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’, some type of imputation must be done. Missing data imputation is complicated. The previous approach, using the mean value to fill in missing data, is not good for multivariate analysis since it hides correlations. 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

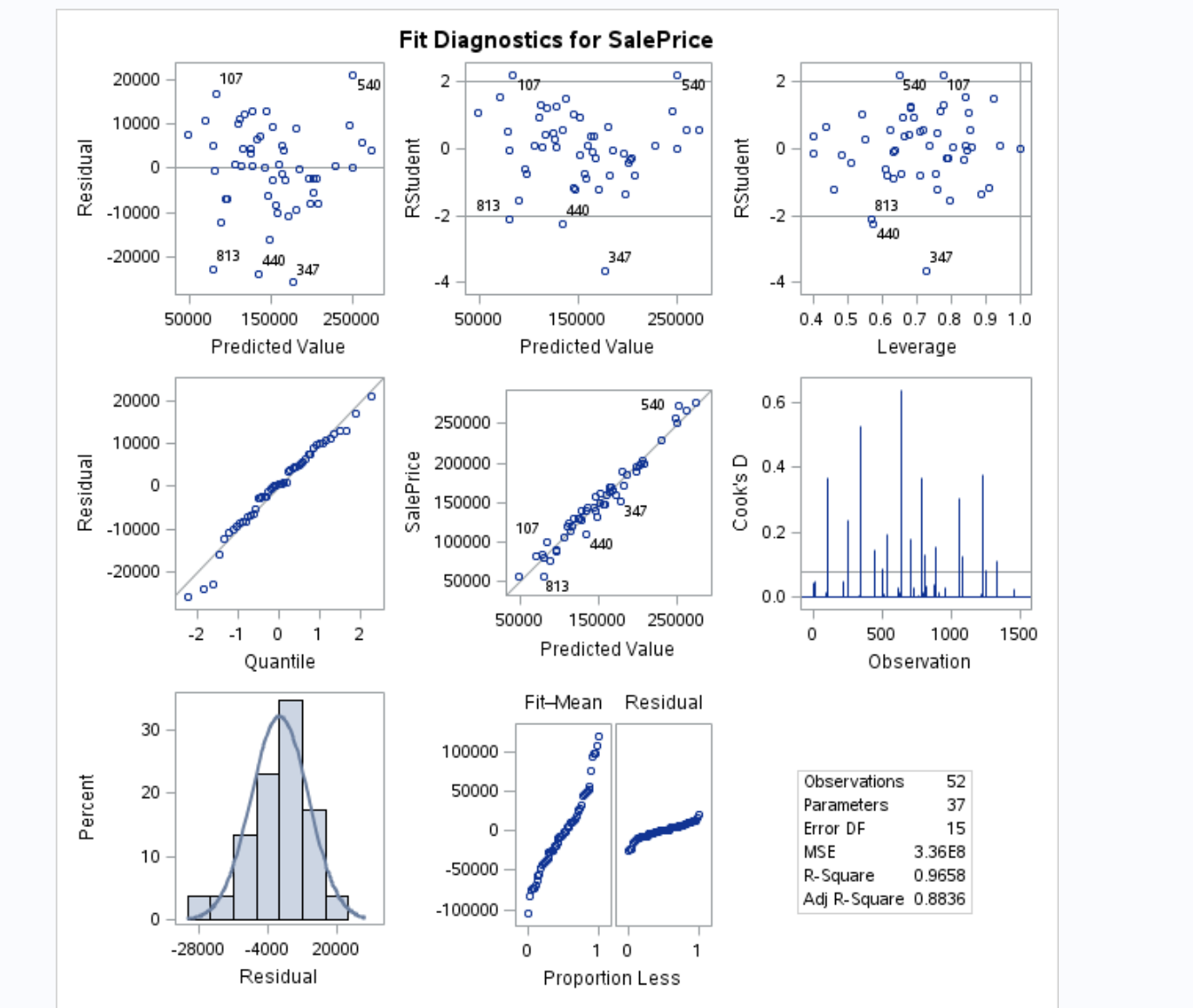
