XIMAN LIU

DATA ANALYSIS AND REGRESSION

Assignment-2 | Total points: 15 for DSC 323 / 20 for DSC 423
Due Date: 4/12/2021 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.**

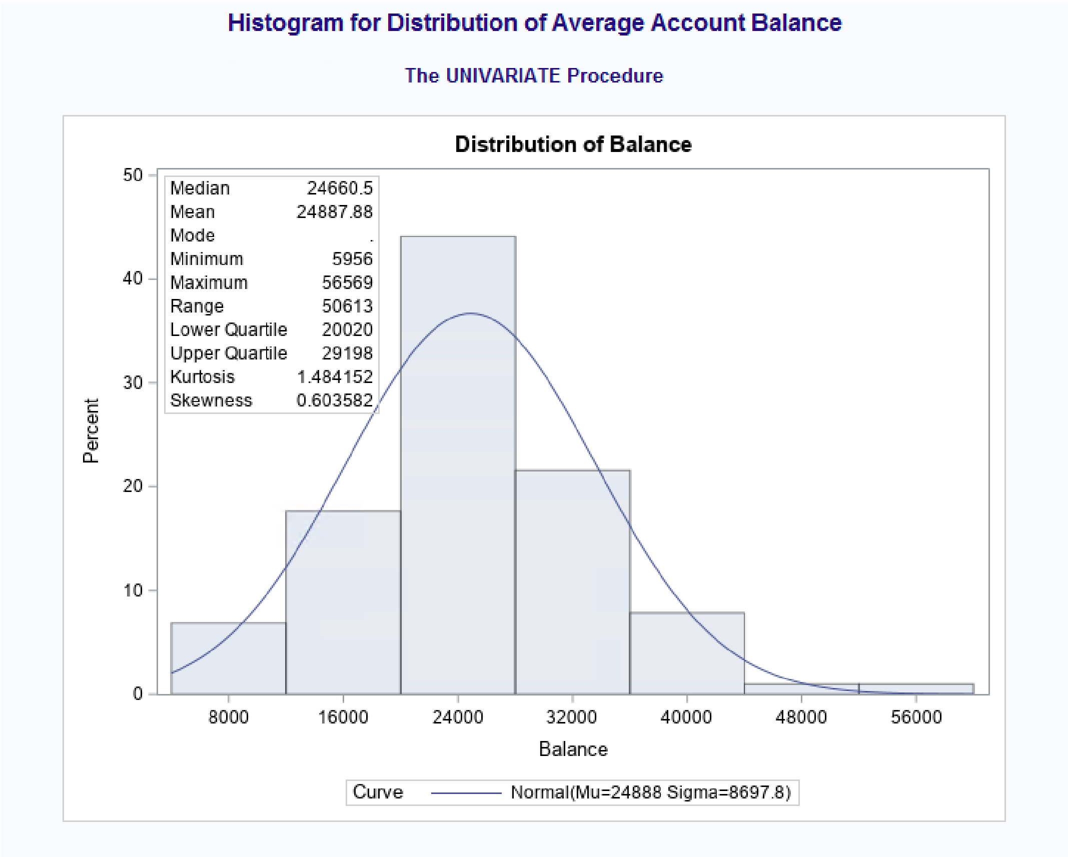
PROBLEM 1 [15 pts] – to be answered by everyone

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.

