Assignment #2

Andrea Bruckner

Introduction

In this report we are using the Ames Housing data set to begin building regression models that can predict the sale price of typical single-family homes in Ames, Iowa. After familiarizing ourselves with the data in Assignment #1, we have selected two continuous variables, GrLivArea and TotalBsmtSF, that indicate a promising ability to predict home sale price, SalePrice. We will be analyzing the goodness-of-fit of each model, as well as refitting each model with a logarithmic transformation of the response variable, in order to determine which model can best predict home sale prices.

Data

The Ames Housing data set is from the Ames Assessor's Office. It contains 2930 observations with 82 variables concerning residential properties sold between 2006 and 2010 in Ames, Iowa. The variables consist of 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables, as well as 2 additional observation variables. A broad overview of the data in Assignment #1 revealed that we have enough data to predict home prices. The data set contains a variety of categorical and numeric variables from which we can draw conclusions, and the data set contains very few missing values.

Sample Population

Because not all of the properties in the Ames Housing data set are representative of a typical home, we created a sample population to focus on just the relevant data for our model. **Table 1** below illustrates the criteria that would cause an observation to drop from our sample. The seventh drop condition, 07: Pricing Error, does not appear in the table because all of the SalePrice data met the condition that the housing price must be greater than $0. The sample population totals 1831 observations and represents 62.5 percent of the original data set.

Drop Condition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01: Not a Single Family Home	505	17.24	505	17.24
02: Non-Residential Zone or RV Park	103	3.52	608	20.75
03: Not Normal Sale Condition	379	12.94	987	33.69
04: Atypically Large House	1	0.03	988	33.72
05: No Basement	44	1.50	1032	35.22
06: No Garage	66	2.25	1098	37.47
08: Not All Public Utilities Available	1	0.03	1099	37.51
09: Sample Population	1831	62.49	2930	100.00

Table 1: Drop Conditions for the Sample Population

We decided on these drop conditions based on what a typical homebuyer would likely consider when purchasing a home. Most homebuyers are seeking single-family, non-mobile homes; are purchasing the home under normal conditions (i.e., not buying the home from a family member

or through a short sale); are buying a normal sized home (i.e., not a mansion); and are looking for homes that include a basement, garage, and all public utilities.

Exploratory Data Analysis

We conducted an initial exploratory data analysis (EDA) in Assignment #1 to familiarize ourselves with the data set, and we continued our EDA in this assignment to determine which variables we should focus on for our models. We chose two continuous variables, GrLivArea and TotalBsmtSF, as the most promising for predicting SalePrice. **Table 2** shows that both GrLivArea and TotalBsmtSF are strongly correlated (0.77 and 0.65, respectively) with the response variable, SalePrice. The histograms in Figures 1 and 2 show that while these variables are not normally distributed, they do not deviate too much from a normal distribution. These findings helped us select these predictor variables from the continuous variables available in the data set. These predictor variables alsoo represent the total amount of usable space within a home, which is something homebuyers typically prioritize and are willing to pay more for when buying a new home.

Pearson Correlation Coefficients, N = 1831 Prob > |r| under H0: Rho=0			
	SalePrice	**GrLivArea**	**TotalBsmtSF**
SalePrice	1.00000	0.76953 <.0001	0.65001 <.0001
GrLivArea	0.76953 <.0001	1.00000	0.39289 <.0001
TotalBsmtSF	0.65001 <.0001	0.39289 <.0001	1.00000

Table 2: Correlation Coefficents for Response and Predictor Variables

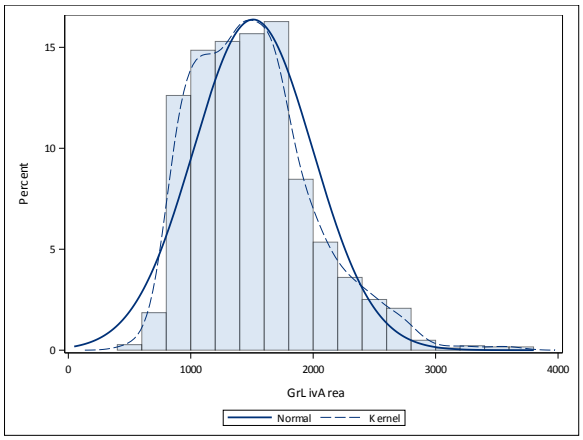

Figure 1: Histogram of GrLivArea

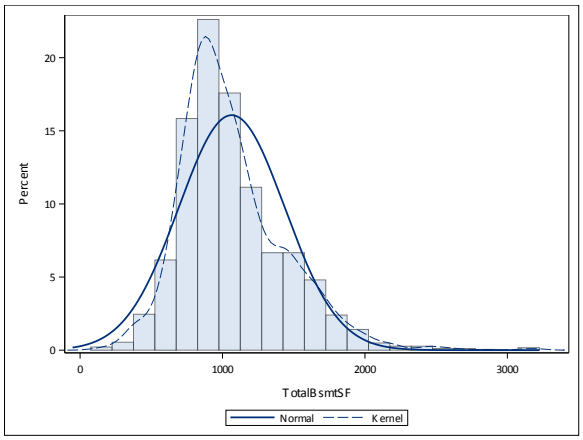

Figure 2: Histogram of TotalBsmtSF

Model 1: Simple Linear Regression of GrLivArea v. SalePrice

The results of Model 1 reveal a mediocre goodness-of-fit (GOF). Because the p-value is very low (<.0001) and the F-statistic is very high (2655.82) (see **Table 3**), we can conclude that GrLivArea has a significant effect on SalePrice. Per **Table 4**, the fitted equation for this model is **SalePrice = 11911 + 112.89*GrLivArea.** For every unit increase in GrLivArea, an approximate 112.89 unit increase in SalePrice is predicted. That is, as the above ground living space increases by 1 square foot, the home sale price increases by $112.89.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	5.53506E12	5.53506E12	2655.82	<.0001
Error	1829	3.811865E12	2084125219		
Corrected Total	1830	9.346925E12			

Table 3: ANOVA Table for Model 1

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	11911	3474.87983	3.43	0.0006	5095.79209	18726
GrLivArea	1	112.89116	2.19059	51.53	<.0001	108.59484	117.18748

Table 4: Parameter Estimates Table for Model 1

Although the results in the above tables indicate a strong correlation between GrLivArea and SalePrice, further analysis of the results show that the predictive model has plenty of room for improvement. The R-squared value (0.5922) and adjusted R-squared value (0.5920) shown in **Figure 3** are nearly equivalent and indicate that approximately 59 percent of the variation can be explained by the model, while 41 percent is not.

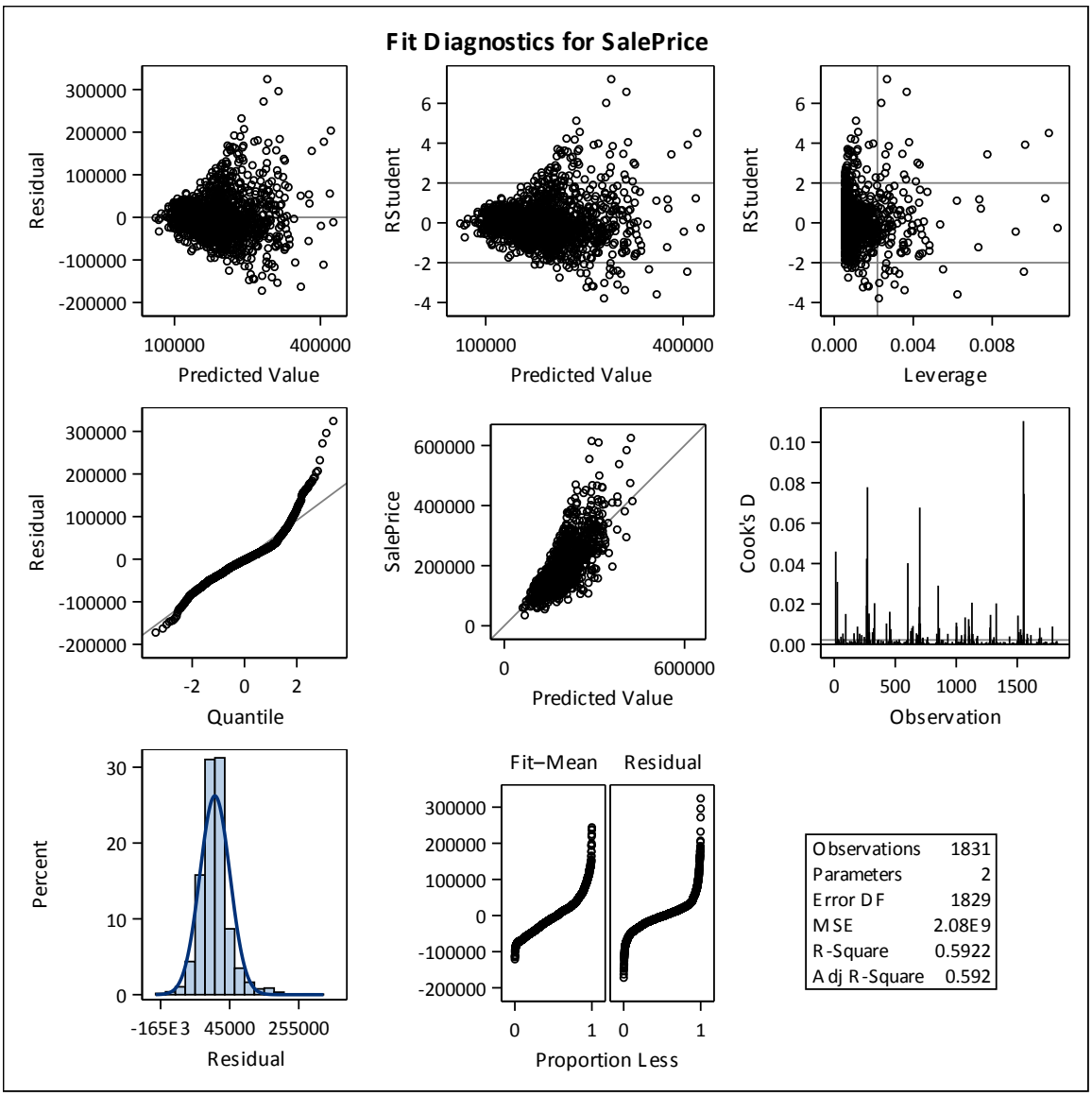

Figure 3: Fit Diagnostics for Model 1

Figure 3 contains a lot of indicators about the GOF of Model 1. The residual-leverage plot indicates that there are several observations that may be outliers (the points that fall outside the horizontal lines), as well as high-leverage points (the points to the right of the vertical line). Because the Q-Q plot falls below the line on the left and above the line of the right, it indicates that the sample distribution has slightly longer tails than a normal distribution. Overall, the plotted points also increase from left to right, indicating a right-skewed distribution. The residual histogram confirms these conclusions.

