Applied Statistics Preliminary Examination
Theory of Linear Models
August 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 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 all the eigenvalues of $n \times n$ matrix $\boldsymbol{A}$ are less than 1 in absolute value, then

$$(\boldsymbol{I}_n - \boldsymbol{A})^{-1} = \boldsymbol{I}_n + \sum_{k=1}^{\infty} \boldsymbol{A}^k.$$

1. Consider the vector of observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ from the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is an $n \times k$ matrix of (full) rank k. Note that, with $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,k})^T$ denoting the vector of covariates for the i-th case, $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, we can write:

$$\boldsymbol{X}^T\boldsymbol{X} = \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T, \qquad \text{and} \qquad \boldsymbol{X}^T\boldsymbol{y} = \sum_{i=1}^{n} y_i\boldsymbol{x}_i.$$

Recall that if $\hat{\boldsymbol{\beta}}$ denotes the usual OLS estimator of $\boldsymbol{\beta}$, the i-th *residual* is $r_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, and we denote by $P_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$ the i-th diagonal element of the *hat matrix*. The goal of this Problem is to obtain an expression for the difference $\hat{\boldsymbol{\beta}}^{(-i)} - \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}^{(-i)}$ denotes the OLS estimator of $\boldsymbol{\beta}$ with the i-th case omitted. (This is leads to the definition of measures of influence such as Cook's distance.)

   (a) Let vectors $\boldsymbol{u}$ and $\boldsymbol{b}$ be such that $|\boldsymbol{b}^T\boldsymbol{u}| < 1$, and assume that the eigenvalues of matrix $\boldsymbol{u}\boldsymbol{b}^T$ are less than 1 in absolute value. Using the "possibly useful results" on page 1, or otherwise, prove that:

$$(\boldsymbol{I}_n - \boldsymbol{u}\boldsymbol{b}^T)^{-1} = \boldsymbol{I}_n + \frac{1}{1 - \boldsymbol{b}^T\boldsymbol{u}}\boldsymbol{u}\boldsymbol{b}^T.$$

   (b) Using (a), or otherwise, prove the following special case of the Sherman-Morrison identity:

$$(\boldsymbol{A} - \boldsymbol{a}\boldsymbol{b}^T)^{-1} = \boldsymbol{A}^{-1} + \frac{1}{1 - \boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{a}}\boldsymbol{A}^{-1}\boldsymbol{a}\boldsymbol{b}^T\boldsymbol{A}^{-1},$$

   where $\boldsymbol{A}$ is invertible and the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are such that $\boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{a} \neq 1$. (Assume all the required conditions for the validity of the result in (a) also hold here.)

   (c) With $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{c} = \boldsymbol{X}^T\boldsymbol{y}$, show that: $\hat{\boldsymbol{\beta}}^{(-i)} = (\boldsymbol{A} - \boldsymbol{x}_i\boldsymbol{x}_i^T)^{-1}(\boldsymbol{c} - y_i\boldsymbol{x}_i)$.

   (d) Use the above results to obtain a (simple) expression for $\hat{\boldsymbol{\beta}}^{(-i)} - \hat{\boldsymbol{\beta}}$ as a function of only: r_i, P_{ii}, $\boldsymbol{x}_i$, and the matrix $\boldsymbol{A}$ defined in (c).

2. In order to estimate the two parameters θ and ϕ, observations y_i, $i = 1, \ldots, N$, are taken, each observation being additively contaminated by IID errors $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. A total of $N = n + m + m$ are observations, $y_i = \mu_i + \epsilon_i$, are thus collected according to the following scheme:

   - n observations having mean $\mu_i = \theta$, for $i = 1, \ldots, n$;
   - m observations having mean $\mu_i = \theta - \phi$, for $i = n + 1, \ldots, n + m$;
   - m observations having mean $\mu_i = \phi - \theta$, for $i = n + m + 1, \ldots, N$.

   In order to answer the following questions, note that this can be set up as the standard linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, whence we denote by $\hat{\boldsymbol{\beta}} = (\hat{\theta}, \hat{\phi})^T$ the usual LSE of $\boldsymbol{\beta} = (\theta, \phi)^T$. To this end, define the following statistics:

$$S_1 = \sum_{i=1}^{n} y_i, \qquad S_2 = \sum_{i=n+1}^{n+m} y_i, \qquad S_3 = \sum_{i=n+m+1}^{N} y_i.$$

   (a) Show that the covariance matrix of $\hat{\boldsymbol{\beta}}$ is:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \frac{\sigma^2}{2mn} \begin{bmatrix} 2m & 2m \\ 2m & 2m + n \end{bmatrix}.$$

   (b) Explicitly compute $\hat{\boldsymbol{\beta}}$.

   (c) Explicitly compute s^2, the usual unbiased estimate of σ^2.

   (d) Construct a 95% confidence interval for $\phi - \theta$.

   (e) Determine a joint 95% confidence region for the two parameters in $\boldsymbol{\beta}$.

