Applied Statistics Preliminary Examination
Theory of Linear Models
August 2018

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 $\boldsymbol{A}$.
  - $\mathrm{tr}(\boldsymbol{A})$ 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 $\mathrm{var}(X)$: 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 $\boldsymbol{\mu}$ 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 non-singular matrix $\boldsymbol{A}$ is given in block form as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix},$$

    where $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are square, then

$$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{A}_{11}^{-1} + \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^T & -\boldsymbol{C}\boldsymbol{B}^{-1} \\ -\boldsymbol{B}^{-1}\boldsymbol{C}^T & \boldsymbol{B}^{-1} \end{pmatrix},$$

    with $\boldsymbol{B} = \boldsymbol{A}_{22} - \boldsymbol{A}_{12}^T\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$, and $\boldsymbol{C} = \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$.

1. Consider the following $n = 3$ observations from the linear model

$$y_1 = 2\beta_1 + \epsilon_1, \qquad y_2 = \beta_1 + \beta_2 + \epsilon_2, \qquad y_3 = \beta_1 - \beta_2 + \epsilon_3,$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3)^T \sim \text{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_3)$. For $i = 1, 2, 3$, let $\hat{\beta}_i$ denote the *minimum variance unbiased estimator (MVUE)* of β_i.

   (a) Find the MVUE of $\beta_1 - \beta_2$.

   (b) Find the joint distribution of $(\hat{\beta}_1 + \hat{\beta}_2, \hat{\beta}_1 - \hat{\beta}_2)$, and hence compute $\text{corr}(\hat{\beta}_1 + \hat{\beta}_2, \hat{\beta}_1 - \hat{\beta}_2)$, the correlation between $\hat{\beta}_1 + \hat{\beta}_2$ and $\hat{\beta}_1 - \hat{\beta}_2$.

   (c) Construct a 95% confidence interval for $\beta_1 - \beta_2$. Be sure to either specifically compute, or show exactly how to compute, all the quantities in your interval.

   (d) Construct the *likelihood ratio test* of $H_0 : \beta_1 = \beta_2$ vs. the alternative $H_1 : \beta_1 \neq \beta_2$, and clearly specify the rejection rule.

   (e) Repeat part (a) if $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2 \boldsymbol{V})$, where

$$\boldsymbol{V} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

2. Let $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be $n \times k_1$ and $n \times k_2$ matrices of full column ranks k_1 and k_2, respectively, and consider the usual linear model specification:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2).$$

Denoting by $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ the LSE of $\boldsymbol{\beta}$, recall that we have $\hat{\boldsymbol{y}} \equiv \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{y}$, where $\boldsymbol{P}$ is the projection matrix onto $C(\boldsymbol{X})$, and $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is the residual vector. The *Frisch-Waugh-Lovell Theorem* gives the exact form of the LSE of $\boldsymbol{\beta}_1$ as:

$$\hat{\boldsymbol{\beta}}_1 = \left[\boldsymbol{X}_1^T(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1\right]^{-1} \boldsymbol{X}_1^T(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y}, \qquad (\star)$$

where $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are the projection matrices onto $C(\boldsymbol{X}_1)$ and $C(\boldsymbol{X}_2)$, respectively. (Note that by symmetry we have an analogous result for $\hat{\boldsymbol{\beta}}_2$.)

   (a) Noting the identity, $\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$, or otherwise, show that:

$$\boldsymbol{X}_1^T(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y} = \boldsymbol{X}_1^T(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_1^T(\boldsymbol{I} - \boldsymbol{P}_2)(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.$$

   (b) Prove equation $(\star)$.

   (c) Show that if the vector $(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y}$ is regressed on the design matrix $(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1$, the *residual* vector is identical to that from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$.

