Statistics 601 **November 6, 2019**

Assignment 3 — Due Nov 17, 2019

1. Consider the multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots \beta_{p-1} X_{p-1} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Suppose σ^2 is unknown and consider n independent observations from the model. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ denote the observed responses; let $\boldsymbol{X} = [\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{p-1}]$ denote the $n \times p$ design matrix, where $\boldsymbol{x}_i$ corresponds to the $(i+1)^{th}$ column of $\boldsymbol{X}$; and let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ denote the vector of regression coefficients. Assume that $\text{rank}(\boldsymbol{X}) = p$. Let $\widehat{\beta}_j$ denote the least squares estimate of β_j, and define $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_{p-1})'$.

   (a) Suppose $\boldsymbol{c} = (c_0, \ldots, c_{p-1})'$ is a vector of known constants. What is the distribution of $\sum_{j=0}^{p-1} c_j \widehat{\beta}_j$? Justify your answer, and find its mean and variance using matrix notation.

   (b) Consider the hypothesis $H_0 : \sum_{j=0}^{p-1} c_j \beta_j = h$ and $H_1 : \sum_{j=0}^{p-1} c_j \beta_j \neq h$, where h is a given constant. Explain how to test theses hypotheses at significance level α. Construct a suitable test statistic, find its distribution, and specify the rejection region.

   (c) Suppose we wish to predict the value of a future observation Y_{n+1}. Let $\boldsymbol{z} = (1, z_1, \ldots, z_{p-1})'$ denote its corresponding vector of predictor variables (i.e., $X_i = z_i$, for $i = 1, \ldots, p-1$), and consider the prediction $\widehat{Y}_{n+1} = \boldsymbol{z}\widehat{\boldsymbol{\beta}}$. Find the distribution of $Y_{n+1} - \widehat{Y}_{n+1}$.

   (d) Show that the MSE (mean square error) of the prediction $\widehat{Y}_{n+1} = \boldsymbol{z}\widehat{\boldsymbol{\beta}}$ is strictly greater than σ^2.

   (e) Given the vector $\boldsymbol{z}$ of predictor variables for the future observation Y_{n+1}, find an interval $\mathcal{I}$ such that $\mathbb{P}(Y_{n+1} \in \mathcal{I}) = 1 - \alpha$.

2. Echolocation in bats: These data, *bats*, have been used to study how do bats make their way about in the dark. Where *lenergy* is the log of the in-flight energy, *lmass* is the log of the body mass, and TYPE is the three-level factor represented by the indicator variables *bird* (which takes on a value of 1 if the type of species is a bird, and 0 if not) and *ebat* (which takes on a value of 1 if the species is an echolocating bat, and 0 if not). The third level of TYPE , non-echolocating bats, is treated here as the reference level, so its indicator variable does not appear in the regression model. You will create those two indicator variables (*bird* and *ebat*).

   Consider these three models:
   $\mu\{lenergy|lmass, TYPE\} = \beta_0 + \beta_1 lmass$
   $\mu\{lenergy|lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$
   $\mu\{lenergy|lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat + \beta_4 (lmass) \times (bird) + \beta_4 (lmass) \times (ebat)$

   (a) Explain why they can be described as representing the single line, parallel lines, and separate lines models, respectively.

   (b) Explain why the second model can be the "reduced model" in one F -test but the "full model" in another.

   (c) Estimate the FULL model and show how to use the extra-sum-of-squares F-test to test (a).

3. Brain Weights. The data are the average values of brain weight, body weight, gestation lengths (length of pregnancy), and litter size for 96 species of mammals. (Data from G. A. Sacher and E. F. Staffeldt, "Relation of Gestation Time to Brain Weight for Placental Mammals; Implications for the Theory of Vertebrate Growth," American Naturalist, 108 (1974): 593–613. The common names for the species correspond to the Latin names given in the original paper, and those followed by a Roman numeral indicate subspecies.)Identifying which mammals have larger brain weights than were predicted by the regression model might point the way to further variables that can be examined.

(a) Fit the regression of brain weight on body weight, gestation, and log litter size, using no transformations. Obtain a set of case-influence statistics. Is any mammal influential in this fit?

(b) Refit the regression without the influential observation, and obtain the new set of case influence statistics. Are there any influential observations from this fit? What are your main noticeable differences from the model in (a)?

(c) Consider all the data and fit the regression of log brain weight on log body weight, log gestation, and log litter size, and compute the studentized residuals.

(d) Which mammals have substantially larger brain weights than were predicted by the model?

(e) Do any mammals have substantially smaller brain weights than were predicted by the model?

(f) What lessons about the connection between the need for a log transformation and influence can be discerned?

4. A, B, and C are three explanatory variables in a multiple linear regression with $n = 28$ cases.

Model variables	**Residual sum of squares**	**Degrees of freedom**
None	8100	27
A	6240	26
B	5980	26
C	6760	26
AB	5500	25
AC	5250	25
BC	5750	25
ABC	5160	24

(a) Calculate the estimate of σ^2 for each model.

(b) Calculate the adjusted R^2 for each model.

(c) Calculate the Cp statistic for each model.

(d) Calculate the BIC for each model.

(e) Which model has (i) the smallest estimate of σ^2? (ii) the largest adjusted R^2 ? (iii) the smallest Cp statistic? (iv) the smallest BIC?

(f) Find the model indicated by forward selection. (Start with the model "None," and identify the single-variable model that has the smallest residual sum of squares. Then perform an extra-sum-of-squares F -test to see whether that variable is significant. If it is, find the two-variable model that includes the first term and has the smallest residual sum of squares. Then perform an extra-sum-of-squares F -test to see whether the additional variable is significant. Continue until no F -statistics greater than 4 remain for inclusion of another variable.)

(g) Calculate $exp(BIC - BIC_{min})$ for each model. Add these and divide each by the sum. What is the resulting posterior distribution on the models?