**Aamir Patel**

**Data Analysis And Regression**

**Assignment-3** | **Total Points: 20**

Note:

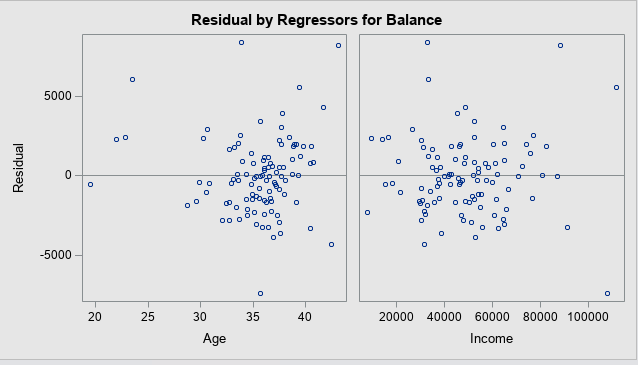
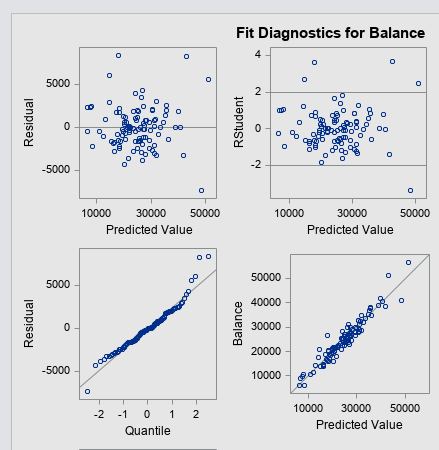
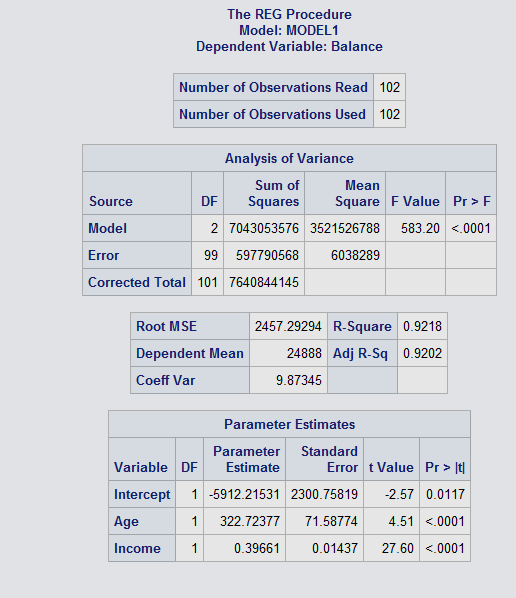
* All assignments should be submitted in a **single MS WORD format**, no PDFs or any other file types will be accepted. If you submit any other file type, it will not be graded.
* No extensions will be given unless for a documented reason specified in the syllabus, no late assignments past the due date even a couple of minutes late will be accepted as you have an extra day (8-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.

***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.***

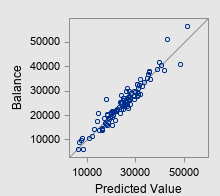
**Problem 1 [5 pts]**

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.

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

* 
* 
* Upon looking at the residual charts, I noticed that both of the charts seem to independent and does not follow any pattern. I also noticed that they are also very linear, and that is very linear.
* 
* Another thing that I noticed was that the Adjusted R^2 was very high at 0.9202 or 92.02%.
* Also, I noticed that the RMSE was not optimal, because the RMSE value was higher than it should be at 2457.29. while the F-value was very good at 583.2 with a p value of < .0001.
* However, even with the RMSE issues, the model assumption is satisfied and meets all the requirements and has a great f value, we can consider this model to be okay to predict data.

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.

* Global F Test: Hypothesis: H0 (Balance) = B1 (Age) + B2 (Income) = 0
* Ha = Bj /= 0
* 

Since our output from SAS has a linear output of the predicted values, we can safely say that the there is a relationship between Y(Balance) and X’s (age and income).

