1. You are very creative in the variables selection which is great.
2. Generally, this is a fairly complete report with most required elements.
3. My suggestion to you is to focus more on the diagnostic plots. Don’t ignore the hint when assumptions are severely violated. You need to think over the interpretation of coefficient estimates. Please check my related comments. Your model performance is not very good, which could be related to insufficient/inappropriate data treatment prior to modeling.

(The best Kaggle score of the class is 0.1261.)

# Introduction

The goal of Project 1 is to predict the final sales price (response) of each home in the Kaggle – Ames, IA dataset. Document your model selection approach and confidence.

# Data Description

The data is from Century 21 realtors. There are 1,460 observations of homes sold between 2006 and 2010 with 80 variables used to determine sales price. This is an observational study; therefore, no causal inference can be implied across the Ames, IA real estate market.

# Exploratory Analysis (updated)

Before Exploratory Data Analysis the issue of significant missing data is addressed. The missing data did not appear to be coding errors, rather just a lot of variables being tracked. Given the nature of this project with Kaggle requiring every row, the approach of deleting missing entries is not an option. With what is already known about the data set requiring transformation, the numerous entries with ‘0’ value, e.g., ‘0 HalfBaths’, cannot be deleted. Missing data imputation is complicated. The previous mean approach is not good for multivariate analysis since it hides the correlation. Given time and current experience, a combination of median and small value imputation[[1]](#footnote-1) is used. For those entries with ‘NA’ or missing values, the median value for the variable is substituted. For those entries that had a ‘0’, like ‘0 HalfBaths’, .0001 replaced the ‘0’ to allow for transformation.

Initial analysis of the numerical variables shows that the data does not sufficiently meet the assumptions of linear regression, like linear relationship, no multicollinearity, multivariate normality, etc. A log transformation is conducted on key variables to meet the linear regression assumptions. Side by side diagrams show the initial model and then the log transformed model.

