Applied Statistics **Preliminary Examination** May 2012

1. Consider the coordinate-free linear model, $Y = \eta + E$, $\eta \in \mathbb{U}$, where $\mathbb{U} \subset \mathbb{R}^n$ is a linear subspace of dimension $0 < \dim \mathbb{U} = p < n$, and $E \stackrel{d}{=} N_n(0, \sigma^2 I_n)$, $0 < \sigma^2 < \infty$. The null hypotheseis is given by $H_0 : \eta \in \mathbb{U}_0$, where $0 \le \dim \mathbb{U}_0 = p_0 < p$. Define $\mathbb{U}_1 = \mathbb{U}_0^{\perp} \cap \mathbb{U}$.
   (a) Let $a, b \in \mathbb{R}^n$ be such that $||a|| = ||b|| = 1$ and $a \perp b$, and let $\mathbb{U} = \{x \in \mathbb{R}^n : a^* x = 0\}$, $\mathbb{U}_0 = \{a^* x = 0, b^* x = 0\}$. Give an explicit expression for the projections P, P_0, and P_1 on the subspaces $\mathbb{U}, \mathbb{U}_0$, and $\mathbb{U}_1$, respectively.
   (b) Compute the test statistic for testing H_0 for the subspace $\mathbb{U}_0$ as in part (a), and specify its distribution under H_0.
   (c) For each $h \in \mathbb{U}_1$ construct a confidence interval for $h^* \eta_1, \eta_1 \in \mathbb{U}_1$, in such a way that these intervals have simultaneous level $1 - \alpha$, for some $0 < \alpha < 1$.

2. Consider the model $Y_{ij} = \mu + \alpha_i + E_{ij}$, for $i = 1, 2, 3$, and $j = 1, 2$, where $\mu \in \mathbb{R}$, $\alpha_1 \in \mathbb{R}$, and the E_{ij} are independent and identically distributed with common distribution $N(0, \sigma^2)$, $0 < \sigma^2 < \infty$.
   (a) Express $Y = (Y_{11}, \ldots, Y_{32})^*$ in terms of the vector $\theta = (\mu, \alpha_1, \alpha_2, \alpha_3)^*$ of parameters by means of a design matrix X.
   (b) Is this design matrix of full rank? Why?
   (c) If not, find a generalized inverse for $X^* X$.
   (d) Compute the least squares estimator $\hat{\theta}$ of θ using this generalized inverse.
   (e) Is α_1 estimable? Why?
   (f) Is $\alpha_1 - \alpha_2$ estimable? Why?
   (g) Find a 95% confidence interval for $\alpha_1 - \alpha_2$.
   (Remark: it suffices to write the generalized inverse of $X^* X$ in the form M^{-1} with M properly specified.)

3. Let X be an $n \times p$ $(1 \le p < n-1)$ design matrix of full rank. Write $1_n = (1, \ldots, 1)^* \in \mathbb{R}^n$ and $P_n = \frac{1}{n} 1_n 1_n^*$ for the projection onto the line $[\![1_n]\!]$. Assuming that $1_n \notin \Re(X)$, where $\Re(X)$ is the range of X, consider the standard regression model

$$Y = (1_n | X) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + E,$$

and its canonical version

$$Y = (1_n | X - P_n X) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + E,$$

where $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p$, and $E \stackrel{d}{=} N_n(0, \sigma^2 I_n)$, $0 < \sigma^2 < \infty$. Let $\widehat{\beta}_S$ denote the least squares estimator of β in the standard and $\widehat{\beta}_C$ in the canonical model. Prove that $\widehat{\beta}_S = \widehat{\beta}_C$.

(Hint: you may use that

$$\left(\begin{array}{c|c} \mathrm{A} & \mathrm{B} \\ \hline \mathrm{C} & \mathrm{D} \end{array}\right)^{-1} = \left(\begin{array}{c|c} * & * \\ \hline * & M^{-1} \end{array}\right),$$

where $M = D - CA^{-1}B$.)

4. The two-way layout model with interactions is given by $Y_{jk\ell} = \mu + \alpha_j + \beta_k + \gamma_{jk} + E_{jk\ell}$, $1 \le j \le J, 1 \le k \le K, 1 \le \ell \le L$, under the restrictions $\sum_{j=1}^J \alpha_j = \sum_{k=1}^K \beta_k = 0, \sum_{j=1}^J \gamma_{jk} = 0$ for all k, $\sum_{k=1}^K \gamma_{jk} = 0$ for all j. In this model the μ, α_j, β_k, and γ_{jk} are all real, and the $E_{jk\ell}$ are all independent with the same $N(0, \sigma^2)$ distribution $(0 < \sigma^2 < \infty)$. Using tensor products, the model says that Y (the vector of all $Y_{jk\ell}$) lies in a subspace $\mathbb{U} \subset \mathbb{R}^J \otimes \mathbb{R}^K \otimes \mathbb{R}^L$.

