**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: 01/30/2018 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**

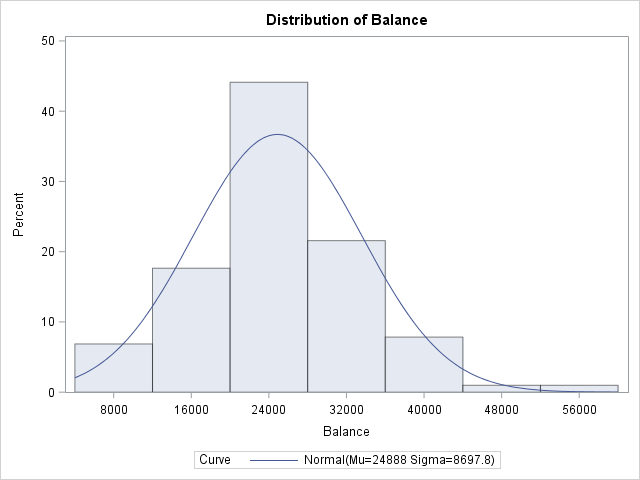
The file banking.txt attached to this assignment provides data acquired from banking and census records for different zip codes in the bank’s current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices. The data show

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

S:\Homeworks\banking.txt

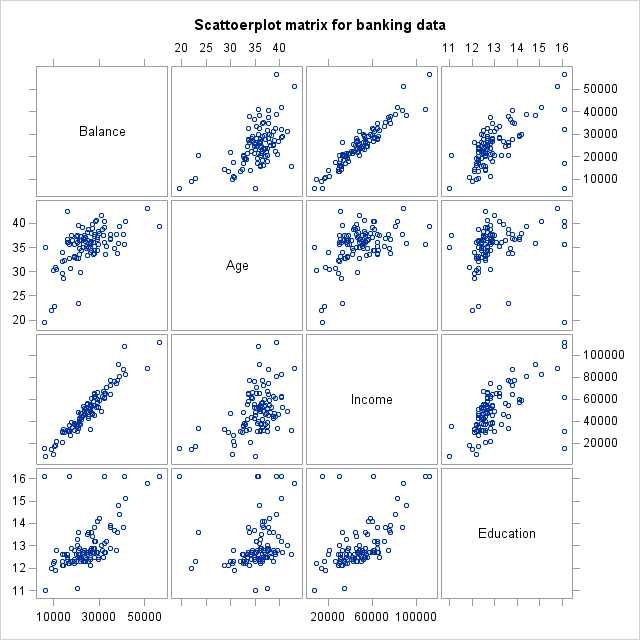
In this exercise you are asked to apply regression analysis techniques to describe the effect of age education and income on average account balance.

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.



The mean of the account balance is 24887.88 which the median is 24660.5. According to the histogram about 45% of the average account balance values are at $24000. Since the median and mean are fairly close in value, this is a normal distribution. The range of values is 50613.

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.



The scatter plot of balance vs income appears to be linear or has high positive correlation.

The scatter plot of balance vs age and balance vs education both appear to be linear with moderately positive correlation.

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.

Income and balance have a high correlation value with 0.95168. Balance/Age and Balance/Education have similar correlation values with 0.56547 and 0.55488 respectively.



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

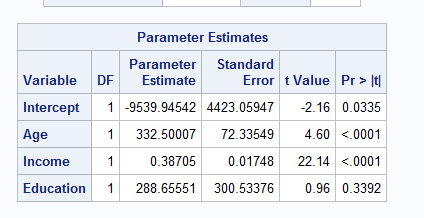
Dependent variable: Balance

Independent variables: Age, Income, Education

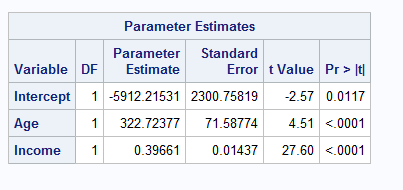
1. Use SAS to fit a regression model to predict balance from age, education and income. Analyze the model parameters. Which predictors have a significant effect on balance? Use the t-tests on the parameters for alpha=0.05. Include the relevant regression output.

Age and Income have a significant effect on balance based on their p-values, both of which are <0.0001.

Education has a p-value of 0.3392, meaning it does not have a significant effect on balance.



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.



Balance = -5912.215 + 322.72Age + 0.396 Income

1. Interpret the value of the parameters for the variables in the model.

For every 1 unit of median Income ($1) the Balance increase by $0.396.

For every 1 unit of Age (1 year) the Balance increase by $322.72

1. 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.



R2 means that the regression accounts for 92.18% of the variability in the data.

Adj-R2 means 92.02 of the variability can be exampled by the age and income variables

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.

Balance = -5912.215+(322.72)\*12.5 + 0.396\*(42201)

= $14,822.28

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

Prediction error = 21572-14822.28 = $6738.62

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

Title 'PArt1: import data';

**PROC** **IMPORT** datafile = 'S:\Homeworks\banking.txt' out=banking replace;

delimiter = '09'x;

getnames=YES;

datarow=**2**;

**run**;

**proc** **print**;

**run**;

Title 'Histogram of avg account balance';

**PROC** **univariate** normal;

var balance;

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

**run**;

**proc** **sgscatter**;

title "Scattoerplot matrix for banking data";

matrix balance age income education;

**run**;

**proc** **corr**;

title "Correlation Coeff";

var balance age income education;

**run**;

title "Regression";

**proc** **reg**;

model balance=age income education;

**run**;

title "Regression v2";

**proc** **reg**;

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.

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?

Yes, because as the year increase, the winning times decrease.

Negative slope

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

It appears that there is a negative linear trend which decreases at a faster rate than the men;s winning times. As the years increase, the winning times decrease drastically.

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

Women’s winning times has a greater slope. The winning time decreases faster over a shorter period of time.

