MSDS-6371
House Prices Advanced Regression Techniques
Halle Purdom & Taylor Bonar

# Introduction

In order to determine a home's worth, numerous factors are taken into account like its location, condition, past repairs, etc. In this work, we are going to be looking specifically at the sale price of different homes in Ames, Iowa to answer two main questions of interest.

The first question is regarding the relationship between square footage of the living area of the house and its sale price. For this question, we will be narrowing the domain of the data to only include houses in the North Ames, Edwards and BrookSide neighborhoods. This will allow us to see if the relationship between square footage and sale price is different when these specific neighborhoods are taken into account where Century's 21 Ames sells houses at.

For the second question of interest, we developed models from the housing dataset that predict a house's sale price using select factors about the home. These were pulled from the 79 explanatory variables given in the housing dataset. Three models were developed using different variable selection techniques: forward selection model, backward selection model, and stepwise selection model. The variables for the three models were selected through automatic processes. Statistics for each model were compared to determine which model would most accurately predict the sale price of a house.

The data given for the houses in Ames, Iowa contains information on 1,460 houses. For each of these houses, there are 79 explanatory variables including components such as year remodeled, number of bedrooms, and the square footage of various rooms around the house. In addressing the first topic, the three variables of interest are sale price, living area square footage, and neighborhood. The second topic expands on that, including a unique set of variables for each model. To find out more about the data, click here: https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

# Analysis of House Prices in Iowa Ames Century 21's Target Areas

## How does the Square Footage of the Living Area Affect Sale Prices for the Houses in North Ames, Edwards, and Brookside?

### Build and Fit the Model

When initially checking the assumptions for how the living area in square footage relates to the sale price of a home, we created a scatterplot to interpret a larger picture of the initial data (Appendix:Model0:Scatter Plot). We, however, noticed that there were two distinct outliers for living area space in North Ames, Edwards and Brookside neighborhoods. These two points exceeded over 4,500 sq. ft. which was well beyond the average and median of the rest of the dataset, when excluding them. Upon examining their ability to leverage, or rather, cause disturbance to the linear regression model, we found that both were relatively low in influence

(Cook's D < 6) and a moderate leverage (between 1-2) that would affect our model's line. Regardless, we chose to exclude them from our dataset as we did not have enough data for sale prices of homes with such large living areas in the neighborhood, and wanted to focus on what an average realtor can expect.

## Checking Assumptions of Our Limited Model

### Normality

As we chose to limit the scope of our living area sq footage to be from 0 to 3,500 sq. ft., we were able to balance out the assumptions and reduce reliance on robustness. As shown in Appendix:Model 0:Residual Plots, the QQ-Plot shows that the residuals are relatively normal, but are right-skewed. To mitigate violations of normalcy, we performed several log-transformations. A log-log transformation, with or without interaction of the Neighborhood, corrects the departure from normality.

### Linear Relationship

Using Pearson's r, we can establish how strong of a linear relationship there might be between variables. The below table captures each models' linear correlation effect and how strong it is:

| Model | Pearson's r Value | Effect |
|---|---|---|
| Original Model | 0.58 | Large and positive |
| Limited Ranged Model | 0.67 | Large and positive |
| Limited Ranged linear-log | 0.6659861 | Large and positive |
| Limited Ranged log-linear | 0.6430006 | Large and positive |
| Limited Ranged log-log | 0.6655912 | Large and positive |
| Limited Ranged log-log w/ Neighbor Categorical | 0.6655912 | Large and positive |
| Limited Ranged log-log w/ Neighbor Categorical and Interaction between Neighborhood and Living Area square footage | 0.6655912 | Large and positive |

As you can observe, all the effects of association are large and positive when examining the Pearson's r, or correlation coefficient. All models meet the linear relationship assumption with some variation on our best fit.

### Equivalent Variation/Standard Deviation

While exploring the initial scatterplot of the ranged model (Appendix:Model 0), there is an expansion and contraction of standard deviation in regards to the sub-populations. With transformations, we found the log-log transformation, with or without the interaction of Neighborhood, was able to lessen this effect and comply with the assumptions much more so.

### Independence

All data provided in this dataset was provided by the Ames City Assessor's Office of house sales between 2006 and 2010. We will assume all sales are independent and are not influential to a given sale price of another.

## Comparing Competing Models

| **Model Type (All Limited from 0-3,500 sqft of Living Area)** | **Adj R^2** | **CV** |
|---|---|---|
| **Model 1: Limited Ranged Model linear-linear** $Sale\ Price\ = \beta_0 + \beta_1(Living\ Area\ Sq\ Ft)$ | .4559 | 8.011034e+08 |
| **Model 2: Limited Ranged linear-log** $Sale\ Price\ = \beta_0 + \beta_1(log(Living\ Area\ Sq\ Ft))$ | .4421 | 8.199520e+08 |
| **Model 3: Limited Ranged log-linear** $log(Sale\ Price)\ = \beta_0 + \beta_1(Living\ Area\ Sq\ Ft)$ | .4119 | 4.432823e-02 |
| **Model 4: Limited Ranged log-log** $log(Sale\ Price)\ = \beta_0 + \beta_1(log(Living\ Area\ Sq\ Ft))$ | .4415 | 4.207065e-02 |
| **Model 5: Limited Ranged log-log w/ Neighbor Categorical** $log(Sale\ Price)\ = \beta_0 + \beta_1(log(Living\ Area\ Sq\ Ft)) + \beta_2 Neighborhood$ | .5002 | 3.794933e-02 |
| **Model 6: Limited Ranged log-log w/ Neighbor Categorical w/ Interaction Variable** $log(Sale\ Price)\ = \beta_0 + \beta_1(log(Living\ Area\ Sq\ Ft)) + \beta_2 Neighborhood + \beta_3 log(LivingAreaSqFt)\ *\ Neighborhood$ | 0.5216 | 3.660911e-02 |

When comparing the various model's R^2 and CV values, it was discovered that the 5th model with the limited range of the Living Area, log-log transformation, and the inclusion of the neighbor interaction term, is able to fit the observed data much better.

Parameters Estimates and Confidence Intervals for Model 6

| | lSalePrice | | | | | |
|---|---|---|---|---|---|---|
| *Predictors* | *Estimates* | *Std. Error* | *CI* | *t-value* | *p-value* | *df* |
| (Intercept) | 5.91 | 0.50 | 4.94 – 6.89 | 11.91 | **<0.001** | 375.00 |
| lGrLivArea | 0.82 | 0.07 | 0.68 – 0.96 | 11.63 | **<0.001** | 375.00 |
| Neighborhood [Edwards] | 1.01 | 0.70 | -0.36 – 2.38 | 1.45 | 0.149 | 375.00 |
| Neighborhood [NAmes] | 2.58 | 0.59 | 1.42 – 3.74 | 4.37 | **<0.001** | 375.00 |
| lGrLivArea * Neighborhood [Edwards] | -0.15 | 0.10 | -0.34 – 0.05 | -1.48 | 0.139 | 375.00 |
| lGrLivArea * Neighborhood [NAmes] | -0.35 | 0.08 | -0.51 – -0.18 | -4.15 | **<0.001** | 375.00 |
| Observations | 381 | | | | | |
| $R^2$ / $R^2$ adjusted | 0.528 / 0.522 | | | | | |

Each neighborhood's sale price can be predicted roughly with the following equations:

- $log(Sale\ Price) = \beta_0 + \beta_1(log(Living\ Area\ Sq\ Ft)) + \beta_2 Edwards + \beta_3 LivingAreaSqFt * Edwa$
- $log(Sale\ Price) = \beta_0 + \beta_1(log(Living\ Area\ Sq\ Ft)) + \beta_2 North\ Ames + \beta_3 LivingAreaSqFt * No$
- $log(Sale\ Price) = \beta_0 + \beta_1(log(Living\ Area\ in\ Square\ Footage))$

## Conclusion

In conclusion, with our linear regression model taking into account the log value of the living area square footage and the neighborhood of the house being sold, we can interpret that we can roughly estimate a sale price with the following equations:

Model 5 (all Neighborhoods have same slope)

$$log(Sale\ Price) = \beta_0 + \beta_1(log(Living\ Area\ in\ Square\ Footage)) + \beta_2 Neighborhood$$

Model 6 (Neighborhoods have different slopes)

$$log(Sale\ Price) = \beta_0 + \beta_1(log(Living\ Area\ Sq\ Ft)) + \beta_2 Neighborhood + \beta_3 log(LivingAreaSqFt) * Neighborhood$$

Although this equation may be useful for a general estimate, it should be taken lightly as we only explained potentially 66.5% of the relationship between sale price and the living area square footage for houses sold in North Ames, Edwards, and Brookside neighborhoods.

The data suggests each 100 square footage increase in living area for homes in Brookside is associated with a increase in the median Sale Price by a factor of $100 * e^{\beta_0}$ = **36,870.62**. In Edwards, this increase is 36,870.62 + $100 * e^{-0.01}$= **99.00498,** and in North Ames, this increase is 36,870.62 + $100 * e^{0.13}$= **113.8828.**

Without taking neighborhood into account (Model 4), a doubling of living area square footage is associated with a $2^{\beta_1} = 2^{\beta_{0.56824}}$2^β1 = 2^(0.56824) = 1.482714 multiplicative change in the median Sale Price. We are 95% confident the multiplicative increase in the mean Sale Price after a doubling in square footage is ($2^{\beta_{0.5010272}}$, $2^{\beta_{0.6354552}}$) = (1.415221 ,1.553428).

Taking the three different neighborhoods into account, this relationship changes slightly. For Brookside, a doubling of living area square footage is associated with a $2^{\beta_1} = 2^{\beta_{0.81965}}$ = **1.764978** multiplicative change in the median Sale Price. In Edwards, this multiplicative change is 1.764978 - ($2^{\beta_1} = 2^{\beta_{-0.29998}}$= 0.8122637) = **0.9527143**. In North Ames this multiplicative change is 1.764978 - ($2^{\beta_1} = 2^{\beta_{-0.34662}}$= 0.7864244) = **0.9785536**. See Model 6 table for confidence interval values.

## Building the most predictive model for sales prices of homes in all of Ames, Iowa.

### Model Selection

In order to build a model that will most closely predict the sales prices of homes, 79 explanatory variables are initially given. In this work, the predictive models use multiple linear regression (MLR) meaning rather than building a model from 79 explanatory variables, we narrow the range down to only numeric variables- 37 explanatory variables. For reference, explanatory variables are factors describing the house such as year built, garage square footage, or pool area. The response variable in this case, or what the model predicts, is the sale price of the house. In the model selection process, we are continuing to use the data cleaning done in analysis one, where sale price and living space square footage were both log transformed and the houses above 3500 living area square footage were removed. For this reason, our models below will only apply to houses with living area square footage under 3,500 square feet.

In situations where the amount of explanatory variables is too large to parse through one-by-one, automatic variable selection is often used to find a smaller set of variables to look further into for the most predictive model. The three models developed to predict the sale prices of homes in Ames were done with automatic variable selection techniques. This means variables are chosen automatically through fixed processes rather than being manually chosen for the MLR model. In this case, we ran all 37 variables through different types of automatic variable selection processes to narrow down which variables would be best for our model: forward selection, backward selection, and stepwise selection.

Because some of the numeric explanatory variables we ran through the model had missing values, we chose to set those values to zero rather than remove them. Both methods were tested and setting null values to zero gave us the most predictive model for the data.

### Forward Selection Model

$$log(SalePrice) = \beta_0 + \beta_1 MSSubClass + \beta_2 LotArea + \beta_3 OverallQual + \beta_4 OverallCond + \beta_5 YearBuilt + \beta_6 YearRemodAdd + \beta_7 BsmtFinSF1 + \beta_8 TotalBsmtSF + \beta_9 X1stFlrSF + \beta_{10} X2ndFlrSF + \beta_{11} BsmtFullBath + \beta_{12} KitchenAbvGr + \beta_{13} Fireplaces + \beta_{14} GarageCars + \beta_{15} GarageArea + \beta_{16} WoodDeckSF + \beta_{17} ScreenPorch + \beta_{18} YrSold + \beta_{19} log(GrLivArea)$$

The forward model selected 19 explanatory variables to best predict the sale price of a home. The MLR created from those variables is the equation above, with estimates $\beta_0 - \beta_{19}$ (Appendix:Forward Selection). Internal cross-validation of the data was set as the indicator to stop adding variables to the model. When the cv press statistic hit the lowest value of 23.22090, the forward selection process stopped and produced the above model. The model has an adjusted R-squared of 0.8986 The RMSE was 0.14223 which demonstrates a relatively close fit and accuracy of the model, but is not the favored one.

### Backward Selection Model

$$log(SalePrice) = \beta_0 + \beta_1 MSSubClass + \beta_2 LotFrontage + \beta_3 LotArea + \beta_4 OverallQual + \beta_5 OverallCond + \beta_6 YearBuilt + \beta_7 YearRemodAdd + \beta_8 BsmtFinSF1 + \beta_9 BsmtFinSF2 + \beta_{10} BsmtUnfSF + \beta_{10} X1stFlrSF + \beta_{11} X2ndFlBsmtrSF + \beta_{12} FullBath + \beta_{13} BedroomAbvGr + \beta_{14} KitchenAbvGr + \beta_{15} TotRmsAbvGrd + \beta_{16} Fireplaces + \beta_{17} GarageCars + \beta_{18} GarageArea + \beta_{19} WoodDeckSF + \beta_{20} EnclosedPorch + \beta_{21} ScreenPorch + \beta_{22} YrSold + \beta_{23} log(GrLivArea)$$

The backward model selected 23 explanatory variables to best predict the sale price of a home. The MLR created from those variables is the equation above, with estimates $\beta_0 - \beta_{23}$ (Appendix:Backward Selection). Internal cross-validation of the data was set as the indicator to stop adding variables to the model. When the cv press statistic hit the lowest value of 23.23167, the forward selection process stopped and produced the above model. The model has an adjusted R-squared of 0.8994. The RMSE was 0.14185 which also demonstrates a relatively close fit and accuracy of the model, but is not the favored one.

### Stepwise Selection Model

$$log(SalePrice) = \beta_0 + \beta_1 MSSubClass + \beta_2 LotArea + \beta_3 OverallQual + \beta_4 OverallCond +$$

$$\beta_5 YearBuilt + \beta_6 YearRemodAdd + \beta_7 BsmtFinSF1 + \beta_8 TotalBsmtSF + \beta_9 BsmtFullBath + \beta_{10} KitchenAbvGr + \beta_{11} Fireplaces + \beta_{12} GarageCars + \beta_{13} GarageArea + \beta_{14} WoodDeckSF + \beta_{15} ScreenPorch + \beta_{16} log(GrLivArea)$$

The stepwise model selected 16 explanatory variables to best predict the sale price of a home. The MLR created from those variables is the equation above, with estimates $\beta_0 - \beta_{16}$ (Appendix:StepwiseSelection). Internal cross-validation of the data was set as the indicator to stop adding variables to the model. When the cv press statistic hit the lowest value of 23.52714, the forward selection process stopped and produced the above model. The model has an adjusted R-squared of 0.8967. The RMSE was 0.14064 which demonstrates a relatively close fit and accuracy of the model, but is the favored fit out of the three automatic variable selection models.

## Checking the Assumptions

### Residual Plots & Influential Point Analysis

All 3 models for the automatic variable selection had relatively low Cook's D values less than .08 and were not influential according to the residual leverage plot (Appendix:Automatic Variable Selection Models) with values less than 0.5. No outliers had significant influence nor leverage on the regression lines.

