CSC 423

ASSIGNMENT – 3

BY

Sasidhar Mukthinuthalapati

1. Problem 1
   1. Dependent Variable: Salary. Independent Variables: College and Sex. For College, we will use three dummy variables. For College using Dummy variables is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| COLLEGE | C1 | C2 | C3 |
| BUSINESS | 0 | 0 | 0 |
| ENGINEERING | 0 | 0 | 1 |
| LIBERAL ARTS | 0 | 1 | 0 |
| NURSING | 1 | 0 | 0 |

For Sex using Dummy variables is as follows:

|  |  |
| --- | --- |
| SEX | S1 |
| MALE | 0 |
| FEMALE | 1 |

* 1. The general regression model relating to salary and both college and sex is as follows:

**Salary = Beta0\*C1 + Beta1\*C2 + Beta3\*C3 + Beta4\*S1 + Beta5(Intercept) +e(error)**

Where C1, C2, C3 AND S1 are dummy variables for College type and Sex Type.

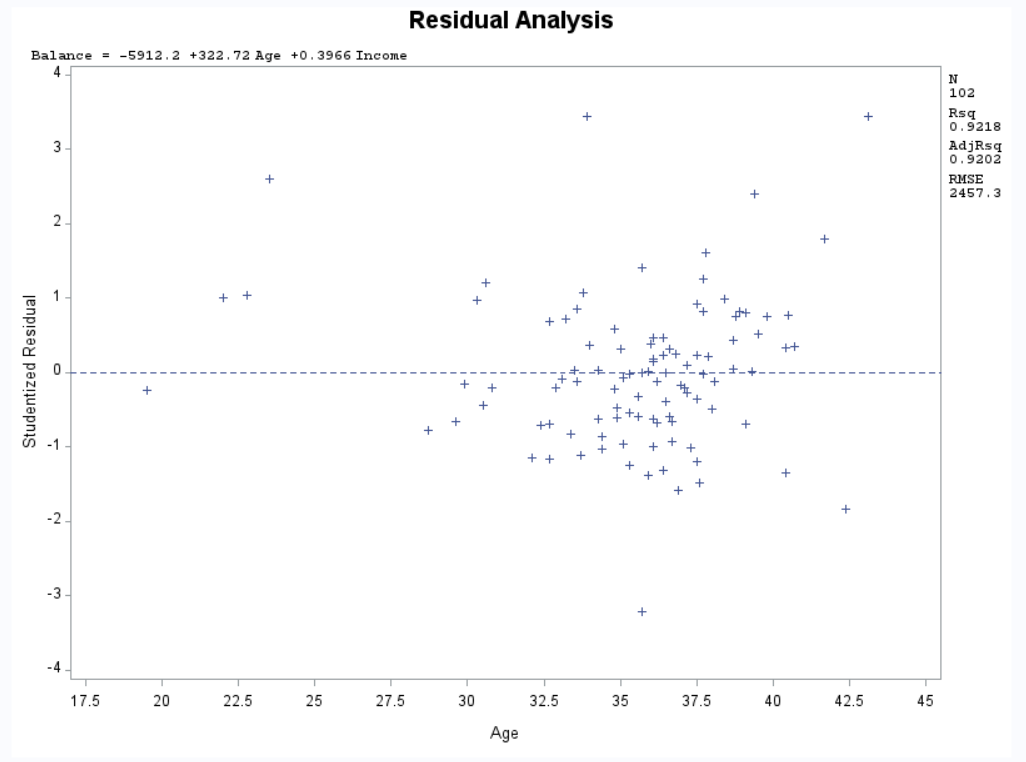
* 1. Then instead of using 3 Dummy variables we will just be using 2 Dummy variables because engineering and business will be treated as the same.

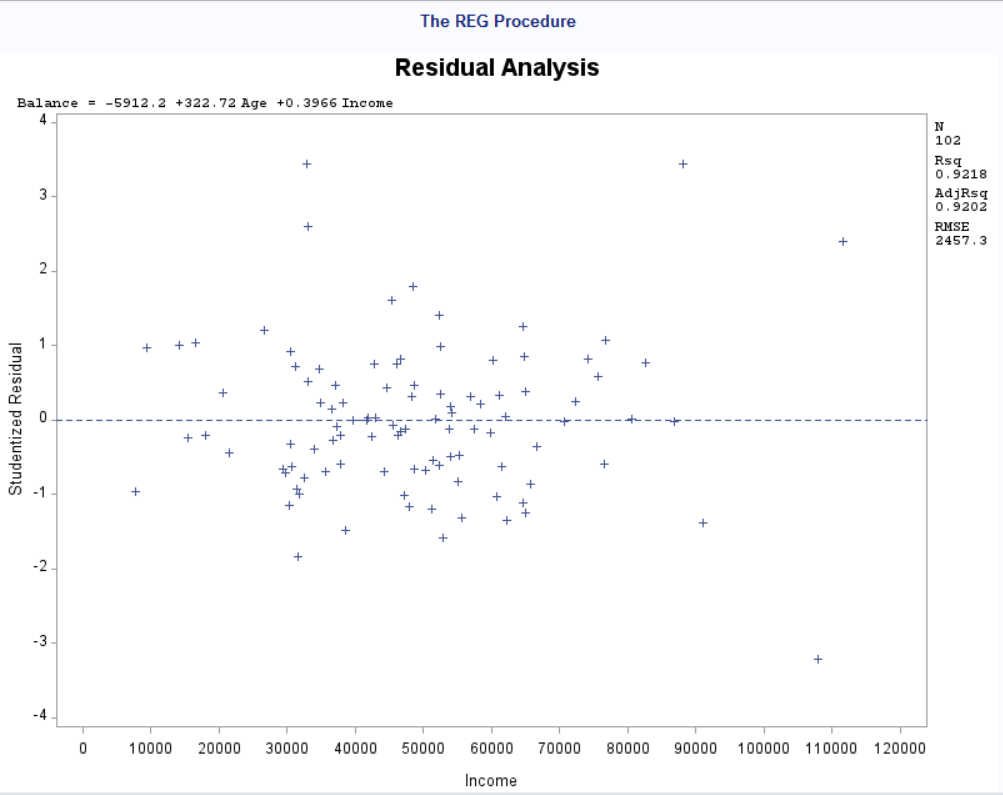
|  |  |  |
| --- | --- | --- |
| COLLEGE | C1 | C2 |
| BUSINESS/ENGINEERING | 0 | 0 |
| LIBERAL ARTS | 0 | 1 |
| NURSING | 1 | 0 |

