**Ramadan Gannud**

**Assignment-3**

**Problem 1**

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**

   Number of levels = 4

   Number of dummy variables = 3

   Dcoll1 = 1 if COLLEGE = Engineering, or otherwise Dcoll1 = 0

   Dcoll2 = 1 if COLLEGE = Liberal Arts, or otherwise Dcoll2 = 0

   Dcoll3 = 1 if COLLEGE = Nursing, or otherwise Dcoll3 = 0

| | Z1 Dcoll1 | Z2 Dcoll2 | Z3 Dcoll3 |
|---|---|---|---|
| Business | 0 | 0 | 0 |
| Engineering | 1 | 0 | 0 |
| Liberal Arts | 0 | 1 | 0 |
| Nursing | 0 | 0 | 1 |

| | Z4 Numsex |
|---|---|
| Male | 1 |
| Female | 0 |

   Numsex = 1 if SEX = male

   Numsex = 0 if SEX = female

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

   Starting salary = $\beta_0 + \beta_1$ Dcoll1 + $\beta_2$ Dcoll2 + $\beta_3$ Dcoll3 + $\beta_4$ Numsex + e

   Starting salary = $\beta_0 + \beta_1 Z1 + \beta_2 Z2 + \beta_3 Z3 + \beta_4 Z4 + e$

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

   Starting salary (Engineering) = Starting salary (Business)

   $\beta_0 + \beta_1 (1) + \beta_2 (0) + \beta_3 (0) + \beta_4 Z4 + e = \beta_0 + \beta_1 (0) + \beta_2 (0) + \beta_3 (0) + \beta_4 Z4 + e$

   $\beta_0 + \beta_1 + \beta_4 Z4 + e = \beta_0 + \beta_4 Z4 + e$

   $\beta_1 = 0$

   Starting salary = $\beta_0 + \beta_2 Z2 + \beta_3 Z3 + \beta_4 Z4 + e$

**Problem 2:**

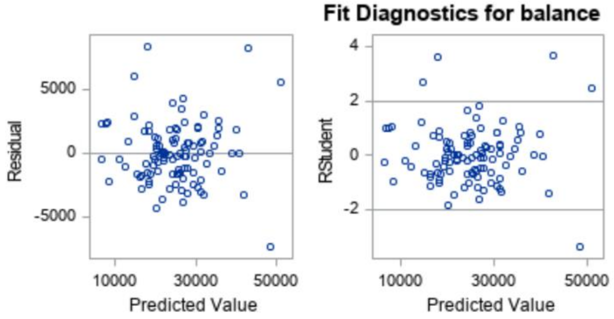
You will continue the analysis of the banking.txt dataset that was analyzed in Assignment 2 – data file is attached. Answer this question based on your final model from assignment-2.

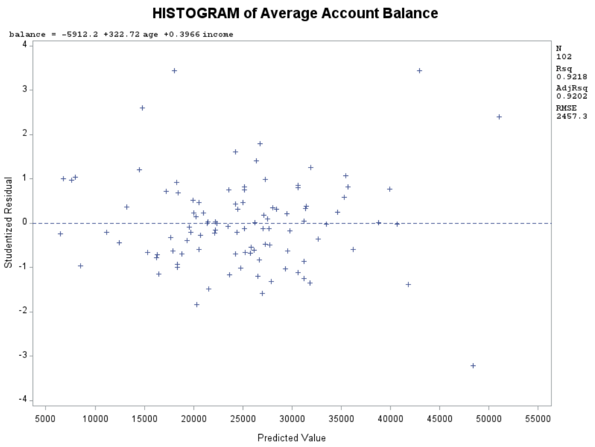
a) **Analyze the residuals of the regression model you found in your previous assignment. Include the residual plots. Discuss your findings.**

Balance = -5912.215 +322.724*Age + 0.397 * Income + e

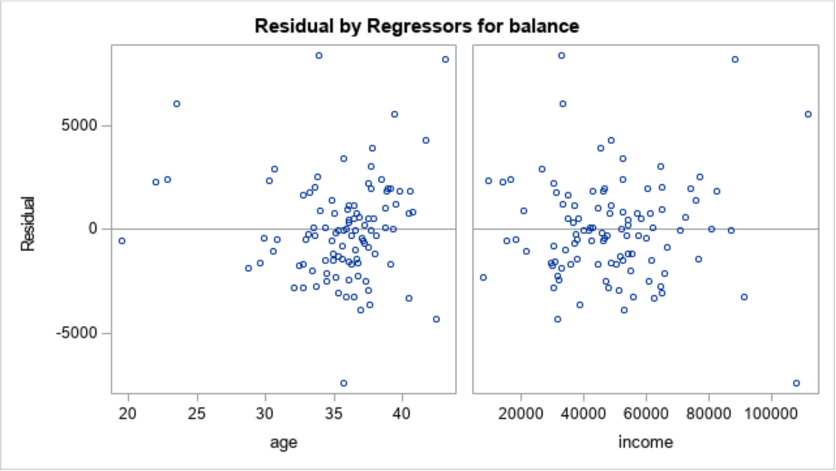
**Assumptions of Constant Variance and Independence**:

