**DATA ANALYSIS AND REGRESSION**

**Assignment-1** | **Total points: 25 pts for DSC 323; 30 pts for DSC 423**

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 (7-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.

• **Make sure to double check your submissions. After you submit the assignment, log out of D2L, log back in, and click on your submission to see if you submitted the right file(s) and it is the correct version. Wrong submissions will not be graded.**

**PROBLEM 1 [10 pts] – To be answered by everyone**

The file election.txt attached to this assignment provides data acquired from census records selected counties in the U.S. who voted in the elections. The data show

County – Name of the county

PctVoted – Percentage of people voted

MedianAge – Median age of the voters in that county

MeanIncome – Mean income in U.S. Dollars in that county

PctUnemployment – Percentage of people unemployed in that county

PopulationDensity – Population density (Population divided by square miles) in that county Gender – Dominant gender of the people voted in that county

***Use SAS to compute the analysis below. All the functions are in either the code for the Lab Session-1 we did in class (see code that was posted on D2L). This is the first assignment, and for many of you it may be the first time you use SAS outside of the first lab session. So if you run into an error, post a message on the discussion board or contact me. Make sure to include your code in the message.***

In this exercise you are asked to get the data into a SAS dataset and perform basic exploratory analysis of the data to analyze the characteristics of people voted.

a) Open the dataset and examine the data. Answer the following:

1. How many Observations are there? 2704

2. How many fields are there? 7

3. Which fields are numerical? 5

4. Which fields are text? 2

b) Write the SAS code to create the SAS dataset using either IMPORT or INFILE statement. If you are using INFILE statement, pay attention to the text fields while writing your code.

**PROC** **IMPORT** datafile = "C:\Users\JDORETTI\Downloads\election.txt" out=election replace;

delimiter = '09'x;

getnames = YES;

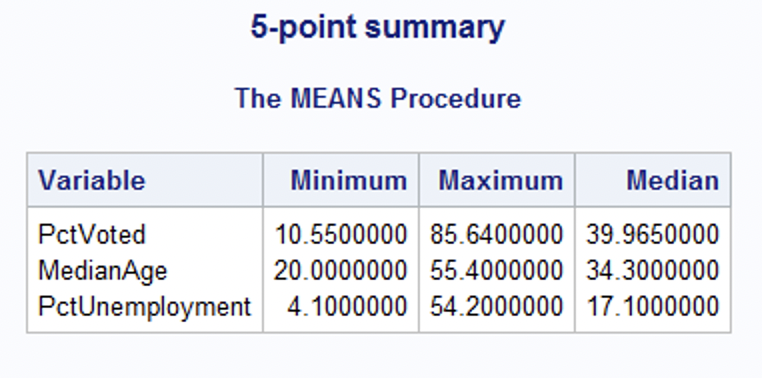
datarow = **2**;

**RUN**;

c) Run a PROC PRINT to print your dataset in SAS. Do a print screen, copy and paste the first 5 observations of the output.

| OBS | County | PctVoted | MedianAge | MeanINCOME | PCTUnemployment | PopulationDensity | Gender |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Autauga, AL | 31.92 | 31.7 | 57621 | 17.4 | 60.4 | M |
| 2 | Baldwin, AL | 27.24 | 35.5 | 84933 | 15.8 | 66.5 | M |
| 3 | Barbour, AL | 47.36 | 32.8 | 83654 | 27.2 | 28.8 | M |
| 4 | Blount, AL | 33.92 | 34.5 | 61247 | 17.2 | 62.4 | M |
| 5 | Bullock, AL | 68.67 | 31.7 | 75723 | 35.2 | 17.6 | M |

d) What is the 5-point summary numbers for percentage of people voted, median age and percent unemployment? The 5-point summary numbers are min, max, median or 50% percentile, Q1 and Q3. Include the output. Discuss your findings based on the values you see.



For the percentage of people who voted, the smallest county only saw 10.6% (rounded) people vote. This is not good because in an ideal election you would have 100% of the people vote. While the maximum shows 85.6% (rounded) people voted. This is higher than the median that is below 50%. This shows that the majority of the 2704 entries had a voting percentage under 50%.

Median Age shows that the majority of voters are in their early adult years to near retirement age. This age range would be from 18 - 65 - supported by the minimum age being 20 and maximum 55.4. That being said, the median age for votes is around 34 years old. This means that most voters are younger. Overall, these counties are seeing younger voters being the majority of voters.

Percent unemployment tells us that it doesn't matter if people are jobless or not - people will still vote. This is supported by the edge cases - minimum unemployment of 4.1% is low and a maximum of 54.2% is really high. Unemployment in these counties do seem to be higher as the median unemployment is 17.1%. The unemployment seems to be higher and this can show a lack of economic prosperity.

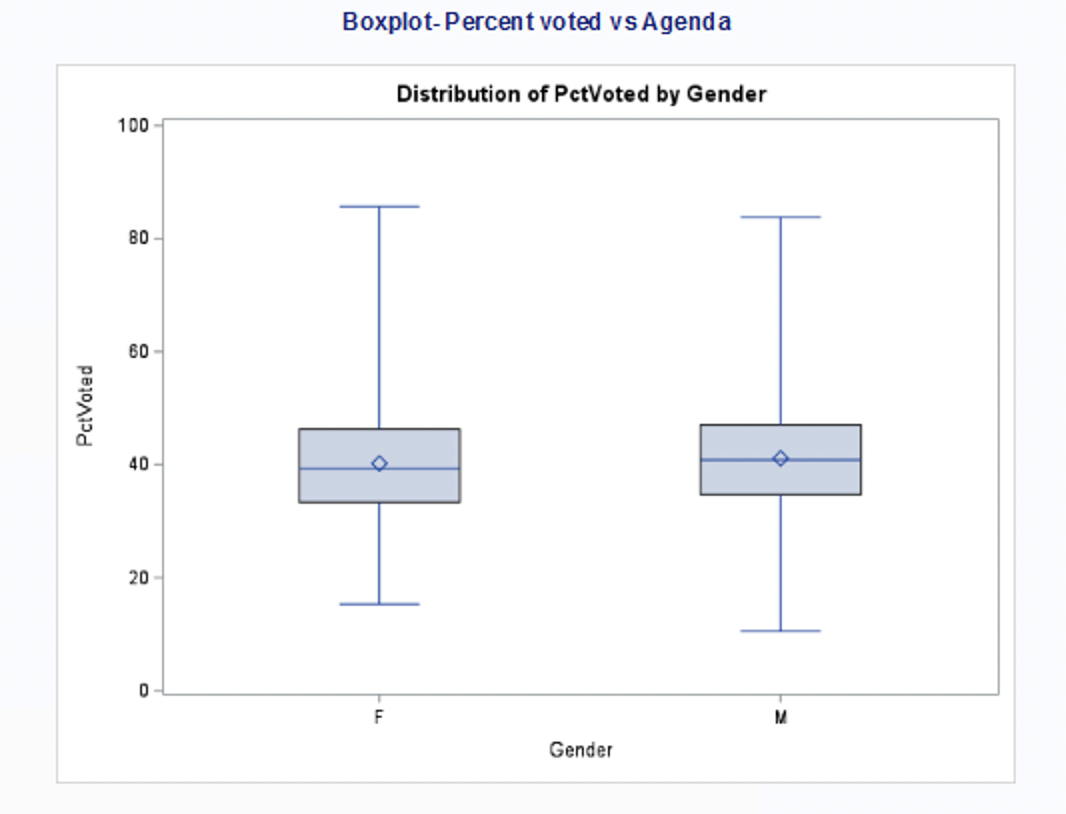
e) Create a histogram to analyze the percent of people voted. Include the histogram output. Using the histogram and the 5-point summary from the previous question, analyze the histogram. Discuss your findings using the 5-point summary, normality (normal or skewed), modality, outliers, etc.?