For Sex there wouldn’t be any change. The new genereal regression equation will be as follows:

**Salary = Beta0\*C1 + Beta1\*C2 + Beta3\*S1 + Beta4(Intercept) + e(Error)**

Where C1, C2 AND S1 are dummy variables for College and Sex Type.

1. Problem 2
   1. The residual plot for the Regressors (age and Income) are the following: 



The code to generate the above mentioned plot is the following:

title "Residual Analysis";

**proc** **reg**;

model balance=age income / stb;

plot student.\*( age income);

**run**;

The findings from the above mentioned residual plot is as follows:

* The age’s residual plot shows the presence of outliers and the shape seems to funnel out which means there is less independence and the variance is also not constant. This residual plot also shows that the spread is more concentrated towards the right-hand side.
* The income’s residual plot also shows he presence of outliers. The residual plot also tells us that income has constant variance and the points are independent of each other i.e. the subsequent points are not associated with each other. There is more randomness in income when compared to age. There are quite few outliers (3) present in Income.

The Code to generate the Residual plot for the predicted values of the model is as follows:

title "Residual Analysis";

**proc** **reg**;

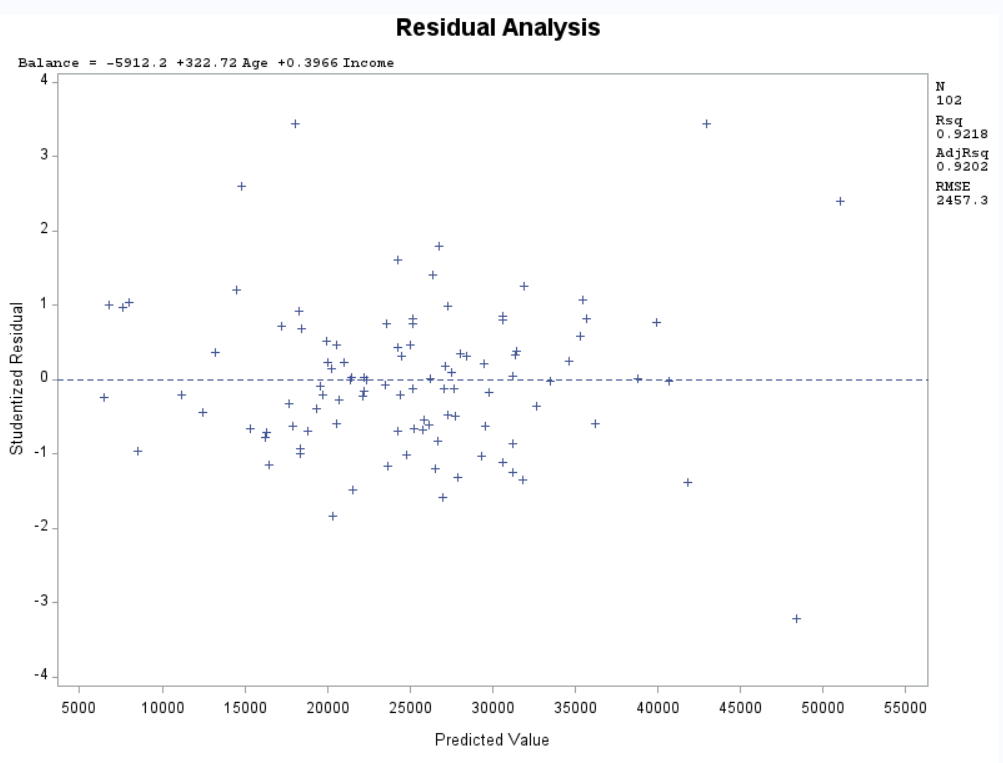
model balance=age income / stb;

plot student.\*predicted.;

plot npp.\*student.;

