# 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 Sales Price from the Dream Data. The approach used was LASSO, FORWARD and STEPWISE selection methods on the entire variable set and the data mined variables.

The use of LASSO, a more modern approach to model selection, is balanced by the use of more traditional 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 mined data was to see if a smaller, statistically significant variables for each logical grouping would improve our ability to predict Sales Price.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test Set Models** | Adjusted R2 | AIC | Root MSE | BIC or SBC[[2]](#footnote-2) | External CVPress | **Kaggle**  **Score** |
| Model 1-Forward Data Mined | 0.4389 | 68841 | 79954 | 66004 | 1.89 | 0.905 |
| Model 1 - Forward | 0.4507 | 68774 | 78930 | 65992 | 1.84 | 0.877 |
| Model 2 – LASSO Data Mined | 0.378 | 69129 | 84098 | 66250 | 1.96 | 0.905 |
| Model 2 - LASSO | 0.394 | 69048 | 82909 | 66181 | 1.86 | 0.909 |
| Model 3 – Stepwise Data Mined | 0.44 | 68841 | 79954 | 66004 | 1.89 | .865 |
| Model 3 - Stepwise | 0.45 | 68774 | 78930 | 65985 | 1.86 | .941 |
| Model 4 – Human Inference & Selection | .45 | 68774 | 78930 | 65985 | 1.84 | .912 |

# Conclusion

Mixed result conclusion.

Evaluation of outliers determined no appearance of errors, just more expensive homes. Therefore, we chose not to eliminate any observations (and the Kaggle process would not let you).

The Kaggle process expected a full dataset and SAS expected no ‘0’. Dataset cleansing of ‘0’ or ‘NA’ was changed to response averages. It is unclear if eliminating observations that were incomplete or heavily 0 would produce better SAS and/or Kaggle score. Given the volume of ‘0’ responses, the dataset could have been reduced past the Central Limit Theorem levels.

The outcome was unexpected in that the best Adjusted R2 and CVPress did not result in best Kaggle score for prediction of home sale price.

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

In summary, data analysis depends heavily on the question(s) being answered. Insight into the data could be one approach and prediction could require a different approach. In the end, there seems to be no “right” answer in data analysis.

Foundation Statistics I course results are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Predictive Models | Adjusted R2 | CV PRESS | Kaggle Score |
| Forward | .7917 | 2.189873E12 | .33181 |
| Backward | .9945 | 1.40454E14 | .37350 |
| Stepwise | .8201 | 2.376408E12 | .25672 |
| CUSTOM | .9836 | 3.400997E13 | .2511 |

How is useful model different than a predictive model. – Depends on the question you are answering determines the “right” model.

The conclusion should reprise the questions and conclusions of the introduction, perhaps augmented by some additional observations or details gleaned from the analysis section. New questions, future work, etc., can also be raised here.

# Appendix

## Question 1: Code

## Question 2: Data Mined 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 & Misc
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, 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.

