**Dr. Zhankun Sun, Kevin**

**MS 5318 – Homework # 4**

**Due: Wednesday, Apr 13**

**Use R to answer the questions for this assignment. Include the R codes you used for each question.**

**Question 1**

**convenience.csv**

It’s rare that you’ll find a gas station these days that only sells gas. It’s become more common to find a convenience store that also sells gas. These data describe the sales over time at two franchise outlets of a major U.S. oil company. These particular stations sell gas, and each of them also has a convenience store. Each row in the data set summarizes sales for one day. The column labeled *Sales* gives the dollar sales of the convenience store, and the column *Volume* gives the number of gallons of gas sold. The column *Site* gives the location of the convenience store (categorical with two groups).

(a) The manager of the chain wants to compare the sales performance of the two service stations. Is it appropriate for him/her to conclude based on only a two-sample t-test on the sales? Would such a comparison be confounded by different levels of traffic (as measured by the volume of gas sold)?

(b) Perform the two-sample t-test to compare the sales of the two service stations. Summarize this analysis, assuming that there are no confounding variables. (You may run a linear regression to perform the two-sample t-test.)

(c) Compare the sales of the two stations while including *Volume* in the analysis. Summarize the comparison of sales based on this analysis. (Assume for the moment that the model meets the conditions for the multiple regression model.)

(d) Compare the results from parts (b) and (c). Do they agree? Explain why they agree or differ.

You should take into account the precision of the estimates and your answer to part (a).

**Question 2**

**UniversalBank.csv**

The data set contains information on 5000 loan applications. The response (*PersonalLoan*) is whether or not an offered loan had been accepted on an earlier occasion. The explanatory variables include

* *Age*: age of customer;
* *Experience*: professional experience in years;
* *Income*: income of customer;
* *Family*: family size of customer (1,2,3,4);
* *CCAvg*: average monthly credit card spending;
* *Mortgage*: size of mortgage;
* *Education*: three categories (1=undergraduate=1, 2=graduate, 3=professional).

(a) Run a logistic regression to predict the probability of loan acceptance using all the given predictors excluding their interactions. **Make sure that *Family* and *Education* are categorical variables.**

(b) With the model in part (a), predict the probability of loan acceptance for the following application:

Age=33, Experience=8, Income=85, Family=3, Education=2, Mortgage=0, CCAvg=1.1