MA 575 – Qualifying Exam

Spring 2012

1. You are given the results from fitting a linear regression with the same response and predictors to two datasets, $(X_1, \mathbf{y}_1)$ and $(X_2, \mathbf{y}_2)$, each with the same number n of observations:

$$\mathbf{y}_i = X_i\beta + \mathbf{e}_i, \quad \mathbf{e}_i \sim N(0, \sigma^2 I_n), \quad i = 1, 2.$$

Unfortunately, the original datasets were lost, but from the results you have the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ from datasets 1 and 2, respectively, and, from their sampling covariances, $(X_1^T X_1)^{-1}$ and $(X_2^T X_2)^{-1}$. Since both datasets aim to estimate the same coefficients, you wish to combine the datasets. To this end, you decide to obtain the (linearly) combined estimate

$$\hat{\beta}_\lambda = \lambda\hat{\beta}_1 + (1-\lambda)\hat{\beta}_2$$

where $0 \le \lambda \le 1$.

(a) Show that $\hat{\beta}_\lambda$ is an unbiased estimator for β for any λ.

(b) Find $\mathrm{Var}[\hat{\beta}_\lambda \,|\, X_1, X_2]$ as a function of λ, σ^2, X_1, and X_2.

(c) You have now observed $\mathbf{x}^*$ and want the *fitted* value $\hat{y}^*$ based on $\hat{\beta}_\lambda$, that is, $\hat{y}^* = \hat{\beta}_\lambda^T \mathbf{x}^*$. Similarly, define $\hat{y}_1^*$ and $\hat{y}_2^*$ as the fitted values based on $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.
If $V_i = \mathrm{Var}[\hat{y}_i^* \,|\, X_i, \mathbf{x}^*]$ is the variance of the fitted value for $\mathbf{x}^*$ based on $\hat{\beta}_i$, for $i = 1, 2$, show that the value λ^* that *minimizes* the variance $V^* = \mathrm{Var}[\hat{y}^* \,|\, X_1, X_2, \mathbf{x}^*]$ of $\hat{y}^*$ is $V_2/(V_1 + V_2)$. As a consequence, also show that the best precision—the inverse of the variance—for the fitted value is the sum of the precisions for each dataset, $1/V^* = 1/V_1 + 1/V_2$.

(d) Defining λ^* as in the previous item gives you the best linear combination of $\hat{\beta}_1$ and $\hat{\beta}_2$ when estimating the fitted value for $\mathbf{x}^*$. What if you wanted a linear combination that is best regardless of the observation?
To achieve this goal, you decide to really pool the datasets together and obtain a least-squares estimator $\hat{\beta}$ by regressing

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \mathbf{e},$$

where $\mathbf{e} \sim N(0, \sigma^2 I_{2n})$. A good friend reminds you of the Searle identity that would enable you to write

$$\begin{aligned}(X_1^T X_1 + X_2^T X_2)^{-1} &= (X_1^T X_1)^{-1}\big[(X_1^T X_1)^{-1} + (X_2^T X_2)^{-1}\big]^{-1}(X_2^T X_2)^{-1} \\ &= (X_2^T X_2)^{-1}\big[(X_1^T X_1)^{-1} + (X_2^T X_2)^{-1}\big]^{-1}(X_1^T X_1)^{-1} \quad \text{(by symmetry)}\end{aligned}$$

Show, using the identity above, that the LSE $\hat{\beta}$ can then be written as

$$\hat{\beta} = \Lambda\hat{\beta}_1 + (I_p - \Lambda)\hat{\beta}_2, \qquad (*)$$

with $\Lambda = (X_2^T X_2)^{-1}\big[(X_1^T X_1)^{-1} + (X_2^T X_2)^{-1}\big]^{-1}$.

(e) Inspired by $(*)$, how would you obtain λ such that $\hat{\beta}_\lambda$ is as "close" as possible to $\hat{\beta}$? (Note: you do not need to fully solve the problem; just comment on your reasoning, how you would set it up and so on.)

2. Based on data from an orthogonal design—say, you are fitting a polynomial regression with an orthogonal basis—you want to select a subset of predictors. More specifically, the dataset contains n observations: a response $\mathbf{y}$ and p *orthogonal* predictors in $X = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_p \end{bmatrix}$, where the $\mathbf{x}_j$ are column vectors (for each predictor.) As usual, assume that $\mathbf{y} = X\beta + \mathbf{e}$ where $\mathbf{e} \sim N(0, I_n)$.

