DUE SUNDAY 10/1

# Introduction

The goal of Project 1 is to predict the final sales price of each home in the Kaggle – Ames, IA dataset, aka “Dream Data”. Per the project guidelines, two (2) specific questions with three (3) competitive models each are to be used to predict final home price.

# Data Description

The Dream Data is from Century 21 realtors and represents the entire population of this observational study. No causal inference can be implied across the Ames, IA housing market based on this study. There are 1,460 observations of homes sold between 2006 and 2010 with 80 variables used to determine sales price.

# Exploratory Analysis

Given the complexity of the data for this team, it was decided to run the models with the Dream Data in its entirety and a subset, called Data Mined. See Appendix

[….The data was first mined by grouping the variables into logical groups to determine which variable in the category was the most influential. The categories are as follows:

* Lot
* Quality
* Basement
* House
* Garage
* Deck & Porch
* Pool & Miscellaneous
* Year Sold

The response variable is the Sales Price.

Blocking was not chosen because of the many levels of the variables. ]

# Question 1

## Question 1: Problem Statement

Create a useful, valid model about the Dream Data that includes confidence intervals for the response variable Sales Price, including a description of the model selection process.

## Question 1: Model Selection

Type of Selection

LASSO, Model Averaging

Stepwise, Forward, Backward, Mallows Cp,

Manual / Intuition

A mix of all of the above.

**At least two of the above required.**

Checking Assumptions

Residual Plots

Influential point analysis (Cook’s D and Leverage)

Comparing Competing Models

AIC, BIC, adj R2 **Required**

Interval CVPress **Required**

External Cross Validation **Required**

Parameter Interpretation

Interpretation (Verbal) Required

Confidence Intervals **Required**

# Question 2

## Question 2: Problem Statement & Model Selection

The goal was to create the most predictive model for the Sales Price from the Dream Data. The approach used was LASSO, FORWARD and STEPWISE selection methods on the Dream Data (entire variable set) and the Data Mine variables (see Appendix I).

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 reading discouraged STEPWISE, the *“Introducing the GLMSELECT PROCEDURE for Model Selection”[[1]](#footnote-1)* demonstrated that STEPWISE selection can be the most powerful for predictive analytics.

The use of the Data Mine set was to see if smaller, statistically significant variables for each logical grouping would improve our ability to predict Sales Price.

Data set cleansing consisted of changing ‘0’ and ‘NA’ to response averages, mainly due to the fact that Kaggle requires all observations in the data set and SAS will not allow a ‘0’. Given the volume of ‘0’ responses, eliminating the ‘0’ could have reduced the size of the data set past the Central Limit Theorem levels. Both data sets met the assumptions of Normal (enough) distribution, scattered residual plots, and decent QQplots, therefore no transformations were conducted. Evaluation of Outliers for both data sets, Dream Data (Figure 1) and Data Mine (Figure 2), did not show the appearance of errors, just more expensive homes. The sheer volume of the Dream Data variables masked the finding of any significant individual outlier (see Appendix 1).

