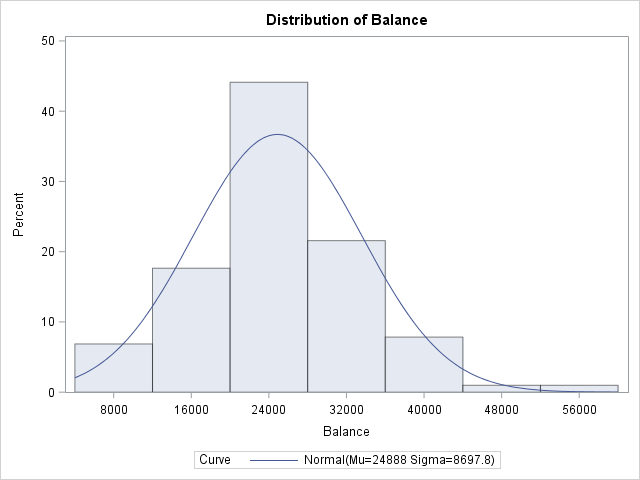
CSC-423 ASSIGNMENT 2

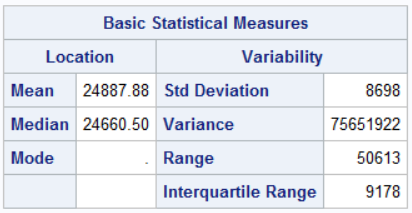
1. Histogram for Account Balance and its Analysis:

* The first thing that we can observe from the Histogram is that the distribution looks to be symmetric but on evaluation of the Basic Statistical Measures we can say that the **Mean > Median** which means that the **distribution is Positively Skewed or Right Skewed**.
* From the Quantiles table, we can infer that **Q1 = 20020** and **Q3 = 29198.00**.
* The **IQR** is **9178** and hence anything outside the **(Q1-1.5\*IQR, Q3+1.5\*IQR)** is an outlier. i.e Any data point which lies outside this range **(6253,42965)** will be considered as an outlier. By referring to the histogram and the Quantiles table we can say that **outliers are present** in the dataset because the **minimum is $5956.00** and **maximum is $56569.00** which are also considered as Outliers in this case.
* The Kurtosis value for this dataset is **< 3 (1.484)** which means that the **tails are more spread out** which can be seen in the histogram and the distribution curve.
* The peak value from the histogram is also near 24000 which means the **Mean** is around **~24000-25000** which is correct when we compare it with the actual mean of the dataset which is **$24887.88**. The same can be said about the Median also i.e **Median** is also near the peak according to the histogram which is also correct and the value can be inferred from the basic statistical measure table which is **$24660.50**.

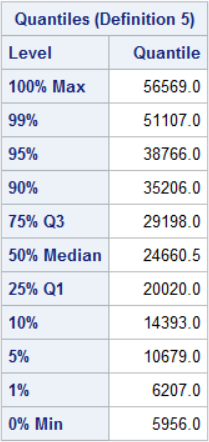


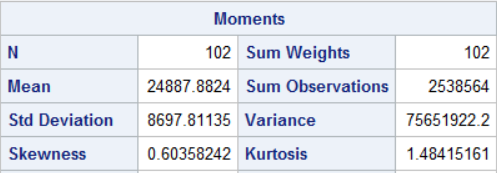
* From the Moments table, we can get the **Skewness** value which is **0.6035** which means that the **Right tail is longer than left** and the **distribution is Positively skewed**.
* The Range for this Data set is Maximum – Minimum which is 56569.00 – 5956.00 = 50604.00.

