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

**Assignment-3** | **Total Points: 25pts for CSC 423 / 20pts for CSC 324**

**Due Date: 02/06/2018 by 11:59 pm**

***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 [5 pts] – ONLY for GRADUATES**

A university career center collects information on the job status and starting salary of graduating seniors. Data recently collected over a two-year period included over 900 seniors who had found employment at the time of graduation. The information was used to model starting salary Y as a function of two qualitative independent variables: COLLEGE at four levels {Business, Engineering, Liberal Arts, Nursing} and SEX (male and female).

1. Define the dummy variables to include college (use Business as your baseline) in a regression model for starting salary Y

|  |  |
| --- | --- |
| Sex | Z1 |
| M | 1 |
| F | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
| College | Z2 | Z3 | Z4 |
| Business | 0 | 0 | 0 |
| Engineering | 1 | 0 | 0 |
| Liberal Arts | 0 | 1 | 0 |
| Nursing | 0 | 0 | 1 |

1. Write down the general regression model relating starting salary Y to both college and sex.

Salary = B0 + B1Z1 + B2Z2 + B2Z3 + B3Z4

1. How would your model change if students in Engineering have the same starting salary as students in Business? Show the final regression model.

There would be no need for Z2, because it would also be a base. Since engineering and business start out the same, liberal arts and nursing would be compared to them both.

Salary = B0 + B1Z1 + B2Z3 + B3Z4

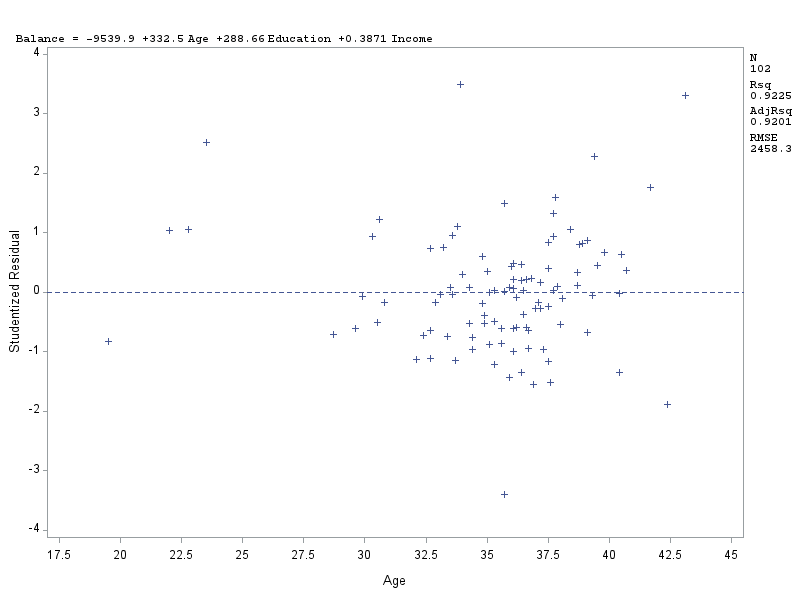
**Problem 2 [5 pts] – to be answered by everyone**

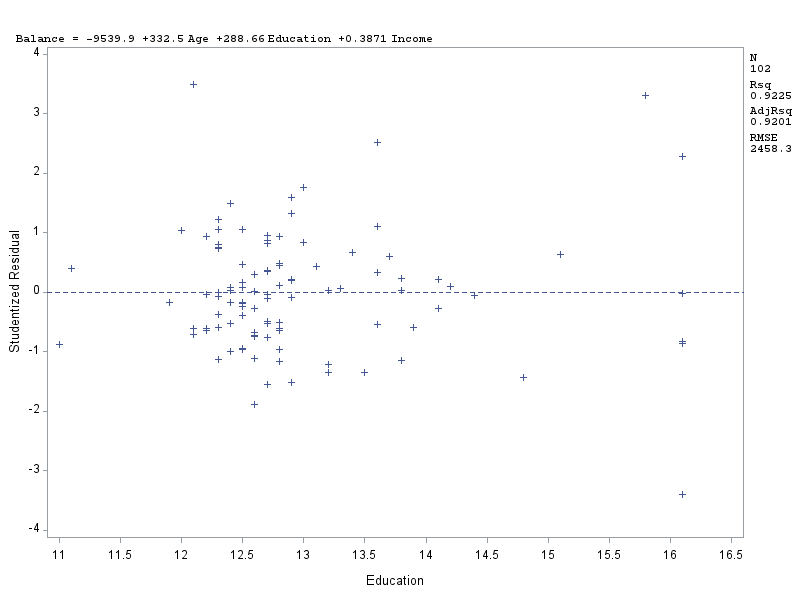
You will continue the analysis of the banking.txt dataset that was analyzed in Assignment 2 – data file is attached.

