**Project 1**

**By: Lizzy Sterling and Scott Herford**

Table of Contents

[Introduction 1](#_Toc506549306)

[Data Description 1](#_Toc506549307)

[Exploratory Analysis 2](#_Toc506549308)

[Analysis Question 1 3](#_Toc506549309)

[The Problem 3](#_Toc506549310)

[Forward Model 3](#_Toc506549311)

[Backward Model 3](#_Toc506549312)

[Stepwise Model 4](#_Toc506549313)

[Table: 4](#_Toc506549314)

[Analysis Question 2 5](#_Toc506549315)

[The Problem 5](#_Toc506549316)

[The Model & Assumptions 5](#_Toc506549317)

[Comparing Competing Models 5](#_Toc506549318)

[Conclusion 5](#_Toc506549319)

[Appendix 5](#_Toc506549320)

[Exploratory Analysis 5](#_Toc506549321)

[Analysis 1 5](#_Toc506549322)

# Introduction

“Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this Kaggle competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence”(De Cock, D., Kaggle 2011). This project, investigates factors that influence sale prices of houses in Ames, Iowa. We will conduct two analyses for this project to examine the influence that neighborhood and square footage have on sale prices of homes as well as other factors that may be significant on the sale price of homes. This project could benefit realtors and homebuyers in Ames that may want to know what explanatory variables influence home sale prices.

# Data Description

This data is from the Kaggle competition “House Prices: Advanced Regression Techniques” (see more at <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>). The data consists of 79 explanatory variables that describe elements of homes in Ames, Iowa. The training data set consists of 1460 homes with the test data set providing 1459 more observations, and observations consisting of 2919 homes.

***Note: All figures and graphs have been placed in the appendix for readability and space conservation purposes. Please refer to the*** [***appendix***](#_Appendix) ***to see figures referenced in this project.***

# Exploratory Analysis

To begin this analysis, we first needed to check all of the variables to make sure that all of the data was entered in correctly. We found that LotFrontage was saved in the data as a character datatype rather than numeric, so we altered the datatype in order to run a proper analysis.

Next we took into consideration three main assumptions: **Normality**, **Equality of Variance**, **linearity,** and **Independence**. We started to look at this by comparing our response variable, SalePrice, side-by-side with every other quantifiable variable. Looking at [figure 1](#_Exploratory_Analysis), the data appeared to be very skewed and non-linear, so we attempted a log transformation on SalePrice, the results of which can be found in [figure 2](#_Exploratory_Analysis_1). We are going to keep the log transformation of SalePrice for the remainder of our analysis because it improved the residuals contained in the diagnostic plot output; thus, the log transformation of SalePrice decreased chances of violating any of the above mentioned assumptions for regression analysis.

Once we decided to log our response variable, we decided to look at regressions of each explanatory variable on logSalePrice as well as the output of residuals. We observed a few variables that appeared to be skewed due to outliers along with residuals that contained outliers with high leverage points and some that appeared to violate the assumption of linearity, so we decided to look into transformations. We found that the variables GrLivArea, LowQualFinSF, TotalBsmtSF, PoolArea, and MiscVal all had major issues with equality of variances, so we decided to log them and rerun the individual analysis with logSalePrice. We found that GrLivArea performed a lot better when logged as well as PoolArea and MiscVal. With the latter two variables we kept them logged, but keep in mind that both of these have very few data points so they might not meet other assumptions. Although we saw these improvements, we are not going to dwell too much here since logging the explanatory variables will not have too large of an effect on our models due to the fact that we are fitting so many variables into a linear model.

Because each model is unique, we are going to conduct further analysis as we create new models.

In terms of interpretation, we are going to look at a few key parameters. We are going to use our forward model as an example:

logSalePrice = .8455 + .0788\*OverallQual + .4588\*log(GrLivArea) …

If we were using the untransformed version of our response variable, the intercept indicates that as all other parameters are 0, estimated SalePrice = .8455. But Bbcause our response variable is transformed, we are going to focus on the variables rather than the intercept. Looking at OverallQual, as OverallQual increases by one unit, estimated SalePrice is estimated to increase by (e^.0788 = 1.0820) 8.2%, holding all other variables constant. Looking at GrLivArea, because it is also transformed we need to look at this a little differently. As GrLivArea doubles in square feet, estimated sales price is estimated to increase by $.14. You can do the same thing for all of the remaining variables. When looking at the confidence intervals, all you have to do is take each value and raise it over e. For example, we will look at OverallQual. From [figure 6](#_Forward_Selection_12) we see that the 95% confidence interval for the estimated value of OverallQual is [.070,.088]. Therefore we are going to say that as OverallQual increases by one unit, we can say with 95% confidence that SalePrice will increase by [e^.070,e^.088], or [7.25%, 9.20%]. Interpretations of the rest of the models will be conducted in the same manner.

# Analysis Question 1

## The Problem

We want to create a model that will help buyers, real estate agents, and contractors gain insight into what variables influence housing prices in Ames, Iowa. By doing this, we are going to perform three basic methods of model selection: **Forward**, **Backward**, and **Stepwise** selection.

## Forward Model

We began our model by including all qualitative and quantitative variables into a glmselect procedure. After running a procglmselect procedure with forward model selction a few times, we ended up deciding that this model ([figure 3](#_Forward_Selection_7)) was the best fit. To ensure that we do not have collinearity, we checked the VIFs and found that although Yearbuilt, logGrLivArea and OverallQual have a higher VIF than the rest, they are still pretty low with values around 2, so we kept both of these parameters in our model. Now we are going to check the assumptions on this model in particular:

**Normality:** Looking at the histogram ([figure 4](#_Forward_Selection_9)) the data appears to be normal, although the QQ-plot shows a slight curvature pattern. We are going to assume normality due the large number of data points and even if the data is not normal, we can assume that the central limit theorem will kick in.

**Equality of Variance:** Looking at the residual plot ([figure 5](#_Forward_Selection_10)) , there still seems to be a few outliers, but looking at the cook’s D and leverage, these points don’t seem to be anything to worry about, so we are going to assume equality of variance.