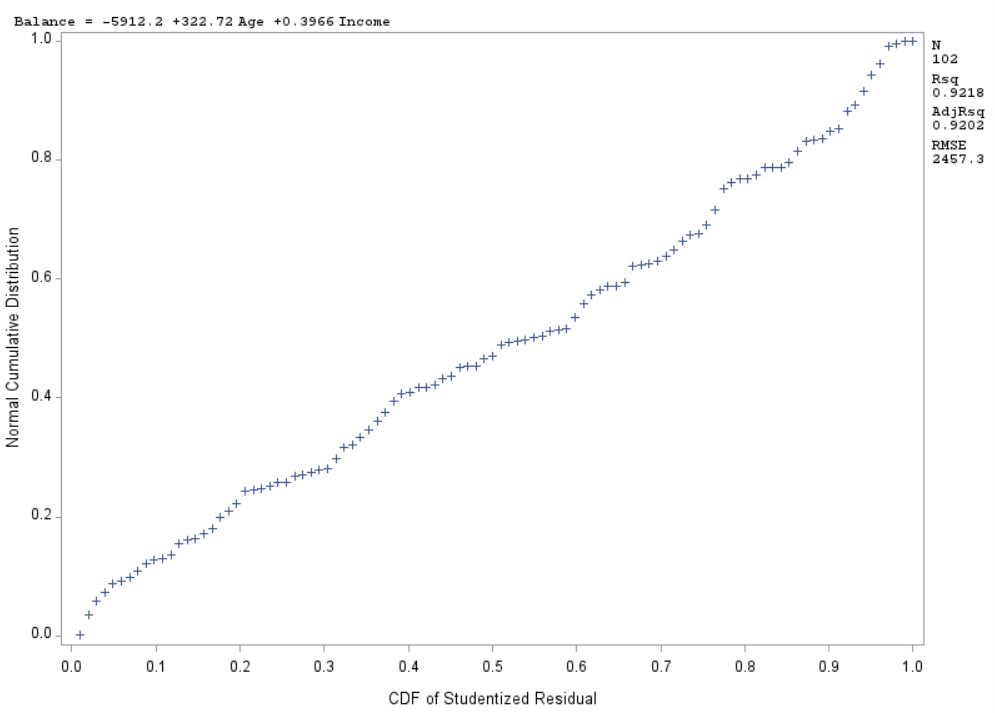
**run**;

The plot which is generated is as follows:

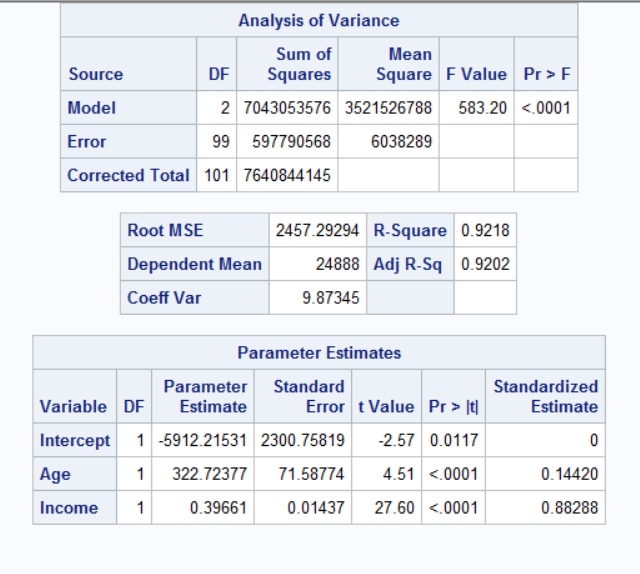


The Residual plot of the predicted values shows constant variance and the subsequent points are independent of each other. But there are few outliers (3) present in this dataset.

The normality plot is also Linear which means that the error is normally distributed. The slight curve at the ending in the **Top Right** **also tells us that outliers are present.** The normality plot is as follows:



* 1. The F-test:



The code mentioned in the previous bit generates the above-mentioned table under the analysis of variance.

In this the **F-value** for the current model is also present which is **583.20**. the **p-value** associated with this F-test score is less than **<.0001** which means that the model is good.

The test hypothesis is as follows:

**H0**: all the Beta coefficients are zero including the intercept.

**Ha**: all the beta coefficients are not equal to zero is the alternate hypothesis.

**F-value: 583.20**

**P-value: <.0001**

We can conclude that the null hypothesis is rejected and that there is at least one variance which explains the variance in y. the p-value also tells us that the model is pretty good and based on the R-squared value we can say that the current model explains **92.025 of variance** in Balance based on the predictors we chose.

* 1. The code for the above-mentioned plots and tables is as follows:

title "Residual Analysis";

**proc** **reg**;

model balance=age income / stb;

plot student.\*( age income);

plot student.\*predicted.;

plot npp.\*student.;

**run**;

1. Problem-3
   1. We need two dummy variables one for home and one for region. To create them in SAS we run the following code:

title "housesales using infile method";

**data** housesale\_infile;

infile "HouseSales.txt" delimiter = '09'x missover firstobs=**2**;

input Region $ Type $ Price Cost;

regionNew = **1**;

if Region = 'M' then regionNew = **0**;

typeNew = **0**;

if type = 'SF' then typeNew = **1**;

**proc** **print**;

**run**;

From the above code the regionNew variable is a dummy variable for Region. typeNew is a dummy variable for Type.

* 1. For performing the analysis between the dependent variable (Selling Price) and the Independent Variable (RegionNew, TypeNew and Cost) we run the code for Correlation analysis between the variables.