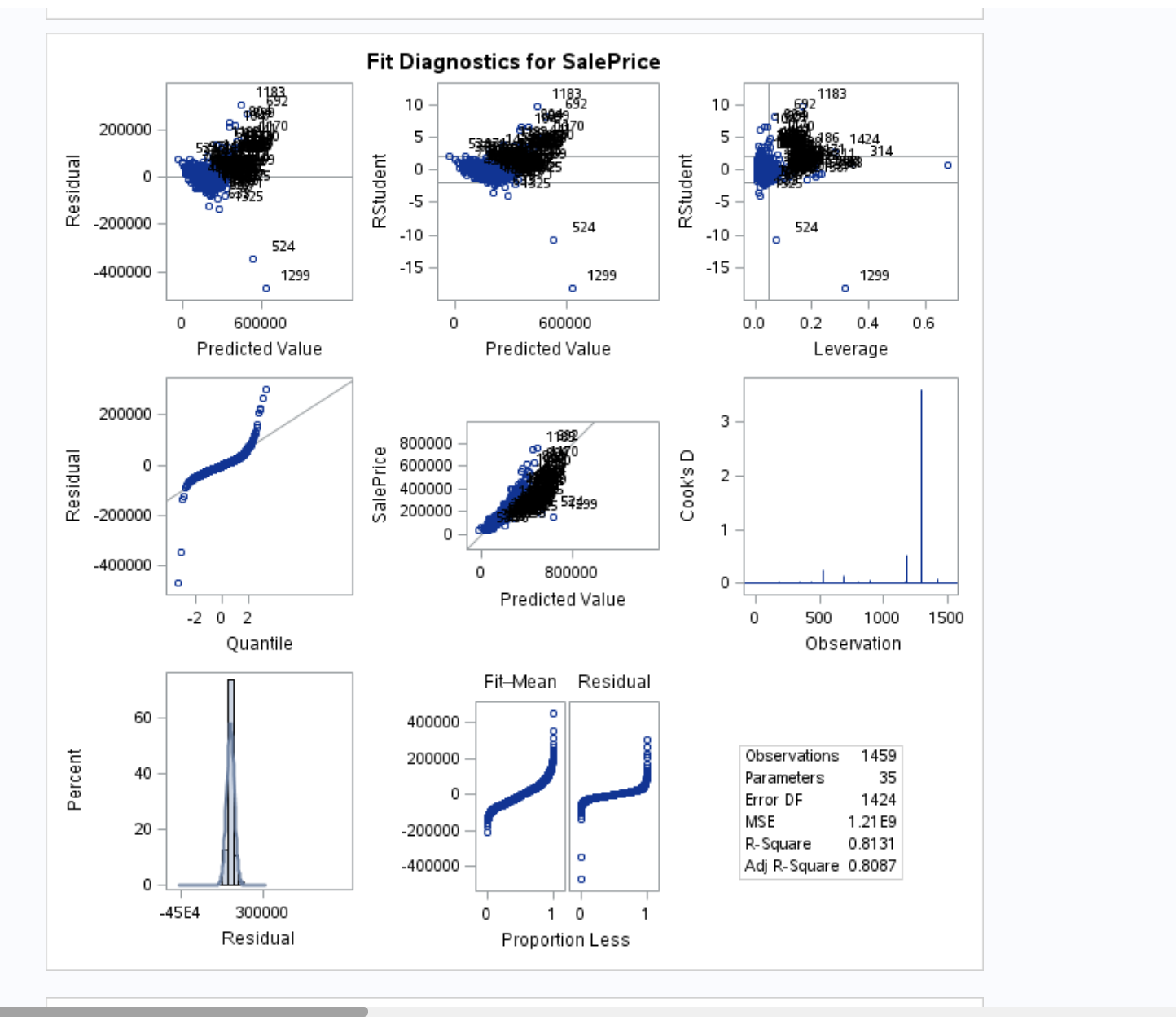
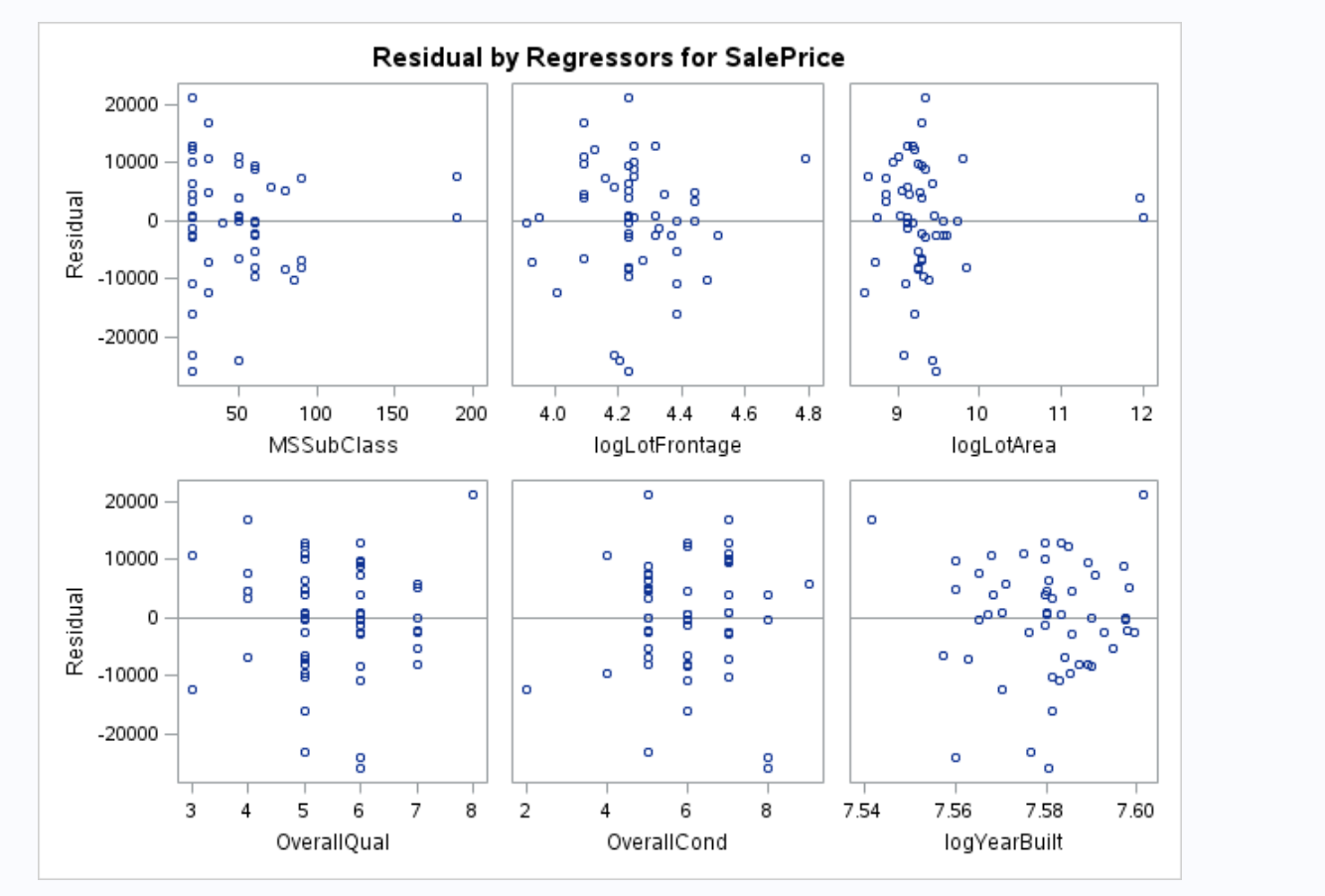
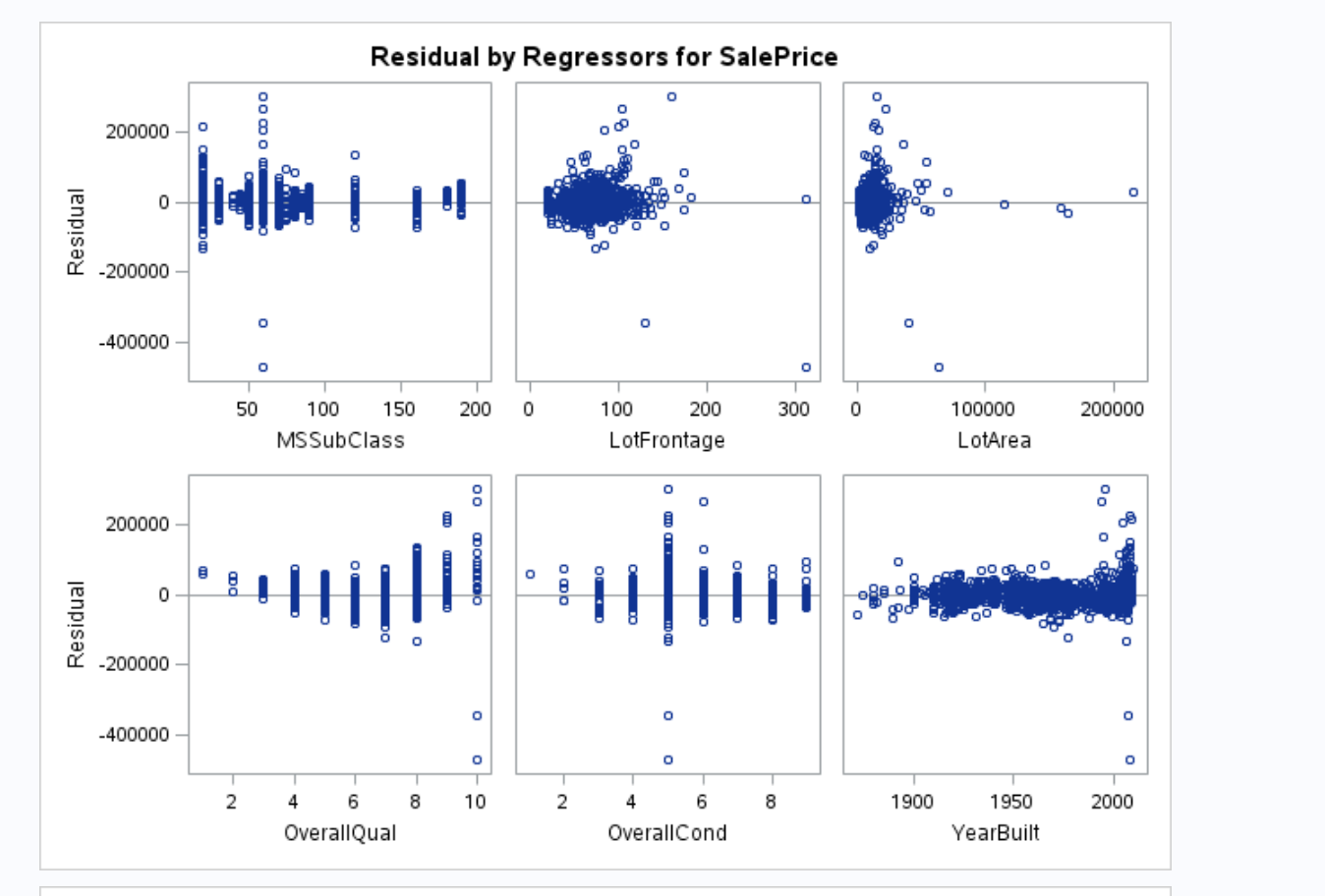
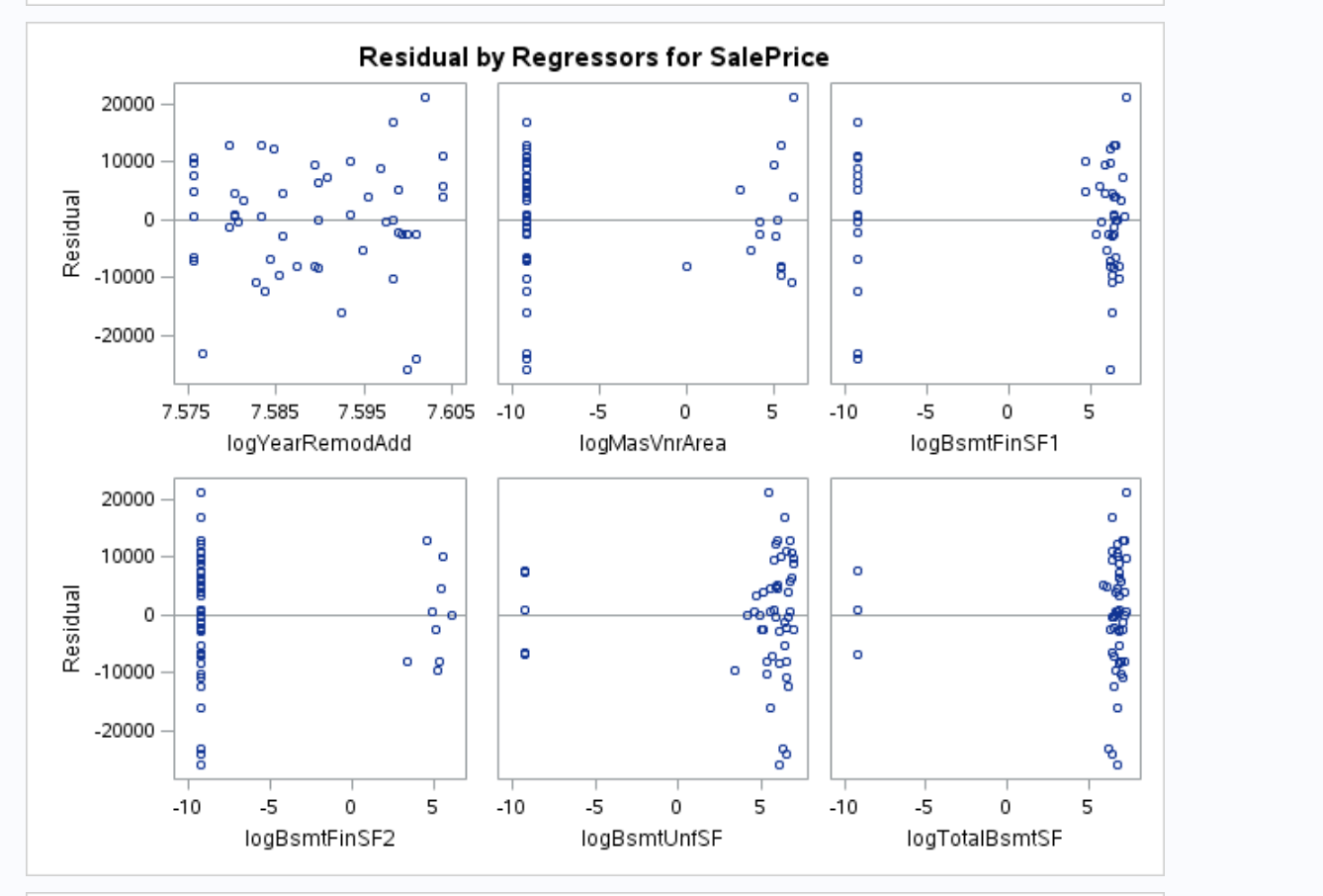
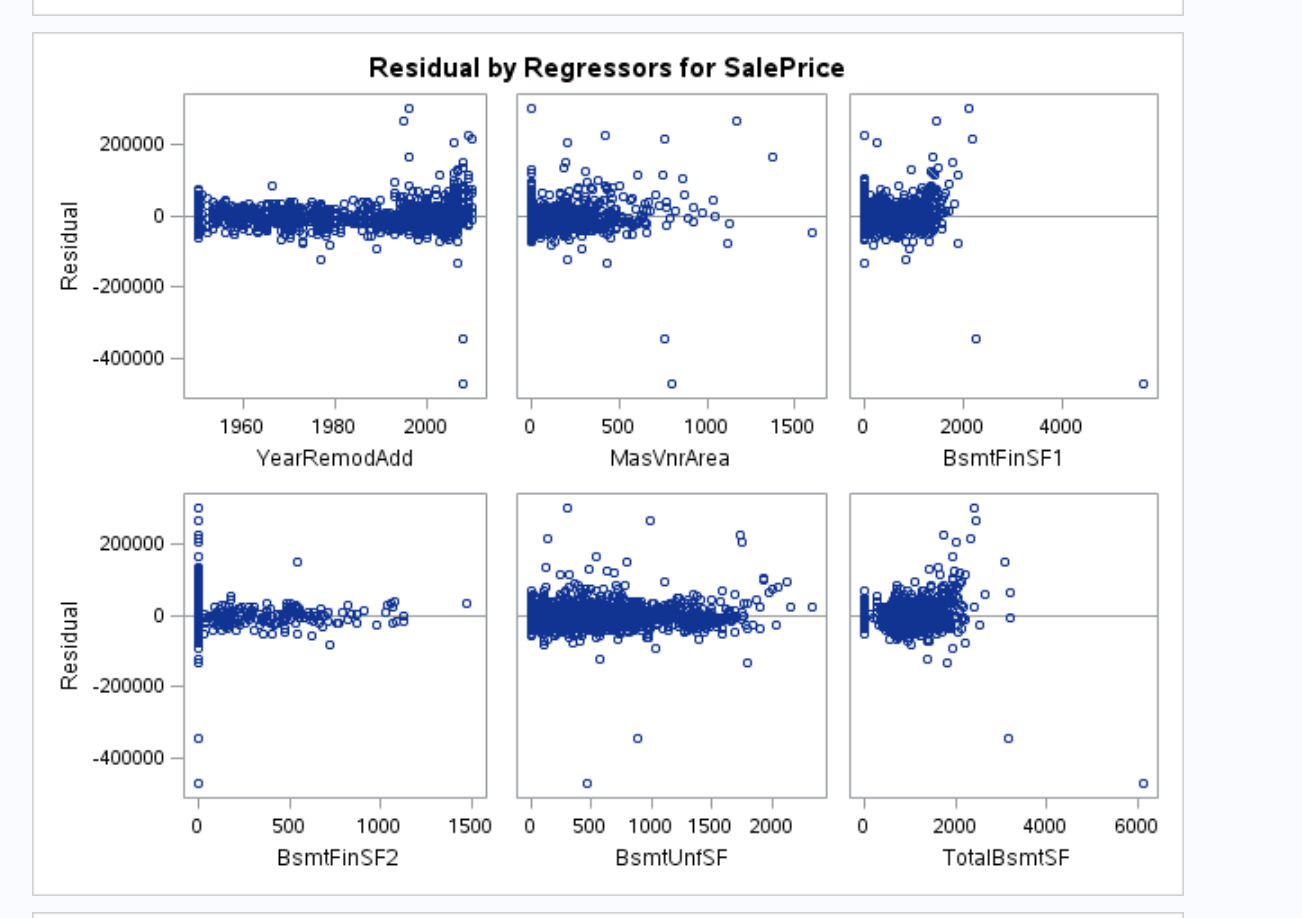
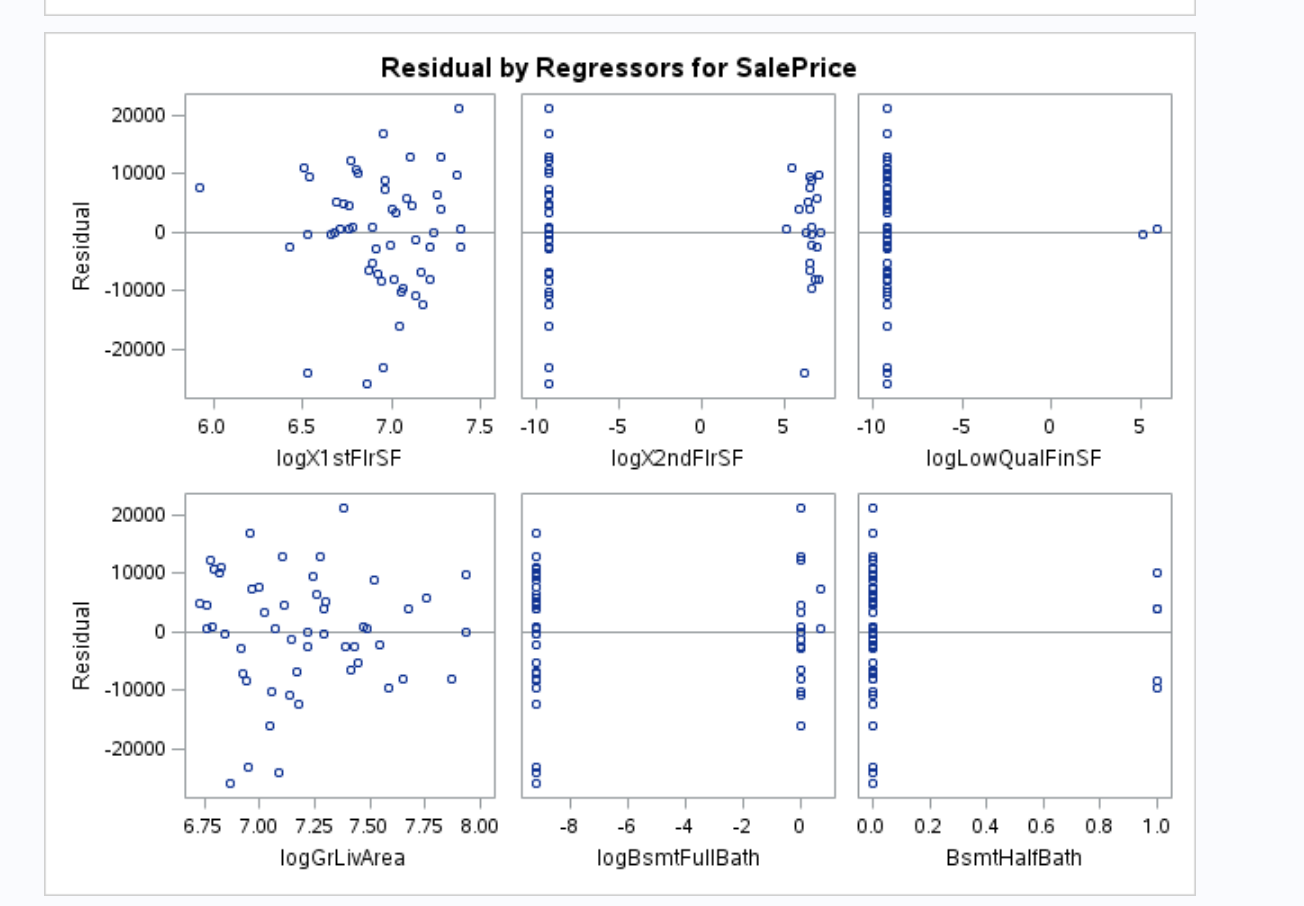
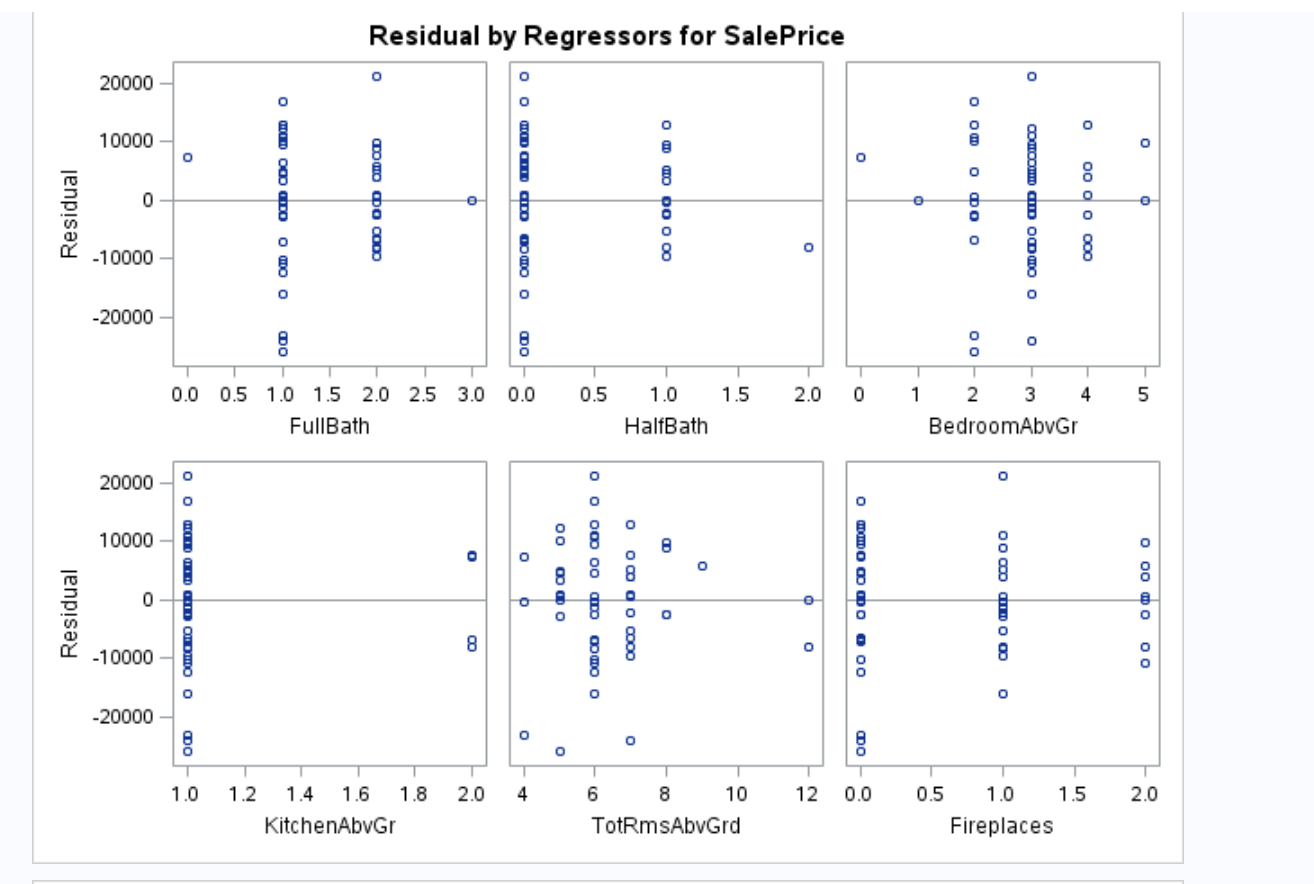
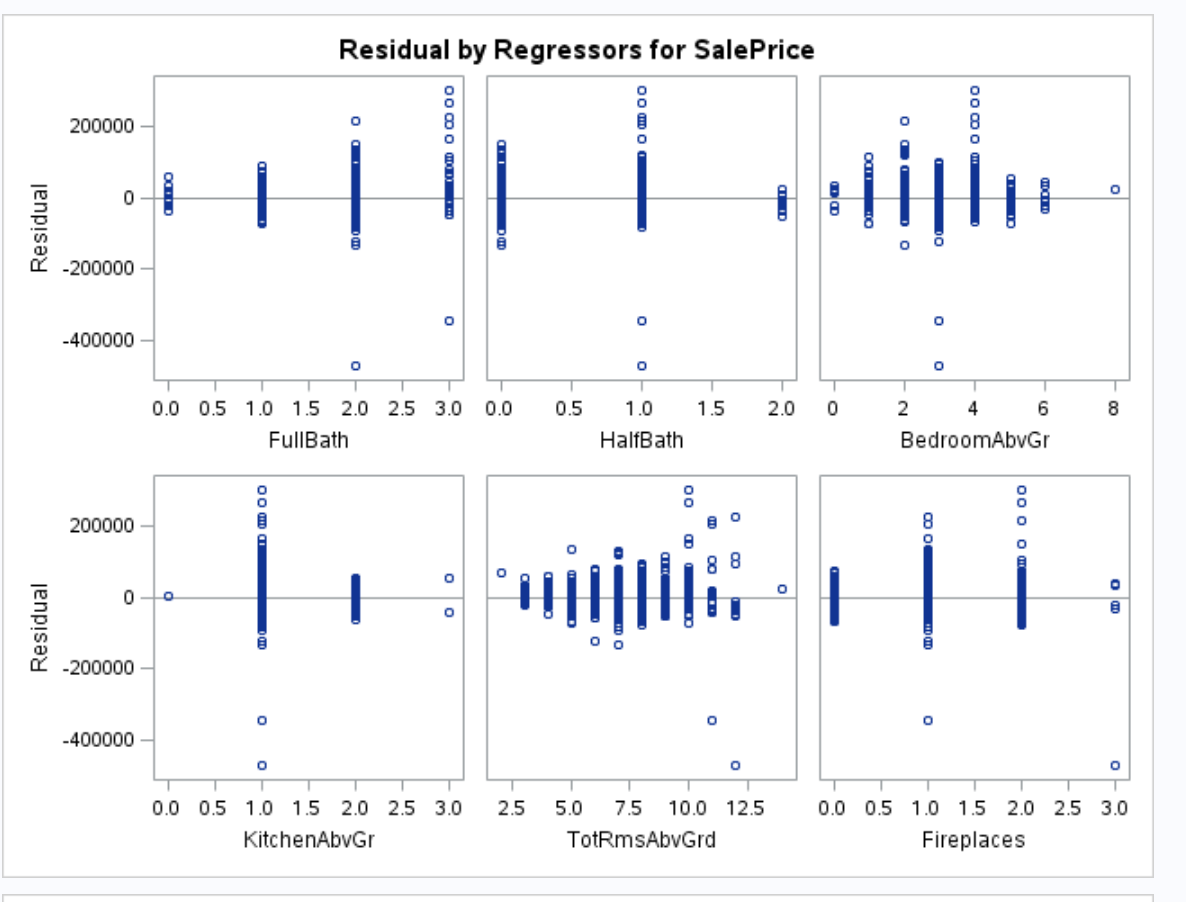