The plot of the response variable v. predicted value in the center of **Figure 3** indicates that the model is not the best fit. The points should ideally all fall on the line, or be very close to it, and not exhibit an upward curve. In the residual-fit plot right below it, the residual side is taller than the fit side, indicating that there is still some unexplained variation in the model. Additionally, the Cook's D plot shows several high peaks, which indicate that some observations require further investigation since they might have a large effect on the coefficients.

Both the residuals and studentized residuals v. predicted values plots in **Figure 3** indicate that the assumption of constant variance is not likely true since the plots each show an increasing trend. This pattern is also apparent in the residuals v. predictor plot in **Figure 4**. As the predictor variable increases, the error variance also increases.

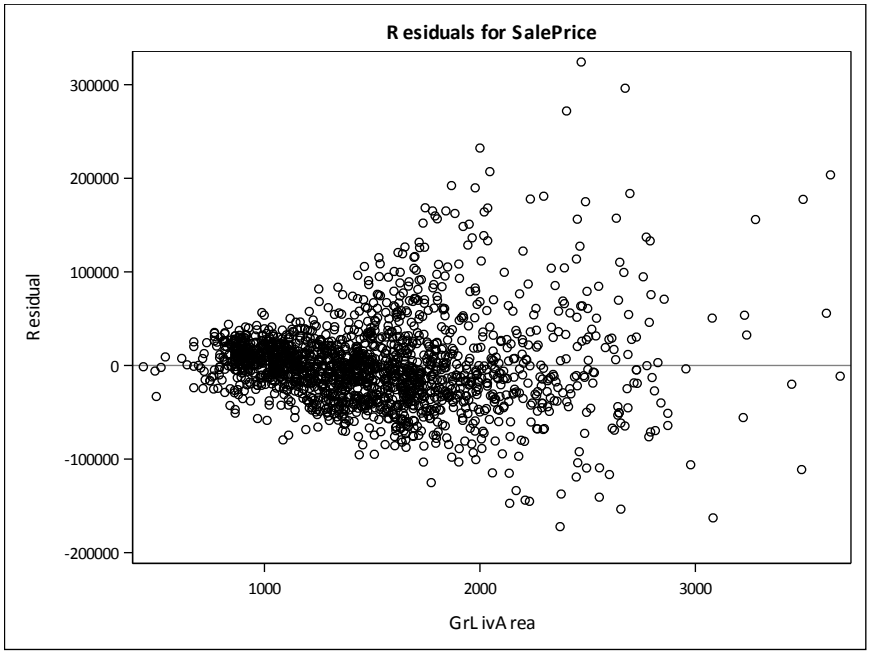

Figure 4: Model 1 GrLivArea v. Residuals for SalePrice

Figure 5 shows the overall fit of the model. While many of the data points are along the regression line, there are many data points that fall outside both the Confidence Limits and Prediction Limits. This indicates that Model 1 should undergo modifications that will hopefully improve the model's GOF.

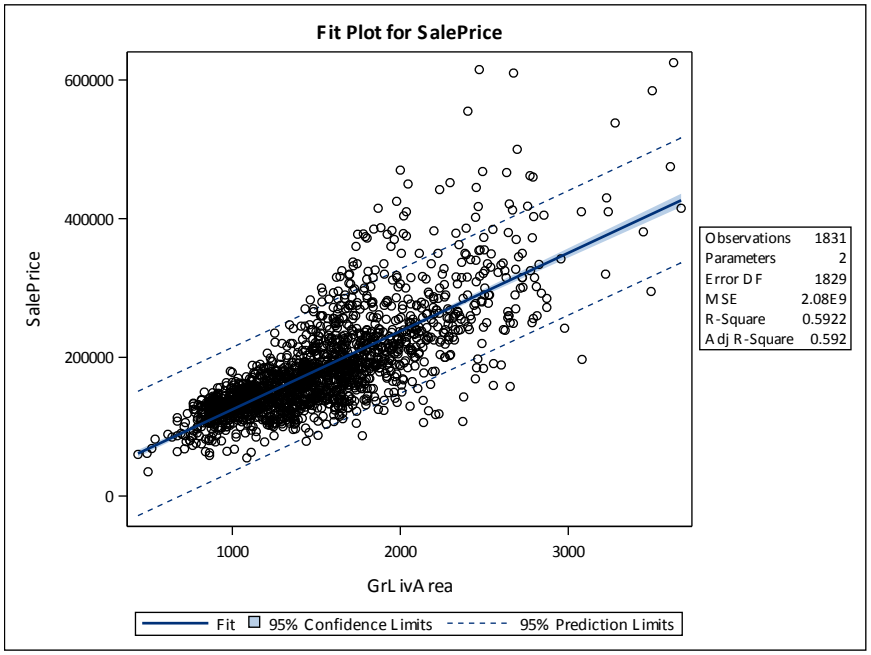

Figure 5: Fit Plot for Model 1

Model 2: Simple Linear Regression of TotalBsmtSF v. SalePrice

The regression output for Model 2 has some similarities to Model 1, and it also has a mediocre GOF. **Table 5** reveals that Model 2, like Model 1, has a very low p-value and a very high F-statistic, indicating that TotalBsmtSF has a significant effect of SalePrice. In **Table 6** we can determine that the fitted equation for this model is **SalePrice = 49547 + 124.75*TotalBsmtSF.**

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	3.949152E12	3.949152E12	1338.14	<.0001
Error	1829	5.397773E12	2951215198		
Corrected Total	1830	9.346925E12			

Table 5: ANOVA Table for Model 2

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	49547	3845.70511	12.88	<.0001	42004	57089
TotalBsmtSF	1	124.75046	3.41029	36.58	<.0001	118.06200	131.43893

Table 6: Parameter Estimates Table for Model 2

In **Figure 6**, we can see that the R-squared and adjusted R-squared values are approximately 0.42, which indicates that this model accounts for 42 percent of the variation in the data. This is 17 percent less than Model 1, which suggests that Model 1 is a better fit than Model 2.

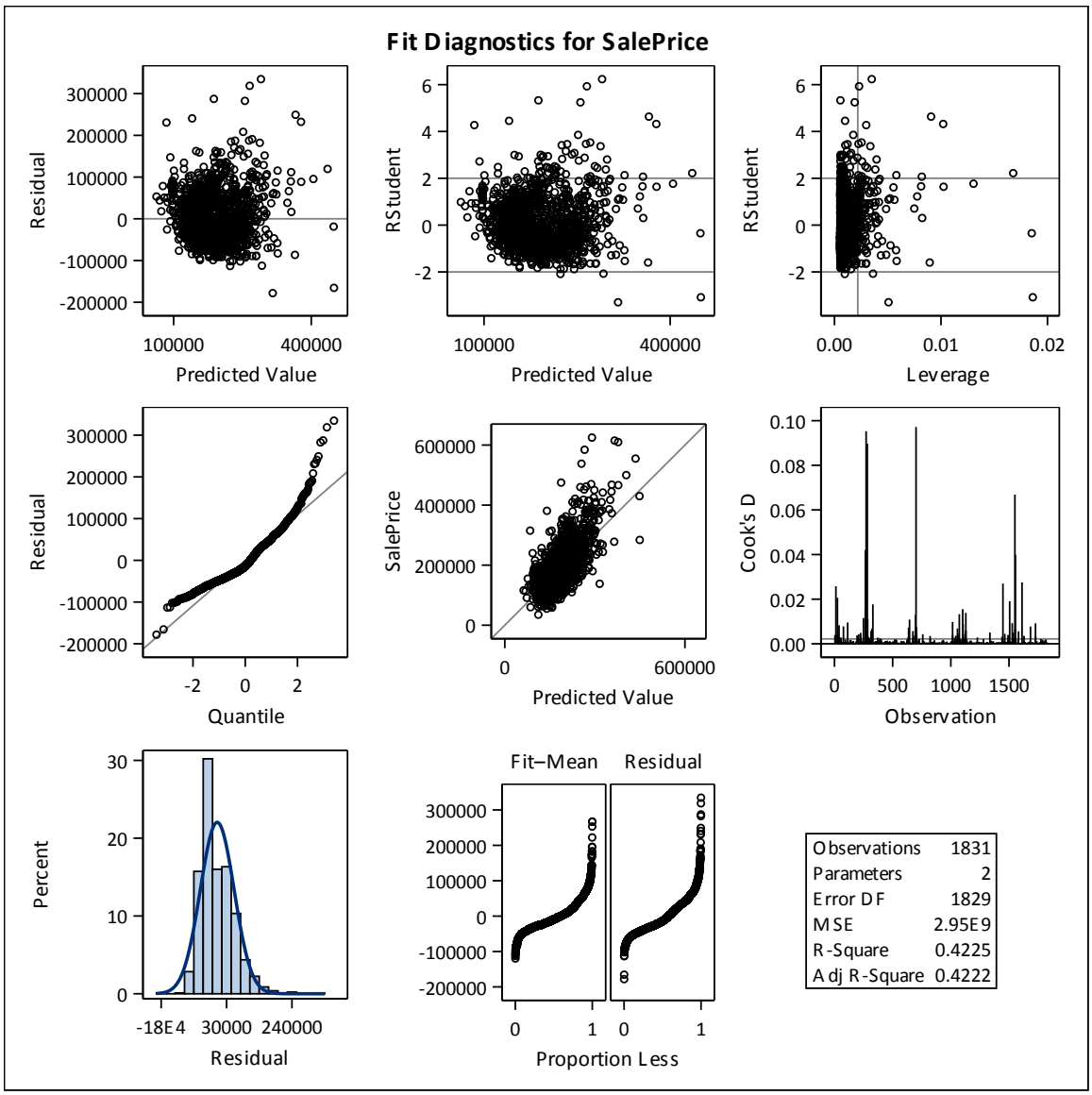

Figure 6: Fit Diagnostics for Model 2

The residual-leverage plot in **Figure 6** suggests the same conclusions as the equivalent plot in **Figure 3**: There are observations that are indicative of being outliers, high-leverage point, or both. The Q-Q plot and histogram both show a right-skewed distribution with several potential outliers. The plot of the response variable v. predicted value in the center of **Figure 6** indicates that the model generally fits but could fit better to avoid the slight upward curve shown in the plot. Similar to Model 1's plot in **Figure 3**, the Cook's D plot in **Figure 6** shows several high peaks, indicating further investigation into some observations is necessary. The residual-fit plot in **Figure 6** also is similar to the same plot in **Figure 3** since the residual side is taller than the fit side, indicating there is unexplained variance in the model.