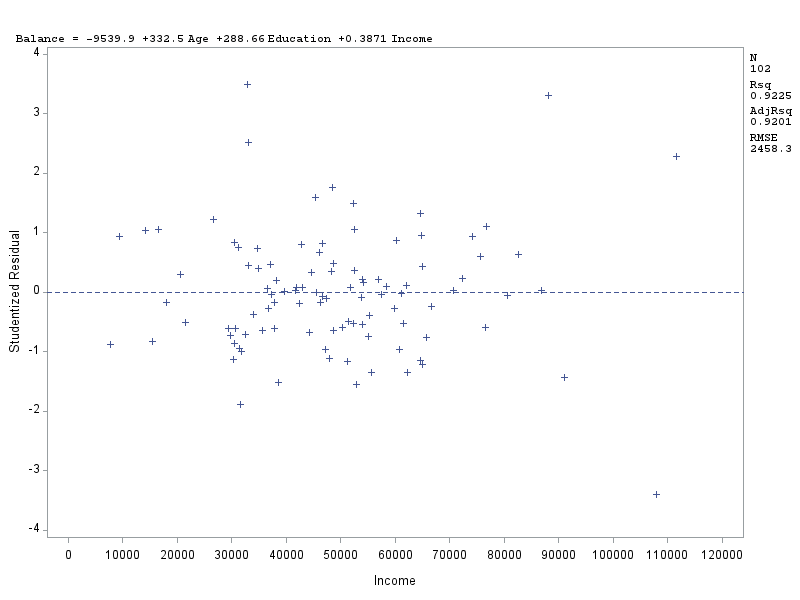
1. Analyze the residuals of the regression model you found in your previous assignment. Include the residual plots. Discuss your findings.

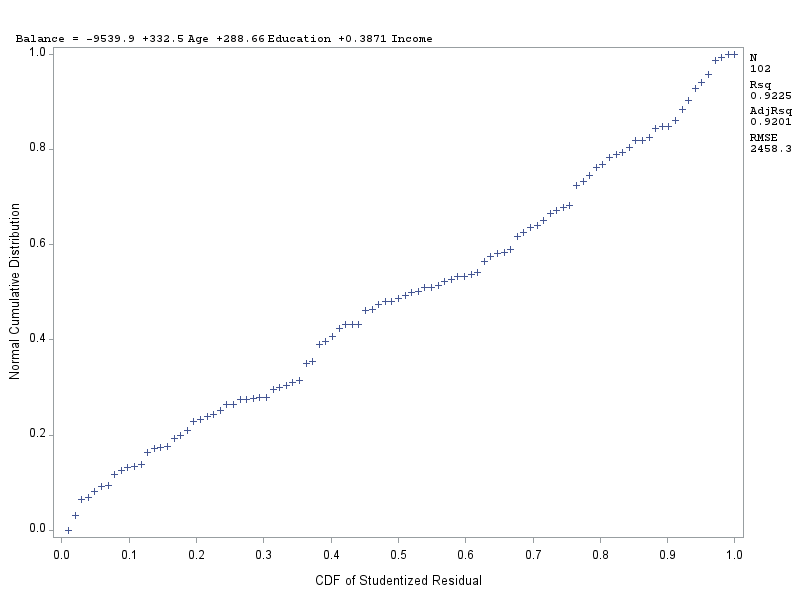
Assumptions:

1. Constant variance is violated
2. Linearity is violated which can be seen easily in the Age and Education graphs.
3. Independence is violated because the graphs show a pattern and because constant variance is violated.
4. Normal distribution is NOT violated.