## Comparing Competing Models

| Predictive Models | Adjusted R squared | CV Press | Kaggle Score |
|---|---|---|---|
| **Forward** | 0.8986 | 23.22090 | 0.14223 |
| **Backward** | 0.8994 | 23.23167 | 0.14185 |
| **Stepwise** | 0.8967 | 23.52714 | 0.14064 |

## Conclusion

When comparing the three automatic selection variable models, we found our stepwise model was the most accurate out of the three (adjusted R^2 of 89.67 and CV Press of 23.52714). Furthermore, when comparing the RSME generated by the inclusion of the test dataset, we found the RSME to be the lowest, demonstrating the model's fit to be the favored one. We hypothesized that since the stepwise model has overall less variables, it was able to exclude some noise that that backward model introduced with more variables.

As we are no experts within this field, we ideally would like to confer with Ames Century 21 to further refine the model. This would include discussion of the different explanatory variables in

their point of view as their experience over the years will help enlighten potential patterns and significant variables to include into the model.

# Appendix

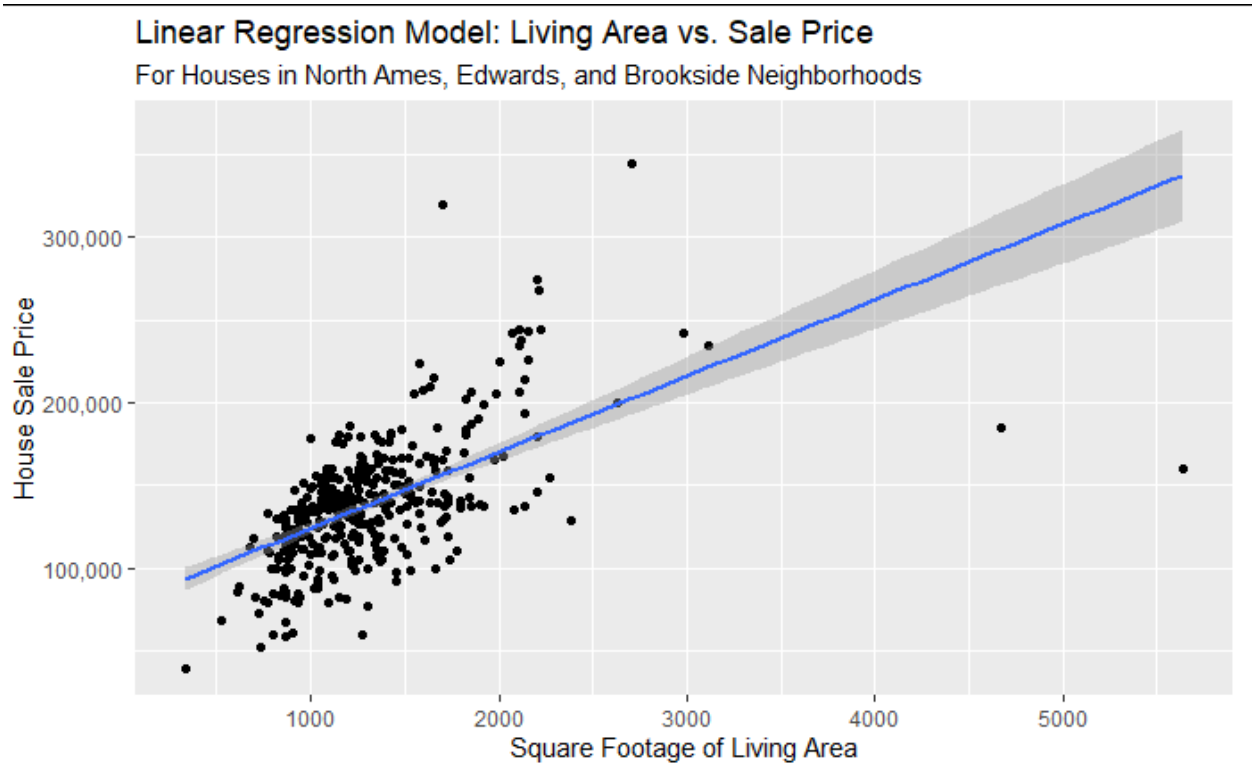
## Linear Regression Model Assumptions

1. There is a normally distributed sub-population of responses for each value of the explanatory variable (Normalcy)
2. The means of the sub-populations fall on a straight line function of the explanatory variable (Linear Relationship)
    a. Appears to be a positive upward trend. Additionally, Adjusted R^2 shows a medium positive linear trend between SalePrice vs GrLivArea (Adjusted R^2 =.34)
    b. Only 34% of the variance of the Sale Price is being explained by the GrLivArea by initial models
    c. Log transformations does not increase adjusted R^2 values
    d. Restriction of R^2 does
3. The sub-population standard deviation are all equal (to $\sigma$) (Equivalent Variation)
    a. Note: Widening of trend line shows potential deviation from equivalent variation assumption.
4. The selection of an observation from any of the sub-populations is independent of the selection of any other observation (Independence)
    a. Data was gathered from Assessor's office as a data dump from their record system in Ames, Iowa, between 2006 and 2010, assuming there are no sales for the same house between this 4 year period.

## Model 0: linear-linear Transformation

w/o Range Limitations on GrLivArea / Living Area square footage Statistical Images

Scatter Plot




Coefficient Table

| | SalePrice | | | | |
|---|---|---|---|---|---|
| *Predictors* | *Estimates* | *Std. Error* | *CI* | *t-value* | *p-value* |
| Intercept | 78205.58 | 4536.05 | 69286.74 – 87124.41 | 17.24 | **<0.001** |
| Sq. Ft. of Living Area | 45.98 | 3.27 | 39.56 – 52.40 | 14.08 | **<0.001** |
| Observations | 383 | | | | |
| $R^2$ / $R^2$ adjusted | 0.342 / 0.341 | | | | |

**Summary Statistics of Linear Model; Confidence Interval; CV Press Statistics**

```
Call:
lm(formula = SalePrice ~ GrLivArea, data = train_filtered)

Residuals:
    Min      1Q  Median      3Q     Max
-177619  -17918     919   15227  163722

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 78205.578   4536.054   17.24   <2e-16 ***
GrLivArea      45.979      3.265   14.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30980 on 381 degrees of freedom
Multiple R-squared:  0.3423,    Adjusted R-squared:  0.3406
F-statistic: 198.3 on 1 and 381 DF,  p-value: < 2.2e-16

                  2.5 %      97.5 %
(Intercept) 69286.74463 87124.41142
GrLivArea      39.55885    52.39907
```

## Model 1: linear-linear

With Range Limitations on GrLivArea / Living Area of Square Footage Statistical Images

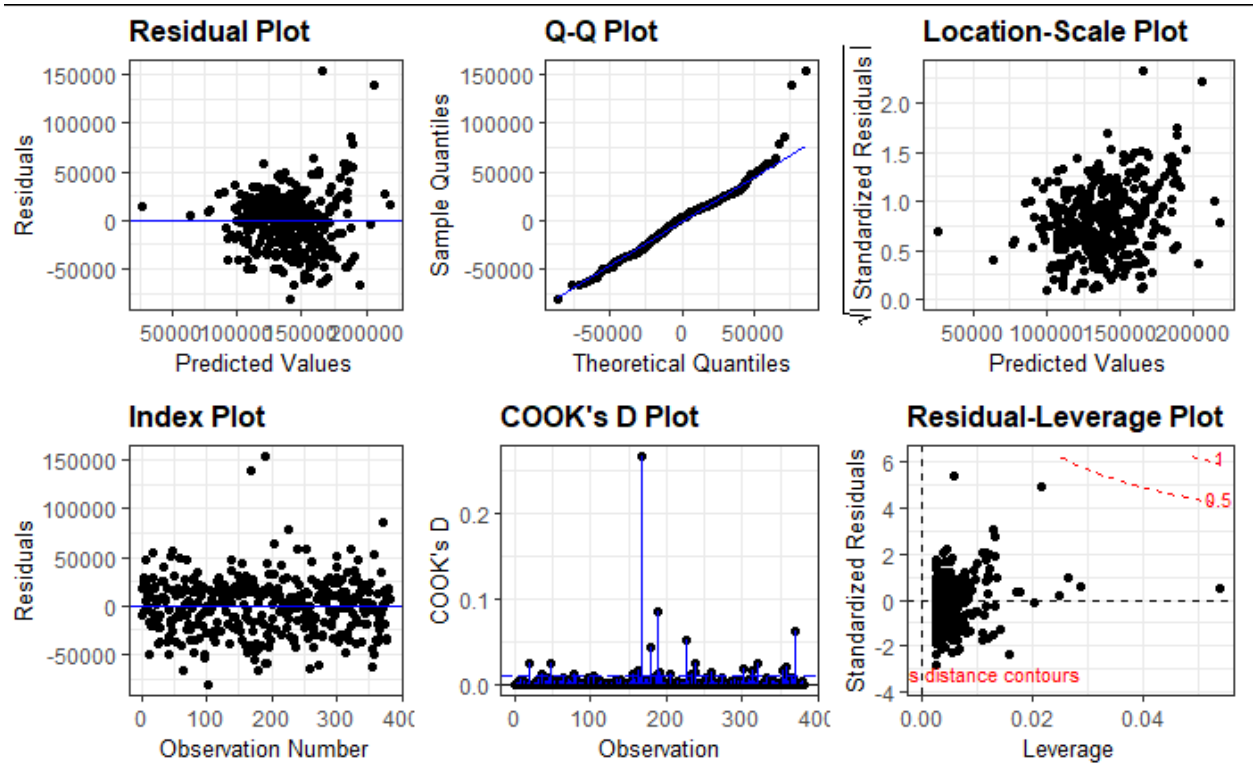
Coefficient Table

| | | | SalePrice | | |
|---|---|---|---|---|---|
| *Predictors* | *Estimates* | *Std. Error* | *CI* | *t-value* | *p-value* |
| Intercept | 54415.16 | 4888.35 | 44803.48 – 64026.84 | 11.13 | **<0.001** |
| Sq. Ft. of Living Area | 65.13 | 3.64 | 57.96 – 72.29 | 17.87 | **<0.001** |
| Observations | 381 | | | | |
| $R^2$ / $R^2$ adjusted | 0.457 / 0.456 | | | | |

Summary Statistics of Linear Model; Confidence Interval; CV Press Statistics

```
Call:
lm(formula = SalePrice ~ GrLivArea, data = train_filtered_ranged)

Residuals:
   Min     1Q Median     3Q    Max
-80420 -19186   2318  17146 154997

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 54415.162   4888.348   11.13   <2e-16 ***
GrLivArea      65.128      3.644   17.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28150 on 379 degrees of freedom
Multiple R-squared:  0.4573,    Adjusted R-squared:  0.4559
F-statistic: 319.3 on 1 and 379 DF,  p-value: < 2.2e-16

                  2.5 %      97.5 %
(Intercept) 44803.48173 64026.84149
GrLivArea      57.96214    72.29407
          CV           AIC          AICc           BIC        AdjR2
8.011034e+08 7.810980e+03 7.811043e+03 7.822808e+03 4.558573e-01
```

Residual Plots



## Model 2: Linear-Log Transformation

Range Limitations on GrLivArea / Living Area of Square Footage  Statistical Images

Coefficient Table

| | | SalePrice | | | |
|---|---|---|---|---|---|
| *Predictors* | *Estimates* | *Std. Error* | *CI* | *t-value* | *p-value* |
| Intercept | -474672.63 | 35273.68 | -544029.26 – -405316.00 | -13.46 | **<0.001** |
| Sq. Ft. of Living Area | 86131.79 | 4955.60 | 76387.87 – 95875.71 | 17.38 | **<0.001** |
| Observations | 381 | | | | |
| $R^2$ / $R^2$ adjusted | 0.444 / 0.442 | | | | |

Summary Statistics of Linear Model; Confidence Interval; CV Press Statistics

```
Call:
lm(formula = SalePrice ~ lGrLivArea, data = train_log)

Residuals:
   Min     1Q Median     3Q    Max
-81298 -18915   2866  16085 154093

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -474673      35274  -13.46   <2e-16 ***
lGrLivArea     86132       4956   17.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28510 on 379 degrees of freedom
Multiple R-squared:  0.4435,    Adjusted R-squared:  0.4421
F-statistic: 302.1 on 1 and 379 DF,  p-value: < 2.2e-16

                 2.5 %     97.5 %
(Intercept) -544029.26 -405316.00
lGrLivArea    76387.87   95875.71
          CV           AIC          AICc           BIC        AdjR2
8.199520e+08 7.820514e+03 7.820577e+03 7.832342e+03 4.420693e-01
```

Residual Graph

Linear Regression Model

## Model 3: Log-Linear Transformation

Range Limitations on GrLivArea / Living Area of Square Footage  Statistical Images

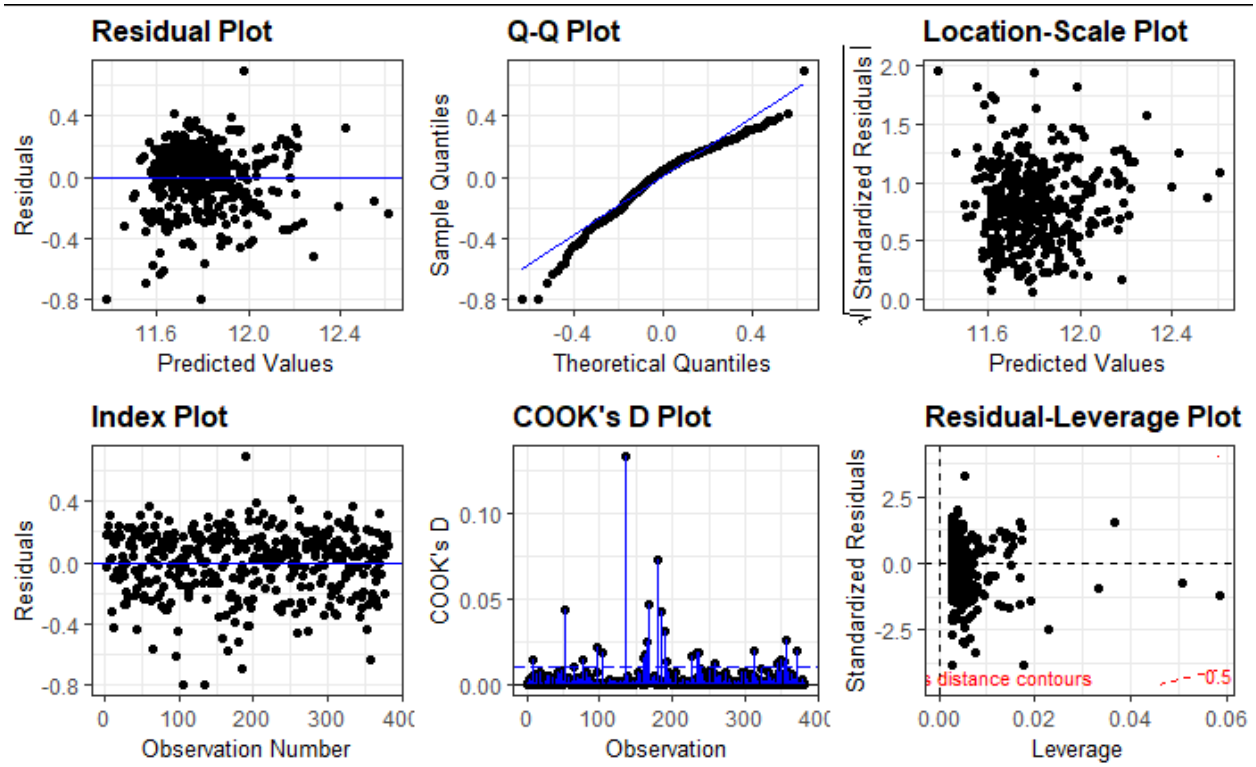
Coefficient Table

| | lSalePrice | | | | |
|---|---|---|---|---|---|
| *Predictors* | *Estimates* | *Std. Error* | *CI* | *t-value* | *p-value* |
| (Intercept) | 7.41 | 0.25 | 6.91 – 7.91 | 29.30 | **<0.001** |
| lGrLivArea | 0.62 | 0.04 | 0.55 – 0.69 | 17.36 | **<0.001** |
| Observations | 381 | | | | |
| $R^2$ / $R^2$ adjusted | 0.443 / 0.442 | | | | |