3. Consider the vector of 3 observations $\boldsymbol{y} = (y_1, y_2, y_3)^T$ from the linear model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, given in expanded form as:

$$\boldsymbol{y} = \begin{pmatrix} \beta_1 + \beta_2 + \beta_3 \\ \beta_1 + \beta_3 \\ \beta_2 \end{pmatrix} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_3).$$

Define the following linear combinations of the parameters:

$$\eta_1 = \beta_1, \qquad \eta_2 = \beta_2, \qquad \eta_3 = \beta_3, \qquad \eta_4 = \beta_1 - 2\beta_2 + \beta_3, \qquad \eta_5 = \beta_1 - 3\beta_3.$$

In answering the following question parts, be sure to carry all calculations explicitly in order to yield results in as simple a form as possible.

(a) Show that only η_2 and η_4 are *estimable*.

(b) By working with the original rank-deficient parametrization, explicitly compute the BLUE of η_4, and determine its distribution.

(c) Reparametrize the model to full rank, i.e., construct a matrix $\boldsymbol{U}$ such that $\boldsymbol{\gamma} = \boldsymbol{U\beta}$ is the new parameter vector in the full rank linear model $\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\epsilon}$. Then find the BLUE of η_4 in this context, and show that it is the same as in (b).

(d) Is it possible to find other *linear and unbiased* estimators of η_4 that are different from the BLUE? If it is possible, find one such estimator and compare its variance to that of the BLUE. If it is not possible, give a clear justification (proof) of why that is the case.

(e) Construct a level α two-sided test of the null hypothesis $H_0 : \eta_4 = 0$, and clearly state the rejection rule.

Design of Experiment: Prelim Problems

Aug 2021

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

1. An experiment is performed to determine the effect of temperature(factor A) and heat treatment time(factor B) on the strength of normalized steel. Two temperatures and three times are selected. Assuming factor A has fixed effect and factor B has random effect. Please write down the statistical model for the following designs. Assume temperature and time interaction is the only interaction effect under consideration.

   a) Suppose 24 steel specimens are randomly assigned to the 6 treatment combinations, so that each treatment combination has 4 observations.

   b) Suppose the experiment is run on 4 days. On each day, 6 steel specimens are randomly assigned to the 6 treatment combinations.

   c) Suppose the experiment is performed by heating the oven to a randomly selected temperature 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.

   d) Suppose the experiment is performed by heating the oven to a randomly selected temperature and inserting 12 specimens. After 10 minutes four specimens are removed, after 20 minutes another four specimen are removed, and after 30 minutes the final 4 specimens are removed. Then the temperature is changed to the other level and the process is repeated.

2. An experiment concerned the evaluation of eight drugs (factor A at a = 8 levels) for the treatment of arthritis. A second factor was the dose of the drug (factor B at b = 2 levels), and the third factor was the length of time (factor C at c = 2 levels) that a measurement was taken after injection by a substance known to cause an inflammatory reaction. The experimental units used in the study were n = 64 rats. The response was the amount of fluid (in milliliters) measured in the pleural cavity of an animal after having been administered a particular treatment combination.

   In many pharmacological studies, time of day has an effect on the response due to changing laboratory conditions, etc. Consequently, the experiment was divided into 2 days (blocks), whole plots and split plots. The blocks were of size 32, each set of 32 observations being measured on a single day. Each treatment combination was measured once per day. Each day was then subdivided into 4 whole plots of size 8, where the eight measurements within a whole plot were taken fairly close together in time.

   Since the effect of the drug (A) was of primary importance, and since the effects of B and C were of interest only in the form of an interaction with A, the main effects of B and C and the BC interaction were confounded with the whole plots. The levels of factor A are randomly assigned to the split plots. Assume all the three treatment factors' main effects and all their interaction effects have fixed effects. Assume day has random effect. Assume there are no interaction effects between day and treatment factors.

   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)
Day	_____	0.0402	_____
B	_____	0.05796	_____
C	_____	1.39358	_____
B*C	_____	0.03881	_____
Error(W)	_____	0.00213	_____
Total(W)	_____		_____
A	_____	0.10682	_____
A*B	_____	0.25033	_____
A*C	_____	0.13437	_____
A*B*C	_____	0.12976	_____
Error(S)	_____	0.38048	_____
Total(S)	_____		

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

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

e) Calculate E(Mean Square for B) and E(Mean Square for Day).

f) Please give the formula to construct 95% confidence intervals for the differences between pairs of drugs.

g) If we want to determine appropriate number of days to control the width of confidence intervals in f) at most 0.1 milliliters, write down the related inequation that you need to solve.

3. An experiment was described to determine the viscosity of a polymeric material. The material was divided into two samples. The two samples were each divided into ten "aliquots." After preparation of these aliquots, they were divided into two subaliquots and a further step in the preparation made. Finally, each subaliquot was divided into two parts and the final step of the preparation made. The viscosity for each part was measured.

   a) Write down a model for the viscosity determinations allowing for variability in the samples, aliquots, subaliquots and parts.

   b) Fill in the ANOVA table below.

Source	d.f.	MS	E(MS)
Sample		30.0781	
	18	27.9708	
		22.7364	
Error		2.7279	
Total			

   c) Estimate the variances of all the random effects in the model.

   d) Construct 95% confidence interval for the variance component introduced in the subaliquote preparation step.

   e) Test whether the variance component introduced in the subaliquote preparation step is higher than 3 times of the variance of random error at 0.05 significance level.