Applied Statistics Preliminary Examination
Theory of Linear Models
August 2015

Instructions:

- Do all 4 Problems. Neither calculators nor electronic devices of any kind are allowed. Clearly state any theorem or fact that you use. Each of the 13 parts is worth 10 points, except for 2(c) which is worth 5 points, and 4(e) which is worth 15 points.
- Abbreviations/Acronyms.
  - IID (independent and identically distributed); LSE (least squares estimator); BLUE (best linear unbiased 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.
  - $\mathbb{E}(X)$ 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 $X \sim F(p, q)$, then:

$$\frac{\left(\frac{p}{q}\right) X}{1 + \left(\frac{p}{q}\right) X} \sim \text{Beta}\left(\frac{p}{2}, \frac{q}{2}\right).$$

1. Let $\mathbf{A}$ and $\mathbf{B}$ be $n \times n$ symmetric idempotent matrices with $\text{rank}(\mathbf{A}) = r > 0$, $\text{rank}(\mathbf{B}) = s > 0$, and satisfying the condition $\mathbf{AB} = \mathbf{0}$. Letting $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, with $\varepsilon_1, \ldots, \varepsilon_n \sim$ IID $\text{N}(0, \sigma^2)$, define $Q_{\mathbf{C}} = \boldsymbol{\varepsilon}^T \mathbf{C} \boldsymbol{\varepsilon}$ to be a generic quadratic form in $\boldsymbol{\varepsilon}$ formed from the $n \times n$ symmetric matrix $\mathbf{C}$.

   (a) Show that:
   $$\frac{sQ_{\mathbf{A}}}{rQ_{\mathbf{B}}} = \frac{s\boldsymbol{\varepsilon}^T \mathbf{A} \boldsymbol{\varepsilon}}{r\boldsymbol{\varepsilon}^T \mathbf{B} \boldsymbol{\varepsilon}} \sim F(r, s).$$

   (b) Find the distribution of:
   $$\frac{(n-r)Q_{\mathbf{A}}}{rQ_{\mathbf{I}_n - \mathbf{A}}} = \frac{(n-r)\boldsymbol{\varepsilon}^T \mathbf{A} \boldsymbol{\varepsilon}}{r\boldsymbol{\varepsilon}^T (\mathbf{I}_n - \mathbf{A}) \boldsymbol{\varepsilon}}.$$

   (c) Find the distribution of:
   $$\frac{Q_{\mathbf{A}}}{Q_{\mathbf{A}+\mathbf{B}}} = \frac{\boldsymbol{\varepsilon}^T \mathbf{A} \boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T (\mathbf{A} + \mathbf{B}) \boldsymbol{\varepsilon}}.$$

2. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, $\mathbf{y} = (y_1, \ldots, y_n)^T$, and $\mathbf{X}$ an $n \times p$ design matrix of full rank. Let $\hat{\boldsymbol{\beta}}$ denote the LSE of $\boldsymbol{\beta}$, $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ the residual vector, and s^2 the usual unbiased estimator of σ^2.

   (a) Find the distribution of the vector $\mathbf{u} = (\hat{\boldsymbol{\beta}}^T, \mathbf{e}^T)^T$, and hence show that $\hat{\boldsymbol{\beta}}$ and $\mathbf{e}$ are independent.

   (b) Suppose that a new observation $y_0 = \mathbf{x}_0^T \boldsymbol{\beta} + \epsilon_0$ is made, where $\epsilon_0 \sim N(0, \sigma^2)$ is independent of $\boldsymbol{\epsilon}$. Derive from first principles the distribution of the predictive pivot:
   $$w = \frac{y_0 - \mathbf{x}_0^T \hat{\boldsymbol{\beta}}}{s\sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}}.$$

   (c) Using the result from (b), construct a $(1-\alpha)100\%$ prediction interval for y_0.

3. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, $\mathbf{X}$ is an $n \times p$ design matrix, $\mathbf{Z}$ is an $n \times q$ design matrix, and $\mathbf{M} = (\mathbf{X}, \mathbf{Z})$ is hence $n \times (p+q)$ of full rank.

   (a) Construct the F-test of $H_0 : \boldsymbol{\delta} = \mathbf{0}$ vs. $H_1 : \boldsymbol{\delta} \neq \mathbf{0}$.