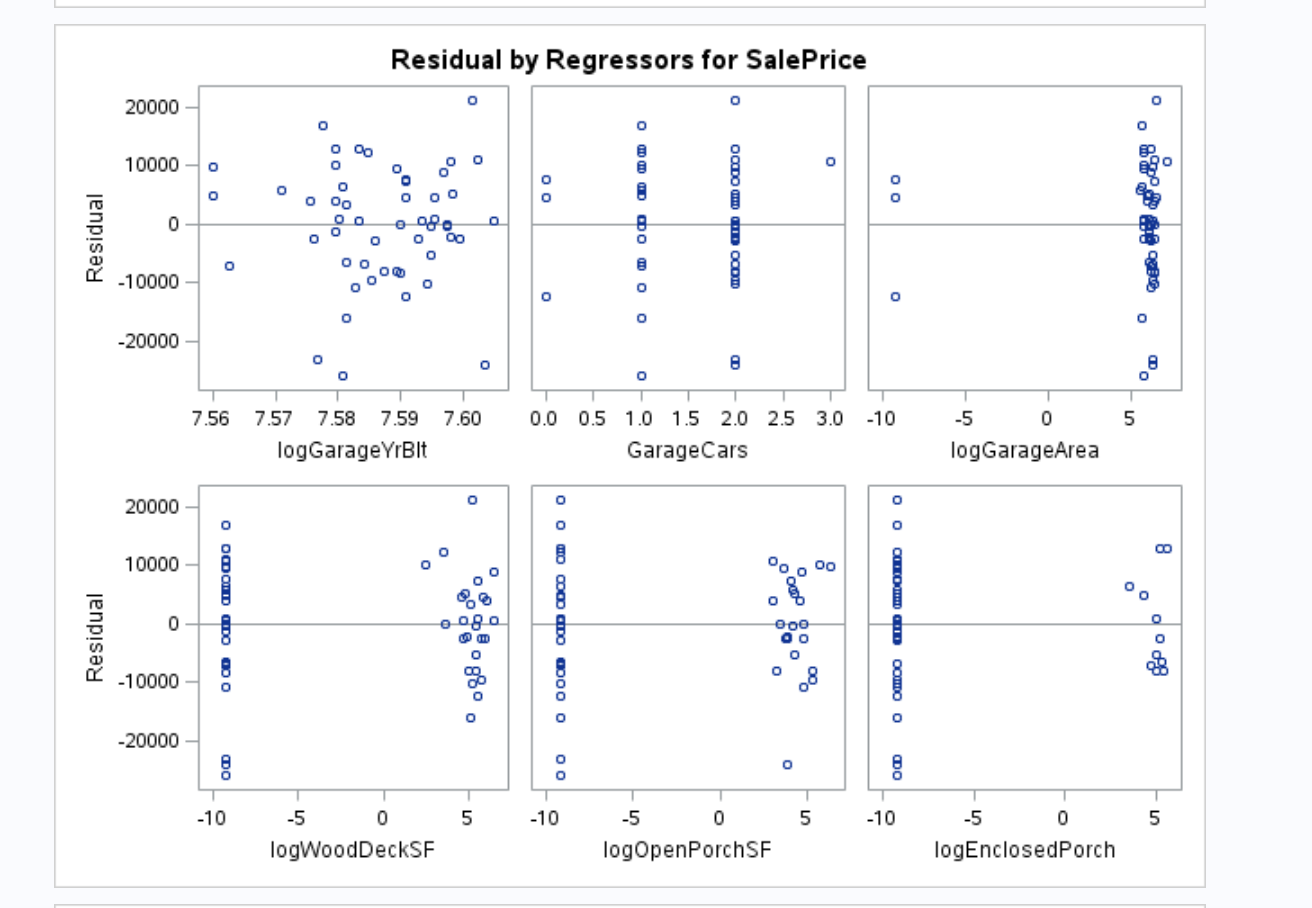
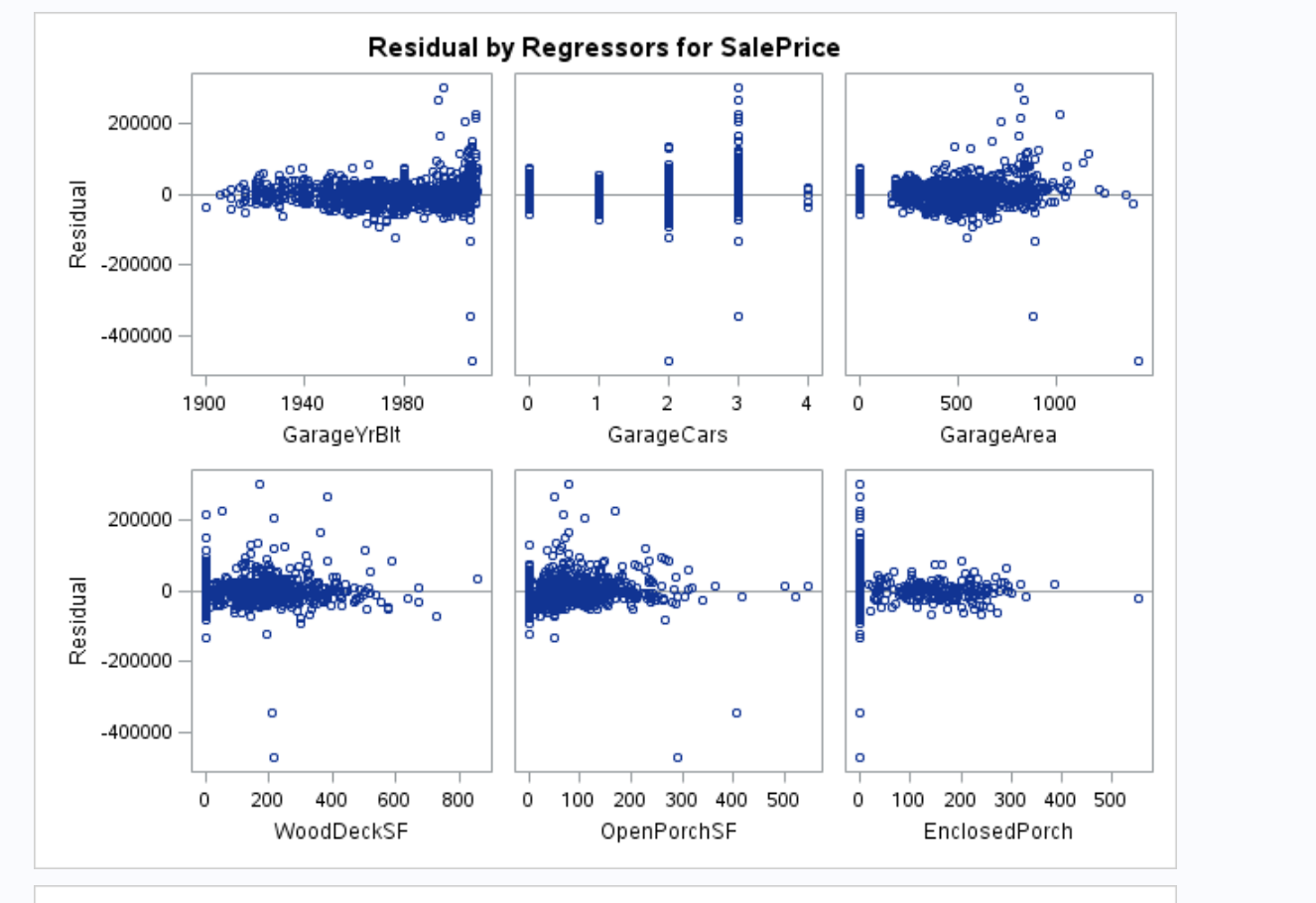
Figure 2 – Residual by Regressor - Before Log on left and After Log on right