This is a bimodal, with peaks at 36% and 40%, slightly positively skewed histogram. This makes the right tail longer. The outliers are 84% or more of people who voted in these counties. The histogram has the same minimum and maximum value as the 5-point summary. This supports my claim that the majority of counties see a voter turnout of less than 50%.



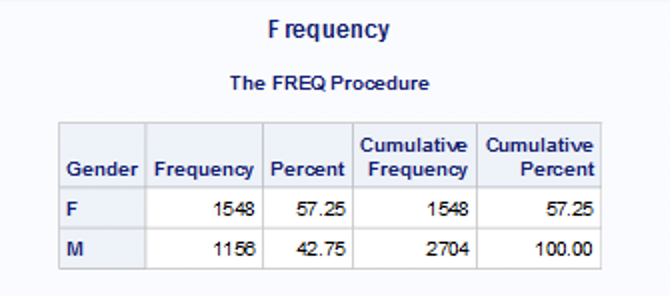
f) Create a boxplot to analyze percentage of people voted by gender. Include the output. What can you say about the gender and voting patterns? Discuss your findings using the boxplot.

With the box plot you can see that both genders perform almost the same. It seems that the male gender is performing a little bit better. The median percentage of voters is slightly lower than the male median percentage of voters. The male percentage of voters is symmetric while the female percentage of voters is slightly skewed to the right.



g) What is the gender breakdown in this dataset? (Hint: use PROC FREQ). Include the output. Which is the predominant gender in this dataset? Explain how you came to this conclusion.

In the data set - females are the majority of the voters. The data consists of 57.25% female and 42.75% male. This is close enough to a 50-50 split between genders to consider the data reliable. However, the closer to 50-50 the better.



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

\* Import election data;

**PROC** **IMPORT** datafile = "C:\Users\JDORETTI\Downloads\election.txt" out=election replace;

delimiter = '09'x;

getnames = YES;

datarow = **2**;

**RUN**;

\* Print command;

**PROC** **PRINT;**

**RUN**;

\*5-point summary;

TITLE "5-point summary";

**PROC** **MEANS** min max median;

VAR PctVoted MedianAge PctUnemployment;

**RUN**;

\*Histogram;

TITLE "Histogram";

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

VAR PctVoted;

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

**RUN**;

\*Boxplot;

TITLE "Boxplot- Percent voted vs Agenda";

\*sort;

**PROC** **SORT**;

BY Gender;

**RUN**;

\*Boxplot;

**PROC** **BOXPLOT**;

PLOT PCtVoted \* Gender;

**RUN**;

\*Frequency;

TITLE "Frequency";

**PROC** **FREQ**;

TABLES Gender;

**RUN**;

**PROBLEM 2 [15 pts] – To be answered by everyone**

The file unemployment.txt attached to this assignment provides data acquired from census records for different zip codes. The data show

− median age of the population (Age)

− median income (Income) in $

− average bank balance (Balance) in $

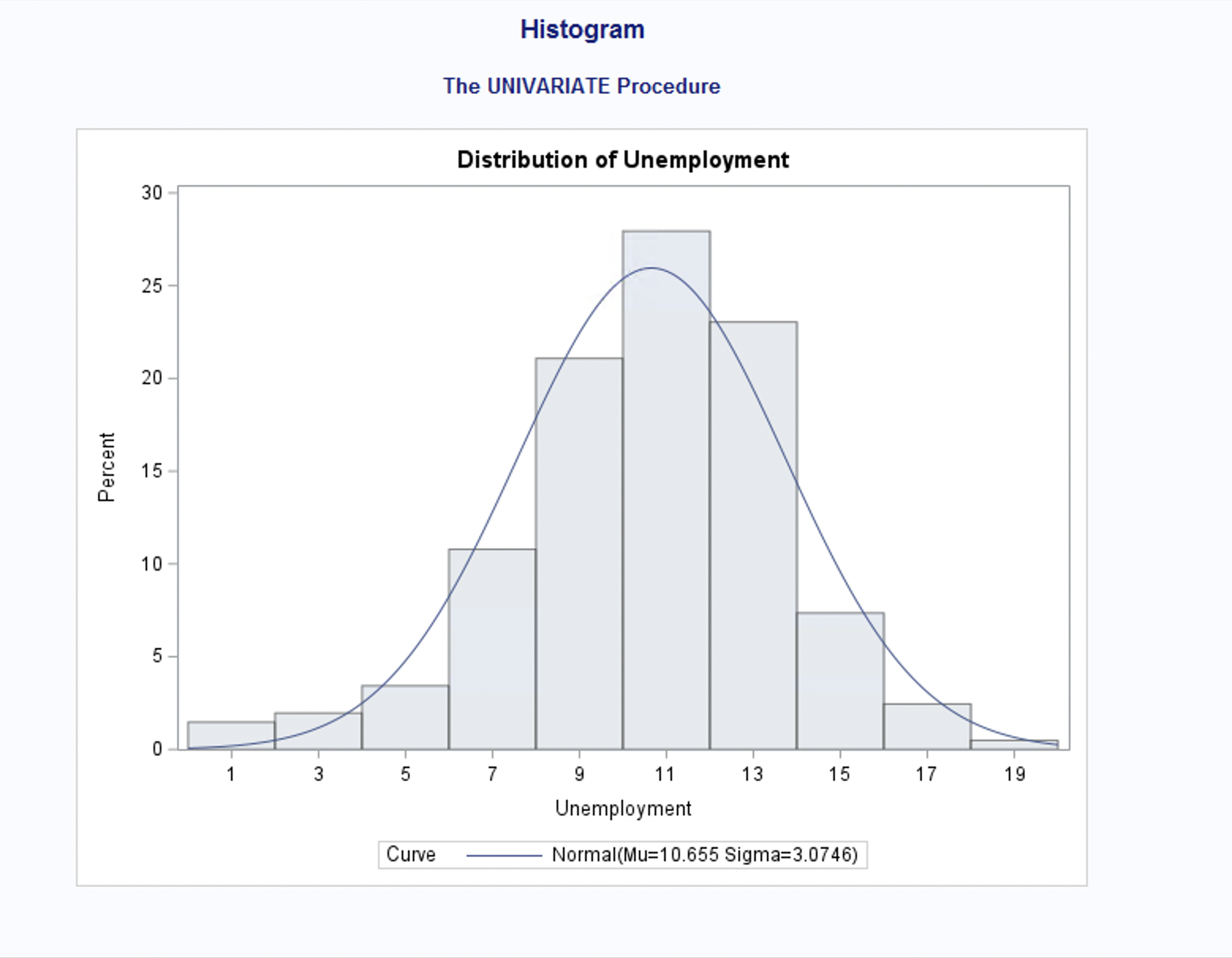
− median years of education (Education) in years