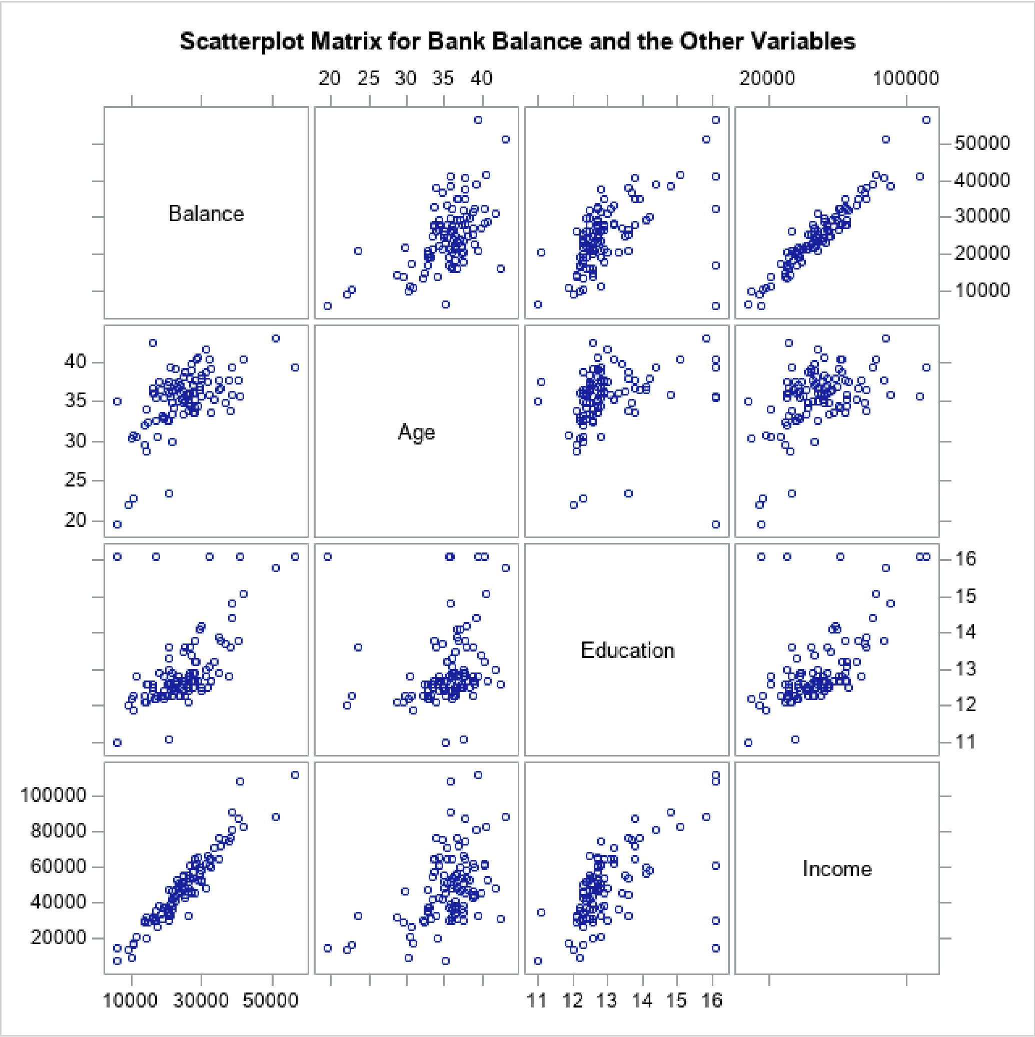
a) Analyze the distribution of average account balance using histogram, and compute appropriate descriptive statistics. Write a paragraph describing distribution of Balance and use appropriate descriptive statistics to describe center and spread of the distribution. Discuss your findings. Also, do you see any outliers? Include the histogram.


Histogram for Distribution of Average Account Balance

The UNIVARIATE Procedure

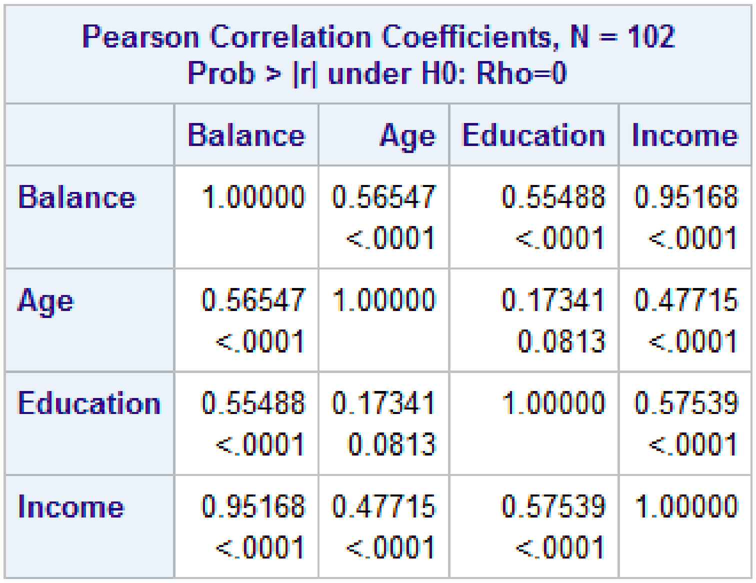
The mean for distribution of average account balance is $24,887.88. The minimum account balance is $5,956 which can be considered as very low. Vice versa, the maximum account balance is $56,569. 50% of account balance is between $20,020 and $29,198 among all census records for different zip codes. The median is between of lower quartile Q1 and upper quartile Q3. The range for distribution is $50,613 which means it is a large number. Besides, we can tell that the mean is greater than the median, it represents that the graph is right skewed, positively skewed and unimodal. There are outliers because the kurtosis is 1.484152 which less than 3. The graph also shows a flat top and heavy tails.

b) 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.



The scatterplot matrix represents positive linear relationship between bank balance and the other variables. The relationship between bank balance and income has stronger linear correlation. Besides, the relationship between bank balance and age shows a relatively weak linear correlation. And the relationship between bank balance and education also shows the same situation.