In **Figure 6**, the residuals and studentized residuals v. predicted values plots indicate that the assumption of constant variance is likely true since there is no discernable pattern. The residuals v. predictor plot in **Figure 7**, however, shows a curved clustering of values, as well as stacked values near the 300 value mark. Both of these observations suggest that further

analysis is necessary. While more of the data points fall within the Prediction Limits in **Figure 8** than in **Figure 5**, the clustering of the data points in **Figure 8** and the low R-square and adjusted R-square values of Model 2 indicate that there is still some unexplained variation in the model that warrants modification in attempt to produce a better GOF.

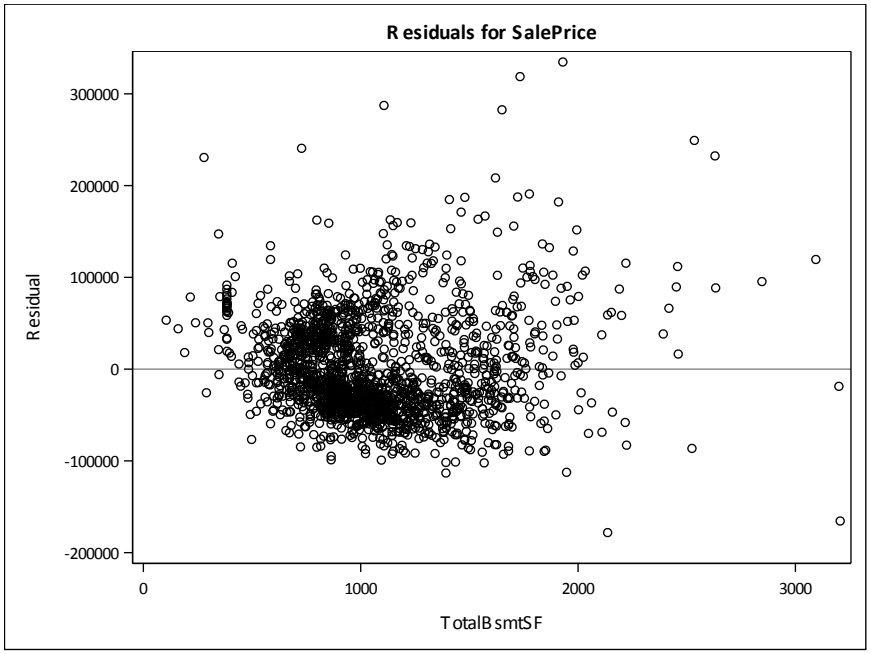

Figure 7: Model 2 TotalBsmtSF v. Residuals for SalePrice

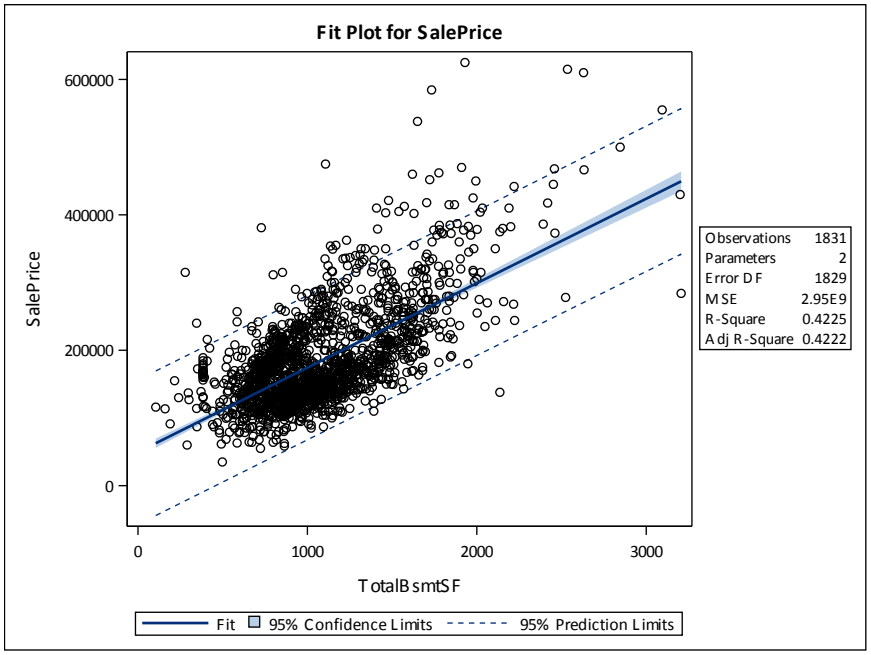

Figure 8: Fit Plot for Model 2

Model 3: Multiple Linear Regression of GrLivArea and TotalBsmtSF v. SalePrice

Model 3 fits better than Models 1 and 2, but it still does not have a very strong GOF. As in the simple linear regression models, Model 1 and Model 2, explored above, this multiple linear regression model shows that the combination of the predictor variables (GrLivArea and TotalBsmtSF) have a significant effect on SalePrice, as indicated by the p-value and F-statistic in **Table 7**. The fitted equation for this model is **SalePrice = -36306 + 89.19*GrLivArea + 78.90*TotalBsmtSF** (see **Table 8**).

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	6.871044E12	3.435522E12	2536.53	<.0001
Error	1828	2.47588E12	1354420168		
Corrected Total	1830	9.346925E12			

Table 7: ANOVA Table for Model 3

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	-36306	3194.37188	-11.37	<.0001	-42571	-30041
GrLivArea	1	89.19477	1.92037	46.45	<.0001	85.42843	92.96111
TotalBsmtSF	1	78.90409	2.51232	31.41	<.0001	73.97676	83.83141

Table 8: Parameter Estimates Table for Model 3

The R-squared and adjusted R-squared values in **Figure 9** both show that the model explains approximately 73.5 percent of the variation in the data. This is higher than the result for either Model 1 or Model 2, which indicates that Model 3 might be a better fit. An examination of the other outputs in **Figure 9**, however, suggests that this model has several similar issues as the simple linear regression models. Neither the residuals nor studentized residuals v. predicted values plots indicate constant variation since both plots show an increasing trend. The residual-leverage plot suggests the presence of outliers and high-leverage points. The Q-Q plot and residual histogram indicate an approximately normal distribution with potential outliers. The plot of the response variable v. predicted value indicates that the model, in general, fits, with the exception of some possible outliers. As in the output for Models 1 and 2, the Cook's D plot for Model 3 shows several high peaks, representing observations that require closer analysis. The residual-fit plot for Model 3 differs from the plots in Models 1 and 2 in that the fit side is taller than the residual side, indicating that the model generally accounts for the variation in the data.