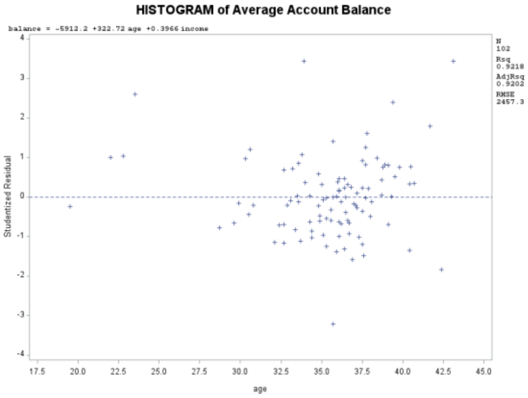
1. Plot residuals vs predicted values.

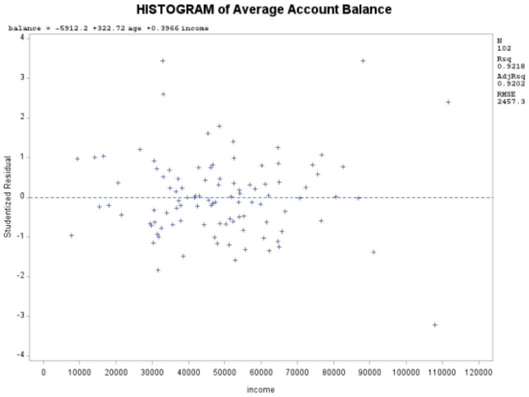




2. Plot residuals vs each x-variable.





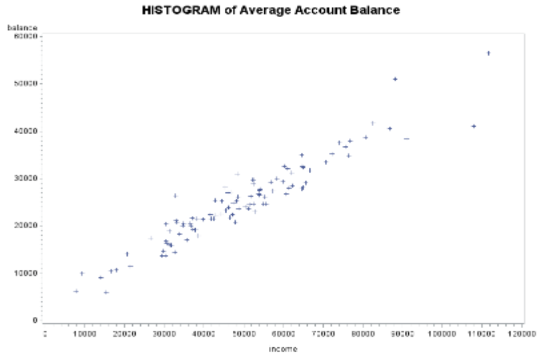


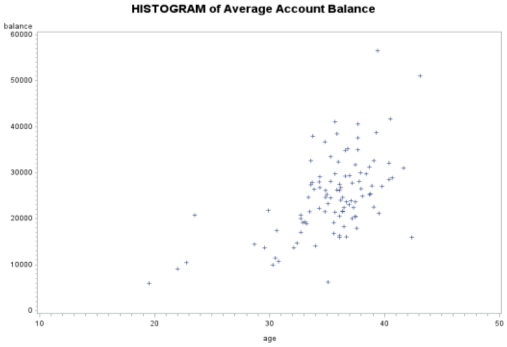
Points are randomly scattered, and residual analysis show no concern for the model fit.

**Assumptions for linearity:**

1. Scatterplot for each x-variable.

   Only the income scatterplot shows some linearity and the association appear to be linear as it's shown in the next graphs.



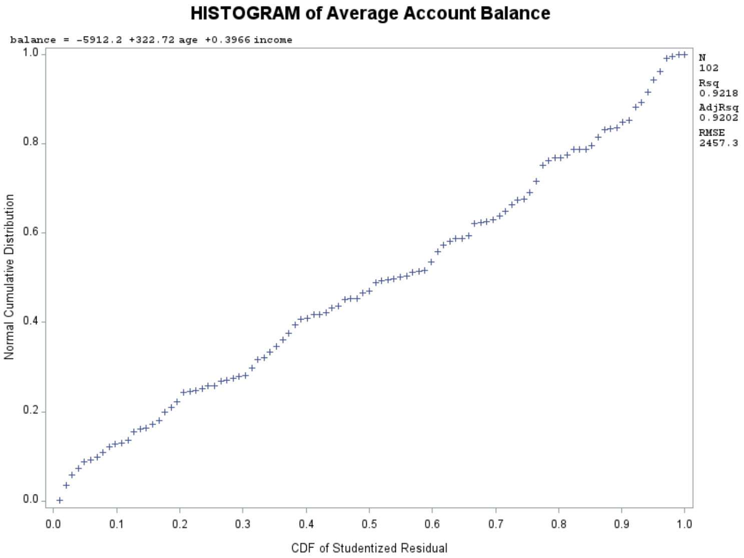


2. Plot residuals vs each x-variable.

   None of them show almost a straight line as it was shown in the residuals vs each x-variable graphs in the previous page. Therefore, there is no linearity.