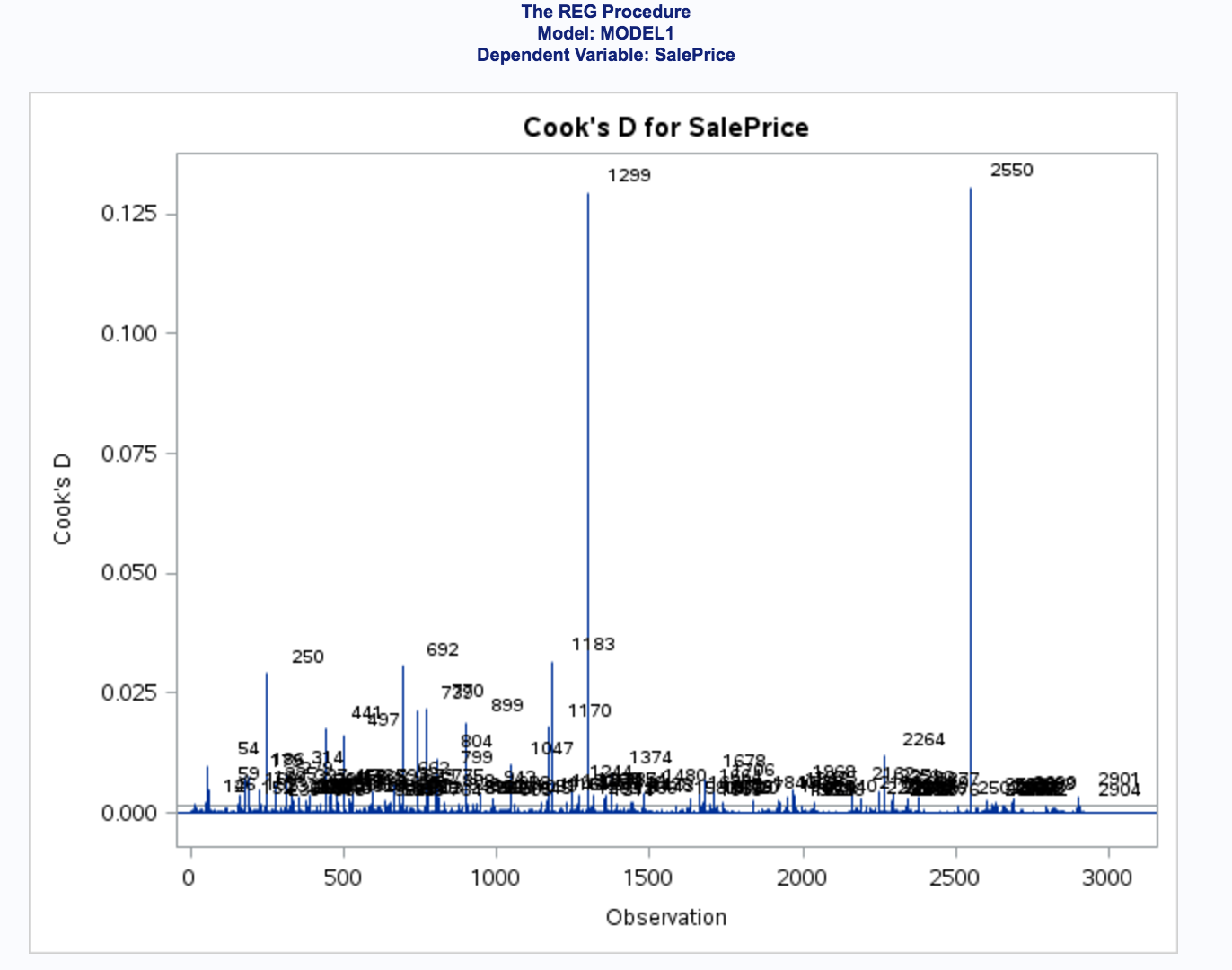
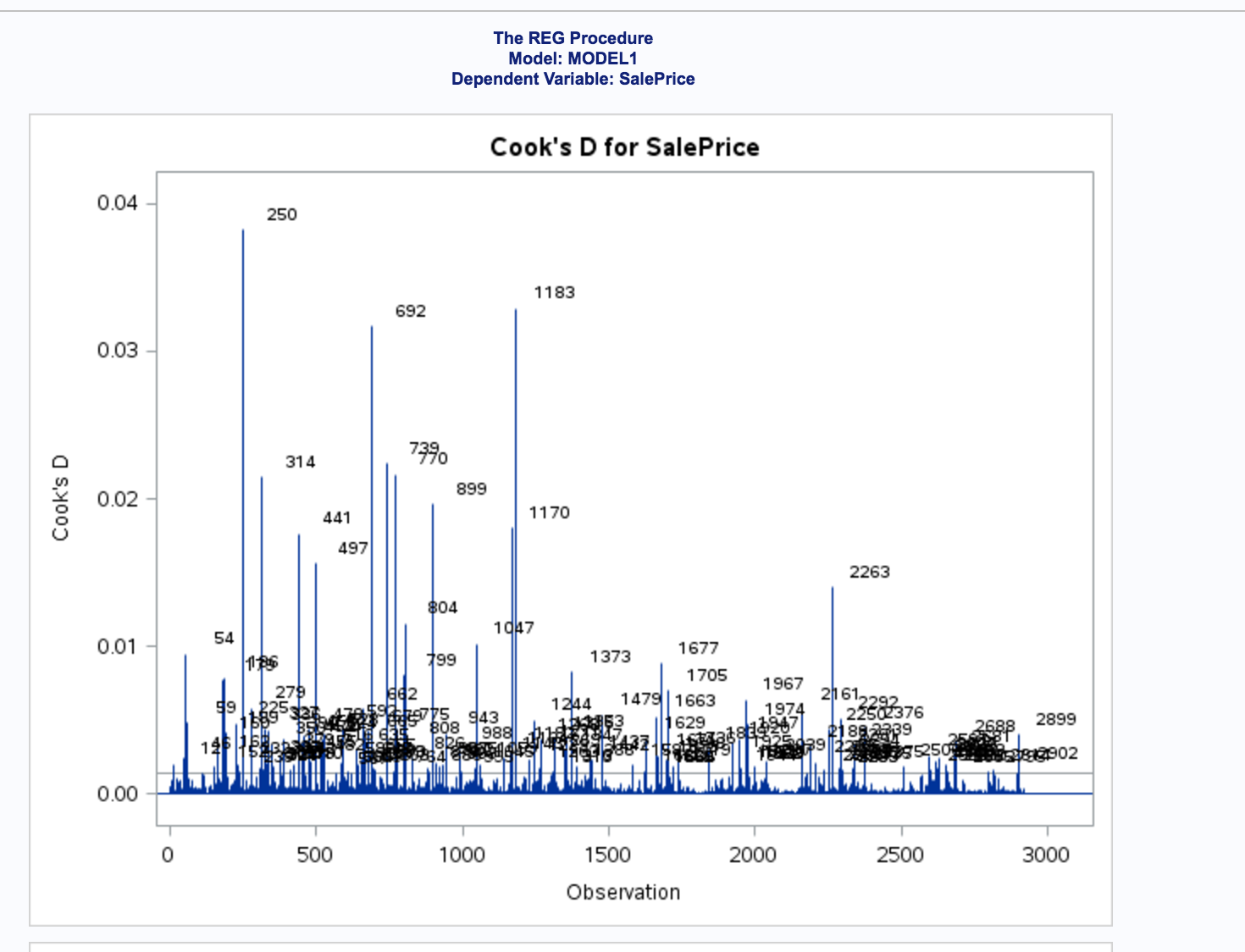


