Assignment #3

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 Assignments #1 and #2, we selected two continuous variables, GrLivArea and TotalBsmtSF, that indicate a promising ability to predict home sale prices, SalePrice. We are building the models in SAS and analyzing the summary statistics and diagnostic plots to determine the goodness-of-fit (GOF) of each model. In order to assess the GOF, we are looking to confirm the model assumptions: the relationship between the response variable and predictor variable(s) is linear, the errors are normally distributed and uncorrelated, and the errors have a mean of zero and show constant variance. We created a variety of models and discovered that removing outliers or refitting the multiple linear regression model with a logarithmic transformation of the response variable improved the overall fit of the model. Removing outliers showed the greatest improvement in GOF and suggests that scrutinizing the data for outliers will help strengthen the model's ability to accurately 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[1]. 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. In the next section we explain how we created a sample population from this data set in order to build predictive models for home sale prices.

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. For instance, we did not want to include data about apartment sales or commercial properties in our sample. **Table 1** below illustrates the criteria that caused observations to drop from our sample population. The seventh drop condition, 07: Pricing Error, does not appear in the table because all of the SalePrice data met the condition that the house sale price must be greater than $0. We decided to include this drop condition to eliminate any obvious errors. As shown in the table below, the sample population totals 1831 observations and represents 62.5 percent of the original data set.

[1] Source: http://www.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt

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 Assignments #1 and #2 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 variables 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. A plot matrix of these variables in **Figure 1** provides a visual means of understanding the Pearson correlation coefficients presented in **Table 2**.

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 Coefficients for Response and Predictor Variables

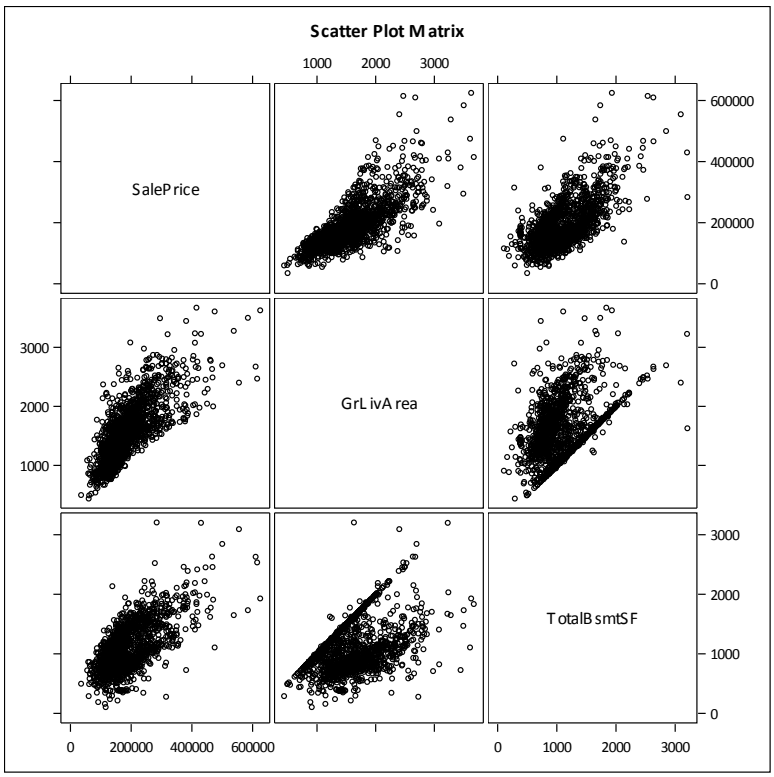


Figure 1: Scatter Plot Matrix of Response and Predictor Variables

The bivariate plots in **Figure 1** reveal that the variables are all positively correlated and suggest a mostly linear relationship. The two response variables do not have as strong of a correlation when they are paired together as when they are each paired with the response variable. The plots showing the relationship of the predictor variables also indicate that there is some collinearity between observations for GrLivArea and TotalBsmtSF since several of the data points fall on a straight line. These observations are something we may want to investigate later when building a multiple linear regression model.