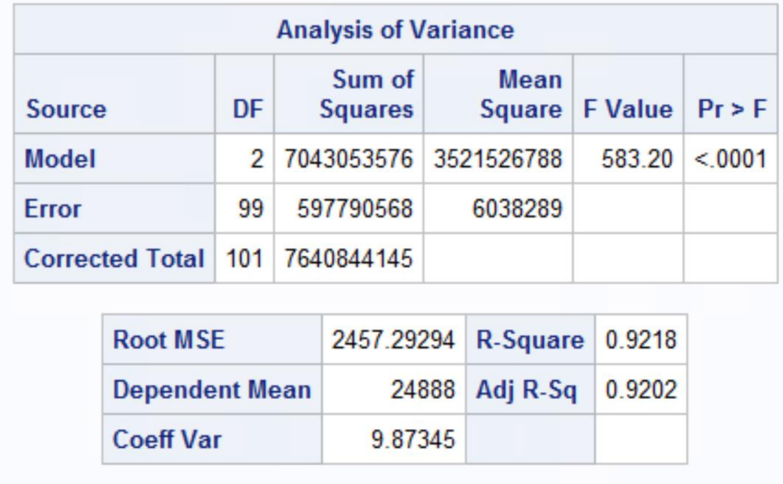
**Assumptions for Normality:**

Plot the normal probability plot of the residuals.



Almost a straight which means it's normal.

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



| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 7043053576 | 3521526788 | 583.20 | <.0001 |
| Error | 99 | 597790568 | 6038289 | | |
| Corrected Total | 101 | 7640844145 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 2457.29294 | R-Square | 0.9218 |
| Dependent Mean | 24888 | Adj R-Sq | 0.9202 |
| Coeff Var | 9.87345 | | |

**Null hypothesis:**

**Ho:** $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \ldots = 0$

No association between Y and x-variables

**Alternative hypothesis:**

**Ha:** At least one coefficient.

**Test Statistics:** F = MS(Regression)/MS(Residual) = MSR/MSE

MSR = 3521526788 MSE = 6038289

F = 3521526788/6038289 = 583.2

F = 583.2 and with p-value less than 0.05.

The null hypothesis of no association between y and x is rejected and the F-test gives strong support to the fitted model. Linear regression explains variation in Y because SSR >>SSE, thus F statistic is large.

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

```
proc reg;
model balance=age income;
* Residual plot: residuals vs x-variables;
plot student.*(age income);
* Residual plot: residuals vs pred. values;
plot student.*predicted.;

* Normal probability plot or QQ plot;
plot npp.*student.;
run;
```

**Problem 3**

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.

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

| | Z1 numtype |
|---|---|
| SF | 1 |
| T | 0 |

| | Z2 numregion |
|---|---|
| M | 1 |
| S | 0 |

numtype = 1 if HOME = SF

numtype = 0 if HOME = T

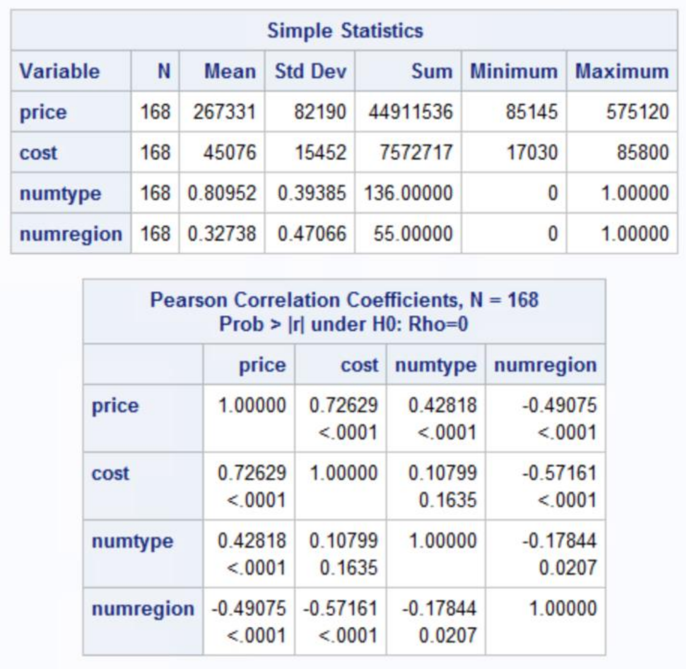
numregion = 1 if REGION = M

numregion = 0 if REGION = S

***code;**

```
data HouseSlaes;
infile "S:\HW3\HouseSales.txt" delimiter = '09'x missover firstobs=2;
input region $ type $ price cost;
numtype = 1;
if type = 'T' then numtype = 0;
numregion = 1;
if region = 'S' then numregion = 0;
run;
proc print;
run;
```