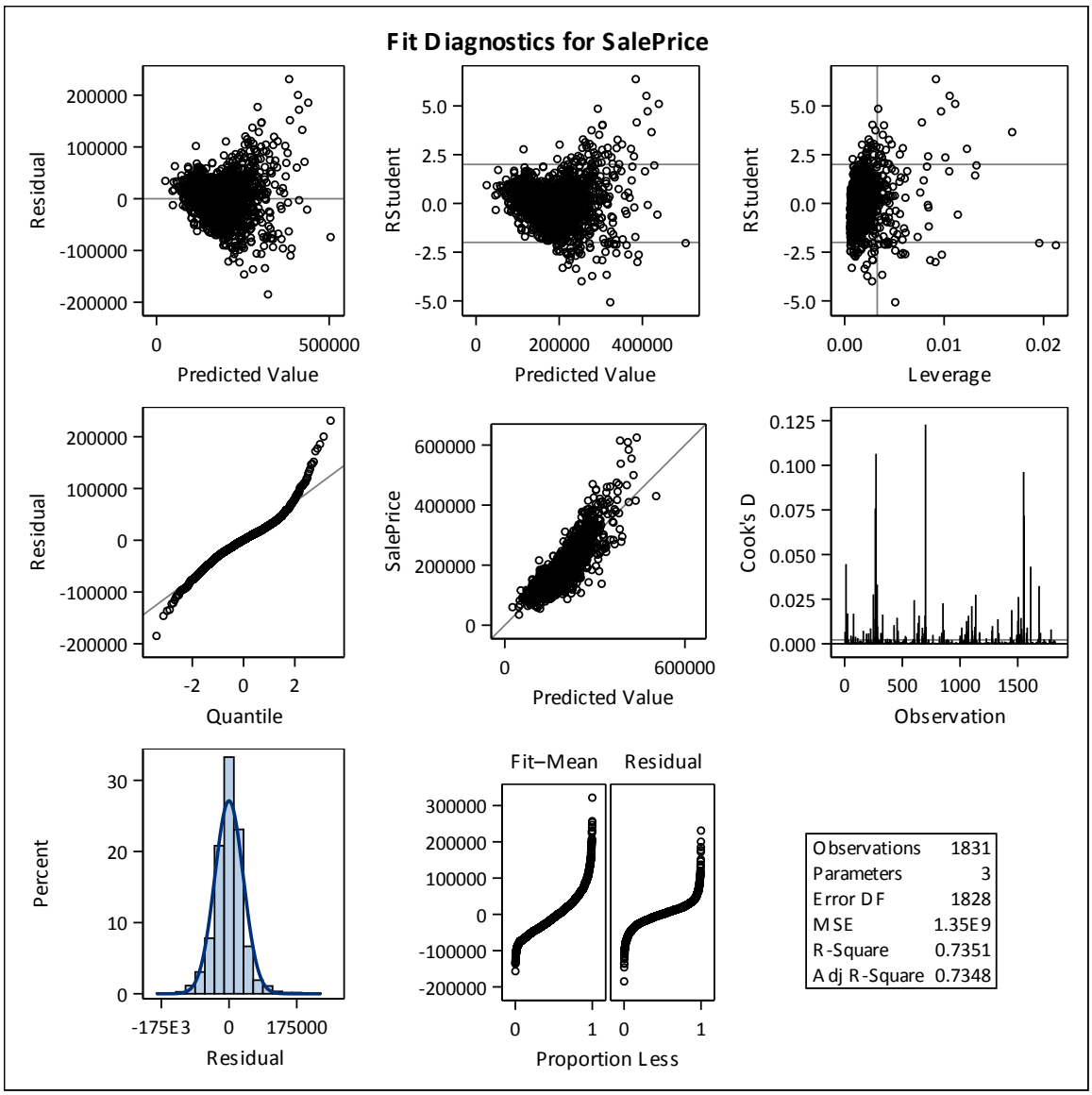

Figure 9: Fit Diagnostics for Model 3

Multiple linear regression models do not produce a fit plot similar to **Figures 5 and 8**, but Model 3 does include the residuals v. predictors plots. The plots in **Figure 10** reveal that the data is not as random and linear as assumed.

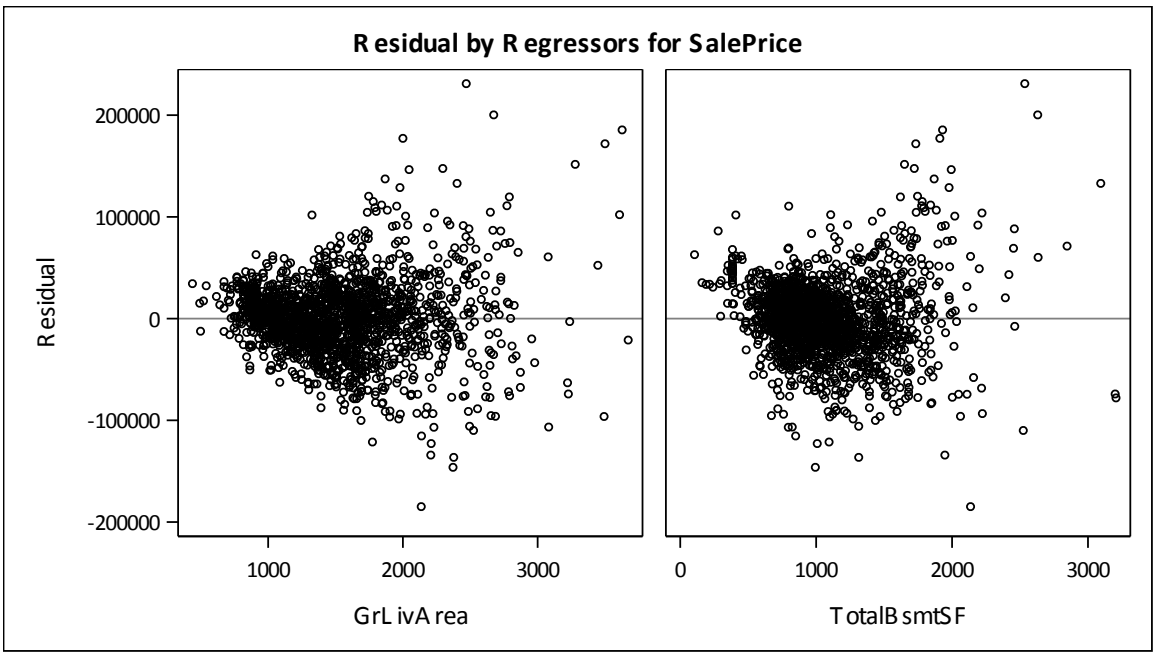

Figure 10: Residuals for Model 3

Logarithmic Transformation of the Response Variable SalePrice

During our EDA, we created a couple histograms of SalePrice and the logarithmic transformation of SalePrice so that we could get a sense of whether a log transformation of the response variable might be helpful in creating our model. The log transformation of SalePrice, as shown in **Figure 12**, follows a much more normal distribution that the untransformed response variable shown in **Figure 11**.

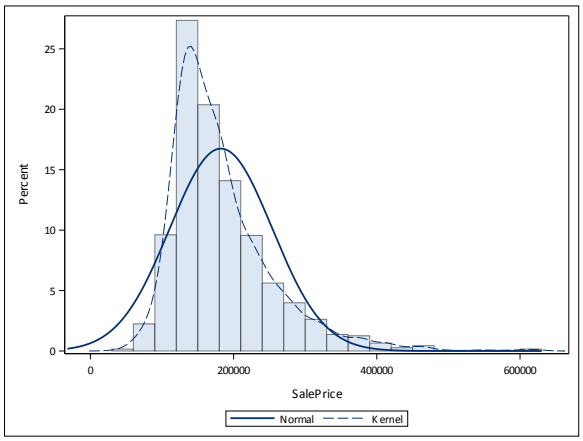

Figure 11: Histogram of SalePrice

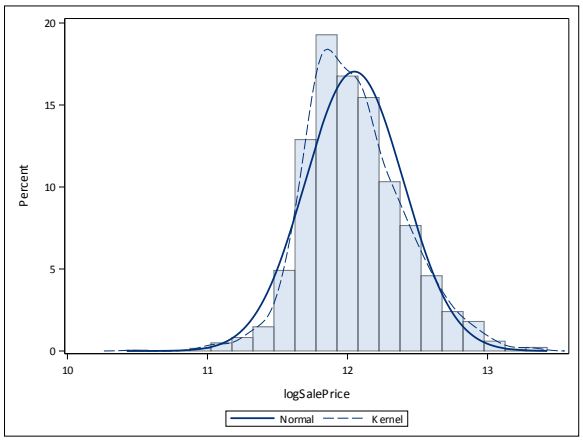

Figure 12: Histogram of log(SalePrice)

In the next few sections we will examine the models created with the same predictor variables but substitute log(SalePrice) for SalePrice as the response variable. We created the variable logSalePrice to represent log(SalePrice), and the figures are labeled as such. We will compare the refitted models with the original models in order to determine the GOF after the log transformation, and we will try to determine which model fits best overall.

Model 4: Simple Linear Regression of GrLivArea v. log(SalePrice)

Performing a log transformation of SalePrice resulted in a more normal distribution of the residuals for Model 4. This is evident when we compare **Figures 13 and 14** below. The log transformation also caused the residuals distribution to shift from a right skew to a left skew, as indicated by the curvature of the Q-Q plots in **Figures 15 and 16**, as well as in the histograms.

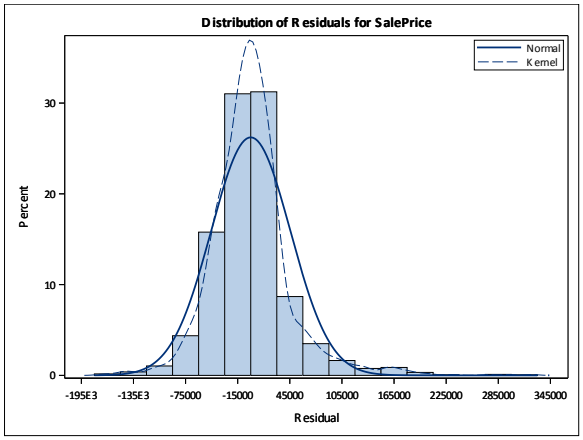

Figure 13: Residuals Histogram for Model 1

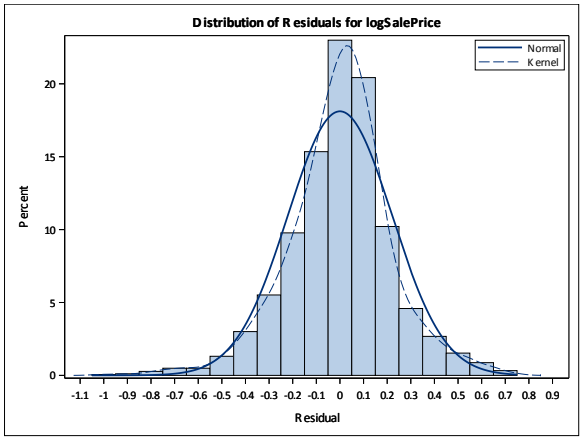

Figure 14: Residuals Histogram for Model 4

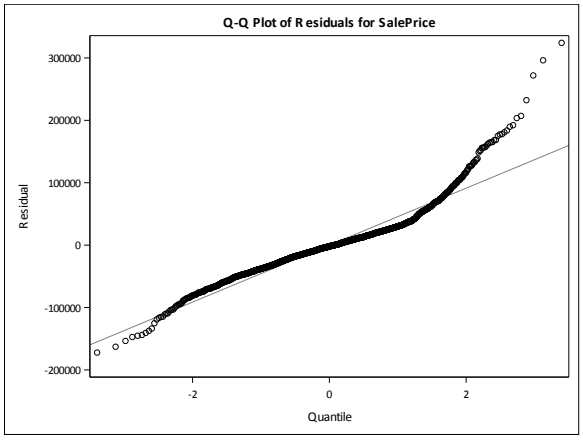

Figure 15: Residuals for Model 1

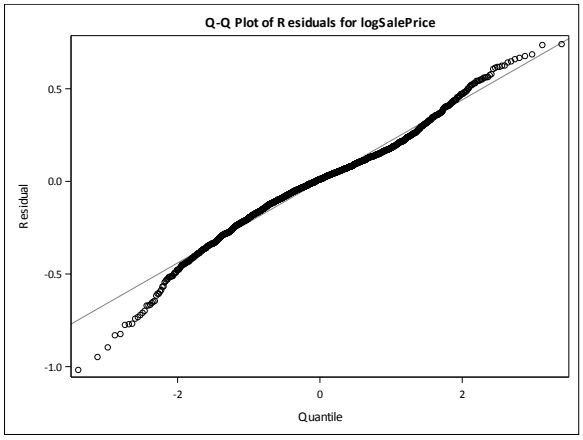

Figure 16: Residuals for Model 4

When comparing the residual plots for both the original and refitted models, we can see that the refitted model, **Figure 18**, exhibits more randomness than the original model in **Figure 17**. This suggests that the assumption of constant variance is more likely true in the refitted model than in the original model.