The histograms in **Figures 2 and 3** reveal that these variables are skewed right and do not quite follow a normal distribution. However, these predictor variables are more normally distributed than the other continuous variables we looked at during our EDA, which is one of the reasons we chose these variables as our predictor variables. Examining the graphical output and correlation coefficients during our EDA led us to select GrLivArea and TotalBsmtSF as the most promising predictors of SalePrice. Together these predictor variables also 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.

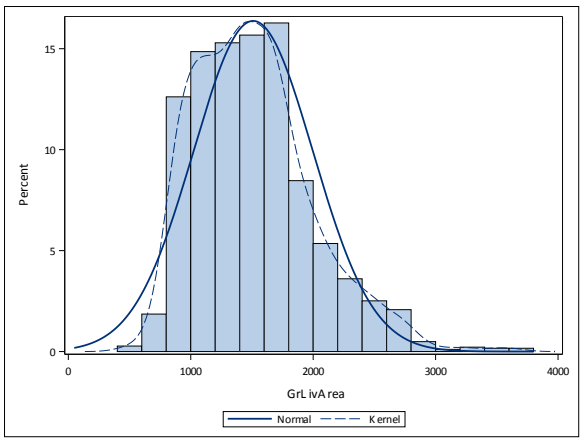


Figure 2: Histogram of GrLivArea

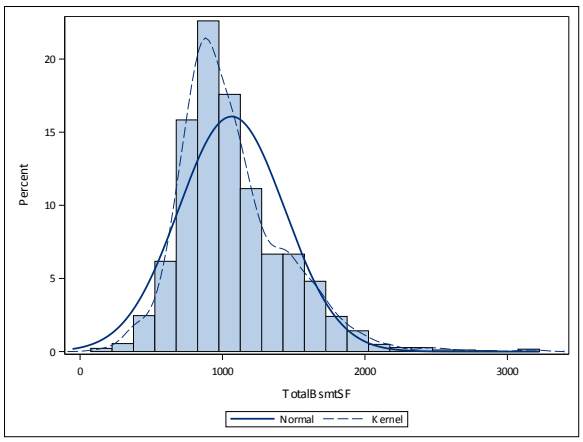


Figure 3: Histogram of TotalBsmtSF

Model 1: Simple Linear Regression of GrLivArea v. SalePrice

In order to evaluate the GOF of a model, we must examine the results to see if the residuals support the ordinary least squares (OLS) assumptions that the errors are normally distributed and homoscedastic.

Based on the output in **Table 3**, we can conclude that GrLivArea has a significant effect on SalePrice since the p-value is very low (<.0001) and the F-statistic is very high (2655.82). The R-squared value (0.5922) and adjusted R-squared value (0.5920) shown in **Table 4** are nearly equivalent and indicate that approximately 59.2 percent of the variation in the response variable can be explained by the predictor variable in the model. Per **Table 5**, 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. The parameter estimates seem reasonable since they indicate an increase in SalePrice for an increase in GrLivArea. If the parameter estimates indicated a decrease in SalePrice for an increase in GrLivArea, we would have to investigate further since this is not a logical relationship.

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

Root MSE	45652	**R-Square**	0.5922
Dependent Mean	182338	**Adj R-Sq**	0.5920
Coeff Var	25.03710		

Table 4: Overall Model Fit 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 5: Parameter Estimates Table for Model 1

Although the results in the above tables indicate a strong linear correlation between GrLivArea and SalePrice, further analysis of the diagnostic plots reveal that Model 1 has a poor GOF and has plenty of room for improvement.