Also, We can see that the F value is very accurate at 92% with a p value of <0.0001, so we can safely say that it passes the Global F Test.

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

\*Import Statement;

**PROC** **IMPORT** datafile="banking.txt" out=main\_data replace;

delimiter='09'x;

getnames=YES;

datarow=**2**;

**RUN**;

\*Print dataset main\_data;

TITLE 'Dataset of banking.txt File';

**PROC** **PRINT**;

**RUN**;

\*Print Updated Regression model with no Education variable";

Title "Regression Model without Education Variable.";

**PROC** **reg** data= main\_data;

model Balance = Age Income;

plot residual.\*( Age Income);

plot test.\*predicted.;

**RUN**;

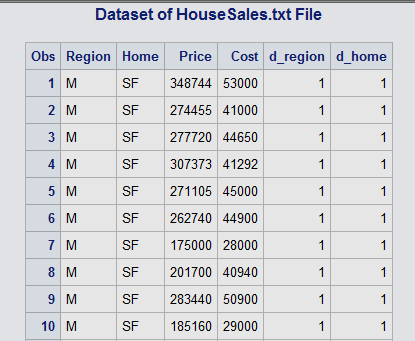
**Problem 2 [15pts]**

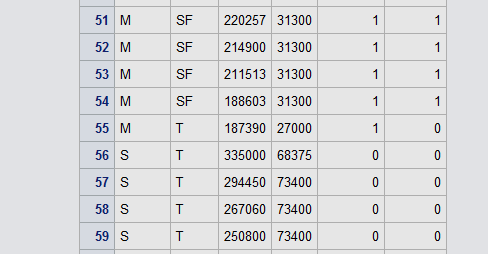
A national homebuilder builds single-family homes and condominium style townhouses.

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.

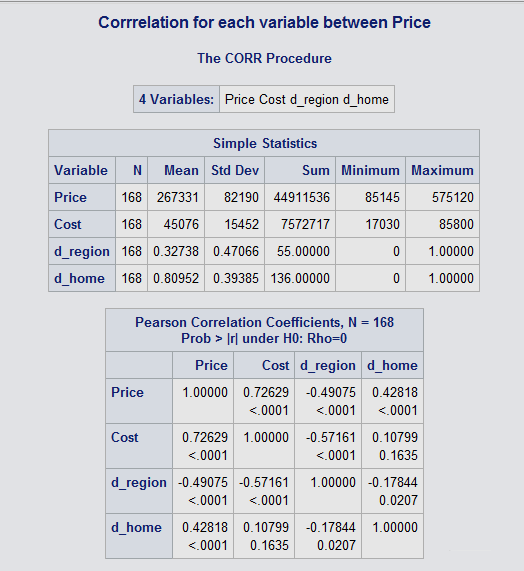
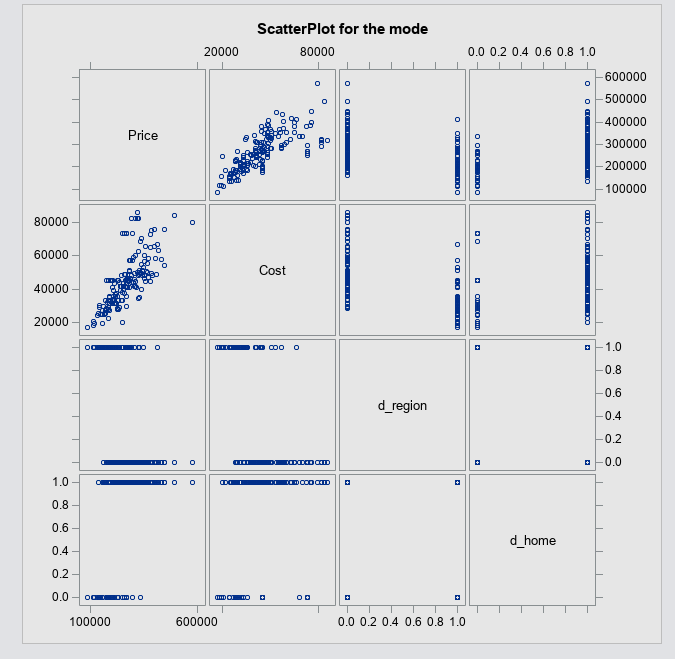
1. Define the dummy variables for region and home (write them down here), and create them in SAS.