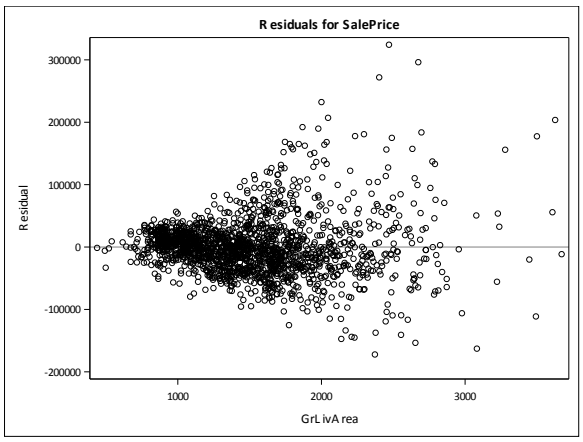

Figure 17: Residuals for Model 1

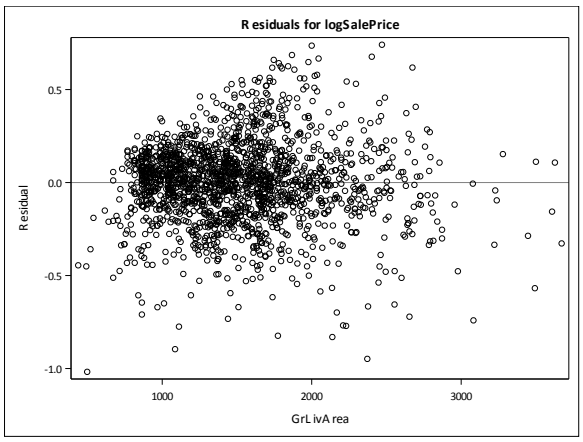

Figure 18: Residuals for Model 4

In the residual-fit plots, the residual side in the refitted model is taller than the fit side (see **Figure 20**), indicating that the model mostly accounts for the variation in the data, unlike Model 1 (see **Figure 19**), which suggests that there is still some unexplained variation in the model.

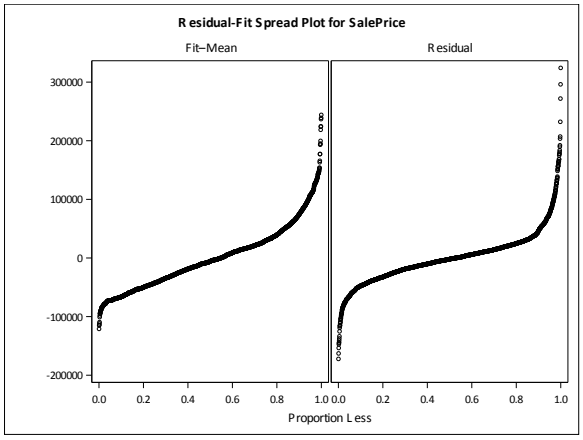

Figure 19: Residual-Fit Spread for Model 1

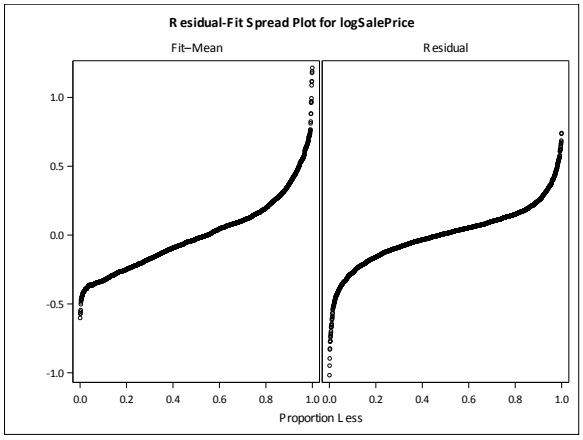

Figure 20: Residual-Fit Spread for Model 4

The fit plots for the original model (**Figure 21**) and the refitted model (**Figure 22**) both show that the data points cluster around the regression line, yet some of them extend beyond the Prediction Limits. **Figure 22** reveals that Model 4 fits much better than Model 1, despite having only slightly higher R-squared and adjusted R-squared scores. Per **Figure 22**, approximately 60.7 percent of the variation in the data can be explained by Model 4. Though it fits better than Model 1, Model 4 has only a mediocre GOF.

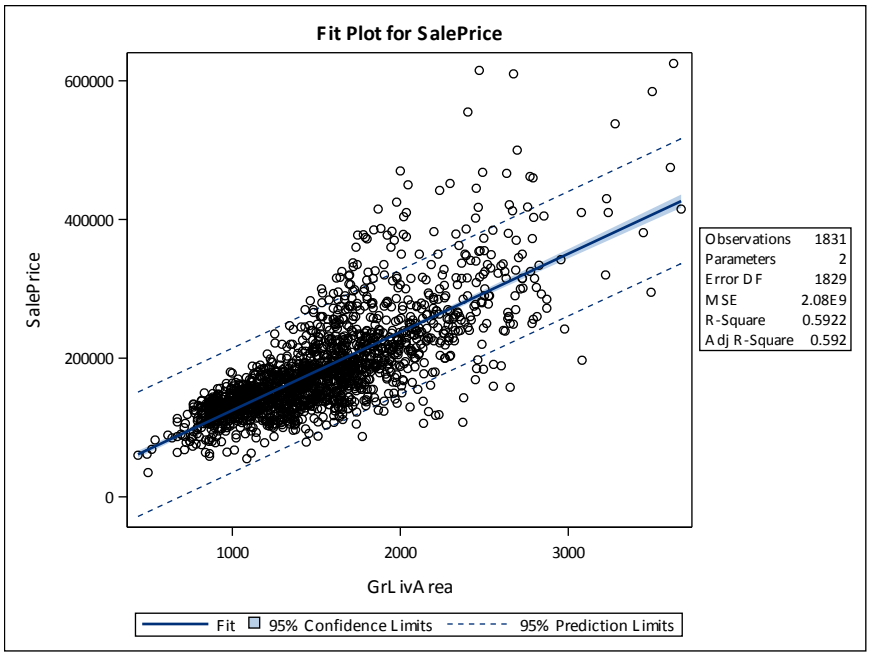

Figure 21: Fit Plot for Model 1

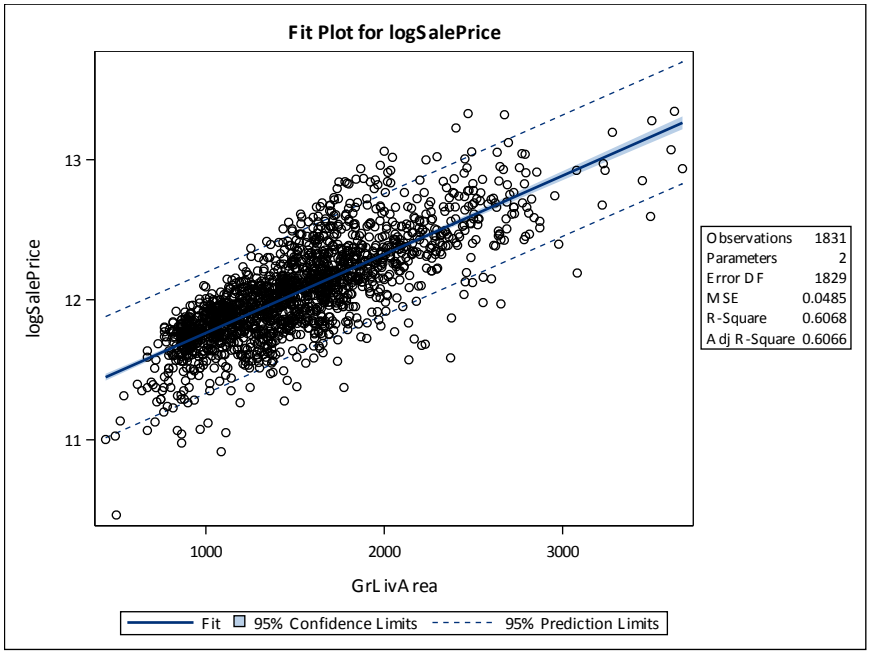

Figure 22: Fit Plot for Model 4

Model 5: Simple Linear Regression of TotalBsmtSF v. log(SalePrice)

Performing a log transformation of SalePrice resulted in a more normal distribution of the residuals for Model 5. The log transformation caused the residuals distribution to lose its left skew that was present in Model 2. This change is evident in the Q-Q plots in **Figures 25 and 26**, as well as in the histograms, **Figures 23 and 24**.

