Applied Statistics Preliminary Examination
Theory of Linear Models
August 2017

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 15 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 $\boldsymbol{\mu}$ 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 $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and the random vector $\boldsymbol{z}$ given by:

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{z} = \begin{pmatrix} y_1 \\ \overline{y} \end{pmatrix},$$

where $\overline{y} = (y_1 + y_2 + y_3)/3$ is the usual sample mean. In addition, let $w = 2y_1 - 3y_2 + y_3$ and $Q = 9y_1^2 + 30y_1y_2 + 6y_1y_3 + 25y_2^2 + 10y_2y_3 + y_3^2$.

   (a) Find the distribution of w.

   (b) Are w and Q independent? Justify.

   (c) Find the distribution of $y_1 | y_2, y_3$, i.e., the conditional distribution of y_1 given y_2 and y_3.

   (d) Find the distribution of $\boldsymbol{z}$.

   (e) If $v = ay_1 + by_2 + cy_3$, is it possible to find $a \neq 0$, $b \neq 0$, and $c \neq 0$, such that v and $\boldsymbol{z}$ are independent? If so, find such a set of values for $\{a, b, c\}$; if not, justify your reasoning.

2. For $n_1 \times p$ matrix $\boldsymbol{X}_1$ and $n_2 \times p$ matrix $\boldsymbol{X}_2$, both matrices being of (full) rank p, and for $\boldsymbol{\epsilon}_1$ independent of $\boldsymbol{\epsilon}_2$, consider the two linear models, Model (1) and Model (2):

$$\boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1, \qquad \boldsymbol{\epsilon}_1 \sim \mathrm{N}(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I}_{n_1}), \tag{1}$$

$$\boldsymbol{y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2, \qquad \boldsymbol{\epsilon}_2 \sim \mathrm{N}(\boldsymbol{0}, \sigma_2^2 \boldsymbol{I}_{n_2}), \tag{2}$$

and denote by $\hat{\boldsymbol{\beta}}_i$ the LSE of $\boldsymbol{\beta}_i$ and s_i^2 the usual unbiased estimator of σ_i^2, in Model (i), $i = 1, 2$. Defining $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ and $n = n_1 + n_2$, consider also the model that combines these two:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \qquad \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{pmatrix}. \tag{†}$$

Define also $\boldsymbol{G}_i = (\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}$, for $i = 1, 2$, and $\boldsymbol{A} = [\sigma_1^2\boldsymbol{G}_1 + \sigma_2^2\boldsymbol{G}_2]^{-1}$, with the asumption that $\text{rank}(\boldsymbol{A}) = p$. The ultimate goal of this question is to devise a test of $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

   (a) Show that $\boldsymbol{X}$ is of (full) rank $2p$.

   (b) Show that the *generalized least squares* (GLS) estimate of $\boldsymbol{\beta}$ in model (†) is $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$.

   (c) Find the distribution of $w = (n_1 - p)s_1^2/\sigma_1^2 + (n_2 - p)s_2^2/\sigma_2^2$.

   (d) Find the distribution of $v = (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)^T\boldsymbol{A}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)$ under $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

   (e) Hence deduce that for an appropriate constant c, $F = cv/w$ has an F-distribution under $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, and specify the parameters of F as well as the value of c.

3. Let $n \geq 6$ be an <u>even number</u>, and consider the vector of observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ from the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, given in expanded form as:

$$\boldsymbol{y} = \begin{pmatrix} 1 & 1 & 2 & 0 & -1 \\ 1 & 2 & 3 & -1 & -2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 2 & 0 & -1 \\ 1 & 2 & 3 & -1 & -2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n).$$

Note carefully the repeating block structure of $\boldsymbol{X}$: all odd-numbered rows are identical, and all even-numbered rows are identical (recall that n is even).

(a) Show that $\boldsymbol{X}$ has rank 2.

(b) Let $\boldsymbol{X}^T\boldsymbol{X}$ be given in block form as shown below, where $\boldsymbol{A}$ is a 2×2 matrix. Compute $\boldsymbol{A}$, and use it to construct a generalized inverse $\boldsymbol{G}$ of $\boldsymbol{X}^T\boldsymbol{X}$.

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{C} \end{pmatrix}.$$

(c) By characterizing all the *estimable* functions of the form $\boldsymbol{\lambda}^T\boldsymbol{\beta}$, show that $\eta = \beta_2 + \beta_3 - \beta_4 - \beta_5$ is estimable.