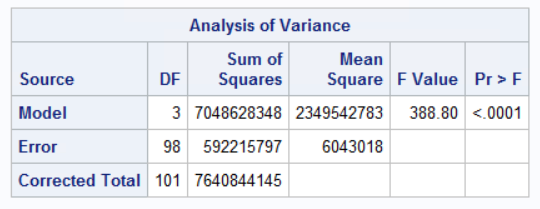
1. Conduct a global F-test for overall model adequacy. Write down the test hypotheses and test statistic and discuss conclusions. Include the relevant output.

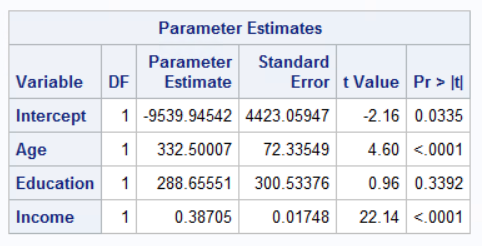
Null Hypothesis: Bj = 0

Alt Hypothesis: Bj <> 0

F-value = 388.80.

Based on the P-value there is enough evidence to reject the null hypothesis. We can determine that this model is a good fit. However, based on the x-variable p-values, the Education variable could be removed from the model to increase the Adj-R^2 value.





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;

/\*regression analysis fitting a linear model

and residual analysis\*/

proc reg corr;

model balance = age education income;

\* Residual plot: residuals vs x-variables;

plot student.\*( age education income);

\* Residual plot: residuals vs predicted values;

plot student.\*predicted.;

\* Normal probability plot or QQ plot;

plot npp.\*student.;

run;

**Problem 3 [15pts] – to be answered by everyone**

A national homebuilder builds single-family homes and condominium style townhouses.

The file housesales.txt provides information on the selling price (PRICE), lot cost (COST), type of home (HOME) (SF=single family home or T=condominium style) and region of the country (REGION) (M=Midwest, S=south) for closings during one month.

1. Define the dummy variables for region and home (write them down here), and create them in SAS.

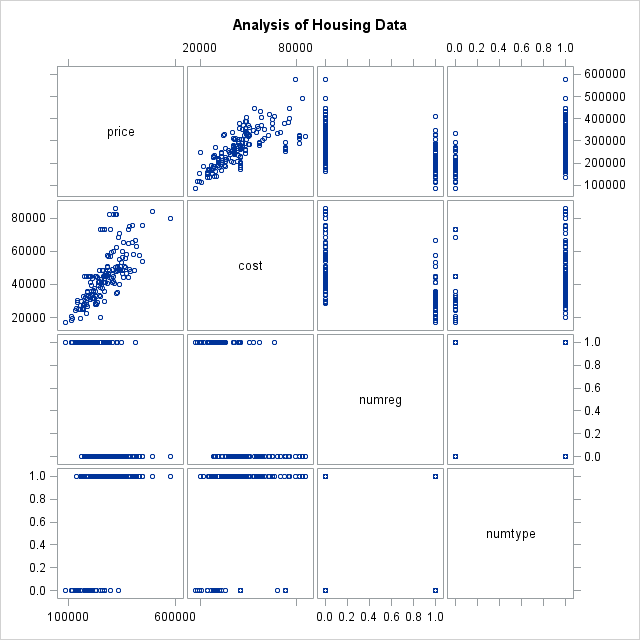
|  |  |
| --- | --- |
| Home | Z1 |
| SF | 0 |
| T | 1 |

|  |  |
| --- | --- |
| Region | Z2 |
| Midwest | 0 |
| South | 1 |

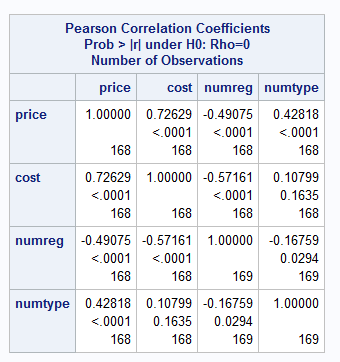
Numhome = (SF = 1)

NumReg = (M = 1)

1. Analyze the association between selling price and each individual attribute (cost, home and region) using appropriate statistics and graphs. Discuss your findings. Include the relevant output.

