**Ximan Liu**

**Data Analysis And Regression**

**Assignment-4** | **Total Points: 46 pts for DSC 423 and 26 pts for DSC 323**

**Due Date: 4/26/2021 by 11:59 pm**

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 (8-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.

***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 bankingfull.txt attached to this assignment contains the full dataset. You analyzed a smaller set for a previous assignment. 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.

1. 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)

A picture containing calendar

Description automatically generated

**The graph represents positive linear relationship between balance and other variables (age, education, income, homeval and wealth). The relationship between balance and income, and the relationship between balance and wealth represent strong associations. There are possible potential outliers at the top right corner.**

**The relationship of age and balance is weak association because there is a cloud in the graph.**

**The relationship of education and balance is weak association because there is a cloud in the graph.**

**The relationship of homeval and balance is strong association. There are possible potential outliers in the plot.**

1. 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.

Table

Description automatically generated

**The higher the correlation r value (between 0 and 1), the stronger the linear the linear relationship.**

**r(balance, income) = 0.95168 and r(balance, wealth) = 0.94871 represent highly strong perfect positive association between balance and income, and balance and wealth.**

**r(balance, homeval) = 0.76639 represents a relatively strong association between balance and homeval.**

**r(balance, age) = 0.56547 and r(balance, education) = 0.55488 represent a moderate weak association between balance and age, and balance and education.**

1. 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.

Table

Description automatically generated

**Multicollinearity is when two or more of the predictors in a regression model are moderately or highly correlated with each other. If VIF of a variable is greater than 10, then it is considered to be multicollinearity. The R-Square is 0.9469 which is a quite high number. In the graph, income and wealth are both greater than 10, which means high possibility of multicollinearity. Hence, the estimation of regression coefficients is weak.**

**The corresponding solution is, remove either income or wealth from the model. The reason for removing income is, income might not be a sufficient predictor for balance. Vice versa, we could also remove wealth here, because holders with a lot of money might have properties that aren't recorded in the bank.**

1. 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?

**Drop Wealth**

Table

Description automatically generated

**Drop Income (M2)**

Table

Description automatically generated

**Compared the dataset removing income and the one removing wealth, we can tell that the regression model removing income has a higher R-Square (0.9398) and adj R-Square (0.9373). In that case, we choose the model removing income (M2).**

* 1. 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.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

**The first five graphs (age, education, home value, wealth and predicted value) have scattered around zero line randomly. And it represents independence and constant variance.**

**The first four graphs (age, education, home value, wealth) are linearly.**

**The last one shows a linearly line and it is normal distributed.**

* 1. 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.

Table

Description automatically generated

A picture containing text, thermometer, device

Description automatically generated

Table

Description automatically generated with medium confidence

**There are four outliers at obs 38, 85, 91, 102 marked as red color, they are over +3 or -3 range. Besides, the influential points marked as blue are obs 9, 38, 59, 77, 82, 84, 85, 91 and 102.**

**Firstly, let's say obs 102 is marked with both outlier and influential point. We will remove it. Then we retry it and see whether it helps to improve the result until nothing will be changed at the end.**

**All in all, I tried removing 6 obs and see the changes on R-Square and adj R-Square.**

**At this time, with R-Square is 0.9688 and adj R-Square is 0.9675, we can tell that we can stop deleting and this is the final model.**

* 1. Compute the standardized coefficients for M2 and discuss which predictor has the strongest influence on balance? Include the relevant output.

Table

Description automatically generated

**The largest absolute value of the standardized coefficient represents the x-variable that has the greatest influence on Y, and can be considered the strongest predictor of Y.**

