Applied Statistics Preliminary Examination

Theory of Linear Models

August 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 11 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. Consider a linear model with $n = 3$ observations Y_1, Y_2, and Y_3 and four parameters μ, α_1, α_2, and α_3. Suppose that

$$\begin{aligned} E(Y_1) &= \mu + \alpha_1 + \alpha_2 + \alpha_3 \\ E(Y_2) &= \mu + \alpha_1 + \alpha_2 \\ E(Y_3) &= \mu + \alpha_1. \end{aligned}$$

Denote the parameter vector $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3)'$.

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

(b) Characterize each of the following functions as estimable or nonestimable: $\mu,\ \alpha_1,\ \mu + \alpha_1,\ \mu + \alpha_1 + \alpha_2 + \alpha_3$. For each one, justify your answer.

(c) Write down the normal equations for this model.

(d) Find two solutions $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ to the normal equations.

(e) Find the least squares estimates of the estimable functions in part (b). Verify that these estimates are invariant to which of $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ is used.

2. Consider the linear regression model

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{1}\beta_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}, \end{aligned} \tag{1}$$

where $\mathbf{Y}$ is $n \times 1$, $\mathbf{1}$ is an $n \times 1$ vector of ones, $\mathbf{X}_1$ is an $n \times p$ matrix of (full) rank p, $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1)$ is an $n \times (p+1)$ matrix with $\mathbf{1}'\mathbf{X}_1 = \mathbf{0}$. β_0 is a fixed (scalar) parameter, $\boldsymbol{\beta}_1$ is a $p \times 1$ vector of fixed parameters, and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')'$. Suppose that $\boldsymbol{\epsilon}$ has zero mean and covariance matrix

$$\mathbf{V} = \sigma^2\{(1-\rho)\mathbf{I} + \rho\mathbf{J}\}, \tag{2}$$

where $\mathbf{J} = \mathbf{1}\mathbf{1}'$, $\mathbf{I}$ is an $n \times n$ identic matrix, $-1 < \rho < 1$ and $\sigma^2 > 0$.

(a) Show

$$\mathbf{V}^{-1} = \left\{\sigma^2(1-\rho)\right\}^{-1}(\mathbf{I} + b\mathbf{J}), \tag{3}$$

where $b = -\rho/(1 - \rho + n\rho)$.

(b) Suppose that ρ is known. Show that the ordinary least squares (OLS) and the generalized least squares (GLS) estimators for $\boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta}_1'\right)'$ are identical.

(c) Find the covariance matrix of the OLS estimator of $\boldsymbol{\beta}_1$.

(d) Let $\widehat{\mathbf{Y}}$ denote the vector of predicted values obtained by fitting the model by OLS, and let $\widehat{\mathbf{e}} = \mathbf{Y} - \widehat{\mathbf{Y}}$. Show that the residual mean square

$$\text{MSE} = \frac{\widehat{\mathbf{e}}'\widehat{\mathbf{e}}}{n-p-1} \tag{4}$$

is an unbiased estimator for $\sigma^2(1-\rho)$.

3. Let X_1, X_2, and X_3 be independent $N(0, \sigma^2)$ random variables. For each of the following indicate whether the statistic has the stated distribution. If the distribution is correct, give the associated parameters. If the distribution is not correct give a reason.

   (a) $\frac{2X_1 - X_2 - X_3}{\sqrt{(X_1+X_2+X_3)^2 + (3/2)(X_2-X_3)^2}}$ has a Student t-distribution.

   (b) $\frac{(4/3)(X_1+X_2+X_3)^2}{(X_3-X_2)^2+(X_3+X_2)^2}$ has an F-distribution.

Design of Experiment: Prelim Problems

August 2019

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

1. A researcher wants to collect 30 observations to compare the effects of 3 types of drugs on blood pressure. 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) 30 subjects were used for the study and each drug was applied to 10 subjects randomly selected.

   b) 10 hospitals were randomly selected, and each drug was assigned to one subjects selected at random from each hospital.

   c) 6 subjects were randomly selected from each of 5 hospitals, and each drug was assigned to two subjects selected at random from each hospital.

   d) 15 subjects were used, each drug was assigned to 5 subjects selected at random, and 2 observations were made on each subject.

   e) 10 subjects were used. Every subject take all the three drugs in a random order. An observation was made for each subject after taking each drug.

