Homework 2

STAT448 - Advanced Data Analysis

Due: Thursday, Sep 28, 2017 at 11:00 pm

To complete this assignment, you must use `Program_HW2_Data.sas` which is an existing SAS program file containing the data steps needed to analyze the data sets in the following problems.

1. (2 parts) Use the `diy` data set for this problem. This data set is a modification of a data set we will use in Chapter 8. You can read more of the background of the data there. The variables are included in the table below.

agegrp	age group; younger than 45 years or older than 45 years
response	Last year did you handle home improvements yourself that you would have previously hired someone to do? (yes or no)
n	count for particular response

   (a) Construct a contingency table for **agegrp** and **response** and comment on any apparent associations between age group and decision to hire someone to do home improvements.

   (b) Perform and comment on appropriate tests of association, and interpret the results.

2. (3 parts) Use the `crime100` data set to answer the questions below. This data set is a modification of the uscrime data set provided with the textbook. For more information regarding the data set, read Chapter 7 of the textbook. The crime100 data set contains the following variables:

Greater100	whether the crime rate is greater than 100 known offenses per million population (yes or no)
South	whether or not the state is in the South (yes or no)

(a) Construct a contingency table for **Greater100** and **South**. Comment on any apparent associations between crime rates greater than 100 crimes per million population, and whether or not the state is in the South.

(b) Perform and comment on appropriate tests of association, and interpret the results.

(c) Test (using risk differences) if Southern states have a significantly higher probability of crime rates greater than 100 crimes per million population than other states do.

3. (6 parts) For this problem use the bupa data set. Check the footnote [1] and UCI Machine Learning Repository for more information. The mean corpuscular volume and alkaline phosphatase are blood tests thought to be sensitive to liver disorder related to excessive alcohol consumption. Both association test and ANOVA will be performed on this dataset.

Variable Name	Description
mcv	mean corpuscular volume
alkphos	alkaline phosphatase
drinkgroup	categorization of the half-pint equivalents of alcoholic beverages drunk per day: group 1 $\sim$ less than 1 drink. group 2 $\sim$ at least 1 but fewer than 3 drinks. group 3 $\sim$ at least 3 but fewer than 6 drinks. group 4 $\sim$ at 6 but fewer than 9 drinks. group 5 $\sim$ 9 or more drinks.

(a) Create a new binary variable that contains two groups of **mcv** using the sample median as the cutoff. Subjects with mean corpuscular volume greater than or equal to the sample median are 1's while the rest are 0's. Construct a contingency table for the newly added grouping variable and **drinkgoup**. Comment on appropriate tests of association, and interpret the results.

(b) Now perform a one-way ANOVA for **mcv** as a function of drinking group. Use continuous **mcv** as in dataset. Comment on statistical significance of the model, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

(c) Compare the answers from (a) and (b). Do they give the same conclusion? What are the differences, if any, in the two approaches (contingency table analysis vs. ANOVA)?

[1] Bache, K. and Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

(d) Using the model in part (b) identify and comment on any significantly different groups. What can we infer about differences in mean corpuscular volume among the five drinking groups based on this analysis?

(e) Perform a one-way ANOVA for **alkphos** as a function of drinking group. Comment on statistical significance of the model, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

(f) Comment on whether there are more significant differences in alkaline phosphatase or mean corpuscular volume across the drinking groups.