− percent unemployment (Unemployment) in %

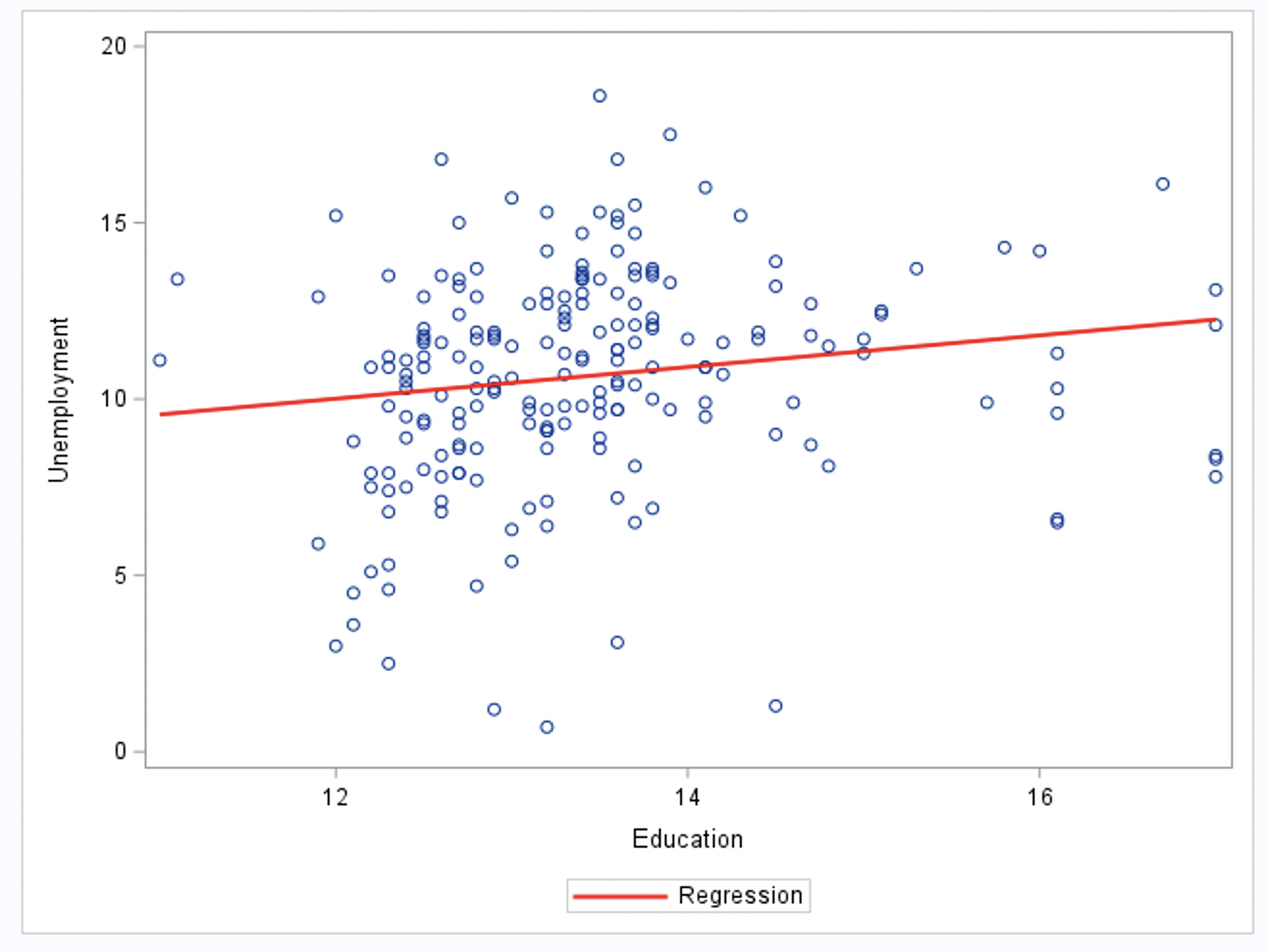
In this exercise you are asked to apply regression analysis techniques to describe the effect of age education, income and average account balance on unemployment.

