Applied Statistics Preliminary Examination
Theory of Linear Models
May 2022

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 $\boldsymbol{A}$.
  - $\mathrm{tr}(\boldsymbol{A})$ 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.
  - If $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given in partitioned form as

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

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

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

1. Let the n-dimensional vector $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$, where $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{b}$ for an $(n \times k)$ full-rank matrix of constants $\boldsymbol{X}$ with $k < n$. Define the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, and the quadratic forms $Q_1 = \boldsymbol{y}^T\boldsymbol{H}\boldsymbol{y}$ and $Q_2 = \boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y}$.

   (a) Show that both of the matrices $\boldsymbol{H}$ and $\boldsymbol{I}_n - \boldsymbol{H}$ are symmetric and idempotent.

   (b) Find all the eigenvalues and ranks of each of $\boldsymbol{H}$ and $\boldsymbol{I}_n - \boldsymbol{H}$.

   (c) Find a constant c such that cQ_1 and cQ_2 have familiar distributions, and specify what these resulting distributions are.

   (d) Find the distribution of
   $$W = \frac{(n-k)Q_1}{kQ_2}.$$

2. Consider fitting the linear model where the full-rank $(n \times k)$ model matrix is decomposed as $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, with $\boldsymbol{X}_1$ $(n \times k_1)$, $\boldsymbol{X}_2$ $(n \times k_2)$, $k = k_1 + k_2$, $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$, and the normal spherical errors assumption, $\boldsymbol{\epsilon} \sim N(0, \sigma^2\boldsymbol{I}_n)$:
   $$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}.$$
   Suppose however that the vector of observations $\boldsymbol{y}$ comes from the reduced model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$, and it is known that the column spaces of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal, i.e., $\mathcal{C}(\boldsymbol{X}_1) \perp \mathcal{C}(\boldsymbol{X}_2)$. Thus, by fitting the model with model matrix $\boldsymbol{X}$ we are in fact *overfitting*. Let $\widehat{\boldsymbol{\beta}}^T = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)$, denote the resulting LSE of $\boldsymbol{\beta}$, and $s^2 = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/(n-k)$ the usual estimator of σ^2.

   (a) Find $\mathbb{E}(\widehat{\boldsymbol{\beta}}_1)$ and $\mathbb{E}(\widehat{\boldsymbol{\beta}}_2)$. Is $\widehat{\boldsymbol{\beta}}_1$ unbiased for $\boldsymbol{\beta}_1$?

   (b) Compute $\mathbb{V}(\widehat{\boldsymbol{\beta}}_1)$, $\mathbb{V}(\widehat{\boldsymbol{\beta}}_2)$, and $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$.

   (c) If $\mathcal{P}_{\mathcal{C}(\boldsymbol{X})}$, $\mathcal{P}_{\mathcal{C}(\boldsymbol{X}_1)}$, and $\mathcal{P}_{\mathcal{C}(\boldsymbol{X}_2)}$ denote the respective projection matrices onto the column spaces $\mathcal{C}(\boldsymbol{X})$, $\mathcal{C}(\boldsymbol{X}_1)$, and $\mathcal{C}(\boldsymbol{X}_2)$, show that $\mathcal{P}_{\mathcal{C}(\boldsymbol{X})} = \mathcal{P}_{\mathcal{C}(\boldsymbol{X}_1)} + \mathcal{P}_{\mathcal{C}(\boldsymbol{X}_2)}$.

   (d) Show that even though the model is overfit, s^2 is still unbiased for σ^2.

3. Consider the linear model $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \epsilon_i$, where $i = 1, \ldots, n$, $\{\epsilon_i\} \sim$ IID $\mathrm{N}(0, \sigma^2)$, and the $x_{i,j}$ are elements of the $(n$ by $p+1)$ design matrix $\boldsymbol{X}$ (of rank $k < p+1 < n$). Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\overline{y} = \boldsymbol{j}_n^T\boldsymbol{y}/n$ be the sample mean of the y_i, and let $\widehat{\boldsymbol{\beta}} = \boldsymbol{G}\boldsymbol{X}^T\boldsymbol{y}$ denote the LSE of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ corresponding to the generalized inverse $\boldsymbol{G}$ of $\boldsymbol{X}^T\boldsymbol{X}$. Additionally, let $R^2 = (\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} - n\overline{y}^2)/(\boldsymbol{y}^T\boldsymbol{y} - n\overline{y}^2)$ be the usual *coefficient of determination*, and define the quadratic forms:
   $$Q_1 = (\overline{y}\boldsymbol{j}_n - \boldsymbol{y})^T(\overline{y}\boldsymbol{j}_n - \boldsymbol{y}), \qquad Q_2 = (\overline{y}\boldsymbol{j}_n - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\overline{y}\boldsymbol{j}_n - \boldsymbol{X}\widehat{\boldsymbol{\beta}}), \qquad Q_3 = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

   (a) Without using calculus, prove that $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$. (Your argument should use the form of $\widehat{\boldsymbol{\beta}}$ and derive any properties of generalized inverses that are deemed necessary in the proof.)

   (b) Prove that $Q_1 = Q_2 + Q_3$.

   (c) If $\mathcal{C}(\boldsymbol{X})$ denotes the column space of $\boldsymbol{X}$, prove or disprove the statement: $R^2 = 1$ if and only if $\boldsymbol{y} \in \mathcal{C}(\boldsymbol{X})$.

   (d) Assuming that $\eta = 3\beta_1 - \beta_0$ is *estimable*, show how to construct a test of $H_0 : \eta = 2$ vs. $H_1 : \eta \neq 2$. Clearly define the test statistic, its distribution under both H_0 and H_1, and the rejection rule.