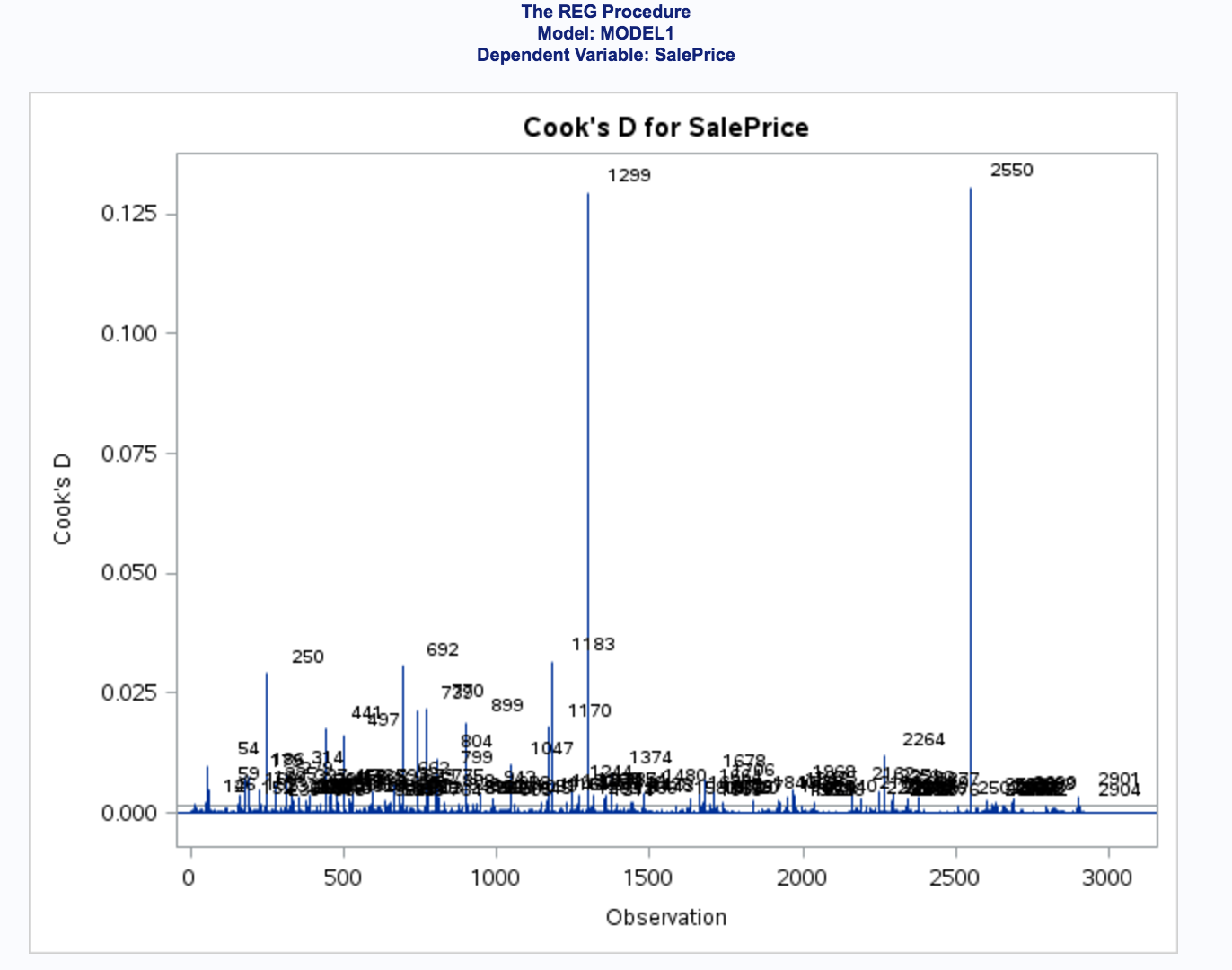


Figure 1 - CooksD highlighting outliers

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

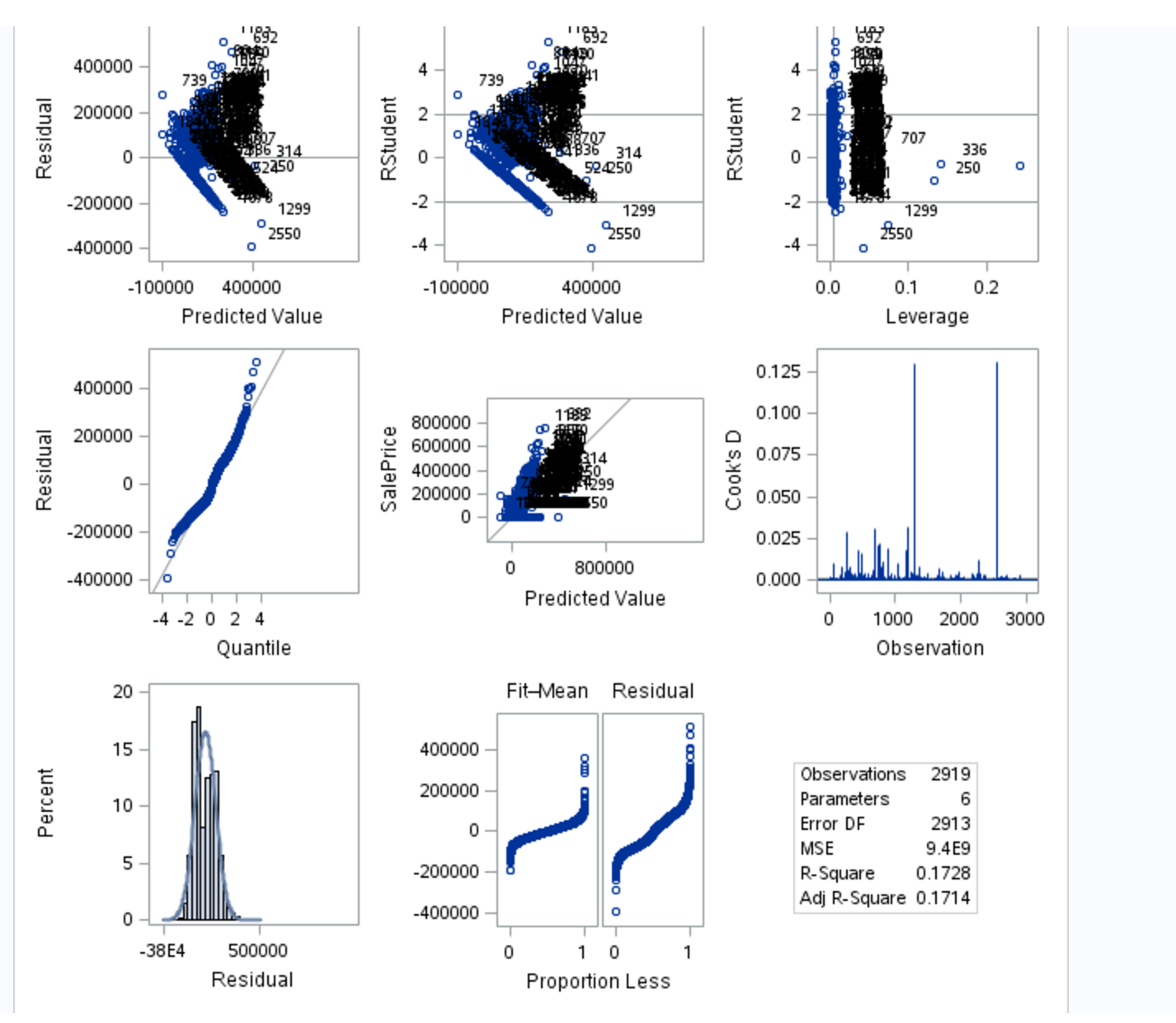
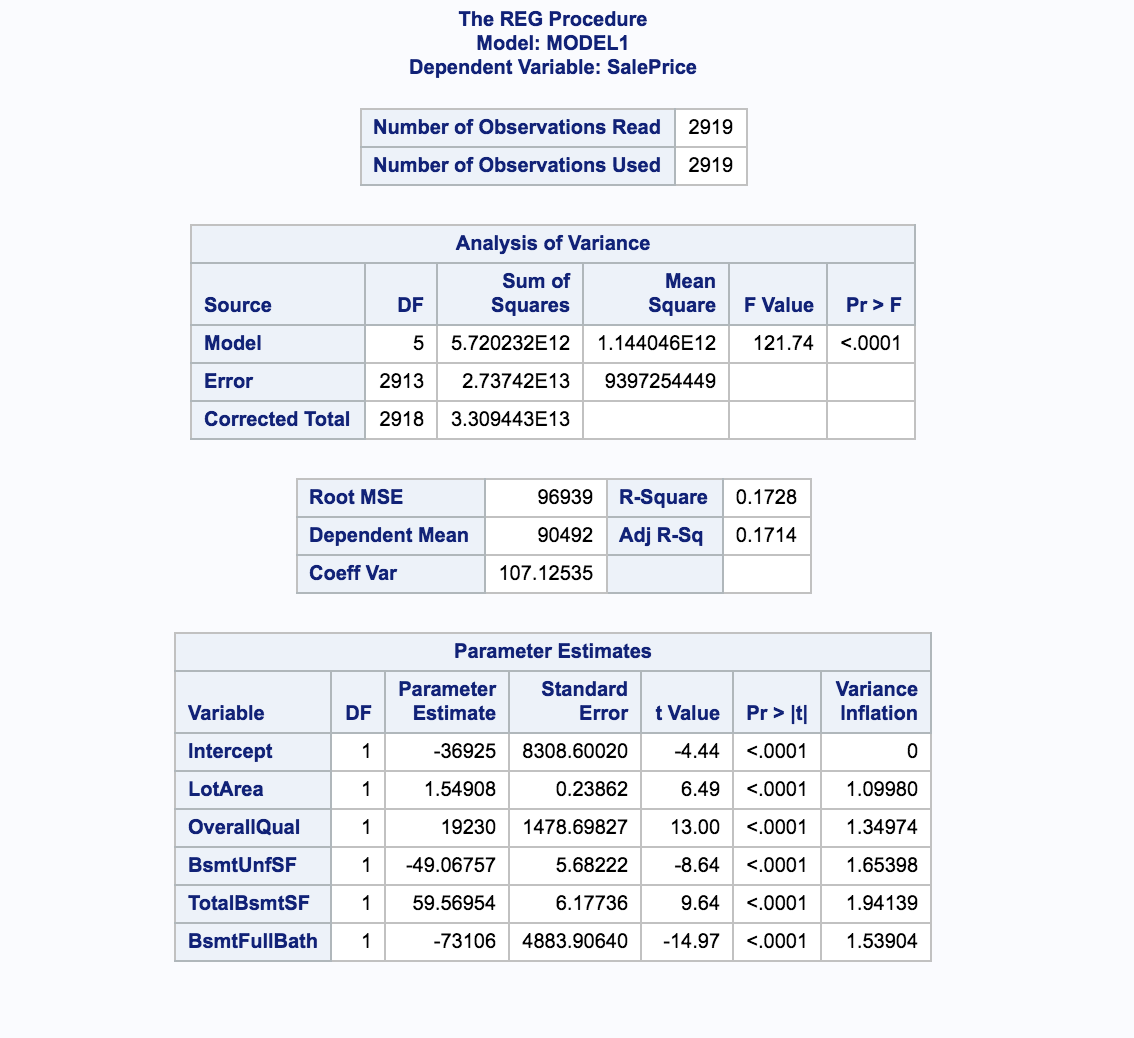


Figure 2 – Output from Proc Reg with plots

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



Figure 3 - 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 all 80 variables, aka Dream Data.

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)