b) **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.**

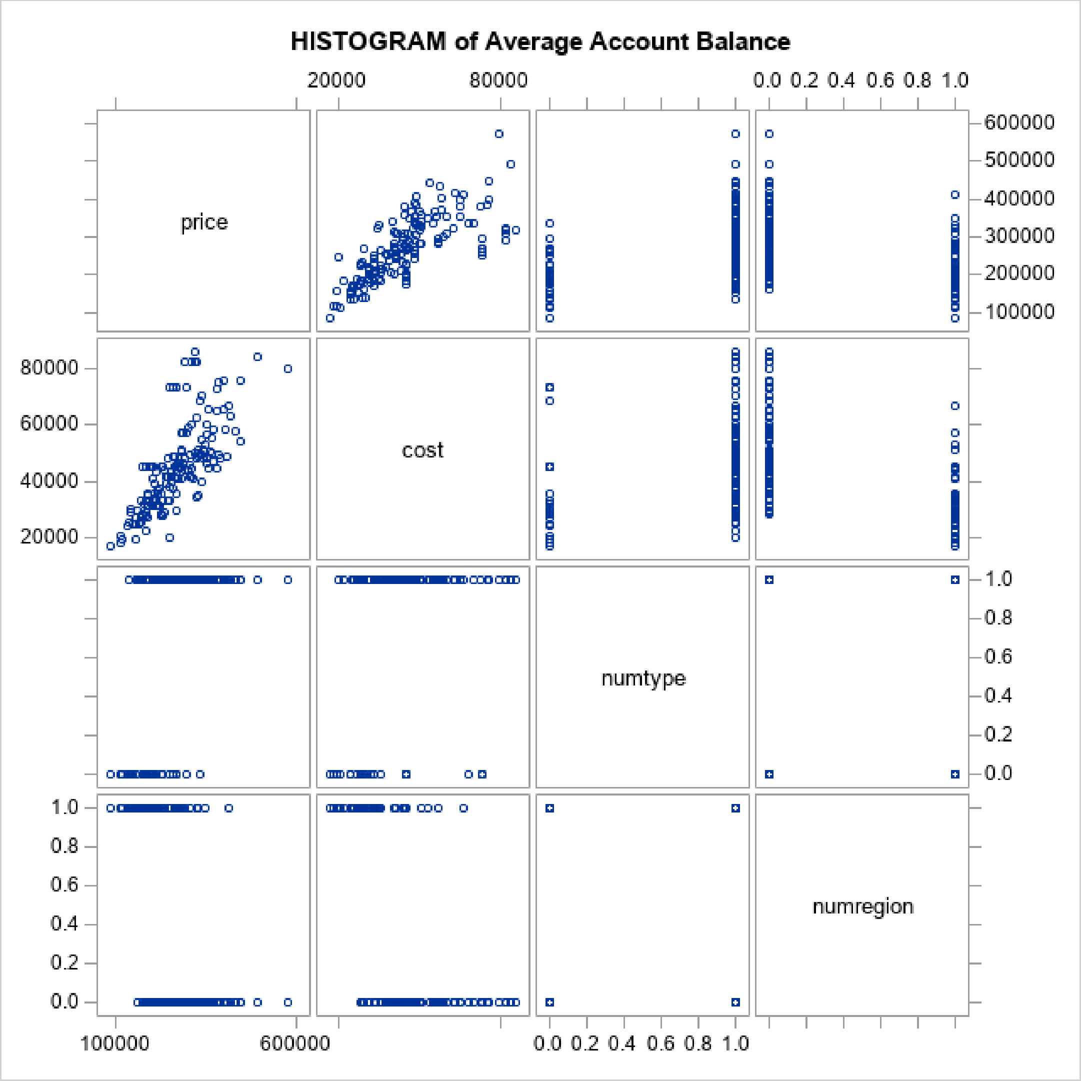


**Simple Statistics**

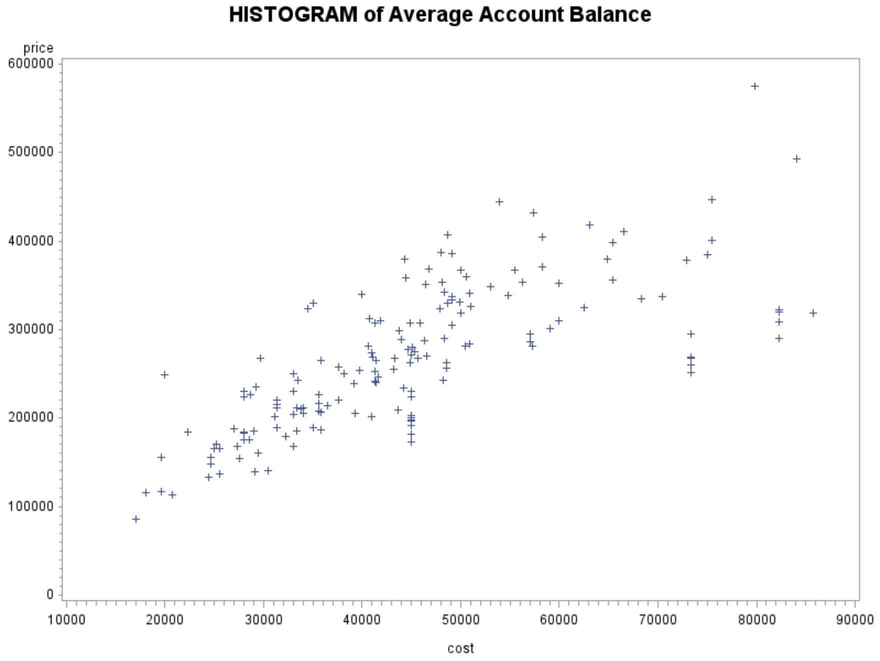
| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| price | 168 | 267331 | 82190 | 44911536 | 85145 | 575120 |
| cost | 168 | 45076 | 15452 | 7572717 | 17030 | 85800 |
| numtype | 168 | 0.80952 | 0.39385 | 136.00000 | 0 | 1.00000 |
| numregion | 168 | 0.32738 | 0.47066 | 55.00000 | 0 | 1.00000 |

**Pearson Correlation Coefficients, N = 168**
**Prob > |r| under H0: Rho=0**

| | price | cost | numtype | numregion |
|---|---|---|---|---|
| price | 1.00000 | 0.72629<br><.0001 | 0.42818<br><.0001 | -0.49075<br><.0001 |
| cost | 0.72629<br><.0001 | 1.00000 | 0.10799<br>0.1635 | -0.57161<br><.0001 |
| numtype | 0.42818<br><.0001 | 0.10799<br>0.1635 | 1.00000 | -0.17844<br>0.0207 |
| numregion | -0.49075<br><.0001 | -0.57161<br><.0001 | -0.17844<br>0.0207 | 1.00000 |

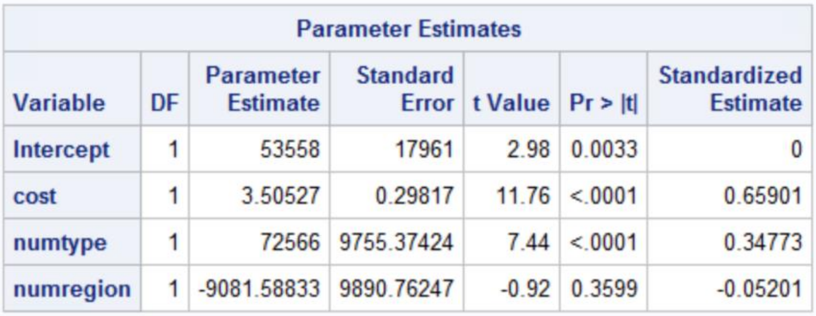