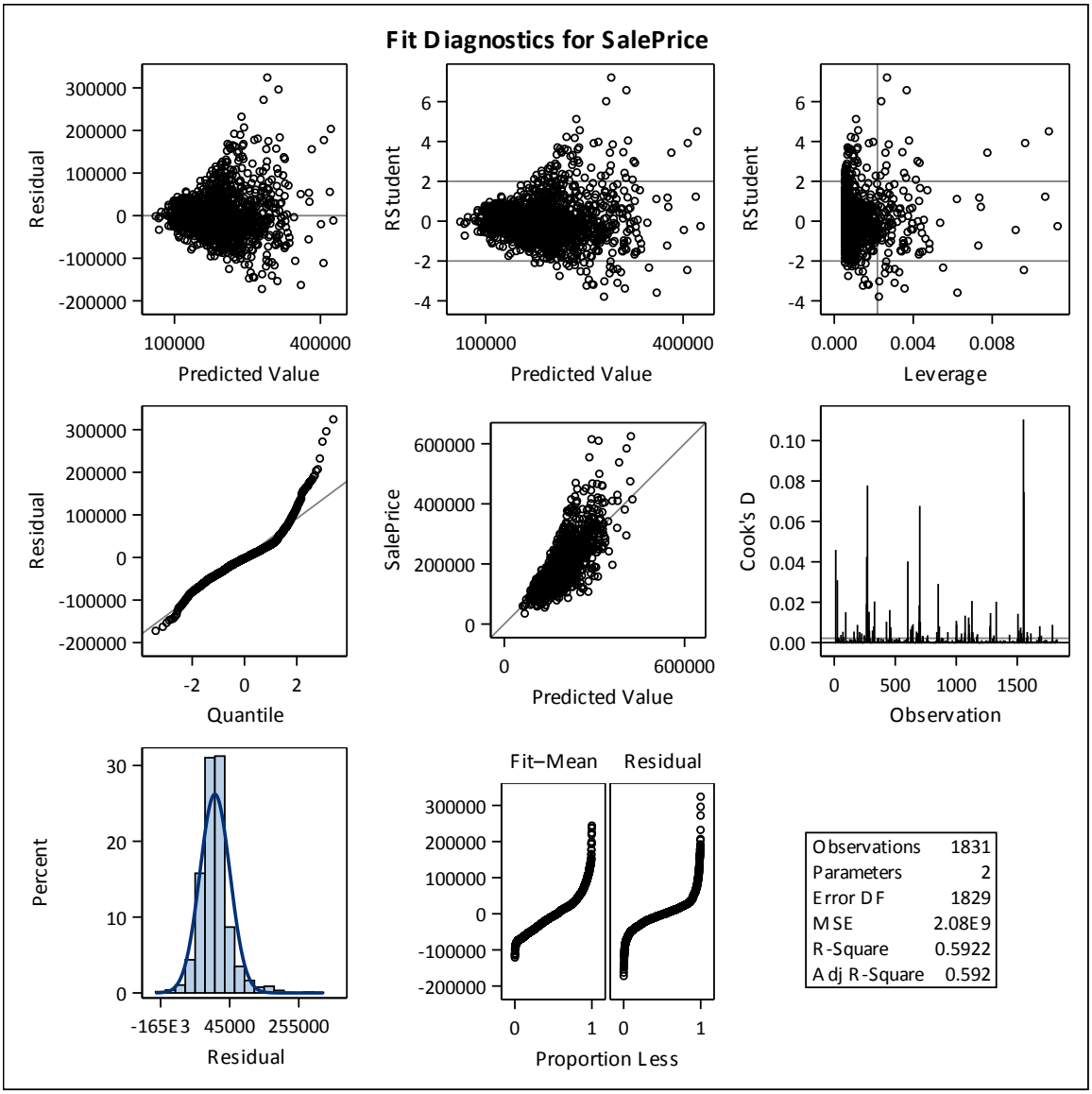


Figure 4: Fit Diagnostic Plots for Model 1

Figure 4 contains a lot of indicators about the GOF for 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).

Examining these observations closer and removing some of them may help improve the GOF. The curvature of the Q-Q plot and the heavy tail indicate a right-skewed distribution. The residuals histogram confirms the conclusion that the model violates the normality assumption.

