Applied Statistics Preliminary Examination
Theory of Linear Models
August 2020

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.
  - $\boldsymbol{x}^T$ or $\boldsymbol{A}^T$: indicates transpose of vector $\boldsymbol{x}$ or matrix $\boldsymbol{A}$.
  - $\text{tr}(\boldsymbol{A})$ 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 $\boldsymbol{A}$ be a positive semidefinite symmetric matrix of dimension n, and suppose there exists $\boldsymbol{x}_0 \neq \boldsymbol{0}$ such that for every real vector $\boldsymbol{x} \neq \boldsymbol{0}$ we have the following inequality:

$$\lambda_0 := \frac{\boldsymbol{x}_0^T \boldsymbol{A} \boldsymbol{x}_0}{\boldsymbol{x}_0^T \boldsymbol{x}_0} \leq \frac{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}.$$

(Note: as the notation suggests, λ_0 is defined to be the ratio of quadratic forms $\boldsymbol{x}_0^T \boldsymbol{A} \boldsymbol{x}_0$ and $\boldsymbol{x}_0^T \boldsymbol{x}_0$.)

(a) Show that any eigenvalue/eigenvector pair $(\lambda, \boldsymbol{x})$ of $\boldsymbol{A}$ satisfies the relation:

$$\lambda = \frac{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}.$$

(b) Show that if $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \boldsymbol{0}$ for some $\boldsymbol{x} \neq \boldsymbol{0}$, then it must follow that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$.

(c) Show that $\boldsymbol{A} - \lambda_0 \boldsymbol{I}_n$ is a positive semidefinite matrix.

(d) Putting all the above together, or otherwise, deduce that λ_0 is the minimum eigenvalue of $\boldsymbol{A}$.

(e) Find the minimum value of the quadratic form $\boldsymbol{z}^T \boldsymbol{A} \boldsymbol{z}$ over all real vectors $\boldsymbol{z} \in \mathbb{R}^3$ of unit length (i.e., $\|\boldsymbol{z}\| = 1$), when $\boldsymbol{A}$ is given by:

$$\boldsymbol{A} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix}.$$

2. Consider the $n = 6$ observations $(y_{11}, y_{12}, y_{21}, y_{22}, y_{31}, y_{32}) = (3, -1, 6, 1, 4, 0)$ from the linear model:

$$y_{1j} = \mu + \theta_1 + \epsilon_{1j}, \quad y_{2j} = \theta_2 + \epsilon_{2j}, \quad y_{3j} = \theta_1 + \theta_2 + \epsilon_{3j}, \qquad \text{for } j = 1, 2,$$

with $\epsilon_{ij} \sim$ IID $N(0, \sigma^2)$ for all i and j. The model can be written in the usual vector and matrix form, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y} = (y_{11}, y_{12}, \ldots, y_{31}, y_{32})^T$ and $\boldsymbol{\beta} = (\mu, \theta_1, \theta_2)^T$.

(a) Verify that

$$(\boldsymbol{X}^T \boldsymbol{X})^{-1} = \frac{1}{2} \begin{pmatrix} 3 & -2 & 1 \\ -2 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix}.$$

(b) Show that the projection matrix onto the column space of $\boldsymbol{X}$ is given by:

$$P_{C(\boldsymbol{X})} = \frac{1}{2} \begin{pmatrix} \boldsymbol{J}_{2,2} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}_{2,2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{J}_{2,2} \end{pmatrix}.$$

(c) Find the MVUE (minimum variance unbiased estimator) of σ^2.

(d) Find the MVUE of $\theta_1 + \theta_2$, and give a 95% confidence interval for it.

(e) If it is known that $\theta_1 + \theta_2 = 4$, is there an MVUE of σ^2? If so, find it; it not, find any unbiased estimator.