The code the generate the scatter plot and the Pearson’s Coefficient values are as follows:

title "Scatter plot Matrix";

**proc** **sgscatter**;

matrix Price Regionnew Typenew Cost;

**run**;

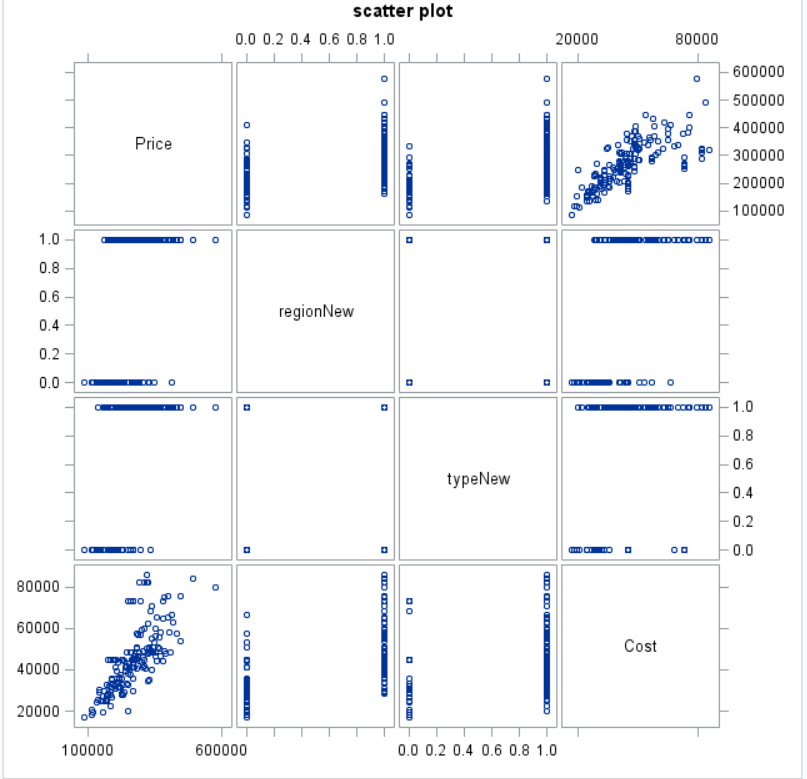
title "Correlation analysis among the variables";

**proc** **corr**;

var Price Regionnew Typenew Cost;

**run**;

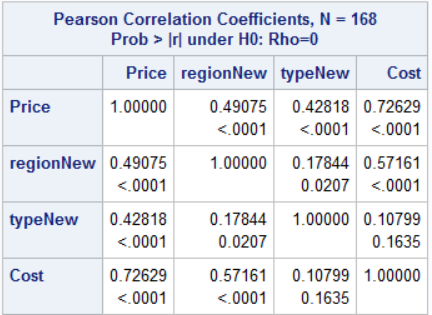
The scatter plot matrix which is generated using the above code is as follows:



Based on the above scatter plots we can make the following observations:

* The plots for Price vs (TypeNew and RegionNew) are a bit distinct because these two variables can only take the possible of two values 0 or 1. It’s hard to provide any conclusion using these plots and hence we would refer to the Pearson’s Coefficient values.
* The scatter plot between Cost and Price shows somewhat strong positive correlation. To determine the magnitude of strength of the correlation we need to refer to the Pearson’s coefficients.

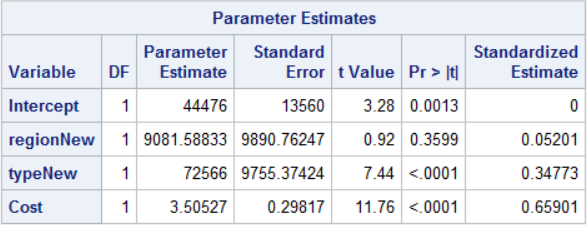
To further support the above points, we will perform correlation analysis among the variables. The Pearson’s coefficient values for the variables are given below:



Based on the Pearson Coefficient values we can infer the following things:

* Cost and Price are positively correlated and using Cost alone we can explain up to 72.62% of variance in Price of the house. The correlation can be said to be the strongest among the available predictors.
* Price is positively correlated with RegionNew but the correlation isn’t very strong. 49.07% of the variance in Price can be explained using the RegoinNew as the predictor.
* Price and TypeNew are strongly correlated but not as much as Cost. Using only TypeNew we can explain 42.81% of the variance in the Price of the house.
  1. To come up with an adequate regression model firstly we will make a model using all the available independent variables and based on the parameter estimates values we will remove the variables to come up with the final model.

First Model is as follows:



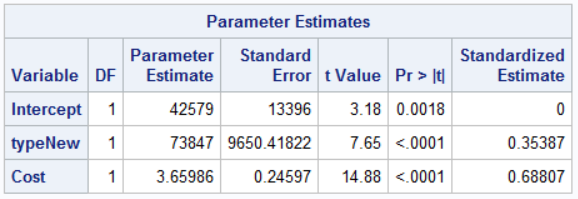


**Price = 44476+ 9081.588\*regionNew + 72566\*typeNew + 3.505\*Cost**

But in this model, the p-value for **regionNew is .3599 which is greater than .05** and hence in the new model we will **remove this variable** as it is not significant.

Based on the Adjusted R-Squared value we can say that **64.67% of the variance in Price** could be explained using the above-mentioned equation.