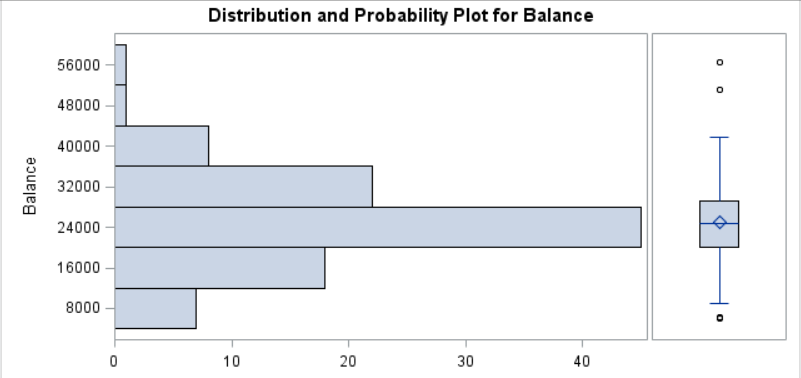
Basic Statistical Measure is:



Quantiles for Balance are:







It is also clear from the above boxplot the presence of outliers in this dataset. The outliers are represented using small black hollow circles in the box plot. They are above and below the Upper and Lower Whiskers.

The code to generate the above tables and plot is as follows:

title "Assignment 2 Dataset";

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

delimiter = '09'x;

getnames = yes;

datarow = **2**;

**run**;

title "Histogram on Account Balance";

**proc** **univariate** normal;

var Balance;

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

**run**;

title "Boxplot for Balance to check for outliers";

**proc** **univariate** plots;

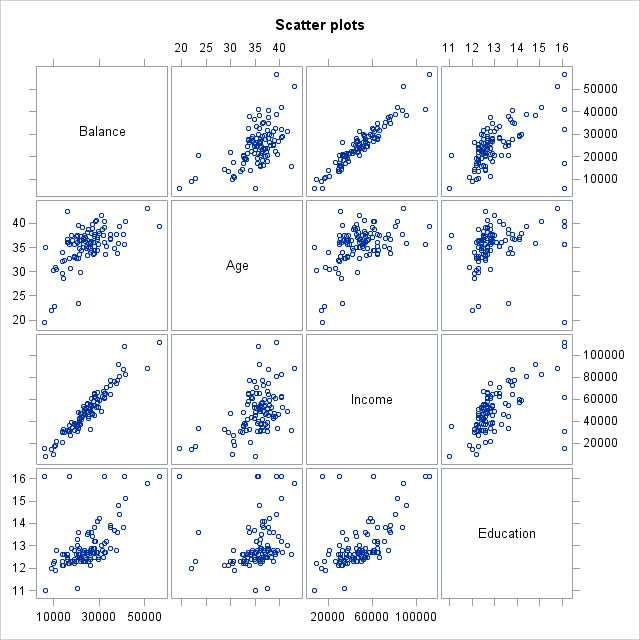
var Balance;

**run**;

1. Scatter plots of the variable. Here the **Dependent Variable is BALANCE** and the **Independent Variable are AGE, INCOME and EDUCATION**.

From the scatter plots given below, we can make the following observations:

* Balance and Education are slightly and positively correlated as the plot corresponding to these two variables is somewhat Random but they still have positive correlation which is weak.
* Balance and Income have very high positive correlation (.95). This means that the as Income increases the Balance also increases.
* Balance and Age are also positively correlated. Age is less strong when compared to Income but is stronger than Education when it comes to positive correlation to Balance.
* Based on the scatter plots we can come to an initial assumption that all the independent variable (Income, Age and Education) have influence over Balance and that the Independent variables are also positively related to the Dependent variable which is Balance.
* The correlation between Balance and Income, Age and Education appears to be **linear.**



The code for generating the scatter plot is as follows:

title "Scatter plots";

**proc** **sgscatter**;

MATRIX balance age income education;

**RUN**;

title "Individual plot";

**proc** **plot**;

plot balance\*(age income education);

**RUN**;

1. The code for generation of the Pearson’s Correlation coefficients is as follows:

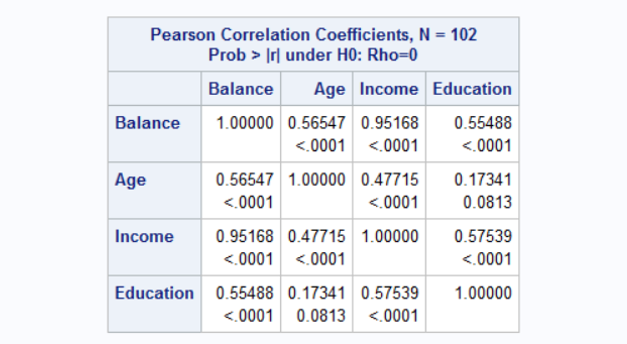
title "Correlation values between the Dependent and Independent Variables";