   (b) Show that the likelihood ratio test of $H_0 : \boldsymbol{\delta} = \mathbf{0}$ vs. $H_1 : \boldsymbol{\delta} \neq \mathbf{0}$ is equivalent to the F-test in (a).

4. Consider the linear model $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, for $i = 1, 2, 3$ and $j = 1, 2, 3$, with the $\epsilon_{ij} \sim$ IID $\text{N}(0, \sigma^2)$ for all i and j. The model can be written in vector and matrix form as:
   $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \mathbf{y} = (y_{11}, \ldots, y_{13}, \ldots, y_{31}, \ldots, y_{33})^T, \qquad \boldsymbol{\beta} = (\mu, \tau_1, \tau_2, \tau_3)^T.$$

   (a) Write down the normal equations as a set of 4 equations in 4 unknowns.

   (b) Propose a full set of linearly estimable functions for this model.

   (c) Propose a set of side conditions in order to reparametrize this model to full rank, and find the LSE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

   (d) Show that $H_0 : \tau_1 = \tau_2 = \tau_3$ is a *testable* hypothesis.

   (e) Compute the ANOVA table, and especially the F-statistic, for testing the hypothesis in (d).

Design of Experiments: Prelim Problems
August 2015

Please Do All Problems

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

1. A process engineer is testing the yield of a product manufactured on three machines. Each machine can be operated at two power settings. Furthermore, each machine has three stations on which the product is formed. An experiment is conducted in which each machine is tested at both power settings, and three observations on yield are taken from each station. The runs are made in random order, and the results are shown in the following table. Assume that all three factors are fixed.

	Machine 1			Machine 2			Machine 3		
Station	1	2	3	1	2	3	1	2	3
Power	34.1	33.7	36.2	31.1	33.1	32.8	32.9	33.8	33.6
Setting 1	30.3	34.9	36.8	33.5	34.7	35.1	33.0	33.4	32.8
	31.6	35.0	37.1	34.0	33.9	34.3	33.1	32.8	31.7
Power	24.3	28.1	25.7	24.1	24.1	26.0	24.2	23.2	24.7
Setting 2	26.3	29.3	26.1	25.0	25.1	27.1	26.1	27.4	22.0
	27.1	28.6	24.9	26.3	27.9	23.9	25.3	28.0	24.8

   (a) Write a model for the data and clearly define all the terms and state all relevant assumptions. (10 points)

   (b) Write the expressions for the sums of squares corresponding to the effects in the linear model in part (a). No actual calculations are required. (10 points)

   (c) Write the ANOVA table and conduct all appropriate tests for the significance of effects associated with the model in part (a). Include the actual values of the degrees of freedom but no other calculations are required. (10 points)

   (d) Suppose that a large number of power settings could have been used and that the two selected for the experiment were chosen randomly. Obtain the expected mean squares for this situation assuming the unrestricted form of the mixed model, and appropriately modify the analysis in part (c). (10 points)

   (e) Let $\bar{Y}_{m1} - \bar{Y}_{m2}$ denote the difference of the means of machines 1 and 2. What is the estimated variance of this difference under the assumptions of part (d)? (10 points)

2. A balanced two-factor experiment yields responses Y_{ijk} for replication k, at level i of factor A and level j of factor B, for $i = 1, 2,\ j = 1, 2,\ k = 1, 2, 3, 4$. The data are analyzed using the model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

(a) Assuming that both factors are fixed, write the assumptions associated with the linear model. (10 points)

(b) Assuming that we have the following values of the least squares estimates of the model parameters:

$$\widehat{\mu} = 10, \quad \widehat{\alpha}_1 = 6, \quad \widehat{\beta}_1 = 2, \quad \widehat{(\alpha\beta)}_{11} = 1$$

Find the estimates of the remaining parameters. (10 points)

(c) Assuming that $\sum_{ijk} y_{ijk}^2 = 2400$, calculate the sums of squares SSA, SSB, SSAB, SSE, and SST. (10 points)

(d) Write the analysis of variance table. Include the expected mean squares column. Perform all tests at the $\alpha = 0.05$ level. (10 points)

(e) Find a 95% confidence interval for $\mu + \alpha_1$. Is $\mu + \alpha_1$ a contrast? (10 points)

(f) Find a 95% confidence interval for $\alpha_1 - \alpha_2$. Is $\alpha_1 - \alpha_2$ a contrast? (10 points)

(g) Assume that factor B is random. State the changes in the model assumptions in part (a) and estimate the variance components. (10 points)