Applied Statistics Preliminary Examination
Theory of Linear Models
January 2021

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.
  - $\mathbb{E}(X)$ and $\mathbb{V}(X)$: expectation and variance of random variable 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 $\boldsymbol{y} = (y_1, y_2, y_3)^T \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $y_1 \sim N(0, 2)$, $y_2 \sim N(1, 6)$, $y_3 \sim N(2, 2)$, and where the covariance matrix $\boldsymbol{\Sigma}$, whose elements are all non-negative, is specified in terms of the following marginal conditional variances:

$$\mathbb{V}(y_2 \mid y_1) = \frac{11}{2}, \qquad \mathbb{V}(y_3 \mid y_2) = \frac{1}{2}, \qquad \mathbb{V}(y_1 \mid y_3) = 2.$$

In addition, let $x_1 = 2y_1 - y_2$, $x_2 = ay_2 + by_3$, for some constants a and b, and let $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{\mu}$ denote the centered $\boldsymbol{y}$. Further define the following quadratic form in the elements of $\boldsymbol{z} = (z_1, z_2, z_3)^T$:

$$Q = (3z_1^2 - 4z_1z_2 + 6z_1z_3 + 4z_2^2 - 12z_2z_3 + 11z_3^2)/4.$$

(a) Find the covariance matrix $\boldsymbol{\Sigma}$.

(b) Find the joint distribution of x_1 and x_2, and hence determine (with justification) whether it is possible to choose $a \neq 0$ and $b \neq 0$ such that x_1 and x_2 are independent.

(c) Compute the value of the partial correlation coefficient $\rho_{12|3}$ for the random vector $\boldsymbol{y}$, i.e., the correlation coefficient between y_1 and y_2, given y_3.

(d) Show that it is impossible for $\boldsymbol{Az}$ to be independent of Q for any non-zero matrix $\boldsymbol{A}$. Hence, or otherwise, find the distribution of Q.

2. Consider the vector of observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ from the linear model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{V}),$$

where $\boldsymbol{X}$ is an $n \times p$ matrix of (full) rank p, and $\boldsymbol{V}$ is a known positive definite symmetric matrix. Let $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ be the usual OLS estimator of $\boldsymbol{\beta}$.

(a) Find the bias and variance of $\hat{\boldsymbol{\beta}}_{OLS}$. If the OLS is biased, propose an unbiased estimator of $\boldsymbol{\beta}$ and find its variance.

(b) Find the BLUE of $\boldsymbol{\beta}$, and prove that the *best linear unbiased* property follows from an appropriate application of the Gauss-Markov Theorem. Also find the variance of your BLUE.

(c) Find the maximum likelihood estimators (MLEs) of $\boldsymbol{\beta}$ and σ^2. (It is not enough to state the MLEs, you need to derive them starting from the likelihood function.)

(d) Find the bias of each of your MLEs in (c). For each estimator that is biased, propose a corresponding unbiased estimator.

3. Consider the vector of 4 observations $\boldsymbol{y} = (y_1, \ldots, y_4)^T$ from the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, given in expanded form as:

$$\boldsymbol{y} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 3 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_4).$$

The ultimate goal here is to make inference on the function $\eta = (1, 1, 1)^T \boldsymbol{\beta} = \beta_1 + \beta_2 + \beta_3$.

(a) Characterize all the *estimable* functions of the form $\boldsymbol{\lambda}^T \boldsymbol{\beta}$, and thus show that η is estimable.

(b) Explicitly show how to *reparametrize* the model to full rank. That is, construct a matrix $\boldsymbol{U}$ such that $\boldsymbol{\gamma} = \boldsymbol{U}\boldsymbol{\beta}$ is the new parameter vector in the full rank linear model $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, and find the new (full rank) design matrix $\boldsymbol{Z}$.

(c) Sketch out the general approach to imposing *side conditions* in order to render the model to full rank. Apply the approach to this problem, and thus show how to obtain the corresponding OLS estimator of $\boldsymbol{\beta}$. (Note: it is sufficient to give the form of the OLS as a function of matrices and $\boldsymbol{y}$, but the matrices should be completely specified.)

(d) Taking any of the above approaches, or otherwise, construct a $(1 - \alpha)100\%$ confidence interval for η. (Note: once again it is sufficient to give the form of the interval as a function of matrices and $\boldsymbol{y}$, but the matrices should be completely specified.)

Design of Experiment: Prelim Problems
January 2021

Please Do All Problems. Each of the 15 parts carries an equal weight of 10 points.

1. A state highway department studied the wear characteristics of five different paints in the state. The standard, currently used paint (paint 1) and four experimental paints (paints 2, 3, 4, 5) were included in the study. Assuming there is no interaction effect, please write the model under each of the following scenarios. Clearly define all the terms and state all relevant assumptions.

   a) Forty randomly selected locations were separated into 5 groups. Each paint was randomly assigned to one group of locations without replacement. After a suitable period of exposure to weather and traffic, a measure of wear for each paint at each related location was obtained.

   b) Eight locations were randomly selected, thus reflecting variations in traffic densities throughout the state. At each location, a random ordering of the paints to the chosen road surface was employed. After a suitable period of exposure to weather and traffic, a measure of wear for each paint was obtained. Note that as each paint was removed before the next paint was applied, we don't consider carry-over effect here. Assume location effect is random.

   c) Four locations at each of two cities were randomly selected, thus reflecting variations in traffic densities throughout the state. At each location, a random ordering of the paints to the chosen road surface was employed. After a suitable period of exposure to weather and traffic, a measure of wear for each paint was obtained. Note that as each paint was removed before the next paint was applied, we don't consider carry-over effect here. Assume both city effect and location effect are random.