**proc** **corr**;

var balance age income education;

**run**;

The Correlation values which are generated for the dataset are as follows:

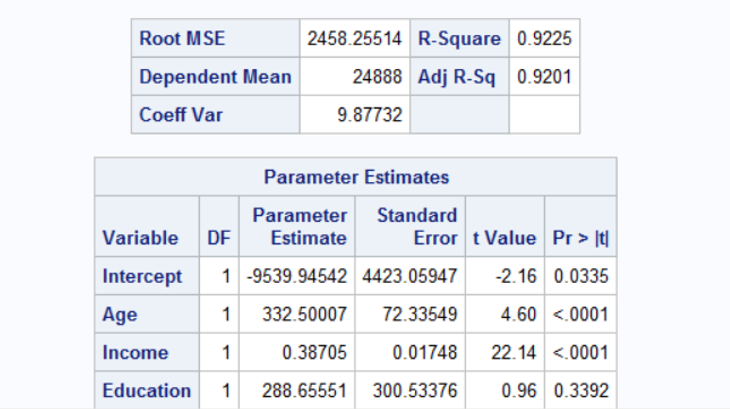


The following things can be inferred from the Correlation values which are mentioned above:

* Age and Balance are positively correlated but the correlation isn’t very strongly associated. 56.54% of Variance in Balance can be explained using Age.
* Income and Balance are strongly and positively associated. Based on the given dataset we can say that Income has the strongest influence over Balance. 95.16% of variance in Balance can be explained using Income.
* Education and Balance are positively associated, but among the three independent variables Education is the weakest in terms of correlation. 55.48% of variance in Balance can be explained using Education.

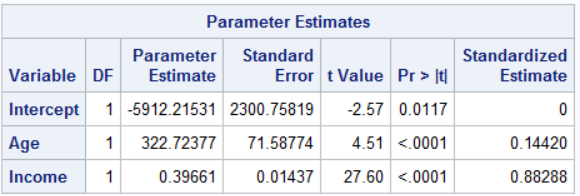
1. **Dependent variable** in this case is **BALANCE** and the **Independent Variables** are **AGE**, **INCOME** and **EDUCATION.** This is so because we are trying to create a model which will predict the Balance when the values for Age, Income and Education are provided.
2. When we first create the Regression model using the three independent variable we get the following table which has the parameter estimates for all the three variables and also the intercept value. **The initial model without taking Alpha into consideration** is as follows:

**BALANCE = -9539.9454 + 332.50007\*AGE + 0.38705\*INCOME + 288.655\*EDUCATION**



But it is mentioned that the Alpha value is 0.05 but for Education the p-value is .3392 which is greater than the Alpha value which means that education is not significant for predicting the Balance and hence we won’t consider it in the next model which we create.

After creating the new model, we get the following parameter estimates:



The **new regression model** is as follows:

**BALANCE = -5912.21531 + 322.723\*AGE + 0.39661\*INCOME**

By adding standardized estimate, we can say that the **highest influencer** between Age and Income is **Income**.

The code to generate the above models is as follows:

title "Regression model to fit the variables";

**proc** **reg**;

model balance=age income education;

**run**;

title "Regression model to fit the variables excluding Education";

**proc** **reg**;

model balance=age income/ stb;

**run**;

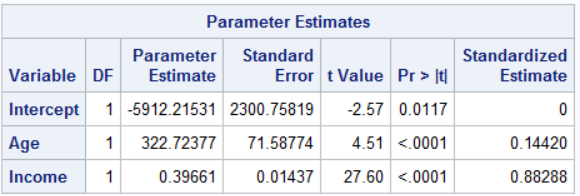
1. As mentioned in the previous question, the Education variable is not a significant predictor because it had p-value (0.3392) which was greater than the Alpha value (0.05) and hence it was removed and a new model was created using the following code which is followed by the parameter estimates value and then the newly fitted Regression model:

title "Regression model to fit the variables excluding Education";

**proc** **reg**;

model balance=age income/ stb;

**run**;

Parameter Estimates: 

Based on the values obtained from the Parameter Estimates we get the following Fitted Regression Model:

**BALANCE = -5912.21531 + 322.723\*AGE + 0.39661\*INCOME**

In this I’ve also included the Standardized Estimate to show that Income has the strongest influence over Balance.

