Preliminary Examination: LINEAR MODELS

Answer all questions and show all work.

1. Three groups of n observations are fitted using the following fixed-effects model:

$$Y_{ij} = \mu + \theta_i + \epsilon_{ij},$$

where $\{\epsilon_{ij}\}$ are independent and identically distributed $N(0, \sigma^2)$ random variables, for $i = 1, 2, 3$ and $j = 1, \ldots, n$. To avoid identifiability issues, we set $\sum_{i=1}^{3} \theta_i = 0$ and remove θ_3 from the above formulation, that is, the parameters in our model are μ, θ_1, and θ_2.

   a. Specify the design matrix and compute the *correlation* between the least squares estimators (LSE) $\hat{\theta}_1$ and $\hat{\theta}_2$.

   b. After fitting the model, you obtain LSE $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\sigma}^2$. Construct $100(1-\alpha)\%$ confidence intervals for θ_1 and θ_2 based on these estimates, respectively.

   c. Show that a $100(1-\alpha)\%$ joint confidence region for θ_1 and θ_2 can be specified by

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq \frac{2\hat{\sigma}^2}{n} F(\alpha; 2, 3(n-1))$$

   where $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, and $F(\alpha, 2, 3(n-1))$ is the $1-\alpha$ quantile of an $F_{2,3(n-1)}$ distribution.

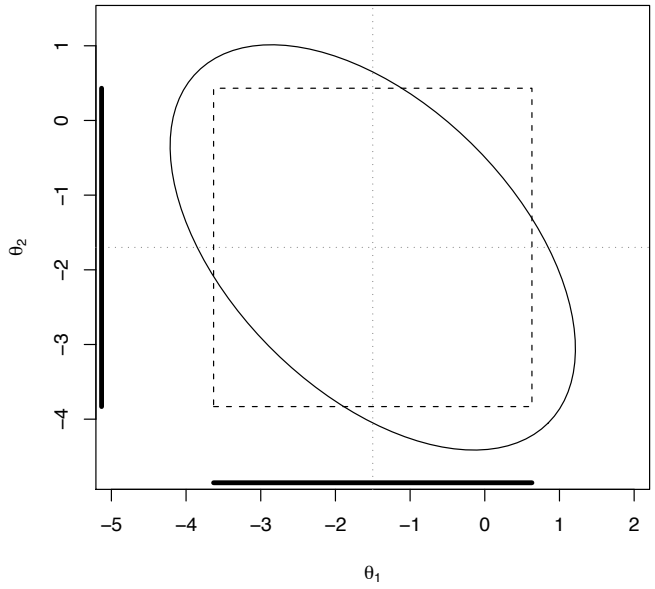
   d. The confidence regions from the two previous items are pictured below: the bold lines mark the separate confidence intervals, the dashed square is the (Cartesian) product of these intervals, and the ellipsoid represents the joint confidence region. If you were to test the hypothesis of no difference across groups, what would be your conclusions from this figure using: (i) the separate confidence intervals and (ii) the joint confidence region? Explain why your conclusions from (i) and (ii) are not consistent.

2. In the following, $\mathbf{I}_m$ is an $m \times m$ identity matrix, $\mathbf{0}_m$ is an $m \times 1$ vector of zero elements, and $\mathbf{J}_m = \mathbf{1}_m \mathbf{1}_m'$ where $\mathbf{1}_m$ is an $m \times 1$ vector of 1's. You may use, without proof, the fact that

$$[\mathbf{I}_m + \phi \mathbf{J}_m]^{-1} = \left[\mathbf{I}_m - \frac{\phi}{1 + m\phi} \mathbf{J}_m\right].$$

   a. Consider the model:

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(\mathbf{0}_n, \sigma^2 \mathbf{V})$$



where $\boldsymbol{\theta}$ is a vector of p unknown parameters, σ^2 is a constant, and $\mathbf{W}$ and $\mathbf{V}$ are known matrices of sizes $n \times p$ and $n \times n$, respectively. Give a formula for a joint confidence region for a set of contrasts $\mathbf{C}'\boldsymbol{\theta}$ using Scheffé's method. Here, $\mathbf{C}'$ is $q \times p$ of rank q (and $q \geq 1$).