Figure 1 - Fit Diagnostics Before Log on left and After Log on right

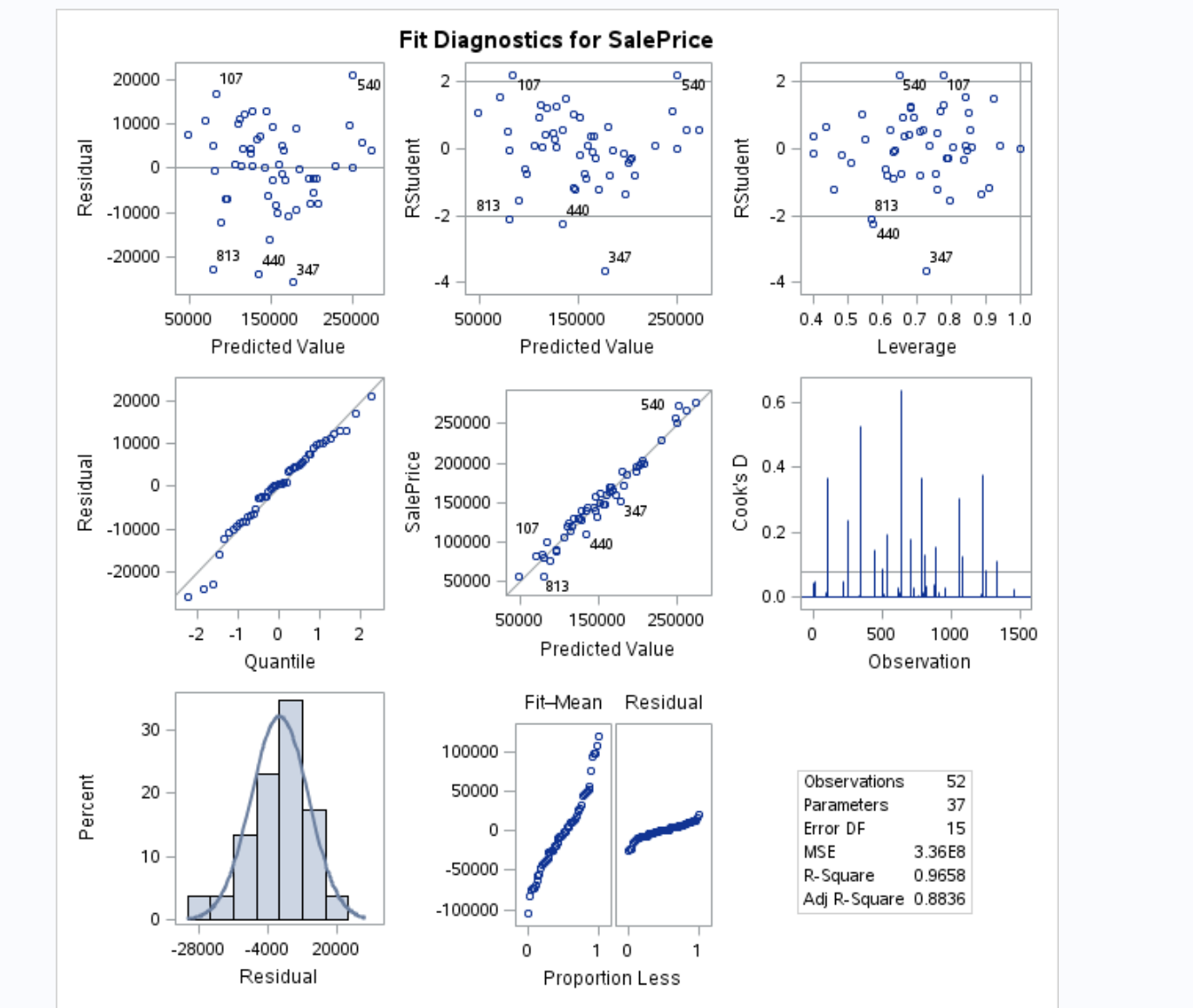
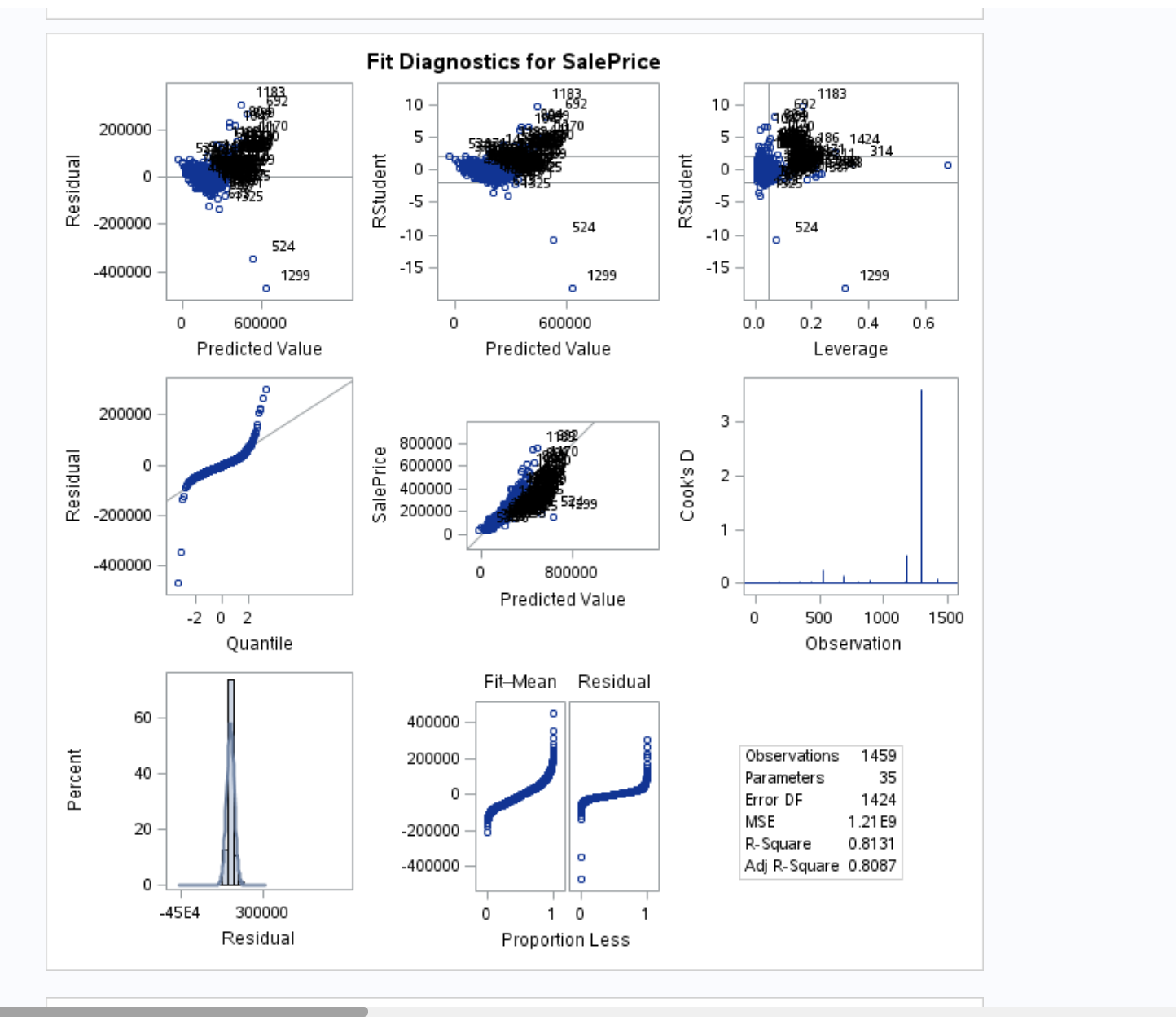
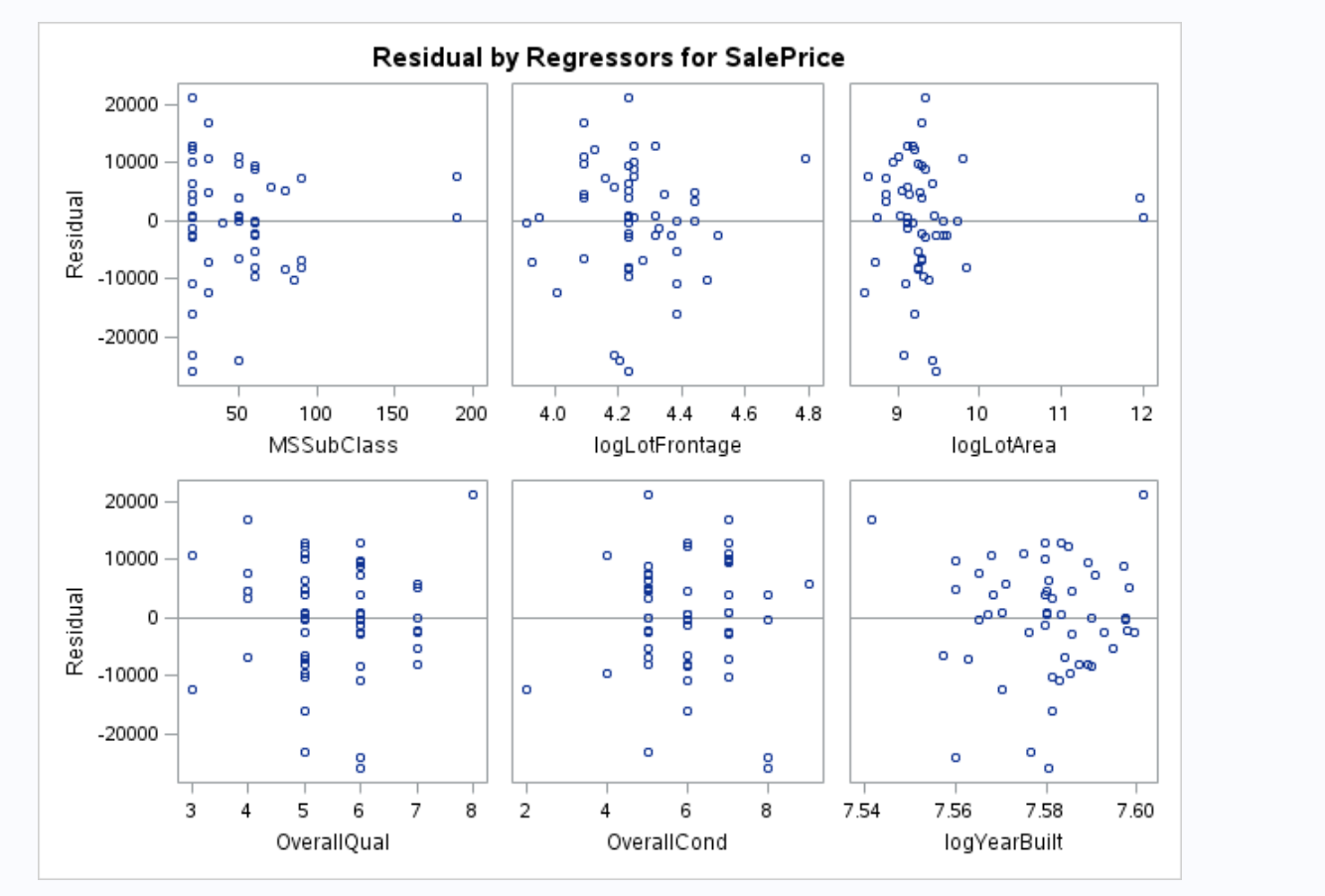
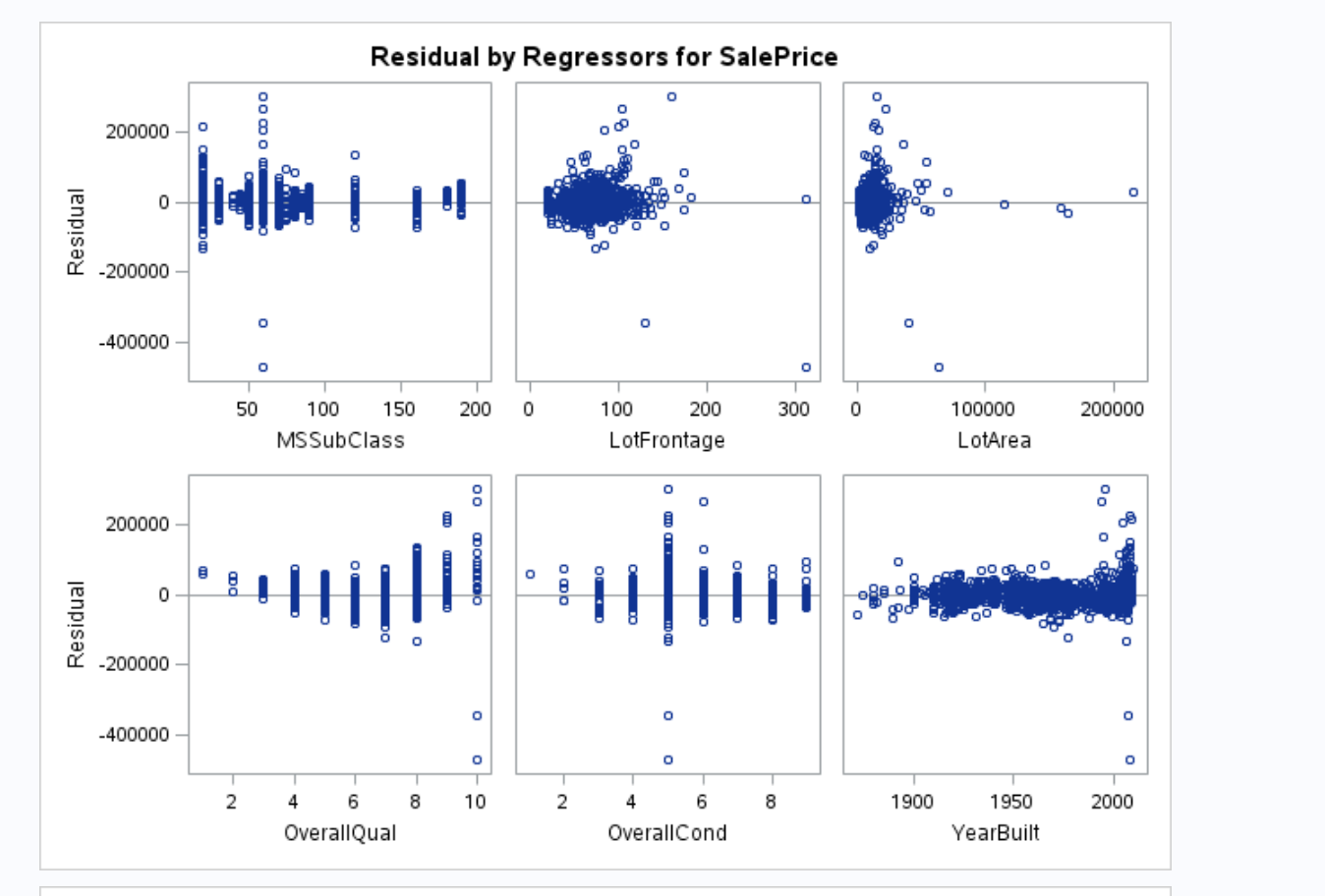
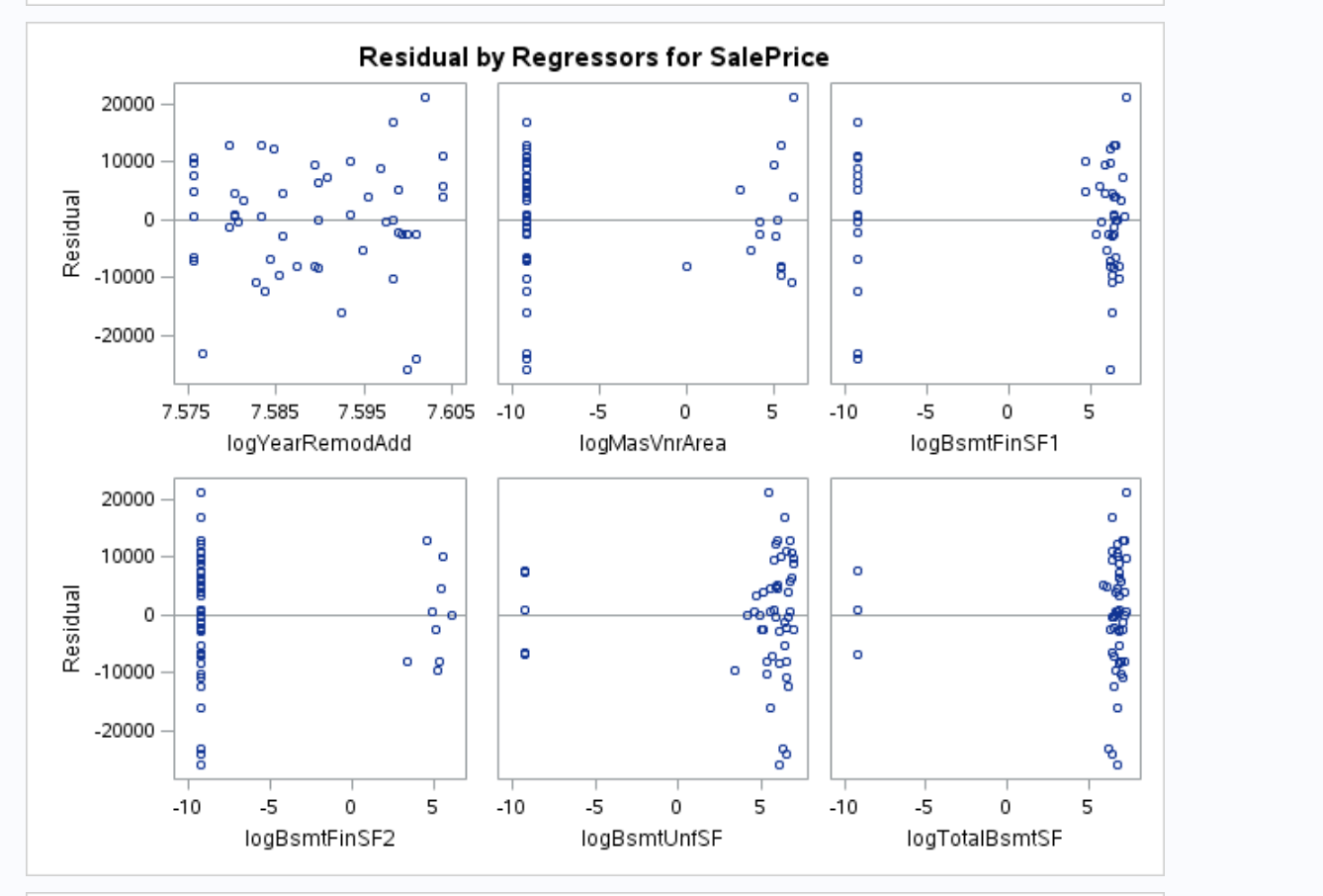
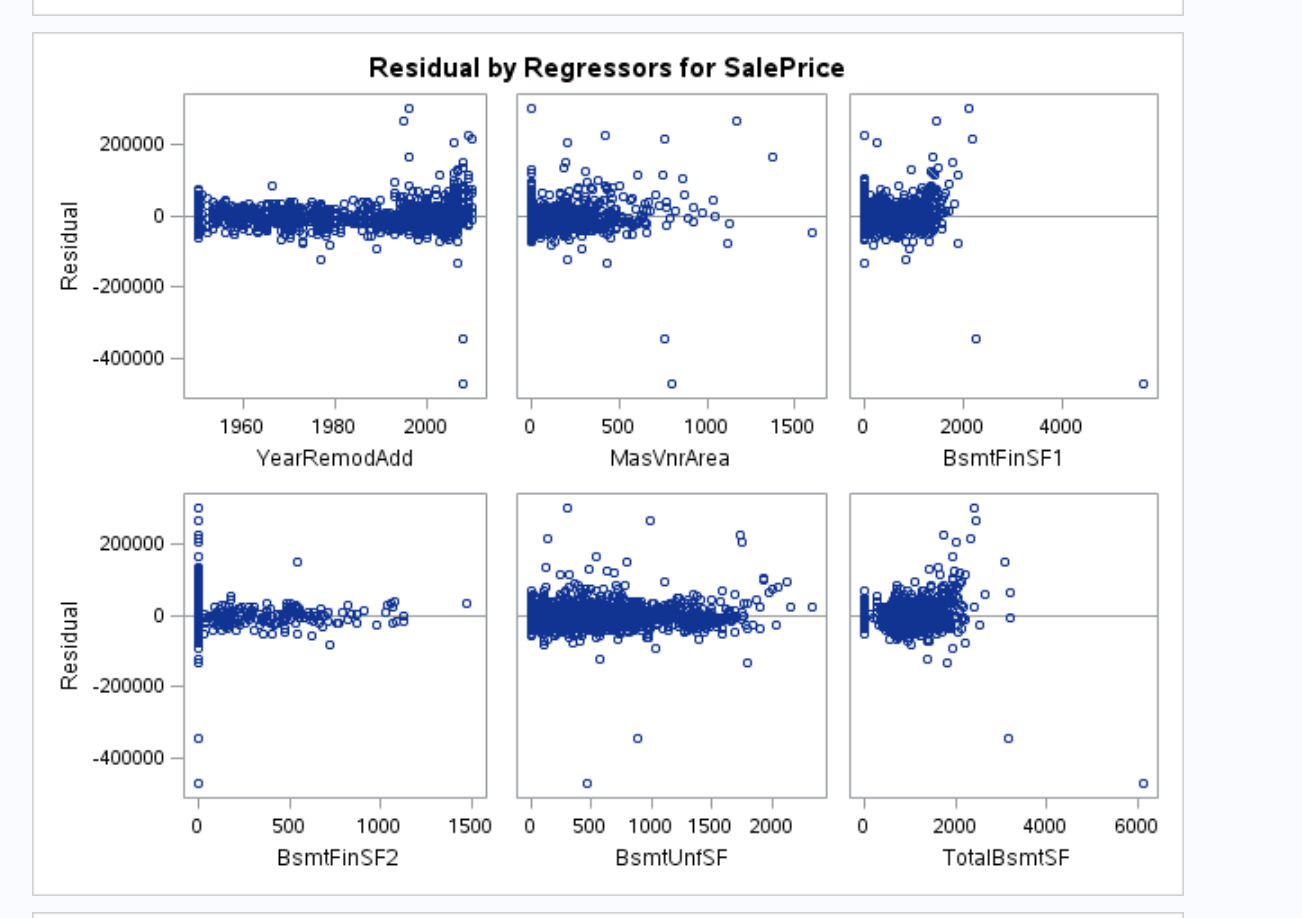
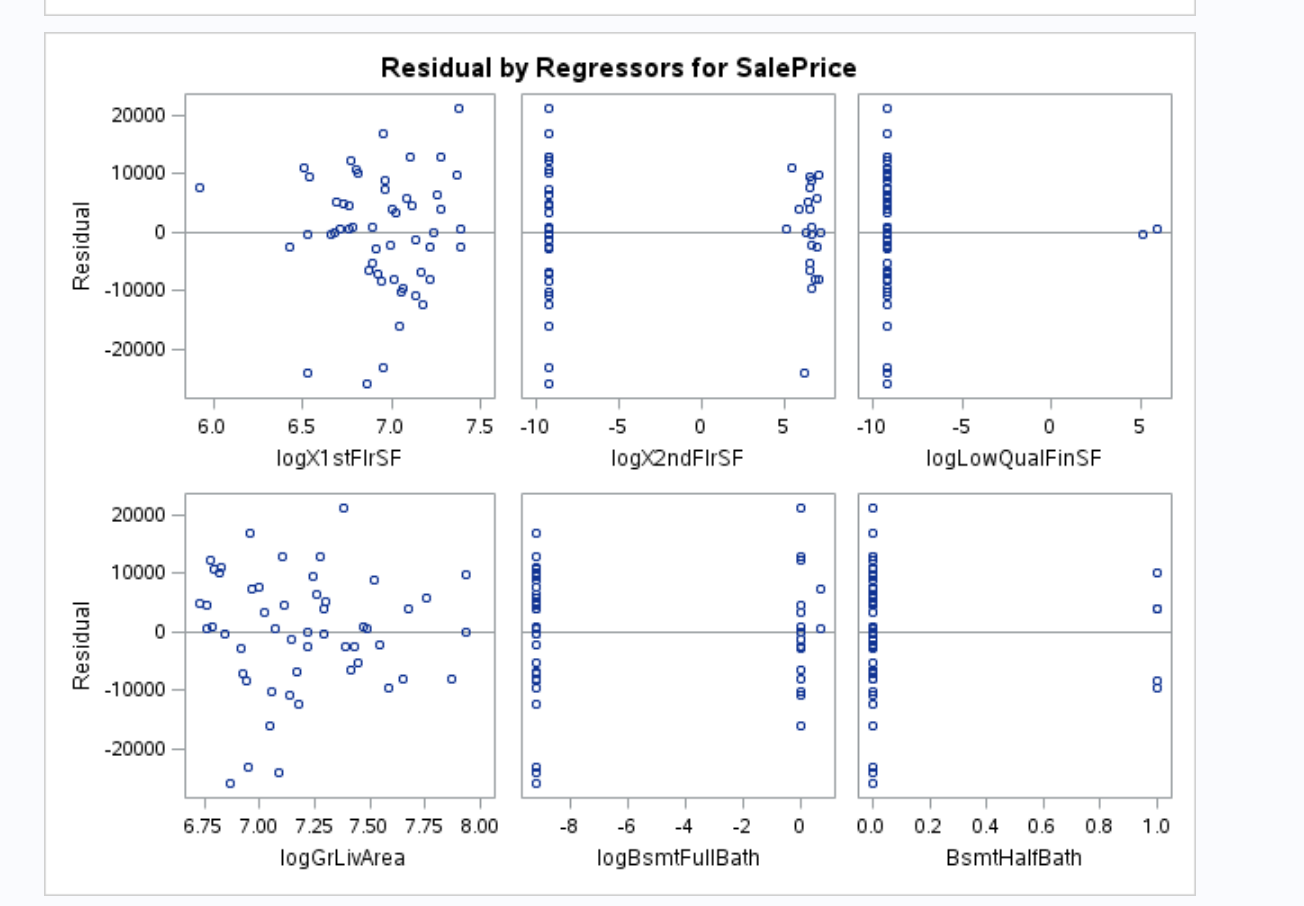
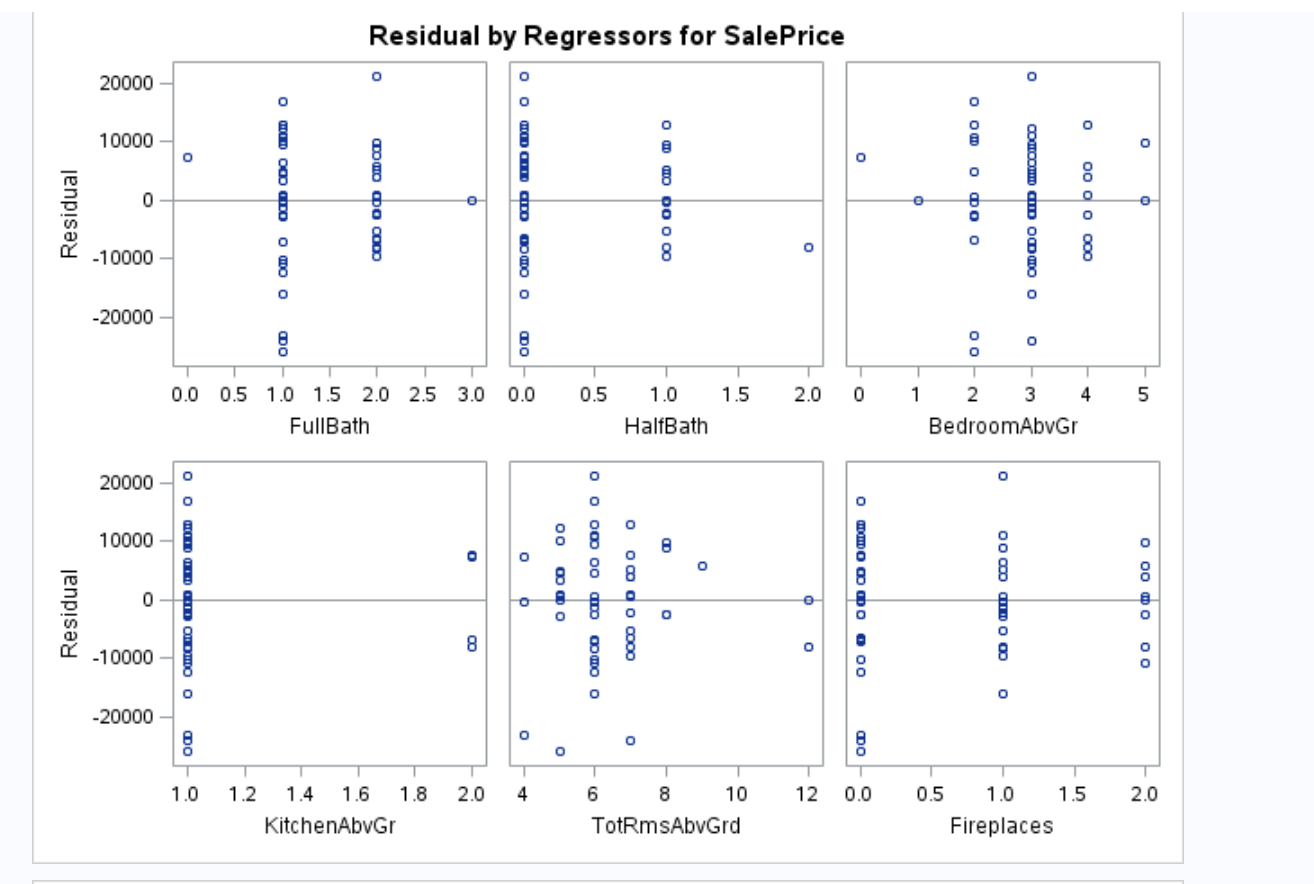
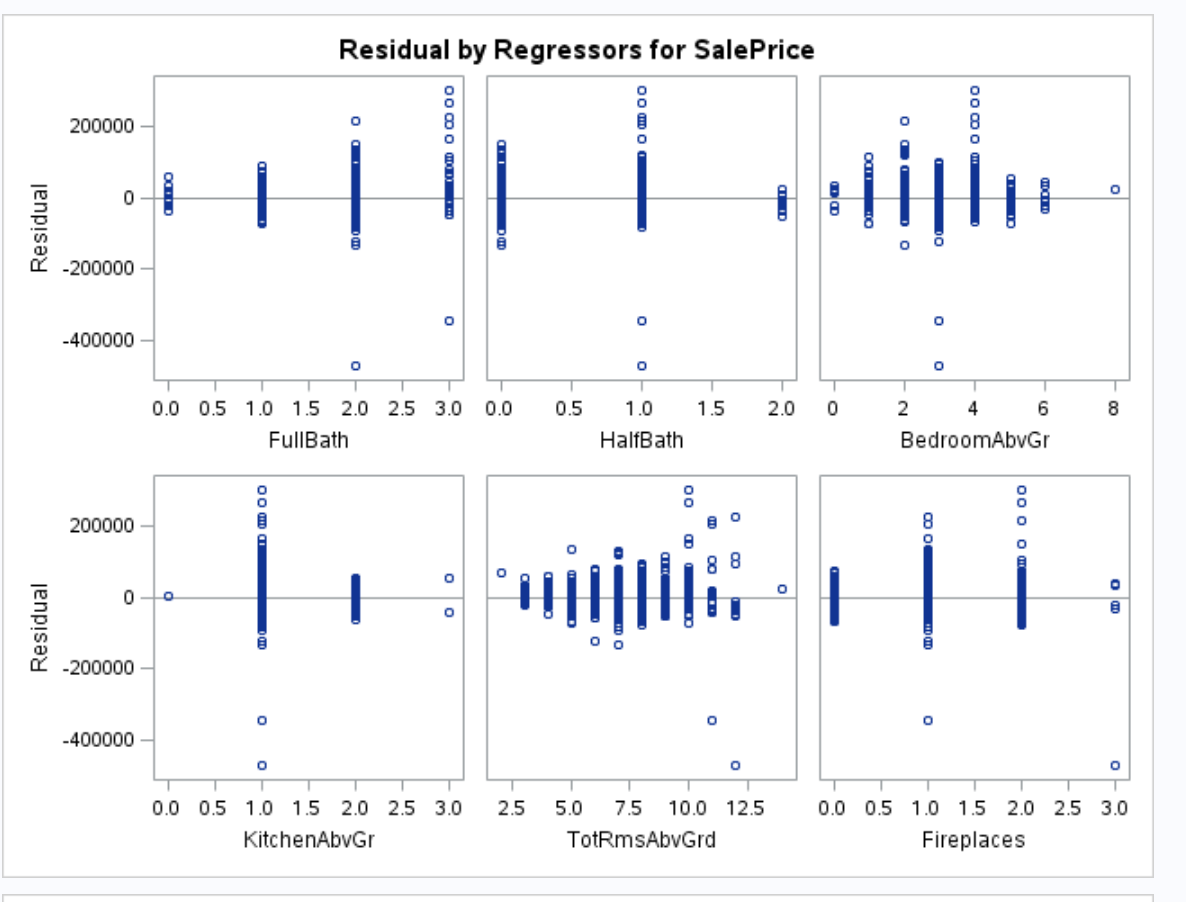