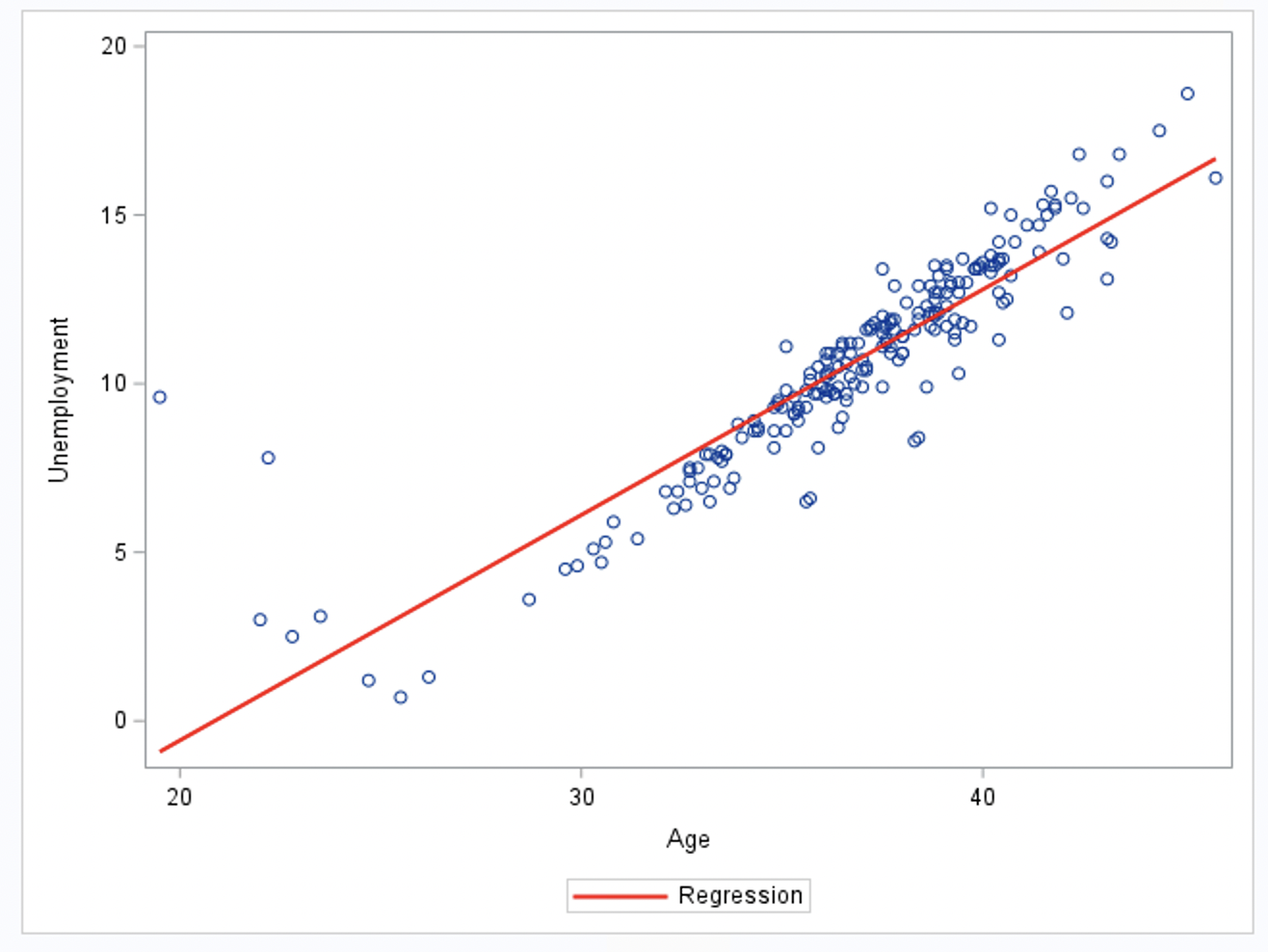
a) Analyze the distribution of unemployment using histogram, and compute appropriate descriptive statistics. Write a paragraph describing distribution of unemployment and use appropriate descriptive statistics to describe center and spread of the distribution. Discuss your findings. Also, explain if you see any outliers, if so, why do you think they are outliers? Include the histogram.

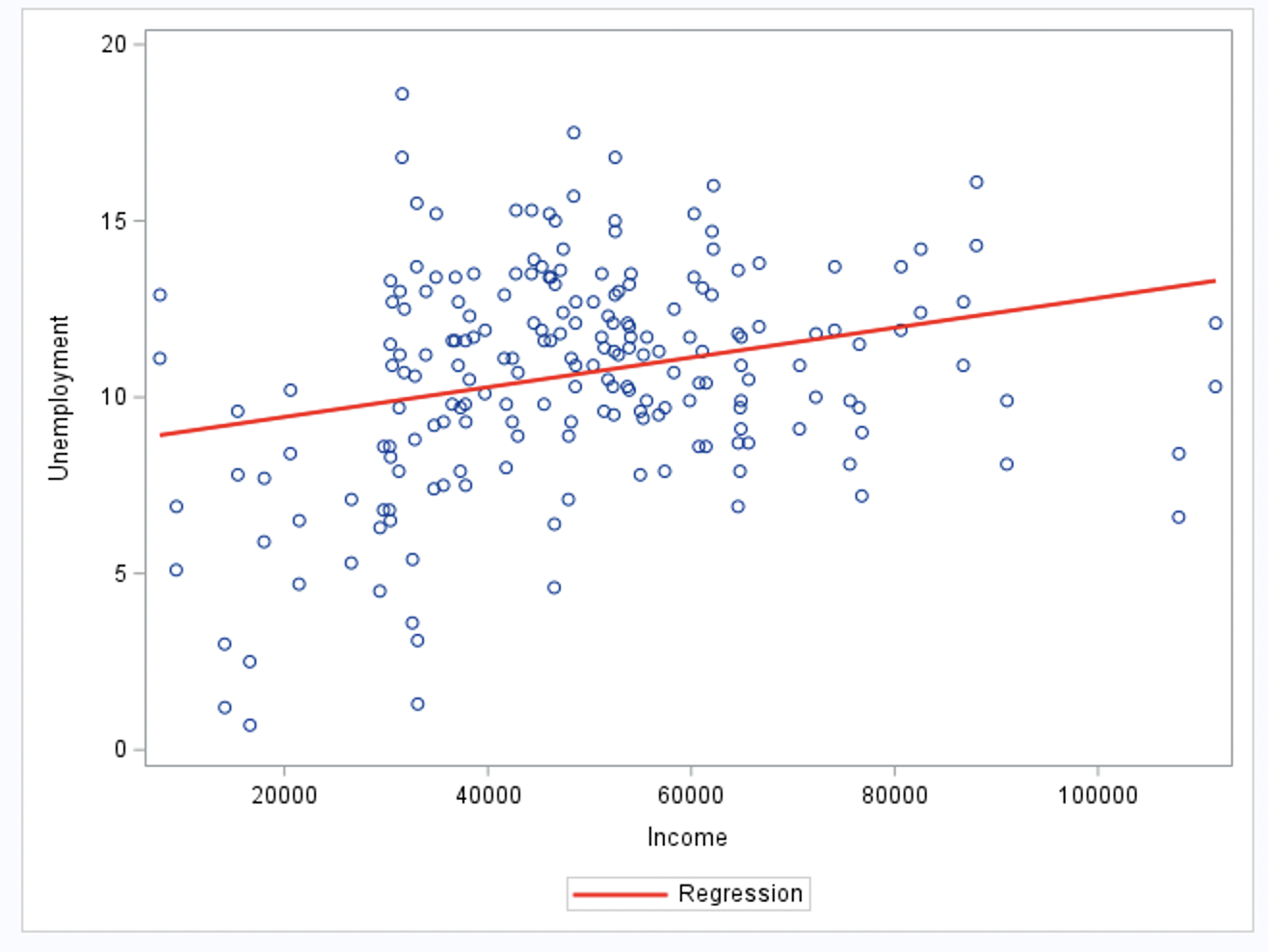
This histogram is a singular modal and very slightly skewed negative. The outliers seem to be 1-5 and 17-19 as those values do not correlate to the 25th 50th and 75th quartiles. This shows that the mean and the mode will be closer in numbers as the spread of unemployment is centralized within these percentiles. This also shows that out of the zipcodes studies that the percentage of unemployment is around the ranges of approximately 10% to 27%. This is a large range and is well above the average unemployment for the USA.