Figure 1 - CooksD for Dream Data Figure 2 CooksD for Data Mine data set

The results are as follows below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test Set Models** | **Adjusted R2** | **AIC** | **Root MSE** | **BIC or SBC[[2]](#footnote-2)** | **External CVPress** | **Kaggle**  **Score** |
| Model 1-Forward Data Mine | 0.44 | 68841 | 79954 | 66004 | 1.89 | 0.905 |
| Model 1 – Forward Dream Data | 0.45 | 68774 | 78930 | 65992 | 1.84 | 0.877 |
| Model 2 – LASSO Data Mine | 0.38 | 69129 | 84098 | 66250 | 1.96 | 0.905 |
| Model 2 – LASSO Dream Data | 0.40 | 69048 | 82909 | 66181 | 1.86 | 0.909 |
| Model 3 – Stepwise Data Mine | 0.44 | 68841 | 79954 | 66004 | 1.89 | .865 |
| Model 3 – Stepwise Dream Data | 0.45 | 68774 | 78930 | 65985 | 1.86 | .941 |
| Model 4 – Human Inference & Selection Dream Data | 0.45 | 68774 | 78930 | 65985 | 1.84 | .912 |

Figure - Results Summary Table

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). The unexpected was the low Adjusted R2.

# Conclusion

Clearly the Data Mine data set did not generate an improvement in the model. In each case, this data set scores were not as good as the Dream Data.

Model 4 – Human Inference & Selection Dream Data 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.

Is there such a thing as data fatigue or does Occam Razor win out? In 6371 Kaggle effort, results of Adjusted R2 was much higher for the best model.

|  |  |  |  |
| --- | --- | --- | --- |
| Backward | .9945 | 1.40454E14 | .37350 |

Figure - Class 6371 Kaggle Results

Was this the result of not knowing as much about how to approach the data? Or are standard, simple approaches the best?

# Appendix

## Question 1: Code

## Question 2: Data Mine Data Set

Beyond the straight model fitting, a data mining and manual/intuition approach was 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 were 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 was used in the Data Mined dataset, which included:

* 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, turned out not to be significant.

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

While model’s assumptions looked 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 was 0.17