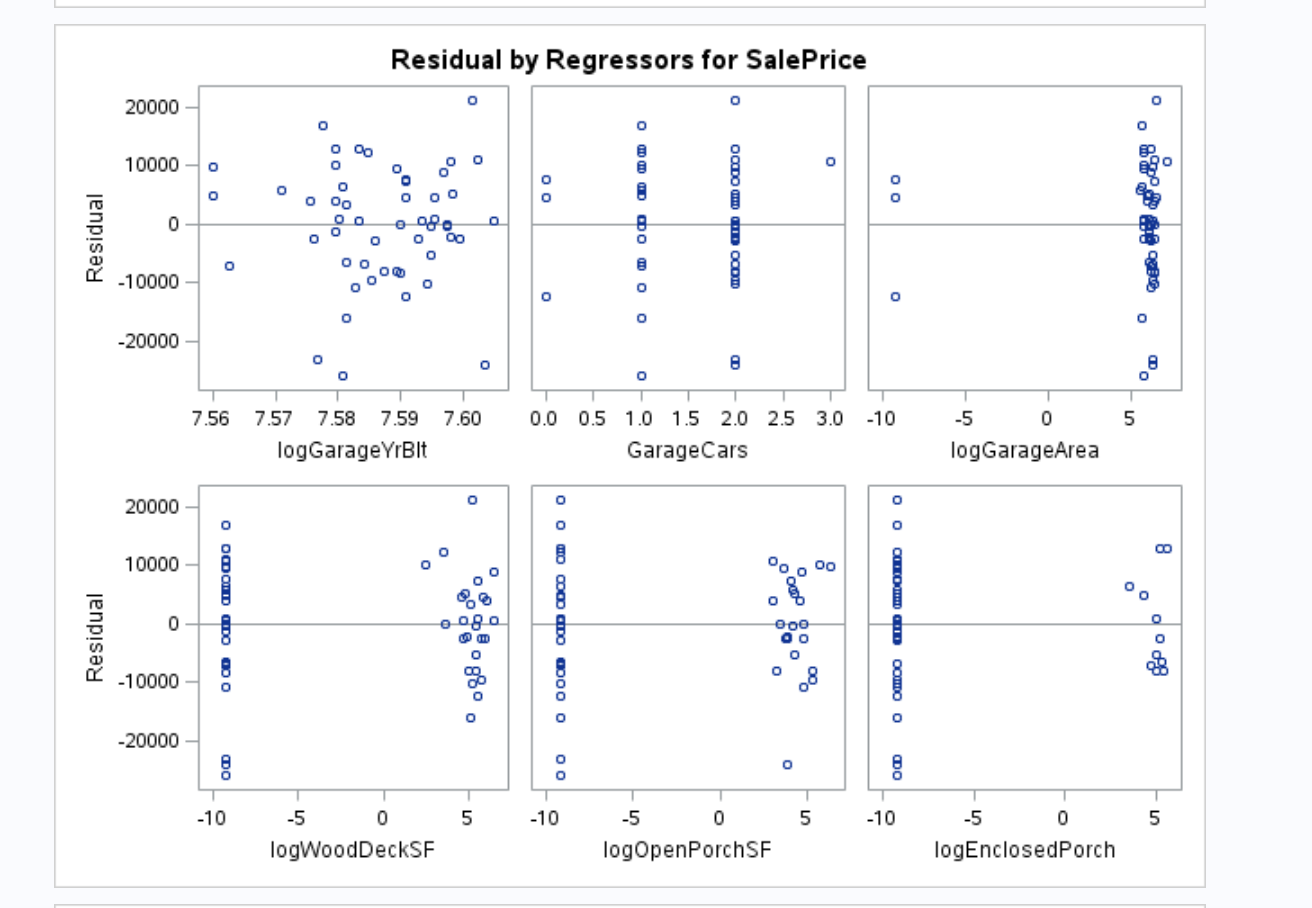
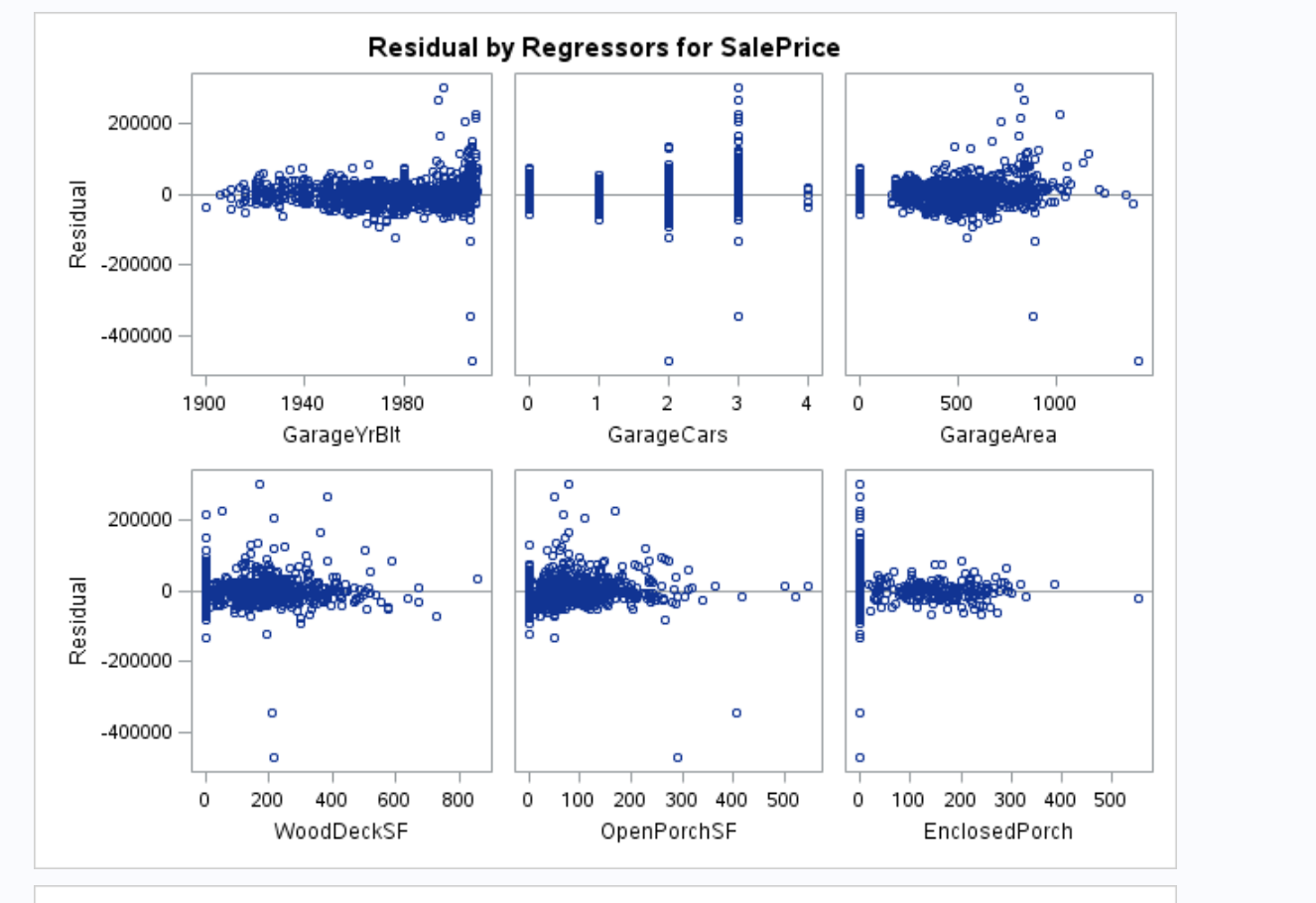
Figure 2 – Residual by Regressor - Before Log on left and After Log on right