The new model Parameter Estimate values are as follows:

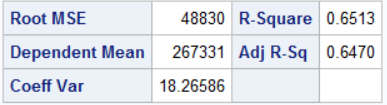


From these values of Parameter Estimate we come up with a new linear regression equation for predicting Price which is as follows:

**Price = 42579 + 73847\*typeNew + 3.659\*Cost**

Here, **typeNew is a dummy variable** which we had created **typenew = 0 for Condominium** and **typenew = 1 for Single Family Home**.

To evaluate the performance of this model we will refer to the Adjusted R-Squared values.



Based on the Adjusted R-Squared values we can say that **64.70% of the variance in Price** can be explained using this model in which we have not taken Region into consideration.

The code to generate the two models is as follows:

title "Regression model for Sales Price with all variables";

**proc** **reg**;

model price = regionNew typenew cost / stb;

**run**;

title "Regression model for Sales Price excluding region";

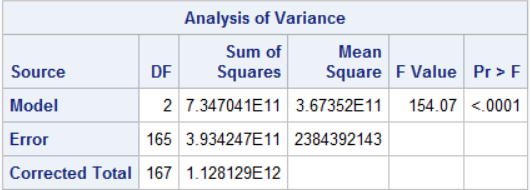
**proc** **reg**;

model price = typenew cost / stb;

**run**;

* 1. Overall Goodness of the model can be determined using the F-test and the p-value associated with it. When we run the code for the above model, SAS also generates a table called Analysis of Variance in which F-Test value is present and the p-value associated with the F-test score.

The table is as follows:



Here the **F-Test value is 154.07** and the p-value associated with it is **<.0001** which means that the model is a very good model.

Here, the Null Hypothesis is **H0**: which means that all the Beta coefficients will be zero.

Alternate Hypothesis **Ha:** The Beta coefficients are not equal to zero.

**F-Value: 154.07**

**P-Value: <.0001**

Based on the above information we can conclude that the model is a **decent fit for this problem**. We also **reject the Null Hypothesis** because the **p-value is <.0001.**

Based on the Test Statistics values we can say that the Beta values are not equal to zero. The beta values for this model are mentioned in the previous bits.

* 1. The studentized residual plot with respect to each variable and the predicted values and the normality curve for this problem can be generated using the following code:

title "Residual Analysis vs xvar";

**proc** **reg**;

model price= typeNew cost;

plot student.\*( typeNew cost);

**run**;

title "Residual Analysis normality";

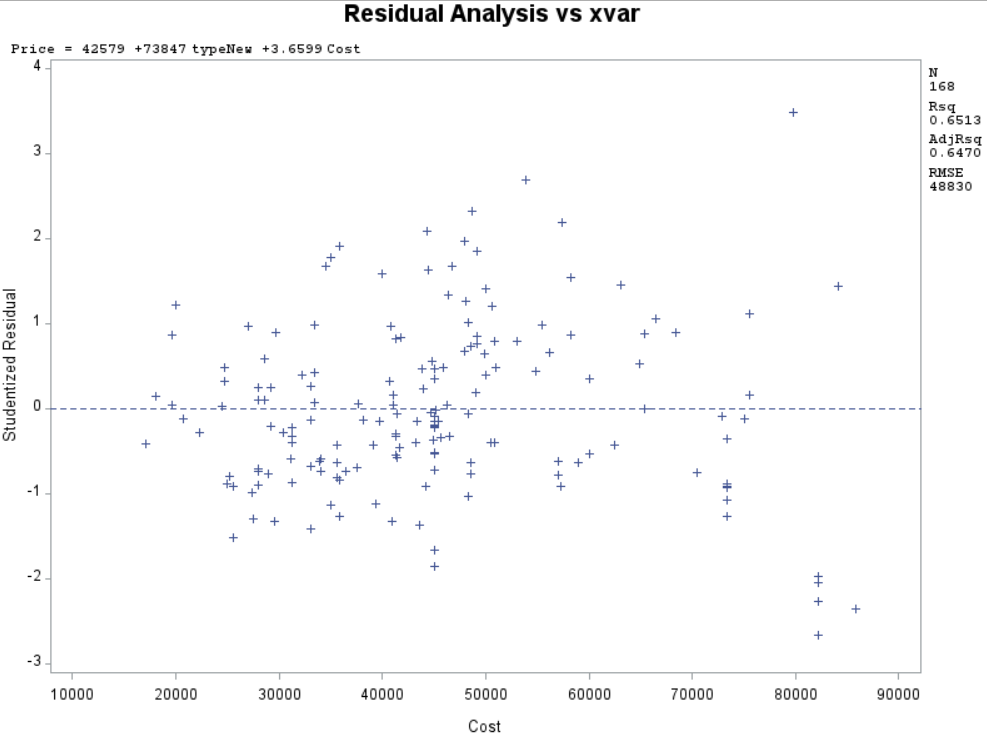
**proc** **reg**;

model price= typeNew cost;

plot npp.\*student.;

