**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.

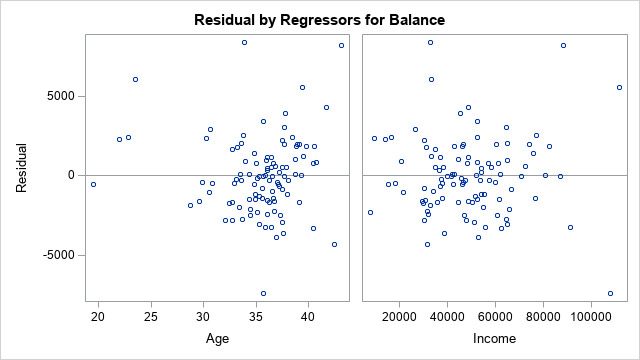
***Note: For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.***

**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.

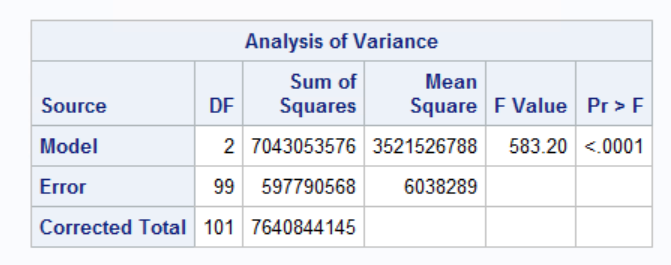
**Residual Plots of banking.txt:**



**My findings on the Age residual plot:** For the Age residual plot I think there is a pattern with the residual plot meaning that at the beginning there is almost to no data plotted but as the graph goes on to the Age 30 to 40 where almost all data is being plotted meaning that over time it will start to increase opening up the data. The spread increases as the age gets older making this funnel shape of data. I have drawn what I see as this spread increasing over time.After drawing what I believe to be an increase in spread shows that in the lines drawn with some outliers but the main clusters of data is plotted mostly after Age in 30-40 with nothing being plotted much at the beginning of this data making a pattern.

**My findings on the Income residual plot:** For the Income residual plot I believe that there is a pattern plotted which forgive my drawing but this I would say is this U shape where at the bottom of the graph under 0 where it starts staying constant but after income starts to dip down at the $100,000 mark to -5000 making that curved shaped pattern. Likewise with positive over 0 on this data where at $20,000 it will be constant positive but by the end at $100,000 that data will start to curve it up over 5000 at $100,000 making this symmetric like curve pattern for both sides. Thus, making this U-shaped curved pattern from what I see in this residual plot and what I have drawn.

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: None of the x-variables included in the model have any association with Y

Alternative hypothesis: At least one X-variable has a significant effect on changes in Y

Test statistic:

F=MSR/MSE

=3521526788/6038289 (circle 1 in Mean square)

=583.19 (circle 2 in the F value just did not round up)

F=583.19 and with p-value at <.001 less than 0.05 (at alpha=0.05)

Conclusion:

The null hypothesis says there is no association between any x-variable and y and should be rejected. The F-test gives strong support to the fitted model meaning since the F-test found that 583.19 is <.0001 when testing alpha=0.05, <.0001 is < 0.05 meaning there is at least one X-variable that has a significant effect on Y accepting the alternative hypothesis.

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

This is added code to Banking.txt:

\*Regression model to predict Balance;

**PROC** **REG** data=Balance;

MODEL Balance= Age Income;

\*Residual Plots;

PLOT student.\*predicted.;

PLOT student.\*(Age Income);

PLOT npp.\*student;

