UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
ADVANCED EXAM - LINEAR MODELS
August 31, 2011

Work all problems. 70 points are needed to pass.

1. (10 points) Consider the usual linear model $Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, $E(\boldsymbol{\varepsilon}) = \underset{\sim}{0}$ and $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I$ $(\sigma > 0)$. X is a known $n \times p$ matrix with rank r $(1 \leq r \leq n)$.

    (a) Define estimability of a linear function $l'\boldsymbol{\beta}$.

    (b) State the theorem (called Gauss-Markov theorem) about obtaining the MVU linear estimator for an estimable linear function of $\boldsymbol{\beta}$. (Specify the notations that you used in your statement.)

2. (30 points) This problem concerns data from a simple linear regression model of the form:

$$\text{Given } X_i = x_i, \text{ then } Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i \in 1 \ldots n.$$

We wish to estimate the parameter $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. We treat two conditions on the distribution of the X's and ϵ_i's.

For given estimate $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)'$, define the sum of squared errors as:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

   (a) **Condition A:** For the first three parts, consider $x_1, \cdots, x_n$ fixed predictors and $\epsilon_i \sim iid \;\; N(0, \sigma^2)$. Suppose we observe data $y_1, \ldots y_n$, and $x_1, \ldots x_n$.

      i. Under Condition A, derive estimates in $\widehat{\boldsymbol{\beta}}$ which minimize SSE. (It is easier to use matrix approach.)

      ii. Under Condition A, find the variance of $\widehat{\beta}_1$ in the previous part. Denote this variance $V(\widehat{\beta}_1 | x_1, \cdots, x_n)$.

      iii. Given a new unit with value X_{new} for the predictor, under Condition A, find a point prediction $\widehat{Y}_{new}$ for Y_{new}. What is the variance of $\widehat{Y}_{new}$?

      iv. What is the variance of $\widehat{Y}_{new} - Y_{new}$? Then, find a 95% prediction interval for Y_{new}.

      v. Suppose a new unit is observed, but the corresponding value of X_{new} is not observed. Under Condition A, is it possible to predict Y_{new}? If not, briefly explain why not. If so, what value would you predict?

   (b) **Condition B:** For the next three parts, consider $\varepsilon_i \sim iid \;\; N(0, \sigma^2)$ as above, but now

$$X_i \sim iid \;\; N(\mu_x, \sigma_x^2) \quad i \in 1 \ldots n.$$

   You may assume the ϵ_i's and X_i's are independent of each other. Suppose we observe data $y_1, \ldots y_n$, and $x_1, \ldots x_n$.

      i. Under Condition B, the estimates in $\widehat{\boldsymbol{\beta}}$ which minimize SSE are the same as those under Condition A. Why? Either show the algebra to demonstrate this or explain briefly.

ii. Under Condition B, give an expression for the variance of $\widehat{\beta}_1$ in terms of the conditional variance in (a, ii), $V(\widehat{\beta}_1|x_1, \cdots, x_n)$.

iii. Suppose a new unit is observed from the same model. Given value X_{new} associated with the new unit, under Condition B, find a point prediction $\widehat{Y}_{new}$ for Y_{new}.

iv. Suppose a new unit is observed, but the corresponding value of X_{new} is not observed. Under Condition B, is it possible to predict Y_{new}? If not, briefly explain why not. If so, what value would you predict? Would you expect the variance of $\widehat{Y}_{new}$ to be greater than or less than the variance of $\widehat{Y}_{new}$ in part (iii)? (with justification, but you do not need to calculate the variances.)

3. (60 points) Consider the two-way fixed effects model where $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, for $i = 1, \dots, I$ (indicating levels of factor A), $j = 1, \dots, J$ (indicating levels of factor B) and $k = 1, \dots, n_{ij}$. The μ_{ij} are fixed but unknown means and ε_{ijk} are assumed i.i.d. $N(0, \sigma^2)$. This is called the cell means model. Assume $n_{ij} > 2$ for at least one cell.

(a) Express this as a general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \varepsilon$, where $\mathbf{X}$ is a known design matrix and $\boldsymbol{\mu}$ contains the μ_{ij}'s. Then, using general results for the linear model (you can state what you are using but no need to prove them), find explicit expressions for

i. the least square estimators of the μ_{ij}'s.

ii. SSE and an unbiased estimator for σ^2.

(b) State a general result about the distribution of SSE multiplied by an appropriate term, and then use this result to derive a confidence interval for σ^2.

(c) Consider estimating $\theta = \sum_i \sum_j c_{ij}\mu_{ij}$.

i. Give the UMVU linear estimator, $\widehat{\theta}$, for θ and state what you know about the variance of $\widehat{\theta}$ and its distribution. (Again, stating and using general results without proof is okay.)

ii. Argue that $\widehat{\theta}$ is independent of $\widehat{\sigma}^2$.

iii. Use the previous parts to come up with a pivotal quantity involving θ and derive a confidence interval for θ.

(d) State Scheffe's general results for simultaneous confidence intervals, and apply it here to set-up simultaneous confidence intervals for all linear combinations of the cell means.

(e) Define what it means for there to be no interaction between factors A and B.

(f) Suppose we wanted to test for no interaction using the cell meansformulation. Write no interaction as $\mathbf{H}\boldsymbol{\mu} = \underline{0}$, and then show how to compute the F-statistic in general matrix form (involving $\widehat{\boldsymbol{\mu}}$,$\mathbf{H}, \mathbf{X}$ and $\widehat{\sigma}^2$). What is the distribution of F under the null hypothesis?

(g) The effects model specifies $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$, where the α's and β's are main effects and the γ's are interaction terms. Assuming all n_{ij} are equal (say, $n_{ij} = K > 2$), explain how to get to the effects model from the cell means model, including specifying any restrictions that exist on the α, β and γ.

(h) Using the effects model above, show how to derive the F-statistic for no interaction using the full-reduced model approach. Go as far as you can in writing out the sums of squares involved. Note that to do this, you'll have to describe how things get estimated in the null model with $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$.