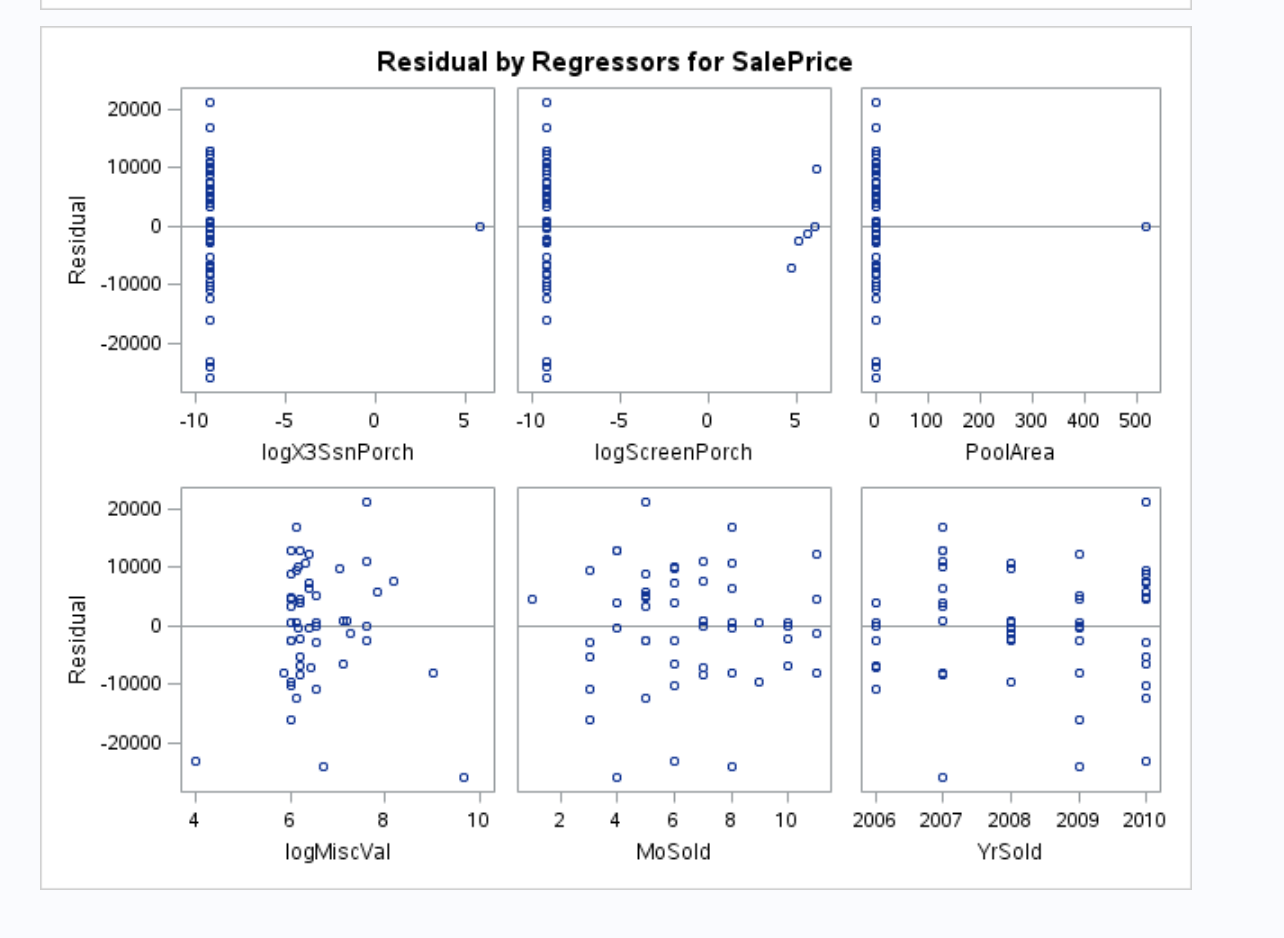
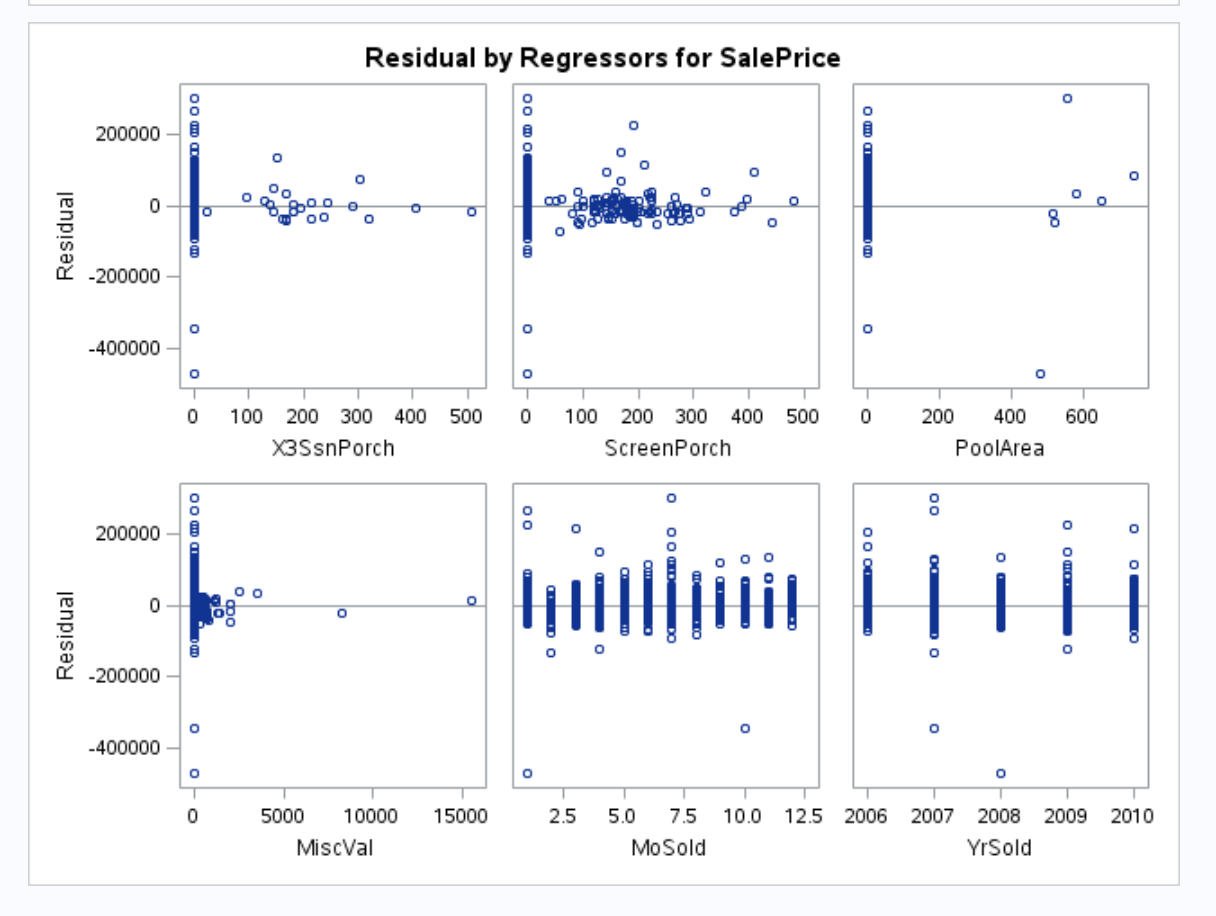