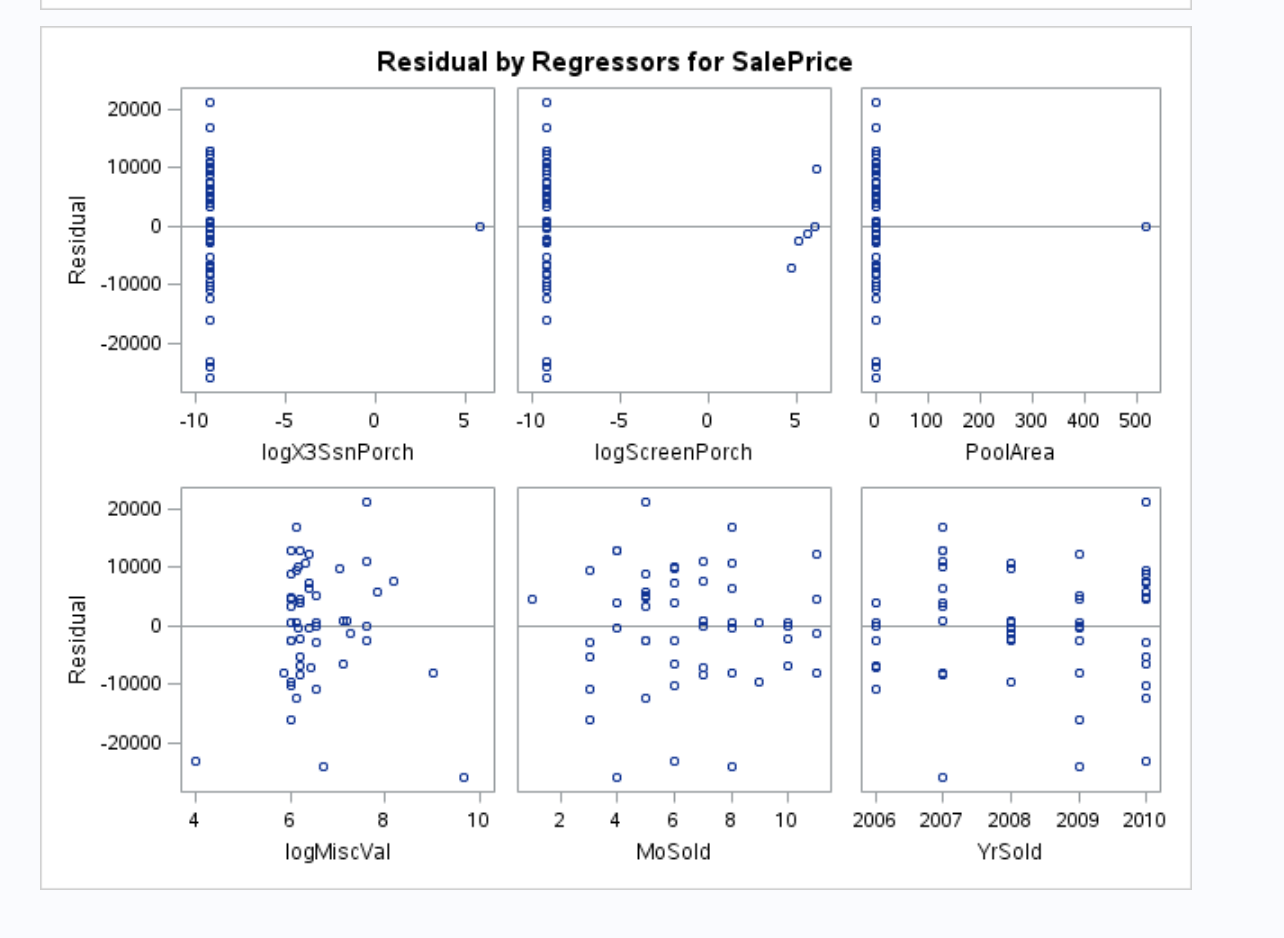
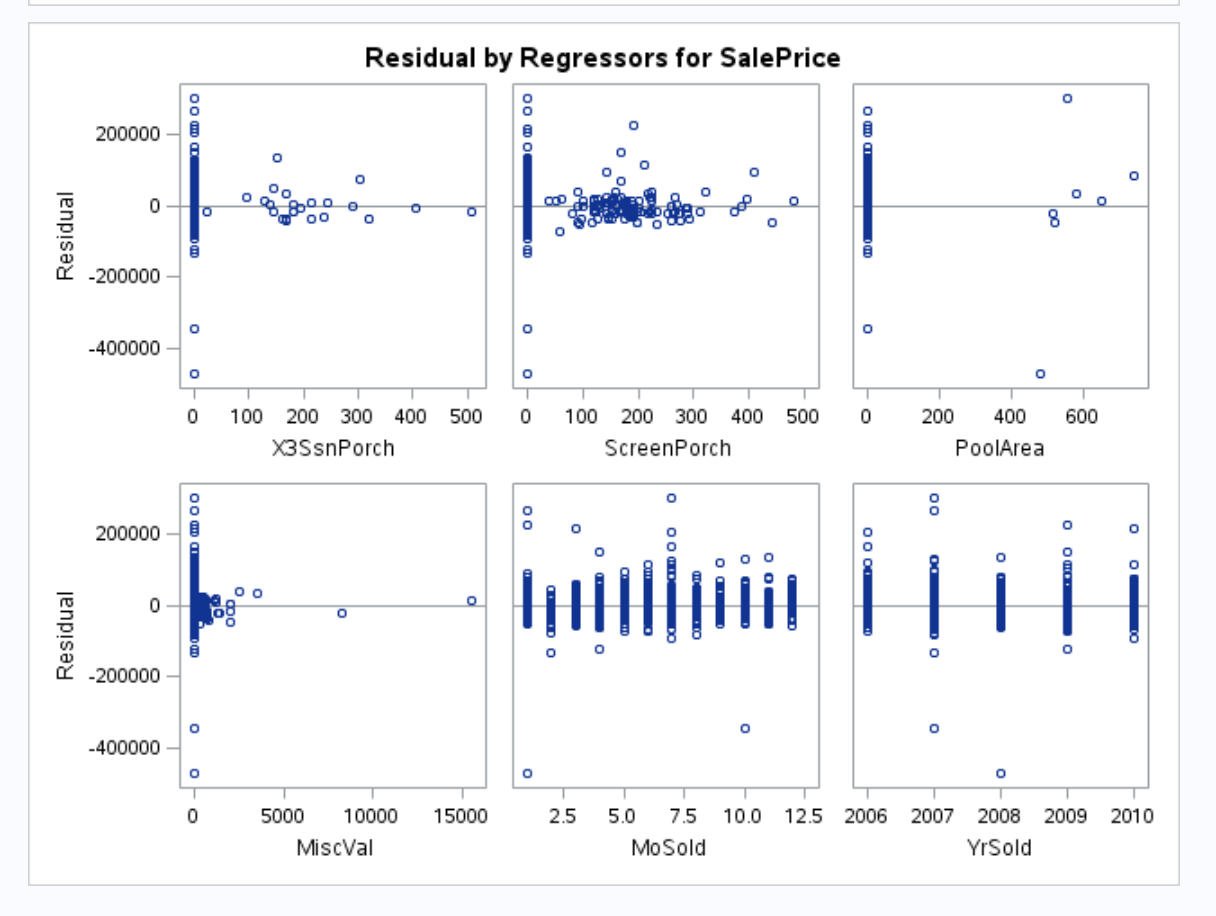












In the first project a second model was built to manage the complexity of 80 variables. Subsequent chapters in class taught rigorous statistical approaches to grouping, which will be used going forward.

## Model Selection

The goal is to create the most predictive model possible for Sales Price from the data that was also “useful.” The approach used was FORWARD, LASSO, STEPWISE, and HUMAN INFERENCE selection methods on the Dream Data (all variables), and the Data Mine data set (logical grouping subset variables).

The use of LASSO, a modern approach to model selection, is balanced by the use of traditional statistical approaches of FORWARD and STEPWISE. While some of our research discouraged STEPWISE, the paper *“Introducing the GLMSELECT PROCEDURE for Model Selection”[[2]](#footnote-2)* demonstrated that STEPWISE selection can be powerful when used for predictive analytics.

To save the reader’s time and space here, only two (2) model selection processes will be outlined on selection.