**Independence:** For the sake of this analysis, we are going to assume independent data.

Since all of our assumptions are met, we are going to look into the statistics of the model, which can be found in the [table](#_Table:). For information on the estimates, look to [figure 6](#_Forward_Selection_11).

## Backward Model

We began our model by including all qualitative and quantitative variables into a glmselect procedure. After running the backward selection procedure a few times, we ended up deciding that this model ([figure 7](#_Backward_Selection_3)) was the best fit. To account for collinearity, we checked variance inflation and found that non of the variables are collinear. Now we are going to check the assumptions on this model in particular:

**Normality:** From the histogram ([figure 8](#_Backward_Selection_3)) we can say that data appear to be relatively normal. Even though the QQ-plot shows evidence of a slight pattern, the sample size is large enough that we would assume the central limit theorem to kick in. For these reasons we are going to assume normality.

**Equality of Variance:** the residual plot ([figure 9](#_Backward_Selection_4)) does not look great – there are a few points in the bottom right that seem to be skewing the data. This is most likely due to the MiscVal variables, which we noted earlier were very few data points, so we are going to try logging MiscVal, which makes the residual plot more of a random cluster ([figure 10](#_Backward_Selection_5)), but when we try to fit the model again with the logged variable, it becomes more trouble than it’s worth because of the huge number of zeros. Therefore we are going to keep MiscVal in the model even though it skews the variance. Because we are only looking at one particular outlier, we are going to assume equality of variance.

**Independence:** For the sake of analysis, we are going to assume the data to be independent.

Since all of our assumptions are met, we are going to look into the statistics of the model, which can be found in the [table](#_Table:). For information on the estimates, look to [figure 11](#_Backward_Selection_6).

## Stepwise Model

We began our model by including all qualitative and quantitative variables into a glmselect procedure. After running the forward selection procedure a few times, we ended up deciding that this model ([figure 12](#_Stepwise_Selection_1)) was the best fit. We checked the variance inflation factors and although OverallQual, logGrLivArea, and YearBuilt all had higher VIF, we are not going to worry about collinearity since they are relatively small (all < 5). Now we are going to check the assumptions on this model in particular:

**Normality:** from the histogram ([figure 13](#_Stepwise_Selection_2)) the data is obviously normal. Again, the QQ-plot appears to be slightly skewed, but the sample size is large enough that the central limit theorem will kick in, so we can assume normality.

**Equality of Variance:** Although the residuals do not look terribly skewed ([figure 14](#_Stepwise_Selection_3)), there does seem to be an outlier or two that are significantly altering the model. We looked into this and found that TotalBsmtSF has a value that is much higher than its next highest value (look at this individual home, the basement area is the same as the General Living Area, so it is likely that this is a large, one-story home that was entered in as basement). For this reason, we are going to leave in the outlier and just run the analysis assuming equality of variance since the variances are otherwise relatively equal.

**Independence:** for the sake of analysis, we are going to assume independence of the data.

Since all of our assumptions are met, we are going to look into the statistics of the model, which can be found in the [table](#_Table:). For information on the estimates, look to [figure 15](#_Stepwise_Selection_4).

## Table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Adjusted R2 | AIC | CVPress | Kaggle Score |
| Forward | .8631 | -3166.34605 | 26.67304 | .17533 |
| Backward | .8701 | -3210.84885 | 25.83838 | 1.24987 |
| Stepwise | .8631 | -3166.34605 | 26.05902 | .17533 |

# Analysis Question 2

## The Problem

We would like to build the most predictive model for sale prices of homes in all of Ames, Iowa. We will use four model selection techniques to analyze the influential factors on sale prices of homes in Ames.

## The Model & Assumptions

## Comparing Competing Models

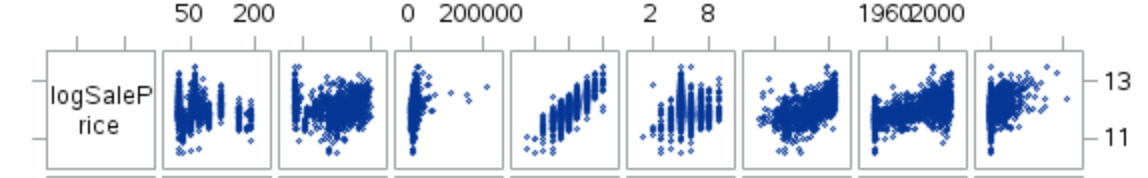
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kaggle Competition Models | Adjusted R2 | AIC | ASE | Kaggle Score |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

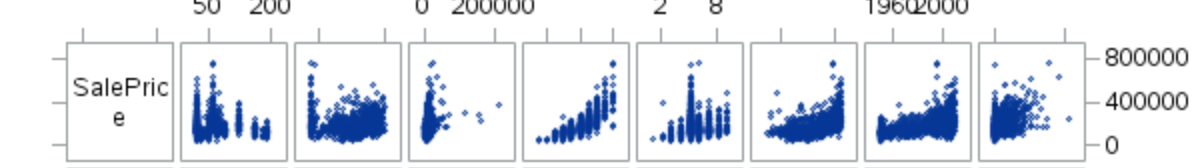
# Conclusion

Given our relatively small knowledge of statistics and model fitting, we are proud of our ability to produce some not-terrible kaggle scores. That being said, there are a few things that we realized too late that we could have done to help our models. One of the biggest, and most time consuming, problems we came across dealt with transforming our explanatory variables. Two variables in particular, MiscVal and PoolArea, had a large number of zeros and only a few actual data points. This caused issues because the data was obviously skewed, but we could not do a transformation due to high number of zeros. We tried to fix this problem by changing all of the zeros to 1s to hopefully keep the same general idea of the data, but we decided that this was more trouble than would be worth it since the process of model fitting is so rigid.