**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.

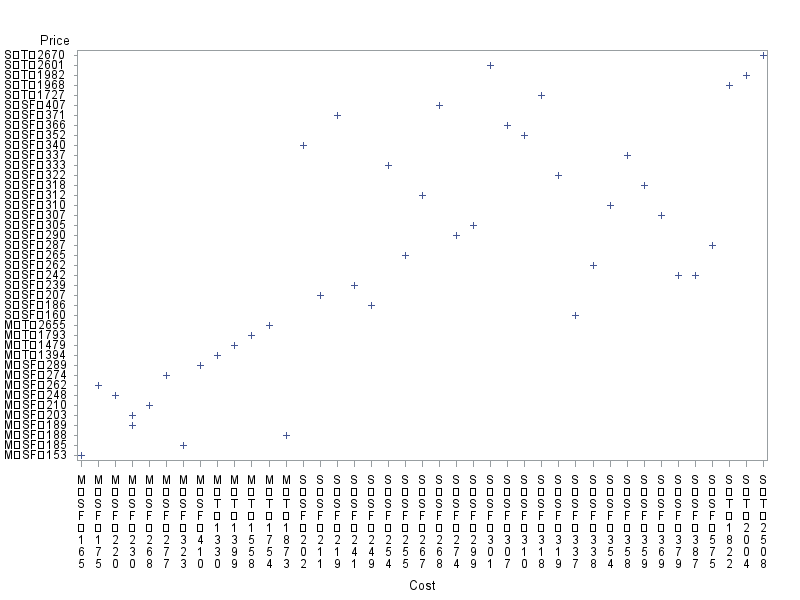
**The dummy variables for Region and Home I created are:**

d\_region which is the Region variable= The region I put the dummy variable on was M or Midwest. d\_region=1 for M (Midwest) and d\_region=0 for S (South)

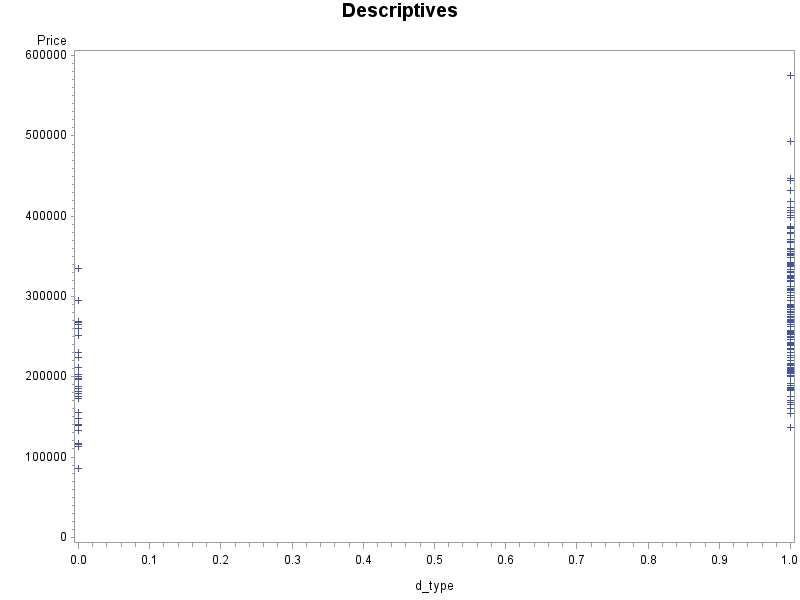
d\_type which is the Type variable=The House Type I put the dummy variable on was SF or Single-Family Home. d\_type=1 for SF (Single family home) and d\_type=0 for T (Condominium style)

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.

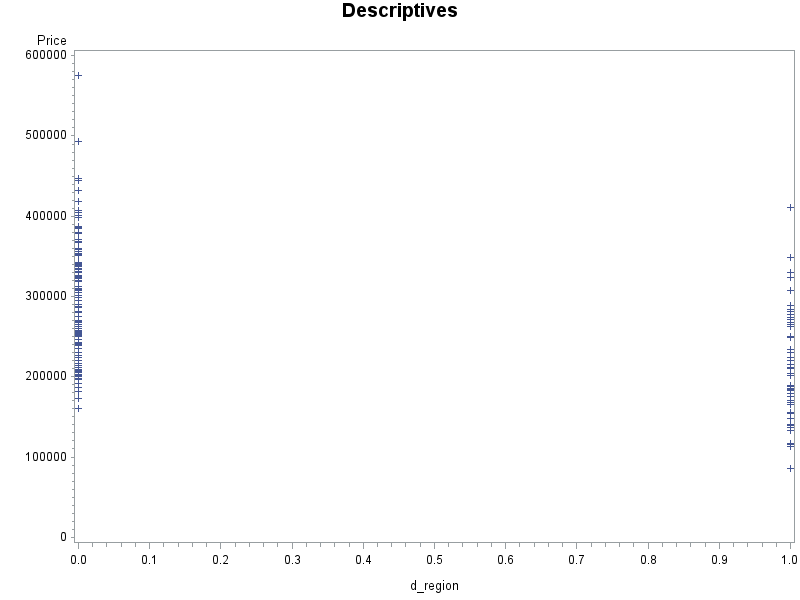
**Scatterplot for Selling Price and Cost:**