3. Consider the linear model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T, \qquad \text{where} \quad \boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -1 \\ -1 & 0 & 1 \end{pmatrix}.$$

(a) Find a generalized inverse, $\boldsymbol{G}$, of $\boldsymbol{X}^T\boldsymbol{X}$.

(b) Characterize all the *estimable* functions of the form $\eta = \boldsymbol{\lambda}^T\boldsymbol{\beta}$.

(c) Using your $\boldsymbol{G}$, find the exact distribution of the BLUEs for the functions in the following list that are estimable:

$$\eta_1 = \beta_1 + \beta_2 + \beta_3, \qquad \eta_2 = \beta_2 + \beta_3, \qquad \eta_3 = \beta_1 + 2\beta_2 + \beta_3.$$

(d) Determine if the hypothesis stated below is *testable*, carefully justifying your answer. If it is testable, show how to construct a test statistic for it, and state the distribution of the test statistic under both H_0 and H_1:

$$H_0 : \beta_2 + \beta_3 - 2 = 0 \quad \text{and} \quad \beta_1 + 2\beta_2 + \beta_3 - 1 = 0.$$

Design of Experiment: Prelim Problems

Aug 2020

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

1. An agricultural experiment was designed to compare the effect of five row spacings (18, 24, 30, 36 and 42 inches) on the yield of two soybean varieties (OM=Ottawa Mandarin; B=Blackhawk). Assume row spacings have random effects, and soybean varieties have fixed effects. Assuming there is no interaction effect, please write the model under each of the following scenarios. Clearly define all the terms and state all the relevant assumptions.

   a) A field was partitioned into 60 plots. Each of 10 treatment combinations (2 soybean variety × 5 row spacings) was randomly assigned to 6 plots. The number of plants per plot was kept fixed, as was the plot size. The yield (in bushels per acre) for each of the plots were collected.

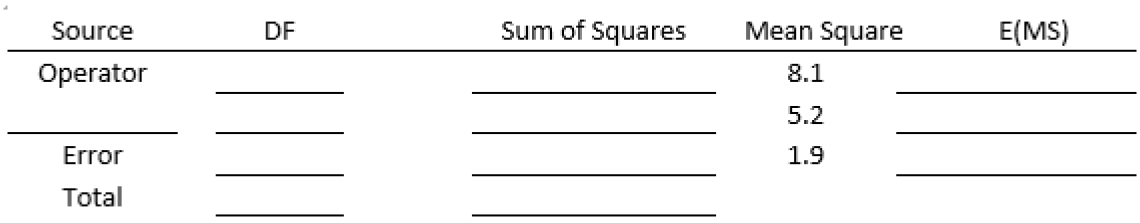
   b) A field was partitioned into 6 long blocks, each was subdivided into 10 plots. Within each block, each of 10 treatment combinations (2 soybean variety × 5 row spacings) was randomly assigned to 1 plot. The number of plants per plot was kept fixed, as was the plot size. The yield (in bushels per acre) for each of the plots were collected.

   c) A field was partitioned into 12 long blocks, each was subdivided into five plots. Each variety of soybean was randomly assigned to six blocks. Within each block, the five plots were randomly assigned a spacing between rows of the planted seeds. The number of plants per plot was kept fixed, as was the plot size. The yield (in bushels per acre) for each of the plots were collected.

2. A textile company weaves a fabric on a large number of looms. The process engineer wants to investigate whether there are significant variations in strength between looms and between operators. To investigate this, she selects sixteen looms and four operators at random. Each of four operators were randomly assigned to four different looms. Four strength determinations are made on the fabric manufactured on each loom. This experiment is run in random order. Assume both operator effect and loom effect are random.

   a) What type of design is this? Write down the model with all the assumptions.

   b) Fill in the ANOVA table bellow.



