**Data Analysis And Regression**

**Assignment-2** | **Total points: 15**

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.***

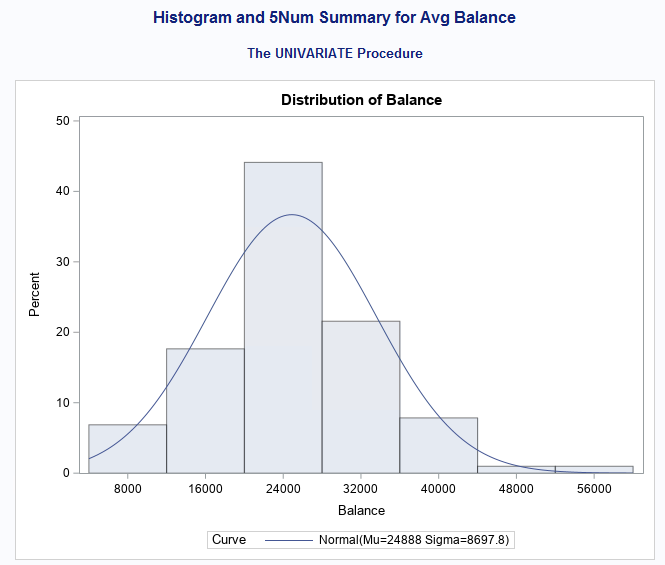
**PROBLEM 1 [15 pts]**

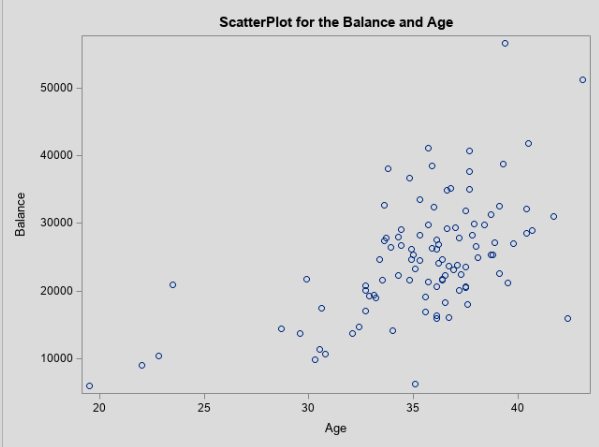
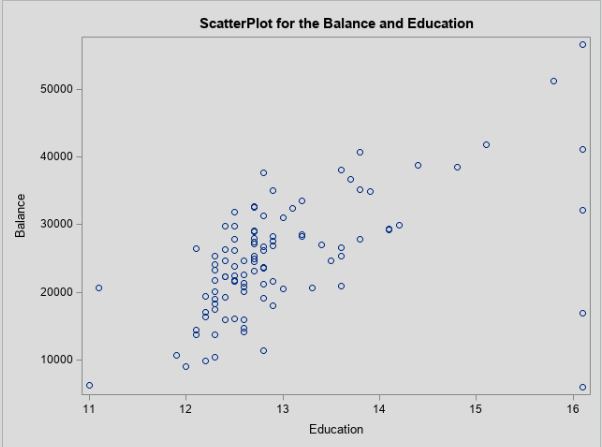
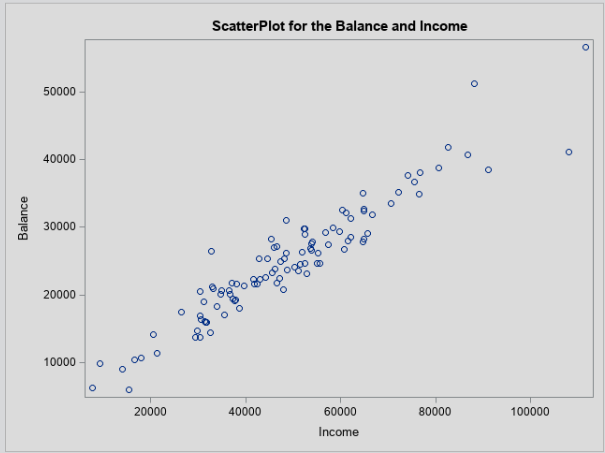
The file banking.txt attached to this assignment provides data acquired from banking and census records for different zip codes in the bank’s current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices. The data show