2. The purpose of this study was to develop basic agronomic data for cultivation of selected medicinal plant species in northeastern United States, including the appropriate level of nitrogen for use on these crops.

   Three plant species, Agastache rugosa, Schizonepeta tenuifolia, and Leonurus japonicus, were grown at a University Research Farm with either 0, 50, 100, or 150 kg/ha of nitrogen supplied as soy bean meal. Twenty small fields were allocated to this experiment and 18 were used. Six of the 18 small fields were randomly assigned to each species and entirely planted with the assigned species. Each field was then divided into quarters and each quarter was randomly assigned one of four levels of nitrogen. The 10 center-most plants in the field quarter were harvested at the appropriate time of year. Agastache rugosa plants were harvested when plants were in full flower. Schizonepeta tenuifolia were harvested when plants were in full bloom. Leonurus were harvested in the fall, prior to frost. The 10 harvested plants were dried and their combined dry weight biomass (in grams) was measured. Of interest are differences in variety and nitrogen levels. Assume both variety and nitrogen have fixed effects. Assume there is interaction effect between variety and nitrogen.

   a) Please write the model under the scenario presented above. Clearly define all the terms and state all relevant assumptions.

   b) Fill in the ANOVA table below.

Source	DF	Sum of Squares	Mean Square	E(MS)
Variety			18459.9	
Error(W)			30148.8	
Total(W)				
Nitrogen			137871.3	
Variety × Nitrogen			1108.2	
Error(S)			5879.7	
Total(S)	71			

   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 nitrogen effect at 0.05 significance level, and interpret the test results.

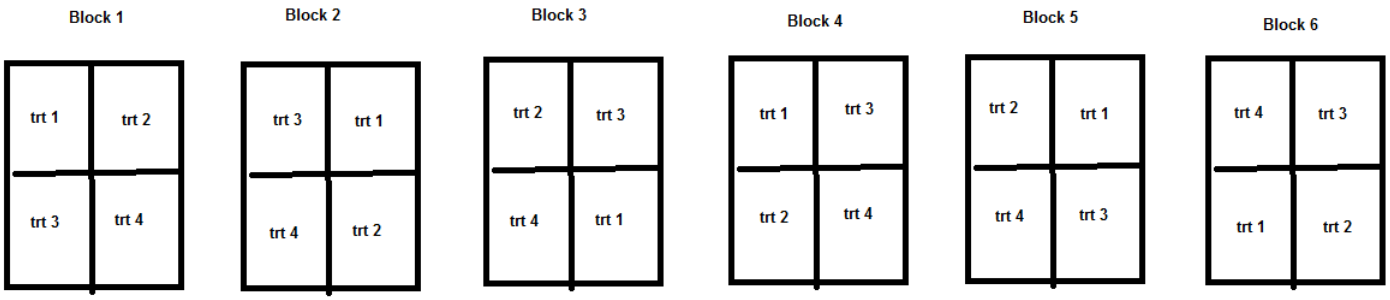
   e) Calculate E(Mean Square for Variety) and E(MSE(W)).

   f) Please give the formula to perform pairwise comparisons for effects of four nitrogen levels at 0.05 significant level.

   g) If we want to estimate the Variety effect more accurately compared to Nitrogen effect, what adjustment do you want to make about the experimental design.

   h) If we assume Nitrogen has random effect, then will you change your answer to question 2d)? If you will, please explain how.

3. A researcher studied the effects of nitrogen supplement and phosphorus supplement on the cotton yield. According to the quality of the field, twenty-four pieces of fields were grouped into six blocks equally. Within each block, four treatment conditions: (1) no supplement, (2) nitrogen supplement only, (3) phosphorus supplement only, and (4) both nitrogen and phosphorus supplements, are randomly applied to the four fields without replacement. Data on yield are collected for each field. The sketch of the design is as follows:



Let Y_{ij} denote the yield of field from j^{th} block which is assigned treatment condition i, for $i = 1, \ldots, 4$, and $j = 1, \ldots, 6$. Suppose the following summary data is available. $\overline{y}_{1.} = 5$, $\overline{y}_{2.} = 15$, $\overline{y}_{3.} = 20$, $\overline{y}_{4.} = 20$, $\sum_j (\overline{y}_{.j} - \overline{y}_{..})^2 = 79$, $\sum_i \sum_j y_{ij}^2 = 6931$.

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

b) Write the ANOVA table appropriate for your model including E(MS).

c) Test whether the effects of two types of nutritional supplements on the responses Y_{ij} could be additive, i.e. whether or not there appears to be interaction between the effects of the two nutritional supplements.

d) Define a contrast to represent the effect of nitrogen. Define another contrast to represent the effect of phosphorus. Find 95% simutaneous confidence intervals for these contrasts.