   Let us introduce the notation

$$a = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_J \end{pmatrix}, b = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}, \Gamma = \begin{pmatrix} \gamma_{11} \cdots \gamma_{1K} \\ \vdots \\ \gamma_{J1} \cdots \gamma_{JK} \end{pmatrix},$$

   for the parameters. 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 projection on $[\![1_m]\!]$, and $P_m^\perp$ for the projection on $[\![1_m]\!]^\perp$; in this model these vectors and projections play a role for $m = J, K$, and L.

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

   (b) The subspace $\mathbb{U}$ corresponding to the model can be written as $\mathbb{U} = \mathbb{U}_1 \oplus \mathbb{U}_2 \oplus \mathbb{U}_3 \oplus \mathbb{U}_4$. Specify the $\mathbb{U}_j$.

   (c) Compute the projection of Y onto the subspace $[\![1_J]\!] \otimes [\![1_K]\!] \otimes [\![1_L]\!]$ and onto the subspace $[\![1_J]\!] \otimes [\![1_K]\!]^\perp \otimes [\![1_L]\!]$.

   (d) Are the projections in part (c) stochastically independent? Why?

   (e) Give a detailed computation of the statistic for testing $H_0 : \Gamma = 0$ (i.e. no interactions present), and specify its distribution under this null hypothesis.

Design of Experiments: Preliminary Examination
Spring 2012

Please Do All the Problems

1. Let Y_{ij} denote an observation from a balanced one-factor random effects model and let $\bar{Y}_{..}$ be the estimator of the overall mean.

   a) Find $\text{Cov}(Y_{ij}, Y_{i'j'})$, for $i \neq i' = 1, \cdots, a;\ j, j' = 1, \cdots, n$

   b) Find $\text{Cov}(Y_{ij}, Y_{ij'})$, for $i = 1, \cdots, a;\ j \neq j' = 1, \cdots, n$

   c) Derive $\text{Var}(\bar{Y}_{..})$

   d) Derive the expression for the $100(1-\alpha)\%$ confidence interval for μ.

2. In an agricultural experiment, the effects of two irrigation methods and two fertilizers on yield of a crop are being studied. The experiment is to be carried out on five randomly selected fields. The experiment is conducted as follows: irrigation methods are randomly assigned to the two plots on each field, the two fertilizers are then randomly assigned to two sub-plots on each plot.

   The following summary statistics are available:

	Irrigation Method		Fertilizer		Field					
	1	2	1	2	1	2	3	4	5	Overall
Mean	37.30	54.00	43.80	47.50	56.00	47.00	40.00	38.75	46.50	45.65

   SS(IrrMethod) = 1394.45, SS(Fertilizer) = 68.45, SS(Field) = 756.80,

   SSE = 12.00, SS(Total) = 2312.55

   a) Write the linear model for the design including all two factor interactions.

   b) Write the ANOVA table including the E(MS) column.

   c) Test whether the two factors interact at the 0.05 level of significance.

   d) Test for significance of treatment effects at the 0.05 level of significance.

3. The factors that influence the breaking strength of a synthetic fiber are being studied. Four production machines and three operators are chosen and a factorial experiment is run using fiber from the same production batch. Two readings are obtained for each combination of factors. Both factors, machines and operators are chosen at random. The following summary statistics are obtained:

   SS(operator) = 160.333 SS(machine) = 12.458

   SS(operator*machine) = 44.667 SSE = 45.500

   a) Write the appropriate linear model for the data.

   b) Write the ANOVA table including the $E(MS)$ column.

   c) Test the relevant hypotheses at the 0.05 level of significance each.

d) If a variance component is statistically significant, then find its unbiased estimate.

4. A production engineer studied the effects of machine model and operator on the output in a bottling plant. Three bottling machines were used, each a different model. Twelve operators were chosen at random. Four operators were assigned to a machine and worked six-hour shifts each. Data on the number of cases produced by each machine and operator per hour were collected for five days. The following summary data were obtained:

Mean (*Cases of Bottles per Hour*)				
	Machine			
Operator	1	2	3	
1	61.80	75.80	76.80	
2	67.80	75.20	69.60	
3	62.60	55.80	74.40	
4	52.60	77.00	73.40	
Overall	61.20	70.95	73.55	68.57

SS(Machine) = 1695.633 SSE = 1132.800 SSTotal = 5100.733

a) Write the appropriate linear model for the data.

b) Write the ANOVA table including the $E(MS)$ column.

c) Test the relevant hypotheses at the 0.05 level of significance each.

d) Calculate the main effects of machines.

e) Conduct pairwise comparisons of the means for machines using Tukey's simultaneous confidence intervals.

f) Estimate the operator variance component.