UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
ADVANCED EXAM - LINEAR MODELS
Monday, August 27, 2012

Work all problems; 75 points are required to pass.

1. (*25 points*) Let $\mathbf{Y}$ be an $n \times 1$ random vector, $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is an $n \times p$ known matrix with non-full rank r $(1 \leq r < p)$, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters. We assume that elements in $\mathbf{Y}$ are uncorrelated. Let $\psi = \mathbf{a}'\boldsymbol{\beta}$ for some p-vector $\mathbf{a}$.

   (a) Give the definition for ψ to be estimable, in terms of an expected value involving $\mathbf{Y}$.

   (b) Derive (with justifications) a necessary and sufficient condition for $\mathbf{a}$ so that $\mathbf{a}'\boldsymbol{\beta}$ is estimable in terms of the matrix $\mathbf{X}$ and how it relates to $\boldsymbol{a}$. You need to show both necessity and sufficiency.

   (c) Write down the normal equation for finding least squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Explain why the solution to the normal equations is not unique. Recall that $\mathbf{X}$ is not of full column rank.

   (d) State carefully what constraints must be put on $\boldsymbol{\beta}$ in order to force a unique solution to the normal equations. You don't need to prove that this yields a unique solution but state carefully what the conditions are.

   (e) Suppose that $\mathbf{a}'\boldsymbol{\beta}$ is an estimable function of $\boldsymbol{\beta}$. State the Gauss-Markov theorem concerning the "best estimator" of $\mathbf{a}'\boldsymbol{\beta}$ and, in light of the fact that there are many solutions to the normal equations, how you would compute the best estimator of $\boldsymbol{a}'\boldsymbol{\beta}$.

   (f) In the above discussions, does $\mathbf{Y}$ need to have a multivariate normal distribution? Use one or two sentences to explain your answer.

2. (*25 points*) Consider the general linear model $Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $E(\boldsymbol{\epsilon}) = 0$ and $D(\boldsymbol{\epsilon}) = \sigma^2 I$, where $\mathbf{X}$ is $n \times p$ of rank p.

   Let $r_i = Y_i - \hat{Y}_i = Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ be the ith residual, where $\widehat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\beta}$.

   (a) Show that $\sum_i \hat{Y}_i(Y_i - \hat{Y}_i) = 0$. (Hint: much work in terms of vectors and matrices, first writing an expression for the vector of residuals.)

   (b) Now consider the linear model with an intercept term included explicitly. That is, the first column of $\mathbf{X}$ is all 1's so $Y_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}^* \beta_j + \epsilon_i$, $i = 1, \ldots, n$, with the same assumptions on the ϵ's as above. Show that in this case $\sum_i r_i = 0$ and

$$\sum_i (Y_i - \overline{Y})^2 = \sum_i (\hat{Y}_i - \overline{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

   (c) Returning to the general linear model at the start of the problem, let $\boldsymbol{\theta} = \mathbf{H}\boldsymbol{\beta}$, where $\mathbf{H}$ is $q \times p$ of rank q and

$$Q = (\mathbf{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta})'(\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}')^{-1}(\mathbf{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta})/\widehat{\sigma}^2.$$

i. What is the distribution of Q? State what general result you are applying and how it applies here. (You can use the fact that $(n-p)\widehat{\sigma}^2/\sigma^2$ is distributed chi-square with $n-p$ degrees of freedom and $\widehat{\sigma}^2$ is independent of $\widehat{\boldsymbol{\beta}}$.)

ii. In order to generate simultaneous Scheffe confidence intervals, Q can be written as a maximum over all $q \times 1$ vectors $\mathbf{c} \neq \mathbf{0}$ of $f(\mathbf{c}, \widehat{\boldsymbol{\beta}}, \mathbf{H}, \boldsymbol{\theta}, \mathbf{X}, \widehat{\sigma}^2)$ for some function f. Write out what that function is and then show carefully how this, along with the distributional result of the previous part, can be used to generate simultaneous confidence interval for all linear combinations of the form $\mathbf{c}'\boldsymbol{\theta}$.

3. (*25 points*) Consider the two-factor model with one observation per cell with

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, i = 1 \text{ to } A, j = 1 \text{ to } B,$$

with no restrictions put on the α_i's and β_j's.

(a) Write out the normal equations for estimating the parameters (hint: you can do this without having to write out $\mathbf{X}'\mathbf{X}$ explicitly).

(b) Characterize what linear combinations of the parameters are estimable. Use this to argue that $\theta_{ij} = \mu + \alpha_i + \tau_j$ is estimable.

(c) Use your previous result to prove what linear combinations of the α_i's are estimable.

(d) Explain what constraints you can use to force a unique solution to the normal equations? (Note: the constraints are not unique).

(e) Use constraints to find a unique solution to the normal equations and from this obtain the BLUE of θ_{ij} as given in part b).

(f) Use your solution to the previous part to find an explicit expression for SSE and then for $\widehat{\sigma}^2$.

4. (*25 points*) Consider the one-factor random effects model $Y_{ij} = \mu + A_i + \epsilon_{ij}$ where $i = 1, \ldots I > 2$, $j = 1, \ldots, J > 2$, μ is a fixed unknown parameter and the A_i and ϵ_{ij} are independent normal random variables with mean 0, $Var(A_i) = \sigma_A^2 > 0$ and $Var(\epsilon_{ij}) = \sigma_\epsilon^2 > 0$.

Define

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_I \end{bmatrix} \text{ with } \mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iJ} \end{bmatrix}.$$

Let

$$F = \frac{S_1^2/(I-1)}{S_2^2/I(J-1)}$$

be the usual F-statistic for testing $H_0 : \sigma_A^2 = 0$, where $S_1^2 = J\sum_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$ and $S_2^2 = \sum_i \sum_j (\bar{Y}_{ij} - \bar{Y}_{i.})^2$.

(a) Find $Cov(\mathbf{Y})$ which you can give by describing $Cov(\mathbf{Y}_i)$ for each i and $Cov(\mathbf{Y}_i, \mathbf{Y}_k)$ for $i \neq k$.

(b) Find the distribution of each of S_1^2 and S_2^2. You can do this by getting constants c_1 and c_2 so that $c_k S_k^2$ has a well known distribution, being sure to state the parameters involved.

If for some reason you cannot get the above part you can "buy" the answer, that is get it from the exam monitor, but give up the points on that question. This will let you continue on.

(c) Argue that S_1^2 and S_2^2 are independent. (Hint: First prove that $\bar{Y}_{i.} - \bar{Y}_{..}$ has 0 covariance with $\bar{Y}_{ij} - \bar{Y}_{i.}$.)

(d) Use the previous two parts to find the distribution of F, which again you could describe by getting a constant c so that cF follows a known distribution (state the parameters involved).

(e) Use the result of (c) to first derive a test of size α of $H_0 : \sigma_A^2/\sigma_\epsilon^2 = g$ versus $H_A : \sigma_A^2/\sigma_\epsilon^2 \neq g$ and then use this to derive a $100(1-\alpha)\%$ confidence interval for $\sigma_A^2/(\sigma_A^2 + \sigma_\epsilon^2$.

(f) Now special the previous part to get a test of $H_0 : \sigma_A^2 = 0$ and show how you would get the power function for this test. Your power calculation can be left in the form of an integral.