Stat 359/Biol 563 Assignment 3 (28 marks)

1. In a study of patients with or without lung cancer, whether or not they smoked was recorded. A '1' is given if the person had lung cancer (case) and a '0' without. Similarly, a '1' indicates a smoker and a '0' indicates a nonsmoker. The data are found in the file lungcancer.csv, which is a comma deliminated file. To read this file in, you will have to use read.csv which specifies a ',' as the separator.

   Use table() to produce a contingency table and then determine whether there is an association between smoking and lung cancer.

   (a) (5 marks) State the hypotheses, test statistic and its distribution under the null hypothesis, p-value of the test, conclusion (in terms of evidence against H_0, and R codes.

   (b) (2 marks) What are the assumptions for the test in (a)? Are the assumptions valid?

2. (3 marks) Modify your function from your in class assignment to calculate AIC or AICc of a model depending on size of the sample. Inputs to the function are only a model object (eg. fit1, where fit1<-lm(y x)). Note that you can get at Sums of Squares by making an object b<-anova(fit1) and then accessing b$"Sum Sq" (found through names(b)). The number of data points is the same as the number of fitted values or the number of residuals (fit1$fitted.values, fit1$residuals). The number of parameters is the number of coefficients (fit$coefficients). Use your new AIC function for the rest of the assignment. Provide your R code for this function.

3. Anscombe produced 4 data sets which are located in anscombe.csv.

   (a) (2 marks) Produce 4 skatter plots (on the same page) of each of the data sets, describe each of these briefly.

   (b) (3 marks) Perform 4 linear regressions and produce a table (in your text editor) that shows the R^2 value as well as AIC (using your function from above). Discuss.

4. (5 marks) Beef consumption (in pounds per capita) in the United States between 1922 and 1941 are given in the data set beef.txt. Other variables of interest are beef price (in cents per pound divided by CPI), income (disposable income per capita in dollars divided by the CPI), and pork consumption (pounds per capita). [CPI= Consumer price index]. Find a model that "best" describes beef consumption in the United States. Complete a full analysis of the data (initial plots, model selection, residual plots etc.). Discuss your results.

5. Nurses employed in the emergency departments of four hospitals were asked to rate the quality of care provided by their facility. The scale ranged from 0 (worst) to 20 (best). The ratings are provided in hospital.csv.

   (a) (2 marks) Write out the model you will use to analyse the data including defining any dummy variables.

   (b) (2 marks) Do the true average ratings for each of the hospitals differ? Clearly state both the null and alternative hypotheses (in terms of your model coefficients) and make a formal conclusion based on your analysis.

(c) (2 marks) Plot the data in an informative way, does your plot support your conclusion?

(d) (2 marks) If you decide that the mean hospital qualities differ, use Tukey's multiple comparisons (TukeyHSD()) to come up with simultaneous confidence intervals to determine which means differ. Do any means differ?