The plot of the response variable v. predicted value in the center of **Figure 4** 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. This plot suggests that the model output and the observations of the response variable do not have as linear of a relationship as assumed. In the residual-fit plot directly 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 suggest 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 4** indicate that the OLS assumption of constant variance is not true since the plots reveal a funnel shape. It shows that the residuals do not have a mean of zero and show an increasing trend rather than display independence and identical distribution. This pattern is also apparent in the residuals v. predictor plot in **Figure 5**. As the predictor variable increases, the error variance also increases.

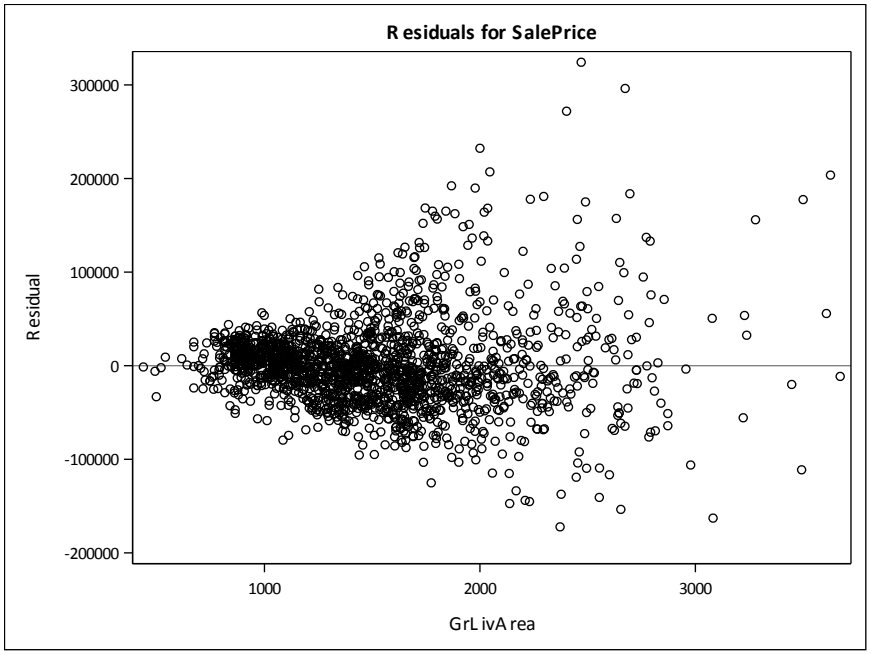


Figure 5: GrLivArea v. Residuals for SalePrice for Model 1

Figure 6 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. **Figure 6** shows that, despite the high R-squared value, Model 1 does not validate the OLS assumptions of normality and homoscedasticity and is therefore a poor fit.

