Applied Statistics **Preliminary Examination** August 2013

Do all three problems. The asterisk * indicates the transpose of a vector or matrix, $\oplus$ is the symbol for direct sum (of two subspaces of a vector space), and $\otimes$ is the symbol for tensor product (of two vectors or vector spaces). For any $m \in \mathbb{N}$, $\mu \in \mathbb{R}^m$, and $m \times m$ matrix Σ, the notation $N_m(\mu, \Sigma)$ refers to an m-variate normal distribution with mean μ and covariance matrix Σ. The symbol I_n stands for $n \times n$ identity matrix.

1. Let A be an $n \times q$-matrix of full rank $q \in \mathbb{N}$, and B an $r \times n$-matrix of full rank $r \in \mathbb{N}$, with $q < r < n$. Introduce the subspaces $U = \{x = Ay : y \in \mathbb{R}^q\}$ and $V = \{x \in \mathbb{R}^n : Bx = 0\}$. Assume that $\mathcal{R}(A) \subset \mathcal{R}(B^*)$.

   (a) Show that $U \perp V$.

   Consider the coordinate-free linear model $Y = \eta + E$, $\eta \in \mathbb{L} = U \oplus V$, where $E \stackrel{d}{=} N_n(0, \sigma^2 I_n)$, $0 < \sigma^2 < \infty$. The null hypothesis is given by $H_0 : \eta \in \mathbb{L}_0 = V$.

   (b) Determine $\mathbb{L}_1 = \mathbb{L}_0^\perp \cap \mathbb{L}$.

   (c) Express the orthogonal projections P, P_0, and P_1 on $\mathbb{L}$, $\mathbb{L}_0$, and $\mathbb{L}_1$ respectively in terms of the matrices A and B.

   (d) Compute the test statistic for testing $H_0 : \eta \in \mathbb{L}_0$, exploiting the results in 1(c).

   (e) What is the distribution of this test statistic under H_0?

2. Let $m, n \in \mathbb{N}$ and $m < n$. Consider the vectors $1_n = (1, \ldots, 1)^*$, $x = (x_1, \ldots, x_n)^*$, $s = (s_1, \ldots, s_m, 0, \ldots, 0)^*$, and $t = (0, \ldots, 0, t_{m+1}, \ldots, t_n)^* \in \mathbb{R}^n$. None of these vectors is the zero vector and we assume the subspaces $[1_n]$, $[x]$, and $[s, t]$ to be mutually orthogonal. Consider the regression model

$$Y_i = \alpha + \beta x_i + \gamma s_i + \delta t_i + E_i, \tag{1}$$

   $i = 1, \ldots, n$, where the $E_1, \cdots, E_n$ are independent and identically distributed $N(0, \sigma^2)$ for some $0 < \sigma^2 < \infty$. This model can be concisely written in the form $Y = X\Theta + E$ for a certain design matrix X, where $Y = (Y_1, \cdots, Y_n)^*$, $E = (E_1, \cdots, E_n)^*$, and $\Theta = (\alpha, \beta, \gamma, \delta)^*$.

   (a) Write X as a partitioned matrix, exploiting the vectors 1_n, x, s, t.

   (b) Compute $(X^*X)^{-1}$.

   (c) Compute the maximum likelihood estimators of α, β, γ, δ and σ^2.

   (d) Compute the test statistic for testing $H_0 : \gamma = \delta = 0$.

   (e) Find a confidence interval (of arbitrary level) for the linear parameter function $\beta - \gamma - \delta$.

Notational convention. The following notation will be useful: for any $m \in \mathbb{N}$ write $1_m = (1, \ldots, 1)^* \in \mathbb{R}^m$, $P_m = \frac{1}{m} 1_m 1_m^*$ for the orthogonal projection onto $[1_m]$, and $P_m^\perp$ for the orthogonal projection onto $[1_m]^\perp \cap \mathbb{R}^m$.

