Tutorial 7

The questions for this tutorial have been revised directly from the class text "The Statistical Sleuth"

Last Updated: Wed Sep 20 18:19:30 2017

Question 3 in Tutorial 6 (Con'd, revised based on the exercise in Chapter 12 from class text "The Statistical Sleuth")

Blood-Brain Barrier. Please use "install.packages('Sleuth3')" and "library(Sleuth3)" to call the dataset in this problem. Using the data stored in the object "case1102" of the R library "Sleuth3", perform the following variable selection techniques to find a subset of the covariates-days after inoculation (Days), tumor weight (Tumor), weight loss (Loss), initial weight (Weight), and sex (Sex)-for explaining log of the ratio of brain tumor antibody count (Brain) to liver antibody count (Liver).

a) Cp plot among all subsets (showing at most 3 subsets for each size).

b) Adjusted R^2 plot among all subsets (showing at most 2 subsets for each size).

c) Forward selection by using AIC.

d) Backward elimination by using BIC.

Question 1 (revised based on the exercise in Chapter 12 from class text "The Statistical Sleuth")

Pollution and Mortality. Dataset is stored in the object "ex1217"" of the R library "Sleuth3". The 15 variables for each of 60 cities are (1) Precip: mean annual precipitation (in inches); (2) Humidity: percent relative humidity (annual average at 1 P.M.); (3) JanTemp: mean January temperature (in degrees Fahrenheit); (4) JulyTemp: mean July temperature (in degrees Fahrenheit); (5) Over65: percentage of the population aged 65 years or over; (6) House: population per household; (7) Educ: median number of school years completed by persons of age 25 years or more; (8) Sound: percentage of the housing that is sound with all facilities; (9) Density: population density (in persons per square mile of urbanized area); (10) NonWhite: percentage of 1960 population that is nonwhite; (11) WhiteCol: percentage of employment in white-collar occupations; (12) Poor: percentage of households with annual income under $3,000 in 1960; (13) relative pollution potential of hydrocarbons (HC); (14) relative pollution potential of oxides of nitrogen (NOX); and (15) relative pollution potential of sulphur dioxide (SO2). It is desired to determine whether the pollution variables (13, 14, and 15) are associated with mortality, after the other climate and socioeconomic variables are accounted for. (Note: These data have problems with influential observations and with lack of independence due to spatial correlation; these problems are ignored for purposes of this exercise.)

With mortality as the response, use Cp and adjusted R^2 plots among all subsets (showing at most 2 subsets for each size) to select a good-fitting regression model involving weather and socioeconomic

variables as explanatory, respectively. To the model with the lowest Cp (highest adjusted R^2), add the three pollution variables (transformed to their logarithms) and obtain the p-value from the extra-sum-of-squares F-test due to their addition. Based on the R output, are the pollution variables (13, 14, and 15) associated with mortality or not, after the other climate and socioeconomic variables are accounted for? Are the results different for Cp and adjusted R^2?

Question 2 (revised based on the exercise in Chapter 12 from class text "The Statistical Sleuth")

Pollution and Mortality. For the dataset in Question 1, please include all the 15 explanatory variables (the three pollution variables transformed to their logarithms) and the squares of them in the multiple linear regression model with mortality as the response. Perform the following variable-selection techniques to find a subset of the covariates. a) Cp plot among all subsets (showing at most 2 subsets for each size); b) forward selection by using BIC; c) stepwise regression by using AIC. Compare the selection results first and then use the R function "system.time()" to compare the computation time for each of the techniques.