### FORWARD

#### Assumptions Dream Data model

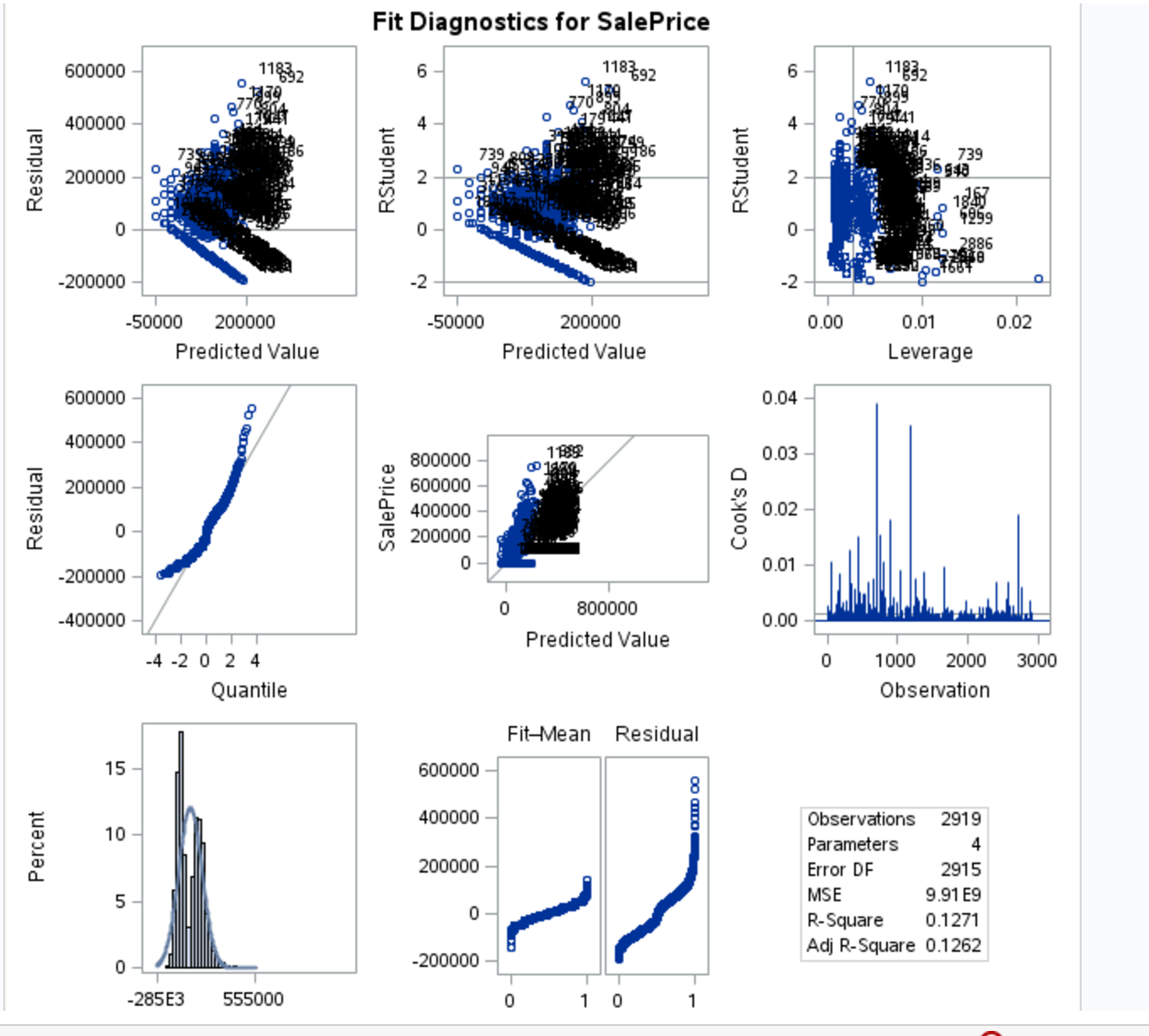


Figure 3 - FORWARD Dream Data Model Assumptions

This model exhibits enough of a Normal distribution with adequate scatter. Evaluation of outliers determines no errors. No transformation was chosen to be used since experimentation with various transformations does not produce measurably improved results.

#### LVB: Dr Xing clearly wants to see transformation. I liked the one async video where Dr McGee automatically transformed the data with log transformation w/o looking at plots. It is the topic we are suppose to discuss in class tonight.

#### Variables in FORWARD Dream Data Model & Analysis

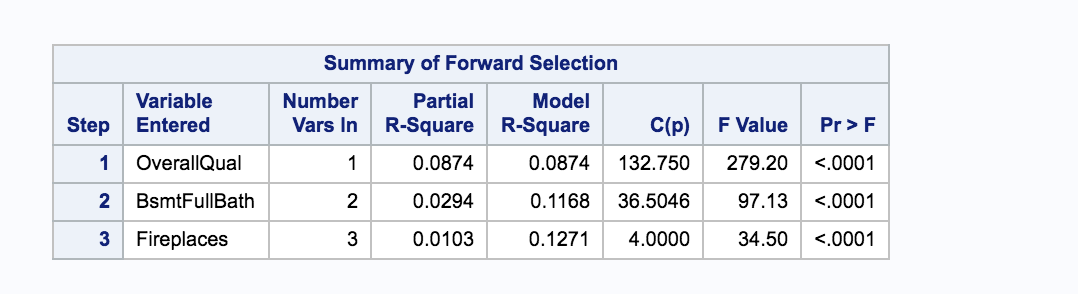


Figure 4 - FORWARD Dream Data Model Selection Result

Based on FORWARD selection using all 80 variables with a SPLIT cross validation using two (2) folds.

*LVB: After doing some additional research, I did not fully understand the mechanism of run the tests on internal and report on external. I just reported. This issue can be corrected in the next “go around”.*

Two folds is used because it was determined that four (4) folds did not yield better results. Resulting variables are OverallQual (p <.0001), BsmtFullBath (p <.0001), Fireplaces (P<.0001).

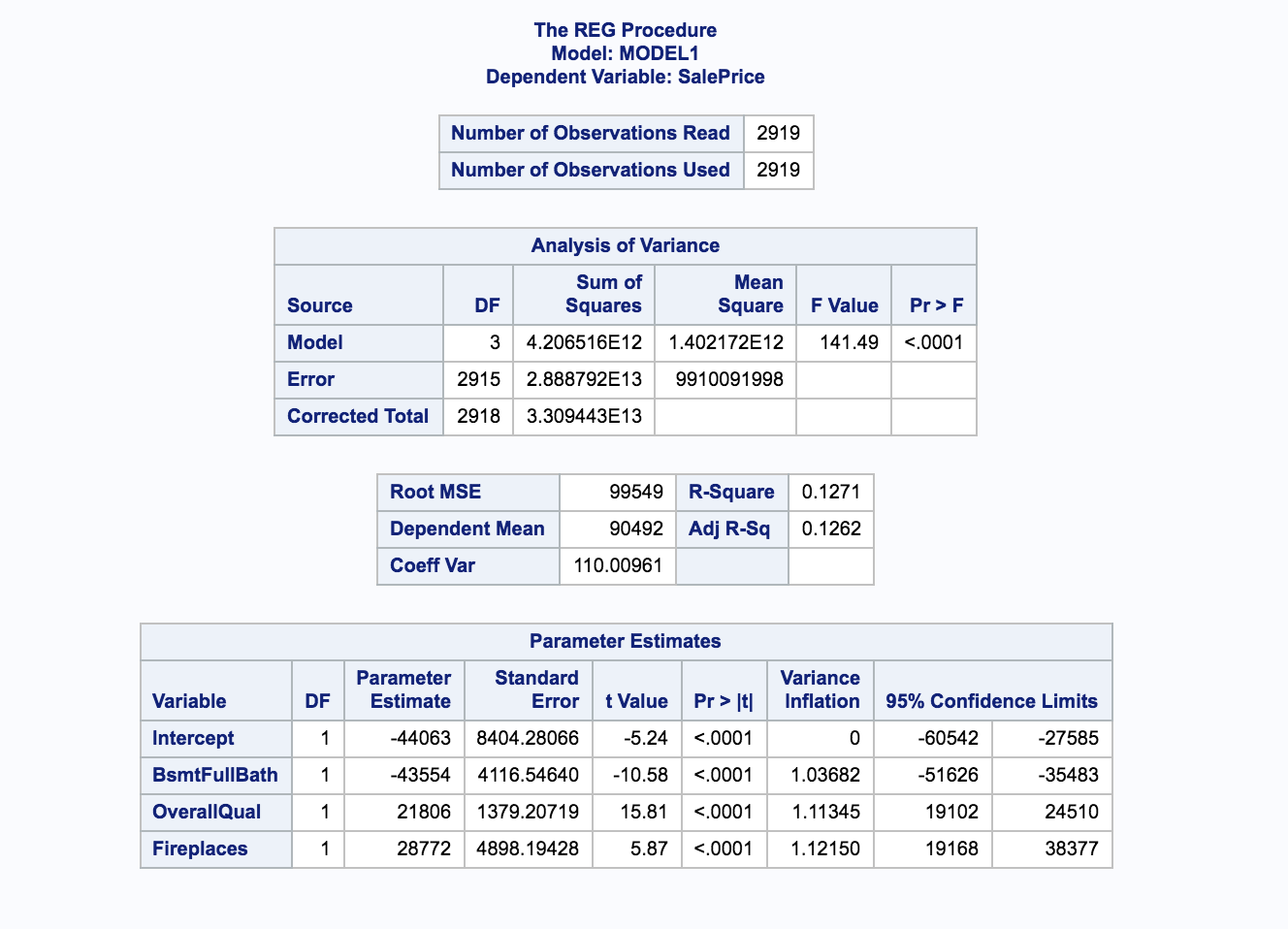


Figure 5 - FORWARD Dream Data Model - Regression Analysis

Sales Price = (-44063) + BsmtFullBath (-43554) + OverallQual (21806) + Fireplaces (28772)

Interpret the slope: If the Sales Price increases by $1, the Model predicts Basement Full Bath (BsmtFullBath) will decrease Sales Price by approximately $43,554 [confidence interval -$51,626, -$35,483]; Overall Quality (OverallQual) will increase Sales Price by approximately $21,806 [confidence interval $19,102, $24,510]; and Fireplaces will increase by approximately $28,772 [confidence interval $19,168, $38,377].

*LVB: This is pretty easy to switch around and can be done on the next go around.*

Interpret the Intercept: If the Sales Price is $0, the Model predicts the value decreases by -$44,063. “The interpretation of the intercept does not make sense in the real world. It is not reasonable to expect” a house having a $0 Sales Price (or value) decrease by -$44,063.[[3]](#footnote-3)

*LVB: As you can tell I struggled with this a bit. Will use her comments to switch it around.*

Adjusted R2 = 0.1262 indicates that 13% of the variability in the Sales Price is explained by this Model. Coefficient of Variation of 110.0 seems high as a unit-less number.

*LVB: I thought coefficient of variation should be small. If 110.0 was in dollars, a $110 variation on a house costing hundreds of thousands is small. By itself, it is rather large.*

### STEPWISE

#### Assumptions Dream Data model

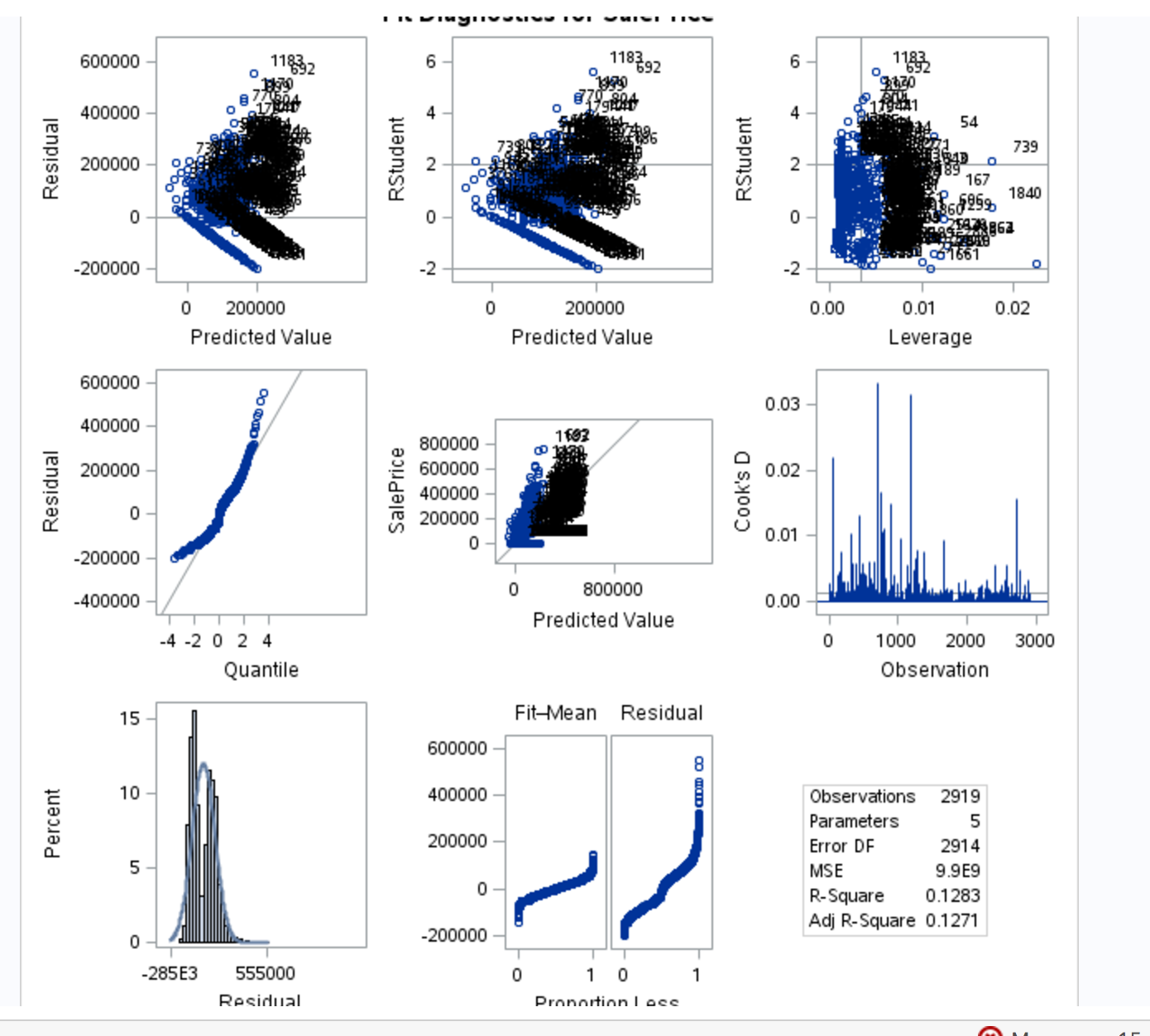


Figure 6 - STEPWISE Dream Data Model Assumptions

This model exhibits enough of a Normal distribution with adequate scatter. Evaluation of outliers determines no errors. No transformation was chosen to be used since experimentation with various transformations does not produce measurably improved results.