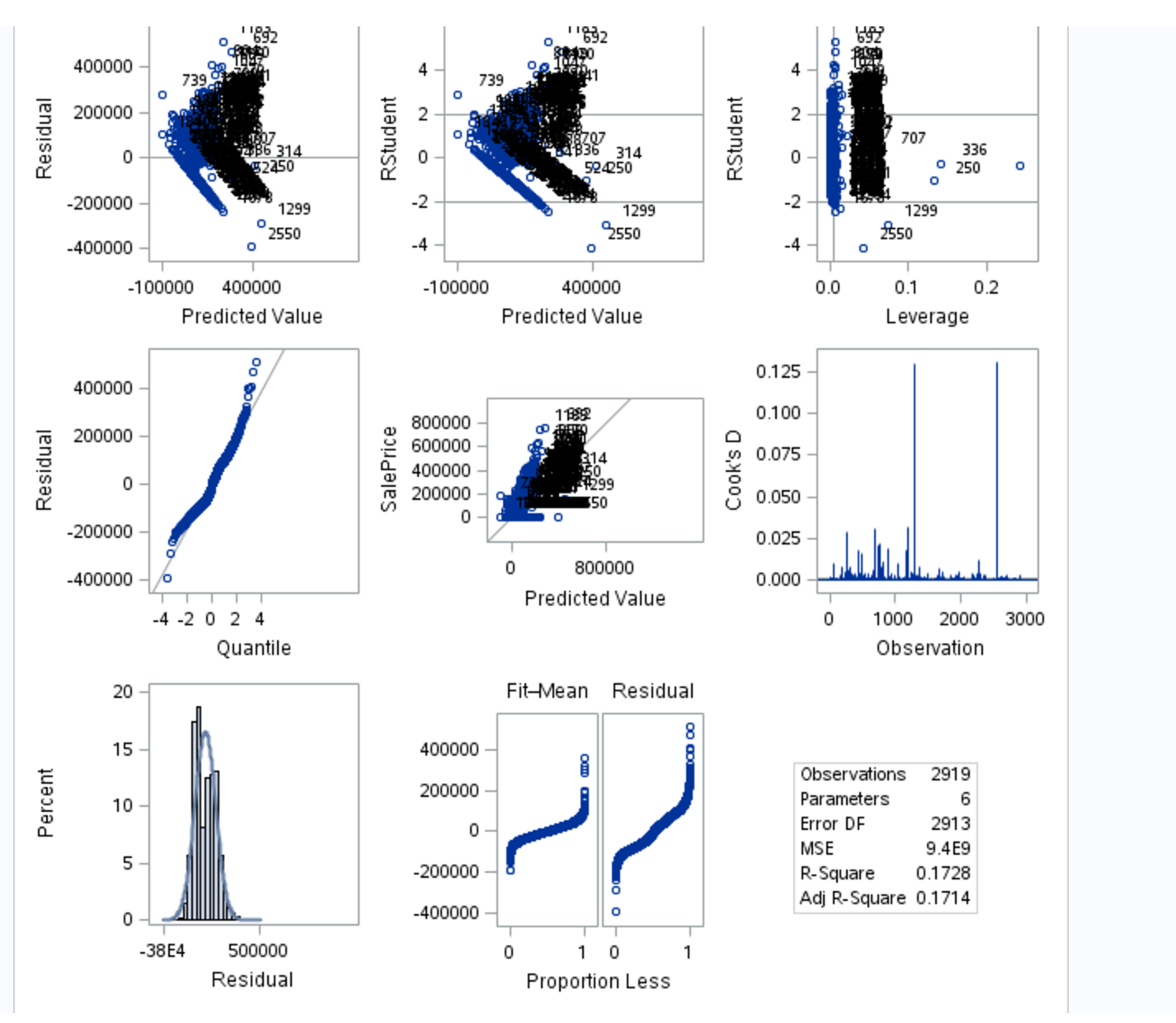
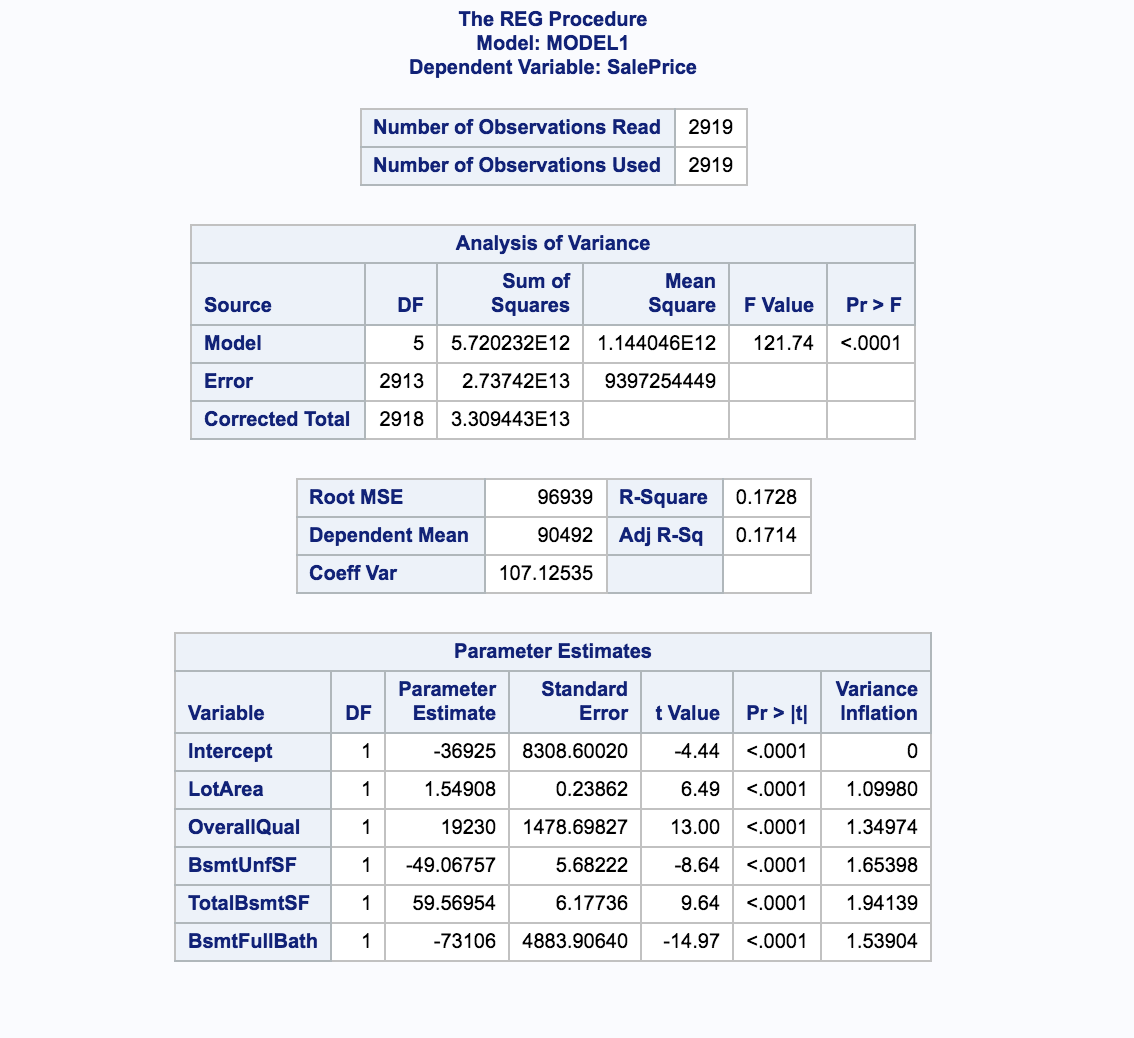


