Tutorial 6

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

QUESTION 1 (revised based on the exercise in Chapter 11 from class text "The Statistical Sleuth")

Does pollution kill people? Data in one early study designed to explore this issue came from 5 Standard Metropolitan Statistical Areas (SMSA) in the United States, obtained for the years 1959-1961. Total age-adjusted mortality from all causes (mortality), in deaths per 100,000 population, is the response variable. The explanatory variables include mean annual precipitation (in inches); median number of school years completed, for persons of age 25 years or older; percentage of 1960 population that is nonwhite; relative pollution potential of oxides of nitrogen, NO_x; and relative pollution potential of sulphur dioxide, SO_2. "Relative pollution potential" is the product of the tons emitted per day per square kilometre and a factor correcting for SMSA dimension and exposure. The data is contained in "pollution". [Data from: G.C McDonald and J.A. Ayers, "Some applications of the 'Chernoff Faces': A technique for graphically representing multivariate data", in Graphical Representation of Multivariate Data, New York, 1978]

a) Is there evidence that mortality is associated with either of the pollution variables, after the effects of the climate and socioeconomic variables are accounted for? To answer this question you must first decide upon an appropriate model. **[Do not worry about higher order terms or interactions in your modelling process.**] After finding an appropriate model make sure to look at the case-influence statistics.

b) Find a 95% confidence interval for the mean mortality for a city with the following characteristics precipitation=30, education=10, nonwhite=20, NO_x=1, and SO_2=1.

c) Conduct an appropriate hypothesis test(s) to determine whether the variables education and nonwhite provide important information about mortality, over and above what is explained by precipitation, NO_x and SO_2.

QUESTION 2 (revised based on the exercise in Chapter 10 from class text "The Statistical Sleuth")

The file "birds" contains measurements on breeding pairs of land bird species collected from 16 islands around Britain over the course of several decades. For each species, the dataset contains average time of extinction on those island where it appeared; the average number of nesting pairs; the size of the species (large or small); and the migratory status of the species (migrant or resident). It is expected that species with larger numbers of nesting pairs will tend to remain longer before becoming extinct. Of interest is whether, after adjusting for number of nesting pairs, size or migratory status has any effect. There is also some interest in whether the effect of size differs depending on the number of nesting pairs. Analyse the data to answer these two questions of interest. Your answer must include appropriate documentation.
[Data from: S.L. Pitman, H.L. Jones, and J. Diamond, "On the risk of extinction", American Naturalist 132 (1988): 757-85].

QUESTION 3 (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 stoblack 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) Forward selection by using F-statistic; (b) backward elimination by using F-statistic; (c) stepwise selection by using F-statistic. Are the above results the same?