   (a) Show that the least-squares estimator $\hat{\beta}$ for β can be obtained component-wise using

   $$\hat{\beta}_j = \frac{\mathbf{x}_j^T \mathbf{y}}{\mathbf{x}_j^T \mathbf{x}_j},$$

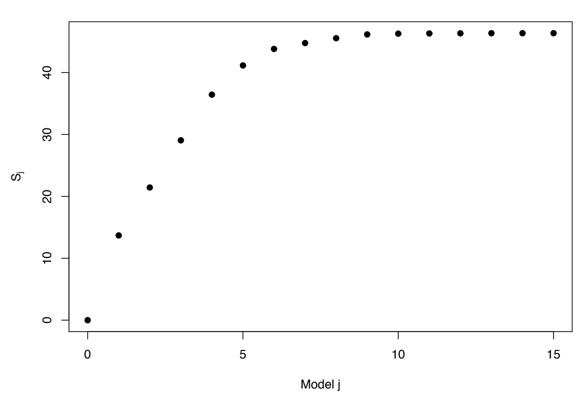
   for $j = 1, \ldots, p$.

   (b) Since RSS $= \hat{\mathbf{e}}^T \hat{\mathbf{e}} = \mathbf{y}^T \mathbf{y} - \hat{\beta}^T X^T X \hat{\beta}$, show that Mallow's C_p for a subset $\mathcal{C}$ of candidate predictors can be written as

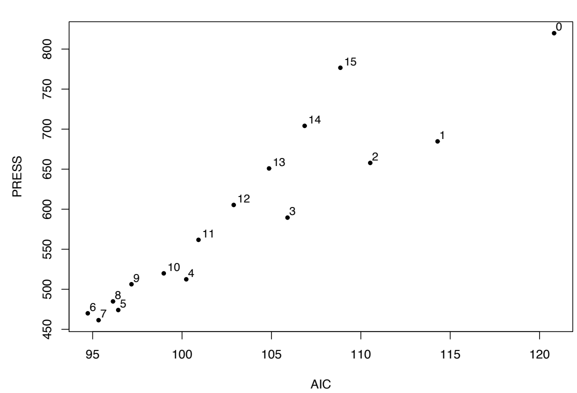
   $$C_p = \frac{\mathbf{y}^T \mathbf{y}}{\hat{\sigma}^2} - n - \sum_{j \in \mathcal{C}} \left(\frac{(\mathbf{x}_j^T \mathbf{x}_j)\hat{\beta}_j^2}{\hat{\sigma}^2} - 2 \right), \qquad (*)$$

   where $\hat{\sigma}^2$ is the least-squares estimate of σ^2 under the full model, that is, when $\mathcal{C} = \{1, \ldots, p\}$.

Based on the expression for C_p from the last item you decide to select a model in a greedy approach by first including the predictors that reduce C_p fastest: you order the predictors decreasingly by $R_j = \mathbf{x}_j^T \mathbf{x}_j \hat{\beta}_j^2 / \hat{\sigma}^2$. So, model 0 contains no predictor, model 1 contains only the predictor with largest R_j, model 2 contains the two predictors with largest R_j and so on. Now define $S_j = \sum_{k=1}^{j} R_k$; from your dataset you plot S_j against j, the number of predictors in the model, in Figure (a) below, in the left.



(a)



(b)

(c) What criterion would you use to select variables based on the expression $(*)$ if you wanted to minimize C_p? How would you use this criterion in Figure (a)?

(d) Still following the order defined by R_j, you now compute AIC and PRESS for each of the models from $j = 0$ to $j = p$. The AIC and PRESS scores are depicted in Figure (b) above, in the right. The labels next to each point list the model (the value of j.) If you were to select predictors based on these two criteria, would they agree? Explain and report the best model according to each criterion.

(e) Suppose now that you want to select predictors by using forward selection with *BIC* as criterion. How similar would your results be to the approach in the last item—using only the models defined by the order on R_j—with AIC as criterion? What if you wanted to use backward elimination instead? Explain.