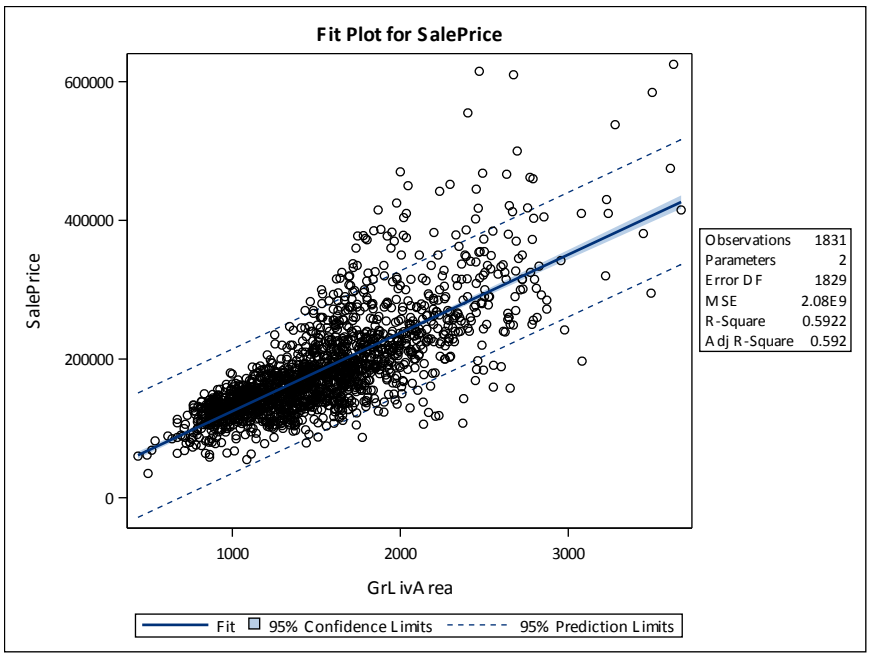


Figure 6: Fit Plot for Model 1

Model 2: Simple Linear Regression of TotalBsmtSF v. SalePrice

The regression output for Model 2 has some similarities to Model 1, but it has a worse fit. **Table 6** 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 on SalePrice. The Mean Square Error (MSE) in **Table 6** is 2.95E9, which is higher than the MSE for Model 1, which was 2.08E9. The higher MSE for Model 2 reveals that it does not fit as well as Model 1. In **Table 7**, we can see that the R-squared and adjusted R-squared values are approximately 0.42, which indicates that this model accounts for about 42 percent of the variation in the response variable. Since this value is 17 percent less than Model 1, this statistic also suggests that Model 1 is a better fit than Model 2. In **Table 8** we can determine that the fitted equation for this model is **SalePrice = 49547 + 124.75*TotalBsmtSF.** As the basement space increases by 1 square foot, the home sale price increases by $124.75. These parameter estimates are reasonable since a home with a larger basement will likely sell for a higher price.

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 6: ANOVA Table for Model 2

Root MSE	54325	**R-Square**	0.4225
Dependent Mean	182338	**Adj R-Sq**	0.4222
Coeff Var	29.79357		

Table 7: Overall Model Fit 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 8: Parameter Estimates Table for Model 2

The summary statistic results shown in the tables above indicate that TotalBsmtSF and SalePrice are strongly related; however, the diagnostic plots reveal that Model 2 violates the OLS assumptions of normality and equal variance.