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. Include the relevant output that shows the correlation values.



Pearson Correlation Coefficients, N = 102
Prob > |r| under H0: Rho=0

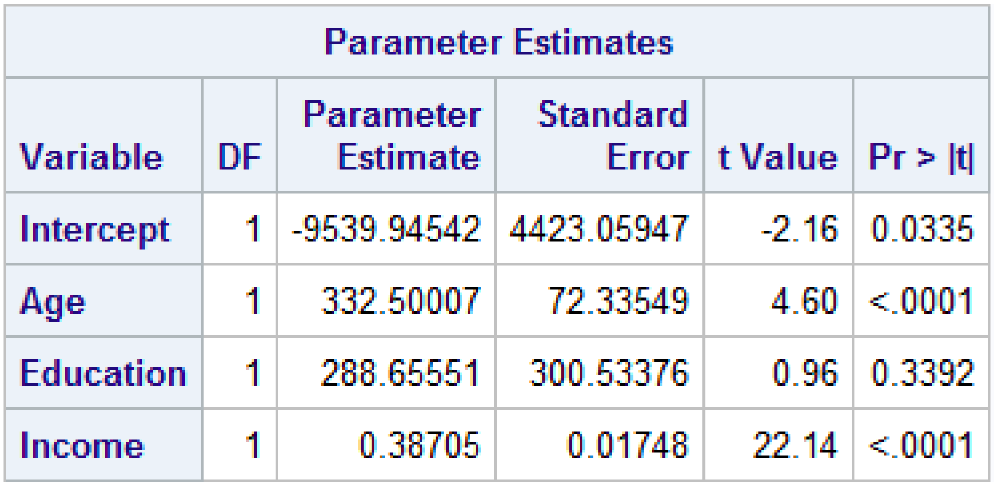
	Balance	Age	Education	Income
Balance	1.00000	0.56547 <.0001	0.55488 <.0001	0.95168 <.0001
Age	0.56547 <.0001	1.00000	0.17341 0.0813	0.47715 <.0001
Education	0.55488 <.0001	0.17341 0.0813	1.00000	0.57539 <.0001
Income	0.95168 <.0001	0.47715 <.0001	0.57539 <.0001	1.00000

The larger the absolute value of correlation coefficient r, the stronger the linear relationship. r(age, balance) = 0.56547 and r(education, balance) = 0.55488 measure a relatively positive correlation of age and education confronting to balance.
r(income, balance) = 0.95168 indicates a perfect positive correlation between balance and income.

d) What is the dependent variable and what are the independent variables in this regression analysis?

The dependent variable is average bank balance (BALANCE) and the independent variables are median age of the population (AGE), median income (INCOME) and median years of education (EDUCATION).

e) 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.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-9539.94542	4423.05947	-2.16	0.0335
Age	1	332.50007	72.33549	4.60	<.0001
Education	1	288.65551	300.53376	0.96	0.3392
Income	1	0.38705	0.01748	22.14	<.0001

Average Bank Balance = -9,539.94542 + 332.50007 * Age + 288.65551 * Education + 0.38705 * Income

As analyzing the model parameters, median age of the population (AGE) has a significant effect on balance, after that is median years of education (EDUCATION) and median income (INCOME).
As median age of the population (AGE) increases for one unit, the average bank balance (BALANCE) increases $332.50007.
As median income (INCOME) increases for one unit, the average bank balance (BALANCE) increases $0.38705.
As median years of education (EDUCATION) increases for one unit, the average bank balance (BALANCE) increases $288.65551.
If we use the t-tests on the parameters for alpha = 0.05, then median age of the population (AGE) and median income (INCOME) have a significant effect on balance, at the same time p-value of both of them are less than 0.0001. However, the p-value for median years of education (EDUCATION) is 0.3392, which is greater than 0.05. In that case, median years of education (EDUCATION) still effects on average bank balance (BALANCE).

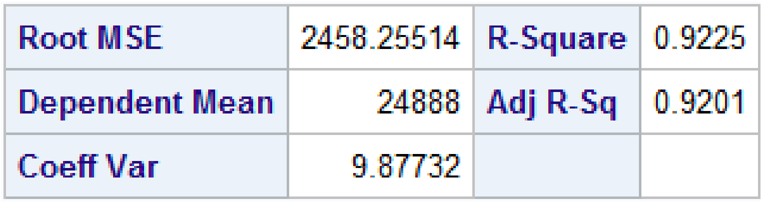
f) If one of the predictors is not significant, remove it from the model and refit the new regression model. Write the expression of the newly fitted regression model.