* median age of the population (AGE)
* median income (INCOME) in $
* average bank balance (BALANCE) in $
* median years of education (EDUCATION)

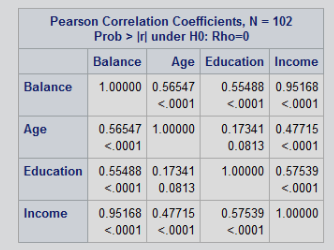
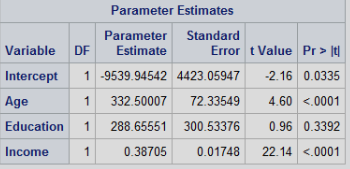
In this exercise you are asked to apply regression analysis techniques to describe the effect of age education and income on average account balance.

1. 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. Discuss your findings. Also, do you see any outliers? Include the histogram.
   * Upon analyzing the dataset provided, I can confidently say that the balance column in the dataset is distributed normally with the average balance (mean) of $24887.8, and a standard deviation of $8697.81. I also noticed that over 40% of bank balances fell between $20,000 to $28,000. The difference between the highest and the lowest value (Range) is about $50613. All this information from the graph tells us that the average population has a balance of $24000, and it is best if we focus the advertisement to target the zip codes with balances ranging from $16,000 and $34,000, because majority of balances lies between this region, and advertisements will reach to most people. Another thing that I noticed was that I did not see any outliers, as most of the balances seemed to be increasing slowly and highest balance did not just appear out of the Blue.



1. Create scatterplots to visualize the associations between bank balance and the other variables. Discuss the patterns displayed by the scatterplot. Also, do the associations appear to be linear? (You can create scatterplots or a matrix plot). Include the scatterplots.
   * Balance and Age – ****
     + Upon creating a scatter plot for Balance and Age, we can see that it is a weak positive linear relationship. As age increases the Balance also increases, although the rate amount of balance at a certain age can vary but majority of the people that are between the ages of 33 and 43, seem to have a balance from $18,000 to $40,000. This might be because majority of the data recorded seem to be between the ages 30 to 40. So, to get a more accurate understanding and correlation between age and balance, there should be more data for ages 20 – 30 and 50 – 70.
   * Balance and Education - 
     + The scatterplot for Balance and Education appears to have a weak positive linear relationship we can see that as the Education increases, the balance also increases for the most part. We can also see that majority of the people in the dataset are either high school graduates or college dropouts, as 80% of the observations seems to be between 12 (high school senior year) or 13 (college first year). We also see that the people with higher education seem to have more balance in the account when compared to the people that are just high school graduates whose balance ranges from $32,000 to $8,000. This is quite low compared to the range of a college dropouts (education 13+) or college graduates (education 16+), which ranges from $55,000 to $20,000. Although there are some outliers in this area, I have not considered those because they do not fit the trend that is obviously present in the graph.
   * Balance and Income – 
     + The scatter plot between Balance and income has a strong positive linear relationship. Income and Balance coloration is a lot more obvious when compared to age and education. We can clearly see that the balance is very dense as the income increases. We can say that people have half of their income saved as balance. As we can see people who make $20,000 has about $10,000 in their bank account and people who make $80,000 has about $40,000 in their bank account.

This raises a question are people following a rule that tells them to save at least half of their income? Or is it just a coincidence.