* The Dummy variables I used in my SAS model are d\_region as a dummy variable for Region, and d\_home as a dummy variable for Home.
* I have attached couple screen shots to show that they work.

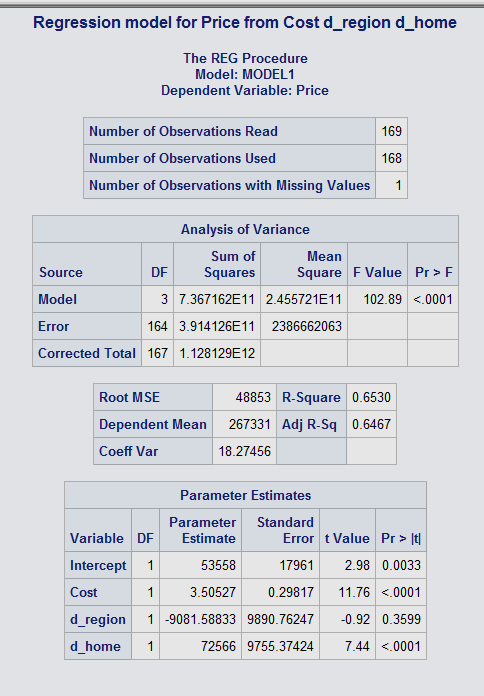


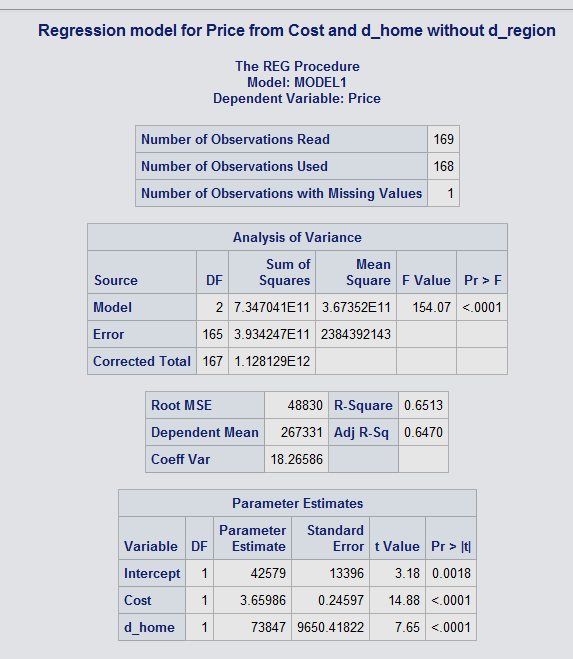


1. Analyze the association between selling price and each individual attribute (cost, home and region) using appropriate statistics and graphs. Discuss your findings. Include the relevant output.

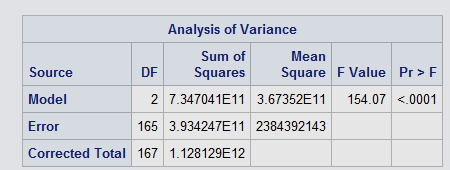
* 
* 
* Upon looking at the data generated using scatterplots and Proc Corr on sas. I can confidently say that the correlation between price and cost is positive at 0.72629, while the correlation between price and region isn’t very accurate since it is only in 0 and 1, so we can not use that as a point to refer back to. This applies to the correlation between Price and Home type, because it is also only in 0 and 1.
* None the less the correlations are as follows, price and Home Type has a positive correlation at 0.42818. while price and Region has negative correlation at -0.49075.

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.