Design of Experiment: Prelim Problems

May 2022

Please Do All Problems. Each of the 16 parts carries an equal weight of 10 points.

1. In an experiment on the preparation of chocolate cakes, 3 recipes (factor A) for the batter were compared. Recipes I and II differed in that the chocolate was added at 40 degrees Celsius and 60 degrees Celsius, respectively, while recipe III contained extra sugar. Six different baking temperatures (factor B), ranging from 175 to 225 degrees Celsius were tested with each recipe. Assume factor A has fixed effect and factor B has random effect. Interaction effects between factors A and B are the only interaction effects. The moisture level of each cake is measured and used as response variable. Please write down the statistical models and assumptions for each of the following designs.

   a) Suppose for each of the 18 treatment combinations, three cakes are made.

   b) Suppose 2 bags of flour (factor D with random effect) were used. Each bag of flour is used to make two cakes for each of the 18 treatment combinations.

   c) Suppose 3 bags of flour (factor D with random effect) are randomly assigned to each of 18 treatment combinations. In total, 54 bags of flour are used. Two cakes are made from each bag.

   d) There were 10 replications (factor C with random effect) conducted over time. Each time that a mix was made by any recipe, enough batter was prepared for 6 cakes, each of which was baked at a different temperature.

2. In a study, the research interest is the total concentration of a specific alkaloid for Herb A as measured by infusion extraction (steeping the tea bag). The process of herb drying, mixing, and bagging was found to be highly reliable (accurate and precise), but the method of growing and harvesting Herb A was found to affect alkaloid concentration. The focus of this problem is to analyze the experimental growing data for Herb A to assess the resulting alkaloid concentration measured in controlled lab conditions.

   Three varieties of Herb A (A1, A2, and A3), were grown at research greenhouses with either 0, 23, 45, or 68 g/m^2 of fertilizer (primarily nitrogen supplied as soy bean meal). Eighteen temporary fields were created by placing standard top soil in previously unplanted areas, with small experimental green houses placed over each one. Each of the 3 varieties were randomly assigned to 6 of the 18 green houses, so that each green house had only one variety (to protect from confusion at harvest). The four levels of fertilizer were randomly assigned to equal-sized quarters of each greenhouse. At harvest time, 3 leaves were collected from each of several center plants in each quarter and combined as the sample. Each sample was uniformly processed (dried, crushed, and blended), then portions shipped to three labs for alkaloid concentration measurements. We will analyze the data from one of the labs. Of interest are differences in variety and fertilizer levels. Assume there are no interaction effects.

   Denote y_{iuj} as the alkaloid concentration measurement for the sample from i^{th} Herb A variety (factor A, assumed with fixed effect), u^{th} assigned field, and j^{th} fertilizer level (factor B, assumed with random effect). We have some summary information about the data as follows: $\sum_j(\bar{y}_{..j} - \bar{y}_{...})^2 = 0.119$; $\bar{y}_{1..} = 1.667$; $\bar{y}_{2..} = 1.725$; $\bar{y}_{3..} = 1.588$.

   a) Please write down a model for this experiment. Clearly define all the terms and state all relevant assumptions.

   b) Fill in the ANOVA table below.

Source	d.f.	MS	E(MS)
variety			
Error(w)		0.128	
Total(w)		0.126	
fertilizer			
Error(S)		0.0173	
Total(S)			

   c) State and conduct statistical tests for Variety effect at 0.05 significance level, and interpret the test results.

   d) State and conduct statistical tests for Fertilizer effect at 0.05 significance level, and interpret the test results.

   e) Calculate E(Mean Square for Variety) and E(MSE(W)).

f) Use Tukey's method to construct simultaneous 95% confidence intervals for the pairwise effect difference of the three varieties.

g) Between Variety effect and Fertilizer effect, which one is more accurately estimated? What adjustment do you want to make about the experimental design, if we want to estimate the other effect more accurately?

3. Suppose that a furniture manufacturer has developed a new method for assembling a particular type of rocking chair at its assembly plant and wishes to determine if the new method (method 2) produces a smaller mean time to assemble the rocking chair than does the current method (method 1). Because industrial research shows that different training techniques are sometimes needed to effectively train different age groups to use a new assembly method, the plant supervisor also wishes to assess whether the plant's training techniques for the new assembly method will effectively train the different age groups of workers in the plant. To this end, the plant supervisor defines three age groups for the plant's workers, where group 1 represents younger workers, group 2 represents intermediate age workers, and group 3 represents older workers. Three workers are randomly selected from the workers in each age group and are trained to use the new assembly method, where a training technique that is thought to be appropriate for a particular age group is used to train each randomly selected worker in that age group. Each trainee then assembles two rocking chairs using method 1 (the current method) and two rocking chairs using method 2 (the new method). (The order in which the methods are used is randomly selected for each trainee.)

   y_{ijkl} represents the l th assembly time used by trainee k nested inside age group j and using assembly method i to assemble the rocking chair. Assume assembly method (factor A) and age group (factor B) both have fixed effect. Trainee (factor C) has random effect. Assume there are interaction effects between assembly method and age group. We know that $\bar{y}_{1\ldots} - \bar{y}_{2\ldots} = 0.85$.

   a) Write a linear model equation appropriate for analyzing the responses y_{ijkl} from this experiment, clearly state the assumptions and conditions for this model.

   b) Fill in the ANOVA table below.

Source	d.f.	MS	E(MS)
A			
B		8.03	
		6.543	
AB		0.59	
Error		2.31	
Total			

   c) Test if the new method (method 2) produces a smaller mean time to assemble the rocking chair than does the current method (method 1) at 0.05 significance level.

   d) State and conduct statistical test for age group effects at 0.05 significance level, and interpret the test results.

   e) State and conduct statistical test for whether the variance for trainee effect is greater than the variance of random error.