Applied Statistics Preliminary Examination
Theory of Linear Models
August 2016

Instructions:

- Do all 3 Problems. Neither calculators nor electronic devices of any kind are allowed. Clearly state any theorem or fact that you use. Each of the 14 parts carries an equal weight of 10 points. Note: 2b(i) and 2b(ii) each count as one part.
- 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 $|\mathbf{A}|$: denotes the trace and determinant, respectively, of matrix $\mathbf{A}$.
  - $\mathbf{I}_n$: the $n \times n$ identity matrix.
  - $\mathbf{j}_n = (1, \ldots, 1)^T$ is an n-vector of ones, and $\mathbf{J}_{m,n}$ is an $m \times n$ matrix of ones.
  - $\mathbb{E}(X)$ and $\mathbb{V}(X)$: 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: $X \sim F(n_1, n_2)$.
- Possibly useful results.
  - If $\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given in partitioned form as

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

    with $m_1 = \dim(\mathbf{x}_1)$, then the conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$ is

$$\mathbf{x}_1 | \mathbf{x}_2 \sim N_{m_1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2),\ \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right).$$

1. Let $\mathbf{y} = (y_1, y_2, y_3)^T \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where we define $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and the vector $\mathbf{z} = (z_1, z_2)^T$ as:

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 6 & 3 \\ 0 & 3 & 2 \end{pmatrix}, \qquad \text{and} \qquad \mathbf{z} \equiv \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} y_1 - 2y_2 + y_3 \\ y_2 \end{pmatrix}.$$

In addition, let $W = 3z_1^2 + 10z_1z_2 + 10z_2^2$, and note that W is defined in terms of the elements of $\mathbf{z}$.

(a) Find the distribution of $\mathbf{z}$.

(b) Does there exist a constant c such that cW has a χ^2 distribution? Justify, and if so, find c.

(c) Find the distribution of $y_1|\overline{y} = 1$, where $\overline{y} = (y_1 + y_2 + y_3)/3$ is the usual sample mean.

(d) Find $\rho_{1,2|3}$, the partial correlation coefficient between y_1 and y_2, given that y_3 is constant.

(e) Suppose that there exists another random variable, y_4, such that (y_1, y_2, y_3, y_4) is multivariate normal, i.e. $(y_1, y_2, y_3, y_4)^T \sim N_4(\boldsymbol{\eta}, \boldsymbol{\Omega})$. Assuming that the conditional distribution of y_4 given (y_1, y_2, y_3) is $N(y_1 - 2y_2 - y_3 + 1, 4)$, compute all the elements of the covariance matrix $\boldsymbol{\Omega}$.

2. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$, and $\mathbf{X}$ is an $n \times k$ design matrix with $\text{rank}(\mathbf{X}) = k < n$. Further suppose that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$, where $\mathbf{V}$ is a known positive definite matrix, and $\sigma^2 > 0$ is an unknown parameter. Let $\hat{\boldsymbol{\beta}}_{\text{BLUE}}$ denote the BLUE of $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}_{\text{OLSE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ denote the OLS estimator of $\boldsymbol{\beta}$.

(a) Prove that the BLUE of $\boldsymbol{\beta}$ is given by: $\hat{\boldsymbol{\beta}}_{\text{BLUE}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$.

(b) Show that the OLSE and BLUE of $\boldsymbol{\beta}$ are the same *if and only if* there exists a non-singular (square) matrix $\mathbf{B}$ such that $\mathbf{VX} = \mathbf{XB}$. Proceed as follows:

i. First, assuming that the OLSE and BLUE are the same, explicitly compute $\mathbf{B}$.

ii. Then, assuming that there exists a non-singular $\mathbf{B}$ such that $\mathbf{VX} = \mathbf{XB}$, show that the OLSE and BLUE must coincide.

(c) Prove that: $\text{tr}\left[\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\right] = k$.

(d) Show that the following is an *unbiased* estimator of σ^2:

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T\left[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\right]\mathbf{y}}{n-k}.$$

3. Consider the vector of observations $\mathbf{y} = (y_1, \ldots, y_4)^T$ from the linear model:

$$\mathbf{y} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -1 & 0 \\ -1 & 1 & 1 \\ -1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_4).$$

(a) Find a solution to the *normal equations*.

(b) Determine which of the following functions are *estimable*:

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

(c) Find BLUEs for the estimable functions in (b), and compute their corresponding distributions.

