DSC423 - Data Analysis and Regression
Assignment 2
Total points: 52
Due Tuesday October 17th 2023 at 11:59 PM

PROBLEM 1 [28 pts]

The file bankingfull.txt attached to this assignment contains the full dataset. You analyzed a smaller set for Assignment 1. It 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 years of education (EDUCATION)
- median income (INCOME) in $
- median home value (HOMEVAL) in $
- median household wealth (WEALTH) in $
- average bank balance (BALANCE) in $

The goal of this exercise is to define a regression model to predict the average bank balance as a function of the other variables.

a) Create scatterplots to visualize the associations between bank balance and the other five variables. Discuss the patterns displayed by the scatterplot. Do the associations appear to be linear? (you can create scatterplots or a matrix plot) **[1 pt R code, 1 pt scatterplots, 2 pts answer = 4 pts]**

b) Compute correlation values of bank balance vs the other variables. Interpret the correlation values, and discuss which variables appear to be strongly associated. **[1 pt R code, 2 pts for answer = 3 pts]**

c) Fit a regression model of balance vs the other five variables (model M1). Compute the VIF statistics for each x-variable and analyze whether there is a problem of multicollinearity. **[2 pts R code, 1 pt answer = 3 pts]**

d) Apply your knowledge of regression analysis to define a better model M2, and answer the following questions:
   - Analyze the Coefficient of Determination R2 values and the adjusted adj-R2 values for both models M1 and M2. Which model has the largest adj-R2 value? **[1 pt selecting better model M2, 1 pt R code, 3 pts answer = 5 pts]**
   - Create residual plots (standardized residuals vs predicted; standardized residuals vs x-variables; and normal plot of residuals). Analyze the residual plots to check if the regression model assumptions are met by the data. **[1 pt R code, 1 pt plots, 1 pt answer = 3 pts]**
   - Analyze if there are any outliers and influential points for your model. If so, what are your recommendations? **[2 pts answer, 2 pts R code = 4 pts]**
   - Compute the standardized coefficients and discuss which predictor has the strongest influence on balance? **[1 pt R code, 1 pt answer = 2 pts]**

e) Use the fitted regression model from d) without removal of influential points to predict the average bank balance for a specific zip code area where there is a plan to open a new branch. Census data in that area show the following values: median age is 34 years, median education is 13 years, median income is $64,000, median home value is $140,000, median wealth is 160,000. (Note that you may not need all these values in your model). Provide predicted average bank balance and 95% confidence interval for your estimate. **[2 pts R code, 2 pts answer = 4 pts]**

Submit your R file for the problem indicating in comments which block of code solves what part of the problem, and your answers in a Word or PDF file. You can also submit an R Markdown file and the HTML/PDF file of the output.

Problem 2 [24 pts]

Analytics is used in many different sports and has become popular with the Money Ball movie. The pgatour2006.csv dataset contains data about 196 tour players in 2006. The variables in the dataset are:

- Player's name
- PrizeMoney = average prize money per tournament

And a set of metrics that evaluate the quality of a player's game.

- DrivingAccuracy = percent of times a player is able to hit the fairway with his tee shot
- GIR = percent of time a player was able to hit the green within two or less than par (Greens in Regulation)
- BirdieConversion = percentage of times a player makes a birdie or better after hitting the green in regulation
- PuttingAverage = putting performance on those holes where the green was hit in regulation.
- PuttsPerRound= average number of putts per round (shots played on the green)

You are asked to build a model for PrizeMoney using the remaining predictors, and to evaluate the relative importance of each different aspects of a player's game on the average prize money.

For the non golfers in the class, you can refer to this page for an explanation of the terms:
http://en.wikipedia.org/wiki/Glossary_of_golf

a) Create scatterplots to visualize the associations between PrizeMoney and the other five variables. Discuss the patterns displayed by the scatterplot. Do the associations appear to be linear? (you can create scatterplots or a matrix plot) **[1 pt R code, 1 pt scatterplots, 2 pts answer = 4 pts]**

b) Analyze distribution of PrizeMoney, and discuss if the distribution is symmetric or skewed. **[1 pt R code, 1 pt answer = 2 pts]**

c) Apply a log transformation to PrizeMoney and compute the new variable ln_Prize=log(PrizeMoney). Analyze distribution of ln_Prize, and discuss if the distribution is symmetric or skewed. **[2 pts R code, 1 pt answer = 3 pts]**

d) Fit a regression model of ln_Prize using the remaining predictors in your dataset. Apply your knowledge of regression analysis to define a valid model to predict ln_Prize. Hint: use scatterplots and correlation **[3 pts R code, 1 pt answer = 4 pts]**

   - If necessary remove not significant variables. Remember to remove one variable at a time (variable with largest p-value is removed first) and refit the model, until all variables are significant. **[2 pts R code, 1 pt answer = 3 pts]**
   - Analyze residual plots to check if the regression model is valid for your data. **[1 pt R code, 1 pt answer = 2 pts]**
   - Analyze if there are any outliers and influential points. If there are points in the dataset that need to be investigated, give one or more reason to support each point chosen. **[1 pt R code, 1 pt answer = 2 pts]**

e) Interpret the regression coefficients in the final model to answer the following question: How does an increase in 1% for GIR affect the average Prize money? **[1 pt R code, 1 pt answer = 2 pts]**

f) Compute the prediction and 95% prediction interval for average prize money for a player that has a GIR of 67%, driving accuracy of 64%, putting average of 1.77, Birdie Conversion of 28% and 29.16 average putts per round. **[1 pt R code, 1 pt answer = 2 pts]**

Submit your R file for the problem indicating in comments which block of code solves what part of the problem, and your answers in a Word or PDF file. You can also submit an R Markdown file and the HTML/PDF file of the output.