appropriate constants a and b . Find a and b , and hence \mathbf{V}^{-1} .
 - If $\mathbf{X} = (\mathbf{j}_n \quad \mathbf{X}_c)$, show that:

$$\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} = \begin{pmatrix} nb & \mathbf{0} \\ \mathbf{0} & a\mathbf{X}_c^T \mathbf{X}_c \end{pmatrix},$$
 where a and b are the constants computed in (b).
 - Find the *generalized least squares* (GLS) estimators of α and $\boldsymbol{\beta}$.
 - Find the *ordinary least squares* (OLS) estimators of α and $\boldsymbol{\beta}$ (corresponding to the assumption that $\rho = 0$), and compare with the GLS estimators from (d).
3. Consider the two-way anova model $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$, for $i = 1, 2, 3$ and $j = 1, 2$, with $\epsilon_{ij} \sim \text{IID } N(0, \sigma^2)$ for all i and j . The model can be written in vector and matrix form as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbf{y} = (y_{11}, y_{12}, \dots, y_{31}, y_{32})^T, \quad \boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2)^T.$$

Assume that a symmetric generalized inverse \mathbf{G} of $\mathbf{X}^T \mathbf{X}$ is available, whose (i, j) -th element is g_{ij} .

- Give the form of \mathbf{X} , and compute its rank.
- Determine which of the following functions of $\boldsymbol{\beta}$ are *estimable*:

$$\eta_1 = \mu, \quad \eta_2 = \mu + \alpha_1, \quad \eta_3 = \mu + \alpha_1 + \beta_1, \quad \eta_4 = \alpha_1 - \alpha_2.$$

- Find BLUEs for the estimable functions in (b), and compute their corresponding distributions.
- Show that the hypothesis $H_0 : \alpha_1 = \alpha_2 = \alpha_3$ is *testable*, and give a test for it. State the distribution of the test statistic under H_0 .
- Find conditions on the constants $\{c_1, c_2, c_3\}$ such that the hypothesis $H_0 : c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 = 0$ is *testable*. Give a suitable test, and find the distribution of the test statistic under H_0 .

Design of Experiments: Prelim Problems
May 2016

Please Do All Problems

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

Note 1: Tukey's Studentized range distribution: If $\bar{Y}_1, \dots, \bar{Y}_n$ are independent random variables with $N(\mu, \sigma^2/n)$ distribution then for $\hat{\sigma}^2$ being an unbiased estimator of σ^2 , the statistic $(\max_i \bar{Y}_i - \min_i \bar{Y}_i)/(\hat{\sigma}/\sqrt{n})$ is said to have Studentized range distribution.

Note 2: Satterthwaite approximation of degrees of freedom for a "mean square". If $\hat{\sigma}_0^2 = \sum_i c_i MS_i/k$, then the approximate degrees of freedom is given by

$$r = \frac{(\sum_i c_i MS_i)^2}{[\sum_i c_i^2 MS_i^2 / f_i]}$$

where c_i 's are constants and f_i is degrees of freedom of MS_i

1. A medical study seeks to understand the effects of a new prescription drug and a nutritional supplement for treating a rare disorder based on a single response variable. The study uses four treatment conditions: (1) placebo, (2) drug only, (3) supplement only, and (4) both drug and supplement. Human patients enter the study at different times. Once 4 patients have entered the study, the 4 treatments are randomly assigned to these patients (each patient receiving a different treatment). This process continues for each new block of 4 consecutive patients, until a total of 24 patients have entered the study. Let Y_{ij} denote the response of patient from j -th block who is assigned treatment condition i , for $i = 1, \dots, 4$, and $j = 1, \dots, 6$

Suppose the following summary data is available.

$$y_{1.} = 120, \quad y_{2.} = 90, \quad y_{3.} = 120, \quad y_{4.} = 30,$$

$$\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 = 79, \quad \sum_i \sum_j y_{ij}^2 = 6931$$

- Write a linear model equation appropriate for analyzing the responses Y_{ij} from this experiment, clearly state the assumptions and conditions for this model. (10 points)
- Write the ANOVA table appropriate for your model including E(MS). (20 points)
- Test whether there are any differences between the treatment conditions (Use $\alpha = 0.05$). (10 points)
- Test whether the effects of the drug and supplement on the responses Y_{ij} could be additive, i.e. whether or not there appears to be interaction between the effects of the drug and supplement (Use $\alpha = 0.05$). (10 points)
- Define two contrasts Γ_1 that compares represents the effect of drug and Γ_2 that represents the effect of supplements. Find 95% simultaneous confidence intervals for these contrasts. (20 points)
- Can you assert that the response mean is significantly smaller when the drug is given without the supplement versus when only a placebo is given? Test at level $\alpha = 0.05$. (10 points)

2. Consider the model

$$Y_{ijkl} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk} + \varepsilon_{ijkl}; \quad i = 1, \dots, a; \quad j = 1, \dots, b; \quad k = 1, \dots, c; \quad l = 1, \dots, n$$

$$\beta_{ij} \stackrel{iid}{\sim} N(0, \sigma_\beta^2), \quad \gamma_{ijk} \stackrel{iid}{\sim} N(0, \sigma_\gamma^2), \quad \varepsilon_{ijkl} \stackrel{iid}{\sim} N(0, \sigma^2); \text{ all mutually independent}$$

where $\alpha_1, \dots, \alpha_a$ are fixed values summing to zero, $\sigma^2 > 0$, and $a \geq 2$, $b \geq 2$, $c \geq 2$, $n \geq 2$.

- (a) Find the correlation between Y_{2121} and Y_{2122} . (10 points)
- (b) In terms of the observations Y_{ijkl} , write the expressions for mean squares corresponding to the terms β_{ij} , γ_{ijk} , and ε_{ijkl} . (You may use the dot-and-bar notation.) (10 points)
- (c) Give the expected values of the mean squares from the previous part. (10 points)
- (d) Write the expressions for the unbiased ANOVA (method of moment) estimators of σ_β^2 , σ_γ^2 , and σ^2 . (10 points)
- (e) Construct a test statistic for $H_0 : \sigma_\beta^2 = 0$ versus $H_1 : \sigma_\beta^2 > 0$, and state the condition under which H_0 is rejected. (10 points)
- (f) Form an exact 95% two-sided confidence interval for $\alpha_1 - \alpha_2$. (10 points)
- (g) Form an exact 95% two-sided confidence interval for σ^2 . (10 points)