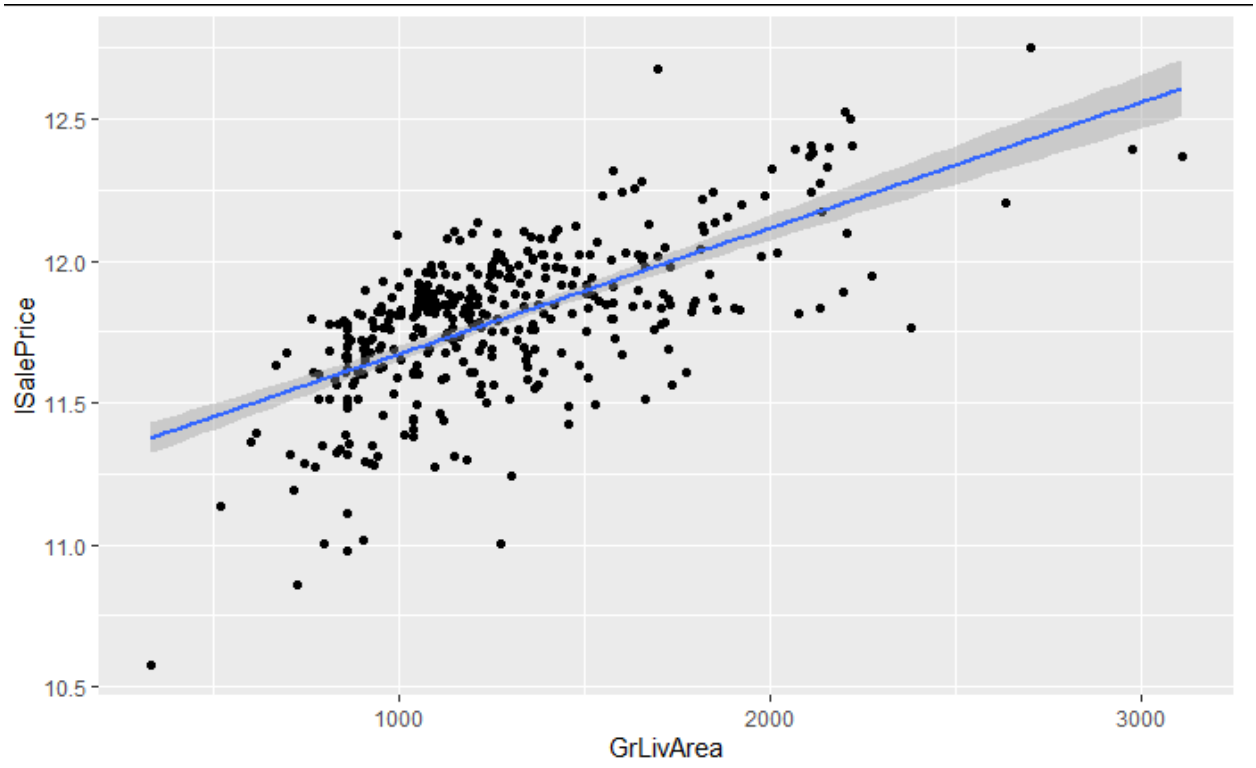
Summary Statistics of Linear Model; Confidence Interval; CV Press Statistics

```
Call:
lm(formula = lSalePrice ~ GrLivArea, data = train_log)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79801 -0.12784  0.04049  0.14489  0.69375

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.123e+01  3.642e-02  308.32   <2e-16 ***
GrLivArea   4.438e-04  2.715e-05   16.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2097 on 379 degrees of freedom
Multiple R-squared:  0.4134,    Adjusted R-squared:  0.4119
F-statistic: 267.2 on 1 and 379 DF,  p-value: < 2.2e-16

                   2.5 %       97.5 %
(Intercept) 1.115716e+01 1.130037e+01
GrLivArea   3.904036e-04 4.971779e-04
           CV           AIC           AICc           BIC          AdjR2
 4.432823e-02 -1.186170e+03 -1.186106e+03 -1.174341e+03  4.119021e-01
```

Residual Graphs




Linear Regression Model



## Model 4 Log-Log Transformation

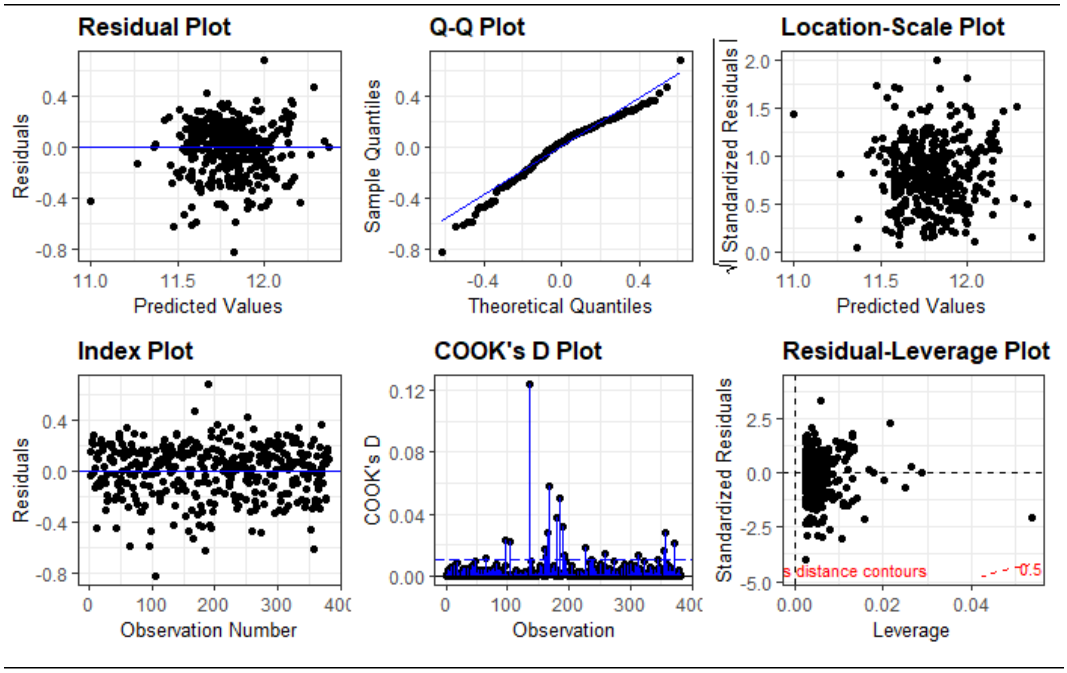
### Range Limitations on GrLivArea / Living Area of Square Footage Statistical Images

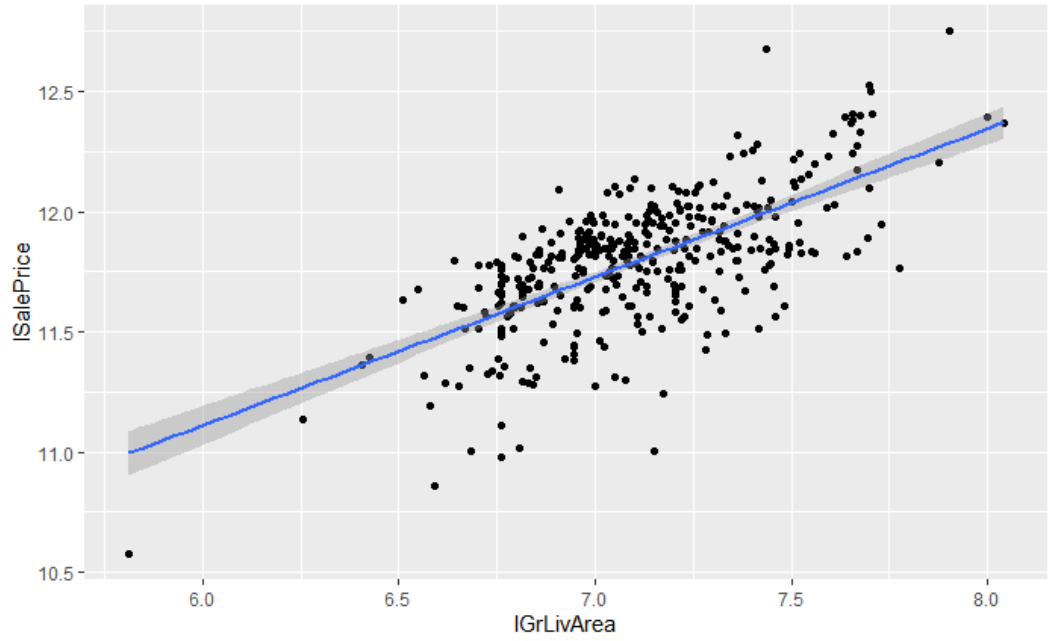
Coefficient Table

| | lSalePrice | | | | |
|---|---|---|---|---|---|
| *Predictors* | *Estimates* | *Std. Error* | *CI* | *t-value* | *p-value* |
| (Intercept) | 7.41 | 0.25 | 6.91 – 7.91 | 29.30 | **<0.001** |
| lGrLivArea | 0.62 | 0.04 | 0.55 – 0.69 | 17.36 | **<0.001** |
| Observations | 381 | | | | |
| $R^2$ / $R^2$ adjusted | 0.443 / 0.442 | | | | |

Summary Statistics of Linear Model; Confidence Interval; CV Press Statistics

```
Call:
lm(formula = lSalePrice ~ lGrLivArea, data = train_log)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81988 -0.12331  0.03752  0.13699  0.67784

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.41038    0.25290   29.30   <2e-16 ***
lGrLivArea   0.61688    0.03553   17.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2044 on 379 degrees of freedom
Multiple R-squared:  0.443,     Adjusted R-squared:  0.4415
F-statistic: 301.4 on 1 and 379 DF,  p-value: < 2.2e-16

                2.5 %    97.5 %
(Intercept) 6.9131180 7.9076462
lGrLivArea  0.5470181 0.6867395
           CV           AIC          AICc           BIC         AdjR2
 4.207065e-02 -1.205873e+03 -1.205809e+03 -1.194044e+03  4.415420e-01
```


Residual Plot
Residuals
Predicted Values
Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Location-Scale Plot
√| Standardized Residuals |
Predicted Values
Index Plot
Residuals
Observation Number
COOK's D Plot
COOK's D
Observation
Residual-Leverage Plot
Standardized Residuals
Leverage
distance contours


lSalePrice
lGrLivArea

## Model 5 Log-Log Transformation + Neighborhood Interaction Term

### Range Limitations on GrLivArea / Living Area of Square Footage Statistical Images

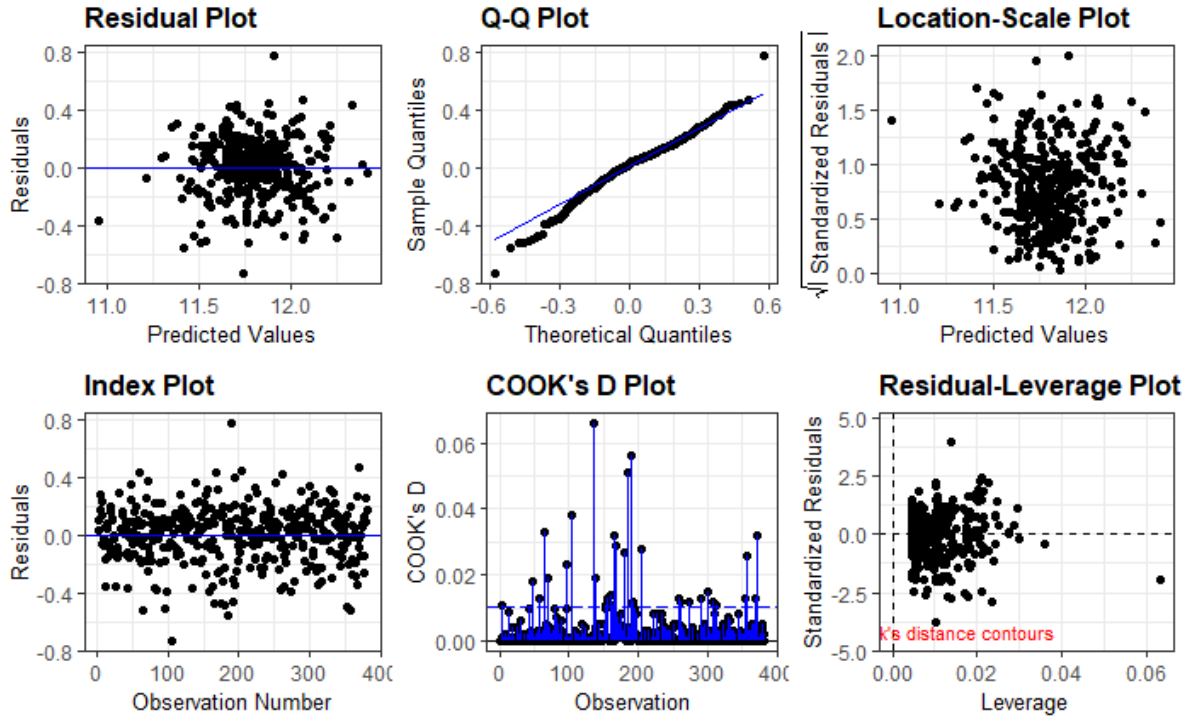
Coefficient Table

| | lSalePrice | | | | |
|---|---|---|---|---|---|
| *Predictors* | *Estimates* | *Std. Error* | *CI* | *t-value* | *p-value* |
| (Intercept) | 7.49 | 0.24 | 7.02 – 7.96 | 31.26 | **<0.001** |
| lGrLivArea | 0.60 | 0.03 | 0.53 – 0.66 | 17.59 | **<0.001** |
| Neighborhood [Edwards] | -0.01 | 0.03 | -0.08 – 0.05 | -0.44 | 0.662 |
| Neighborhood [NAmes] | 0.13 | 0.03 | 0.07 – 0.19 | 4.49 | **<0.001** |
| Observations | 381 | | | | |
| $R^2$ / $R^2$ adjusted | 0.504 / 0.500 | | | | |