   (d) Noting that the fitted regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ can be written as, $\hat{\boldsymbol{y}} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{M}_1\boldsymbol{y} + \boldsymbol{M}_2\boldsymbol{y}$, express the matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ as functions of the matrices $\boldsymbol{X}_i$, $\boldsymbol{P}_i$, $i = 1, 2$, and the vector $\boldsymbol{y}$.

   (e) Does it necessarily follow that $\boldsymbol{M}_1 = \boldsymbol{P}_1$ and $\boldsymbol{M}_2 = \boldsymbol{P}_2$? What relationship between $C(\boldsymbol{X}_1)$ and $C(\boldsymbol{X}_2)$ will guarantee this equality?

3. Consider the two-way cell means anova model without interaction, $y_{ij} = \alpha_i + \beta_j + \epsilon_{ij}$, for $i = 1, 2$ and $j = 1, 2$, with $\epsilon_{ij} \sim$ IID $N(0, \sigma^2)$ for all i and j. The model can be written in vector and matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{y} = (y_{11}, y_{12}, y_{21}, y_{22})^T, \qquad \boldsymbol{\beta} = (\alpha_1, \alpha_2, \beta_1, \beta_2)^T.$$

We are interested in making inference on the linear combination: $\eta = 5\alpha_1 + 2\alpha_2 + 3\beta_1 + 4\beta_2$.

(a) Find conditions on the constants $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ such that $\lambda_1\alpha_1 + \lambda_2\alpha_2 + \lambda_3\beta_1 + \lambda_4\beta_2$ is *estimable*, and hence conclude that η is estimable.

(b) Determine a valid (and simple) set of *side conditions* that can be used to reparametrize the model to full rank, and show that upon substitution of these parameter restrictions, the new (full rank) model becomes:

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\gamma} = (\alpha_1, \alpha_2, \beta_1)^T, \qquad \text{where} \quad \boldsymbol{Z}^T\boldsymbol{Z} = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

(c) For positive integers a, b, and m such that $ab - m \neq 0$, find the form of the inverse $\boldsymbol{H}^{-1}$ for the matrix $\boldsymbol{H}$ expressed in block form as:

$$\boldsymbol{H} = \begin{pmatrix} a\boldsymbol{I}_m & \boldsymbol{j}_m \\ \boldsymbol{j}_m^T & b \end{pmatrix}.$$

(d) Find the BLUE of η, and give a 95% confidence interval for it.

Design of Experiment: Prelim Problems
August 2018
Please Do All Problems. Each of the 14 parts carries an equal weight of 10 points.

1. Agricultural researchers wished to investigate the effects of five different fertilizers on the growth of mangold roots. The five factors were sulphate of ammonia (factor A at levels 0 or 0.6 cwt per acre), superphosphate (factor B at levels 0 or 0.5 cwt per acre), muriate of potash (factor C at levels 0 or 1.0 cwt per acre), agricultural salt (factor D at levels 0 or 5 cwt per acre), and dung (factor E at levels 0 or 10 tons per acre). The researchers grew only 8 mangold roots using 8 different treatment combinations, one per root. Treatments were assigned randomly to roots. Coding the levels of the factors as 0 (for the low level) and 1 (for the high level). Table 1 shows the data, where Y is the length of the roots in cm. The

Table 1: Data of root length for Question 1

Y	A	B	C	D	E
21.7707	0	0	0	0	0
19.1218	0	0	0	1	1
13.4921	1	1	0	0	0
24.5265	1	1	0	1	1
21.3217	0	1	1	1	0
20.5575	0	1	1	0	1
16.8585	1	0	1	1	0
19.5910	1	0	1	0	1

researchers fit a complete 2^5 factorial model to the data and obtained type I sum of squares shown in Table 2.

   (a) Use the information above to determine which effects are aliased.

   (b) Suppose only main effects are not negligible. Create an appropriate ANOVA table. What do you conclude about the effects of factors A, B, C, D, and E on the length of mangold roots? Discuss.