* While looking at the initial regression model for the sales price we can see that the dummy variable d\_region is not significant enough keep in our final model, so we’ll have to create another model without the variable to get an accurate representation on what is going on.
* 
* This new regression model is now a bit more accurate than the previous model with all variables. So now that we know it is all the variables are accurate, we can write the equation for the regression model.
* **Equation: Price = 42579 + 3.659 (Cost) + 71847 (d\_home)**

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.



H0 = 0 (equals to)

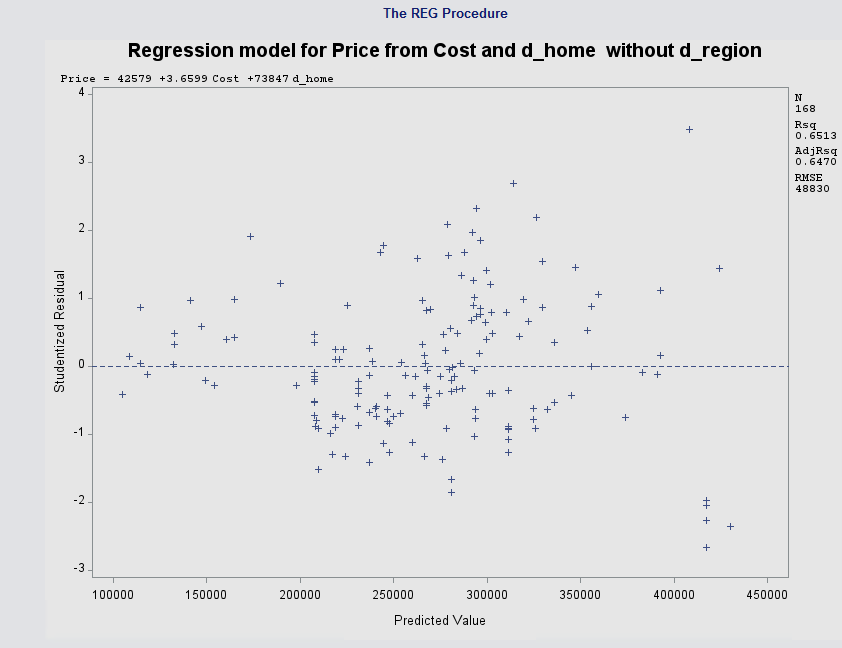
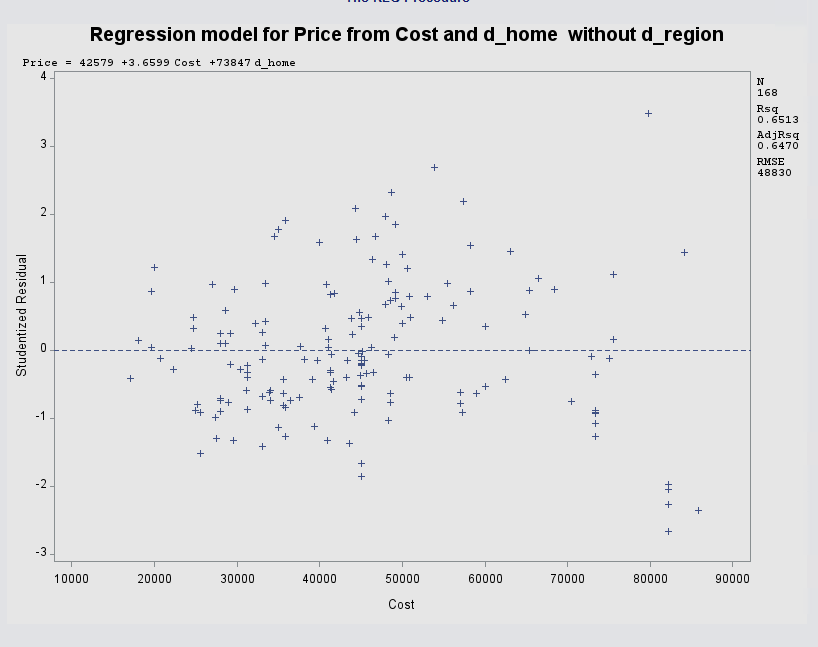
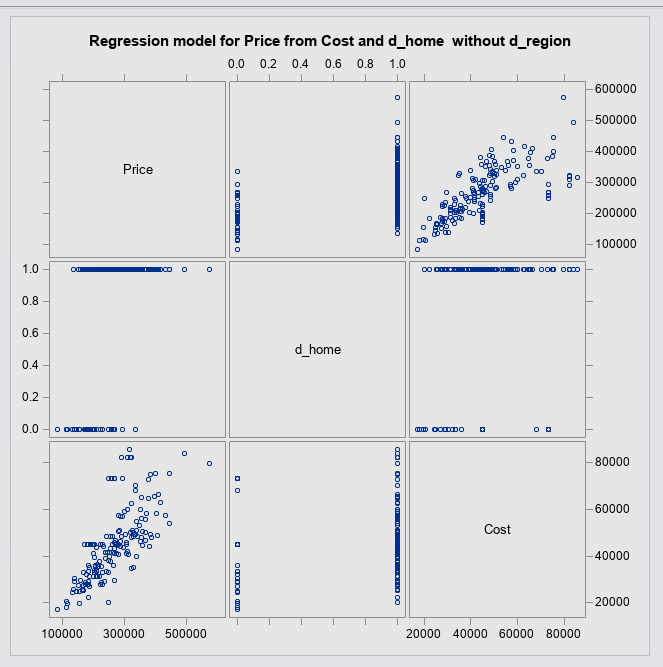
Ha /= 0 (not equals to)