**HISTOGRAM of Average Account Balance**



Since numtype and numregion were qualitative variables and then have been changed to dummy variables, they will not show any linear relationship and dots will be scattered around 0 and 1. Selling price vs cost are both numerical values and have a semi linear relationship as we see in the graph above. The correlation value between price and cost is 0.726 which does not show any significant linear relationship between Y- variable (price) and X-variable (cost).
The association between price and cost is stronger than it’s in home type and region.

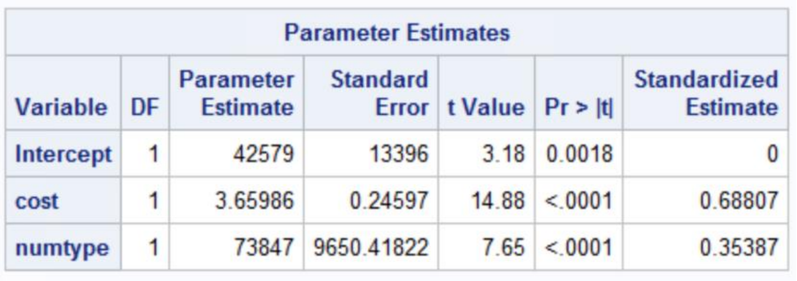
c) **Fit an adequate regression model for sales price as a function of lot cost, region of country, and type of home. Remove the terms that are not significant. The final model should only contain variables that are significantly associated with sale price. Write down the model equation. Include the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



Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
|---|---|---|---|---|---|---|
| Intercept | 1 | 53558 | 17961 | 2.98 | 0.0033 | 0 |
| cost | 1 | 3.50527 | 0.29817 | 11.76 | <.0001 | 0.65901 |
| numtype | 1 | 72566 | 9755.37424 | 7.44 | <.0001 | 0.34773 |
| numregion | 1 | -9081.58833 | 9890.76247 | -0.92 | 0.3599 | -0.05201 |

