DATA ANALYSIS AND REGRESSION

Assignment-4 | Total Points: 26 pts for DSC 323; 40 pts for DSC 423

This assignment is longer to make up for the midterm exam. Make sure to start the assignment the very next day.

Note:

- All assignments should be submitted in a **single MS WORD format**, no PDFs or any other file types will be accepted. If you submit any other file type, it will not be graded.
- No extensions will be given unless for a documented reason specified in the syllabus, no late assignments past the due date even a couple of minutes late will be accepted as you have an extra day (7-days) to submit your assignments.
- Submitting work that is not yours is grounds for an automatic 'F' for the entire course – this includes taking content and ideas from others or consulting others to complete your deliverables other than your instructor.
- SAS software and virtual server stalls, gets slow and crashes; so start early and keep multiple backups in multiple places/mediums. Late submission or inability to do the assignment due to server and/or software issues will not be accepted. Any issues relating with SAS, contact IS using the phone number provided in the syllabus, I won't be able to help you with DePaul software related issues.
- **Make sure to double check your submissions. After you submit the assignment, log out of D2L, log back in, and click on your submission to see if you submitted the right file(s) and it is the correct version. Wrong submissions will not be graded.**

***Note:* For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.**

PROBLEM 1 [16 pts] – to be answered by everyone

The file banking.txt attached to this assignment contains the full dataset. 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. Include the relevant output. Discuss the patterns displayed by the scatterplot. Also, explain if the associations appear to be linear? (you can create either scatterplots or a matrix plot)

b) Compute correlation values of bank balance vs the other variables. Include the relevant output. Interpret the correlation values, and discuss which variables appear to be strongly associated.

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 and take appropriate action. Include the relevant output. Discuss your answer.

d) Apply your knowledge of regression analysis to define a better model M2. Include the SAS output for both models and answer the following questions :

1) Analyze the adj-R2 values for both models M1 and M2. Which model has the largest adj-R2 value?
2) Create residual plots for M2 (Studentized residuals vs predicted; Studentized 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. Include the relevant output and discuss your analysis.
3) Analyze if there are any outliers and/or influential points for your M2 model. If so, what actions would you take to address this issue? Make sure to implement any actions you specify here. Include the relevant output.
4) Compute the standardized coefficients for M2 and discuss which predictor has the strongest influence on balance? Include the relevant output.

e) Copy and paste your FULL SAS code into the word document along with your answers.

Problem 2 [10 pts]– to be answered by everyone

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

Note: 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

SAS Code to Import the data

```
*import data from file;
proc import datafile="golf.csv" out=golf replace;
delimiter=',';
getnames=yes;
run;
```

Note:

- The data file is in CSV format
- It is delimited with a comma
- The SAS dataset it is writing into is golf. You can change the name if you like.

a) Create scatterplots to visualize the associations between PrizeMoney and the other 5 variables. Discuss the patterns displayed by the scatterplot. Also, explain if the associations appear to be linear? (you can create scatterplots or a matrix plot). Include the relevant output.

b) Analyze distribution of PrizeMoney, and discuss if the distribution is symmetric or skewed. Include the relevant output.

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. Include the relevant output.

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. <u>Include the outputs for all the questions below before you analyze them</u>.

1) If necessary remove the non-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) Analyze residual plots to check if the regression model is valid for your data. Discuss your analysis.
3) Analyze if there are any outliers and/or influential points. If there are points in the dataset that need to be investigated, give one or more reason to support each point chosen. Take appropriate action(s) to implement it. Include the relevant outputs. Discuss your answer.
4) Write down the final model equation. Discuss why this is the best model. Include all relevant statistics/values to substantiate your answer.

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?

f) Copy and paste your FULL SAS code into the word document along with your answers.

Problem 3 [14 pts]– – For Graduate Students ONLY

1. Based on the two graphs you see below, which regression line is likely to change due to the point circled in red. Explain why you think that.

Figure-1

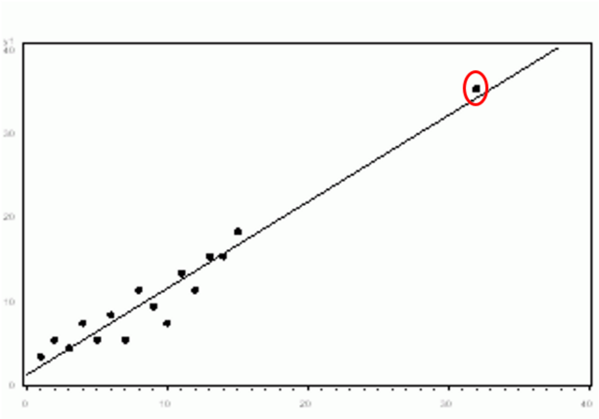
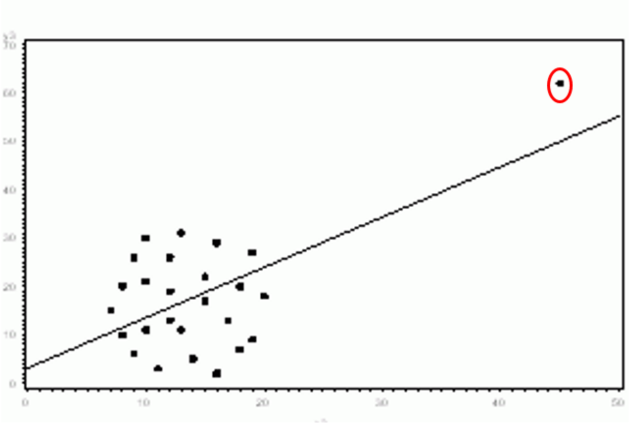


Figure-2



2. What are considered model diagnostics?

3. Why do you check model diagnostics? Explain the reason for each of the diagnostics separately.

4. What are considered model assumptions?

5. Why do you check model assumptions? Explain the reason for each of the assumption separately.