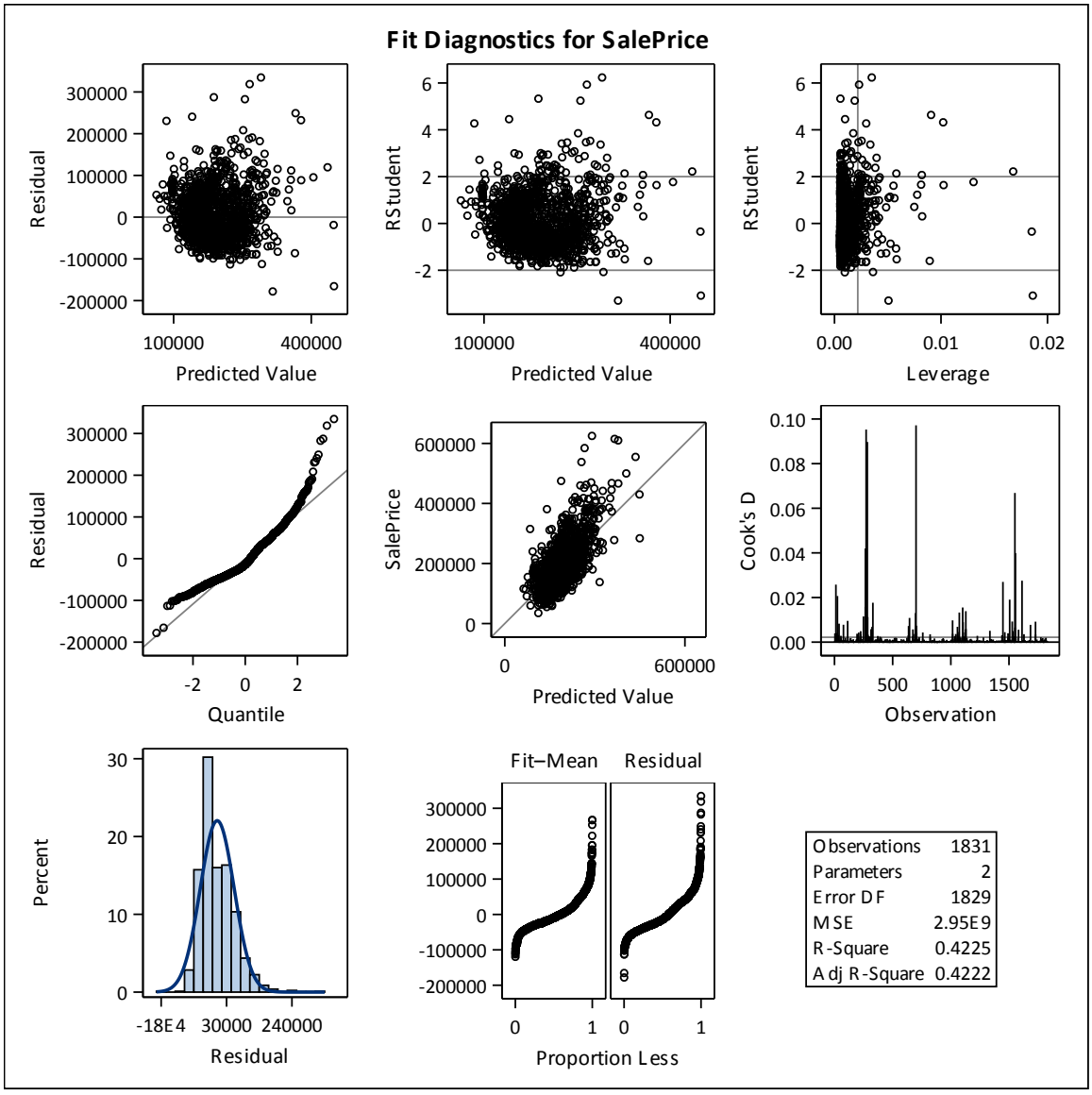


Figure 7: Fit Diagnostic Plots for Model 2

The residual-leverage plot in **Figure 7** suggests the same conclusions as the equivalent plot in **Figure 4**: There are observations that are indicative of being outliers, high-leverage points, or both. We should further analyze these observations later to see if we can improve Model 2's fit

by removing some of them. The Q-Q plot and histogram both show a strongly right-skewed distribution with several potential outliers, which prove that the normality assumption is not valid. The plot of the response variable v. predicted value in the center of **Figure 7** indicates that the model's variables are not perfectly linear and require modification to avoid the slight upward curve in the plot. Similar to Model 1's plot in **Figure 4**, the Cook's D plot in **Figure 7** shows several high peaks, indicating further investigation is necessary to determine if there are high influence points that have caused the model to show a non-normal distribution and non-constant variance. The residual-fit plot in **Figure 7** also is similar to the same plot in **Figure 4** since the residual side is taller than the fit side, indicating there is unexplained variance in the model.

In **Figure 7**, the residuals and studentized residuals v. predicted values plots indicate that the assumption of constant variance is not true since the residuals are a slight double bow shape. The residuals do not display a mean of zero, and some residuals also appear to be potential outliers. The residuals v. predictor plot in **Figure 8** shows a violation of the assumption of homoscedasticity since there is 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 to eliminate possible outliers and influential points.