## Appendix

### Exploratory Analysis

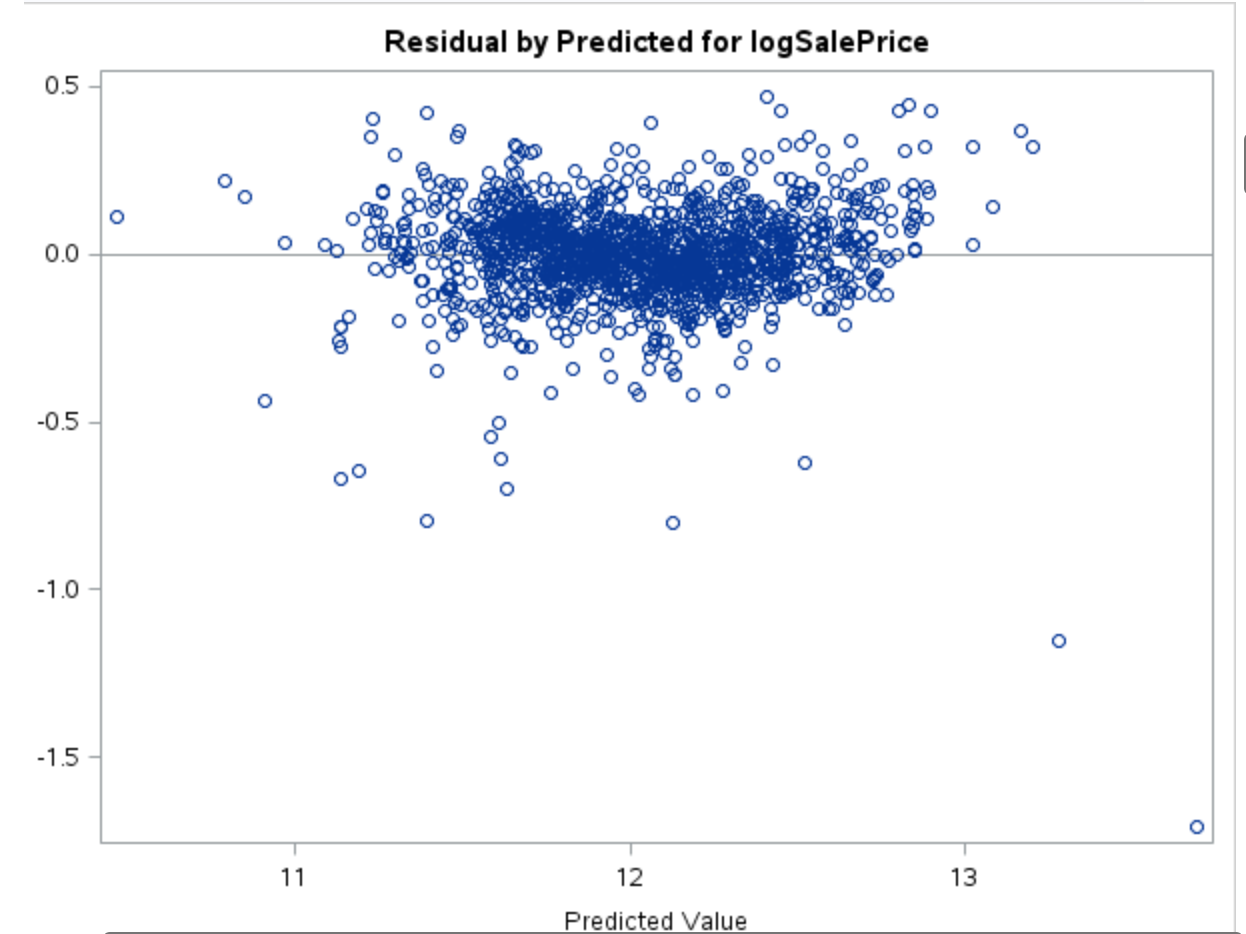