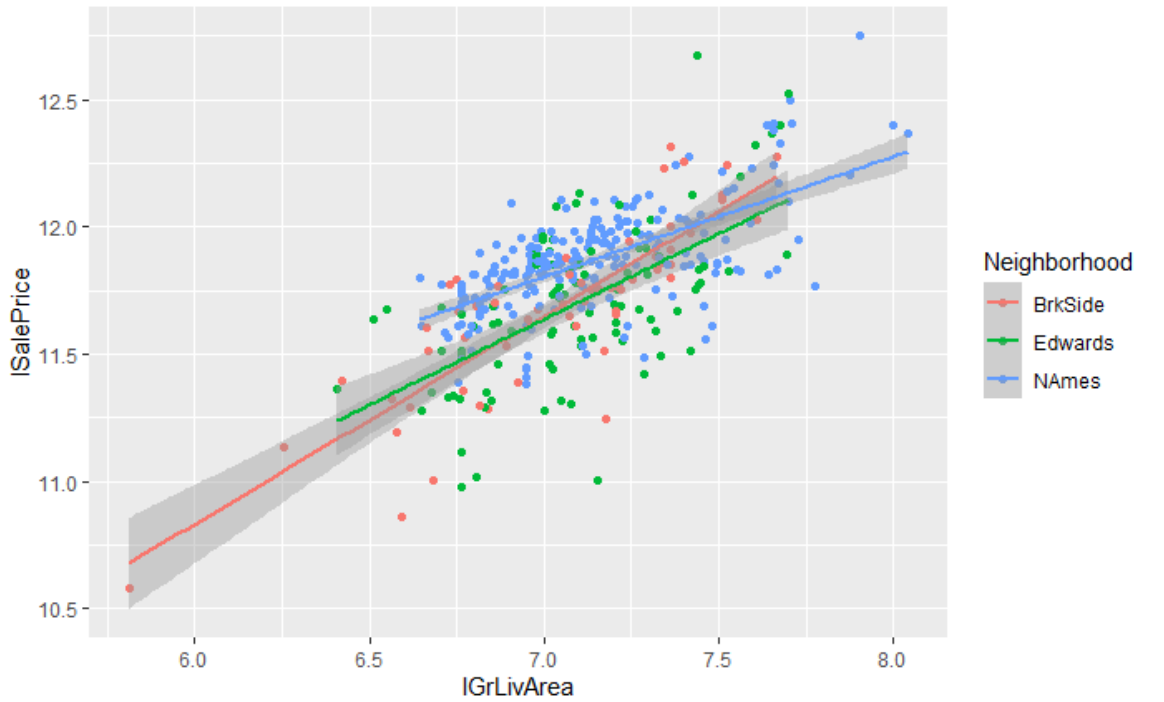
Summary Statistics of Linear Model; Confidence Interval; CV Press Statistics

```
Call:
lm(formula = lSalePrice ~ lGrLivArea + Neighborhood, data = train_log)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73253 -0.10572  0.02277  0.12232  0.77125

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.48892    0.23954  31.263  < 2e-16 ***
lGrLivArea           0.59565    0.03386  17.594  < 2e-16 ***
NeighborhoodEdwards -0.01405    0.03211  -0.438    0.662
NeighborhoodNAmes    0.12881    0.02867   4.492 9.39e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1934 on 377 degrees of freedom
Multiple R-squared:  0.5041,    Adjusted R-squared:  0.5002
F-statistic: 127.8 on 3 and 377 DF,  p-value: < 2.2e-16

                        2.5 %    97.5 %
(Intercept)          7.01791044 7.9599297
lGrLivArea           0.52907953 0.6622165
NeighborhoodEdwards -0.07719457 0.0490859
NeighborhoodNAmes    0.07242562 0.1851868
           CV           AIC          AICc           BIC         AdjR2
 3.794933e-02 -1.246162e+03 -1.246002e+03 -1.226448e+03  5.001920e-01
```

## Residual Plots




## Linear Regression Model



# Model 6: Interaction between Neighborhood and Living Area Square Footage of log log transformation

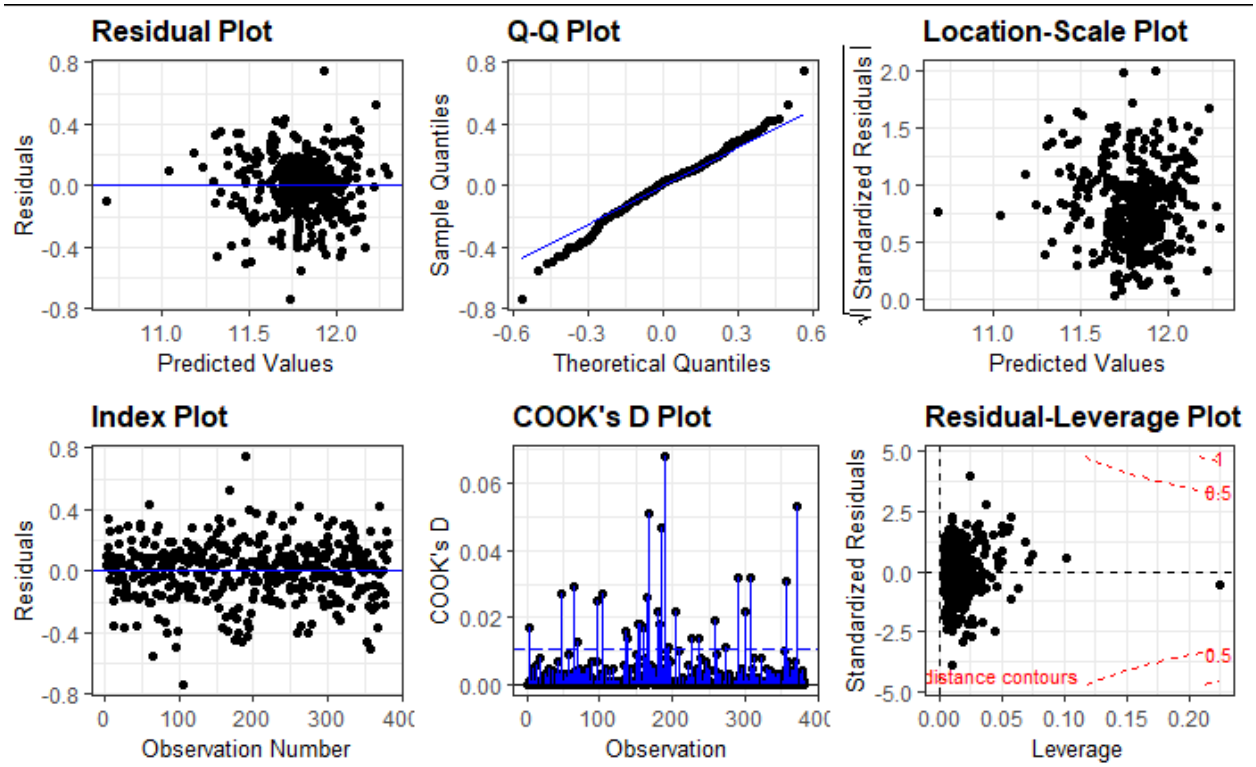
Coefficient Table

| | lSalePrice | | | | | |
|---|---|---|---|---|---|---|
| *Predictors* | *Estimates* | *Std. Error* | *CI* | *t-value* | *p-value* | *df* |
| (Intercept) | 5.91 | 0.50 | 4.94 – 6.89 | 11.91 | <**0.001** | 375.00 |
| lGrLivArea | 0.82 | 0.07 | 0.68 – 0.96 | 11.63 | <**0.001** | 375.00 |
| Neighborhood [Edwards] | 1.01 | 0.70 | -0.36 – 2.38 | 1.45 | 0.149 | 375.00 |
| Neighborhood [NAmes] | 2.58 | 0.59 | 1.42 – 3.74 | 4.37 | <**0.001** | 375.00 |
| lGrLivArea * Neighborhood [Edwards] | -0.15 | 0.10 | -0.34 – 0.05 | -1.48 | 0.139 | 375.00 |
| lGrLivArea * Neighborhood [NAmes] | -0.35 | 0.08 | -0.51 – -0.18 | -4.15 | <**0.001** | 375.00 |
| Observations | 381 | | | | | |
| $R^2$ / $R^2$ adjusted | 0.528 / 0.522 | | | | | |

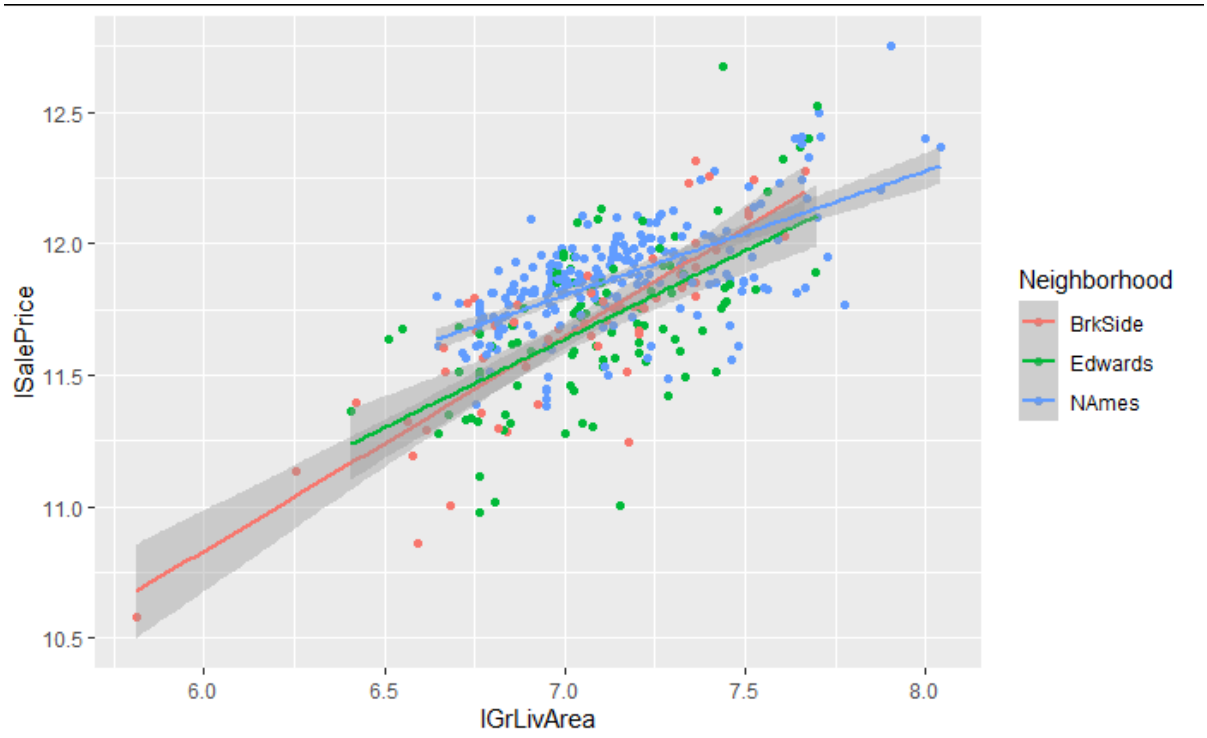
Summary Statistics of Linear Model; Confidence Interval; CV Press Statistics

```
Call:
lm(formula = lSalePrice ~ lGrLivArea + Neighborhood + Neighborhood:lGrLivArea,
    data = train_log)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73636 -0.10679  0.02187  0.10524  0.74523

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   5.91292    0.49642  11.911  < 2e-16 ***
lGrLivArea                    0.81965    0.07047  11.631  < 2e-16 ***
NeighborhoodEdwards           1.01017    0.69821   1.447    0.149
NeighborhoodNAmes             2.57981    0.59016   4.371 1.60e-05 ***
lGrLivArea:NeighborhoodEdwards -0.14631    0.09868  -1.483    0.139
lGrLivArea:NeighborhoodNAmes   -0.34662    0.08345  -4.154 4.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1892 on 375 degrees of freedom
Multiple R-squared:  0.5279,    Adjusted R-squared:  0.5216
F-statistic: 83.87 on 5 and 375 DF,  p-value: < 2.2e-16

                                 2.5 %      97.5 %
(Intercept)                  4.9368110  6.88903046
lGrLivArea                   0.6810853  0.95821077
NeighborhoodEdwards         -0.3627196  2.38306395
NeighborhoodNAmes            1.4193600  3.74025379
lGrLivArea:NeighborhoodEdwards -0.3403440  0.04772312
lGrLivArea:NeighborhoodNAmes   -0.5107056 -0.18254334
          CV           AIC          AICc           BIC         AdjR2
 3.660911e-02 -1.260894e+03 -1.260593e+03 -1.233294e+03  5.216323e-01
           lSalePrice lGrLivArea
lSalePrice  1.0000000  0.6655912
lGrLivArea  0.6655912  1.0000000
```

Residual Plots




Linear Regression Model



# Analysis 1 Code in R

````
---
title: "MSDS 6371 Project"
author: "Halle Purdom & Taylor Bonar"
date: "4/11/2021"
output:
  html_document: default
  pdf_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)

#loading libraries
library(GGally)
library(ggthemes)
library(ggplot2)
library(ggResidpanel) # Residual Plots (e.g. resid_panel())
library(scales) # Scales used to correct some of the scaling on graphs
library(naniar)
library(tidyr)
library(tidyverse)
library(plyr)
library(dplyr)
library(caret)
library(class)
library(e1071)
library(tm)
library(plotly)
library(olsrr)
library(cowplot)
library(IDPmisc)
# HTML CSS - Libraries for Coefficient Table (e.g. tab_model())
library(sjPlot)
library(sjmisc)
library(sjlabelled)
# Cross-Validation Statistic CV() Library
library(forecast)

train = read.csv("./train.csv", header = TRUE)
test = read.csv("./test.csv", header = TRUE)

```
````

# Analysis 1: How does the Square Footage of the Living Area Affect Sale Prices for the Houses in North Ames, Edwards, and Brookside?

## Step 1: Building and Fitting Our Initial Model
Initial Model: $\hat{Sale Price} =  \hat{\beta_0} + \hat{\beta_1} (Living Area ft^2)$
```{r Initial Fit, echo=T}
# Filtering for desired neighborhoods into a separate dataframe
train_filtered = train %>% filter(Neighborhood == "NAmes" | Neighborhood == "Edwards" | Neighborhood == "BrkSide") %>% select(SalePrice, GrLivArea)

# Sales Prices' relation to square footage of the Living Area
## Create linear model of Response (SalePrice) to Explanatory Variables (GrLivArea)
fit_initial = lm(SalePrice~GrLivArea, data=train_filtered)
## Find the overall Summary and Confidence Interval of our Linear Model
# Option 1 for auto-generating summary statistics -- PDF doesn't like, good for HTML quick referencing
tab_model(fit_initial, show.se = T, show.stat = T, string.stat = "t-value", string.p = "p-value", string.se = "Std. Error", pred.labels = c("Intercept", "Sq. Ft. of Living Area"))
# Option 2 - Use for PDF
summary(fit_initial)
confint(fit_initial)
CV(fit_initial)
cor(train_filtered)

## Generate Linear Regression Line
preds = predict(fit_initial)
train_filtered %>% ggplot(aes(y = SalePrice, x = GrLivArea)) +
  geom_point() + geom_smooth(method = "lm", formula = y~x) +
  ggtitle("Linear Regression Model: Living Area vs. Sale Price", "For Houses in North Ames, Edwards, and Brookside Neighborhoods")  +
  xlab("Square Footage of Living Area") +
  ylab("House Sale Price") +
  scale_y_continuous(labels = comma)

# Without two outliers of 3,500+ sq ft
train_filtered_ranged = train %>% filter(Neighborhood == "NAmes" | Neighborhood == "Edwards" | Neighborhood == "BrkSide") %>% filter(GrLivArea <= 3500)
train_filtered_ranged2 = train_filtered_ranged %>% select(SalePrice, GrLivArea)
# Sales Prices' relation to square footage of the Living Area
## Create linear model of Response (SalePrice) to Explanatory Variables (GrLivArea)
fit_ranged = lm(SalePrice~GrLivArea, data=train_filtered_ranged)
## Find the overall Summary and Confidence Interval of our Linear Model
```