**run**;

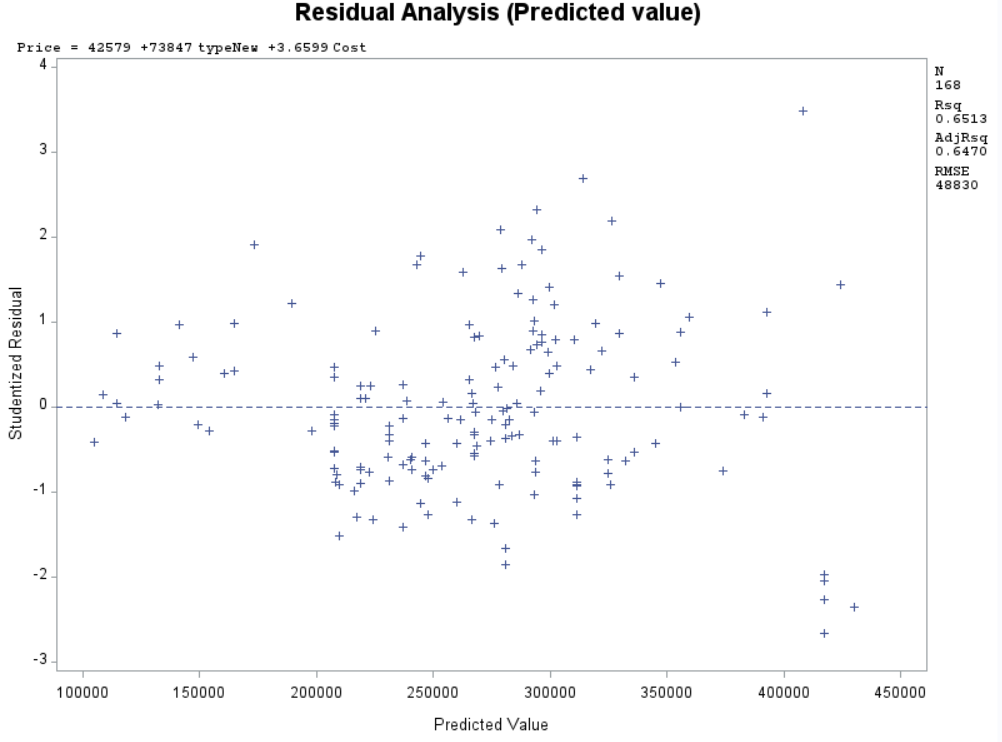
The residual plot generated is as follows:



The following observations can be made based on the residual plot:

* There is **an outlier present** in the dataset. This can be identified as the point which is above +3.
* The plot is seeming to have constant variance till Cost = 40000, but after that the plot seems to funnel out which means that there **isn’t a constant variance**. But we can say that up until cost = 60000 there seems to be somewhat constant variance because there are few points which lie between -2 and -3 region of the residual plot.
* Since there isn’t a proper constant variance we can say that the model could be further improved by introducing various other higher order terms into the dataset.

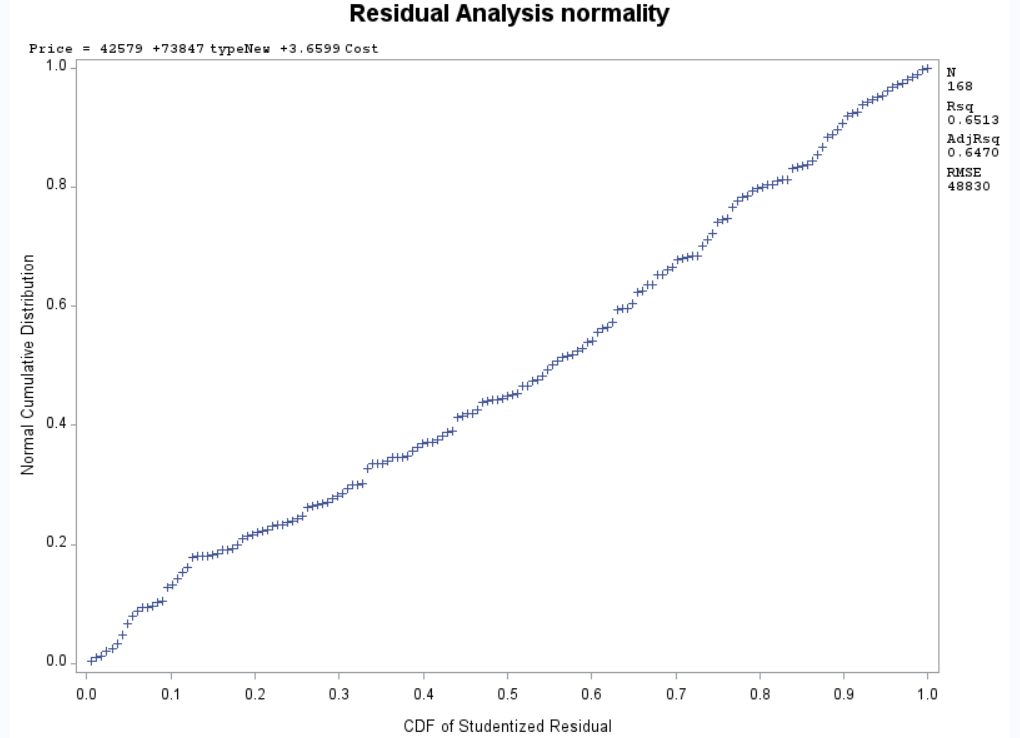
The residual plot for the model wrt to the value it predicted is as follows:



From the above plot we can make the following observations:

* There is **one outlier present** in this dataset based on this model.
* The residual values seem to funnel outwards. We can infer that there isn’t **constant variance** present in it and that the subsequent points are **not totally independent** of each other.
* The Adjusted R-Squared value is .6470 which suggests that we may need to include other higher order terms into the data set to come up with a better model. But with the data which has been provided we can say that this is a good model because we can explain 64.7% of variance in the price using this dataset.