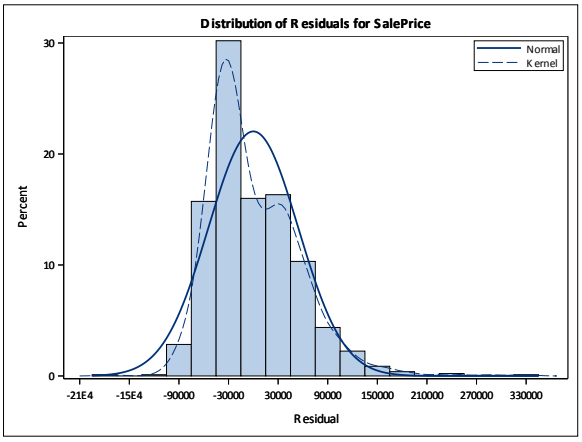

Figure 23: Residuals Histogram for Model 2

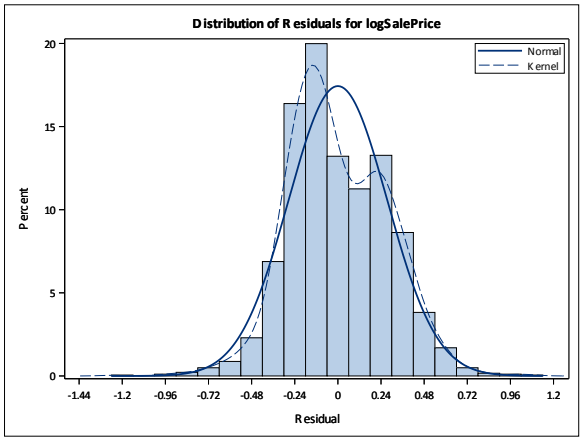

Figure 24: Residuals Histogram for Model 5

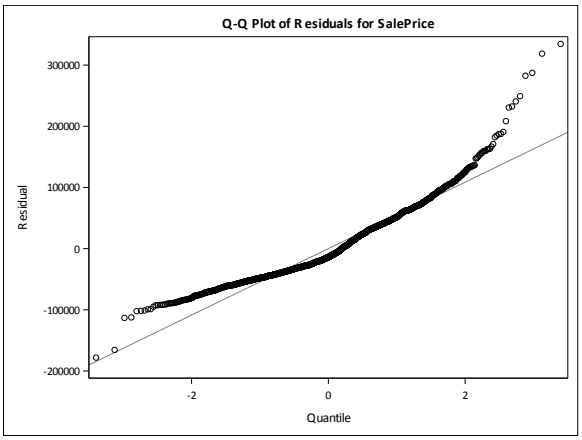

Figure 25: Residuals for Model 2

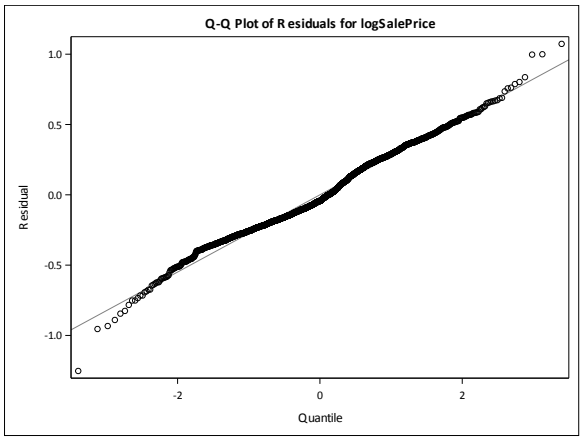

Figure 26: Residuals for Model 5

In the residual plots for both the original and refitted models, we can see that the refitted model, **Figure 28**, exhibits more randomness in variation than the original model in **Figure 27**. The refitted model better supports the assumption of constant variance than in the original model. Both models still, however, show clustering and a stack of points near the 300 mark.

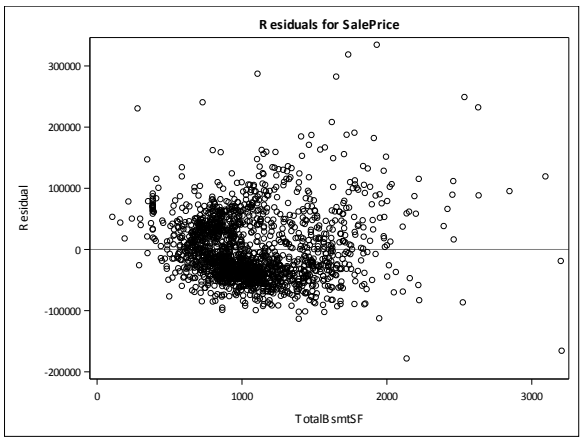

Figure 27: Residuals for Model 2

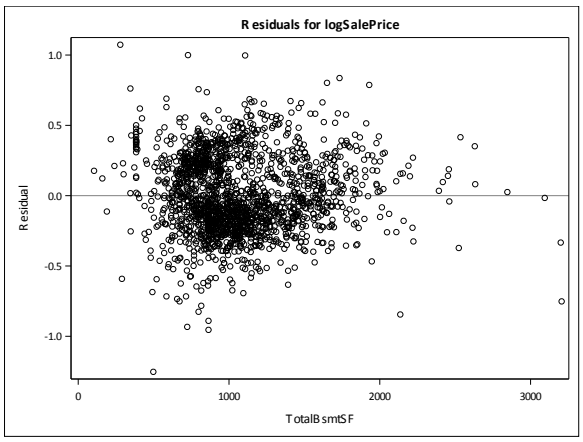

Figure 28: Residuals for Model 5

In the residual-fit plots, the fit side in the refitted model is taller than the residual side (see **Figure 30**), indicating that the model mostly accounts for the variation in the data, unlike Model 2 (see **Figure 29**), which suggests that there is still some unexplained variation in the model.

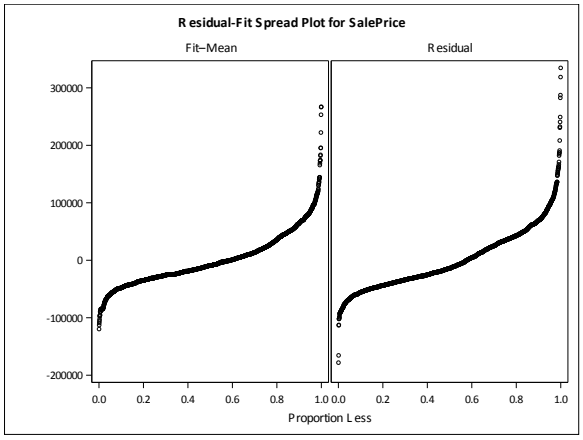

Figure 29: Residual-Fit Spread for Model 2

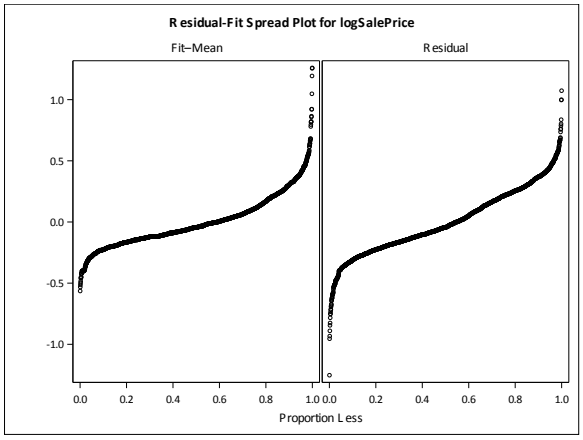

Figure 30: Residual-Fit Spread for Model 5

The fit plots for the original model (**Figure 31**) and the refitted model (**Figure 32**) both show that the data points mostly fall within the Prediction Limits, yet both models show some points beyond the Prediction Limits. Although **Figure 31** contains data points that are farther from the regression line than in Figure 32, the overall R-squared and adjusted R-squared scores are higher for Model 2 than for Model 5. Per **Figure 32**, approximately 38.9 percent of the variation in the data can be explained by Model 5. Although some of the other outputs indicate that Model 5 fits better than Model 2, the lower scores indicate otherwise.

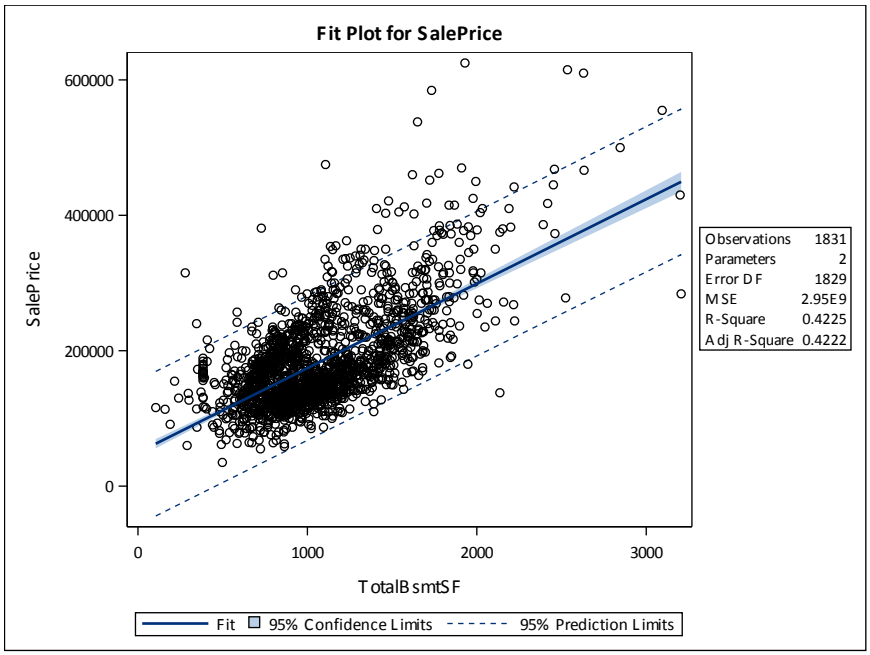

Figure 31: Fit Plot for Model 2

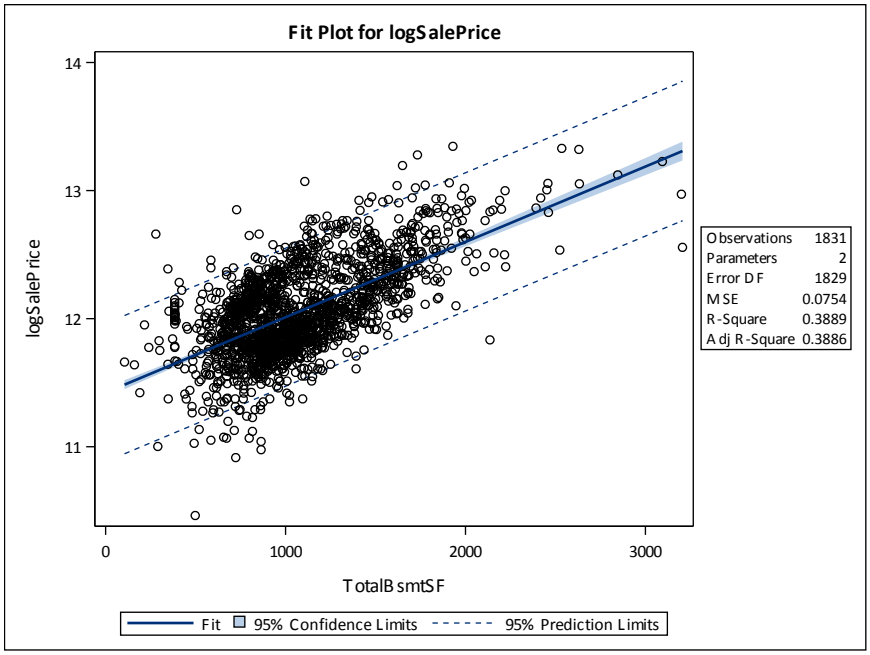

Figure 32: Fit Plot for Model 5

Model 6: Multiple Linear Regression of GrLivArea and TotalBsmtSF v. log(SalePrice)

Performing a log transformation of SalePrice did not change the distribution much. **Figure 33** shows that the residuals resemble a normal distribution for Model 3, and **Figure 34** shows the residuals fall into a similar distribution, though slightly less normal. The residuals distribution shifted from a right skew to a left skew, as indicated by the curvature of the Q-Q plots in **Figures 35 and 36**.