```
# Option 1 for auto-generating summary statistics -- PDF doesn't like, good for HTML quick referencing
tab_model(fit_ranged, show.se = T, show.stat = T, string.stat = "t-value", string.p = "p-value", string.se = "Std. Error", pred.labels = c("Intercept", "Sq. Ft. of Living Area"))
# Option 2 - Use for PDF
summary(fit_ranged)
confint(fit_ranged)
CV(fit_ranged)
cor(train_filtered_ranged2)

preds_filtered_ranged = predict(fit_ranged)
train_filtered_ranged2 %>% ggplot(aes(y = SalePrice, x = GrLivArea)) +
  geom_point() + geom_smooth(method = "lm", formula = y~x) +
  ggtitle("Linear Regression Model: Living Area vs. Sale Price", "Range Limitation (0 to 3,500 sqft")  +
  xlab("Square Footage of Living Area") +
  ylab("House Sale Price") +
  scale_y_continuous(labels = comma)
```

## Step 2: Checking Assumptions

Assumptions:

- There is a normally distributed sub-population of responses for each value of the explanatory variable (Normalcy)
- The means of the sub-populations fall on a straight line function of the explanatory variable (Linear Relationship)
- The sub-population standard deviation are all equal (to $\sigma$) (Equivalent Variation)
- The selection of an observation from any of the sub-populations is independent of the selection of any other observation (Independence)

### Examining Residual Plots & Influential Points

```{r Initial Model Plots, echo=T}
resid_panel(fit_initial, plots=c("resid","qq","ls","index","cookd","lev"))
resid_panel(fit_ranged, plots=c("resid","qq","ls","index","cookd","lev"))
```

Generally speaking, our initial model demonstrates a medium positive linear association between a living room square footage and the sale price of a home in our three neighborhoods ($r^2=.342$, $adjr^2 = .341$).

## Step 3: Comparing Competing Models

```{r Transformations, echo=T}
# Create a new train set with log transformations w/ desired variables
train_log = train_filtered_ranged %>%
as.data.frame(train_filtered_ranged$GrLivArea,train_filtered_ranged$SalePrice,train_filtered_ranged$Neighborhood)
```

```
train_log$lGrLivArea = log(train_log$GrLivArea)
train_log$lSalePrice = log(train_log$SalePrice)
```
### normal-log Model Analysis:
$\hat{Sale Price} =  \hat{\beta_0} + \hat{\beta_1} (log(Living Area ft^2))$
```{r logGrLivArea Analysis, echo=T}
fit1 = lm(SalePrice~lGrLivArea, data=train_log)
# Option 1 for auto-generating summary statistics -- PDF doesn't like, good for HTML quick referencing
tab_model(fit1, show.se = T, show.stat = T, string.stat = "t-value", string.p = "p-value", string.se = "Std. Error", pred.labels = c("Intercept", "Sq. Ft. of Living Area"))
summary(fit1)
confint(fit1)
CV(fit1)

train_filtered_log1 = train_log %>% filter(Neighborhood == "NAmes" | Neighborhood == "Edwards" | Neighborhood == "BrkSide") %>% select(SalePrice, lGrLivArea)
cor(train_filtered_log1)

resid_panel(fit1, plots=c("resid","qq","ls","index","cookd","lev"))
preds1 = predict(fit1)
train_log %>% ggplot(aes(y = SalePrice, x = lGrLivArea)) + geom_point() + geom_smooth(method = "lm",formula = y~x)
train_log %>% ggplot(aes(y = SalePrice, x = lGrLivArea)) + geom_point() + geom_line(data = train_log, aes( x = lGrLivArea, y = preds1))
```
### log-normal Model Analysis:
Model: $log(\hat{Sale Price}) =  \hat{\beta_0} + \hat{\beta_1} (Living Area ft^2)$
```{r logSalePrice Analysis, echo=T}
fit2 = lm(lSalePrice~GrLivArea, data=train_log)
# Option 1 for auto-generating summary statistics -- PDF doesn't like, good for HTML quick referencing
tab_model(fit2, show.se = T, show.stat = T, string.stat = "t-value", string.p = "p-value", string.se = "Std. Error", pred.labels = c("Intercept", "Sq. Ft. of Living Area"))
summary(fit2)
confint(fit2)
CV(fit2)

train_filtered_log2 = train_log %>% filter(Neighborhood == "NAmes" | Neighborhood == "Edwards" | Neighborhood == "BrkSide") %>% select(lSalePrice, GrLivArea)
cor(train_filtered_log2)

resid_panel(fit2, plots=c("resid","qq","ls","index","cookd","lev"))
preds2 = predict(fit2)
```

```
train_log %>% ggplot(aes(y = lSalePrice, x = GrLivArea)) + geom_point() +
geom_smooth(method = "lm",formula = y~x)
train_log %>% ggplot(aes(y = lSalePrice, x = GrLivArea)) + geom_point() + geom_line(data =
train_log, aes( x = GrLivArea, y = preds2))
```
### log-log Model
```{r log-log Model Analysis, echo=T}
fit3 = lm(lSalePrice~lGrLivArea, data=train_log)
# Option 1 for auto-generating summary statistics -- PDF doesn't like, good for HTML quick
referencing
tab_model(fit3, show.se = T, show.stat = T, string.stat = "t-value", string.p = "p-value", string.se =
"Std. Error")
summary(fit3)
confint(fit3)
CV(fit3)

train_filtered_log3 = train_log %>% filter(Neighborhood == "NAmes" | Neighborhood ==
"Edwards" | Neighborhood == "BrkSide") %>% select(lSalePrice, lGrLivArea)
cor(train_filtered_log3)

resid_panel(fit3, plots=c("resid","qq","ls","index","cookd","lev"))
preds3 = predict(fit3)
train_log %>% ggplot(aes(y = lSalePrice, x = lGrLivArea)) + geom_point() +
geom_smooth(method = "lm",formula = y~x)
```

### log-log Model with Neighborhood
```{r Neighborhood Model Analysis, echo=T}
fit4 = lm(lSalePrice~lGrLivArea + Neighborhood, data=train_log)
# Option 1 for auto-generating summary statistics -- PDF doesn't like, good for HTML quick
referencing
tab_model(fit4, show.se = T, show.stat = T, show.df = T, string.stat = "t-value", string.p =
"p-value", string.se = "Std. Error")
summary(fit4)
confint(fit4)
CV(fit4)

train_filtered_log4 = train_log %>% filter(Neighborhood == "NAmes" | Neighborhood ==
"Edwards" | Neighborhood == "BrkSide") %>% select(lSalePrice, lGrLivArea)
cor(train_filtered_log4)


resid_panel(fit4, plots=c("resid","qq","ls","index","cookd","lev"))
preds4 = predict(fit4)


```

```
train_log %>% ggplot(aes(y = lSalePrice, x = lGrLivArea)) + geom_point() +
geom_smooth(method = "lm",formula = y~x)
train_log %>% ggplot(aes(y = lSalePrice, x = lGrLivArea, color=Neighborhood)) + geom_point()
+ geom_smooth(method = "lm",formula = y~x)
train_log %>% ggplot(aes(y = lSalePrice, x = lGrLivArea, color=Neighborhood)) + geom_point()
+ geom_line(data = train_log, aes(group=Neighborhood, x = lGrLivArea, y = preds4))
```

### log-log Model with Neighborhood*GrLivArea
```{r Neighborhood Model Analysis, echo=T}
fit5 = lm(lSalePrice~lGrLivArea + Neighborhood + Neighborhood:lGrLivArea, data=train_log)
# Option 1 for auto-generating summary statistics -- PDF doesn't like, good for HTML quick
referencing
tab_model(fit5, show.se = T, show.stat = T, show.df = T, string.stat = "t-value", string.p =
"p-value", string.se = "Std. Error")
summary(fit5)
confint(fit5)
CV(fit5)

train_filtered_log5 = train_log %>% filter(Neighborhood == "NAmes" | Neighborhood ==
"Edwards" | Neighborhood == "BrkSide") %>% select(lSalePrice, lGrLivArea)
cor(train_filtered_log4)


resid_panel(fit5, plots=c("resid","qq","ls","index","cookd","lev"))
preds4 = predict(fit5)
train_log %>% ggplot(aes(y = lSalePrice, x = lGrLivArea)) + geom_point() +
geom_smooth(method = "lm",formula = y~x)
train_log %>% ggplot(aes(y = lSalePrice, x = lGrLivArea, color=Neighborhood)) + geom_point()
+ geom_smooth(method = "lm",formula = y~x)
train_log %>% ggplot(aes(y = lSalePrice, x = lGrLivArea, color=Neighborhood)) + geom_point()
+ geom_line(data = train_log, aes(group=Neighborhood, x = lGrLivArea, y = preds4))
```

## Model Selection Process SAS Code

```
/*Import train_q3.csv file*/
/* train_q3 is only the numeric variables from train.
It has log transformed SalePrice and GrLivArea
in addition to setting missing variables to zero*/
proc import datafile="/home/u50452984/train_q3.csv"
	out=train_q3
	dbms=csv
	replace;
```

```
run;

/*Forward selection with all variables*/
proc glmselect data=train_q3 plots(stepaxis = number) = (criterionpanel ASEplot);
model lSalePrice = MSSubClass--lGrLivArea
/ selection = forward(choose = cv stop=CV) cvdetails=all;

/*Backward selection with all variables*/
proc glmselect data=train_q3 plots(stepaxis = number) = (criterionpanel ASEplot);
model lSalePrice = MSSubClass--lGrLivArea
/ selection = backward(choose = cv stop=CV) cvdetails=all;