# Model Selection (updated)

The goal is to create the most predictive model possible for Sales Price from the data that was also “useful.” The approaches used were FORWARD, LASSO and BACKWARD selection methods using all of the variables. The partition of the data is 50%. The best model thus far, using *PROC GLM,* is LASSO, as shown below.

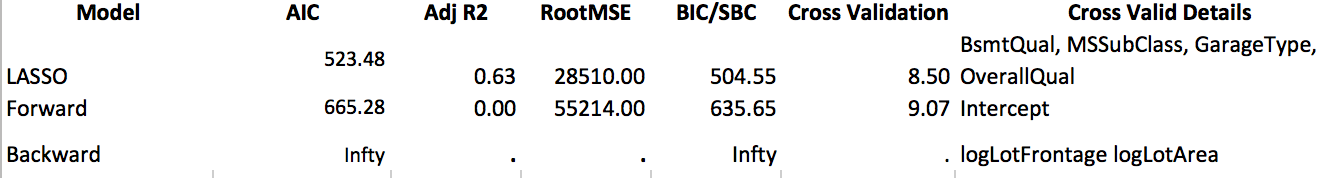
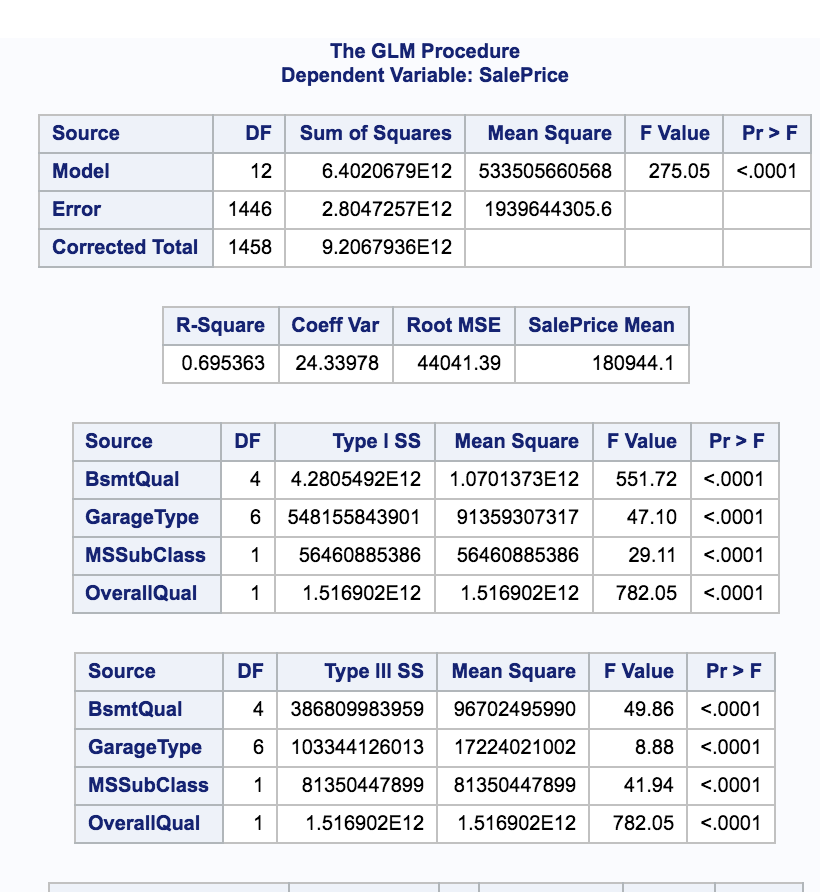