*LVB: Let’s see how transformation works this time around. For last class it was not helpful at all. Transformation did not clear up any issues. Makes me wonder what you do in a case like that.*

#### Variables in Selected Model & Analysis

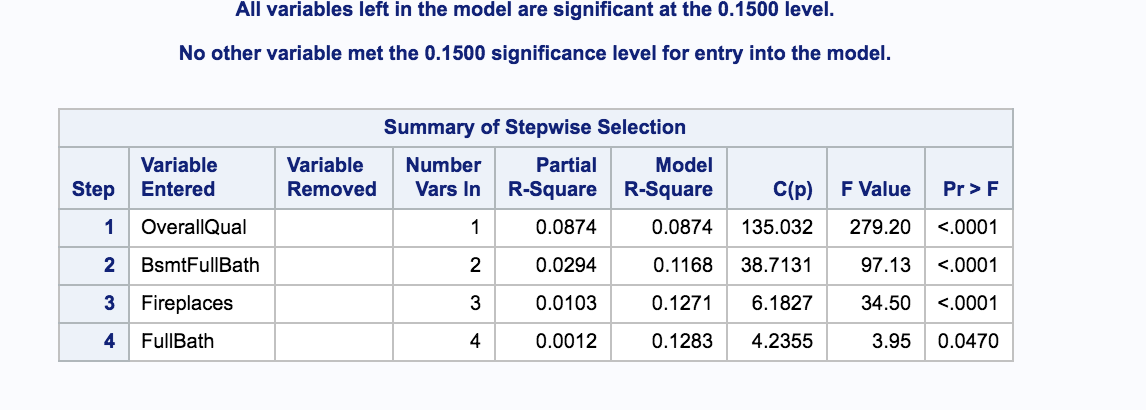


Figure 7 - STEPWISE Dream Data Model

Based on STEPWISE selection using all 80 variables with a SPLIT cross validation using five (5) folds. Made SPLIT change from to two (2) to five (5) after learning in class the typical default is five (5). Resulting variables are OverallQual (p <.0001), BsmtFullBath (p <.0001), Fireplaces (p<.0001) and FullBath (p<.0470).

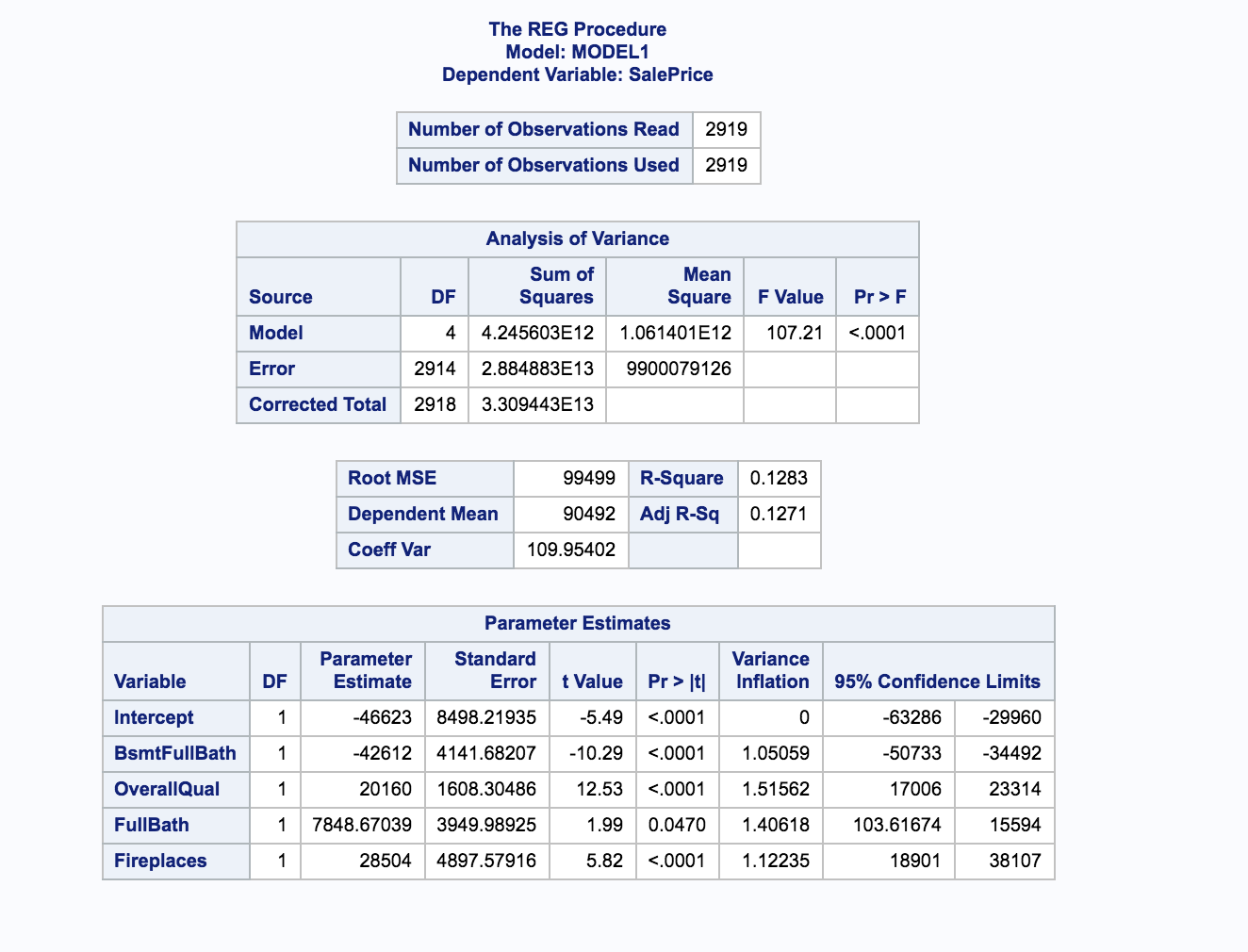


Figure 8 - STEPWISE Dream Data Model Regression Analysis

Sales Price = (-46623) + BsmtFullBath (-42612) + OverallQual (20160) + FullBath 7849 + Fireplaces (28504)

Interpret the slope: If the Sales Price increases by $1, the Model predicts Basement Full Bath (BsmtFullBath) will decrease Sales Price by approximately $42,612 [confidence interval -$63,288, -$29,960]; Overall Quality (OverallQual) will increase Sales Price by approximately $20,160 [confidence interval $17,006, $23,314]; FullBath will increase Sales Price by approximately $7,849 [confidence interval $104, $15594], and Fireplaces will increase by approximately $28,504 [confidence interval $18,901, $38,107].

Interpret the Intercept: If the Sales Price is $0, the Model predicts the value decreases by -$46,623. “The interpretation of the slope does not make sense in the real world. It is not reasonable to expect” a house having a $0 Sales Price (or value) decrease by $46,623.2

Adjusted R2 = 0.1271 indicates that 13% of the variability in the Sales Price is explained by this Model. Coefficient of Variation of 110.0 seems high as a unit-less number.

## Results

The best model was determined by the question being asked. Best Kaggle scores (Model 3- STEPWISE Dream Data) is not the best CV Press (Model 1 – FORWARD Dream Data and Model 4 – HUMAN INFERENCE & SELECTION Dream Data) or Adjusted R2 (Model 1 – FORWARD Dream Data, Model 3 – STEPWISE Dream Data, Model 4 – HUMAN INTERFERENCE & SELECTION Dream Data). What was not expected was the low Adjusted R2 across the board with little impact by variation in the model fitting. All in all, none of these are great models that explain the Sales Price.

For AIC and BIC, the lowest number was found in the FORWARD Dream Data Model [AIC 68769, BIC 78838], but is not much lower than the ‘next runner up’ of STEPWISE Dream Data Model and HUMAN INFERENCE & SELECTION [AIC 68774, BIC 78930].

The results are as follows below:

Table 1 - Results Dream Data and Data Mine Models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Models** | **Adjusted R2** | **AIC** | **Root MSE** | **BIC or SBC[[4]](#footnote-4)** | **External CVPress** | **Cross Valid Details** | **Kaggle**  **Score** |
| Model 1-FORWARD Data Mine | 0.44 | 68841 | 79954 | 66004 | 1.89 | Fold 2. Stop based on SBC.  Method:Split | 0.905 |
| Model 1 – FORWARD Dream Data | 0.45 | 68769 | 78838 | 65992 | 1.84 | Fold 2. Stop based on SBC.  Method:Split | 0.877 |
| Model 2 – LASSO Data Mine | 0.38 | 69129 | 84098 | 66250 | 1.96 | Fold 5. Stop based on SBC.  Method:Random | 0.905 |
| Model 2 – LASSO Dream Data | 0.40 | 69048 | 82909 | 66181 | 1.86 | Fold 5. Stop based on SBC.  Method:Random | 0.909 |
| Model 3 – STEPWISE Data Mine | 0.44 | 68841 | 79954 | 66004 | 1.89 | Fold 5. Stop based on SBC.  Method:Split | .865 |
| Model 3 – STEPWISE Dream Data | 0.45 | 68774 | 78930 | 65985 | 1.86 | Fold 5. Stop based on SBC.  Method:Split | .941 |
| Model 4 – HUMAN INFERENCE & SELECTION Dream Data | 0.45 | 68774 | 78930 | 65985 | 1.84 | Fold 2. Stop based on SBC.  Method:Split | .912 |

# Conclusion

Clearly the Data Mine model does not generate an improvement in the explanation of the linear regression relationship and the variables that explain Sales Price. In each case, the Data Mine Model scored lower than the Dream Data model. See Table 1. The experiment did prove that data sets with too few variables can yield poorer results than those with more variables. Where is equilibrium? Not sure at this point, but recognize that the data mining effort was clearly not worth the effort.

*LVB: Will be my next approach to show more of this background effort.*

Overall, the controversial STEPWISE results of variable explanation were the best, although not great given its low adjusted R2 score and not much better than FORWARD selection model.

Out of curiosity, and observing the low linear relationship scoring, the Model 4 – HUMAN INFERENCE & SELECTION Dream Data was added at the end to see if a human could pick better variables that could predict the Sales Price. While the human approach did not do better, it proved to be just as good as Model 3 – STEPWISE Dream Data. This demonstrates the point that the model and the tests need to make sense; and that human intuition is useful.

Is there such a thing as data fatigue? This analysis seems to be at the saturation point.

# Appendix 1

## Data Mine Model

Beyond the straight model fitting, a data mining and manual/intuition approach is used to determine if fewer attributes of a house sold would provide greater insight into the predicted sales price. To control the complexity, the variables are grouped into 8 groups as follows:

1. Lot
2. Quality
3. Basement
4. House
5. Garage
6. Deck & Porch
7. Pool & Miscellaneous
8. Year Sold

The statistically most significant variable from each group, based on Type I and III sum of squares, is used in the Data Mined dataset, which includes:

* LotArea [ pvalue <.0001, VIF 1.1 ]
* OverallQual [ pvalue < .0001 , VIF1.74 ]
* BsmtUnfSF [ pvalue < .0001, VIF 3.8 ]
* TotalBsmtSF [ pvalue < .0001, VIF 4.6 ]
* BsmtFullBath [ pvalue < .0001, VIF1.65 ]
* FullBath [ pvalue .02 VIF 1.78 ]
* KitchenAbvGrd [ pvalue .006, VIF 1.2 ]
* GarageArea [ pvalue <.0001, VIF 3.5 ]
* WoodDeckSF [ pvalue <.0001, VIF 1.0 ]
* OpenPorchSF [ pvalue < .0001, VIF 1.0 ]
* X3SsnPorch [ pvalue .004, VIF 1.0 ]
* ScreenPorch [ pvalue .005, VIF 1.0]
* PoolArea [ pvalue <.0001, VIF 1.0 ]

Sold variables, as in WhenSold, MoSold, YrSold, turn out not to be significant.

Through the model fitting process, a number of the Data Mine variables are not significant and do not make it into the final Data Mine models. The selection process chosen drives the results of which variables are used as Data Mine models.

While model’s assumptions look good for QQPlot, Normal enough distribution, random scatter and Variance Inflation Factor (VIF) within a good range (1.08 to 1.95), the adjusted R2 is 0.17.