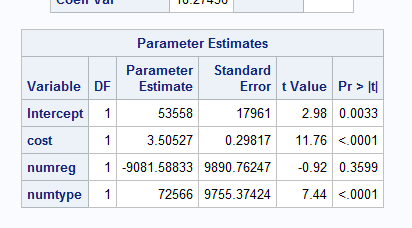


There appears to be a fairly linear relationship between price and cost. The plots for price vs numreg and numtype cannot be used to determine association.

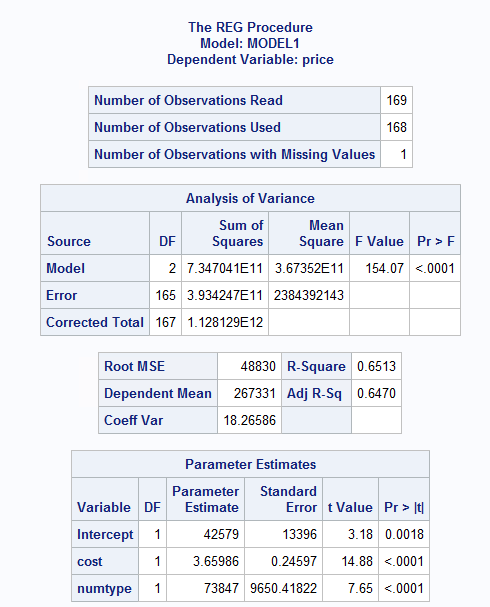


1. Fit an adequate regression model for sales price as a function of lot cost, region of country, and type of home. Remove the terms that are not significant. The final model should only contain variables that are significantly associated with sale price. Write down the model equation. Include the relevant output.

In the first model, we see that numreg is insignificant due to a p-value =.3599



I re-ran the model to exclude numreg.



The equation:

Price = 42579 + 3.66cost + 73847numtype

Where numtype = 1 when type = ‘SF’

1. Conduct a global F-test for overall model adequacy. Write down the test hypotheses and test statistic and discuss conclusions. Include the relevant output.

Null Hypothesis: Bj = 0

Alt Hypothesis: Bj <> 0

F-value = 154.07, and p-value <0.001.

This means that there is significant evidence to reject the null hypothesis, and the model is a good fit.

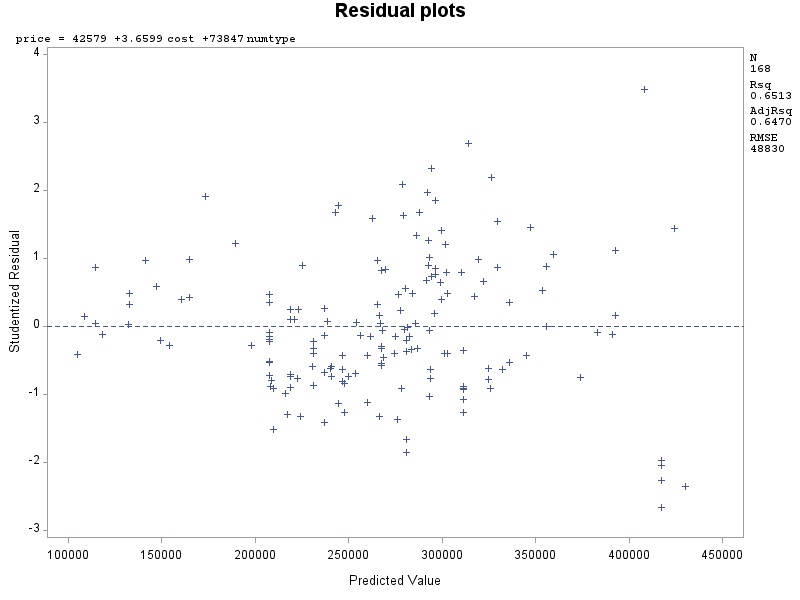
1. Analyze model residuals to check if assumptions on data are satisfied. Discuss your findings. Include the relevant output.

