

|  |
| --- |
| Assignment #3: Regression Model Building Part 2  *PREDICT 410* |

**Data:** The data for this assignment is the Ames, Iowa housing data set. This data will be made available by your instructor.

**Assignment Tasks**

In this assignment we continue building regression models for the home sale price (Y).

*PART A: Transformations – Comparison of Y versus Log(Y)*

1. In assignment #2 –for question #2, you identified the “best” continuous predictor variable X for use in predicting house SalePrice(Y). We will use that variable as the predictor variable (X) in this analysis. Read in the Ames Housing data and, in the data step, create two new variables. The first variable is log(SalePrice) or the log of SalePrice. The second is log(X). You can name these new variables whatever you wish, but usually it is best if you have a descriptive name to remind you what you did to create the new variable. For example, you could use the SAS code:

log\_saleprice = log(SalePrice) ;

log\_X = log(X) ; \\* naturally, you would fill in the variable name you are using for X \*/

Then do a PROC PRINT of your data to be sure SAS appended these two variables to your dataset.

1. Now, you should fit four models, one for each pair-wise combination of these two new variables you constructed. These should be: a) X to predict SalePrice, b) log\_X to predict SalePrice, c) X to predict log\_saleprice, and d) log\_X to predict log\_saleprice. Report each model in equation form and interpret each coefficient in the context of this problem. Based on the output from the regression analysis and the automatically generated ODS output, construct a summative table so that the fit of the four models can be easily compared. You will have to decide on the key metrics to include in the summative table. The models should be the rows and the metrics should be the columns of the summative table. Which of these four is the best fitting model? What concerns do you have about this “best” fitting model? What concerns do you have about using transformed variables? How is the interpretation of the log(SalePrice) model different from the price model?
2. Correlate the continuous variables in the dataset with log\_saleprice. Which continuous variable(s) correlate most with log\_saleprice? Pick an X variable that correlates highly with log\_SalePrice. Make a scatterplot of X with SalePrice and a second scatterplot of X with log\_SalePrice. What do you observe in these plots?
3. Under what conditions does the log transformation of the response variable (Y) improve the model fit? The log transformation is not the only transformation you could use. You can use any mathematical function to transform the response variable Y or a predictor variable X. Create another transformation variable on Y (SalePrice) using some other function, (i.e. square root, square, exponential) that seems appropriate. Fit one more regression model using the response variable X you obtained in part 3) and/or any transformation on that variable that you find appropriate to predict your newly transformed Y. Report each model in equation form and interpret each coefficient in the context of this problem. Use the automatically generated ODS output from SAS to assess the goodness- of-fit of this model. How does this model compare to those you obtained in part 2)?

*PART B: Outliers*

Outliers tend to be a big issue, especially in the newly emerging world of big data. The issue is that outliers can be different from the main body of the data due to data entry errors or not being affiliated with the target population, or they could be legitimate observations that are just different from the other data. In the first case, you would want to remove the outliers because if you retained them in your predictive modeling activity, they would distort the model. Whereas, in the latter case, deleting the outliers would potentially cause you to miss out on valuable information and pattern that needs to be modeled. You should know (and always have in the back of your mind) that some of the world’s greatest discoveries came from examining outliers. Some fields routinely just remove outliers as a matter of principal, regardless of the consequences, while other fields are more concerned about tracking down the original data source to confirm the validity of these data points. Regardless, you need to be able to identify and work with outliers using SAS.

1. Identify observations that are potential outliers for the SalePrice variable, Y. To do this, use a Histogram, scatterplots and PROC SORT on SalePrice(Y) to determine which records are of concern. You will have to decide on criteria to isolate the records that are potential outliers. For example, if SalePrice is negative, this is clearly a problem. The criteria would then be: If (SalePrice < 0), then do something! Often analysts create an if-then-else ladder that assigns each observation to a category. Categories then indicate the need for removal or indicate concern or identify an observation that is fine. A snippet of SAS code that could go into a DATA step to do something like this might be:

outlier\_def = 0 ; /\* a 0 would indicate a valid record \*/

if (SalePrice<0) then outlier\_def=1 ; /\* a code of 1 indicates negative sales data \*/

if ((SalePrice >= 0)&(SalePrice <= 30)) then outlier\_def=2 ;

/\* a code of 2 indicates too small of sale price, or some such thing \*/

If (SalePrice > 1000000) then outlier\_def=3 ;

/\* a code of 3 indicates price is too high and not part of the target population \*/

Define your 'outlier' observations using a rational process. You can use EDA and model validation to identify rules to identify the population that you do not wish to include in your model. What is the 2 standard deviation rule? Should you use that here? Assign an outlier code and an outlier definition for each outlier condition. Which observations did you remove and why? Report a table of counts for each of the outlier definitions using a PROC FREQ statement. NOTE: If you were doing this for a professional activity, you would do similar work on ALL of the variables in your dataset! You are NOT asked to do that here.

1. For modeling purposes you will be able to create a data set without outliers by including a DELETE statement in a subsequent new SAS data step. For example: if (outlier\_def > 0) then delete;

In a new data step, create a “Cleaned” Ames Housing data set by deleting the outliers you identified. Re-fit the models you ran in assignment #2, parts 2, 5 and 6 but this time use the “Cleaned” data to fit the models. Report each model in equation form and interpret each coefficient in the context of this problem. Put together a summative table, like the one you constructed for question 2) above, to assess and compare the goodness- of-fit of these models from Assignment #2 and Assignment #3. Did fit improve by removing the outliers? Discuss, compare, and contrast the model fits.

1. Model based Outliers aka influential points: Even after you have removed ‘outliers’, still you may have unusually large residuals, which you can see from the residual plots. These are called ‘influential’ points. Sometimes, we find that a small subset of ‘influential’ points exerts a disproportionate influence on the model coefficients. These points can be identified by several statistics such as DFFITS. Use the output data set to identify the influential points using the DFFITS statistic. Use the threshold value that is given in the text book on Page 218. Then refit the model after removing them. How many influential points did you find & remove? When you refitted the model, did the model improve?
2. In your conclusion / reflection section, be sure to address the following questions:
   1. In what ways do variable transformation and outlier deletion impact the modeling process and the results? Are these analytical activities a benefit or do they create additional difficulties?
   2. What do you consider to be next steps in the modeling process?

**Assignment Document:**

All assignment reports should conform to the standards and style of the report template provided to you. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary results or information. The document should be submitted in pdf format. Name your file Assignment3\_LastName.pdf.