3. Consider the ANCOVA model

$$Y_{ijk} = \mu + \gamma_{ij} + \delta_k + \xi\, u_i\, v_j + E_{ijk}, \tag{2}$$

$(i = 1, \cdots, I; j = 1, \cdots, J; k = 1, \cdots, K;\ I,\ J,\ K \geq 2)$, where μ, γ_{ij}, δ_k and ξ are real parameters, $u = (u_1, \cdots, u_I)^*$, $v = (v_1, \cdots, v_J)^*$ given vectors, and the E_{ijk} are independent and identically distributed with normal distribution $N(0, \sigma^2)$ for some $0 < \sigma^2 < \infty$. We impose the conditions:

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \gamma_{ij} = 0, \quad \sum_{k=1}^{K} \delta_k = 0, \quad \sum_{i=1}^{I} u_i = 0, \quad \sum_{j=1}^{J} v_j = 0. \tag{3}$$

Let us write $\Gamma = (\gamma_{ij})_{\substack{i = 1, \cdots, I \\ j = 1, \cdots, J}}$ and $d = (\delta_1, \cdots, \delta_K)^*$. Using tensor products the model says that Y (the vector of all Y_{ijk}) lies in a subspace $\mathbb{L}$ of $\mathbb{R}^I \otimes \mathbb{R}^J \otimes \mathbb{R}^K$.

(a) Give a concise description of the model using tensor products.

(b) The subspace $\mathbb{L}$ corresponding to the model can be naturally decomposed as $\mathbb{L} = \mathbb{L}_1 \oplus \mathbb{L}_2 \oplus \mathbb{L}_3 \oplus \mathbb{L}_4$. Specify the subspaces $\mathbb{L}_i$ $(i = 1, \cdots, 4)$.

(c) Suppose we want to test the null hypothesis $H_0 : \xi = 0$, i.e. the model is a pure ANOVA model. Describe the subspace $\mathbb{L}_0$ of $\mathbb{L}$ corresponding to the null hypothesis model.

(d) Compute the projection of Y onto the subspace $\mathbb{L}_0^\perp \cap \mathbb{L}$.

(e) Compute the LR test statistic for testing H_0 and give its distribution under the null hypothesis.

Design of Experiments: Prelim Problems
August 2013

Please Do All the Problems

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

1. The surface finish of metal parts made on four machines is being studied. An experiment is conducted in which each machine is run by three different operators and two specimens from each operator are collected and tested. Because of the location of the machines, different operators are used on each machine, and the operators are chosen at random. The summary data follow.

Machine	Mean
1	68.667
2	83.167
3	62.500
4	49.000

SS(Total) = 7449.33, SSE = 1014.00

(a) Write a linear model for the data. Make sure to clearly define the terms as they relate to this design and state the assumptions.

(b) Write the ANOVA table and include the $E(MS)$ column. Note: if you are unable to calculate $SS(Machine)$ take it to be $SS(Machine) = 3600.00$.

(c) Test for machine effects at the 0.05 level of significance and clearly state your conclusion.

(d) Perform pairwise comparisons of the surface finish measures between the four machines, using Tukey's method. Use $\alpha = 0.05$.

(e) Under the set-up of this experiment, the observations obtained by an operator assigned to a specific machine are correlated. Calculate the estimated value of this correlation coefficient.

(f) Let $\bar{Y}_{i\bullet\bullet}$ denote the mean response for the i-th machine ($i = 1, \cdots, 4$). Calculate the estimated value of $\mathrm{Var}(\bar{Y}_{i\bullet\bullet})$.

(g) Let $\tau_i,\ i = 1, \cdots, 4$, be the main effects of machines. Test $H_0 : \tau_2 - (\tau_1 + \tau_3 + \tau_4)/3 = 0$ vs. $H_1 : \tau_2 - (\tau_1 + \tau_3 + \tau_4)/3 > 0$ at the 0.05 level of significance.

2. Twenty athletes are randomly split into 4 groups, with 5 athletes in each (using a method that makes all such splits equally likely). Three groups are assigned to drink one of three different brands of energy drink, and the fourth group is given none. An hour later, all athletes compete in a 2 mile running race, and their times Y (minutes) to complete the race are recorded. On the previous day, all athletes had run the same 2 mile race without consuming energy drinks, and their completion times X (minutes) were recorded. Analysis of the variance of Y in terms of the group and the covariate X (but not their interaction) produces the following sequential (Type I) sums of squares:

Source	Sequential SS
Group	8.13
Previous Days Time (X)	16.16

Source	Sequential SS
Previous Days Time (X)	17.63
Group	6.66

The sum of the squared residuals is 4.13. (Note the different order of the sources in the two tables.)

(a) Write a linear model equation for the analysis using the covariate, clearly defining all symbols and stating any conditions they satisfy.

(b) Test whether energy drink consumption affects running time. (Your test should account for the effect of the covariate X). Clearly state your conclusion. Use $\alpha = 0.05$.

(c) Perform the same test, but this time ignore the covariate (i.e. based on the simple one-way analysis of variance, as if X had not been recorded). Clearly state your conclusion. Use $\alpha = 0.05$.