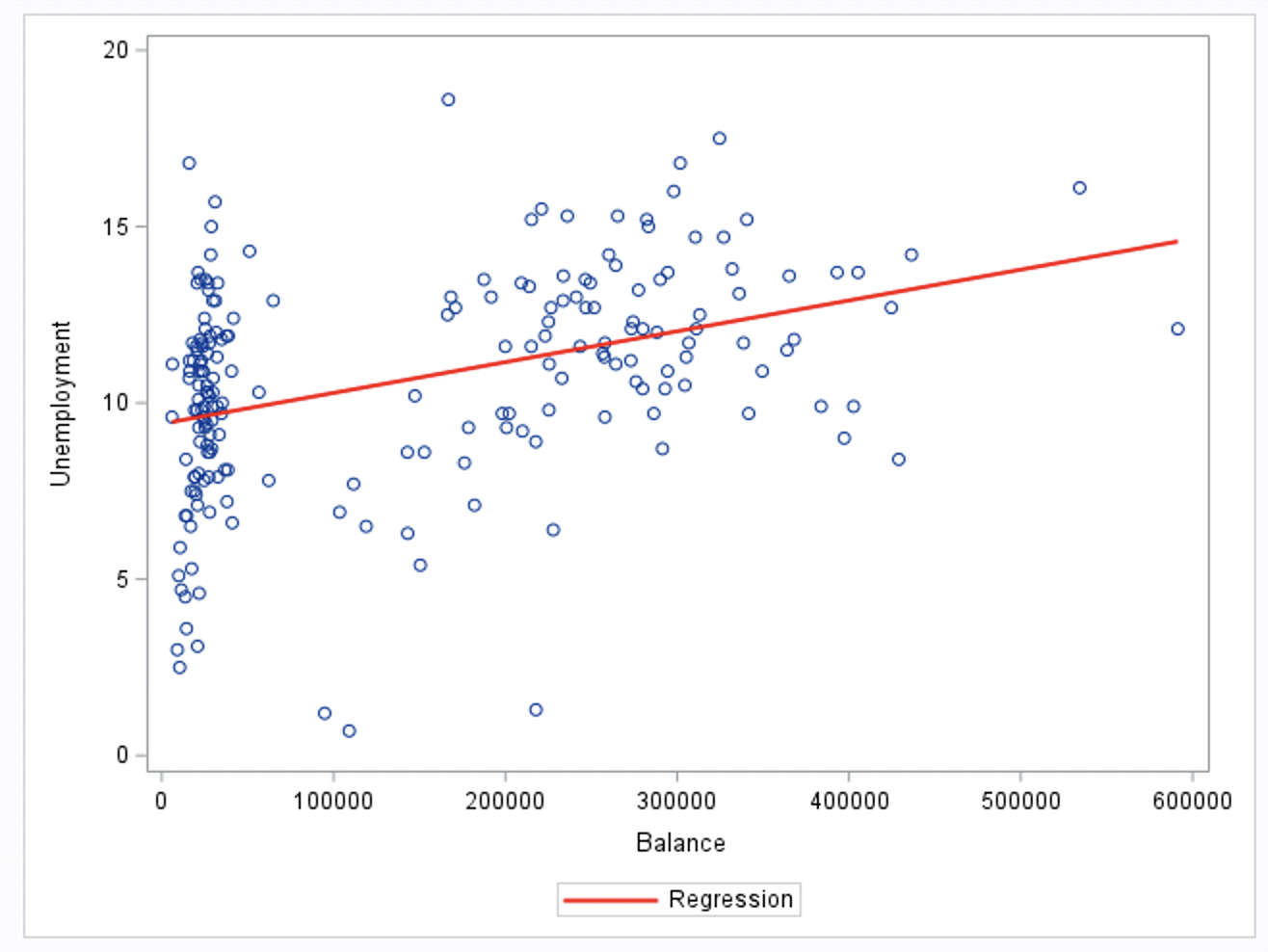


b) Create scatterplots to visualize the associations between unemployment and the other variables. Include the scatterplot. Discuss the patterns and association displayed by the scatterplot.

This scatter plot showing Unemployment to Education shows that these two do not have a correlation. There are lots of outliers at all points. This tells us that anyone regardless of education can be unemployed.

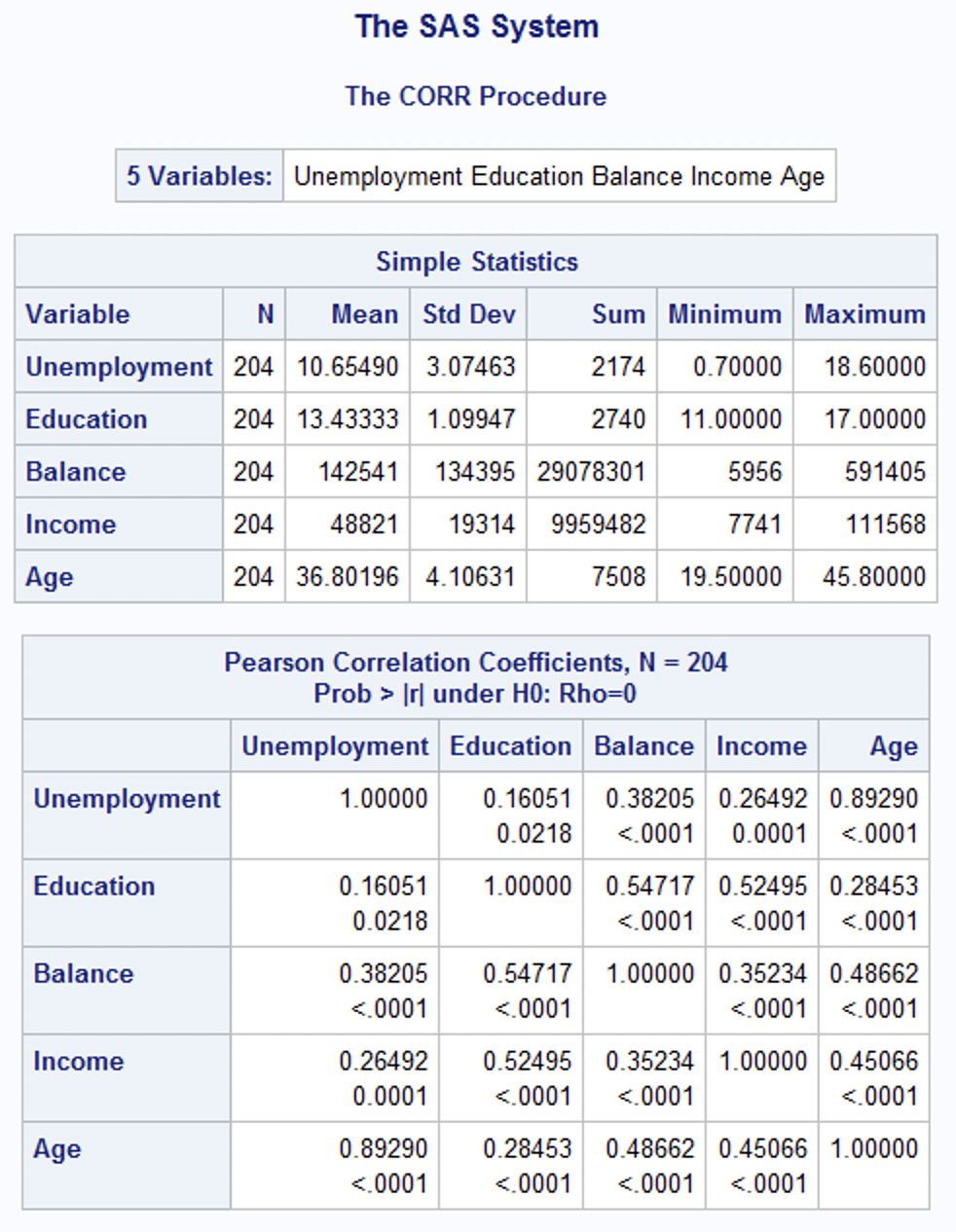
This scatter plot showing Unemployment to age tells us a lot. The line has a high positive correlation; telling us that the older you are the more likely you are to be unemployed. This graph makes sense because as people get older they tend to retire so you will see more unemployment. It also makes sense that the outliers at the younger ages - like 18 years old - see higher unemployment as they are just figuring out to do for work.