/*Stepwise selection with all variables*/
proc glmselect data=train_q3 plots(stepaxis = number) = (criterionpanel ASEplot);
model lSalePrice = MSSubClass--lGrLivArea
/ selection = stepwise(choose = cv stop=CV) cvdetails=all;
```

# Forward Selection Model

**The GLMSELECT Procedure**

| | |
|---|---|
| Data Set | WORK.TRAIN_Q3 |
| Dependent Variable | lSalePrice |
| Selection Method | Forward |
| Select Criterion | SBC |
| Stop Criterion | Cross Validation |
| Choose Criterion | Cross Validation |
| Cross Validation Method | Random |
| Cross Validation Fold | 5 |
| Effect Hierarchy Enforced | None |
| Random Number Seed | 597981612 |

| | |
|---|---|
| Number of Observations Read | 1454 |
| Number of Observations Used | 1454 |

| Dimensions | |
|---|---|
| Number of Effects | 37 |
| Number of Parameters | 37 |

---

**The GLMSELECT Procedure**

**Forward Selection Summary**

| Step | Effect Entered | Number Effects In | SBC | CV PRESS |
|---|---|---|---|---|
| 0 | Intercept | 1 | -2703.2415 | 225.7045 |
| 1 | OverallQual | 2 | -4298.1199 | 75.0770 |
| 2 | lGrLivArea | 3 | -4749.4541 | 54.7387 |
| 3 | BsmtFinSF1 | 4 | -5060.8930 | 44.0061 |
| 4 | YearBuilt | 5 | -5300.0624 | 37.2031 |
| 5 | OverallCond | 6 | -5489.8591 | 32.5171 |
| 6 | TotalBsmtSF | 7 | -5650.8669 | 29.0825 |
| 7 | GarageCars | 8 | -5744.0843 | 27.2225 |
| 8 | LotArea | 9 | -5805.5153 | 26.1697 |
| 9 | MSSubClass | 10 | -5838.5678 | 25.3892 |
| 10 | Fireplaces | 11 | -5865.3930 | 24.8161 |
| 11 | YearRemodAdd | 12 | -5886.0380 | 24.3482 |
| 12 | KitchenAbvGr | 13 | -5893.1995 | 24.1450 |
| 13 | BsmtFullBath | 14 | -5898.8246 | 23.9779 |
| 14 | ScreenPorch | 15 | -5902.1139 | 23.8103 |
| 15 | GarageArea | 16 | -5903.6485 | 23.6856 |
| 16 | WoodDeckSF | 17 | -5903.6589 | 23.6485 |
| 17 | X2ndFlrSF | 18 | -5902.8661 | 23.5710 |
| 18 | X1stFlrSF | 19 | -5913.4570* | 23.2899 |
| 19 | YrSold | 20 | -5912.2478 | 23.2209* |

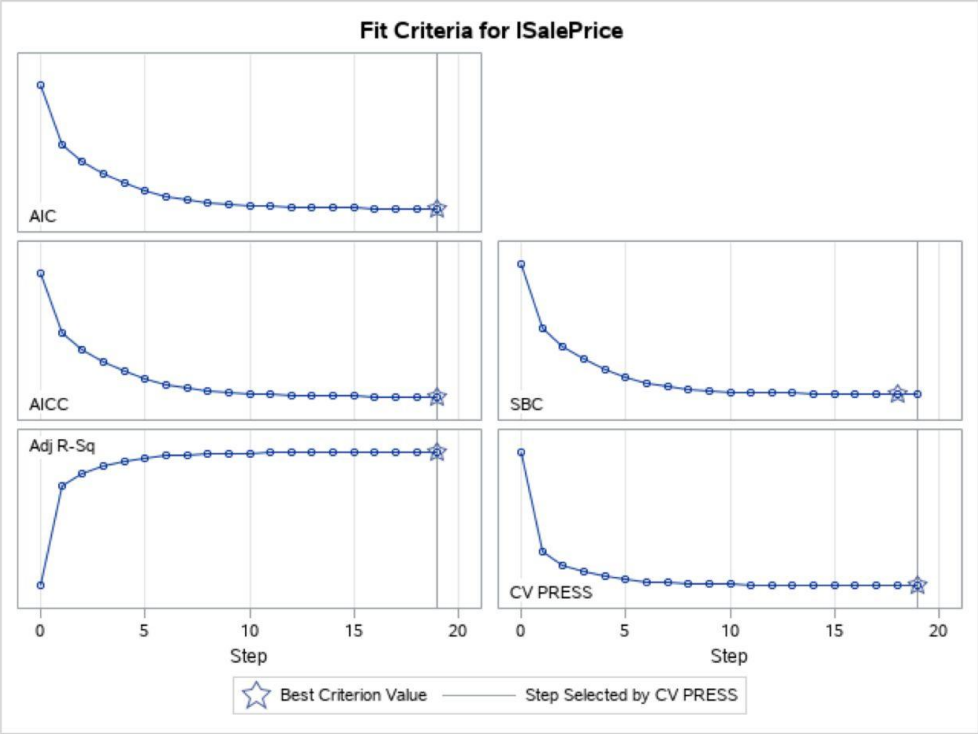
* Optimal Value of Criterion

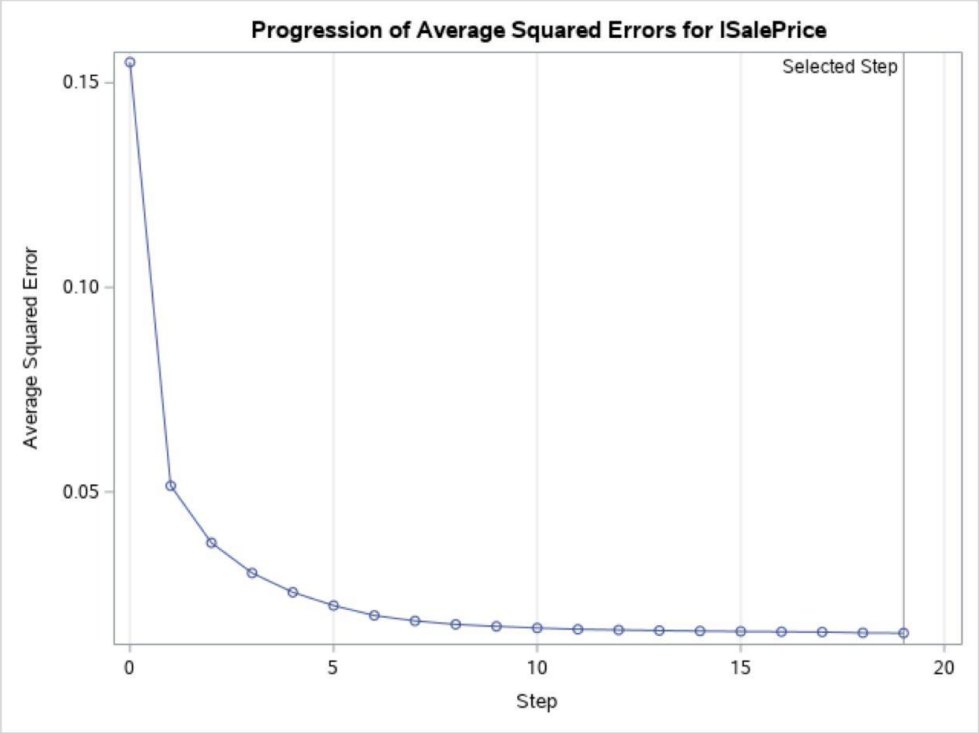
Selection stopped at a local minimum of the cross validation PRESS.

**Stop Details**

| Candidate For | Effect | Candidate CV PRESS | | Compare CV PRESS |
|---|---|---|---|---|

| Stop Details | | | | |
|---|---|---|---|---|
| Candidate For | Effect | Candidate CV PRESS | | Compare CV PRESS |
| Entry | BedroomAbvGr | 23.2488 | > | 23.2209 |






**The GLMSELECT Procedure**
**Selected Model**

**The selected model, based on Cross Validation, is the model at Step 19.**

| | |
|---|---|
| **Effects:** | Intercept MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 TotalBsmtSF X1stFlrSF X2ndFlrSF BsmtFullBath KitchenAbvGr Fireplaces GarageCars GarageArea WoodDeckSF ScreenPorch YrSold lGrLivArea |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Value |
|---|---|---|---|---|
| Model | 19 | 202.85345 | 10.67650 | 678.96 |
| Error | 1434 | 22.54949 | 0.01572 | |
| Corrected Total | 1453 | 225.40294 | | |

| | |
|---|---|
| Root MSE | 0.12540 |
| Dependent Mean | 12.02031 |
| R-Square | 0.9000 |
| Adj R-Sq | 0.8986 |
| AIC | -4561.88924 |
| AICC | -4561.24399 |
| SBC | -5912.24776 |
| CV PRESS | 23.22090 |

**Cross Validation Details**

| Index | Observations | CV PRESS |
|---|---|---|

| Index | Observations | | CV PRESS |
|---|---|---|---|
| | Fitted | Left Out | |
| 1 | 1142 | 312 | 4.3823 |
| 2 | 1159 | 295 | 4.3340 |
| 3 | 1191 | 263 | 4.4033 |
| 4 | 1163 | 291 | 5.2031 |
| 5 | 1161 | 293 | 4.8982 |
| Total | | | 23.2209 |

**Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | t Value | Cross Validation Estimates | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 | 2 | 3 | 4 | 5 |
| Intercept | 1 | 13.125878 | 5.016704 | 2.62 | 1.68E+01 | 1.29E+01 | 1.34E+01 | 9.73E+00 | 1.28E+01 |
| MSSubClass | 1 | -0.000479 | 0.000091848 | -5.22 | -5.28E-04 | -4.35E-04 | -4.60E-04 | -4.82E-04 | -4.90E-04 |
| LotArea | 1 | 0.000002325 | 0.000000362 | 6.42 | 2.64E-06 | 2.18E-06 | 2.49E-06 | 2.25E-06 | 2.28E-06 |
| OverallQual | 1 | 0.073525 | 0.004211 | 17.46 | 7.42E-02 | 7.24E-02 | 7.27E-02 | 7.50E-02 | 7.38E-02 |
| OverallCond | 1 | 0.049910 | 0.003612 | 13.82 | 5.11E-02 | 4.93E-02 | 5.24E-02 | 4.97E-02 | 4.69E-02 |
| YearBuilt | 1 | 0.002992 | 0.000185 | 16.18 | 2.94E-03 | 3.06E-03 | 3.00E-03 | 2.93E-03 | 3.03E-03 |
| YearRemodAdd | 1 | 0.001227 | 0.000230 | 5.32 | 1.43E-03 | 1.19E-03 | 1.11E-03 | 1.25E-03 | 1.14E-03 |
| BsmtFinSF1 | 1 | 0.000075527 | 0.000011228 | 6.73 | 7.17E-05 | 7.73E-05 | 7.06E-05 | 8.44E-05 | 7.32E-05 |
| TotalBsmtSF | 1 | 0.000110 | 0.000015170 | 7.28 | 1.18E-04 | 1.18E-04 | 1.17E-04 | 9.99E-05 | 9.93E-05 |
| X1stFlrSF | 1 | 0.000143 | 0.000033993 | 4.19 | 1.36E-04 | 1.49E-04 | 1.40E-04 | 1.25E-04 | 1.62E-04 |
| X2ndFlrSF | 1 | 0.000148 | 0.000030212 | 4.91 | 1.50E-04 | 1.54E-04 | 1.44E-04 | 1.33E-04 | 1.61E-04 |
| BsmtFullBath | 1 | 0.030891 | 0.008674 | 3.56 | 3.56E-02 | 2.66E-02 | 3.43E-02 | 2.51E-02 | 3.34E-02 |
| KitchenAbvGr | 1 | -0.057010 | 0.017555 | -3.25 | -4.65E-02 | -5.68E-02 | -5.77E-02 | -5.52E-02 | -7.09E-02 |
| Fireplaces | 1 | 0.034277 | 0.006351 | 5.40 | 2.89E-02 | 3.17E-02 | 3.84E-02 | 3.74E-02 | 3.44E-02 |
| GarageCars | 1 | 0.035531 | 0.010333 | 3.44 | 4.18E-02 | 3.36E-02 | 2.97E-02 | 4.06E-02 | 3.19E-02 |
| GarageArea | 1 | 0.000093533 | 0.000034920 | 2.68 | 7.73E-05 | 1.07E-04 | 9.84E-05 | 9.02E-05 | 9.72E-05 |
| WoodDeckSF | 1 | 0.000069641 | 0.000028538 | 2.44 | 7.77E-05 | 5.95E-05 | 7.20E-05 | 3.36E-05 | 1.04E-04 |
| ScreenPorch | 1 | 0.000226 | 0.000061917 | 3.64 | 2.53E-04 | 2.56E-04 | 1.92E-04 | 1.98E-04 | 2.19E-04 |
| YrSold | 1 | -0.006117 | 0.002497 | -2.45 | -8.11E-03 | -6.05E-03 | -6.19E-03 | -4.45E-03 | -5.80E-03 |
| lGrLivArea | 1 | 0.229446 | 0.045737 | 5.02 | 2.31E-01 | 2.26E-01 | 2.39E-01 | 2.50E-01 | 2.03E-01 |

# Backward Selection Model

**The GLMSELECT Procedure**

| | |
|---|---|
| **Data Set** | WORK.TRAIN_Q3 |
| **Dependent Variable** | lSalePrice |
| **Selection Method** | Backward |
| **Select Criterion** | SBC |
| **Stop Criterion** | Cross Validation |
| **Choose Criterion** | Cross Validation |
| **Cross Validation Method** | Random |