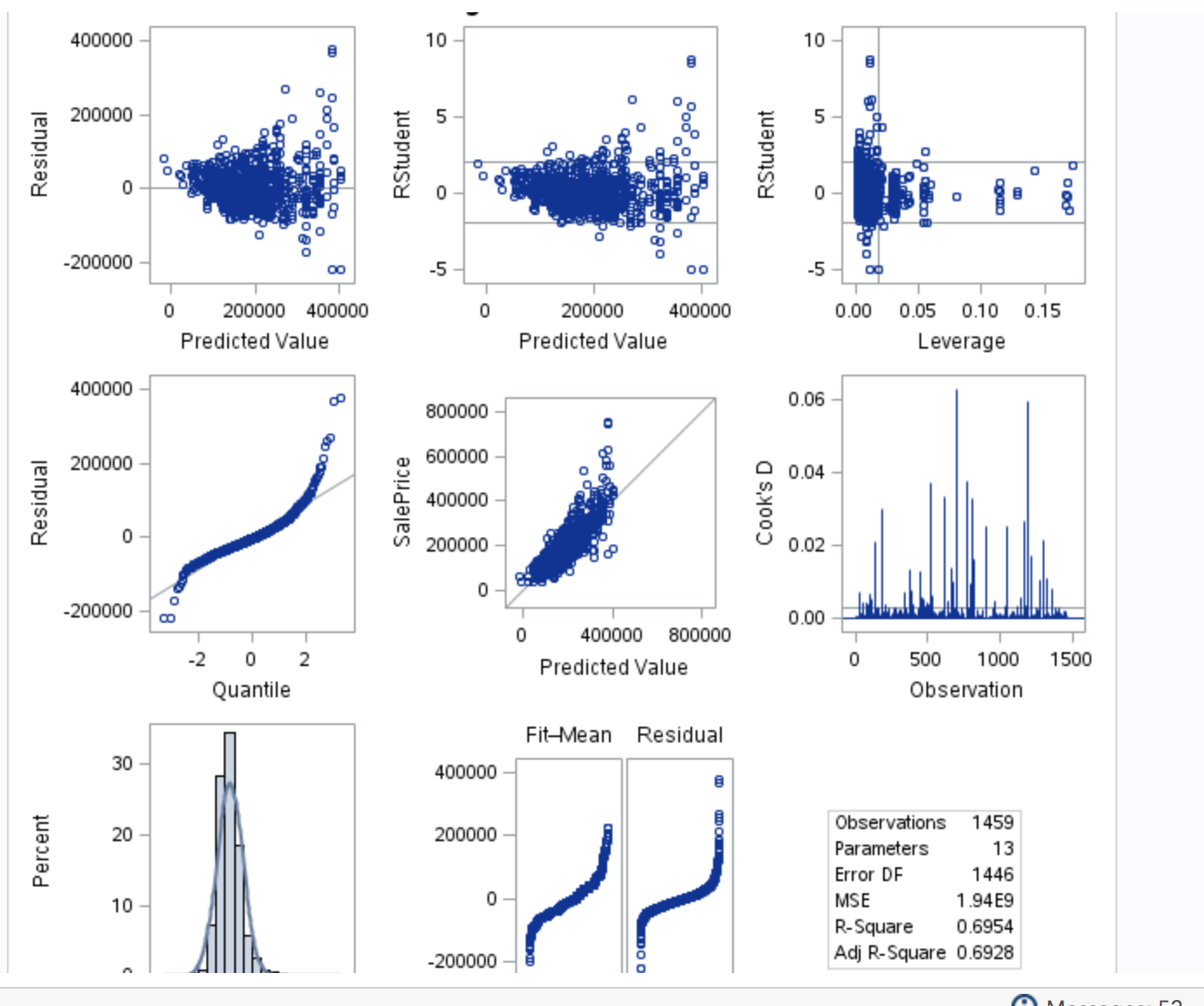
Figure 3 - Model Selection Comparison using partition of Test/Train Data

The Backward Model was a non-starter in that it had no results that were measurable to be considered. LASSO clearly had the smallest AIC and Cross Validation numbers, and the largest Adjusted R2 without excessive number of variables to artificially inflate the result.

### From the LASSO model, *PROC GLM* was used. Interactions were tried but the result was not sufficient enough to keep the interaction (.69 vs .7). Interestingly, the interaction model also degraded the assumptions of Residual Scatter, QQplot and Histogram in the Fit Diagnostics by a small amount.

### Below is the straight model without interaction. While the QQplot is the worst, the Residual Scatter does not show any curve and deemed sufficient. The Histogram is within normal ranges and sufficient.





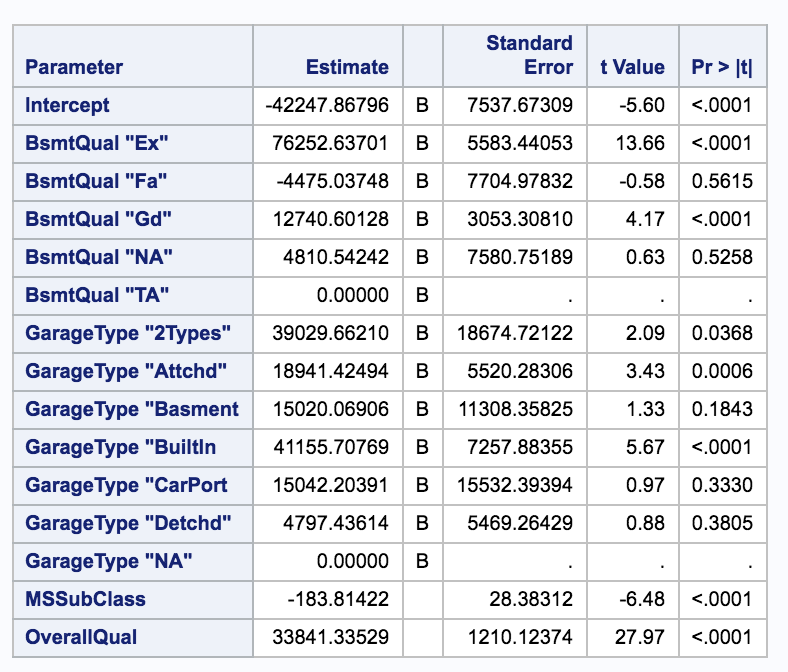


Figure 4 - Linear Regression and Fit Diagnostics for LASSO Model Selection with interaction

### Since this is Project 2, I will not be submitting the LASSO model to Kaggle.

### Project 1 write up has been deleted and moved to Appendix 3.

# Parameter Interpretation (updated)

SalePrice = Intercept (-42248) + MSSubClass (-184) + OverallQual (33841) + BsmtQualEx (76253) + BsmtQualGd (12741) + GarageType2Type (39030) + GarageTypeAttchd (18941) + GarageTypeBuiltin (41156)

If the Sale Price of $0, the prediction for the cost of the house would be - $42,248.

For every extra $1 in Sale Price, the predicted sales price decreases by $8,951:

Intercept -$42,248

Decrease by $184 for MSSubClass

Increase by $33,841 for OverallQual

If the home had the following features, every extra $1 in Sale Price is predicted to be impacted in the following way:

Increase by $76,253 for BsmtQualEx – Basement Quality Excellent

Increase by $12,741 for BsmtQualGd – Basement Quality Good

Increase by $39,030 for GarageType2Type – assume 2 car garage

Increase by $18,941 for GarageTypeAttchd – Attached garage

Increase by $41,156 for GarageTypeBuiltin -- Built in garage

Therefore, if the home had the following features, every extra $1 in Sale Price is predicted to impact in the following manner.

Excellent Basement:

SalePrice = Intercept (-42248) + MSSubClass (-184) + OverallQual (33841) + BsmtQualEx (76253)

= $67,662 increase

Good Basement:

SalePrice = Intercept (-42248) + MSSubClass (-184) + OverallQual (33841) + BsmtQualGd (12741)

= $4,150 increase

Two Car Garage:

SalePrice = Intercept (-42248) + MSSubClass (-184) + OverallQual (33841) + GarageType2Type (39030)

= $30,439 increase

Attached Garage:

SalePrice = Intercept (-42248) + MSSubClass (-184) + OverallQual (33841) + GarageTypeAttchd (18941)

=$10,350 increase

Built in Garage:

SalePrice = Intercept (-42248) + MSSubClass (-184) + OverallQual (33841) + GarageTypeBuiltin (41156)

=$32,565 increase

Since none of the log variables made it to the final model, no back transformation was done.

# CONFIDENCE INTERVALS

# Conclusion (updated)

The intercept of the Sale Price at -42248 is an unrealistic starting home price. The potential of features adding and subtracting to the overall Sale Price are realistic, as shown in the equations above. The coefficient of variation is relatively small at 24.34. The LASSO based model clearly has more linearity than previous attempts.

# Principal Component Analysis

Principal Component Analysis (PCA) will be used to select the optimum model.

Standardized the output and should variability play a role in selection of the variable. Look at high loadings about .4 or larger. Scree Plot look for elbow and # Components. Other diagram tells you the amount the variance explains … .8 or .9… value of all linear combinations on the plots.

As we look at model transformation from correlated variables to linear regression variables, Principal Component Analysis and Regression were used.

Numeric variables submitted into the Principal Component Analysis are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| MSSubClass | logLotFrontage | logLotArea | OverallQual |
| OverallCond | logYearBuilt | logYearRemodAdd | logMasVnrArea |
| logBsmtFinSF1 | logBsmtFinSF2 | logBsmtUnfSF | logTotalBsmtSF |
| logX1stFlrSF | logX2ndFlrSF | logLowQualFinSF | logGrLivArea |
| logBsmtFullBath | BsmtHalfBath | FullBath | HalfBath |
| BedroomAbvGr | KitchenAbvGr | TotRmsAbvGrd | Fireplaces |
| logGarageYrBlt | GarageCars | logGarageArea | logWoodDeckSF |
| logOpenPorchSF | logEnclosedPorch | logX3SsnPorch | logScreenPorch |
| PoolArea | logMiscVal | MoSold | YrSold |
| SalePrice |  |  |  |

# Linear Discriminant Analysis

Fishers. Compare 2 groups.