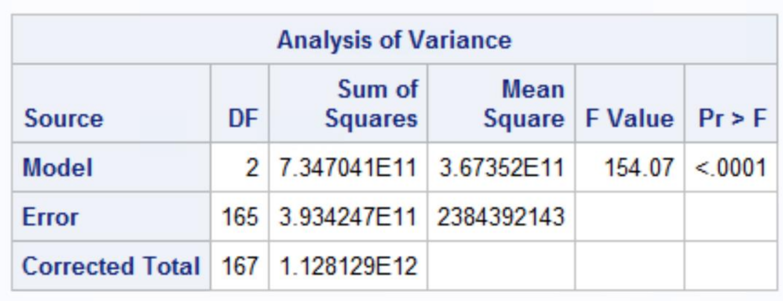
When performing t-test on individual parameters, cost and numtype have p-values that are less than 0.05 which make them significant X variables. The numregion p-value is 0.3599 which makes it insignificant.
After we exclude numregion and rerun the model again, we get the following:



| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
| Intercept | 1 | 42579 | 13396 | 3.18 | 0.0018 | 0 |
| cost | 1 | 3.65986 | 0.24597 | 14.88 | <.0001 | 0.68807 |
| numtype | 1 | 73847 | 9650.41822 | 7.65 | <.0001 | 0.35387 |

**Sale price =** 42579 + 3.659*Cost + 73847*numtype + e

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



| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 7.347041E11 | 3.67352E11 | 154.07 | <.0001 |
| Error | 165 | 3.934247E11 | 2384392143 | | |
| Corrected Total | 167 | 1.128129E12 | | | |

**Null hypothesis:**

**Ho:** $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \ldots = 0$

No association between Y and x-variables

**Alternative hypothesis:**

**Ha:** At least one coefficient.

**Test Statistics:** F = MS(Regression)/MS(Residual) = MSR/MSE

MSR = 3.67352E11 MSE = 2384392143

F = 3.67352E11/2384392143 = 154.065

F = 154.065 and with p-value less than 0.05.

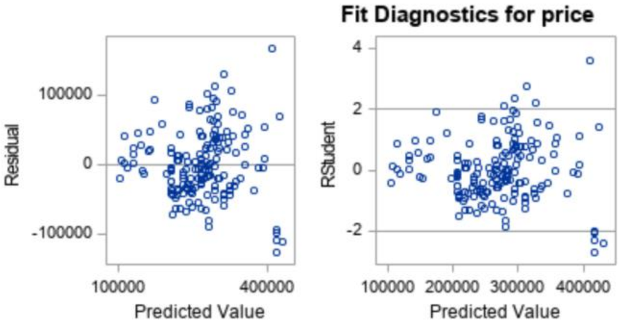
The null hypothesis of no association between y and x is rejected and the F-test gives strong support to the fitted model. Linear regression explains variation in Y because SSR >> SSE, thus F statistic is large. There are two x-variables that have significant effect on Y.

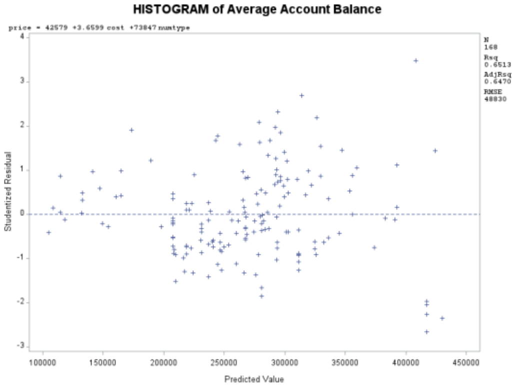
e) **Analyze model residuals to check if assumptions on data are satisfied. Discuss your findings. Include the relevant output.**

**Sale price =** 42579 + 3.659*Cost + 73847*numtype + e

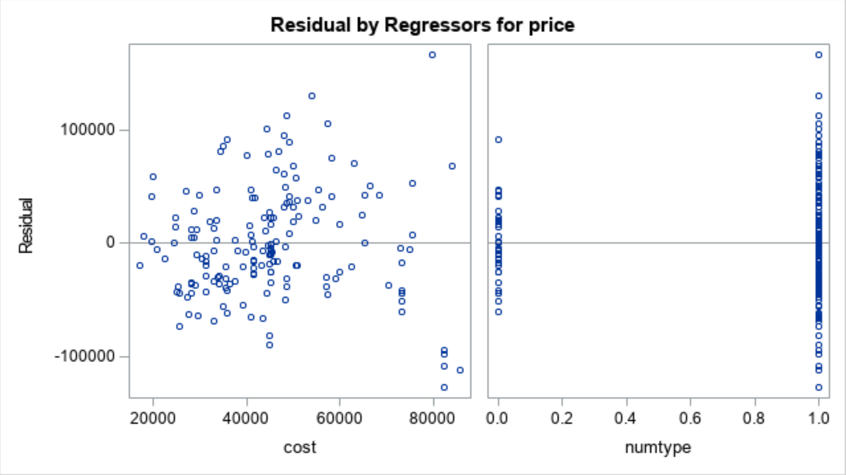
**Assumptions of Constant Variance and Independence**:

1. Plot residuals vs predicted values.





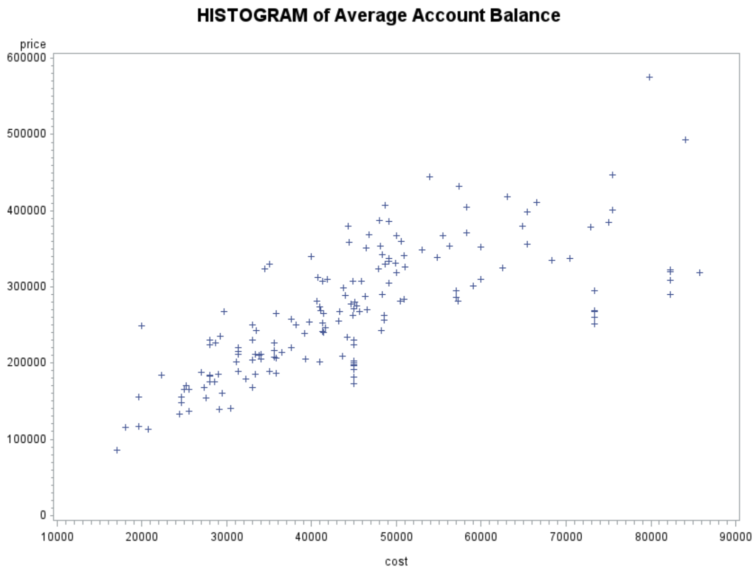
2. Plot residuals vs each x-variable.

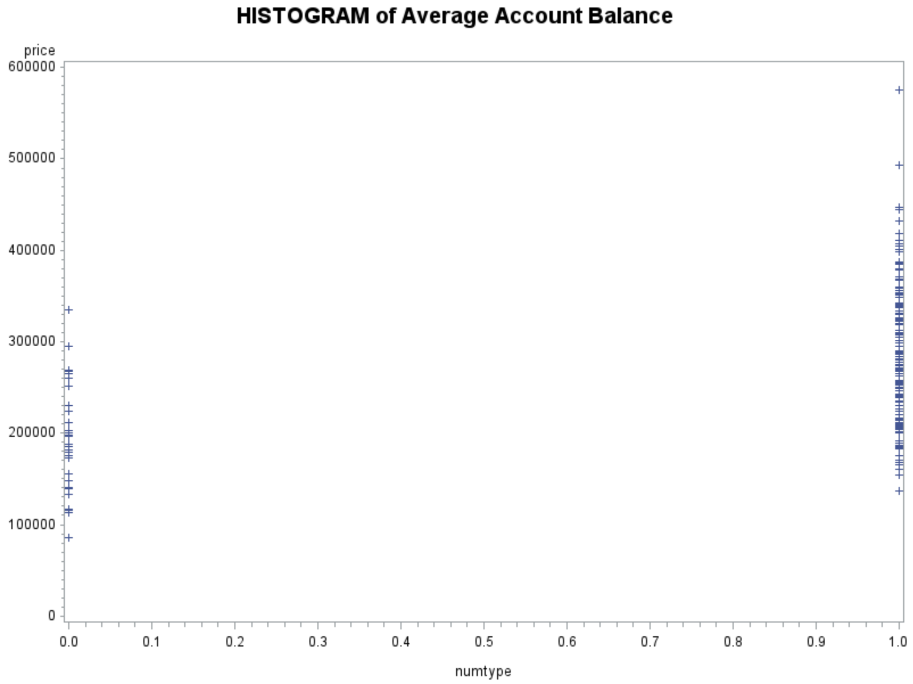


Points are randomly scattered, and residual analysis show no concern for the model fit.