After excluding the education from the model, we need to fit the regression model once again. Then the coefficients of intercept, age and income will be -5912.2, 322.723 and 0.3966 respectively. The model equation is Balance = -5912.215 +322.724*Age + 0.397 * Income + e

g) Interpret the value of the parameters for the variables in the model.

The parameter for age measures that if income is fixed, as per one year added in age, the average account balance increases $332.50007.
The parameter for income measures that if age is fixed, as per one dollar added in income, the average account balance increases $0.38705.

h) Report the value for the R^2 and Adj-R^2 coefficient and describe what it indicates. Include the portion of the output that includes the R^2 and Adj-R^2 coefficient values.



Root MSE	2458.25514	R-Square	0.9225
Dependent Mean	24888	Adj R-Sq	0.9201
Coeff Var	9.87732		

The value for the R2 (0.9225 or 92.25%) and Adj-R2 (0.9201 or 92.01%) coefficient indicates the quantity of the variation in balance explained by the regression line. At this time, R2 (0.9225 or 92.25%) and Adj-R2 (0.9201 or 92.01%) of the variation in balance is explained by age and income. A high value of R2 does not necessarily mean that the regression model is a good fit for the data because R2 will always take a high value even if the variables have no effect on balance.

However, Adj-R2 (0.9201 or 92.01%) does not increase with the addition of a x-variable that does not improve the regression model. A higher Adj-R2 (0.9201 or 92.01%) typically indicates a better model.

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 step (f) above to compute the predicted average balance for the zip code area.

  Required to use final model where, education is not included to compute the predicted average balance.
  Predicted Balance = -5912.215 +322.724*34.8 + 0.397 * 42,401 = $ 22151.78.

- If the observed average balance for the zip code area is $21,572, what's the model prediction error?

  The error is Observed - Predicted = $21572 - $22151.78 = -579.78.

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

```
*1A;
PROC IMPORT datafile="C:\Users\XLIU115\Desktop\Assignment2\banking.txt"
out=salary replace;
getnames=yes;
RUN;

TITLE "Histogram for Distribution of Average Account Balance";
PROC UNIVARIATE normal;
var Balance;
histogram / normal (mu=est sigma=est);
inset median mean mode min max range Q1 Q3 kurtosis skewness;
RUN;


*1B;
PROC SGSCATTER;
  title "Scatterplot Matrix for Bank Balance and the Other Variables";
  matrix Balance Age Education Income;
RUN;

PROC CORR;
  var Balance Age Education Income;
RUN;


*1CDE;
PROC REG CORR;
TITLE "Regression for All the Variables";
model Balance = Age Education Income;
RUN;


*1F;
```

```
PROC REG CORR;
TITLE "Regression for Removing Education Variable";
model Balance = Age Income;
RUN;
```

Problem 2 [5 points] - ONLY for Graduate Students

Historical data about the Boston Marathon can be found on its website. The graph shows winning times (in minutes) for men and women against the year in which the race was run. Men's times are represented by "M" and women's time by "W". The graph also displays two regression lines of winning times vs year for men and women. There is no dataset for this question, but answer the following questions based on the graph.

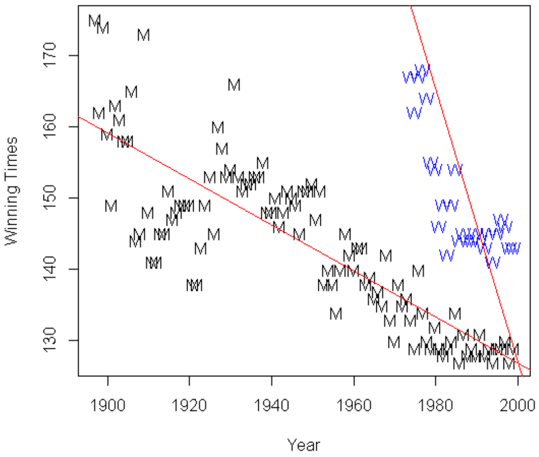
a) Consider the men's winning times, is there evidence of a linear trend? Would you expect the slope of the regression line to be positive or negative?

For men's winning times, there is a low negative linear trend. The slope of the regression line tends to be negative. From the graph, we can tell as the year increases, the winning times for men become shorter.

b) Now let's consider the winning times for women, is there evidence of a linear trend? Discuss.

The sample evidence is not strongly enough to say that there is a low negative linear trend for women's winning times. Observations are quite a few so that it is hard to tell the precise regression line. Also, there are a bunch of data clustered around 1990s, which affects the effectiveness of the regression line.

c) If we fit two separate linear regression models for men's and women's winning times, which slope will be greater in absolute value?



The slope of linear regression model for women's winning times is greater in absolute value. As the year increases, women's winning times drops sharply faster. Vice versa, the decline in men's winning times is more gradual.