This scatter plot has no correlation - telling us that regardless of people’s income, anyone can be unemployed.



This graph has a slight/low positive correlation that can also be looked at as no correlation. That is because people can become unemployed and stay unemployed for many different reasons. Their balance in their bank accounts does not reflect their employment status.

c) Compute correlation values of unemployment vs the other variables. Include the relevant output that shows the correlation values. Interpret the correlation values, and discuss which pairs of variables appear to be strongly/not strongly associated.

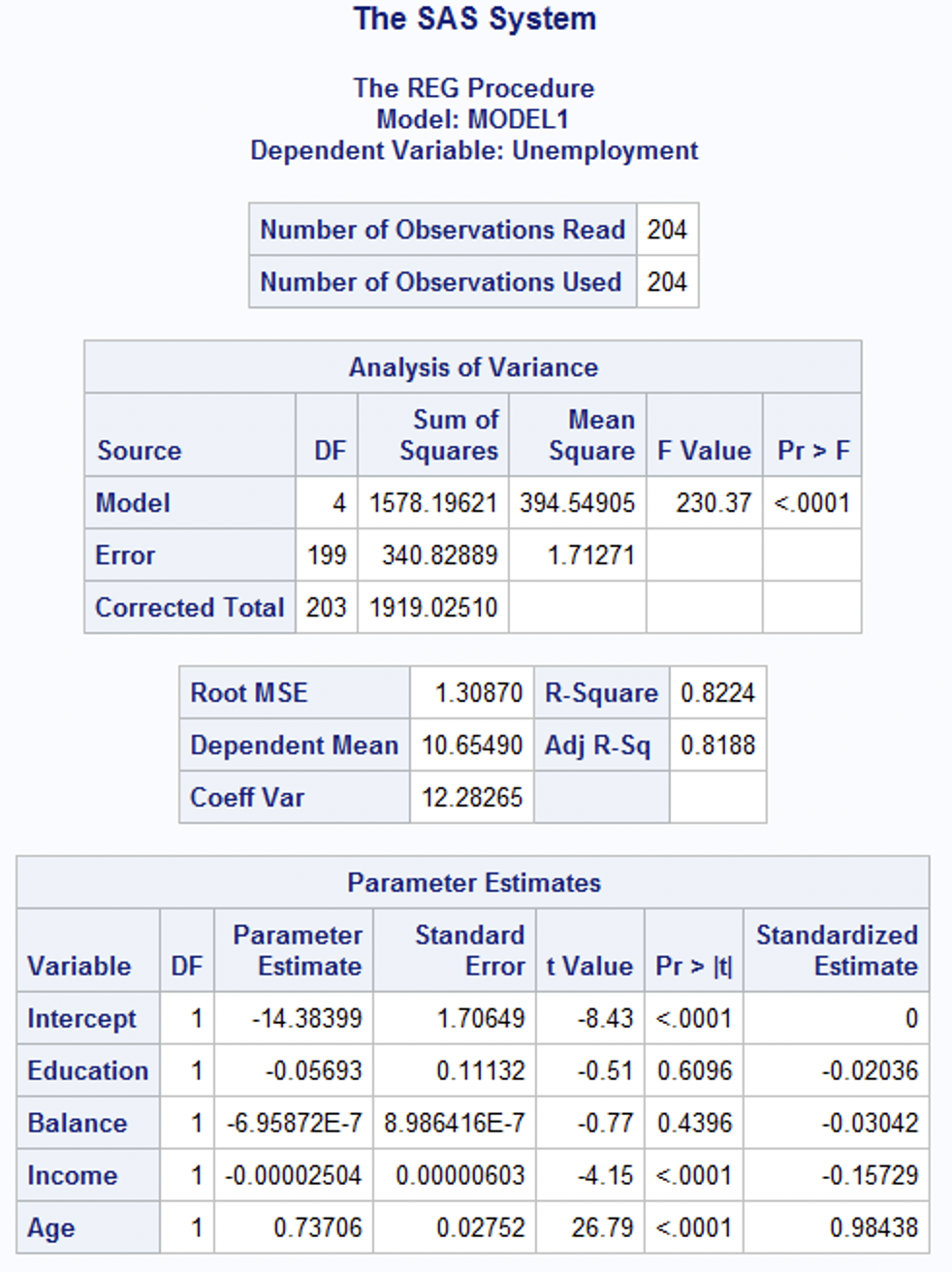


Unemployment seems to be highly correlated to Age and slightly to balance. Education and income also have a small correlation to unemployment. While education is slightly correlated to both balance and income. On top of all that balance and income seem slightly correlated to age.