(d) Justify your reasoning in determining if the following hypothesis is *testable*,

$$H_0 : \{\beta_1 - \beta_2 = 1,\ \beta_3 = 0\}.$$

If it is testable then describe a suitable test of this hypothesis, and state the distribution of the test statistic under H_0 as well as under the alternative hypothesis,

$$H_1 : \{\beta_1 - \beta_2 = 0,\ \beta_3 = 1,\ \sigma^2 = 1\}.$$

Design of Experiments: Prelim Problems
August 2016

Please Do All Problems

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

Note 1: Tukey's Studentized range distribution: If $\bar{Y}_1, \cdots, \bar{Y}_n$ are independent random variables with $N(\mu, \sigma^2/n)$ distribution then for $\hat{\sigma}$ being an unbiased estimator of σ, the statistic $(\max_i \bar{Y}_i - \min_i \bar{Y}_i)/(\hat{\sigma}/\sqrt{n})$ is said to have Studentized range distribution.
Note 2: Satterthwaite approximation of degrees of freedom for a "mean square".

1. Suppose independent data pairs $(x_{ij}, Y_{ij}), i = 1, \cdots, 6, j = 1, \cdots, 4$ are observed, where index i represents a treatment group, and index j represents a replication (within each treatment group). The following four models are fit to the data (using ordinary least squares), with the resulting error sums of squares as specified:

Model 1:	$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$	SSE = 750
Model 2:	$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$	SSE = 600
Model 3:	$Y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$	SSE = 400
Model 4:	$Y_{ij} = \mu + \beta x_{ij} + \varepsilon_{ij}$	SSE = 800

The total sum of squares is 1000. In these models, α_i's are treatment effects satisfying $\sum_{i=1}^{6} \alpha_i = 0$, parameters μ, β, $\beta_1, \cdots, \beta_6$ are unrestricted, and ε_{ij} is the error term. Perform the following F-tests at level $\alpha = 0.05$, clearly stating the null and alternative hypotheses in terms of the parameters. You may assume that the usual normal-theory conditions are valid in each case.

(a) Test for treatment effects on the response variable Y_{ij} as a simple one-way ANOVA (ignoring x_{ij}). (10 points)

(b) Test whether the slope parameter in the simple linear regression of Y_{ij} on x_{ij} is nonzero, ignoring all effects due to different treatment groups. (10 points)

(c) Test whether different treatment groups have different slope parameters for the regression of Y_{ij} on x_{ij} , allowing for separate intercepts for each group. (10 points)

(d) Perform a classical analysis of covariance (ANCOVA) test (common slope model) for treatment effects on Y_{ij} (adjusting for a linear effect in the covariate x_{ij}). (10 points)

2. Often the following model is used for a $p \times p$ Latin square replicated n times.

$$Y_{ijkh} = \mu + \rho_h + \alpha_{i(h)} + \tau_j + \beta_{k(h)} + (\tau\rho)_{jh} + \varepsilon_{ijkh}, \text{ for } (i, j, k = 1, \cdots, p;\ h = 1, \cdots, n)$$

where Y_{ijkh} is the observation on treatment j in row i and column k for the hth square.

(a) Assuming that replication effect is random and all of the other effects are fixed, define the terms in the above model and state all conditions and assumptions. (10 points)

(b) Assume that the following information is available.

Source	SS
Treatment	56.778
Replication	53.389
Row(Rep)	8.889
column(Rep)	7.556
Treatment*Rep	2.111
Error	4.889

Construct the ANOVA table, include the expected mean square column, and test for treatment effects. (20 points)

(c) Formulate Tukey's pairwise comparison procedure for the treatment effects. Assuming that $\bar{y}_{.1..} = 24$, $\bar{y}_{.2..} = 37$, $\bar{y}_{.3..} = 50$ carry out the comparisons. (10 points)

3. Consider the following model for a three-factor factorial design:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + \varepsilon_{ijk},\ i = 1, \cdots, a,\ \ j = 1, \cdots, b,\ \ k = 1, \cdots, c$$

where index i, j, k corresponds to the levels of factor A, B, and C, respectively.

(a) Assuming that all three factors are random, state the assumptions needed for analysis. (10 points)

(b) Develop the analysis of variance table consisting of the information for the source of variation, degrees of freedom, and the expected mean squares. (10 points)

(c) Propose an appropriate test statistics for the variance component of factor B. (10 points)