Figure – Output from Proc Reg with plots

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

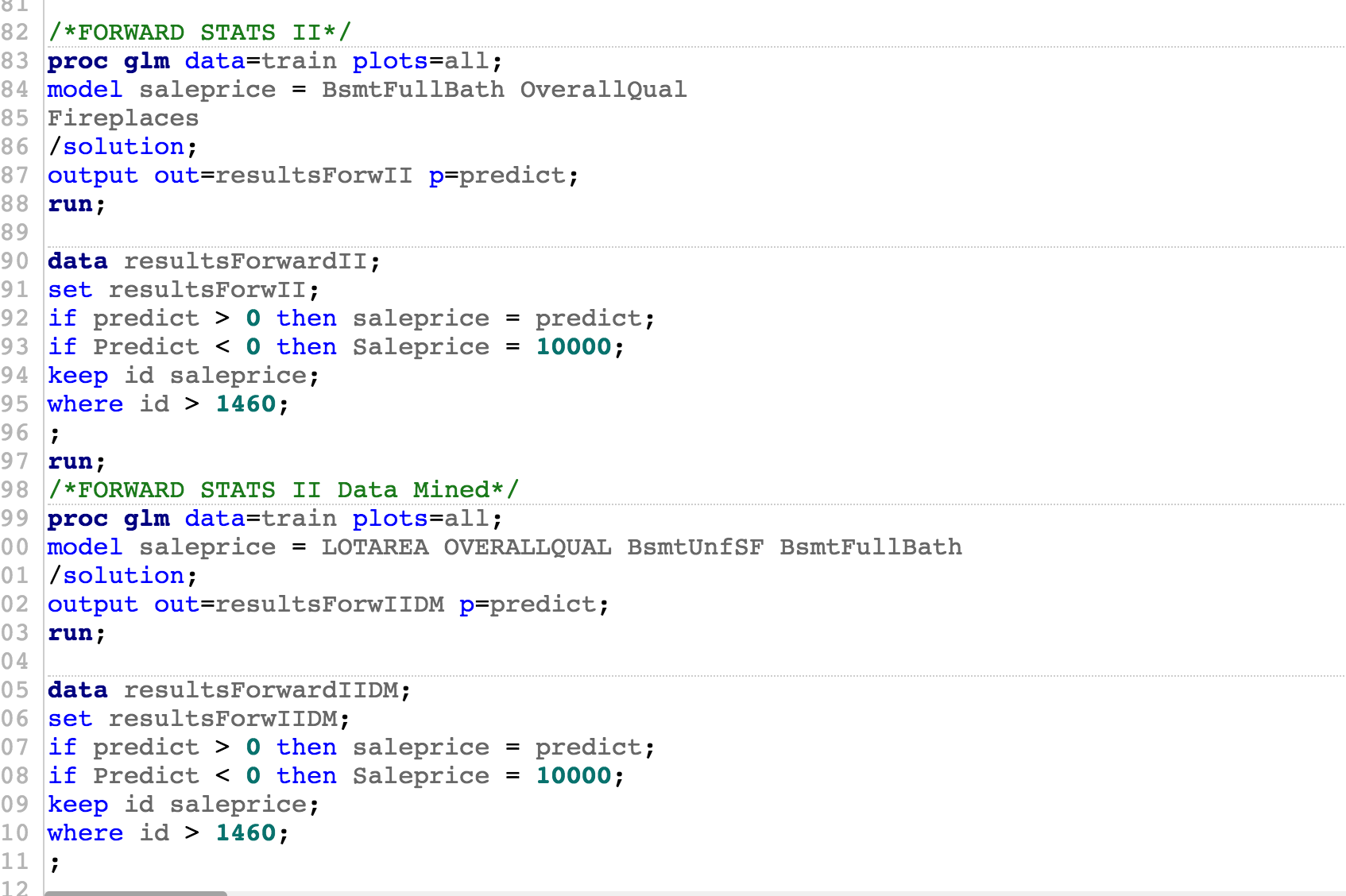


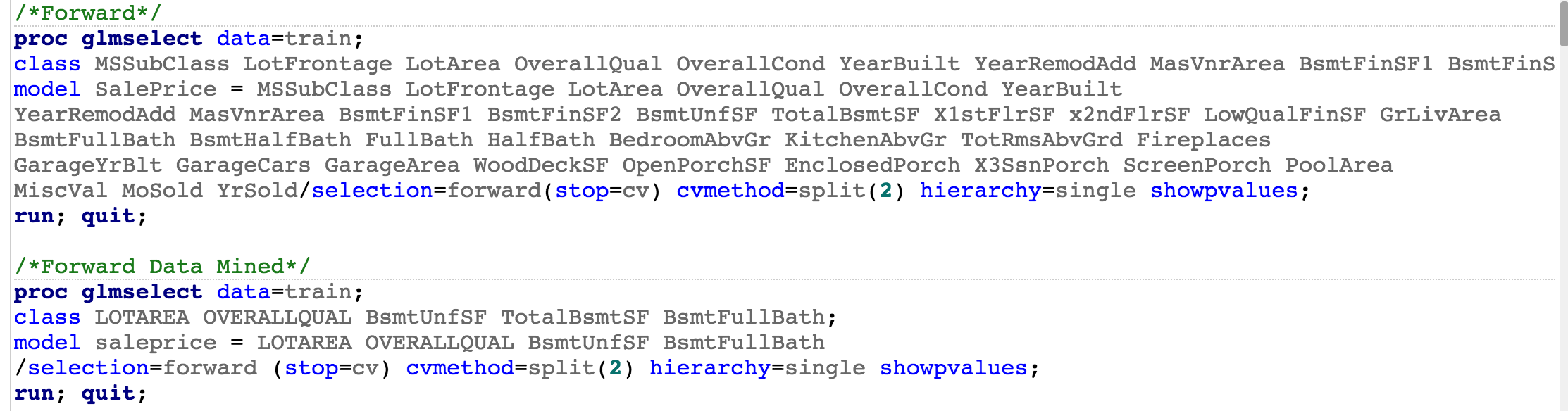
Figure - Proc GLM results after outliers removed