| **Cross Validation Fold** | 5 |
| **Effect Hierarchy Enforced** | None |
| **Random Number Seed** | 759272208 |

| | |
|---|---|
| **Number of Observations Read** | 1454 |
| **Number of Observations Used** | 1454 |

| **Dimensions** | |
|---|---|
| **Number of Effects** | 37 |
| **Number of Parameters** | 37 |

---

**The GLMSELECT Procedure**

**Backward Selection Summary**

| Step | Effect Removed | Number Effects In | SBC | CV PRESS |
|---|---|---|---|---|
| 0 | | 37 | -5818.1229 | 24.0113 |
| | TotalBsmtSF | 36 | -5818.1229 | 24.0113 |
| 1 | MoSold | 35 | -5825.4017 | 23.9851 |
| 2 | LowQualFinSF | 34 | -5832.6458 | 23.9164 |
| 3 | BsmtHalfBath | 33 | -5839.8802 | 23.8932 |
| 4 | PoolArea | 32 | -5846.8390 | 23.5376 |
| 5 | MasVnrArea | 31 | -5853.7846 | 23.5149 |
| 6 | MiscVal | 30 | -5860.4319 | 23.5114 |
| 7 | OpenPorchSF | 29 | -5866.9396 | 23.4384 |
| 8 | HalfBath | 28 | -5873.2070 | 23.4280 |
| 9 | GarageYrBlt | 27 | -5879.5716 | 23.4102 |
| 10 | FullBath | 26 | -5885.7618 | 23.4062 |
| 11 | X3SsnPorch | 25 | -5891.7028* | 23.3662* |

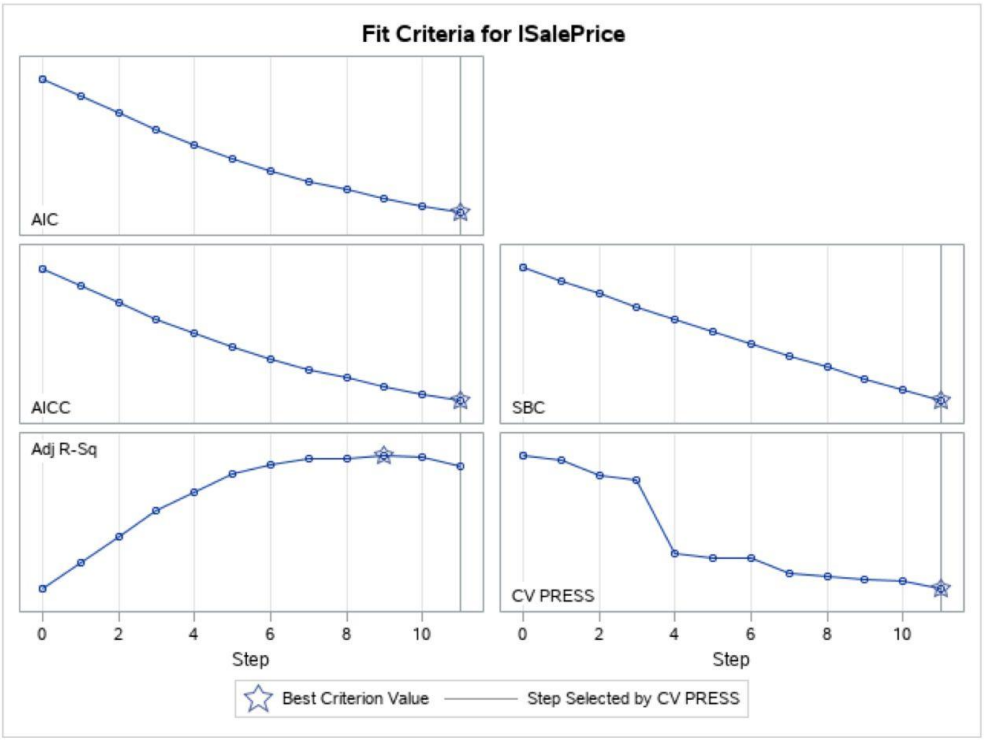
* Optimal Value of Criterion

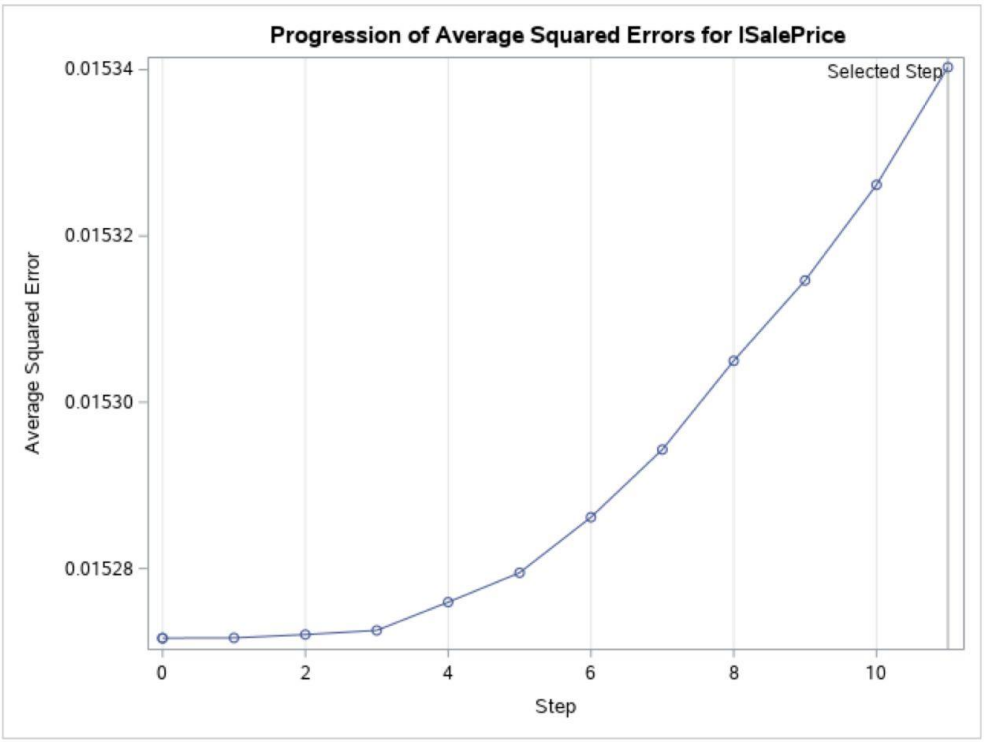
**Note:** Effects dropped at step 0 are redundant.

Selection stopped at a local minimum of the cross validation PRESS.

**Stop Details**

| Candidate For | Effect | Candidate CV PRESS | | Compare CV PRESS |
|---|---|---|---|---|
| Removal | LotFrontage | 23.3841 | > | 23.3662 |


Fit Criteria for lSalePrice
AIC
AICC
SBC
Adj R-Sq
CV PRESS
0
2
4
6
8
10
Step
Best Criterion Value
Step Selected by CV PRESS


Progression of Average Squared Errors for lSalePrice
Selected Step
0.01534
0.01532
0.01530
0.01528
Average Squared Error
0
2
4
6
8
10
Step

**The GLMSELECT Procedure**
**Selected Model**

**The selected model, based on Cross Validation, is the model at Step 11.**

| Effects: | Intercept MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF X1stFlrSF X2ndFlrSF BsmtFullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF EnclosedPorch ScreenPorch YrSold IGrLivArea |
|---|---|

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Value |
|---|---|---|---|---|
| Model | 24 | 203.09816 | 8.46242 | 542.16 |
| Error | 1429 | 22.30478 | 0.01561 | |
| Corrected Total | 1453 | 225.40294 | | |

| | |
|---|---|
| Root MSE | 0.12493 |
| Dependent Mean | 12.02031 |
| R-Square | 0.9010 |
| Adj R-Sq | 0.8994 |
| AIC | -4567.75468 |
| AICC | -4566.77079 |
| SBC | -5891.70284 |
| CV PRESS | 23.36618 |

**Cross Validation Details**

| Index | Observations | | CV PRESS |
|---|---|---|---|
| | Fitted | Left Out | |
| 1 | 1146 | 308 | 5.4426 |
| 2 | 1178 | 276 | 3.8536 |
| 3 | 1160 | 294 | 4.0706 |
| 4 | 1155 | 299 | 6.1294 |
| 5 | 1177 | 277 | 3.8700 |
| Total | | | 23.3662 |

**Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | t Value | Cross Validation Estimates | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 | 2 | 3 | 4 | 5 |
| Intercept | 1 | 12.877565 | 5.004350 | 2.57 | 1.26E+01 | 1.15E+01 | 1.10E+01 | 1.35E+01 | 1.65E+01 |
| MSSubClass | 1 | -0.000463 | 0.000095677 | -4.84 | -4.80E-04 | -4.51E-04 | -4.77E-04 | -4.07E-04 | -4.93E-04 |
| LotFrontage | 1 | 0.000183 | 0.000104 | 1.76 | 1.32E-04 | 1.48E-04 | 2.46E-04 | 1.49E-04 | 2.32E-04 |
| LotArea | 1 | 0.000002358 | 0.000000362 | 6.51 | 2.30E-06 | 2.58E-06 | 2.23E-06 | 2.25E-06 | 2.55E-06 |
| OverallQual | 1 | 0.070273 | 0.004312 | 16.30 | 7.13E-02 | 7.01E-02 | 7.13E-02 | 7.10E-02 | 6.81E-02 |
| OverallCond | 1 | 0.051952 | 0.003663 | 14.18 | 5.01E-02 | 5.34E-02 | 5.63E-02 | 5.07E-02 | 4.95E-02 |
| YearBuilt | 1 | 0.003213 | 0.000200 | 16.10 | 3.25E-03 | 3.10E-03 | 3.39E-03 | 3.15E-03 | 3.19E-03 |
| YearRemodAdd | 1 | 0.001100 | 0.000234 | 4.70 | 1.10E-03 | 1.12E-03 | 1.03E-03 | 1.13E-03 | 1.09E-03 |
| BsmtFinSF1 | 1 | 0.000189 | 0.000017249 | 10.95 | 1.76E-04 | 1.80E-04 | 2.00E-04 | 1.92E-04 | 1.97E-04 |
| BsmtFinSF2 | 1 | 0.000123 | 0.000025256 | 4.86 | 1.06E-04 | 1.05E-04 | 1.46E-04 | 1.24E-04 | 1.35E-04 |
| BsmtUnfSF | 1 | 0.000111 | 0.000015256 | 7.27 | 1.12E-04 | 9.86E-05 | 1.14E-04 | 1.13E-04 | 1.16E-04 |
| X1stFlrSF | 1 | 0.000126 | 0.000034694 | 3.62 | 1.40E-04 | 1.45E-04 | 1.04E-04 | 1.30E-04 | 1.13E-04 |
| X2ndFlrSF | 1 | 0.000137 | 0.000031137 | 4.40 | 1.44E-04 | 1.32E-04 | 1.31E-04 | 1.47E-04 | 1.35E-04 |
| BsmtFullBath | 1 | 0.027957 | 0.008964 | 3.12 | 3.09E-02 | 2.92E-02 | 2.25E-02 | 2.47E-02 | 3.18E-02 |
| BedroomAbvGr | 1 | -0.017688 | 0.006089 | -2.91 | -2.18E-02 | -1.69E-02 | -1.55E-02 | -2.06E-02 | -1.39E-02 |

| Parameter Estimates | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Cross Validation Estimates | | | | |
| Parameter | DF | Estimate | Standard Error | t Value | 1 | 2 | 3 | 4 | 5 |
| KitchenAbvGr | 1 | -0.062633 | 0.018329 | -3.42 | -6.63E-02 | -7.01E-02 | -5.11E-02 | -5.97E-02 | -6.77E-02 |
| TotRmsAbvGrd | 1 | 0.009244 | 0.004451 | 2.08 | 7.44E-03 | 1.02E-02 | 6.58E-03 | 1.08E-02 | 1.06E-02 |
| Fireplaces | 1 | 0.033165 | 0.006394 | 5.19 | 3.59E-02 | 3.24E-02 | 3.53E-02 | 3.00E-02 | 3.24E-02 |
| GarageCars | 1 | 0.034655 | 0.010309 | 3.36 | 3.32E-02 | 3.99E-02 | 3.49E-02 | 3.13E-02 | 3.36E-02 |
| GarageArea | 1 | 0.000085964 | 0.000034934 | 2.46 | 1.13E-04 | 5.59E-05 | 7.71E-05 | 8.62E-05 | 9.84E-05 |
| WoodDeckSF | 1 | 0.000074624 | 0.000028634 | 2.61 | 7.19E-05 | 8.33E-05 | 8.53E-05 | 6.53E-05 | 6.33E-05 |
| EnclosedPorch | 1 | 0.000108 | 0.000060267 | 1.79 | 1.13E-04 | 1.54E-04 | 1.18E-04 | 2.85E-05 | 1.34E-04 |
| ScreenPorch | 1 | 0.000242 | 0.000062546 | 3.87 | 2.14E-04 | 2.51E-04 | 2.51E-04 | 2.06E-04 | 2.82E-04 |
| YrSold | 1 | -0.006123 | 0.002490 | -2.46 | -5.98E-03 | -5.36E-03 | -5.37E-03 | -6.33E-03 | -7.92E-03 |
| lGrLivArea | 1 | 0.242446 | 0.046515 | 5.21 | 2.34E-01 | 2.39E-01 | 2.68E-01 | 2.20E-01 | 2.47E-01 |

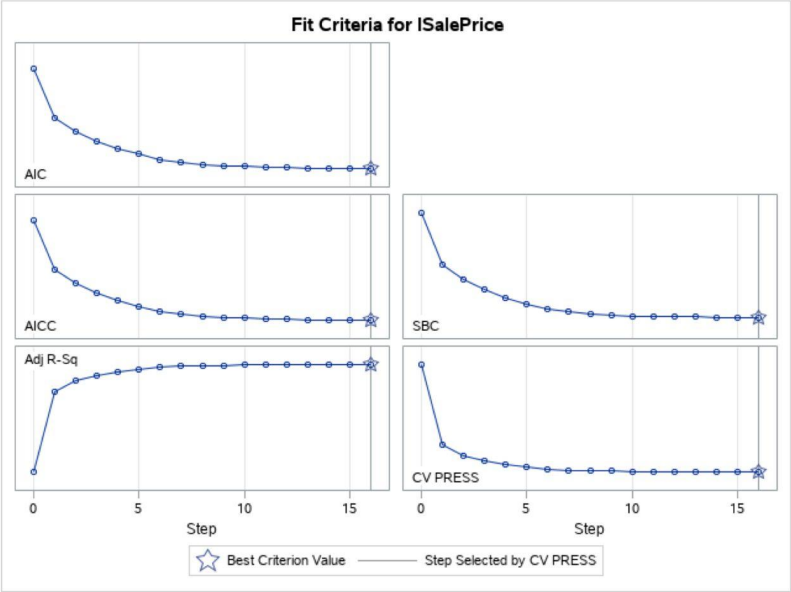
# Stepwise Selection Model

The GLMSELECT Procedure

| | |
|---|---|
| Data Set | WORK.TRAIN_Q3 |
| Dependent Variable | lSalePrice |
| Selection Method | Stepwise |
| Select Criterion | SBC |
| Stop Criterion | Cross Validation |
| Choose Criterion | Cross Validation |