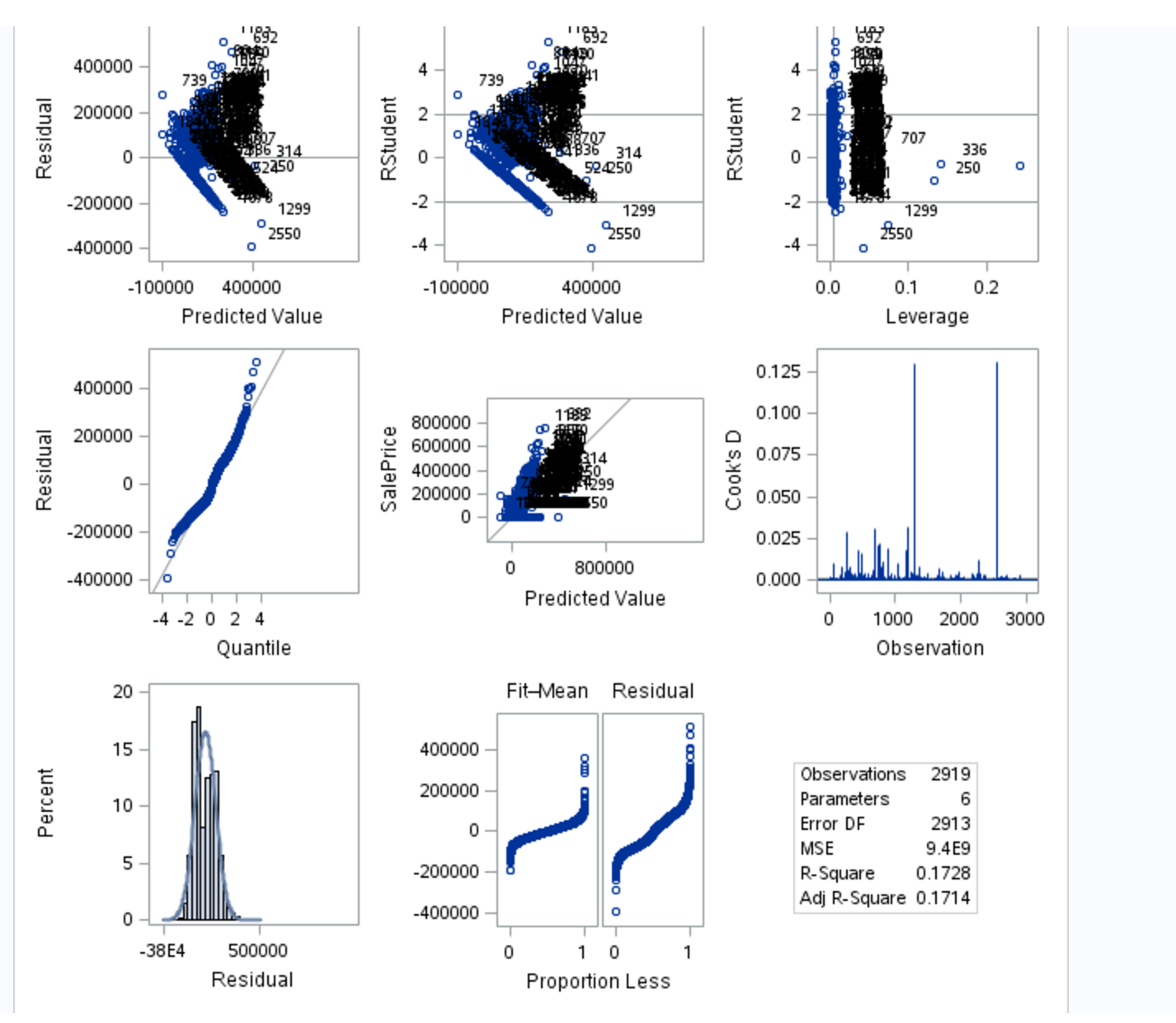
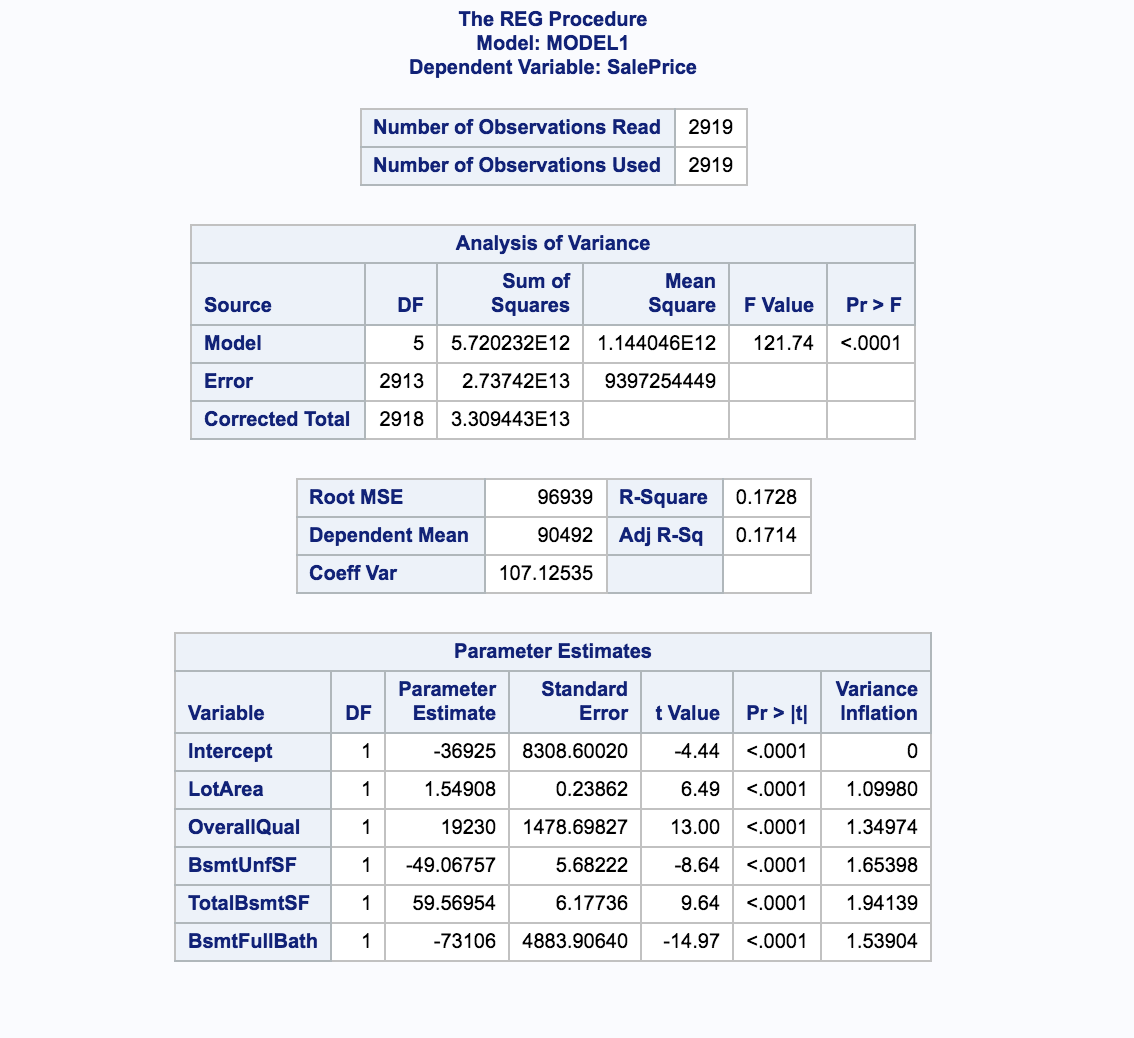


Figure 9 – Output from Proc Reg with plots

After removing outliers 1299 and 2550, the adjusted R2 jumps to .99 using *Proc GLM*, demonstrating the significant impact of outliers 1299 and 2550 for the Data Mine model.

*LVB: No. I never went further because Kaggle requires you have all observations. This is what is a bit frustrating for me about these exercises. Kaggle and SAS have different requirements than the instructors.*



Figure 10 - Proc GLM results after outliers removed

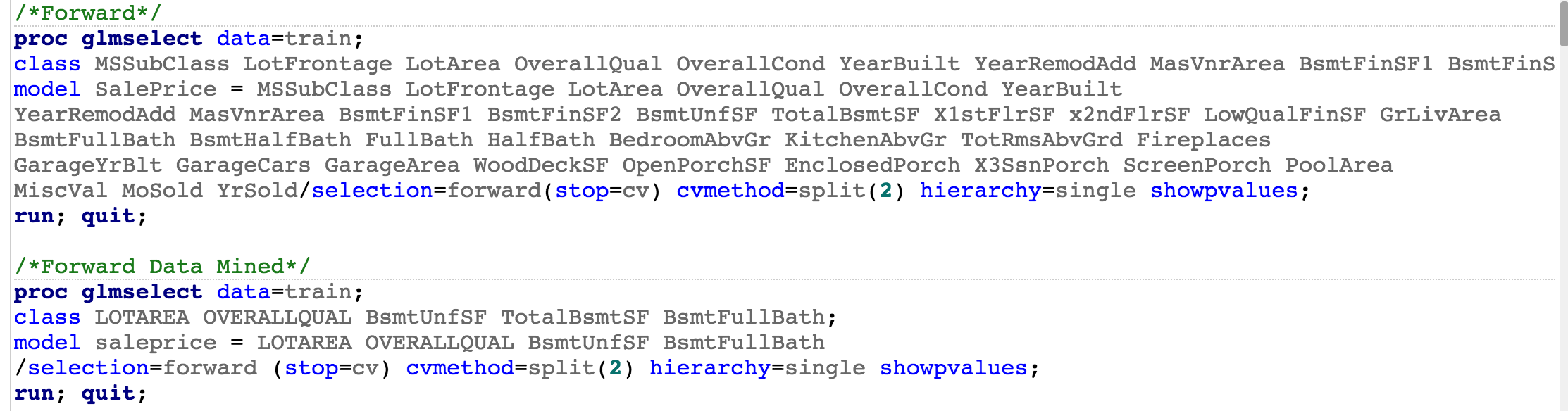
The observations do not appear to have errors, and Kaggle process does not allow removal of outlier data. Therefore, no outliers are removed.

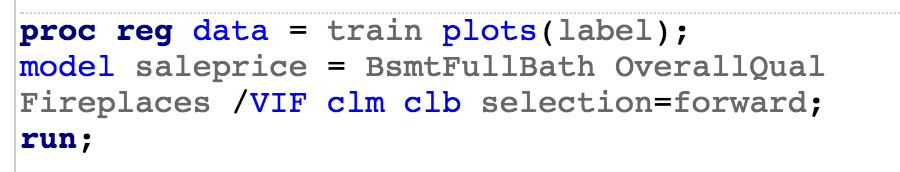
Based on the strong linear relationship of the Data Mine model, it was decided to run the model selection criteria on both the Data Mine data and the Dream Data.