2. A state highway department studied the wear characteristics of five different paints at eight locations in the state. The standard, currently used paint (paint 1) and four experimental paints (paints 2, 3, 4, 5) were included in the study. The 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 combined measure of wear for each paint, considering both durability and visibility, was obtained. The data of wear and part of the ANOVA table are in Table 3 and Table 4.

   (a) State an appropriate statistical model including model assumptions.

   (b) Obtain the complete ANOVA table.

Table 2: Type I sum of squares for Question 1

Source	DF	SS
A	1	8.6188
B	1	0.8165
C	1	0.0424
D	1	5.1474
E	1	13.4005
AB	0	0
AC	0	0
AD	1	12.9709
AE	1	36.8940
BC	0	0
BD	0	0
BE	0	0
CD	0	0
CE	0	0
DE	0	0
ABC	0	0
ABD	0	0
ABE	0	0
ACD	0	0
ACE	0	0
ADE	0	0
BCD	0	0
BCE	0	0
BDE	0	0
CDE	0	0
ABCD	0	0
ABCE	0	0
ABDE	0	0
ACDE	0	0
BCDE	0	0
ABCDE	0	0
Error	0	0
Total	7	77.8904

(c) Test whether the mean wear differs for the five paints; use $\alpha = 0.05$. State the hypothesis, decision rule, and conclusion.

(d) Paints 1, 3, and 5 are white, whereas paints 2 and 4 are yellow. Estimate the difference in the mean wear for the two groups of paints. In addition, compare the difference in the mean wear for the four experimental paints and that for the standard paint. Construct confidence intervals for these two sets of contrasts with overall confidence level to be at

Table 3: Data of wear for Question 2

location	Paint				
	1	2	3	4	5
1	11	13	10	18	15
2	20	28	15	30	18
3	8	10	8	16	12
4	30	35	27	41	28
5	14	16	13	22	16
6	25	27	26	33	25
7	43	46	41	55	42
8	13	14	12	20	13
mean	20.500	23.625	19.000	29.375	21.125

Table 4: ANOVA for Question 2

Source	DF	SS	MS	E(MS)
Location	_____	4826.375	_____	_____
Paint	_____	531.350	_____	_____
Error	_____	122.250	_____	_____
Total				

Table 5: ANOVA with interaction for Question 3

Source	DF	SS	MS	E(MS)
Location				
Paint				
Paint*Location				
Error			3.0	

least 90%. Interpret your findings.

(e) If in practice because of time and money limits to do the experiment, only 4 paints can be applied in random order for each location, and AT MOST 8 locations can be used, then is it possible to have a balanced incomplete block design? If yes, please write down one possible design plan. If not, please give the reason.

(f) Is it reasonable to use p-value for "Location" to evaluate the significance of this factor? If yes, calculate it. If not, explain why not.

(g) One researcher suggests to randomly select 8 locations for each paint to run the experiment. What design is this? What are the advantages and disadvantages of this design compared to the one the state highway department actually used?

3. Assume that in the study in Question 2, 3 observations were collected for each combination and each value in Table 4 was the average of 3 replicates. A two-way ANOVA model with interaction is fitted and the MSE is 3.0.

(a) State an appropriate statistical model including model assumptions.

(b) Complete the ANOVA table in Table 5.

(c) Will the test for paint effect be the same as for Question 2(c)? If not, please give the new test statistics, and draw your conclusion.

(d) Construct a 95% confidence interval for the variance of Paint*Location. (Round d.f. to the nearest integer if it cannot be found in tables.)

(e) One is interested in doing hypothesis testing on whether the variance component for Paint*Location is larger than random error variance at 0.05 significant level. If the analyzer wants the power of the hypothesis testing to be 0.95 when the underlying true variance component for Paint*Location is twice of the random error variance, what is the appropriate sample size? Please write down the inequation to get the sample size, which should be an expression of the number of replicates r for each treatment combination. You don't need to do the calculation.