The normality plot generated from the above-mentioned code is as follows:



The following observations can be made for the above plot:

* The plot is quite normal and seems normal.
* Based on the plot we can also say that the error term is normally distributed.
  1. The linear equation for the association between price and the other variables is as follows:

**Price = 42579 + 73847\*typeNew + 3.659\*Cost**

The change in house type from 0 to 1 i.e. from Condominium style to Single Family Home would result in the **increase** in Price by **$73847.**

If the Cost of the house is **increase by 1$** then the total Price of the house will **increase by $3.65.**

* 1. To determine whether the mean sale price are different for the two regions we would have to create two separate boxplots i.e. Price wrt the regions.

This could be done using the following code:

title "Boxplot based on region";

**proc** **sort**;

by region;

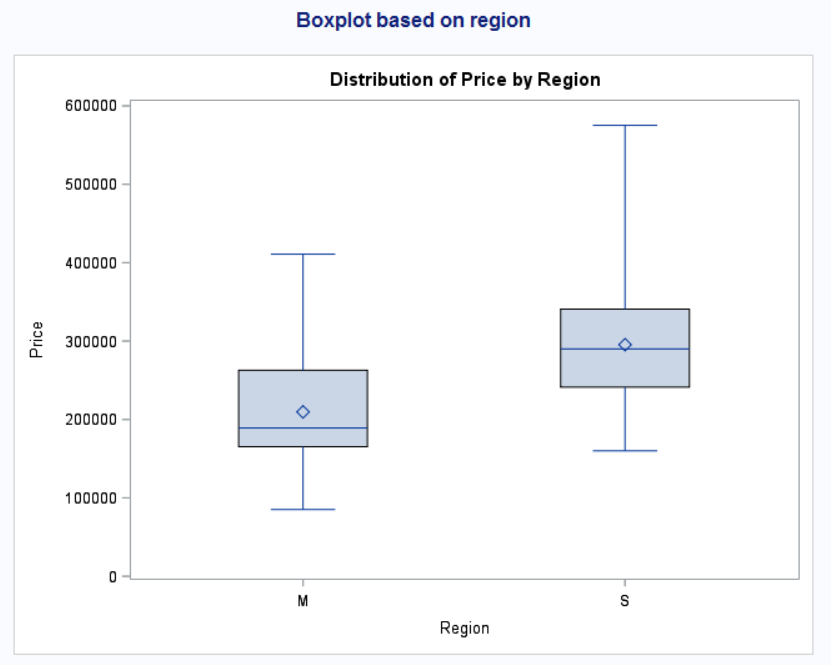
**run**;

**proc** **boxplot**;

plot price\*region;

**run**;

The above code generates the following boxplots:



As it can be seen clearly that the two regions have different set of values.

The following observations could be made from the boxplots without using the model which we created because in the model Region is not considered to be a significant variable/attribute/predictor:

* The mean, median, mode, min and max are all different for both the regions.
* The IQR is also different.

Overall both the regions have all different values and for better understanding its better to run univariate analysis on Price based on the Region.

The code is as follows:

title "Univariate Analysis of Price by Region";

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

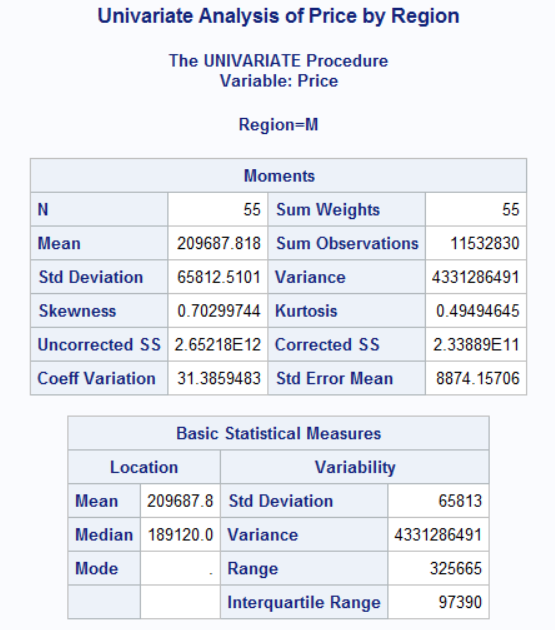
by region;

var Price;

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

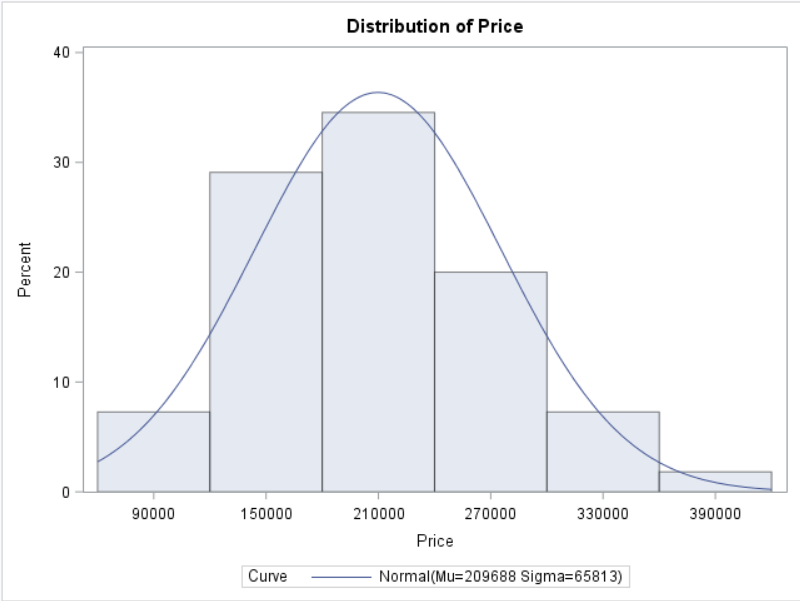
**run**;