1. 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. Include the relevant output that shows the correlation values.
   * Output from SAS - 
   * As I noticed in the scatterplots the correlation between balance and Income is very high, when compared to age or education. Both age and education seemed to have hot spots in a region, rather than a consistent uptrend between the entire graphs. This correlation chart confirms my assumption about age and education having a low correlation, with age having a correlation of +0.56 which is a moderate positive association to balance, and education also has a correlation of +0.55 which is also a moderate positive association to balance. So, we can safely say that people save money based on their income, since the correlation between age and income is +0.95 which is a near perfect positive association.
2. What is the dependent variable and what are the independent variables in this regression analysis?
   * In this regression analysis the dependent variable is Balance, and the independent variables are Age, Education, and Income.
3. Use SAS 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=0.05. Include the relevant regression output.
   * Output from SAS - 
   * Upon looking at the Parameter estimates for the model, I can safely say that the variables Age and Income have a p value of 0.001 which is less than our requirement of 0.05, so we can say that they significant. Therefore, they will go into the model. However, Education is not significant, because it has a p value of 0.33 which is greater than our requirement of 0.05.
   * Significant variables – Age and Income.
   * Non-significant variables – education.
   * Model- ( Balance = -9539.9 + 332.5 Age + 288.6 Education + 0.38 Income.
4. If one of the predictors is not significant, remove it from the model and refit the new regression model. Write the expression of the newly fitted regression model.
   * Code for SAS

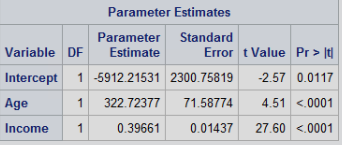
\*Print Updated Regression model with no Education variable";

Title "Regression Model without Education Variable.";

**PROC** **reg** data= main\_data;

model Balance = Age Income;

**RUN**;

* + Output for that Code
  + 
  + Balance = -5912.2 + 322.7 AGE + 0.39 Income

1. Interpret the value of the parameters for the variables in the model.
   * The parameter values for the variables is as follows.
   * Age = 322.7
   * Income = 0.396
   * Equation: Balance = -5912.2 + 322.7 AGE + 0.39 Income
   * So, if we increase 1 in age. we will increase Balance by 322.7.
2. Report the value for the R2and Adj-R2 coefficient and describe what it indicates. Include the portion of the output that includes the R2and Adj-R2 coefficient values.
   * The value for R^2 is 0.9218 or 92.18%.
   * R2 means that the 92.18% of variability of Balance can be explained using Age and Income.
   * The value for Adj-R^2 is 0.9202 or 92.02%
   * The adjusted R^2 tries to do the same thing as R^2. But the only difference is that adj R^2 tried to make it so that it is more accurate even in the case where there is a large number of predictors and less numbers of observations to explain the variance of balance.

1. 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 step (f) above to compute the predicted average balance for the zip code area.
  + Model from Question F - Balance = -5912.2 + 322.7 AGE + 0.39 Income
  + Changes requested in Question I = Age = 34.8yrs, Education = 12.5yrs, Income = $42,401

Age

* + If we increase the age by 1 year, we increase balance by $322.7.
  + 34.8yrs \* $322.7 = $11,229.96
  + Therefore, increasing age to 34.8yrs, while keeping everything else constant will increase Balance by $11,229.96.

Income

* + If we increase the Income by $1, we increase balance by $0.39.
  + $42,401 \* 0.39 = $16,536.39.
  + Therefore, Increasing the Income to $42,401, while keeping everything else constant will increase the Balance by $16,536.9.

Education

* + We determined that Education was not significant enough to keep in our final model, so changing the education level will not have any effect in our final balance.

Final

* + Balance = $ -5912.2 + $16,536.39 + $11,229.96

= $ 21,854.15.

* + If we change out Age to 34.8yrs, and Income to $42,401. We will get a Balance of $21,854.15.
* If the observed average balance for the zip code area is $21,572, what’s the model prediction error?
  + In this case, we can assume that during the person doing the calculation only took the round numbers for the age and left out the 0.8yrs.

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