**The strongest influence on balance is wealth, it has highest standardized estimate as 0.71863.**

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

**TITLE "Analysis of Bankingfull";**

**PROC IMPORT datafile="C:\Users\XLIU115\Desktop\Assignment4\Bankingfull.txt" out=Balance replace;**

**getnames=yes;**

**delimiter='09'x;**

**RUN;**

**PROC PRINT data=Balance;**

**RUN;**

**/\*1A\*/**

**PROC SGSCATTER;**

**title "Scatterplot Matrix - Balance and Other Variables";**

**matrix Balance Age Education Income HomeVal Wealth;**

**RUN;**

**/\*1B\*/**

**PROC CORR;**

**var Balance Age Education Income HomeVal Wealth;**

**RUN;**

**/\*1C\*/**

**PROC REG;**

**title "Regression Model - Balance and Other Variables";**

**model Balance= Age Education Income HomeVal Wealth /vif;**

**RUN;**

**/\*1D\*/**

**/\*Drop Wealth\*/**

**DATA Balance\_1;**

**set Balance;**

**drop Wealth;**

**RUN;**

**PROC PRINT data= Balance\_1;**

**RUN;**

**PROC REG;**

**title "Regression Model - Balance and Other Variables New 1";**

**model Balance= Age Education Income HomeVal /vif;**

**RUN;**

**/\*Drop Income\*/**

**DATA Balance\_2;**

**set Balance;**

**drop Income;**

**RUN;**

**PROC PRINT data= Balance\_2;**

**RUN;**

**PROC REG;**

**title "Residual Plots - Balance and Other Variables New 2";**

**model Balance= Age Education HomeVal Wealth /vif;**

**plot student.\*(Age Education HomeVal Wealth predicted.);**

**plot npp.\*student.;**

**RUN;**

**PROC REG;**

**title "Outliers - Balance and Other Variables New 2";**

**model Balance= Age Education HomeVal Wealth /influence r;**

**plot student.\*(Age Education HomeVal Wealth predicted.);**

**plot npp.\*student.;**

**RUN;**

**DATA Balance\_3;**

**set Balance\_3;**

**if \_n\_=58 then delete;**

**RUN;**

**PROC REG;**

**title "Regression Model - Balance and other variables New 2";**

**model Balance= Age Education HomeVal Wealth /stb;**

**RUN;**

**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 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.

**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="pgatour2006.csv" out=PGATour replace;

delimiter=',';

getnames=yes;

**run**;

Note:

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

1. 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.

A picture containing table

Description automatically generated

Table

Description automatically generated

**The graph represents the linear relationship between PrizeMoney and other variables (DrivingAccuracy, GIR, PuttingAverage, BirdieConversion, PuttsPerRound).**

**From the correlation coefficients, we can tell that there is no association between PrizeMoney and DrivingAccuracy with value 0.02468.**

**GIR has the value as 0.41022 and BirdieConversion has the value as 0.41343. They are moderately positive correlated to PrizeMoney.**

**PuttingAverage has the value as -0.31305 and PuttsPerRound has the value as -0.11249. They represent mildly negative correlation to PrizeMoney.**

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

Chart, histogram

Description automatically generated

**Among 196 players, the mean of PrizeMoney is $50,891.17. The range starts from $2,240 to $662,771, with a large wide $660,531. The median is $36,644.5 which located in the middle of Q1 and Q3. The mean is greater than the median in this graph, hence the distribution is right and positively skewed and unimodal. A pointed top and lighter tails characterize the distribution. The potential outliers may occur at $640,000.**

1. 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.

Chart, histogram

Description automatically generated

**The distribution is symmetric and normal after log transformation. The median ($10.509) is quite close to the mean ($10.37808) now. The range starts from $7.714231 to $13.40418 with a smaller wide $5.689954. The median ($10.509) is located between Q1 ($9.753768) and Q3 ($10.96831). All in all, the distribution has a normal peak.**

1. 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. Include the outputs for all the questions below before you analyze them.

Table

Description automatically generated

Table

Description automatically generatedTable

Description automatically generated

* + 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.

**The first graph is a regression model for In\_Prize and other all variables (DrivingAccuracy, GIR, PuttingAverage, BirdieConversion and PuttsPerRound). As the p-value for DrivingAccuracy is 0.948, it has the largest p-value among all the others. So the next step is deleting DrivingAccuracy.**

**The second graph is removing DrivingAccuracy. Then the situation is telling the regression model for In\_Prize and other variables (GIR, PuttingAverage, BirdieConversion and PuttsPerRound). The PuttingAverage has the largest p-value as 0.0962 among others. So the next step is removing PuttingAverage.**

**Finally, after removing PuttingAverage, we can tell from the graph that all the p-value of variables are less than 0.05. In that case GIR, BirdieConversion and PuttsPerRound are significant variables to In\_Prize.**

* + 1. Analyze residual plots to check if the regression model is valid for your data. Discuss your analysis.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

**The first three graphs (GIR, BirdiesConversion and PuttsPerRound) are linearly.**

**The first four graphs (GIR, BirdiesConversion, PuttsPerRound and predicted value) are scattered around zero line by random. They represent independency and constant variance.**

**The last plot represents a linearly line, it means the model is linearly normal distributed.**

* + 1. 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.