# Conclusion

Needs to be new.

# Addressing Comments

Complimentary comments are being skipped. Thank you for the kind words.

Comment: *I think it will look better without the dot in front of each section title*.

Response: There is no ‘dot’ in front of MSWord Heading1 on my Mac version. Likely this is a compatibility issue. A PDF will be used to eliminate this problem.

Comment: *Scatter plots, histograms, correlation matrix are expected for EDA.*

Response: The Exploratory section has been reassessed and rewritten.

Comment: *I suppose you meant external cross validation is performed but I didn’t see the data partitioning step in your code. This needs a little clarification.*

*I checked your code and see your CV methods in code which is used to build the model using the whole train.csv. But I view it as an internal CV. Our project description says “internal CV Press required” and “We will say that an external cross validation is mandatory to compare the adjusted R2 and AIC of competing models based on the validation dataset.”*

Response: My understanding was using all of the data as “train” to train the model, then use same data in test to do the testing and submit to Kaggle. This was lack of understanding on my part. This was redone for the new model approach in Project 2.

Comment: *The intercept means, when all predictors are 0, the sale price is -$44063.*

Response: See the new Parameter Interpretation section.

Comment: *“*Coefficient of Variation of 110.0 seems high as a unit-less number. *“*

*Don’t really understand this.*

Response: The statement was referring to the interpretation of 110.0. If the Coefficient of Variation represented the variation on housing price, say $110.0 in relation to a home that sells for hundreds of thousands of dollars, it is small. The result of the LASSO model changes the coefficient to 24.34.

Comment: If the Sales Price increases by $1, the Model predicts Basement Full Bath (BsmtFullBath) will decrease Sales Price by approximately $43,554

*When you interpret the coefficient in regression model, focus on how one unit increase of the predictor will impact the response. Don’t start with the change of response.*

Response: See new Parameter Interpretation section.

Comment: *“The model exhibits enough of a Normal distribution with adequate scatter…”*

*Again, the variance is clearly not constant*

Response: Chose to rework the EDA and the approach. See new Exploratory Data Analysis section.

Comment: *Are there two captions given to the same table? There is already one title above this table.*

Response: Proof reading error. Fixed.

Comment: *Where is equilibrium? Not sure at this point, but recognize that the data mining effort was clearly not worth the effort.*

Response: Chose to rework the EDA and the approach.

Comments: Data Mine Model comments.

Response: This approach was eliminated in Project 2 so the comments are understood, but not addressed.