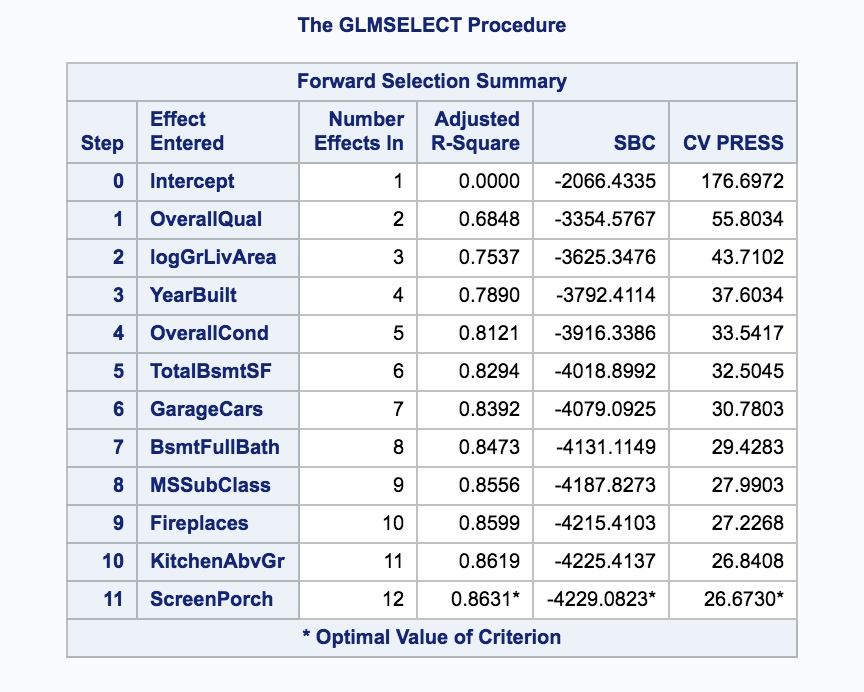
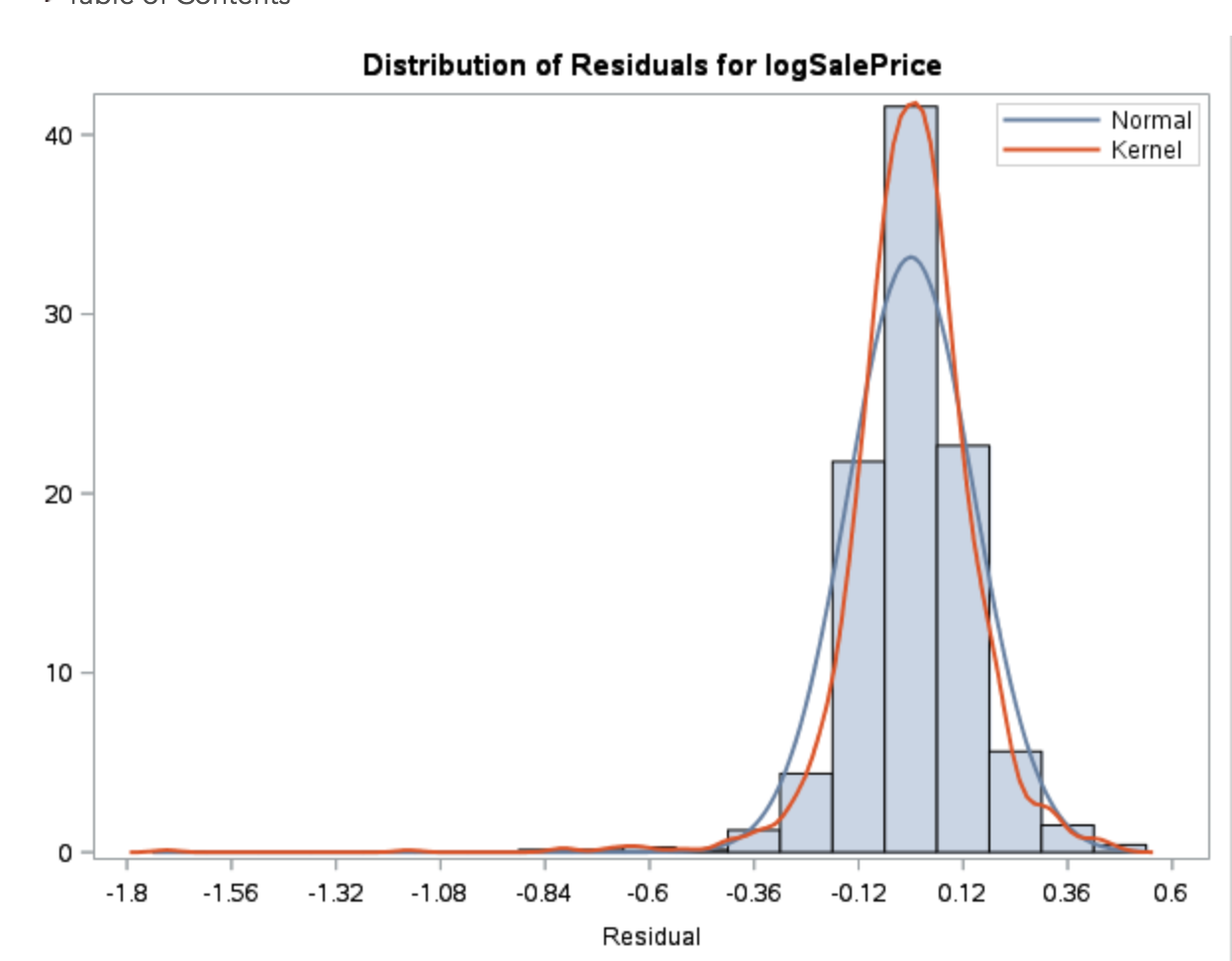
*Figure 1 Figure 2*