| Cross Validation Method | Random |
| Cross Validation Fold | 5 |
| Effect Hierarchy Enforced | None |
| Random Number Seed | 689819207 |

| | |
|---|---|
| Number of Observations Read | 1454 |
| Number of Observations Used | 1454 |

| Dimensions | |
|---|---|
| Number of Effects | 37 |
| Number of Parameters | 37 |



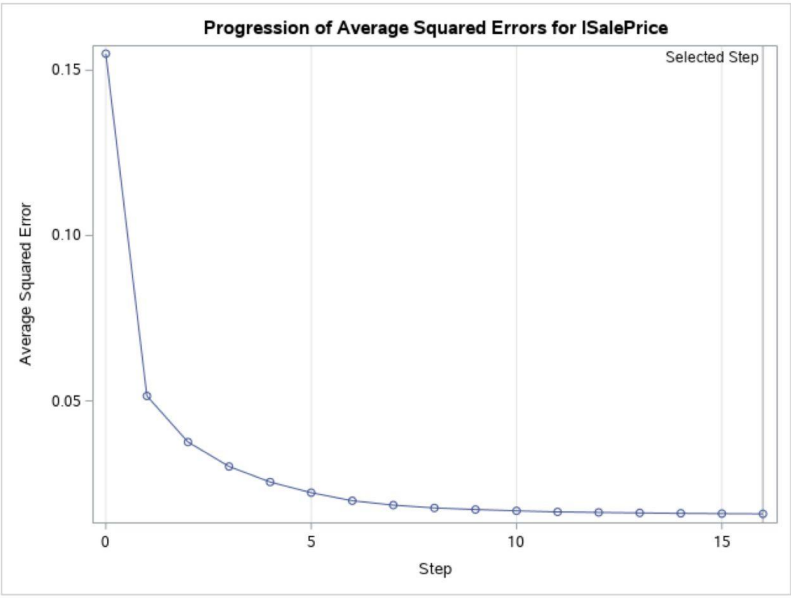

The GLMSELECT Procedure

Stepwise Selection Summary

| Step | Effect Entered | Effect Removed | Number Effects In | SBC | CV PRESS |
|---|---|---|---|---|---|
| 0 | Intercept | | 1 | -2703.2415 | 225.9236 |
| 1 | OverallQual | | 2 | -4298.1199 | 75.0649 |
| 2 | lGrLivArea | | 3 | -4749.4541 | 54.8296 |
| 3 | BsmtFinSF1 | | 4 | -5060.8930 | 44.0163 |
| 4 | YearBuilt | | 5 | -5300.0624 | 37.3025 |
| 5 | OverallCond | | 6 | -5489.8591 | 32.6081 |
| 6 | TotalBsmtSF | | 7 | -5650.8669 | 29.1497 |
| 7 | GarageCars | | 8 | -5744.0843 | 27.1512 |
| 8 | LotArea | | 9 | -5805.5153 | 26.2691 |
| 9 | MSSubClass | | 10 | -5838.5678 | 25.5263 |
| 10 | Fireplaces | | 11 | -5865.3930 | 24.9860 |
| 11 | YearRemodAdd | | 12 | -5886.0380 | 24.5567 |
| 12 | KitchenAbvGr | | 13 | -5893.1995 | 24.3787 |
| 13 | BsmtFullBath | | 14 | -5898.8246 | 24.1827 |
| 14 | ScreenPorch | | 15 | -5902.1139 | 24.0302 |
| 15 | GarageArea | | 16 | -5903.6485 | 23.9444 |
| 16 | WoodDeckSF | | 17 | -5903.6589* | 23.8207* |

* Optimal Value of Criterion

Selection stopped as adding or dropping any effect does not improve the selection criterion.



**The GLMSELECT Procedure**
**Selected Model**

**The selected model, based on Cross Validation, is the model at Step 16.**

| Effects: | Intercept MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 TotalBsmtSF BsmtFullBath KitchenAbvGr Fireplaces GarageCars GarageArea WoodDeckSF ScreenPorch IGrLivArea |
|---|---|

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Value |
|---|---|---|---|---|
| Model | 16 | 202.37647 | 12.64853 | 789.35 |
| Error | 1437 | 23.02647 | 0.01602 | |
| Corrected Total | 1453 | 225.40294 | | |

| | |
|---|---|
| Root MSE | 0.12659 |
| Dependent Mean | 12.02031 |
| R-Square | 0.8978 |
| Adj R-Sq | 0.8967 |
| AIC | -4537.45415 |
| AICC | -4536.97749 |
| SBC | -5903.65890 |
| CV PRESS | 23.82066 |

**Cross Validation Details**

| Index | Observations | | CV PRESS |
|---|---|---|---|
| | Fitted | Left Out | |
| 1 | 1196 | 258 | 4.6819 |
| 2 | 1164 | 290 | 4.1447 |
| 3 | 1167 | 287 | 5.0663 |
| 4 | 1147 | 307 | 5.1720 |
| 5 | 1142 | 312 | 4.7558 |
| Total | | | 23.8207 |

**Parameter Estimates**

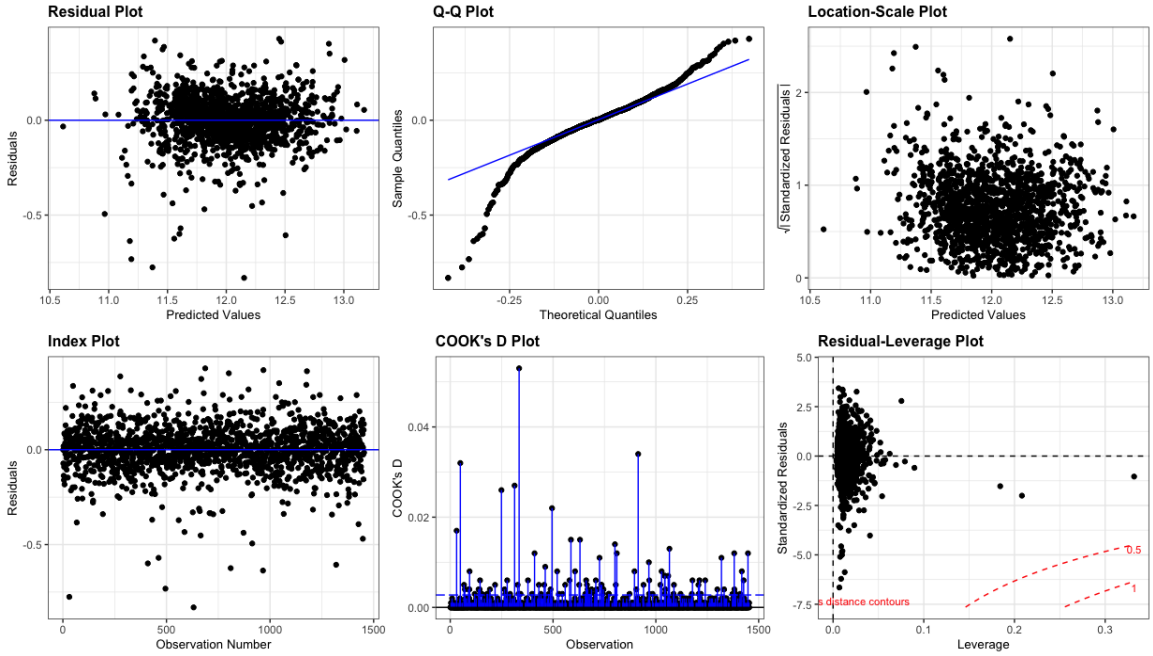
| Parameter | DF | Estimate | Standard Error | t Value | Cross Validation Estimates | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 | 2 | 3 | 4 | 5 |
| Intercept | 1 | -0.327585 | 0.437258 | -0.75 | 8.38E-02 | -3.10E-01 | -6.88E-01 | -0.6095873 | -1.17E-01 |
| MSSubClass | 1 | -0.000467 | 0.000090072 | -5.19 | -4.06E-04 | -3.52E-04 | -5.05E-04 | -0.0005427 | -5.04E-04 |
| LotArea | 1 | 0.000002372 | 0.000000365 | 6.49 | 3.11E-06 | 2.16E-06 | 2.23E-06 | 0.0000026 | 2.24E-06 |
| OverallQual | 1 | 0.074188 | 0.004243 | 17.49 | 7.75E-02 | 7.39E-02 | 7.48E-02 | 0.0721684 | 7.25E-02 |
| OverallCond | 1 | 0.050318 | 0.003642 | 13.82 | 5.14E-02 | 5.00E-02 | 5.02E-02 | 0.0508320 | 4.90E-02 |
| YearBuilt | 1 | 0.002945 | 0.000186 | 15.83 | 2.90E-03 | 2.88E-03 | 3.08E-03 | 0.0028785 | 2.99E-03 |
| YearRemodAdd | 1 | 0.001197 | 0.000232 | 5.16 | 1.05E-03 | 1.23E-03 | 1.28E-03 | 0.0013915 | 1.05E-03 |
| BsmtFinSF1 | 1 | 0.000077875 | 0.000011308 | 6.89 | 8.20E-05 | 7.79E-05 | 8.02E-05 | 0.0000745 | 7.30E-05 |
| TotalBsmtSF | 1 | 0.000106 | 0.000011675 | 9.08 | 1.02E-04 | 1.13E-04 | 1.05E-04 | 0.0001005 | 1.12E-04 |
| BsmtFullBath | 1 | 0.029179 | 0.008735 | 3.34 | 2.67E-02 | 3.18E-02 | 2.94E-02 | 0.0276927 | 3.10E-02 |
| KitchenAbvGr | 1 | -0.059193 | 0.017211 | -3.44 | -6.14E-02 | -5.22E-02 | -7.19E-02 | -0.0575706 | -5.75E-02 |
| Fireplaces | 1 | 0.035210 | 0.006335 | 5.56 | 3.15E-02 | 2.95E-02 | 3.64E-02 | 0.0389308 | 3.98E-02 |
| GarageCars | 1 | 0.035148 | 0.010423 | 3.37 | 3.64E-02 | 3.99E-02 | 2.08E-02 | 0.0382547 | 3.92E-02 |
| GarageArea | 1 | 0.000103 | 0.000035091 | 2.93 | 9.44E-05 | 9.03E-05 | 1.49E-04 | 0.0000829 | 9.93E-05 |
| WoodDeckSF | 1 | 0.000077177 | 0.000028712 | 2.69 | 9.16E-05 | 6.57E-05 | 9.05E-05 | 0.0000684 | 6.28E-05 |

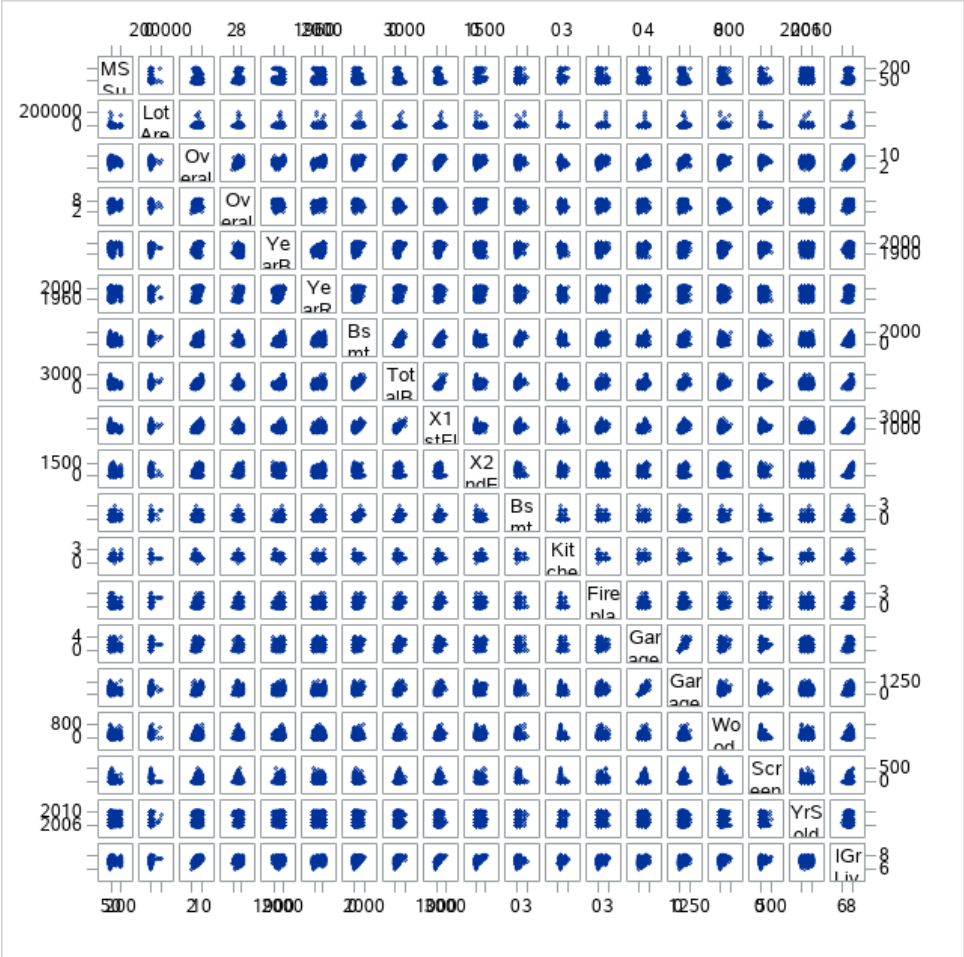
**Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | t Value | Cross Validation Estimates | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 | 2 | 3 | 4 | 5 |
| ScreenPorch | 1 | 0.000222 | 0.000062476 | 3.55 | 1.94E-04 | 2.45E-04 | 2.27E-04 | 0.0002223 | 2.21E-04 |
| IGrLivArea | 1 | 0.440463 | 0.015288 | 28.81 | 4.35E-01 | 4.44E-01 | 4.33E-01 | 0.4472172 | 4.40E-01 |

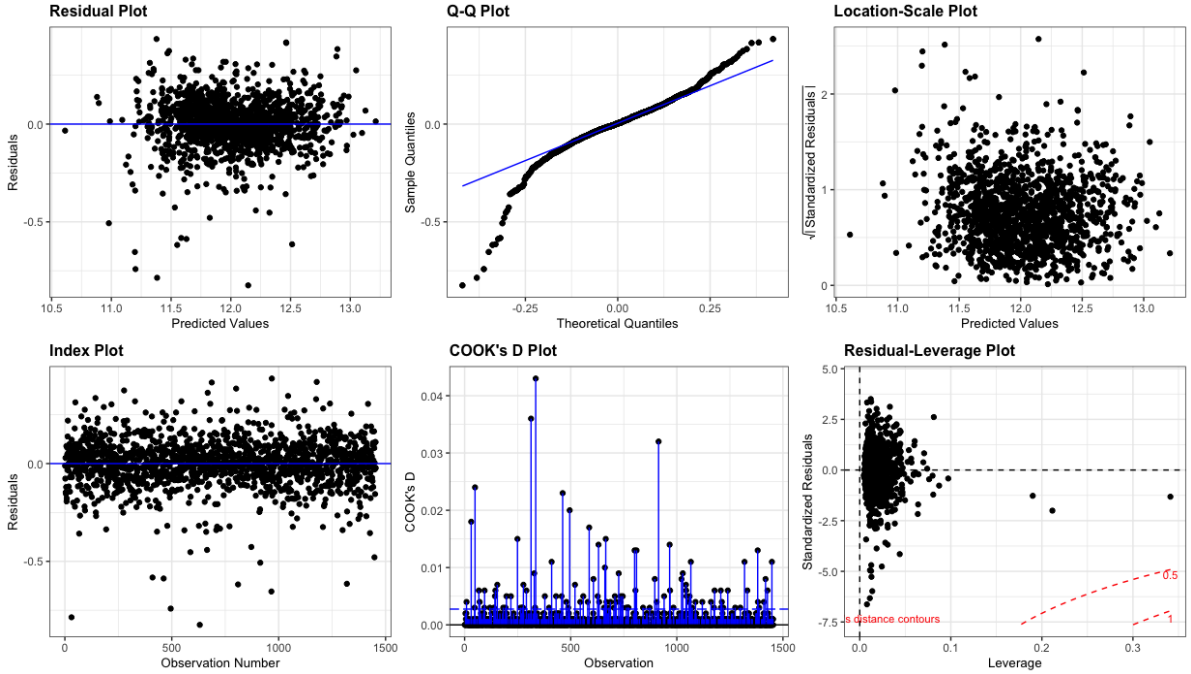
## Checking Assumptions: Residual Plots for Forward, Backward, and Stepwise Models

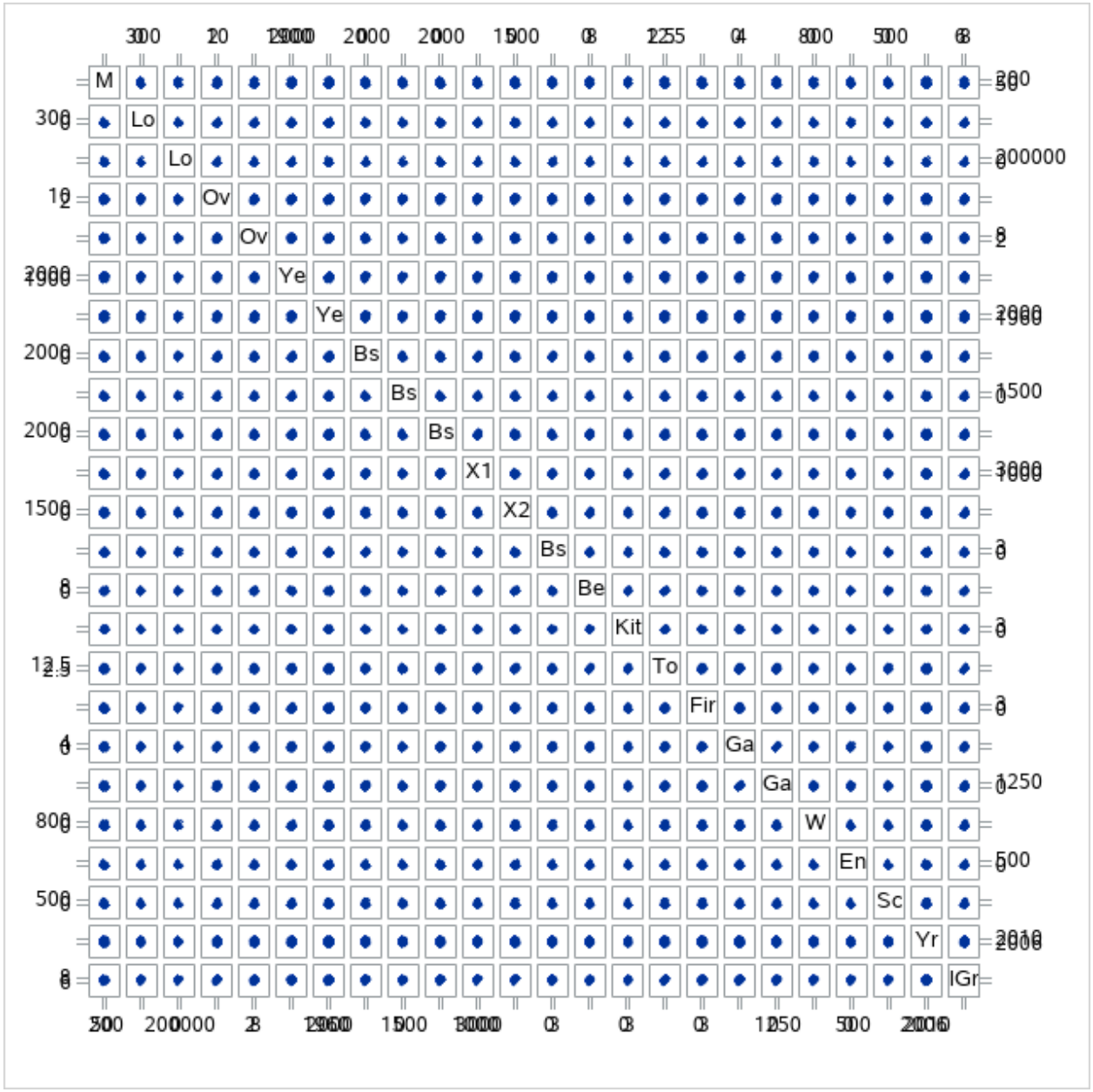
# Forward


Residual Plot
Residuals
Predicted Values
Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Location-Scale Plot
√|Standardized Residuals|
Predicted Values
Index Plot
Residuals
Observation Number
COOK's D Plot
COOK's D
Observation
Residual-Leverage Plot
Standardized Residuals
Leverage
Cook's distance contours
0.5
1

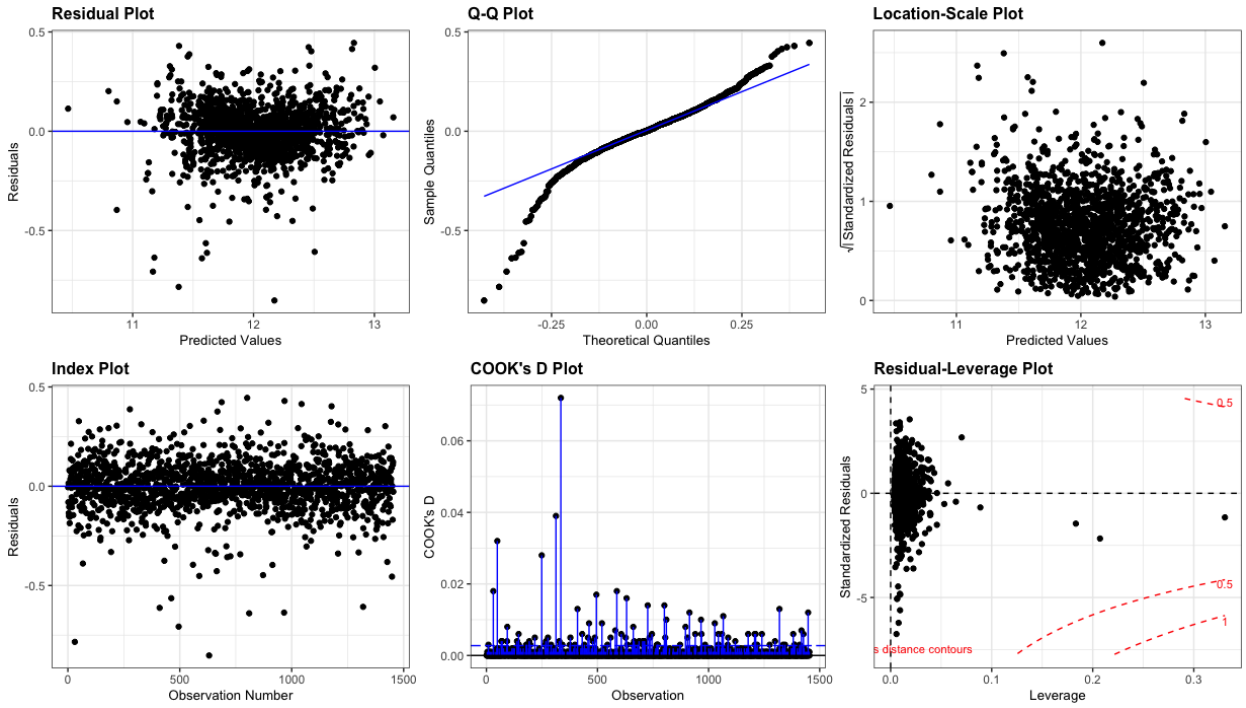

MS Su
Lot Are
Ov eral
Ov eral
Ye arB
Ye arB
Bs mt
Tot alB
X1 stFl
X2 ndF
Bs mt
Kit che
Fire pla
Gar age
Gar age
Wo od
Scr een
YrS old
IGr Liv

# Backward


Residual Plot
Residuals
Predicted Values
Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Location-Scale Plot
√|Standardized Residuals|
Predicted Values
Index Plot
Residuals
Observation Number
COOK's D Plot
COOK's D
Observation
Residual-Leverage Plot
Standardized Residuals
Leverage
0.5
1




Stepwise


Residual Plot
Residuals
Predicted Values
Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Location-Scale Plot
√| Standardized Residuals |
Predicted Values
Index Plot
Residuals
Observation Number
COOK's D Plot
COOK's D
Observation
Residual-Leverage Plot
Standardized Residuals
Leverage
0.5
0.5
1
s distance contours