# Appendix 1 – Code Project 2

## Principal Component Code

## Linear Discrimination Code

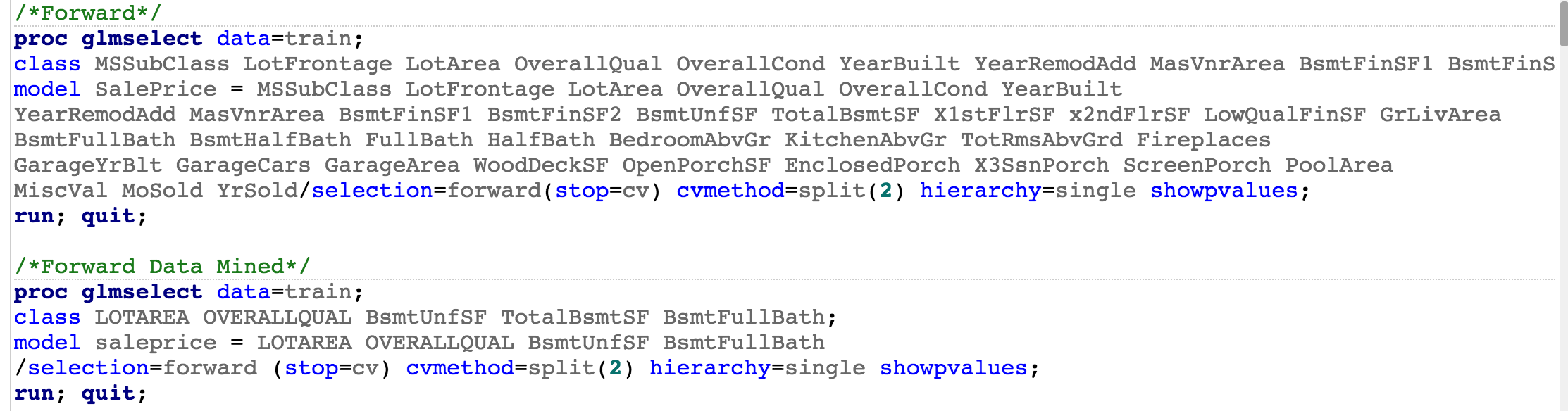
## Code that is ‘redo’ of Project 1

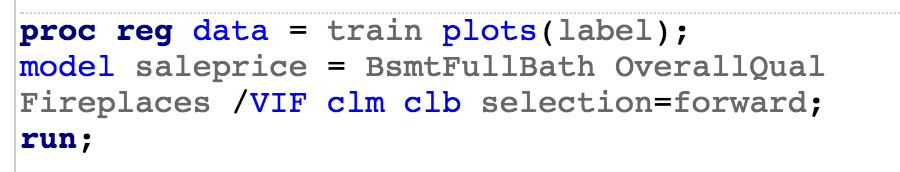
# Appendix 2 – Project 1 CODE

## Code

### Model 1 - Forward

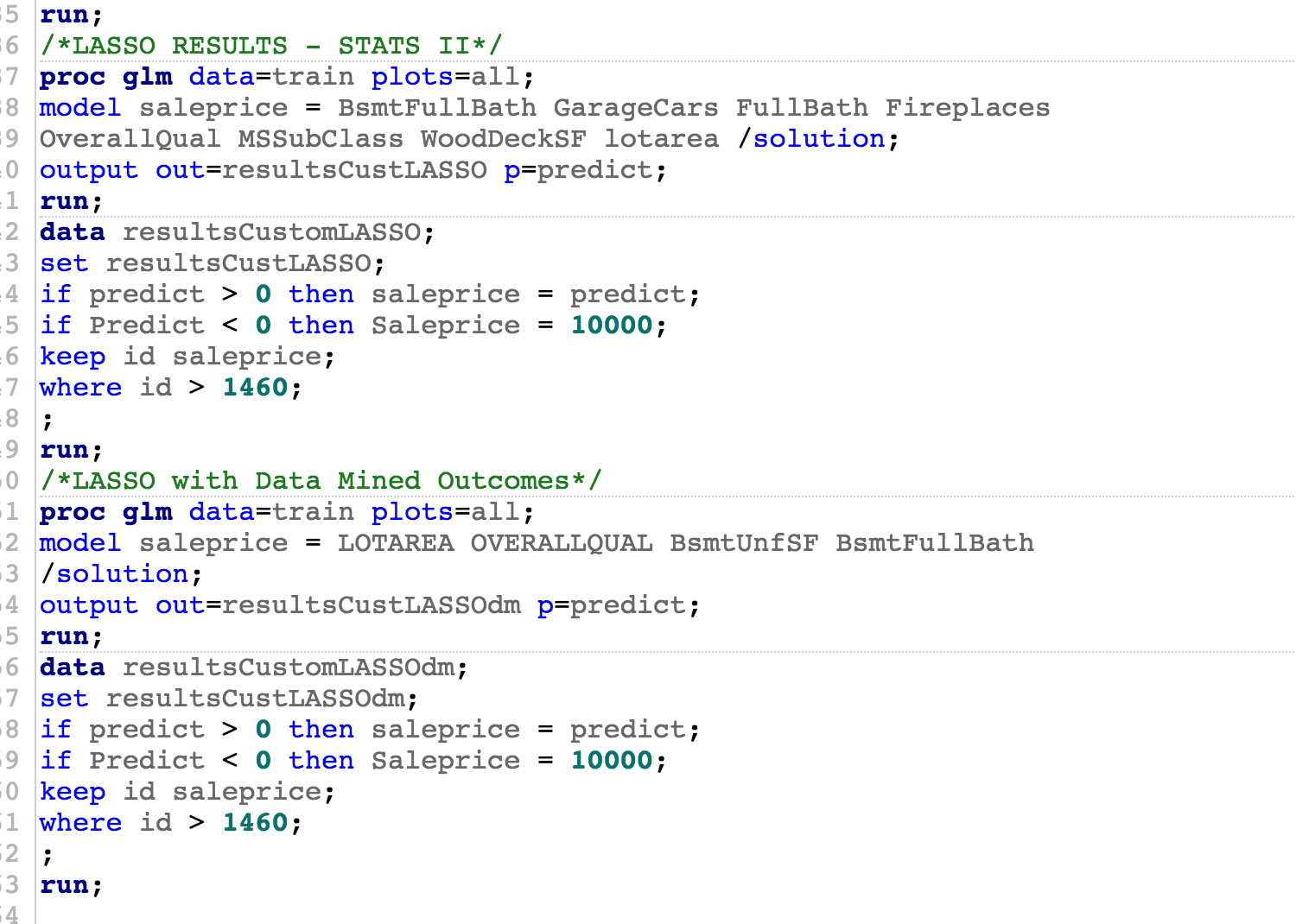


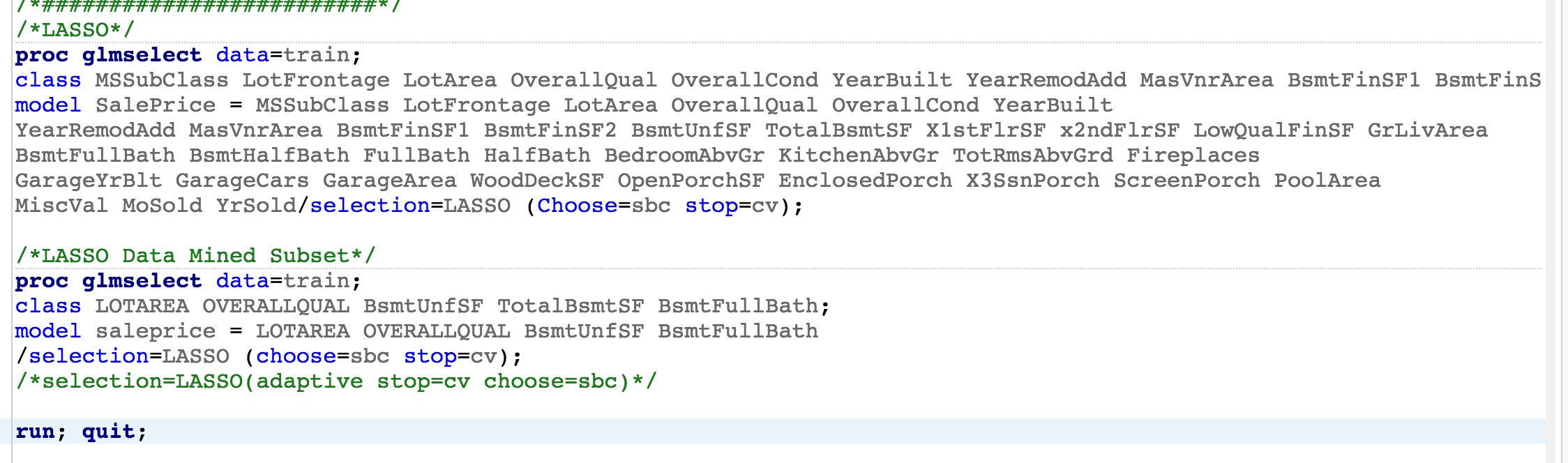




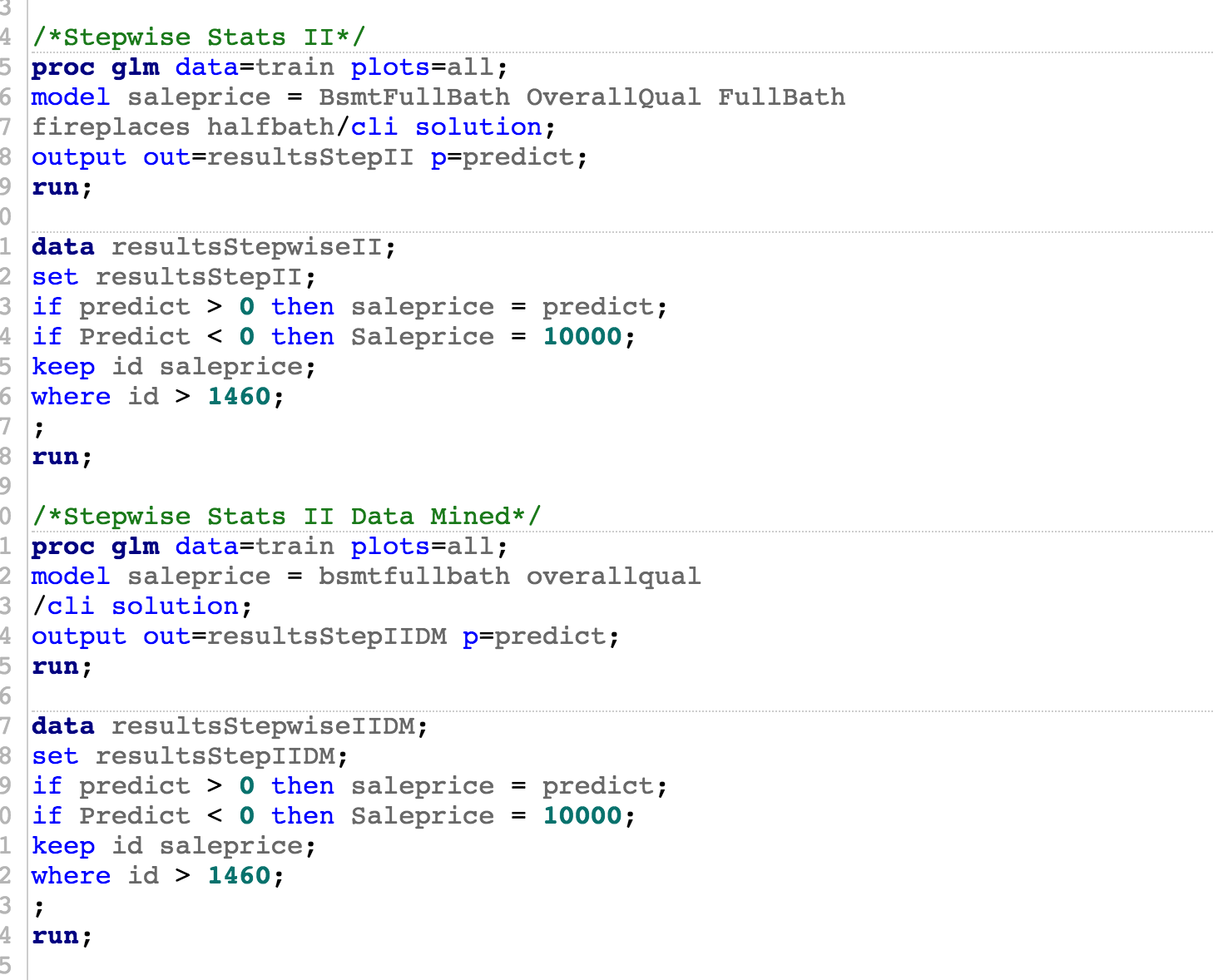


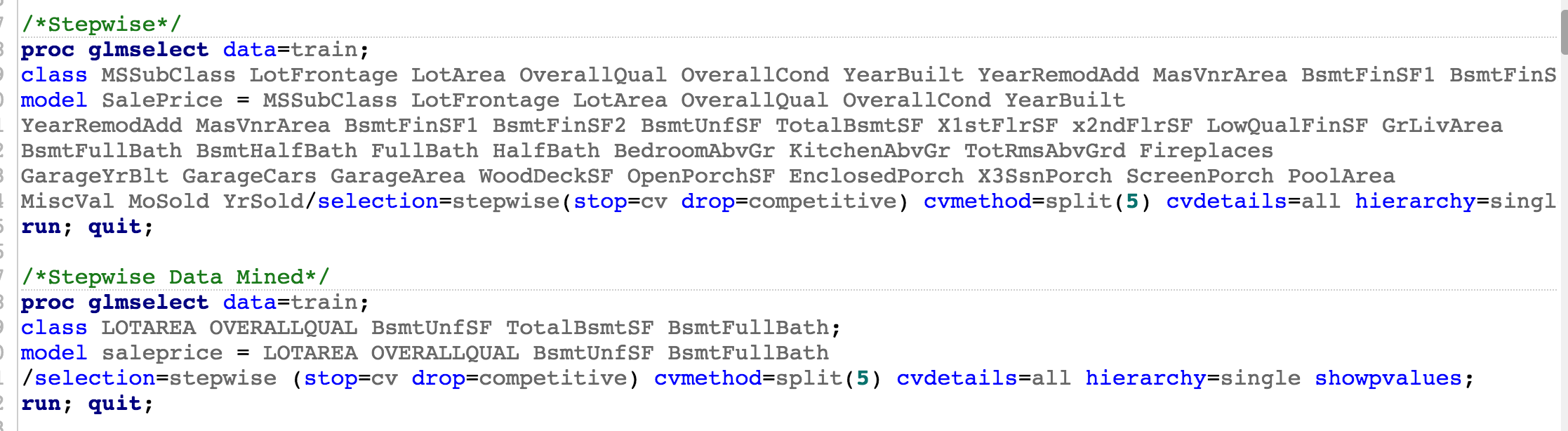
### Model 2 - LASSO

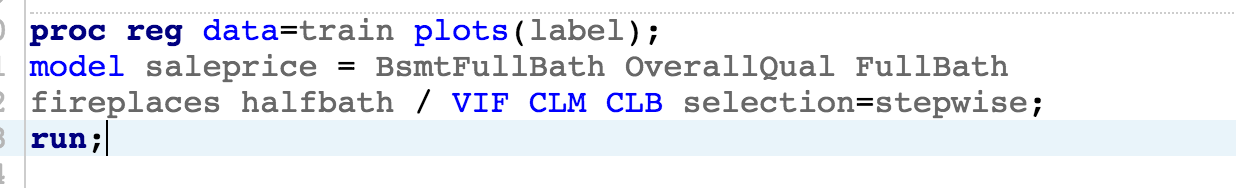


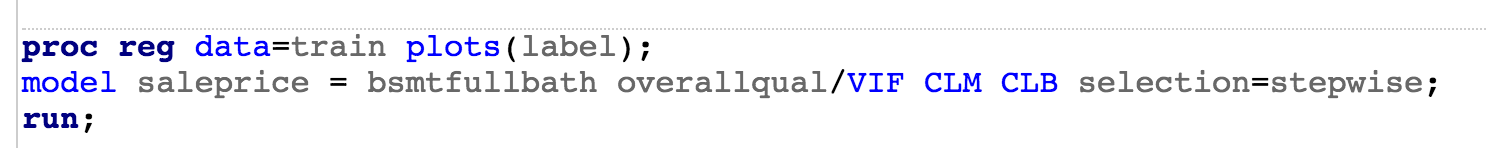


### Model 3 - Stepwise

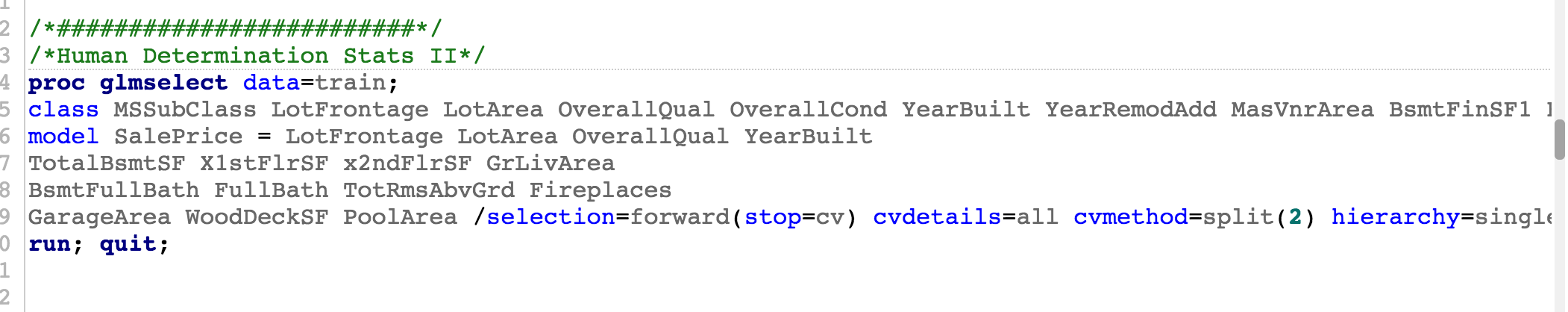








### Model 4 - Human



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)