**Scatterplot for Selling Price and Home or d\_type:**



**Scatterplot for Selling Price and Region or d\_region:**

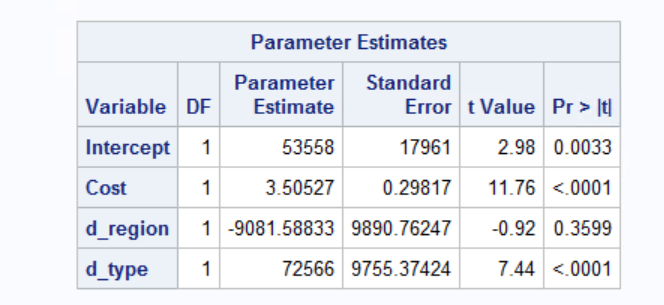


**My findings of these scatterplots:** My findings the first scatterplot that is Selling Price and Cost. This data at minimum is a positive correlation. I would say the association is low positive correlation. On the X-axis being Cost the data values from S, SF, 202 to S, SF, 575 that data ranges is a no observable obvious uniform like correlation and goes from 2601 to 186 on the y-axis being Price in that range being all over in that specified area. There is a positive correlation and actually in a linear fashion before and after that range, but in the areas I stated before it is a low positive because of the data being scattered and going out of unison.

The second scatterplot Price and Type or d\_type with dummy variables, I cannot see a linear relationships since points are scattered along 0 and 1.

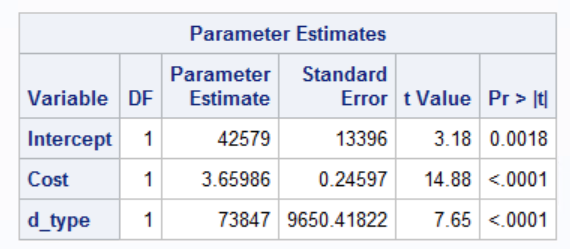
The last scatterplot is between Price and Region or d\_region with dummy variables, I cannot see a linear relationships since points are scattered along 0 and 1.

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.  
   **Parameter Estimates before removal:**



Variables that are not significant are ones that are above alpha test=0.05. That would mean the d\_region is not significant because 0.3599>0.05 which means it needs to be removed from the Parameter estimates and the regression analysis should be ran again.

**After removing d\_region from the regression analysis output:**

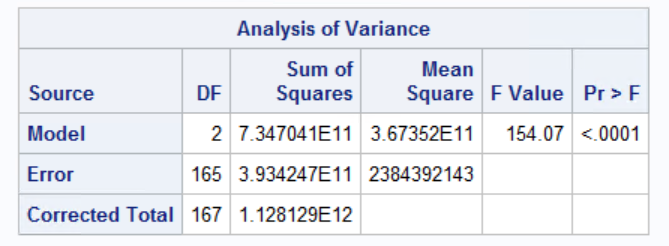


Both Cost and d\_type are significant because they both pass the test alpha=0.05 meaning since both are <.0001 which is less than 0.05.