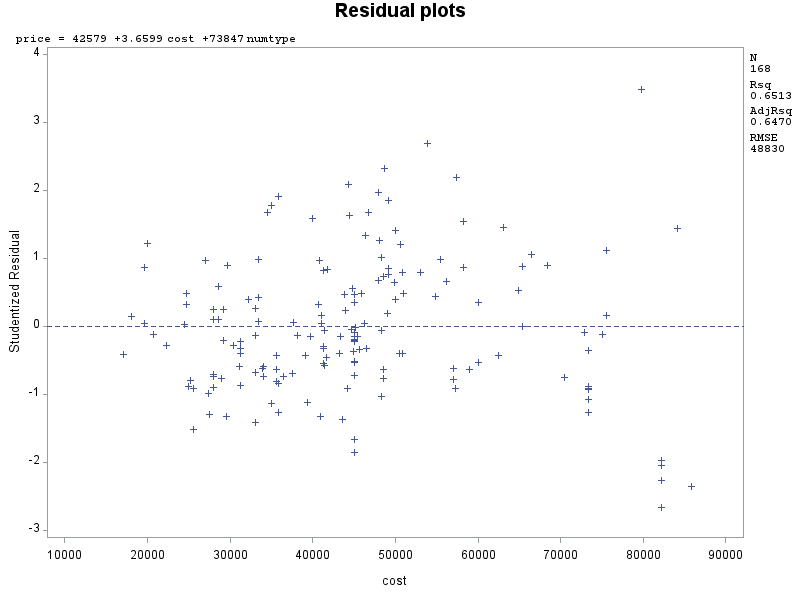
Constant Variance: Not violated.

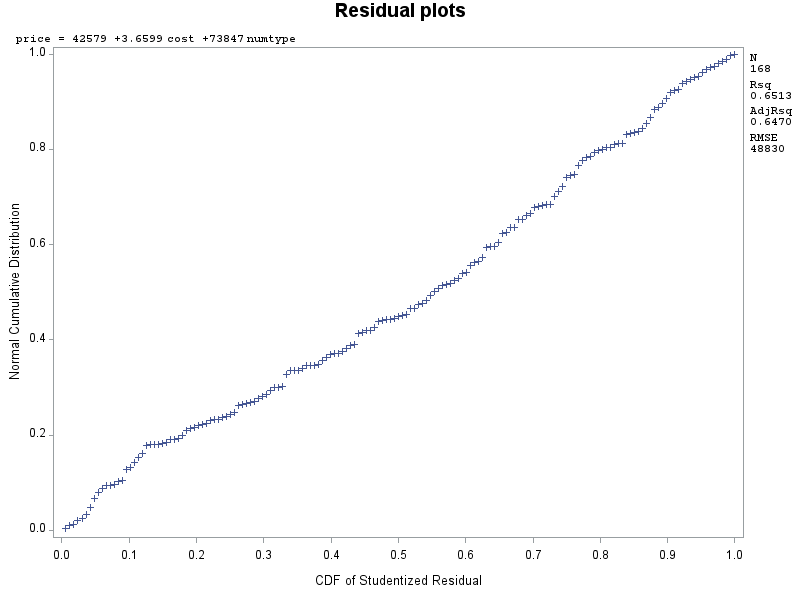
Independence: Does not appear to be violated. The plots don’t look like there is any order to the data

Linearity: violated.

Normal: Not Violated.







1. Discuss what the regression model indicates for the relationship between price and home type (i.e. interpret the coefficient values).

When the numtype = 1, the price of the home increases by $73,847

1. Use the regression analysis to determine whether mean sale prices are different for the two regions? Explain.

We cannot determine this because the numReg variable is insignificant to the sale price. Meaning it has no effect on the sale price.

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

Title "Analysis of Housing Data";

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

delimiter = '09'x;

getnames=YES;

datarow=**2**;

numType = (Type = "SF");

numReg = (Region ="M");

**run**;

**proc** **print**;

**run**;

\*Model 1 scatter plots;

**proc** **sgscatter**;

matrix price cost;

**run**;

\*Correlation coefficients;

**proc** **corr**;

var price cost numreg numtype;

**run**;

\*Model 1;

**proc** **reg**;

model price = cost numreg numtype;

**run**;

\*Model 2;

Title "Model 2";

**proc** **reg**;

model price = cost numtype;

**run**;

\*residual plots;

Title "Residual plots";

plot student.\*predicted.;

plot student.\*(cost);

plot npp.\*student.;

**run**;