Applied Statistics Preliminary Examination

Theory of Linear Models

May 2019

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 $|\mathbf{A}|$: denote the trace and determinant, respectively, of matrix $\mathbf{A}$.
  - $\mathbf{I}_n$: the $n \times n$ identity matrix.
  - $\mathbf{j}_n = (1, \cdots, 1)^T$ is an n-vector of ones, and $\mathbf{J}_{m,n}$ is an $m \times n$ matrix of ones.
  - $\mathbb{E}[X]$ and $\text{var}(X)$: 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: $X \sim F(n_1, n_2)$.

1. Suppose that $Z_1, Z_2, ..., Z_n$ are iid standard normal random variables.

   (a) Derive the joint distribution of $\overline{Z}, Z_1 - \overline{Z}, Z_2 - \overline{Z}, \ldots, Z_n - \overline{Z}$.

   (b) Based on part (a), deduce that $\overline{Z}$ and $\sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2$ are independent.

   (c) Let $\mathbf{Y}$ have an n-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$, where $\text{var}\,(Y_i) = \sigma^2$, for all i, and $\text{cov}\,(Y_i, Y_j) = \sigma^2(1-\rho)$, for $i \neq j$, where $0 < \rho < 1$. Prove that $\overline{Y}$ and $\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$ are independent. This is a generalization of part (b).

2. Consider the linear model defined by $Y_1 = 2\theta + \epsilon_1$, $Y_2 = \theta + \epsilon_2$, where $\epsilon_1 = 2Z_1 - Z_2$ and $\epsilon_2 = Z_1 + 2Z_2$, and Z_1 and Z_2 are independent random variables with zero mean and constant variance σ^2.

   (a) Write this model in $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ form. Find $E(\mathbf{Y})$ and $\text{cov}(\mathbf{Y})$.

   (b) Compute the ordinary least squares (OLS) estimator of θ.

   (c) Compute the generalized least squares (GLS) estimator of θ.

   (d) Show that the OLS and GLS estimators are both unbiased.

   (e) Compute the variance of both estimators and compare.

3. Let $\mathbf{Y}$ be a 6×1 vector,

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & -1 \\ 1 & 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 & -1 \\ 1 & 1 & -1 & 0 & 0 \end{pmatrix}, \tag{1}$$

   and $\boldsymbol{\beta} = (\mu, \theta_1, \theta_2, \theta_3, \theta_4)'$. Assume the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ holds, where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$. Let $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, and $\mathbf{x}_4$ denote the columns of $\mathbf{X}$. Note that $\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 + \mathbf{x}_4 = \mathbf{0}$.

   (a) Find the components of $E(\mathbf{Y})$ in terms of $\mu, \theta_1, \theta_2, \theta_3$, and θ_4.

   (b) Show $\mathbf{x}_0, \mathbf{x}_2, \mathbf{x}_3$, and $\mathbf{x}_4$ are linearly independent.

   (c) Let $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4)'$. Give conditions on $\boldsymbol{\lambda}$, of the form $\boldsymbol{\lambda}'\mathbf{c}_i = 0$, $i = 1, 2, \cdots, s$, that are necessary and sufficient for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable. What is the value of $s = p - r$?

   (d) Show that $\mu + \theta_3 - \theta_4$ is estimable.

(e) Give a nonestimable function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Explain your answer.

(f) Explain briefly how you could use your answer in part (e) to *force* a particular solution to the normal equations $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$. If you did this, would your nonestimable $\boldsymbol{\lambda}'\boldsymbol{\beta}$ in part (e) become estimable? Explain.

Design of Experiment: Prelim Problems

May 2019

Please Do All Problems. Each of the 17 parts carries an equal weight of 10 points.

1. A manufacturing firm investigated the breaking strengths of components made from raw materials purchased from 4 suppliers (A, B, C, D). Data was collected from 2 replicates of a 4×4 Latin square design. The blocking factors were days and operators. Four operators were randomly selected, and were used in both replicates. The two replicates were run over 8 days with the first 4 days assigned to replicate 1 and the second four days assigned to replicate 2.

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

   b) Fill the ANOVA table (Table 1) for this experiment.