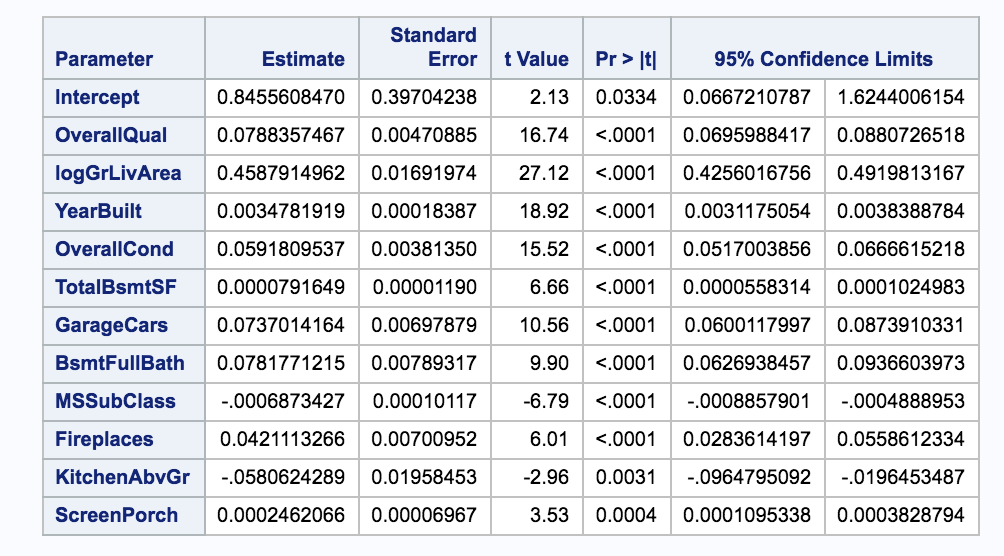


### Analysis 1

#### Forward Selection

*Figure 3 Figure 4 Figure 5*

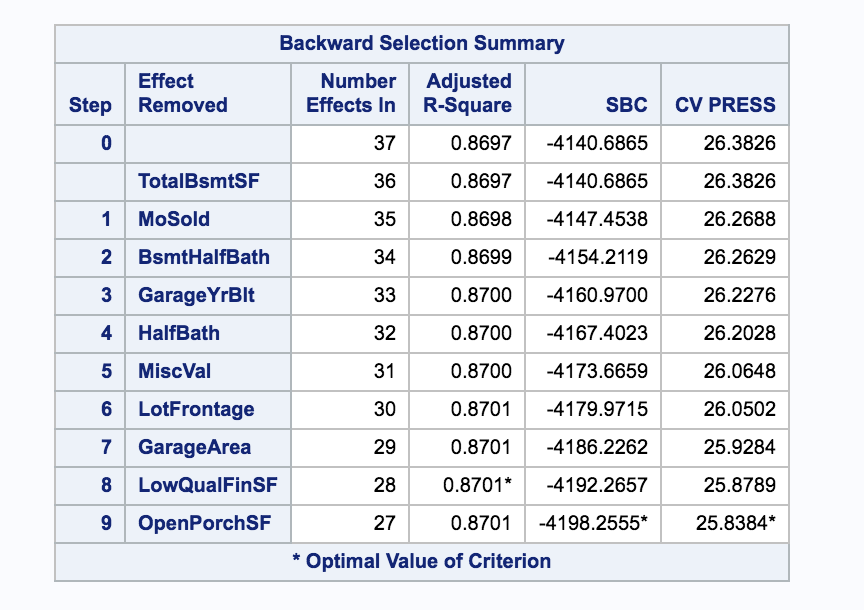
