DSC 423 - Data Analysis and Regression
Assignment 1 - Due on Wednesday September 26, 2023 by midnight
Total points: 43pts

Reading assignment
Review Multiple Linear Regression
- Lecture notes for weeks 1, 2 and 3
- Dummy variables and goodness of fit tests - Chapter 4: sections 4.8, chapter 5: sections 5.1, 5.2, 5.7 and 5.8
- R notes and examples

Review class examples.

For writing the answers to the problems, if R code is required, please write the R code fragment to answer the part of the problem. This R code will be run and if it works, then 1 pt is given.

PROBLEM 1 [19 pts]
The file banking.txt attached to this assignment provides data acquired from banking and census records for different zip codes in the bank's current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices. The data show
- median age of the population (AGE)
- median income (INCOME) in $
- average bank balance (BALANCE) in $
- median years of education (EDUCATION)

Use R to compute the analysis below. *All the functions are in the code for that we covered in class so far and in the exercises. This is the first assignment, and for many of you it may be the first time you use R. So if you run into an error, post a message on the discussion board or contact me. Make sure to include your code in the message.*

In this exercise you are asked to apply regression analysis techniques to describe the effect of age education and income on average account balance.

a) Analyze the distribution of average account balance using histogram, and compute appropriate descriptive statistics. Write a paragraph describing distribution of Balance and use appropriate descriptive statistics to describe center and spread of the distribution (you can review basic statistics concepts at: http://onlinestatbook.com/2/summarizing_distributions/summarizing_distributions.html) [2 pts = 1 pt R code + 1 pt answer]

b) Create scatterplots to visualize the associations between bank balance and the other variables. Discuss the patterns displayed by the scatterplot. Do the associations appear to be linear? (You can create scatterplots or a matrix plot). Do you see any outliers? [3 pts = 1 pt R code + 1 pt scatterplots + 1 pt answer]

c) Compute correlation values of bank balance vs the other variables. Interpret the correlation values, and discuss which pairs of variables appear to be strongly associated. [2 pts = 1 pt R code + 1 pt answer]

d) What's the independent variable and what is the dependent variable in this regression analysis? [1 pt]

e) Use R to fit a regression model to predict balance from age, education and income. Analyze the model parameters. Which predictors have a significant effect on balance? Use the t-tests on the parameters for alpha=.05. [2 pts = 1 pt R code + 1 pt answer]

f) If one of the predictors is not significant, remove it from the model and refit the new regression model. Write the expression of the fitted regression model. [2 pts = 1 pt R code + 1 pt answer]

g) Interpret the value of the parameters for the variables in the model. [1 pt]

h) Report the value for the R^2 coefficient and describe what it indicates. [1 pt]

i) According to census data, the population for a certain zip code area has median age equal to 34.8 years, median education equal to 12.5 years and median income equal to $42,401.
   - Use the final model computed in point (f) to compute the predicted average balance for the zip code area. [1 pt]
   - If the observed average balance for the zip code area is $21,572, what's the model prediction error? [1 pt]

j) Conduct a global F-test for overall model adequacy. Write down the test hypotheses and test statistic and discuss conclusions. [3 pts = 1 pt R code + 1 pt test hypotheses and test statistic + 1 pt conclusion]

PROBLEM 2 [4 pts]

A university career center collects information on the job status and starting salary of graduating seniors. Data recently collected over a two-year period included over 900 seniors who had found employment at the time of graduation. The information was used to model starting salary Y as a function of two qualitative independent variables: COLLEGE at four levels {Business, Engineering, Liberal Arts, Nursing} and SEX (male and female).

1. Define the dummy variables to include college (use Business as your baseline) in a regression model for starting salary Y [1 pt]
2. Write down the general regression model relating starting salary Y to both college and sex. [1 pt]
3. How would your model change if students in Engineering have the same starting salary as students in Business? [2 pts]

PROBLEM 3 [16pts]

A survey of IS managers was used to predict the yearly salary of beginning programmer analysts in a metropolitan area. Managers specified their standard salary (SALARY) in $1000 for a beginning programmer/analyst, the number of employees (NUMEMPL) in the firm's information processing staff, the firm's gross profit margin (MARGIN) in cents per dollar of sales and the firm's information processing cost (IPCOST) as a percentage of total administrative costs.

We'll fit a regression model to evaluate if starting salaries are affected by either the number of IS employees, IP costs and margin profits. The data is in salary_IS.txt.

1) Analyze the interrelationships between variables using scatterplots and correlation values. Is salary linearly related to the three predictors? Which variables are more strongly related? [3 pts = 1 pt answer + 1 pt R code + 1 pt scatterplots]

2) Write down the regression model to predict salary using the other three variables as predictors [2 pts = 1 pt R code + 1 pt regression model]

3) Which of the three predictors have a significant effect on salary? (alpha=.05) [3 pts = 1 pt for test + 1 pt for R code + 1 pt for conclusion]

4) We refit the model using only the two significant attributes. Interpret each of the partial slope coefficients in the new model. [2 pts = 1 pt R code + 1 pt answer]

5) Analyze the Coefficient of Determination R^2 [2 pts = 1 pt R code + 1 pt answer]

6) Analyze the goodness of fit test (write down test hypothesis, test statistic and p-value, and draw conclusions). What can you conclude about the predictive power of the model? [2 pts = 1 pt for the test + 1 pt for conclusion]

7) Analyze the standardized coefficients of the model, which variable has the strongest effect on Y? [2 pts = 1 pt R code + 1 pt answer]

PROBLEM 4 [4 pts]

Laughter is often called ''the best medicine,'' since studies have shown that laughter can reduce muscle tension and increase oxygenation of the blood. In the International Journal of Obesity (January 2007), researchers at Vanderbilt University investigated the physiological changes that accompany laughter. Ninety subjects (18–34 years old) watched film clips designed to evoke laughter. During the laughing period, the researchers measured the heart rate (beats per minute) of each subject with the following summary results: y_bar = 73.5, s = 6. It is well known that the mean resting heart rate of adults is 71 beats/minute.

At $\alpha = .05$, is there sufficient evidence to indicate that the true mean heart rate during laughter exceeds 71 beats/minute? Show your work.

Submission instructions:
Submit the homework at the Course Web page at https:\\d2l.depaul.edu

- Keep a copy of all your submissions!
- Please submit the following: R file or Rmd file for the code, Word/PDF/HTML document for answers to the problems. If using Rmd, please compile and submit the Word/PDF/HTML document
- If you have questions about the homework, email me BEFORE the deadline.
- The assignment will lose 20% of the points if submission is late, unless a valid reason is specified
- Assignments submitted four days after the due date will not be accepted.