STA303/1002H1S: Methods of Data Analysis II **Winter 2017**

Assignment 2

Out: Feb. 5, 2017 *Due: 10pm, Feb 26, 2017*

Reminder : You MUST write your solution independently and turn in your own write-up.

This assignment is due 10 :00pm, Sunday, Feb. 26th, 2017. Submit your solution as instructed by Crowdmark, namely, one pdf file for each question.

Late assignments will be subject to a deduction of 5% of the total marks for the assignment for each day late. Any late assignment after the day I post the solution will get zero mark.

Presentation of solutions is very important. No Rmarkdown solution template for the assignment, but you could simple create one from the template I posted for A1. You should produce a PDF file, split the PDF into different files to get your solution PDF for each question. Also, make sure the source R code at the end is complete. Marks will be deducted if the instructions herein are not followed.

Data 1 : Treatments for improving mental capcity

The file A3Q1.txt contains a sample of 30 subjects data. They were randomly assigned to three treatments/therapies for improving mental capacity. For each subject, a pretest and posttest measurements were recorded.

The variables in the dataset are :

- pretest : pretest measurements before experiment.
- posttest : pretest measurements at the end of experiment.
- trmt : 3 treatments/therapies.You should turn it into factor variable using *as.factor()*

Data 2 : The Donner party

In 1846, the Donner party (Donner and Reed families) left Springfield, Illinois for California in covered wagons. After reaching Fort Bridger, Wyoming, the leaders decided to find a new route to Sacramento. They became stranded in the eastern Sierra Nevada mountains at a place now called Donner Pass (right) when the region was hit by heavy snows in late October. By the time the survivors were rescued on April 21, 1847, 40 out of 87 had died.

Three variables in the dataset of donner.txt are :

- age : age of each subject (years)
- sex : 1=male ; 0=female
- survivorship : 1 = survived, 0 = dead.

Questions

Using R to do all the analysis on **Data 1** for the following questions.

Q1 (10 points)

(1a) (2 pts) Construct the one-way ANOVA analysis for comparing the three treatment means when pretest is ignored. (Show your code, ANOVA output and give your analysis conclusion).

(1b) (5 pts) From the one-way ANOVA in (1a), it involves a F-test for equality of means. Specify a model and a null hypothesis for no therapy effect, then give the formula of the F-ratio and its observed value from data. Use R code to find the critical value of this F-test using $\alpha = 5\%$. Compare the observed F value and the the critical value, what conclusion do you have ? Does it agree with your conclusion based on p-value in (1a) ?

(1c) (3 pts) What is the homogeneity of slopes assumption of ANCOVA ? Why is it important ?

Q2 (10 points)

(2a) (2 pts) Plot posttest versus pretest with a different symbol or color for each treatment. From this plot, does the assumption of homogenous slope look reasonable ?

(2b) (3 pts) Specify a model that can be use to access the homogenous regression slope assumption. Evaluate the assumption for this data. Is the homogenous slopes assumption met ? Evaluate the assumption for the ANCOVA performed to answer the previous question.

(2c) (3 pts) Fit an ANCOVA model to this data. Report the F-test for a treatment effect, after controlling for the effect of the pretest measurement. (Also show your R code and ANOVA output)

(2d) (2 pts) Find unadjusted and adjusted post-test score for 3 treatments.

Q3 (5 points)

(3a) (2 pts) If we define Y=posttest-protest as our new dependent variable, fit the one-way ANOVA model to it. How significant of the treatment effect ?

(3b) (3 pts) Compare the one-way ANOVA with the ANOCA model in Q-2c, which model do you prefer and why ?

Using R to do all the analysis on **Data 2** for the following questions.

Q4 (15 points)

(4a) (3 pts) Fit a logistic regression model to the data with covariates sex and age. Provide the summary output. Give the formula for the estimated curve. What is the fitted male model ? What is the fitted female model ?

(4b) (3 pts) Interpret the intercept and all the slopes from the full fitted model.

(4c) (2 pts)Plot the logistic regression curve versus age that has both male and female curves on it. Look at this plot, what conclusion do you have comparing the estimated survival probability for a male and a female given age =30 ?

(4d) (2 pts) What are the estimated probabilities of survival for men and women of ages 25 and 50 ?

(4e) (2 pts) What is the age at which the estimated probability of survival is 50% for women and for men ?

(4f) (3 pts) Why should one be reluctant to draw conclusions about the ratio of male and female odds of survival for the Donner Party members over 50 ?