**CSC423: Data Analysis And Regression / CSC 324: Data Analysis & Statistical Software II**

**Assignment-2** | **Total points: 15 for CSC 324 / 20 for CSC 423**

**Due Date: 04/10/2018 by 11:59 pm**

**The goal of problem 1 is to apply linear regression to a dataset using a statistical package.**

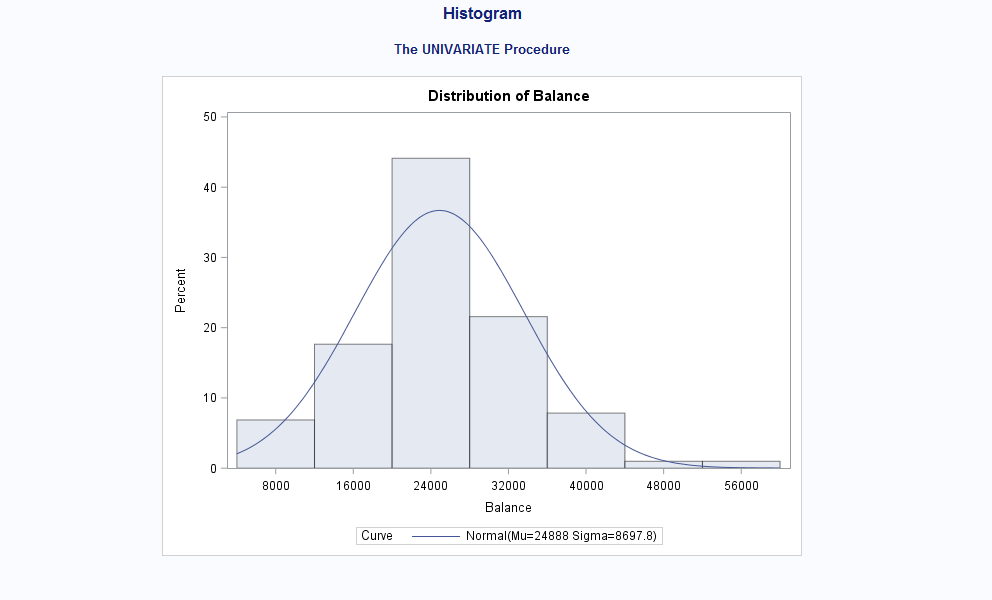
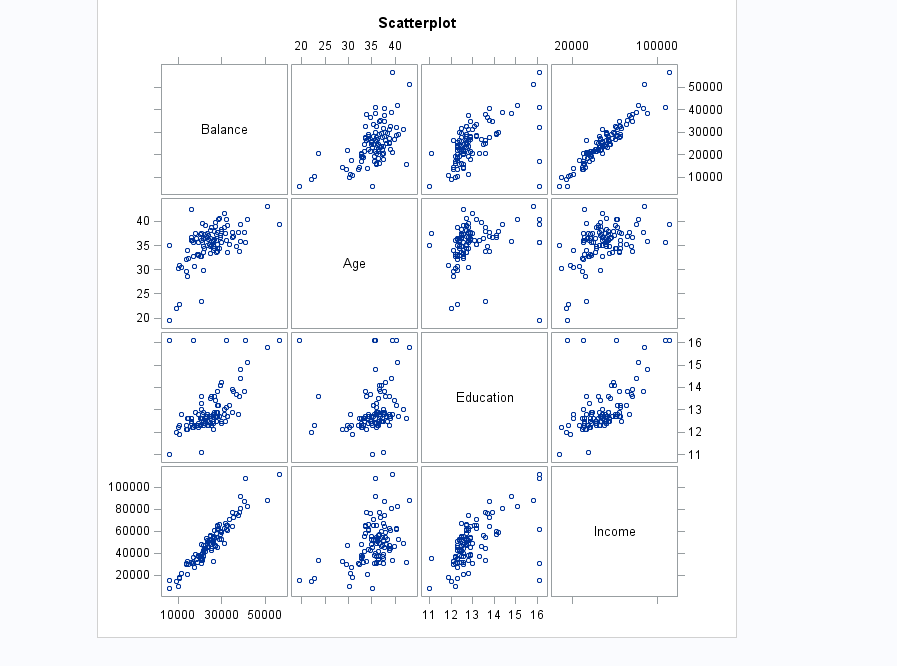
***Note: For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.***

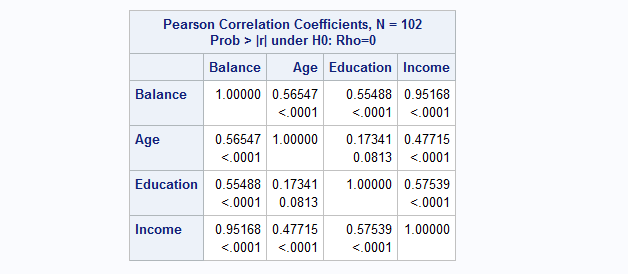
**PROBLEM 1 [15 pts] – to be answered by everyone**

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.

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. 
2. 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. 
3. 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.

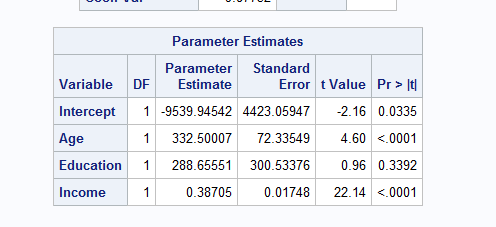


1. What is the dependent variable and what are the independent variables in this regression analysis?

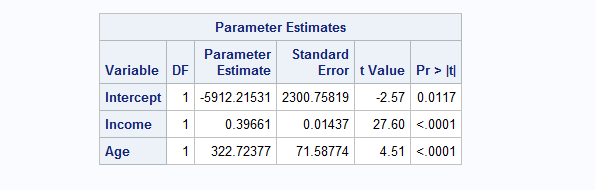
Dependant Variable – Balance

Independent – Age Income Education

1. 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.



1. 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.



1. Interpret the value of the parameters for the variables in the model.
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.
3. 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.
* If the observed average balance for the zip code area is $21,572, what’s the model prediction error?

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

/\* import data using PROC IMPORT\*/

**proc** **import** datafile="Banking.txt" out=Banking replace;

getnames=yes;

**run**;

/\* import data and create a Histogram\*/

title "Histogram";

**PROC** **UNIVARIATE** normal;

var balance;

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

**RUN**;

/\* import data and create a Scatterplot\*/

**proc** **sgscatter**;

title "Scatterplot";

matrix Balance Age Education Income;

**run**;

/\* corelation matrix\*/

**proc** **corr**;

var Balance Age Education Income;

**run**;

/\*regression model\*/

**proc** **reg** corr;

\*full model;

model Balance = Age Education Income;

\*Reduced model;

model Balance = Income Age;

\*Residual Plot : Residuals VS X-Variables;

plot balance.\*(income age);

**run**; **quit**;

/\*regression model 2\*/

**proc** **reg** corr;

\*full model;

model Balance = Age Income;

\*Reduced model;

model Balance = Income Age;

\*Residual Plot : Residuals VS X-Variables;

plot balance.\*(income age);

**run**; **quit**;

**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.

1. 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?
2. Now let’s consider the winning times for women, is there evidence of a linear trend? Discuss.
3. If we fit two separate linear regression models for men’s and women’s winning times, which slope will be greater in absolute value?