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

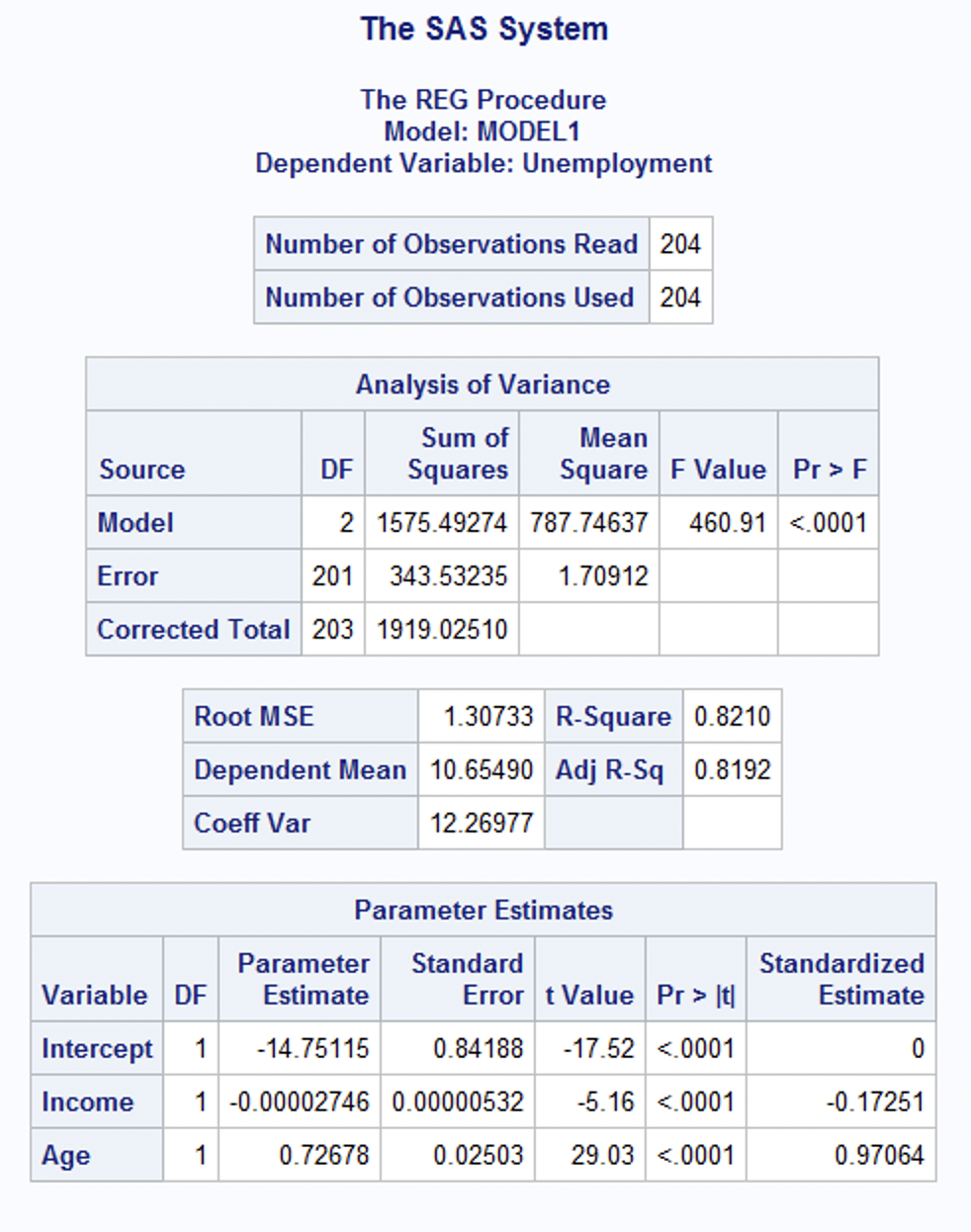
The dependent variable is Unemployment and the independent variable is Education, Age, Income, and Balance.

e) Use SAS to run the full regression model to predict unemployment from age, education, income and balance. Include the relevant regression output. Analyze the model. Which predictors have a significant effect on unemployment, explain why you came to the conclusion that they are significant or insignificant?



To measure impact on unemployment we will be using Standardized Estimate and taking the absolute value of each number. Based on this - Age and Income are the only ones that have any impact. Age with a large impact and income only a small impact. Education and balance have almost no impact at all.

f) If one of the predictors is not significant, remove it from the model and refit the new regression model. Include the relevant regression output. Make sure you include all the steps and outputs at arriving at the final model. Write the expression of the newly fitted regression model.



This model shows the two most impactful measures on unemployment. The new model shows a few changes to the standardized estimate. Age’s impact has drop slightly and Income’s has risen slightly as well.

g) Interpret the value of the parameter estimates for the variables in the model.

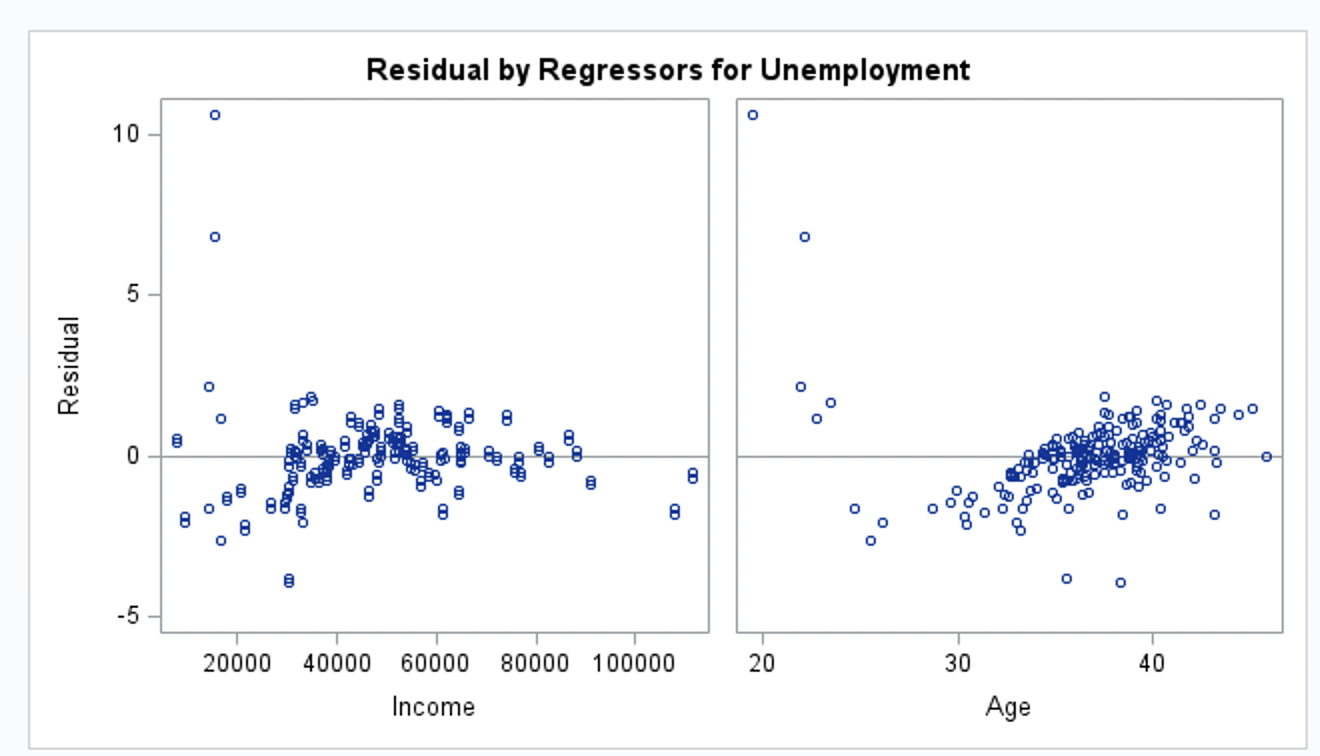
The Parameter estimate is to determine importance as well. Here we will take the absolute value as well. Age ranked the highest in importance to the unemployment rate. Income shows to have no importance to unemployment.

h) Include the portion of the output that includes the R2 and Adj-R2 coefficient values. Report the value for the R2 and Adj-R2 coefficient and describe what it indicates. What can you conclude?