**The model equation after the removal of d\_region is:**

Price=42579+ 3.65 Cost + 73847 d\_type

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: None of the x-variables included in the model have any association with Y

Alternative hypothesis: At least one X-variable has a significant effect on changes in Y

Alternative hypothesis:

F=MSR/MSE

=36,735,200,000,000,000/2,384,392,143 (circle 1 values)

=154.06 (circle 2 values did not round up so value a little different)

F=154.06 with p-value <.001 is less than 0.05 (at alpha=0.05)

Conclusion:

The null hypothesis says there is no association between any x-variable and y is rejected. The F-test gives strong support to the fitted model meaning since the F-test found that 154.07 is <.0001 when testing alpha=0.05, <.0001 is <0.05 meaning there is at least one x-variable that has a significant effect on Y accepting the alternative hypothesis.

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



**My findings of the Cost Residual Plot**: This residual plot I believe does not have a pattern. The best pattern I tried to draw on the residual plot for Cost was a funnel but when doing it made it very wide trying to angle to fit most data makes it not look like a funnel. It is uneven where it fits most of the data on each part of the positive or negative portion of the 0 not making it look like a funnel and very wide to where I think this residual plot does not have pattern and the spread is spread constant.

**My findings of the d-type Residual Plot**: Since it is a dummy variable I can not see a linear relationship since points are scattered along 0 and 1.

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

**Home Type**: It is a positively associated with Price. Model shows that assuming all other variables are constant, for every additional 1 home type, price increases by $73,847.

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

The first way I would check is by separating the data and next time having it in separate files to run comparisons with either hand calculating the means or against with the SAS program to truly find the median number of South and Midwest regions. Instead of having totals combined in one file for this question for better specific results since SAS will calculate a mean for both in individual files and a more accurate conclusion but I believe that there is different mean sale prices for the two regions. Logically it makes sense one region would be higher than another because of terrain, what’s in the region rural areas or urban cities have factors in the prices per region.

From looking at the scatterplots to being with when they were plotted looking at the price and d\_region scatterplot having the dummy variable allowed all the data for it to be plotted at 1 for M or Midwest and 0 for S or South which helped grouped the data. Looking to begin with at the 0 section being for S or South the highest price was ~$60,000 and lowest price was ~$16,000. For Midwest the highest value recorded in that region was ~$45,000 and the minimum was ~$8,000. To me seeing that the S or South had the higher maximum range than the Midwest would mean the mean sale prices would be different than compared and higher because if the highest value is ~$60k and lowest ~$16 it can never be lower than the Midwest which lowest was ~$8,000 and maximum was ~$40,000 meaning even at a high end mean that it will never be the same as the south especially with pull of prices in $60,000 bringing that mean up.

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

\*PROC IMPORT;

**PROC** **IMPORT** DATAFILE="HouseSales.txt" out=Sales\_import replace;

DELIMITER='09'x;

GETNAMES=yes;

DATAROW=**2**;

**RUN**;

**PROC** **PRINT** data=Sales\_import;

**RUN**;

**data** Sales\_import;

set Sales\_import;

d\_type=(Type="SF");\*Dummy variable for SF or single family home;

d\_region=(Region="M");\*Dummy variable for M or Midwest;

**RUN**;

**PROC** **PRINT** data=Sales\_import;

**RUN**;

\*Scatterplot GPLOT for Selling Price and Cost;

**PROC** **GPLOT**;

PLOT Price\*(Cost);

**RUN**;

\*Scatterplot GPLOT for Selling Price and Home;

**PROC** **GPLOT**;

PLOT Price\*(d\_type);

**RUN**;

\*Scatterplot GPLOT for Selling Price and Region;

**PROC** **GPLOT**;

PLOT Price\*(d\_region);

**RUN**;

\*Correlation Values;

**PROC** **CORR**;

VAR d\_region d\_type Price Cost;

**RUN**;

\*Regression Analysis;

**PROC** **REG**;

MODEL Price=Cost d\_region d\_type;

**RUN**;

\*Regression Model without d\_region;

**PROC** **REG**;

MODEL Price=Cost d\_type;

**RUN**;

\*Regression model to Predict Price;

**PROC** **REG** data=Price;

MODEL Price=Cost d\_type;

\*Residual Plot;

PLOT student.\*predicted.;

PLOT student.\*(Cost d\_type);

PLOT npp.\*student;

**RUN**;