Table 1: ANOVA table for Question 1b)

Source	DF	Type III MS
rep	_____	95
operator	_____	799
_____		95575
supplier	_____	32245
error	_____	38807

   c) Test whether or not components made from raw materials purchased from 4 suppliers differ in mean breaking strength at the 0.01 level of significance. Clearly state your null and alternative hypotheses.

   d) Evaluate whether including each of the blocking factors in your model is helpful in testing supplier effect.

   e) If you were to conduct pairwise comparisons for suppliers, what would the minimum significant difference value be with overall 95% confidence level?

   f) Revise your model if it turns out different four operators were used in the two replicates.

2. In an trout experiment reported by Gutsell (Biometrics, 1951), the red blood cell counts in the blood of brown trout were measured. Fish were put at random into eight troughs of water. Two troughs were assigned to each of the four levels of the treatment factor sulfamerazine (0, 5, 10, 15 grams per 100 pounds of fish added to the diet per day). After 42 days, five fish were selected at random from each trough and the red blood cell count from the blood of each fish was measured in two different counting chambers, giving two measurements per fish. The observations reported, when multiplied by 5000, give the number of red blood cells per cubic millimeter of blood.

   a) Write a linear model for this experiment. Clearly define all the terms and state all relevant assumptions. Please use A, B, C to denote sulfamerazine in the diet, trough, and fish factors respectively.

   b) The response variable is counts. Do you have any concerns about the model assumption? How would you check them?

   c) Write out the degrees of freedom and the expected mean squares for each term in the model.(Table 2) Note that you may need to revise the sources of variation in Table 2 if nesting structure exists in your model.

Table 2: ANOVA table for Question 2c)

Source	DF	Mean Square	E(MS)
sulfamerazine	_____	25889	_____
trough	_____	4239	_____
fish	_____	2446	_____
error	_____	351	_____

   d) Test the hypothesis that sulfamerazine has no effect on the red blood cell counts at 0.05 level. Examine the linear and quadratic trends at 95% overall confidence level using the information in Table 3.

Table 3: Data summary table for Question 2d)

Sulfamerazine	Average Response
0	201.9
10	288.2
15	253.05

   e) Give a 95% confidence interval for fish variance component.

   f) Test the hypothesis that trough has no impact on the red blood cell counts at 0.05 significant level. State your hypothesis clearly.

3. An experiment is designed to study pigment dispersion in paint. Four mixes of a particular pigment are studied. Every day, the procedure consists of firstly preparing a particular mix and applying that mix to a panel by three application methods (brushing, spraying, and rolling) in a random order. Then repeat the same procedure for the other three mixes. The response measured is the percentage reflectance of pigment. Three days are required to run the experiment. Assuming that the same four mixes are used on all the three days. Mix and method both have fixed effects. Suppose Y_{hij} is percentage reflectance of pigment for applying jth method to ith mix on hth day.

   A student suggests to use a three-way randomized complete block design to analyze this data as follows.

$$Y_{hij} = \mu + \theta_h + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{hij}$$
$$h = 1, 2, 3. \quad i = 1, 2, 3, 4. \quad j = 1, 2, 3.$$

   θ_h is the effect of the hth day. α_i is the effect of the ith mix level. β_j is the effect of the jth method level. $(\alpha\beta)_{ij}$ is the interaction effects between ith mix and jth method. Part of the student's ANOVA table is shown in Table 4.

Table 4: ANOVA table using the student's model for Question 3

Source	Sum of Square
Day	2.042
Mix	307.479
Method	222.095
Day × Mix	4.527
Mix × Method	10.036
error	10.749

   a) Please evaluate the model used by the student. Assume no other interaction effects are of interest. If you can give a more appropriate model, please write the linear model. Clearly define all the terms and state all relevant assumptions. Then use the appropriate model to answer following questions.

   b) Write out the degrees of freedom and the expected mean squares for each term in the appropriate model.

   c) What is the variance of $\overline{Y_{..1}} - \overline{Y_{..2}}$ in the appropriate model?

   d) Test the hypothesis that different Mixes have the same impact on the response variable at 0.05 significant level. State your hypothesis clearly.

   e) Show E(Mean Square for Mix factor) has the formula as you wrote in 3b) based on the appropriate model.