**For Region “M” we get the following outputs:**

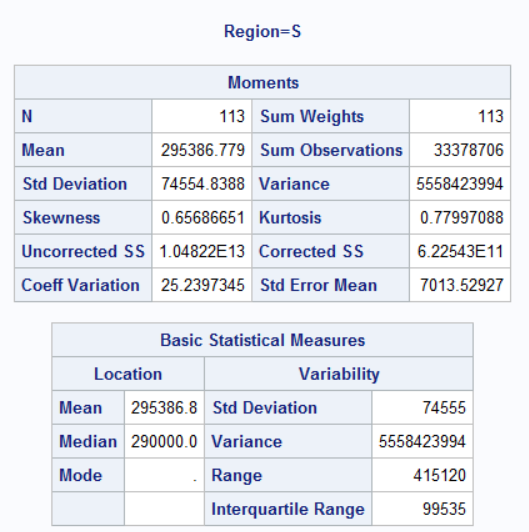


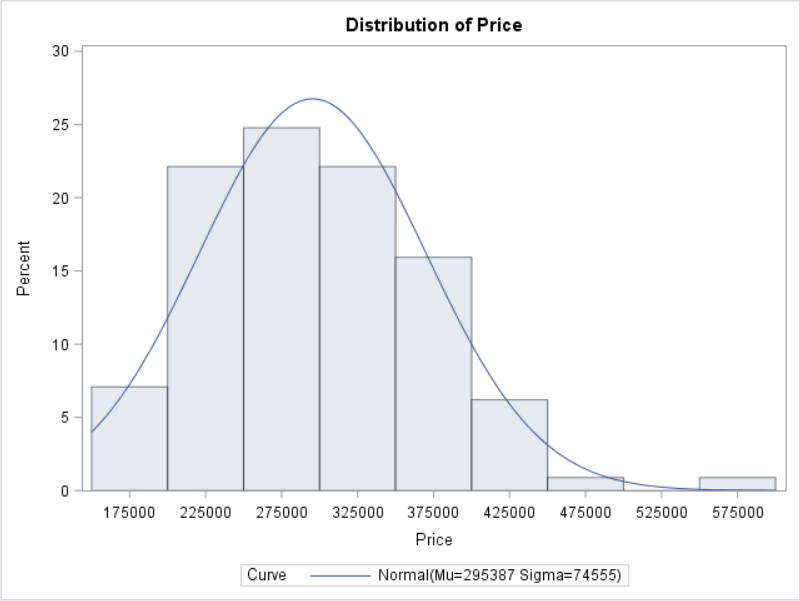
Mean = 209687.81; Meadian = 189120.0 IQR = 97390.

The Mean Price for **Region “M” is 209687.81.** The data is **Positively Skewed.**



**For Region “S” we get the following outputs:**





Here the **Mean Price for the South Region is 295386.8.** The Data again is Positively Skewed.

When we compare the Mean values of both the regions it is evident that **South Region (295386.8) has greater Mean Price than the Midwest Region (209687.81)**.

Both the regions have outliers present in them and **both are positively skewed**.

Based on the two Regions we can say that the South Region has higher Prices for Houses when compared to the Midwest Region this can also be inferred from the boxplots.

* 1. The whole code which was used to generate the above analysis is as follows:

title "housesales using infile method";

**data** housesale\_infile;

infile "HouseSales.txt" delimiter = '09'x missover firstobs=**2**;

input Region $ Type $ Price Cost;

regionNew = **1**;

if Region = 'M' then regionNew = **0**;

typeNew = **0**;

if type = 'SF' then typeNew = **1**;

**proc** **print**;

**run**;

title "scatter plot";

**proc** **sgscatter**;

matrix Price Regionnew Typenew Cost;

**run**;

title "Correlation analysis among the variables";

**proc** **corr**;

var Price Regionnew Typenew Cost;

**run**;

title "Regression model for Sales Price";

**proc** **reg**;

model price = typenew cost / stb;

**run**;

title "Boxplot based on region";

**proc** **sort**;

by region;

**run**;

**proc** **boxplot**;

plot price\*region;

**run**;

title "Univariate Analysis of Price by Region";

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

by region;

var Price;

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

**run**;

title "Residual Analysis vs xvar";

**proc** **reg**;

model price= typeNew cost;

plot student.\*( typeNew cost);

**run**;

title "Residual Analysis (Predicted value)";

**proc** **reg**;

model price= typeNew cost;

plot student.\*predicted.;

**run**;

title "Residual Analysis normality";

**proc** **reg**;

model price= typeNew cost;

plot npp.\*student.;

**run**;