b. Now consider the following linear model:

$$Y_{ijt} = \gamma_i + \tau_j + \epsilon_{ijt},$$

$$\epsilon_{ijt} \sim N(0, \sigma_E^2),\ \gamma_i \sim N(0, \sigma_\gamma^2),$$
$$i = 1, 2;\ j = 1, 2;\ t = 1, 2;$$

where all random variables on the right hand side are mutually independent. Write the model as $\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\tau} + \boldsymbol{\varepsilon}$, where

$$\mathbf{Y} = (Y_{111}, Y_{112}, Y_{121}, Y_{122}, Y_{211}, Y_{212}, Y_{221}, Y_{222})',$$

$$\boldsymbol{\gamma} = (\gamma_1, \gamma_2)', \quad \text{and } \boldsymbol{\tau} = (\tau_1, \tau_2)'$$

and find $\mathbf{Z}$, and $\mathbf{X}$. In addition, derive the variance-covariance matrix of $\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$.

c. State the distribution of $\mathbf{Y}$ and find the best linear unbiased estimator of $\boldsymbol{\tau}$. Assume that $\mathbf{C}'$ is $q \times p$ of rank q (and $q \geq 1$). Give a condition for $\mathbf{C}'\boldsymbol{\tau}$ to be estimable under the model and justify your answer.

d. For given constant vector $\mathbf{d}$ and estimable set of functions $\mathbf{C}'\boldsymbol{\tau}$, state a test statistic for testing

$$H_0 : \mathbf{C}'\boldsymbol{\tau} = \mathbf{d} \text{ versus } H_a : \mathbf{C}'\boldsymbol{\tau} \neq \mathbf{d},$$

where $\mathbf{C}'$ is $q \times p$ of rank q (and $q \geq 1$). Find the expected value of the numerator of the test statistic.

3. Consider a linear regression of the form

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

for $i = 1, \ldots, n$, where $\mathbf{x}$ consists of p predictors, and $\{\epsilon_i\}$ are independent with zero mean and constant variance σ^2. Suppose that this is just one model of many that you wish to compare, and hence you are interested in variable selection. We know that the error sum of squares,

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

is too optimistic as a measure of performance when comparing models. (Here the $\hat{Y}_i$ are the fitted values, based on OLS regression.) Instead, people define the so-called *in-sample prediction error*:

$$Err_{in} = \frac{1}{n}\sum_{i=1}^{n} E_{Y_i^*}(Y_i^* - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2$$

and consider $E_{\mathbf{Y}}(Err_{in})$ to be the preferred quantity of interest. Note: Expectation in the definition of Err_{in} is defined with respect to a new observation Y_i^* at case i. The overall prediction error $E_{\mathbf{Y}}(Err_{in})$ is then defined with respect to the original observations $\mathbf{Y} = (Y_1, \ldots, Y_n)'$.

a. Show that

$$E_{\mathbf{Y}}(Err_{in}) = E_{\mathbf{Y}}(SSE) + \frac{2}{n}\sum_{i=1}^{n} Cov(\hat{Y}_i, Y_i).$$

b. Show that

$$\frac{2}{n}\sum_{i=1}^{n} Cov(\hat{Y}_i, Y_i) = \frac{2p}{n}\sigma^2.$$

c. Explain the relation of the Mallows' C_p statistic

$$C_p = SSE + \frac{2p}{n}\hat{\sigma}^2$$

to the quantity $E_{\mathbf{Y}}(Err_{in})$ and why this suggests C_p as a reasonable criterion for variable selection. Note here $\hat{\sigma}^2$ is the usual unbiased estimate of σ^2 under the full model.

4. Consider the model for three stage nested design,

$$Y_{ijkl} = \mu + \tau_i + \beta_{j(i)} + \gamma_{k(ij)} + \epsilon_{ijkl},$$

$$i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, c;\ l = 1, \ldots, n;$$

where τ_i is the effect of A, $\beta_{j(i)}$ is the effect of B within A, $\gamma_{k(ij)}$ is the effect of C within A and B. Assume that all three factors are random. Construct the ANOVA table including the form of Sum of Squares (SS, e.g., $SST = \sum_i \sum_j \sum_k \sum_l (Y_{ijkl} - \bar{Y}_{....})^2$), degrees of freedom (df), Mean Squares (MS), and Expected Mean Squares (EMS). Also obtain the formulas for estimating the variance components.