2. A structural engineer is studying the strength of aluminum alloy purchased from three vendors. Each vendor submits the alloy in standard-sized bars of 1.0, 1.5, or 2.0 inches. The processing of different sizes of bar stock from a common ingot involves different forging techniques, and so this factor may be important. Furthermore, the bar stock is forged from ingots made in different heats. Each vendor submits two test specimens of each size bar stock from three heats. The heats are randomly selected, and can be different for different vendors. Assume that vendors and bar sizes are fixed and heats are random. Use the unrestricted form of the mixed model. We use Y_{ijkt} to denote the strength of tth aluminum alloy specimen with kth bar stock size purchased from the ith vendor, forged from ingots made in jth heat.

   A student proposed to use a 3-way random complete design as follows:

$$\begin{gathered} Y_{ijkt} = \mu + \alpha_i + B_j + \gamma_k + (\alpha B)_{ij} + (\alpha\gamma)_{ik} + \epsilon_{ijkt} \\ i = 1, 2, 3. \quad j = 1, 2, 3. \quad k = 1, 2, 3. \quad t = 1, 2. \end{gathered}$$

   α_i is the effect of the ith vendor. B_j is the effect of the jth heat for each vendor. γ_k is the effect of kth bar stock sizes. $(\alpha B)_{ij}$ is the interaction effects between ith vendor and jth heat. $(\alpha\gamma)_{ik}$ is the interaction effect between ith vendor and kth size. Table 1 includes part of the student's analysis results.

Table 1: Part of ANOVA for the student's model in Question 2

Source	Mean Square
A	0.0044243
B	0.0167016
C	0.0012631
AB	0.016999
AC	0.0005939
Error	0.000563

   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, mean square, and the expected mean squares for each term in the appropriate model.

   c) Test the hypothesis that size has no effect on the strength of aluminum alloy at 0.05 significant level based on the appropriate model.

   d) If you were to conduct pairwise comparisons for vendors, what would the minimum significant difference value be with overall 95% confidence level based on the appropriate model?

   e) Show E(Mean Square for heat effect) follows the formula you give in 2b).

   f) Give a 95% confidence interval for heat variance component based on the appropriate model.

3. Steel is normalized by heating above the critical temperature, soaking, and then air cooling. This process increases the strength of the steel, refines the grain, and homogenizes the structure. An experiment is performed to determine the effect of temperature and heat treatment time on the strength of normalized steel. Two temperatures and three times are selected. The experiment is performed by heating the oven to one of the two temperatures and inserting three specimens. After 10 minutes one specimen is removed, after 20 minutes the second is removed, and after 30 minutes the final specimen is removed. Then the temperature is changed to the other level and the process is repeated. Four shifts are required to collect the data. Analyze the data assuming temperature and time are both fixed-effect factors.

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

   b) Write out the degrees of freedom and the expected mean squares in table 2 for each term in the model.

Table 2: ANOVA for Question 3

Source	DF	Mean Square	E(MS)
Shift	_____	48.49	_____
Temperature	_____	2340.38	_____
Error^W	_____	80.15	_____
Total^W	_____		_____
Time	_____	79.63	_____
Temperature × Time	_____	397.63	_____
Error^S	_____	40.74	_____

   c) If one wants to compare the effects of the two temperature levels, what would the minimum significant difference value be with 95% confidence level?

   d) Are the averages of the observations at two temperature levels independent from each other? Calculate the variance of the difference of the two averages. Give an unbiased estimation for the variation.

   e) Test the impact of temperature on the strength of normalized steel at 0.05 significant level.

   f) Test the impact of time on the strength of normalized steel at 0.05 significant level. Will the test statistic be different when time is assumed to be random-effect factor? Justify your answer.