**

**

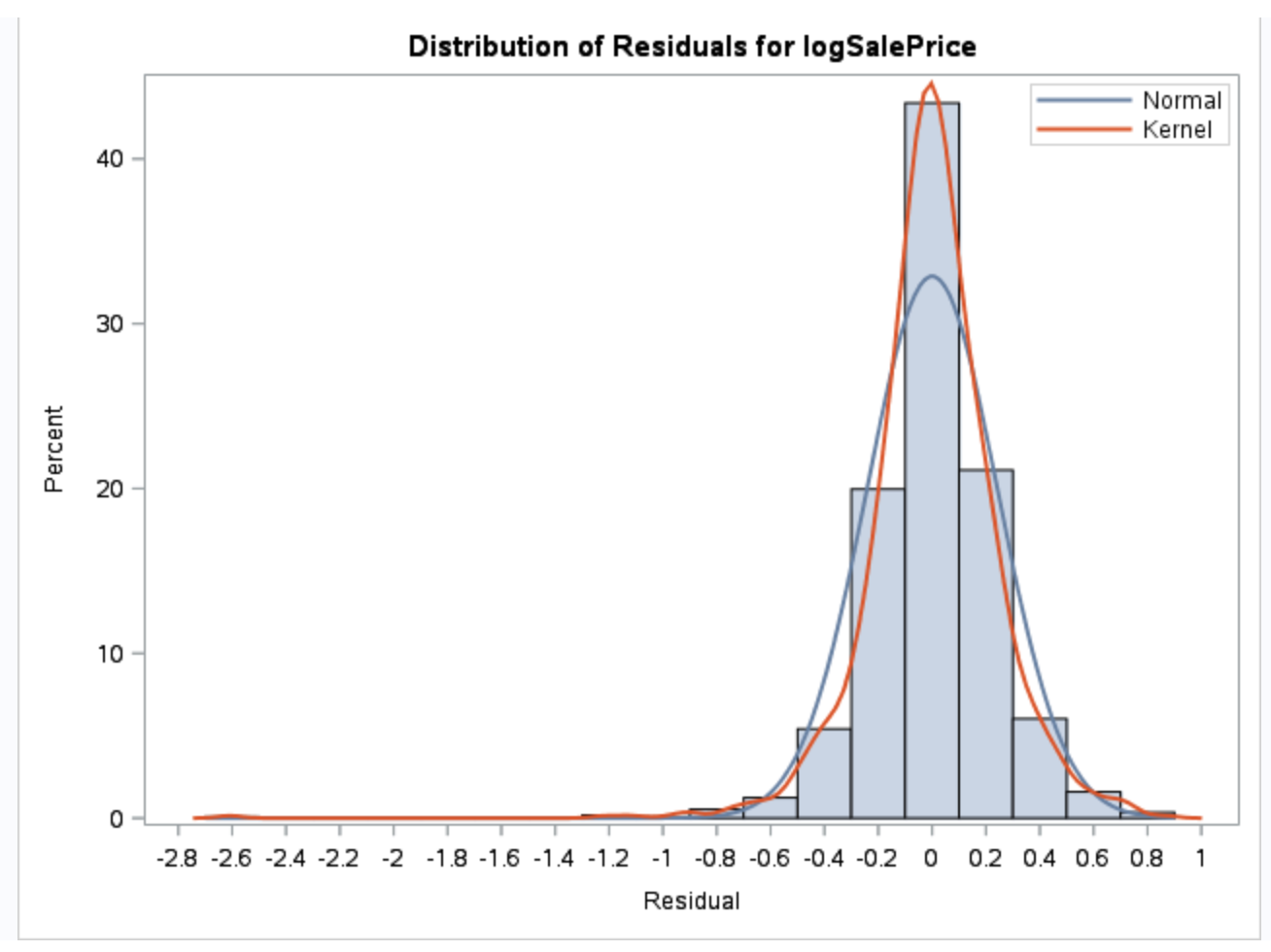
*Figure 6*

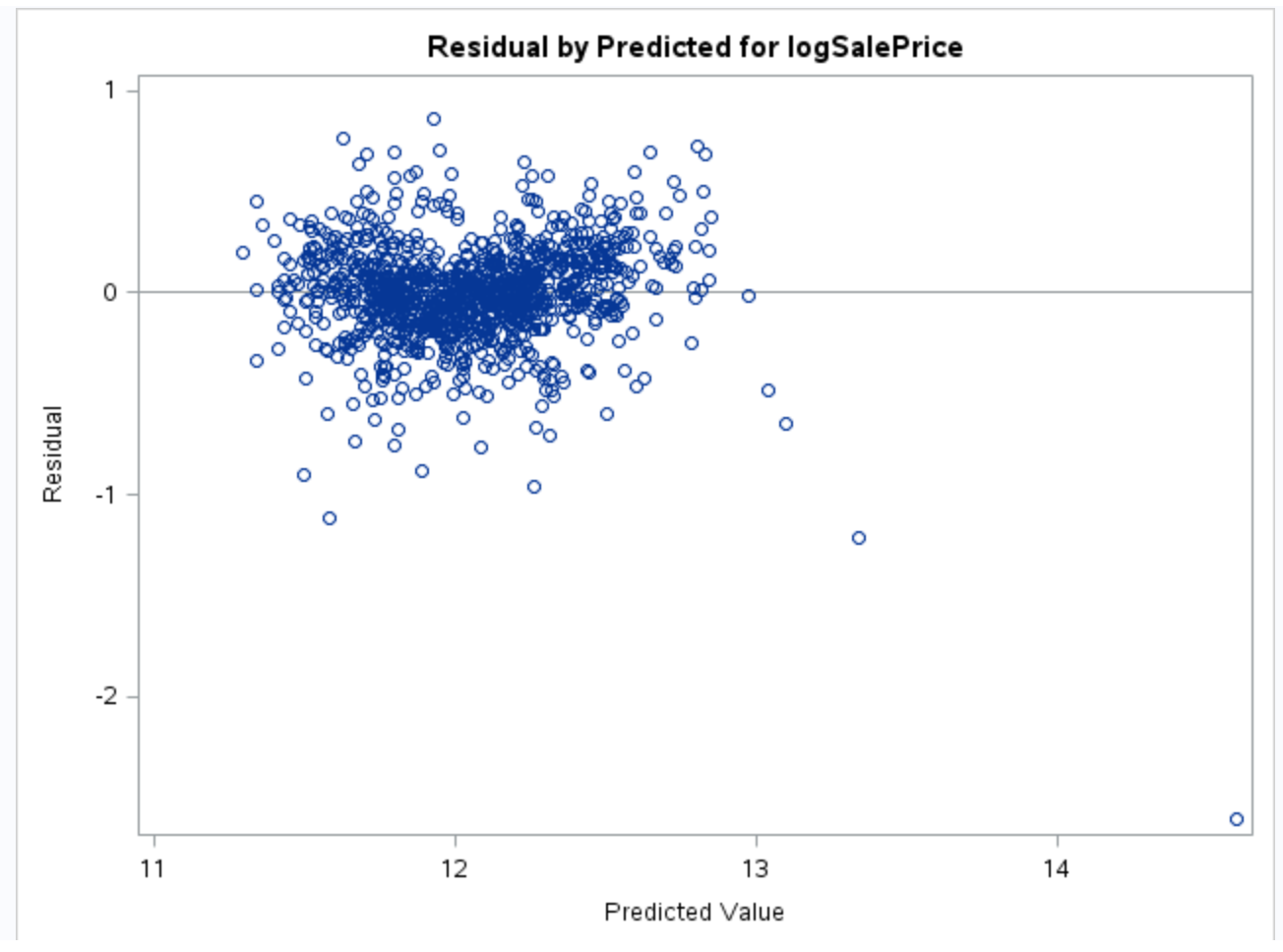
#### 

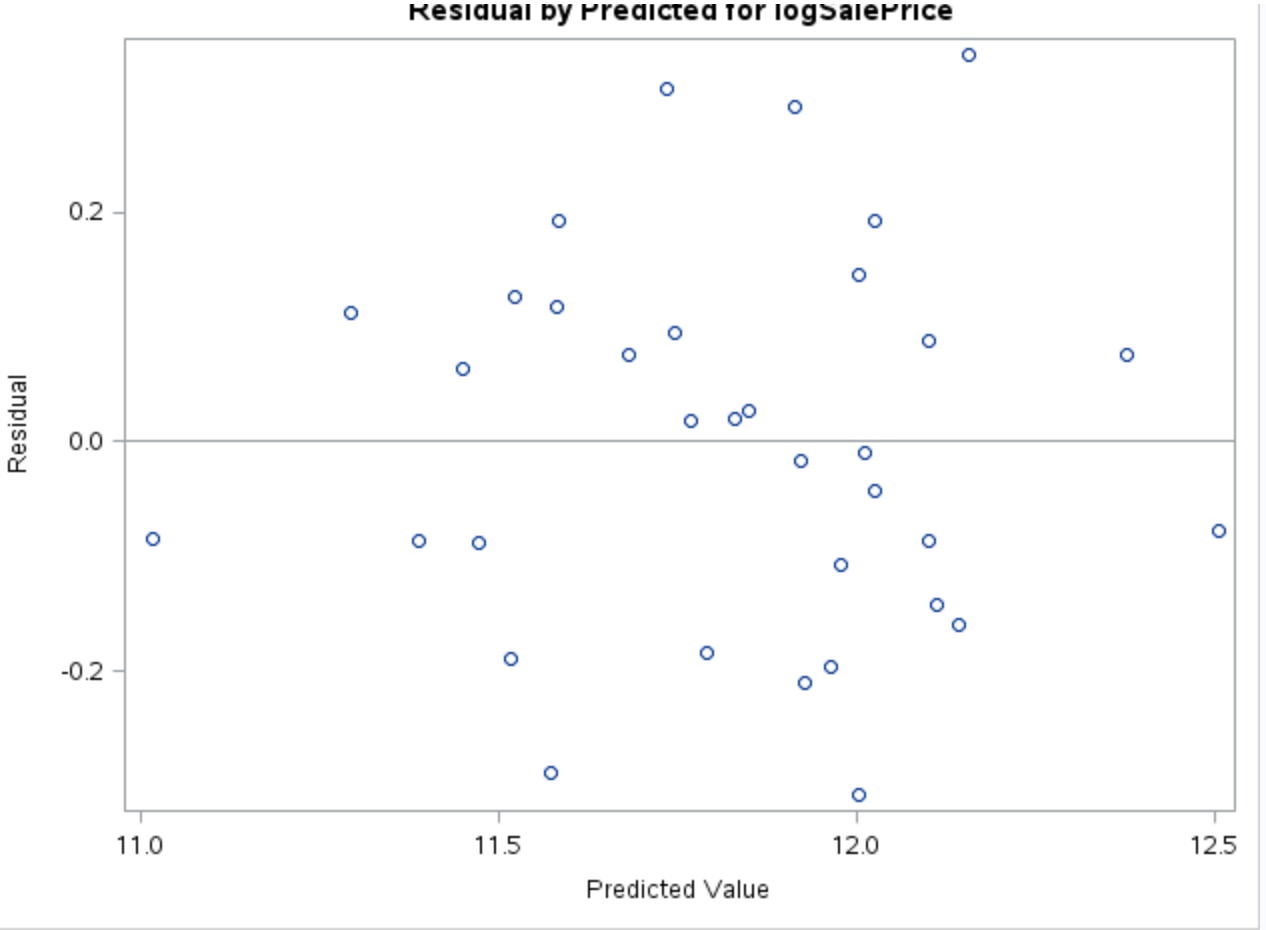