Source	DF	Sum of Squares	Mean Square	E(MS)
Operator			8.1	
			5.2	
Error			1.9	
Total				

   c) Give a 95% confidence interval for the variability due to Loom Differences. (Round d.f. to the nearest integer if it cannot be found in tables.)

   d) Is there significant variability due to Operator Differences (use $\alpha = 0.05$)?

   e) Estimate all variance components.

   f) Write the formula to calculate 95% confidence interval for the ratio of variability due to loom and the variability due to random error.

   g) One suggestion is to select four looms and four operators at random. All four operators work on each loom at random order. Four strength determinations are made on the fabric manufactured by each operator on each loom. What type of design is this? Write down the model with all the assumptions.

   h) For the design in 2g), can the process engineer determine whether there are significant variations in strength between looms and whether there are significant variations in the strength between operators?

3. The cigarette experiment was run by J. Edwards, H. Hwang, S. Jamison, J. Kindelberger, andJ. Steinbugl in 1996 in order to determine the factors that affect the length of time that a cigarette will burn. There were three factors of interest:

   - Tar (factor A) at two levels, regular and ultra-light,
   - Brand (factor B) at two levels, name brand and generic brand (coded 1 and 2),
   - Age (factor C) at three levels, fresh, 24 hour air exposure, 48 hour air exposure.

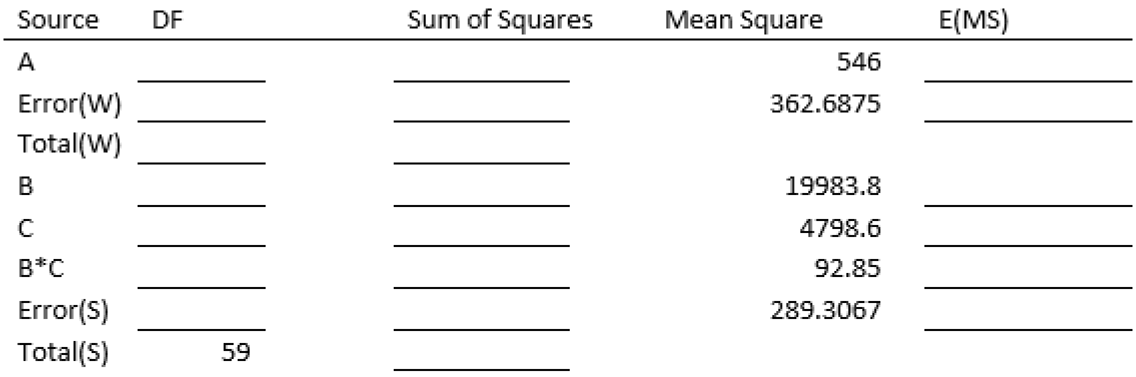
   The cigarettes were to be burned in whole plots of size six. This was to help with the difficulty of recording burning times and to help keep the amount of smoke in the room at a reasonable level.

   There were ten whole plots, and these were assigned at random to the tar levels so that each tar level was assigned five whole plots.

   The six split plots (time slots) in each whole plot were assigned at random to the six brand/age treatment combinations. Marks were made across the seam of each cigarette at a given distance apart. Each cigarette was lit at the beginning of its allotted time slot, and the time taken to burn between the two marks was recorded. Assume all the three factors have fixed effects. Assume B×C is the only interaction effect under consideration.

   a) Write down a model for this experiment.

   b) Fill in the ANOVA table below.



Source	DF	Sum of Squares	Mean Square	E(MS)
A			546	
Error(W)			362.6875	
Total(W)				
B			19983.8	
C			4798.6	
B*C			92.85	
Error(S)			289.3067	
Total(S)	59			

   c) State and conduct statistical tests for factor A effect at 0.05 significant level, and interpret the test results.

   d) State and conduct statistical tests for B×C effect at 0.05 significant level, and interpret the test results.

   e) Calculate E(MSA) and E(MSE^W).

   f) To examine the linear and quadratic trends of burning time due to different ages for name brand, write down the formula to construct confidence intervals for the contrasts with overall 95% confidence level.