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

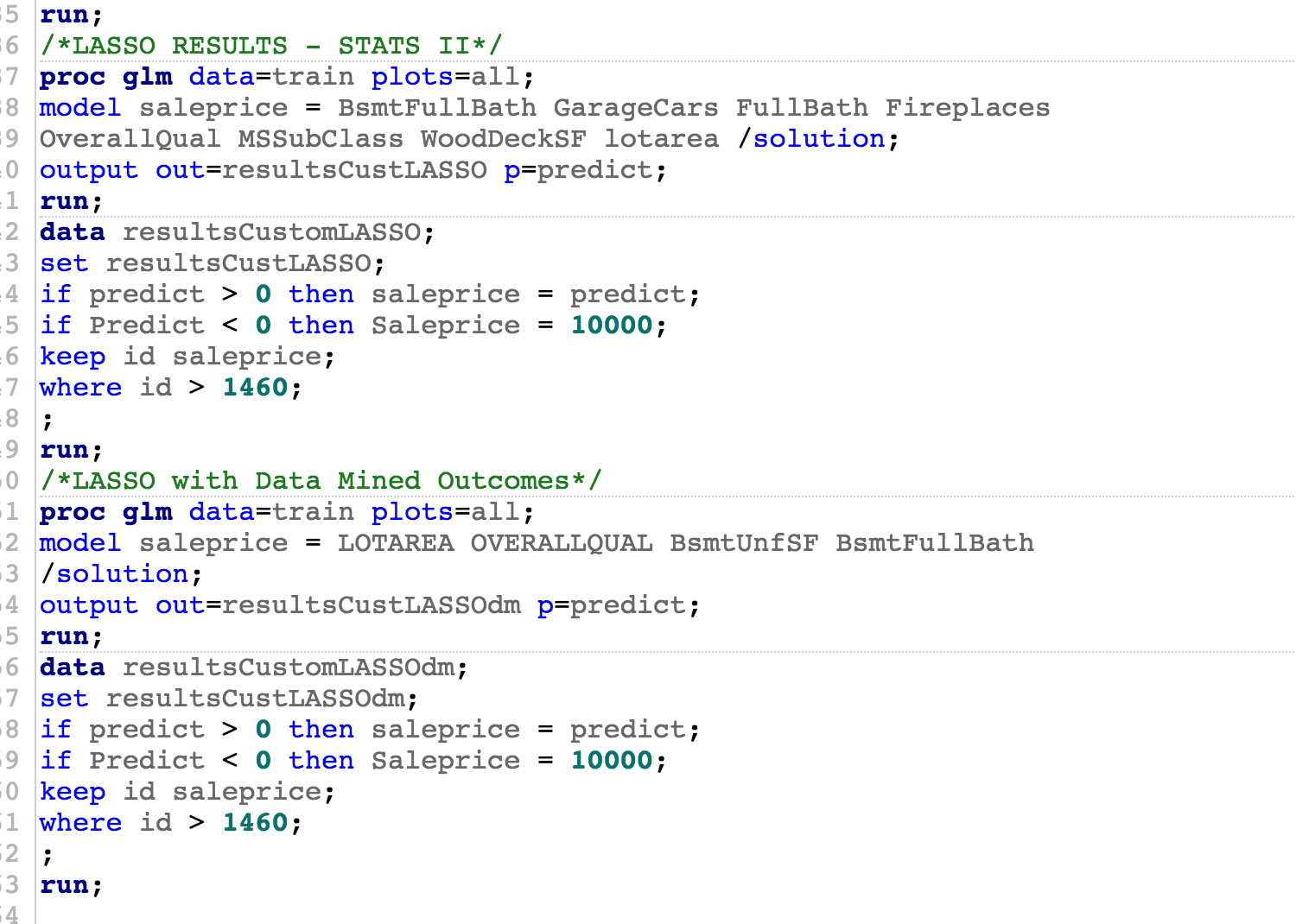
## CODE

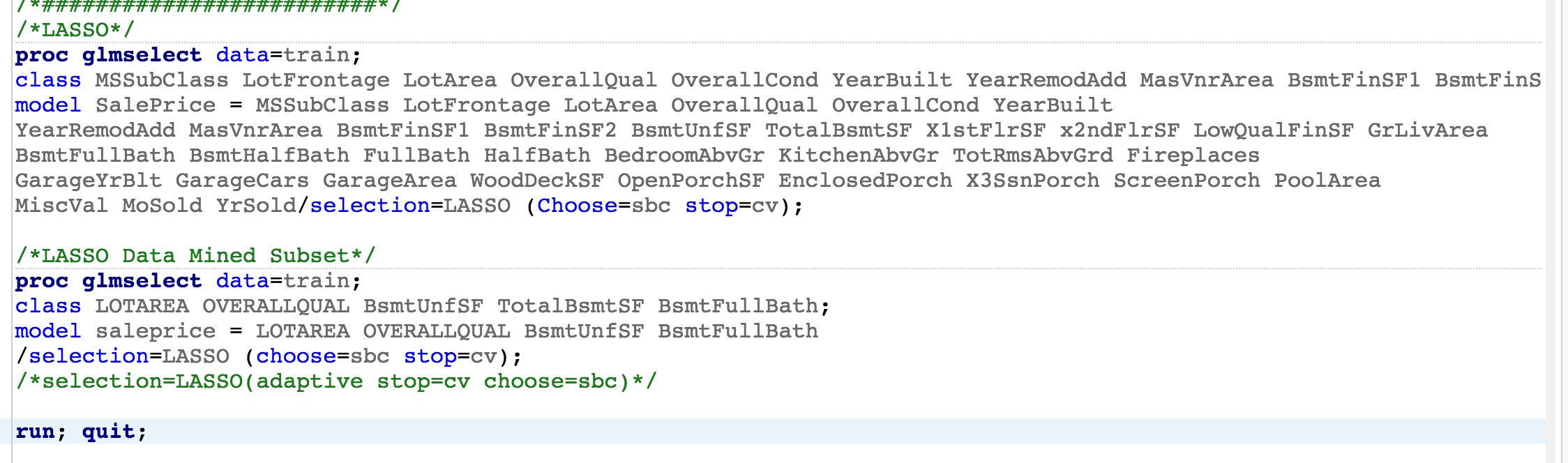
### Model 1 - Forward



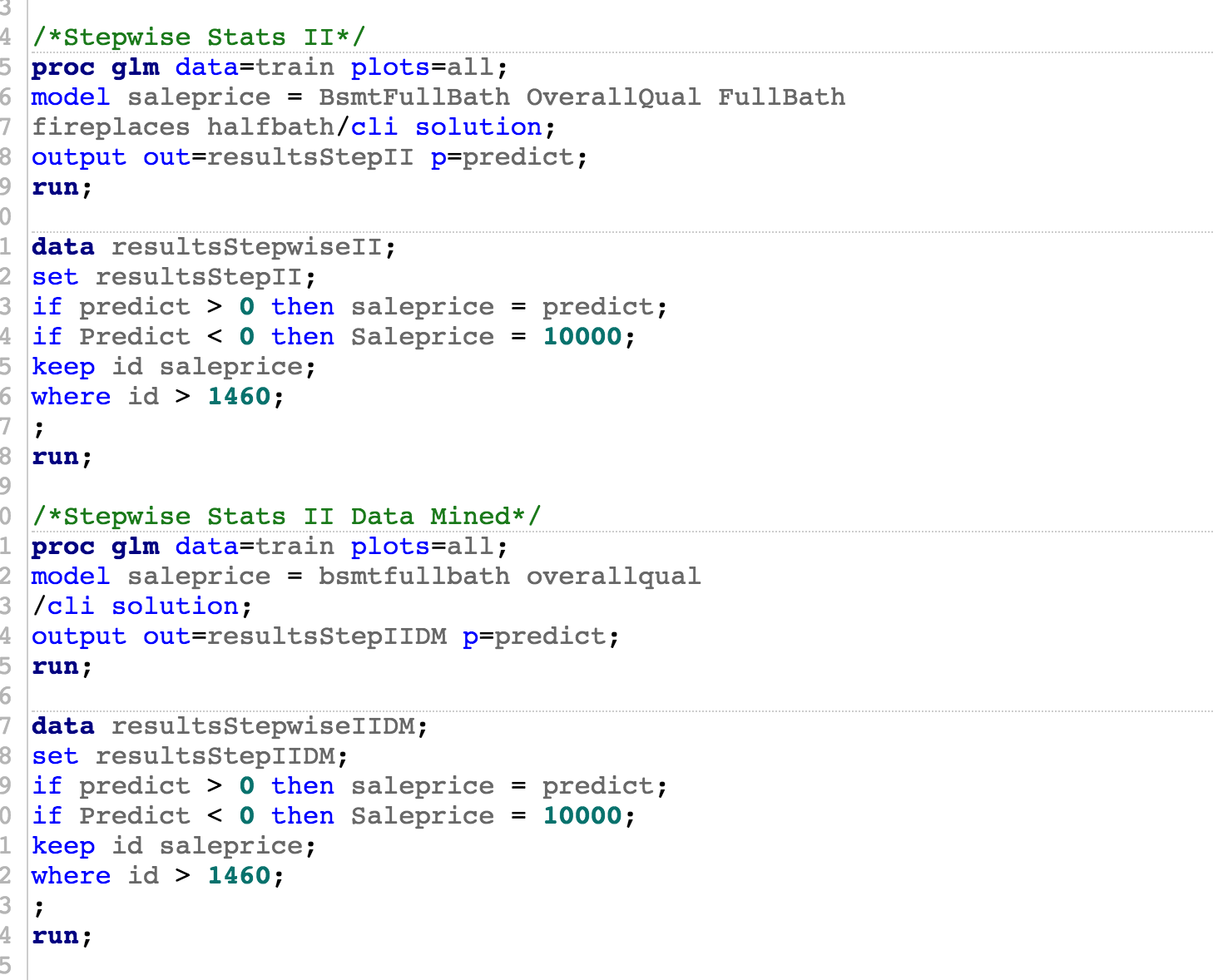