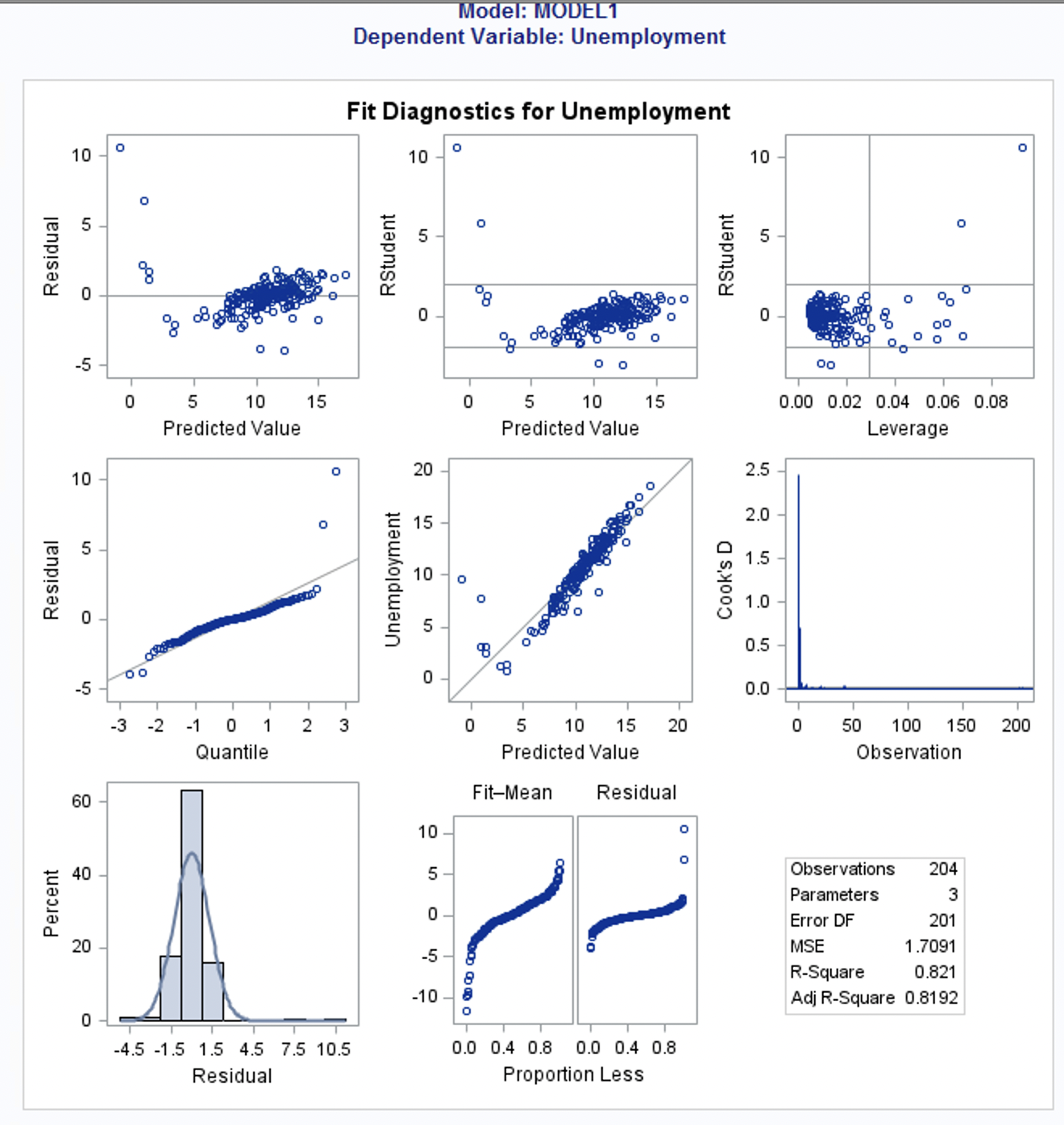
See (f) for R2 and Adj-R2. The closer to 1 R2 and Adj-R2 are the better the Y is at explaining X. So R2 being at .8210 and Adj-R2 being at .8192 means that income and age have a good regression line in response to unemployment.

i) According to census data, the population for a certain zip code area has median age equal to 44.2 years, median education equal to 11.5 years, median income equal to $51,324 and balance is $34,200.

(i) Use the final model computed in step (f) above to compute the predicted unemployment for the zip code area.



(ii) If the observed unemployment for the zip code area is 13.5%, what’s the model prediction error?



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

\* Import unemployment data;

**PROC** **IMPORT** datafile = "C:\Users\JDORETTI\Downloads\unemployment.txt" out=unemployment replace;

delimiter = '09'x;

getnames = YES;

datarow = **2**;

**RUN**;

\*Histogram;

TITLE "Histogram";

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

VAR Unemployment;

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

**RUN**;

**\*Scatter plot - Y:UnEmp X:Edu;**

**Title = "Scatter plot - Y:UnEmp X:Edu";**

**PROC sgplot;**

**reg y=Unemployment x=Education / lineattrs = (color = red thickness = 2);**

**RUN**;

**\*Scatter plot - Y:UnEmp X:Balance;**

**Title = "Scatter plot - Y:UnEmp X:Balance";**

**PROC sgplot;**

**reg y=Unemployment x=Balance / lineattrs = (color = red thickness = 2);**

**RUN**;

**\*Scatter plot - Y:UnEmp X:Income;**

**Title = "Scatter plot - Y:UnEmp X:Income";**

**PROC sgplot;**

**reg y=Unemployment x=Income / lineattrs = (color = red thickness = 2);**

**RUN**;

**\*Scatter plot - Y:UnEmp X:Age;**

**Title = "Scatter plot - Y:UnEmp X:Age";**

**PROC sgplot;**

**reg y=Unemployment x=Age / lineattrs = (color = red thickness = 2);**

**RUN**;

**PROC** **corr**;

VAR Unemployment Education Balance Income Age;

**RUN**;

**PROC** **reg**;

model Unemployment=Education Balance Income Age / stb;

**RUN**;

**PROC** **reg**;

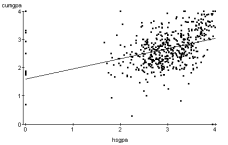
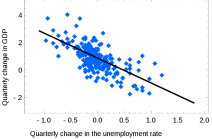
model Unemployment=Income Age / stb;

**RUN**;

**PROBLEM 3 [5 pts] – For Graduate Students ONLY**

1. Based on the two plots you see in figures 1 and 2, which one will produce a more accurate prediction? Need to provide explanation as to why you think it produce a better model.

**Figure-1 Figure-2**

** **2. Explain the type of correlation you see on both figures.