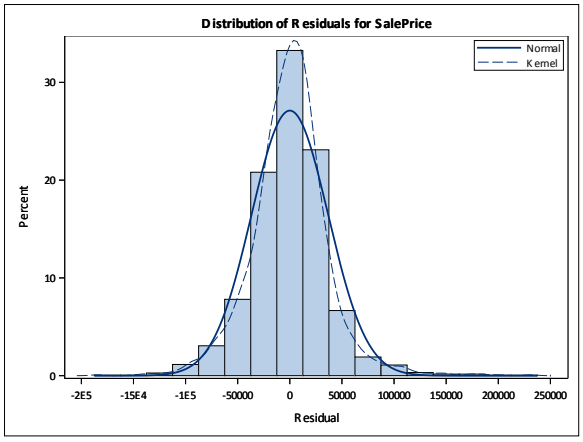

Figure 33: Residuals Histogram for Model 3

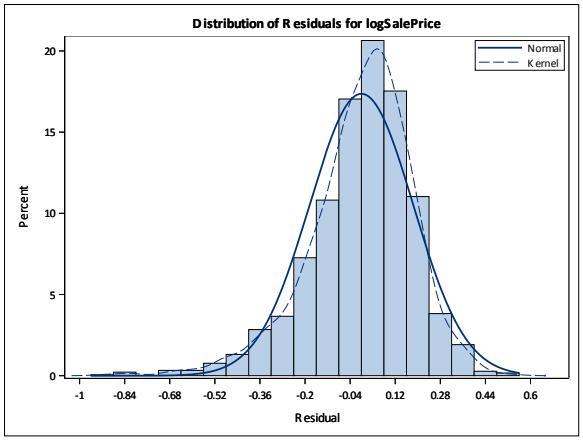

Figure 34: Residuals Histogram for Model 6

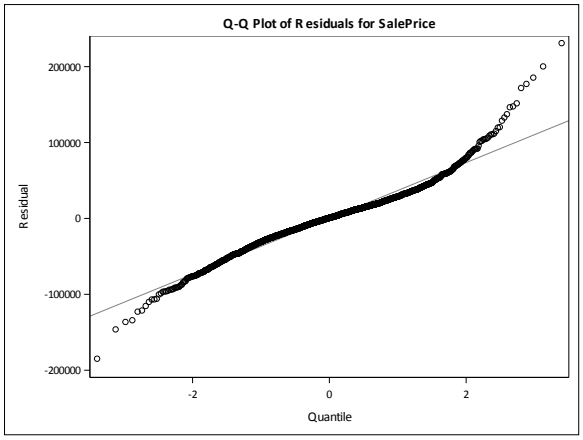

Figure 35: Residuals for Model 3

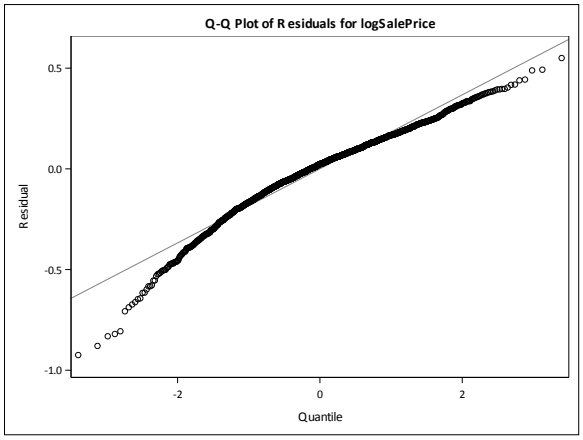

Figure 36: Residuals for Model 6

In the residual plots for both the original and refitted models, we can see that the refitted model, **Figure 38**, exhibits more randomness in variation than the original model in **Figure 37**. The refitted model better supports the assumption of constant variance than in the original model. Both models still, however, show a stack of points near the 300 mark for TotalBsmtSF, which warrants further investigation.

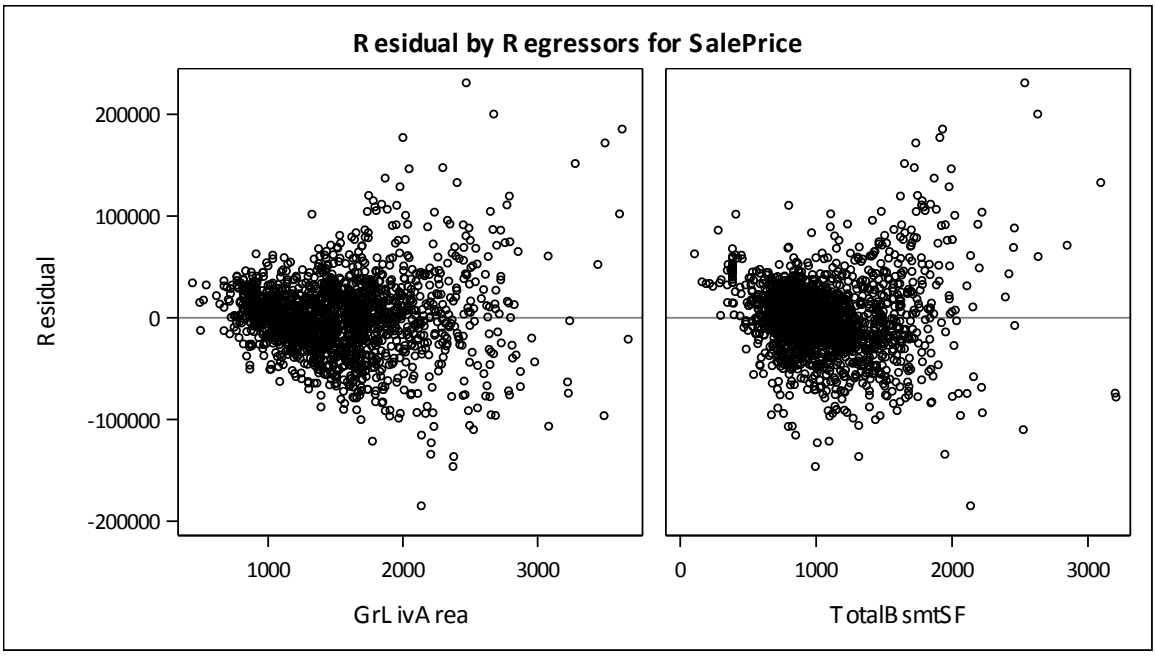

Figure 37: Residuals for Model 3

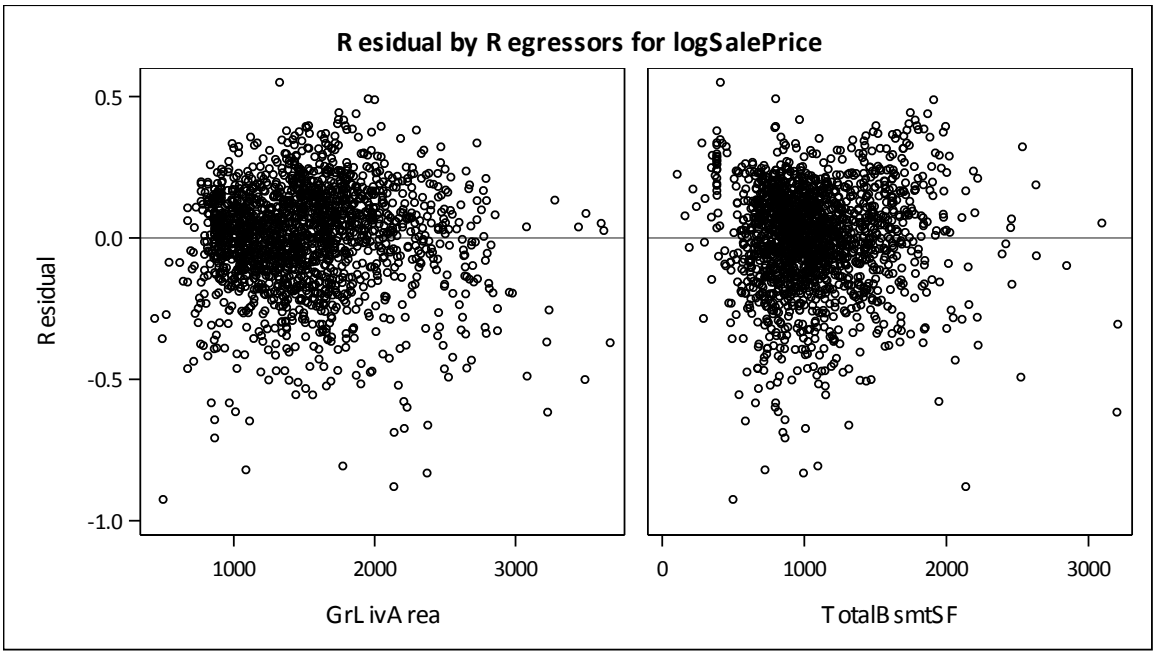

Figure 38: Residuals for Model 6

In both of the residual-fit plots in **Figures 39 and 40**, the fit side is taller than the residual side, suggesting that the model mostly accounts for the variation in the data. The difference in height in **Figure 40** is greater than in **Figure 39** and suggests that Model 6 fits better than Model 3.