1. The parameter estimates from the 2 model will be considered because it is the optimized version of initial model in which Education was also taken into consideration.

For age the beta value is 322.723 i.e. if we increase the value of **Age by 1 year** then the **average balance would be increased by $322.72377.**

The beta value for Income is 0.39661 i.e. if we increase the **income by 1$** then the **average** **balance would increase by $0.396**.

1. The following are the R squared values for the second model in which we ignored the Education variable:



We will consider the adjusted R squared value as it is more accurate because its value is affected if we take into consideration attributes which are not significant. The adjusted R squared value tells us that **92.02% of the variance in Balance** could be explained using the other independent variables like **Income** and **Age**.

The RootMSE (Root Mean Squared Error) gives the overall standard error for the Balance which is the sum of the residuals of the Balance VS the Various Regressors. The lower the RMSE value the better it is.

In the first model that we created RMSE value was 2458.255 and in the second model the RMSE value is 2457.29 and hence this gives us another reason why we should prefer second model in which education is excluded.

The code for this is as follows:

title "Regression model to fit the variables excluding Education";

**proc** **reg**;

model balance=age income/ stb;

**run**;

1. The final model which we had computed is as follows:

**BALANCE = -5912.21531 + 322.723\*AGE + 0.39661\*INCOME**

If Age=34.8 years and Income = $42,401, then the average Balance can be calculated by substituting these values in the above equation.

**BALANCE = -5912.21531 + 322.723\*34.8 + 0.39661\*42,401**

**BALANCE = $22,135.207**

The observed Balance value for the above-mentioned values is **$21,572** and the predicted value is **$22,135.20.**

The model prediction error is the Observed – Predicted which is

**$21,572 - $22,135.20** = **-$563.20**

1. The whole code by which the above things could be generated is as follows:

title "Assignment 2 Dataset";

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

delimiter = '09'x;

getnames = yes;

datarow = **2**;

**run**;

title "Histogram on Account Balance";

**proc** **univariate** normal;

var Balance;

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

**run**;

title "Boxplot for Balance to check for outliers";

**proc** **univariate** plots;

var Balance;

**run**;

title "Scatter plots";

**proc** **sgscatter**;

MATRIX balance age income education;

**RUN**;

title "Individual plot";

**proc** **plot**;

plot balance\*(age income education);

**RUN**;

title "Correlation values between the Dependent and Independent Variables";

**proc** **corr**;

var balance age income education;

**run**;

title "Regression model to fit the variables";

**proc** **reg**;

model balance=age income education;

**run**;

title "Regression model to fit the variables excluding Education";

**proc** **reg**;

model balance=age income;

**run**;

title "Regression model to fit the variables excluding Education and by standardizing the remaining variables";

**proc** **reg**;

model balance=age income / stb;

**run**;

**PROBLEM – 2**

1. Yes, we can say that men’s winning time **follows a linear trend** and the slope of the **regression line is negative** which means that as time progressed the Winning time decreased i.e. both the **time(years) and Winning time are negatively correlated**.
2. Yes, we can say that the winning times for women also **follows a linear trend** and the **slope is also negative** in this case which means as time(years) passed by the winning times kept decreasing i.e. there is **negative correlation between time(years) and winning time**.
3. By visual inspection we can say that the Women’s winning time has a steeper slope and hence its value would be larger when compared to the men’s winning time slope.

The other thing which I would say to justify my stand is that during the time period between 1980 – 2000 Men’s winning time decreased from roughly 135 to 120 which is around ~15 minutes but during the same time Women’s winning time decreased form around 157-160 to around 120 minutes which is around roughly ~35-40 minutes. This also shows that Women’s slope is greater than Men’s.