#### Backward Selection

*Figure 7*

*Figure 8 Figure 9*



**

*Figure 10 Figure 11*

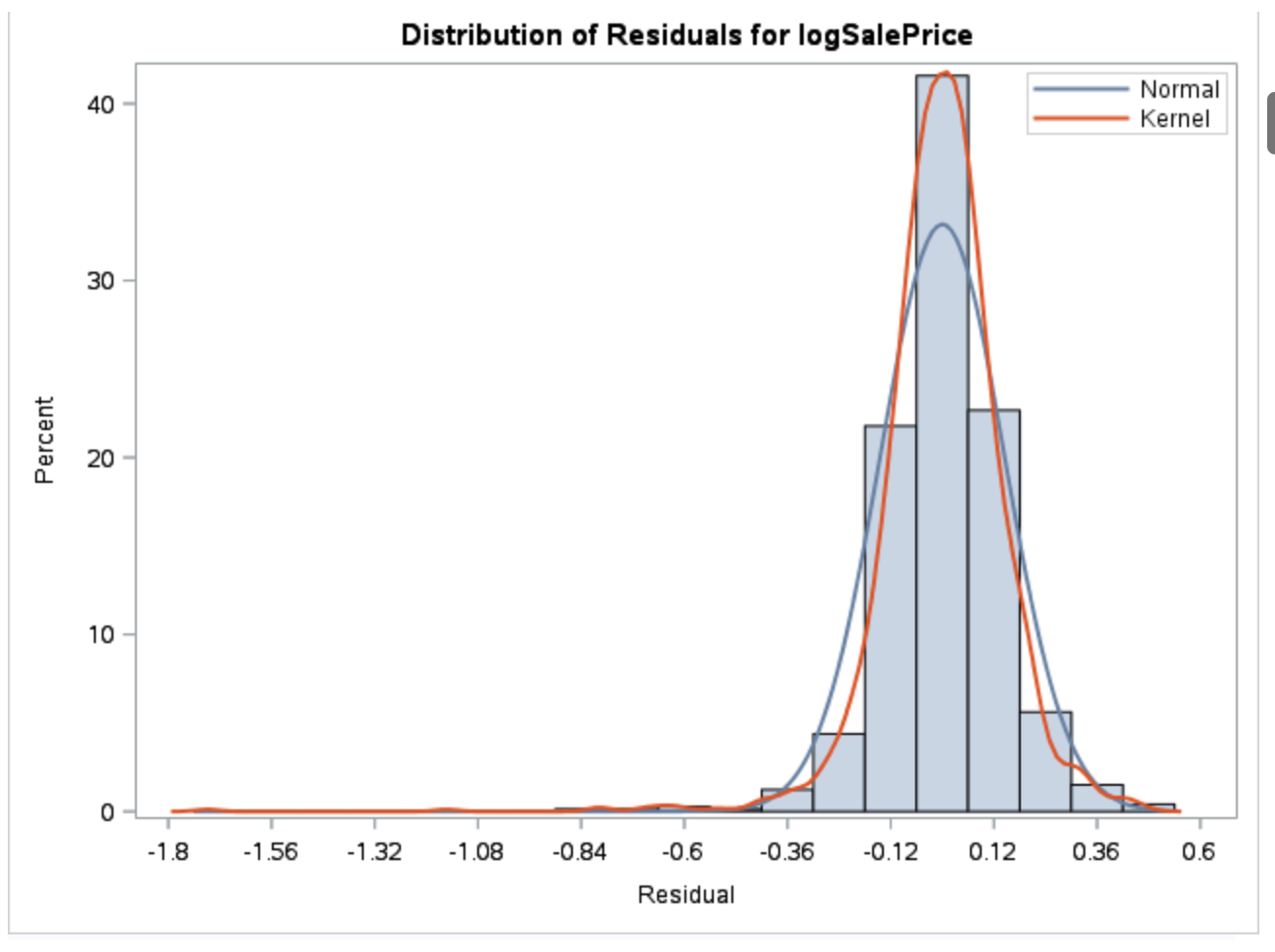
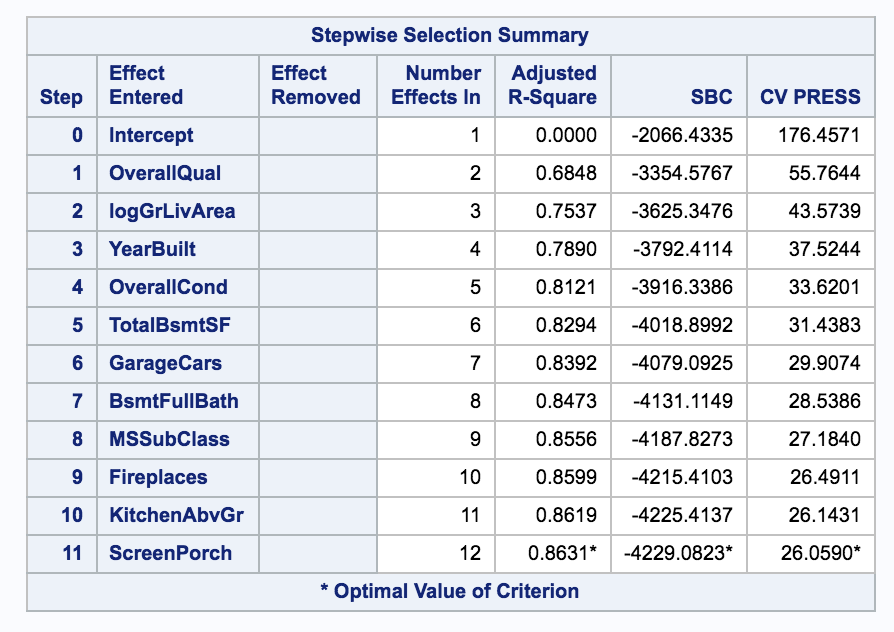
#### 