**Assumptions for linearity:**

1. Scatterplot for each x-variable.



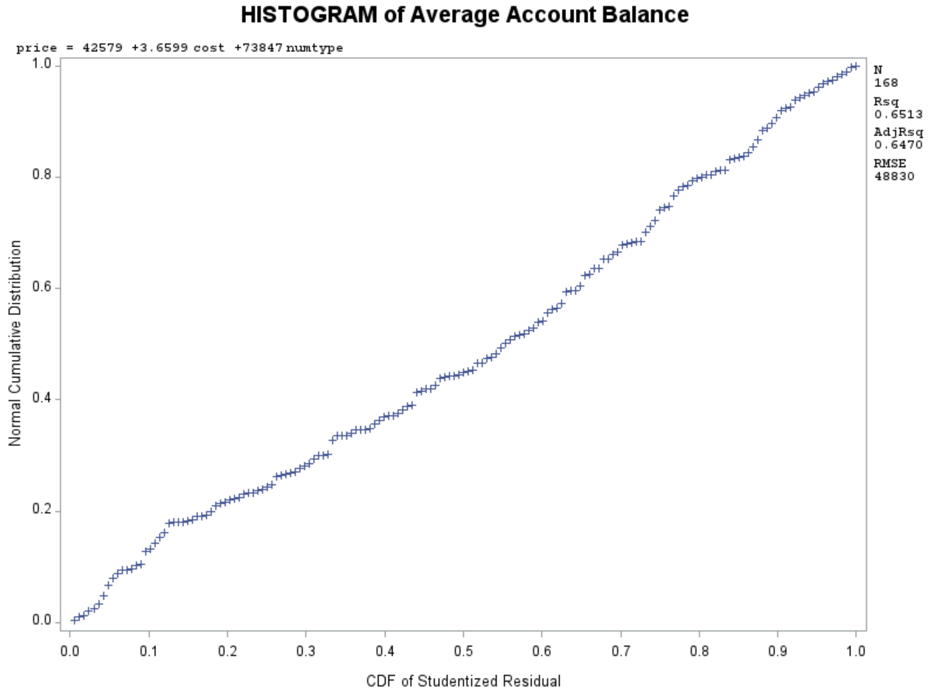


The correlation between price and home type can't show any linearity because it's a dummy variable. Therefore, all dots will be scattered around 0 and 1. On the other hand, the association between price and cost show an insignificant linearity as it's shown in the graph since it starts to spread when the price gets higher

2. Plot residuals vs each x-variable.
   As it's shown in the graph in the previous page, none of them show any linearity.

**Assumptions for Normality:**

Plot the normal probability plot of the residuals. It's almost a straight which means it's normal.



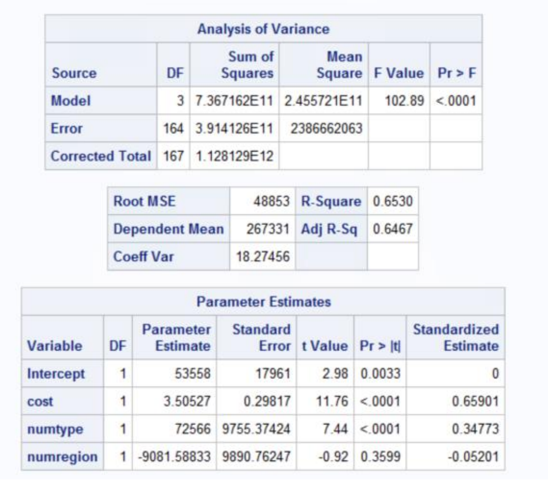
**f) Discuss what the regression model indicates for the relationship between price and home type (i.e. interpret the coefficient values).**

**Sale price =** 42579 + 3.659*Cost + 73847*numtype + e

The coefficient value of the parameter of X measures the predicted change in Y for any unit increase in X while the other independent variables stay constant. Single family home type will increase the price by $73847 compared to condominium style.

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

South and Midwest do not have the same mean sale prices



Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 7.367162E11 | 2.455721E11 | 102.89 | <.0001 |
| Error | 164 | 3.914126E11 | 2386662063 | | |
| Corrected Total | 167 | 1.128129E12 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 48853 | R-Square | 0.6530 |
| Dependent Mean | 267331 | Adj R-Sq | 0.6467 |
| Coeff Var | 18.27456 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
|---|---|---|---|---|---|---|
| Intercept | 1 | 53558 | 17961 | 2.98 | 0.0033 | 0 |
| cost | 1 | 3.50527 | 0.29817 | 11.76 | <.0001 | 0.65901 |
| numtype | 1 | 72566 | 9755.37424 | 7.44 | <.0001 | 0.34773 |
| numregion | 1 | -9081.58833 | 9890.76247 | -0.92 | 0.3599 | -0.05201 |

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

```
* a;
```

```
data HouseSlaes;
infile "S:\HW3\HouseSales.txt" delimiter = '09'x missover firstobs=2;
input region $ type $ price cost;
numtype = 1;
if type = 'T' then numtype = 0;
numregion = 1;
if region = 'S' then numregion = 0;
run;
proc print;
run;

*b;
proc sgscatter;
matrix price cost numtype numregion;
run;

proc gplot;
plot price*(cost numtype numregion);
run;

proc corr;
var price cost numtype numregion;
run;

*Model 1- full model with all predictors;
proc reg;
model price= cost numtype numregion /stb;
run;

*c;
*Model 2: remove numregion because it's not sig;
proc reg;
model price= cost numtype /stb;
run;

*d;
proc reg;
model price=cost numtype;
* Residual plot: residuals vs x-variables;
plot student.*(cost numtype);
* Residual plot: residuals vs pred. values;
plot student.*predicted.;

* Normal probability plot or QQ plot;
plot npp.*student.;
run;

*g;
proc reg;
model price= cost numtype numregion /stb;
run;
```