A picture containing text, device, thermometer

Description automatically generatedA picture containing text, device, thermometer, measuring stick

Description automatically generated

**There is only one outlier at obs 185 marked as red color. Besides, the influential points are marked in blue. Firstly, let's say obs 185 is marked with both outlier and influential point. We will remove it. Then we retry it and see the changes on R-Square and adj R-Square and whether it helps to improve the result until nothing will be changed at the end. All in all, I tried removing 17 obs when R-Square and adj R-Square are the largest.**

* + 1. Write down the final model equation. Discuss why this is the best model. Include all relevant statistics/values to substantiate your answer.

Table

Description automatically generated

**Final model equation:**

**In\_Prize = 11.62487 + 0.29826 \* GIR + 0.13252 \* BirdiesConversion – 0.84227 \* PuttsPerRound + e**

**Null hypothesis:**

**Ho: β1 = β2 = β3 = … = βk = 0**

**Alternative hypothesis:**

**Ha: At least one coefficient βj ≠ 0**

**Test statistic:**

**This model is the final model because it has the highest R-Square (0.6877) and adj R-Square (0.6824). F = 128.46 and with p-value less than 0.001 (at alpha=0.05). The null hypothesis of no association between PrizeMoney and other variables (GIR, BirdiesConversion and PuttsPerRound) is rejected. At least one x-variable has a significant effect on changes in PrizeMoney. F-test gives strong support to the fitted model.**

1. 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?

**GIR has positive association to average prize money. As all the other variables keep constantly, every 1% increase of GIR will make average prize money increases 34.99%.**

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

**TITLE "Analysis of pgatour2006";**

**PROC IMPORT datafile="C:\Users\XLIU115\Desktop\Assignment4\pgatour2006.csv" out=PGATour replace;**

**delimiter=',';**

**getnames=yes;**

**RUN;**

**PROC PRINT data=PGATour;**

**RUN;**

**/\*2A\*/**

**PROC SGSCATTER;**

**title "Scatterplot Matrix - PrizeMoney and Other Variables";**

**matrix PrizeMoney DrivingAccuracy GIR PuttingAverage BirdieConversion PuttsPerRound;**

**RUN;**

**/\*2B\*/**

**TITLE "Histogram - Distribution of PrizeMoney";**

**PROC UNIVARIATE normal;**

**var PrizeMoney;**

**histogram / normal (mu=est sigma=est);**

**inset median mean min max range Q1 Q3 kurtosis skewness /pos = ne;**

**RUN;**

**/\*2C\*/**

**DATA PGATour;**

**set PGATour;**

**ln\_Prize = log(PrizeMoney);**

**PROC PRINT;**

**RUN;**

**TITLE "Histogram - Distribution of ln\_Prize = log(PrizeMoney)";**

**PROC UNIVARIATE normal;**

**var ln\_Prize;**

**histogram / normal (mu=est sigma=est);**

**inset median mean min max range Q1 Q3 kurtosis skewness /pos = ne;**

**RUN;**

**/\*2D\*/**

**/\*Regression Model - All Variables\*/**

**PROC REG;**

**title "Regression Model - All Variables";**

**model ln\_Prize = DrivingAccuracy GIR PuttingAverage BirdieConversion PuttsPerRound;**

**RUN;**

**PROC REG corr;**

**model ln\_Prize = DrivingAccuracy GIR PuttingAverage BirdieConversion PuttsPerRound;**

**RUN;**

**/\*Regression Model - Remove DrivingAccuracy\*/**

**PROC REG;**

**title "Regression Model - Remove DrivingAccuracy";**

**model ln\_Prize = GIR PuttingAverage BirdieConversion PuttsPerRound;**

**RUN;**

**PROC REG corr;**

**model ln\_Prize = GIR PuttingAverage BirdieConversion PuttsPerRound;**

**RUN;**

**/\*Regression Model - Remove DrivingAccuracy & PuttingAverage\*/**

**PROC REG;**

**title "Regression Model - Remove DrivingAccuracy & PuttingAverage";**

**model ln\_Prize = GIR BirdieConversion PuttsPerRound /influence r;**

**plot student.\*(GIR BirdieConversion PuttsPerRound predicted.);**

**plot npp.\*student.;**

**RUN;**

**PROC REG corr;**

**model ln\_Prize = GIR BirdieConversion PuttsPerRound;**

**RUN;**

**DATA PGATour2;**

**set PGATour2;**

**if \_n\_ in (138) then delete;**

**RUN;**

**PROC REG;**

**title "Regression Model - Remove DrivingAccuracy & PuttingAverage New";**

**model ln\_Prize = GIR BirdieConversion PuttsPerRound /influence r;**

**plot student.\*(GIR BirdieConversion PuttsPerRound predicted.);**

**plot npp.\*student.;**

**RUN;**

**Problem 3 [20 pts] – ONLY for GRADUATES**

Answer the following questions:

1. Why is Adj-R2 a better indicator to use than the R2? Explain. (1 point)

**Adj-R2 is the better model when comparing models that have a different amount of variables. R2 always increases when the number of variables increases. Even if adding a useless variable to model, R2will still increase. We should always compare models with different number of independent variables with Adj-R2. Adj-R2 only increases if the new variable improves the model more than would be expected by chance.**

1. Which transformations will you use for ‘Loan Amount’? List all that applies. (1 point)

**1) Log(Y) (only if Y > 0)**

**2) Sqrt(Y) (only if Y >= 0)**

**3) Square Y =Y2**

**4) Cubic Y = Y3**

**5) Inverse Y = 1/Y (only for Y ≠0)**

1. At which stages should you check for assumptions? (2 points)

**Fit the full model (i.e. with all the predictors) and check residuals/assumptions.**

**Final Model (i.e. final fitted model) and check residuals/assumptions.**

1. At which stages should you check for collinearity? (2 points)

**At Data Exploration Stage.**

**At Analysis State – Verify & Confirm.**

1. At which stages should you check for outliers? (2 points)

**At Data Exploration Stage.**

**At Analysis Stage.**

1. At which stages should you check for influential points? (2 points)

**At Data Exploration Stage.**

**At Analysis Stage.**

1. “Outliers and influential points are the same” – explain why this statement is true/false. (2 points)

**False.**

**Outliers are the data points those diverge by good margin from the overall pattern. It can have an extreme X or Y values or both compared to other values.**

**Influential point is an outlier that impacts the slope of the regression line. To test the influence of an outlier is to compute the regression equation with and without the outlier.**

1. When should I ignore outliers and influential points? Explain. (2 points)

**Removing an observation.**

**Remove the outlier obs# and run the model. Check the adj-R2, residual plots and p-values of the predictors. See if they improve.**

**Remove the influential point that got flagged by almost all indicators obs. Check the adj-R2, residual plots and p-values of the predictors. See if they improve. If it doesn’t, keep it as part of observations.**

**Rerun until check adj-R2, goodness of fit test, residuals and p-values of all predictors. If Adj-R2 get improved, f-value is high, and p-value associated with f-statistic is less than 0.05, then overall goodness of fit test shows that at least one predictor is significantly associated with Y. Then we can ignore outliers and influential points.**

1. When encountered with collinearity issue, one option is to drop one of the collinear variable. What other option(s) do I have other than dropping one variable? (2 points)

**Instead of doing nothing, the other options are**

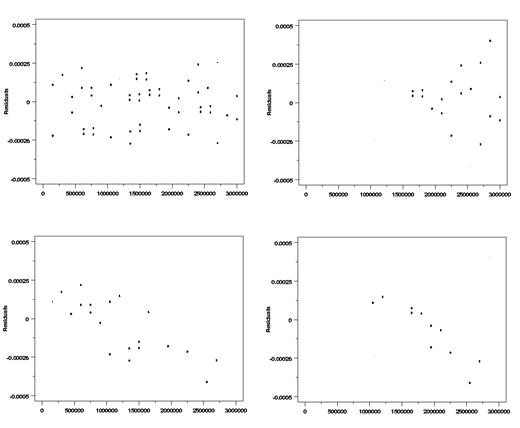
**Create 2 different models, one with one of the collinear variables and the 2nd model with the other collinear variable**

**Reuse the variable by creating another variable**

**Create interaction terms**

**Centering the collinear variables**

1. Based on the residual plots shown for X1, X2, X3 and predicted Y variables, indicate which of these you will need to transform. Explain. (4 points)



Predicted Y

X3

X2

X1

**X2, X3 and Predicted Y need to transformation. Because The residual plots show points that are NOT randomly scattered or the normal probability plot shows “S” shape. In that case we need to apply transformation on the response variable Y to stabilize the variance.**