#### 

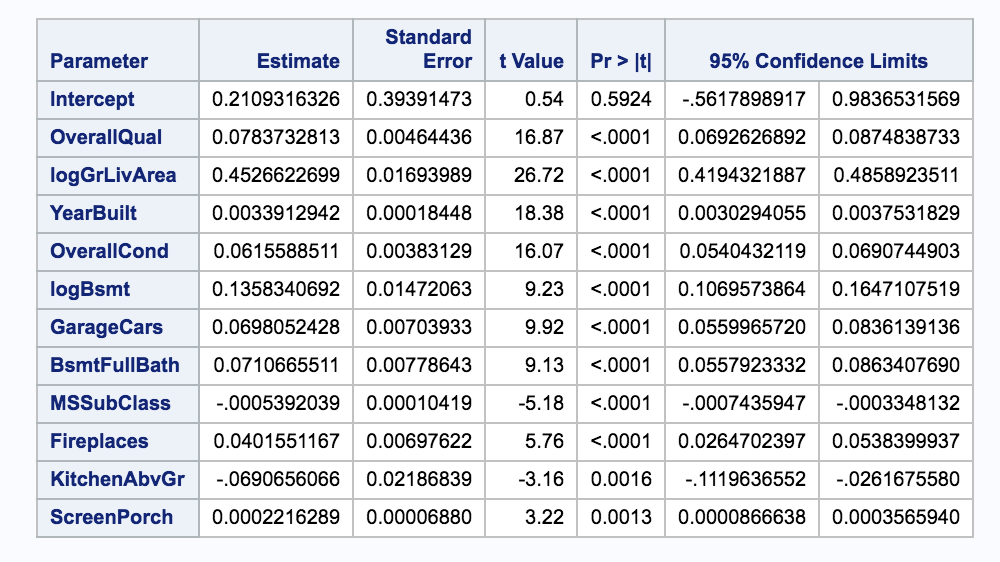
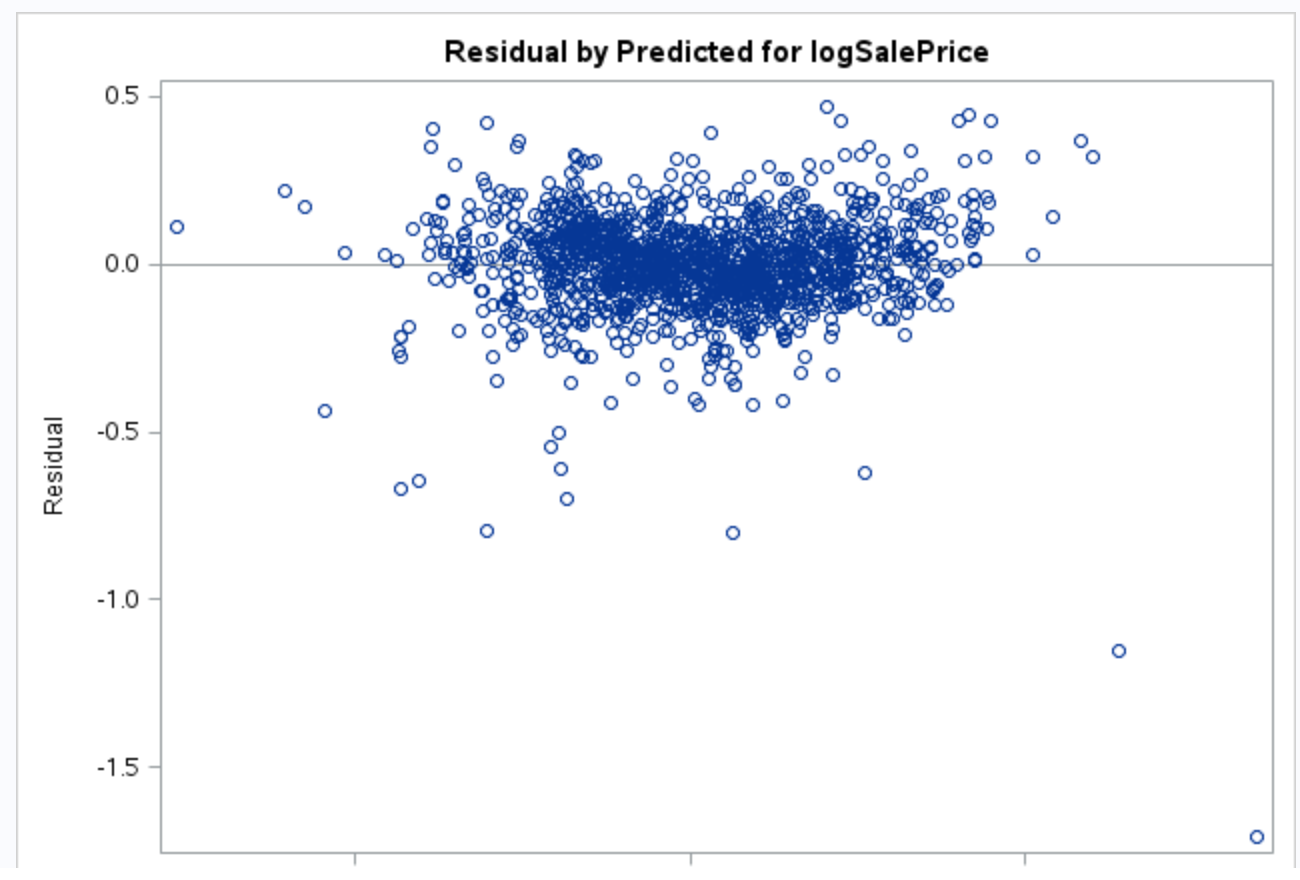
#### 

#### Stepwise Selection

*Figure 12 Figure 13*

**

*Figure 14 Figure 15*

**