## CODE

### Model 1 - Forward

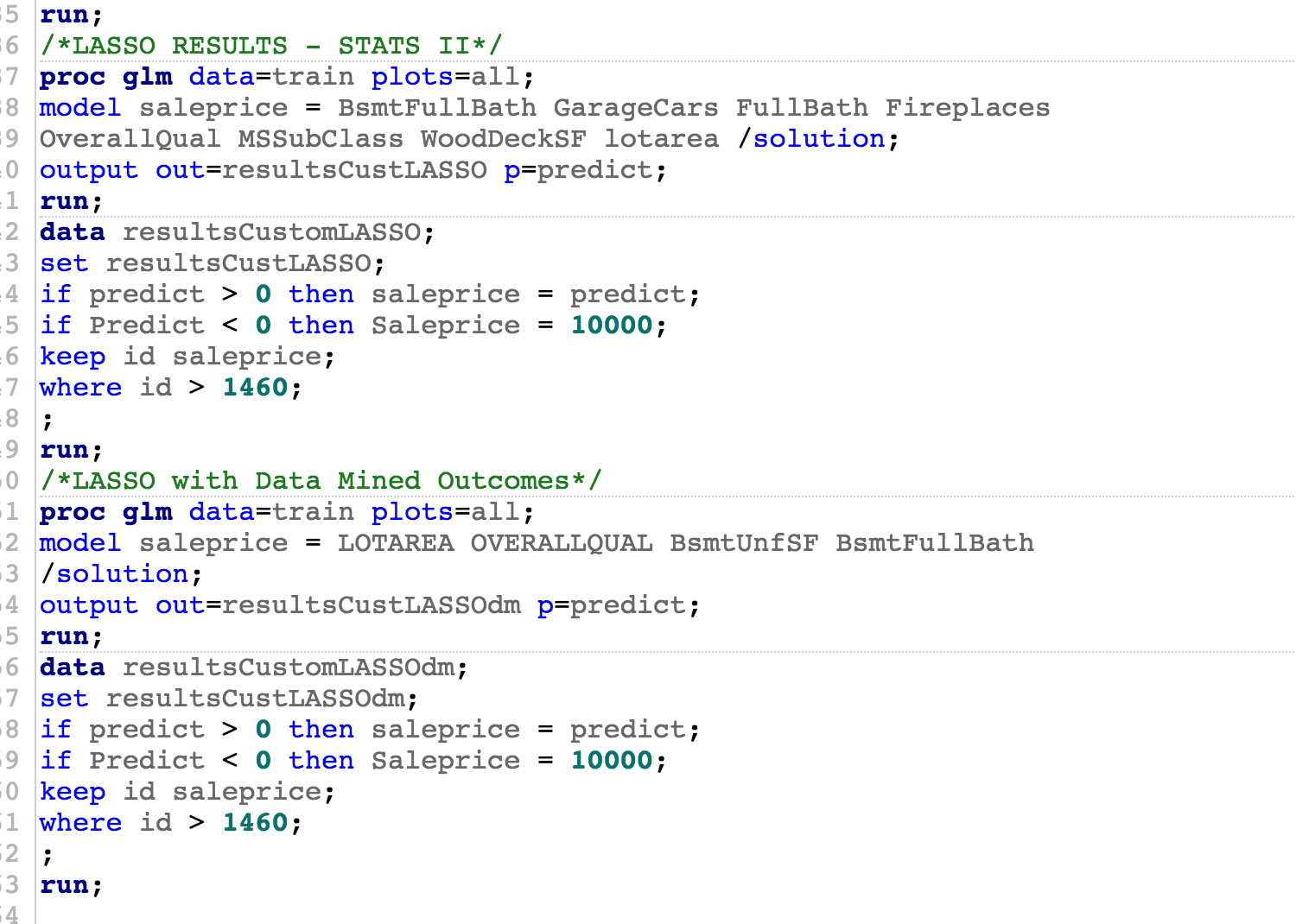


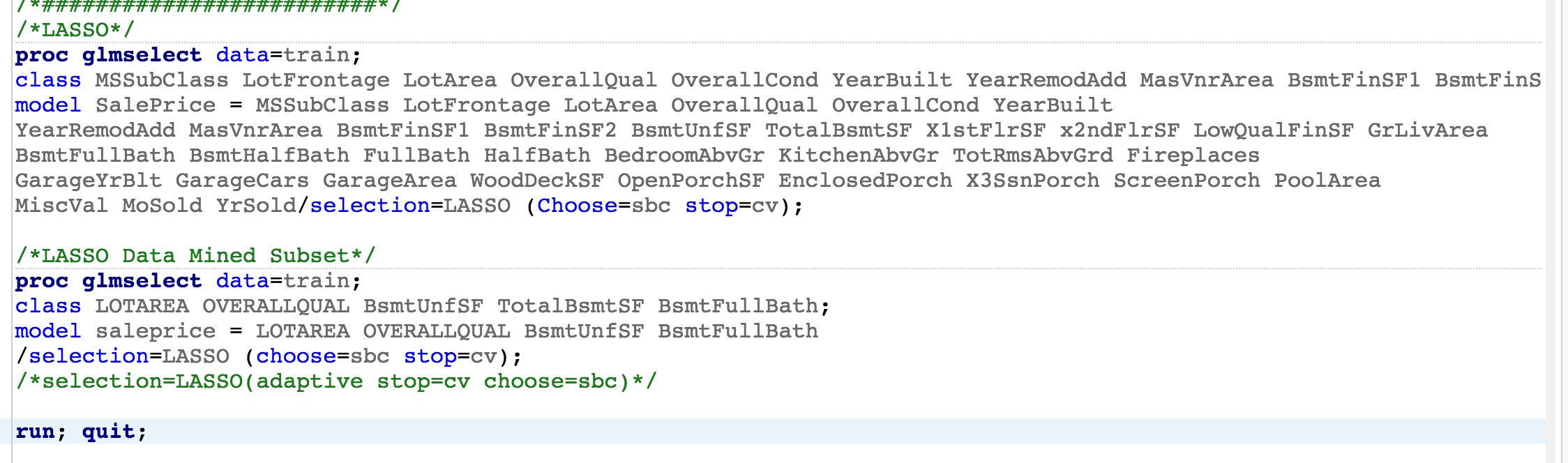




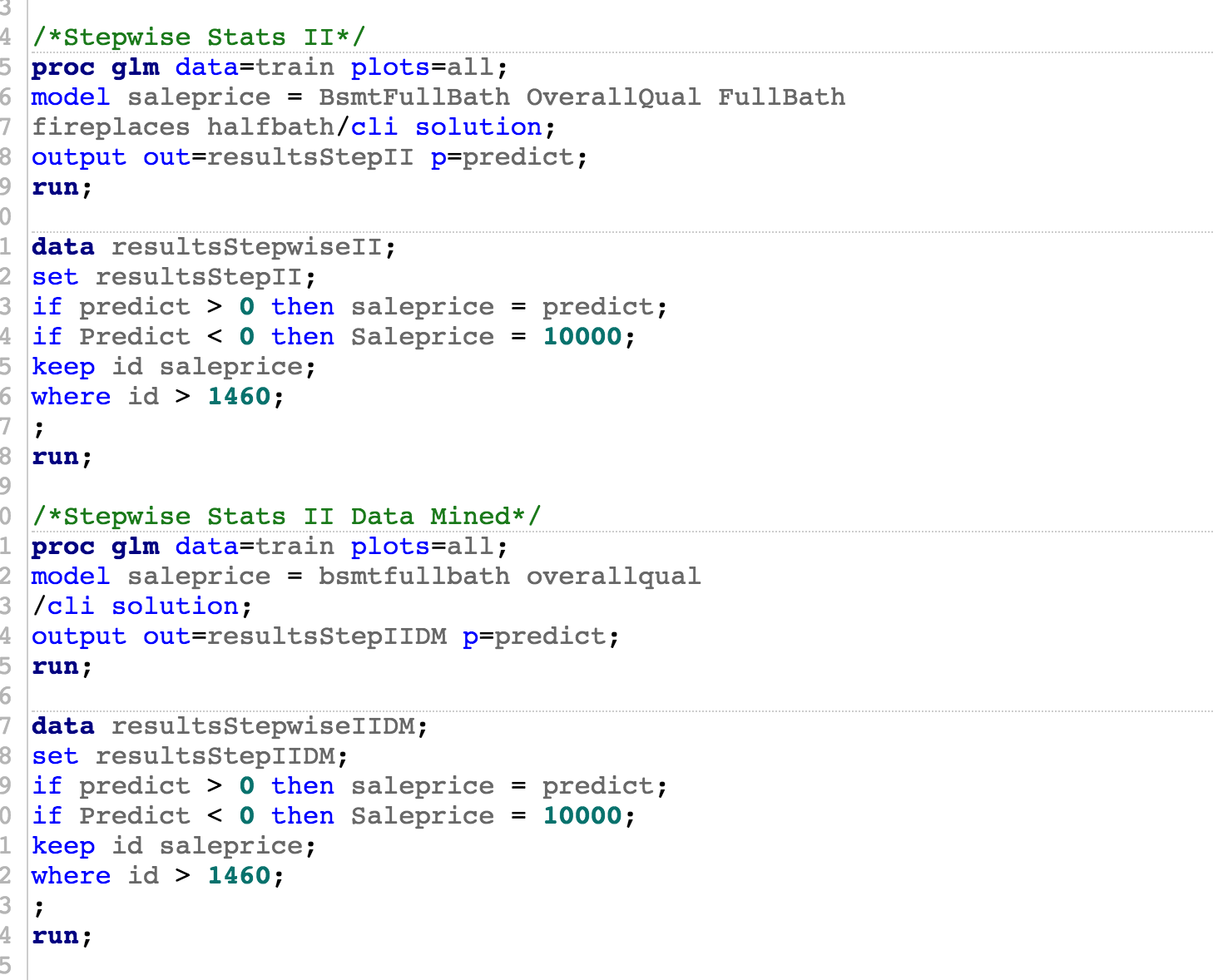


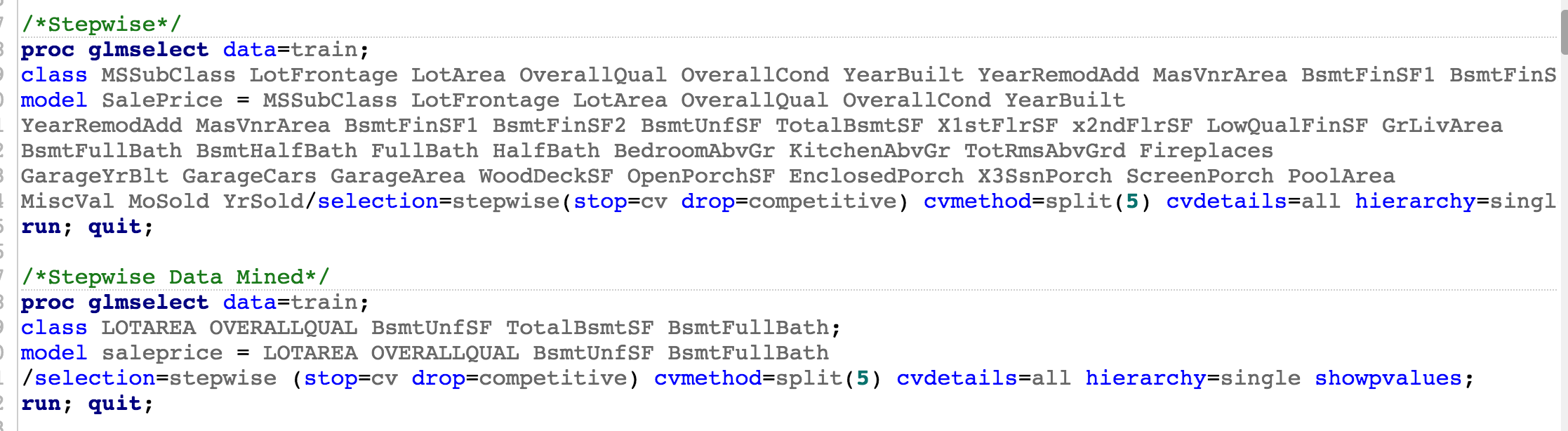
### Model 2 - LASSO

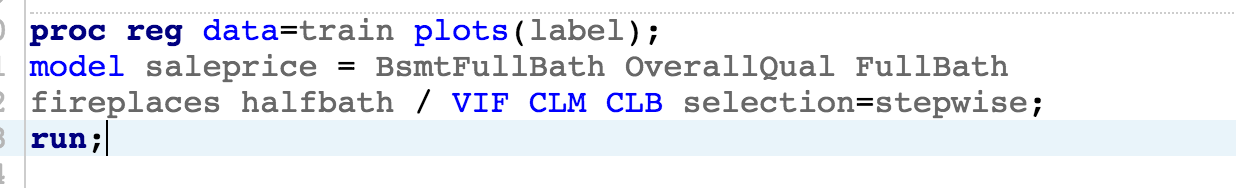


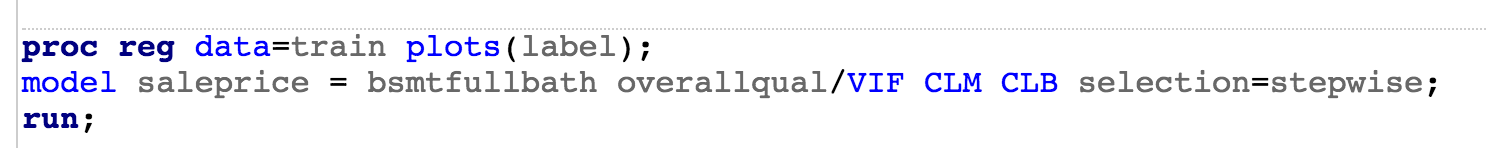


### Model 3 - Stepwise

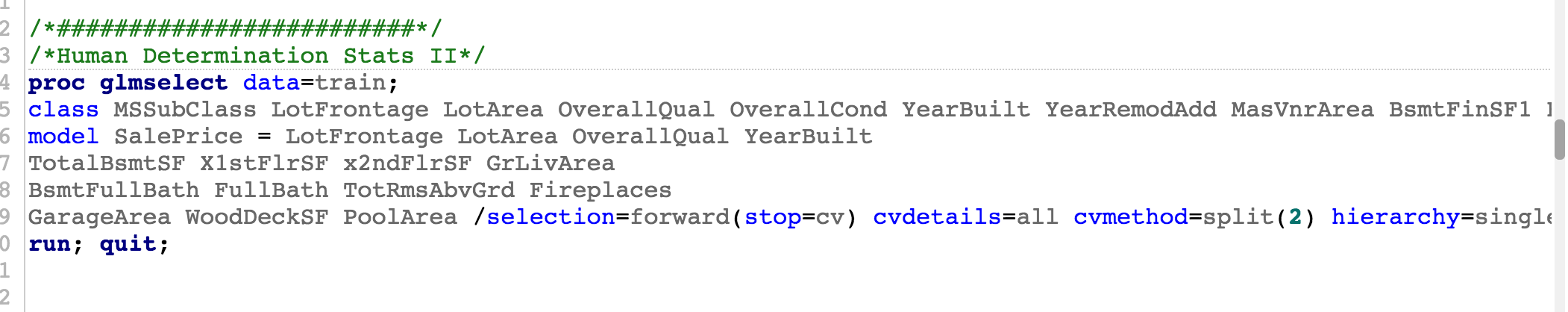








### Model 4 - Human



1. Estimation Methods for Replacing Missing Values, <https://www.ibm.com/support/knowledgecenter/en/SSLVMB_20.0.0/com.ibm.spss.statistics.help/replace_missing_values_estimation_methods.htm> [↑](#footnote-ref-1)
2. Robert A. Cohen, “Introducing the GLMSELECT PROCEDURE for Model Selection”, SAS Institute Paper 207-31, no date, page 17. [↑](#footnote-ref-2)
3. “Interpret the Slope.PDF” - http://www.austincc.edu/mparker/1342/lessons/less5-8/interpret\_slope.pdf. [↑](#footnote-ref-3)
4. Bayesian Information Criteria (BIC) = to Schwartz criterion (SBC) based on https://en.wikipedia.org/wiki/Bayesian\_information\_criterion [↑](#footnote-ref-4)