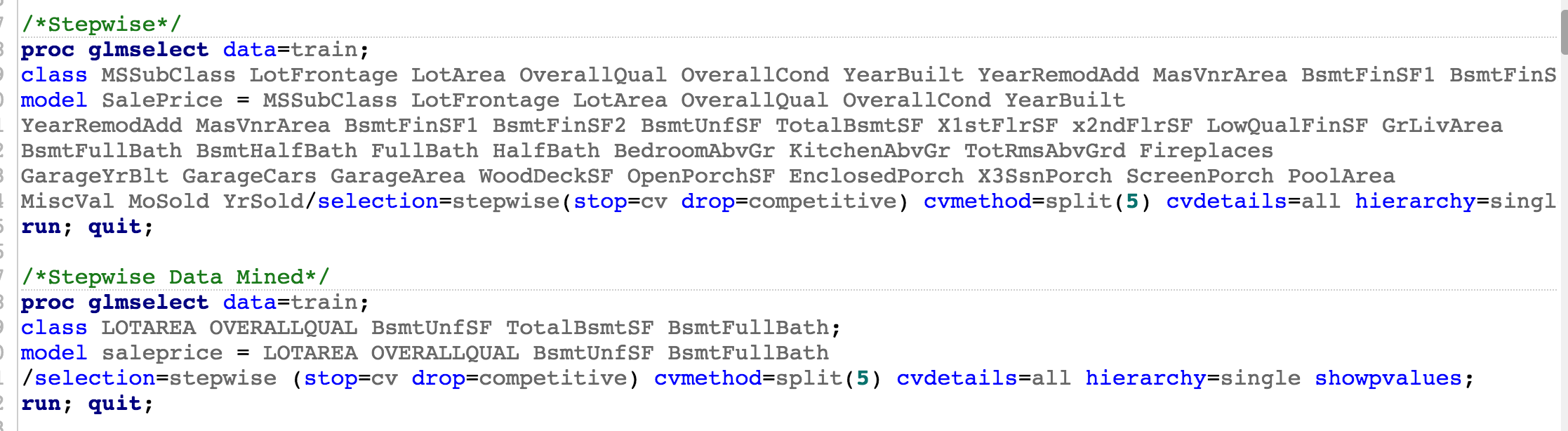
### Model 2 - LASSO



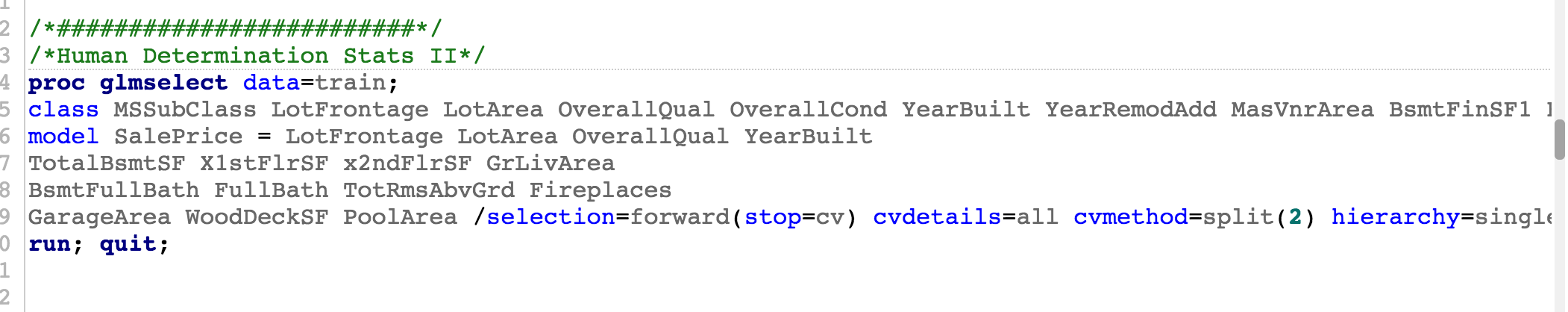


### Model 3 - Stepwise





### Model 4 - Human



1. Robert A. Cohen, “Introducing the GLMSELECT PROCEDURE for Model Selection”, SAS Institute Paper 207-31, no date, page 17. [↑](#footnote-ref-1)
2. Bayesian Information Criteria (BIC) = to Schwartz criterion (SBC) based on https://en.wikipedia.org/wiki/Bayesian\_information\_criterion [↑](#footnote-ref-2)