(d) Completely determine the distribution of any LSE of the estimable function η in (c). That is, compute the distribution of $\hat{\eta} = \hat{\beta}_2 + \hat{\beta}_3 - \hat{\beta}_4 - \hat{\beta}_5$, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_5)^T$ is any solution of the *least squares normal equations*.

(e) Explicitly show how the above results can be used 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 indicate how to compute the new (full rank) design matrix $\boldsymbol{Z}$. (Note: a numerical answer is expected for $\boldsymbol{U}$, but it suffices to then express $\boldsymbol{Z}$ as a function of $\boldsymbol{U}$ and $\boldsymbol{X}$.)

Design of Experiment: Prelim Problems

August 2017

Please Do All Problems. Each of the 12 parts carries an equal weight of 10 points.

For each test, state the null and alternative hypotheses in terms of the model parameters

1. To simplify production scheduling, an industrial engineer is studying the possibility of assigning one time standard to a particular class of jobs, believing that differences between jobs is negligible. To see if this simplification is possible, six jobs are randomly selected. Each job is given to a different group of three operators. Each operator completes the job twice at different times during the week, and the following results are obtained.

	Operator 1		Operator 2		Operator 3	
Job	1	2	1	2	1	2
1	158.3	159.4	159.2	159.6	158.9	157.8
2	154.6	154.9	157.7	156.8	154.8	156.3
3	162.5	162.6	161.0	158.9	160.5	159.5
4	160.0	158.7	157.5	158.9	161.1	158.5
5	156.3	158.1	158.3	156.9	157.7	156.9
6	163.7	161.0	162.3	160.3	162.6	161.8

Computer Output:

```
ANOVA: Time versus Job, Operator

Factor          Type    Levels Values
Job             random       6      1      2      3      4      5      6
Operator(Job)   random       3      1      2      3

Analysis of Variance for Time

Source           DF          SS          MS          F          P

Job              __    ________      29.622    _______    _______

Operator(Job)    __    ________       1.721    _______    _______

Error            __    ________       1.092

Total            35     188.430
```

(a) What design/experiment is this?

(b) Write the statistic model and the corresponding assumptions.

(c) Fill in the missing values for the output.

(d) Estimate the variability between jobs. Write the hypothesis in notation for testing the equality of the jobs. Test the hypothesis at 0.05 significant level.

(e) Estimate the variability between the operators. Construct its 95% confidence interval. (Round d.f. to the nearest integer if it cannot be found in tables).

(f) What are your conclusions about the use of a common time standard for all jobs in this class?

2. An experiment was conducted to establish whether two coolants had different effects on the performance of lathes. Five lathes were used for the experiment. Two pins were manufactured with each lathe, with each coolant. The diameters of the pins were measured. The analyst decided to treat these lathes as a random sample from the population to which inference was to be made, and so the lathe effects were treated as random effects. The analyst fit the model with no interaction between coolant and lathe. The model for Diameter is

$$Y_{ijk} = \mu + \alpha_i + B_j + \epsilon_{ijk},$$

Where $i = 1, 2; j = 1, 2, 3, 4, 5; k = 1, 2.$ μ and the α_i(Coolant effects) are considered fixed effects. The B_j(Lathe effects) follow $N(0, \sigma_B^2)$ and the ϵ_{ijk} follow $N(0, \sigma^2)$. The ϵ_{ijk} and B_j are mutually independent. The ANOVA table below shows output for an analysis of this model:

Analysis of Variance for Diam

Source	DF	SS	MS
Coolant	____	________	0.121897
Lathe	____	________	________
Error	____	0.125515	________
Total	____	0.333850	

(a) Test $H_0^B: \sigma_B^2 > 2\sigma^2$ *and* $H_1^B: \sigma_B^2 \leq 2\sigma^2$ at 0.05 significant level.

(b) Under this model, provide an unbiased estimator for $\mu + \alpha_1$, and an expression for the variance of this estimator. (You do not need to provide a numerical answer).

(c) Provide a formula to estimate the contrast $\alpha_1 - \alpha_2$. Show that your estimator is unbiased.

(d) Construct 95% confidence interval for $\alpha_1 - \alpha_2$, assuming it is known that $\widehat{\alpha_1} - \widehat{\alpha_2} > 0$ based on the data.

(e) An alternative model allows an interaction between Lathe and Coolant (using unrestricted model). Provide an outline of the ANOVA table for this model, giving sources of variation, degrees of freedom, and expected mean squares. Fill in as many of the sums of squares as possible.

(f) Under (e), is the test statistic in (a) still valid? Please explain.