F Value: MSR / MSE

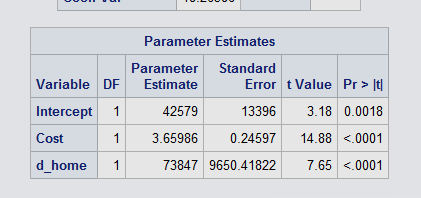
**: 102.89 at p value of <0.001**

In other words, with the F value we can eliminate the H0 because the p value is less than < 0.05.

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

* 
* 
* 
* 
* Upon looking at the all the charts created using SAS, we can say that all of the variables are independent and they also seem linear as shown in the scatter chart. So, we can safely say that we can use this model, besides the d\_home because it is a dummy variable and all of the values are in 0 and 1.

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

* 
* To determine the indication we can go back and look at the equation that we created which was:

Equation: Price = 42579 + 3.659 (Cost) + 71847 (d\_home)

**So, we can say that each time a new home gets added the price goes up by 71,847.**

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

* Upon taking a closer look at the equation that was generated, and parameter estimates we can see that region is **not a significant variable** in the price.

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

**data** HouseSales;

infile 'HouseSales.txt' delimiter='09'x missover;

input Region $ Home $ Price Cost;

d\_region =**1**;

if region = 'S' then d\_region = **0**;

d\_home =**1**;

if home = 'T' then d\_home = **0**;

**RUN**;

\*Print dataset main\_data;

TITLE 'Dataset of HouseSales.txt File';

**proc** **print**;

**run**;

\* Print ScatterPlot for the data;

title "ScatterPlot for the mode";

**PROC** **sgscatter**;

MATRIX Price Cost d\_region d\_home;

**RUN**;

\* Print Correllation to Balance for all variables;

title "Corrrelation for each variable between Price";

**PROC** **corr**;

var Price Cost d\_region d\_home;

**RUN**;

\*Print Regression Model;

Title "Regression model for Price from Cost d\_region d\_home";

**proc** **reg** ;

model Price = Cost d\_region d\_home;

**RUN**;

\*Print Regression Model;

Title "Regression model for Price from Cost and d\_home without d\_region";

**proc** **reg** ;

model Price = Cost d\_home;

plot student.\*(predicted. d\_home Cost);

plot npp.\*student;

**RUN**;

\*Print Scatterplot for model;

**proc** **sgscatter**;

matrix Price d\_home Cost;

**RUN**;