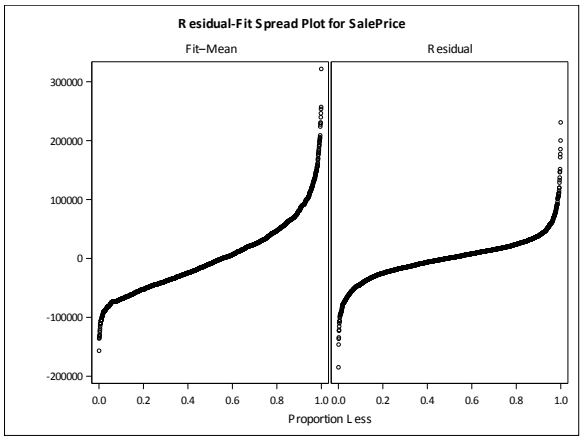

Figure 39: Residual-Fit Spread for Model 3

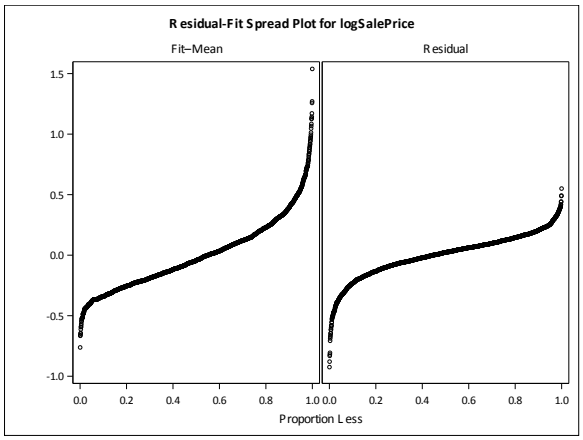

Figure 40: Residual-Fit Spread for Model 6

However, a comparison of the adjusted R-squared values in Tables 9 and 10 suggest that Model 3 fits slightly better than Model 6. According to the tables Model 3 accounts for approximately 73.5 percent variation in the data while Model 6 accounts for 72.6 percent.

Root MSE	36802	**R-Square**	0.7351
Dependent Mean	182338	**Adj R-Sq**	0.7348
Coeff Var	20.18361		

Table 9: Overall Model Fit Table for Model 3

Root MSE	0.18393	**R-Square**	0.7261
Dependent Mean	12.04898	**Adj R-Sq**	0.7258
Coeff Var	1.52649		

Table 10: Overall Model Fit Table for Model 6

Conclusions

Although more predictor values do not necessarily mean a better fitting model, both of the multiple linear regression models (Models 3 and 6) fit better than any of the simple linear regression models. The results from the log transformation of SalePrice in Model 6 suggests that this might be the best fitting model, even if its adjusted R-squared value is approximately 0.9 percent lower than the original model. Building these models has shown that while there is a strong correlation between the response and predictor variables, none of models have a very strong GOF yet. Additional data manipulation is necessary to try to create better fitting models.

Code

```
* Andrea Bruckner
  PREDICT 410, Sec 55
  Winter 2016
  Assignment 2
;

*************************************************************************;
* Preliminary Steps and Data Survey;
*************************************************************************;

* Access library where Ames housing dataset is stored;
libname mydata '/scs/crb519/PREDICT_410/SAS_Data/' access=readonly;
proc datasets library=mydata; run; quit;

* Set Ames housing dataset to short name;
data ames;
set mydata.AMES_HOUSING_DATA;
run;

* Explore Ames housing dataset contents;
proc contents data=ames; run;

* Print first 10 observations as preview of data;
Title "Preview of Dataset";
options obs=10;
proc print data=ames; run;
options obs=max;

* Print scatterplot of GrLivArea v. SalePrice to find obvious outliers or unusual data;
ods graphics on;
proc sgplot data=ames;
        scatter X=GrLivArea Y=SalePrice;
        title "GrLivArea v. SalePrice Scatter Plot – No Smoothers";
run;
ods graphics off;

* Making the previous scatterplot more readable;
proc sgplot data=ames;
reg x=GrLivArea y=SalePrice / clm;
loess x=GrLivArea y=SalePrice / nomarkers;
title "GrLivArea v. SalePrice Scatter Plot – Smoothers";
run; quit;

*************************************************************************;
* Define the Sample Population;
*************************************************************************;

* Create drop conditions;
```

```
data temp;
	set ames;
	format drop_condition $40.;
	if (BldgType ne '1Fam') then drop_condition='01: Not a Single Family Home';
	else if (Zoning = 'A' or zoning = 'C' or zoning = 'FV' or zoning = 'I' or zoning = 'RP')
		then drop_condition='02: Non-Residential Zone or RV Park';
	else if (SaleCondition ne 'Normal') then drop_condition='03: Not Normal Sale Condition';
	else if (GrLivArea >= 4000) then drop_condition='04: Atypically Large House';
	else if (TotalBsmtSF < 1) then drop_condition='05: No Basement';
	else if (GarageArea < 1) then drop_condition='06: No Garage';
	else if (SalePrice <= 0) then drop_condition='07: Pricing Error';
	else if (Utilities ne 'AllPub') then drop_condition='08: Not All Public Utilities Available';
else drop_condition='09: Sample Population';
run;

* View frequency of drop conditions;
proc freq data=temp;
tables drop_condition;
title 'Sample Population Waterfall';
run; quit; * (Reveals that all Sale Prices are positive);

* Subset data for just Sample Population and define logSalePrice variable;
data sample;
set temp;
if (drop_condition='09: Sample Population');
logSalePrice = log(SalePrice);
run;

* View scatterplot of sample data;
ods graphics on;
proc sgplot data=sample;
	scatter X=GrLivArea Y=SalePrice;
	title "Sample Population GrLivArea v. SalePrice Scatter Plot – No Smoothers";
run;
ods graphics off;

*****************************************************************;
* Data Quality Checks;
*****************************************************************;

* Create summary statistics for continuous data;
proc univariate data=sample;
var GrLivArea GarageArea TotalBsmtSF;
title;
run; quit;

proc means data=sample min q1 mean median q3 max std;
var GrLivArea GarageArea TotalBsmtSF;
run; quit;

proc means data=sample min q1 mean median q3 max std;
```

```
class SubClass;
var GrLivArea GarageArea TotalBsmtSF;
run; quit;

*************************************************************************;
* Initial EDA;
*************************************************************************;

proc means data=sample min mean median max;
class Neighborhood;
var SalePrice;
var GrLivArea;
var GarageArea;
var TotalBsmtSF;
run; quit;

proc sgplot data=sample;
histogram SalePrice / transparency=0.5;
density SalePrice / type=normal;
density SalePrice / type=kernel;
run; quit;

proc sgplot data=sample;
histogram GrLivArea / transparency=0.5;
density GrLivArea / type=normal;
density GrLivArea / type=kernel;
run; quit;

proc sgplot data=sample;
histogram GarageArea / transparency=0.5;
density GarageArea / type=normal;
density GarageArea / type=kernel;
run; quit;

proc sgplot data=sample;
histogram TotalBsmtSF / transparency=0.5;
density TotalBsmtSF / type=normal;
density TotalBsmtSF / type=kernel;
run; quit;

ods graphics on;
proc corr data=sample plots=matrix;
var SalePrice GrLivArea TotalBsmtSF;
run;
ods graphics off;

*************************************************************************;
* EDA for Modeling;
*************************************************************************;

proc sgplot data=sample;
```

```
reg x=GrLivArea y=SalePrice / clm;
loess x=GrLivArea y=SalePrice / nomarkers;
run; quit;

proc sgplot data=sample;
reg x=TotalBsmtSF y=SalePrice / clm;
loess x=TotalBsmtSF y=SalePrice / nomarkers;
run; quit;

proc sgplot data=sample;
histogram logSalePrice / transparency=0.5;
density logSalePrice / type=normal;
density logSalePrice / type=kernel;
run; quit;

proc sgplot data=sample;
reg x=GrLivArea y=logSalePrice / clm;
loess x=GrLivArea y=logSalePrice / nomarkers;
run; quit;

proc sgplot data=sample;
reg x=TotalBsmtSF y=logSalePrice / clm;
loess x=TotalBsmtSF y=logSalePrice / nomarkers;
run; quit;

* Create simple linear regression models;

proc reg data=sample alpha=0.05;
model SalePrice=GrLivArea / clb;
run;

proc reg data=sample alpha=0.05;
model SalePrice=TotalBsmtSF / clb;
run;

proc reg data=sample;
model SalePrice=TotalBsmtSF;
run;

* Create multiple linear regression model;

proc reg data=sample alpha=0.05;
model SalePrice=GrLivArea TotalBsmtSF / clb;
run;

* Create log simple linear regression models;

proc reg data=sample alpha=0.05;
model logSalePrice=GrLivArea / clb;
run;
```

```
proc reg data=sample alpha=0.05;
model logSalePrice=TotalBsmtSF / clb;
run;

* Create log multiple linear regression model;

proc reg data=sample alpha=0.05;
model logSalePrice=GrLivArea TotalBsmtSF / clb;
run;

* Create all regular (not log) models unpacked;

ods graphics on;
proc reg data=sample alpha=0.05 plots(unpack);
model SalePrice=GrLivArea / clb;
model SalePrice=TotalBsmtSF / clb;
model SalePrice=GrLivArea TotalBsmtSF / clb;
run;
ods graphics off;

* Create all log models unpacked;

ods graphics on;
proc reg data=sample alpha=0.05 plots(unpack);
model logSalePrice=GrLivArea / clb;
model logSalePrice=TotalBsmtSF / clb;
model logSalePrice=GrLivArea TotalBsmtSF / clb;
run;
ods graphics off;

*************************************************************************;
* END;
*************************************************************************;
```