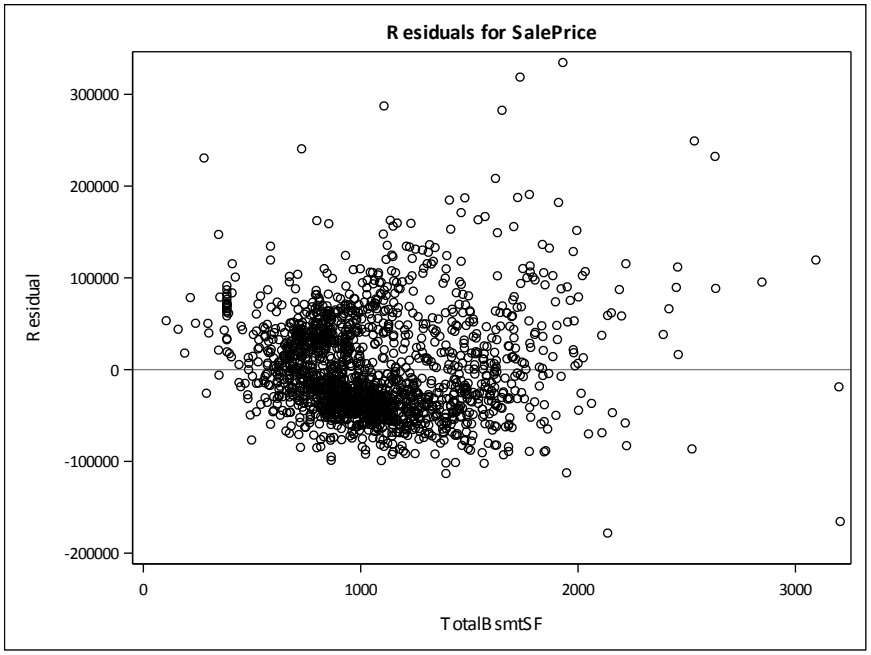


Figure 8: Model 2 TotalBsmtSF v. Residuals for SalePrice

While more of the data points fall within the Prediction Limits in **Figure 9** than in **Figure 6**, the clustering of the data points and the low R-square and adjusted R-square values of Model 2 indicate that the variables do not have as perfect of a linear relationship as assumed. Overall the plots reveal that there is still some unexplained variation in the model that contribute to Model 2's poor fit.

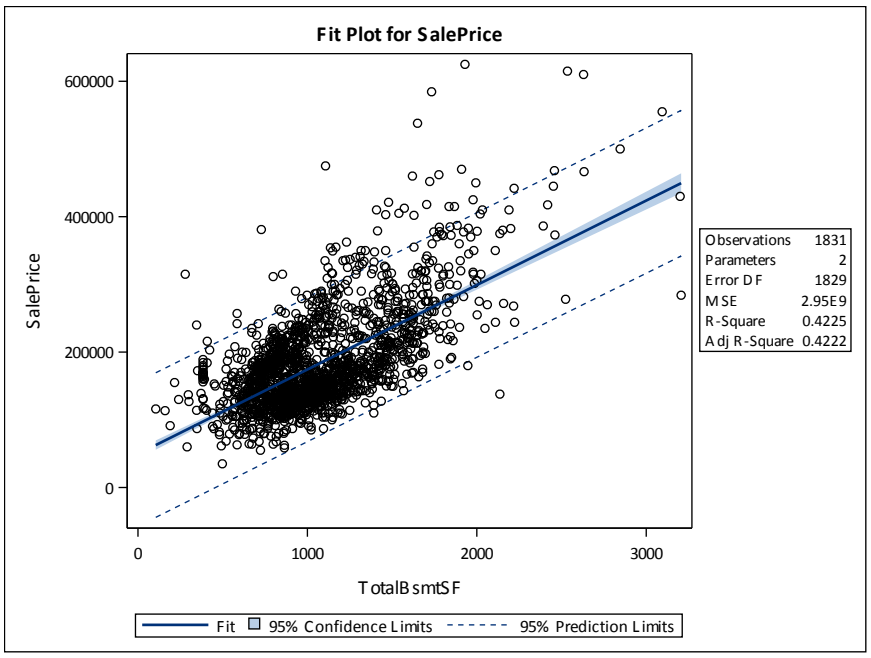


Figure 9: Fit Plot for Model 2

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

As in the simple linear regression models 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 low p-value and high F-statistic in **Table 9**. The MSE in **Table 9** is 1.35E9, which is lower than the MSE for both Models 1 and 2, suggesting that Model 3 fits better than the other models. The R-squared and adjusted R-squared values in **Table 10** both show that the model explains approximately 73.5 percent of the variation in the response variable. This is much higher than the result for either Model 1 or Model 2, which indicates that Model 3 might be a better fit. Per **Table 11**, the fitted equation for this model is **SalePrice = -36306 + 89.19*GrLivArea + 78.90*TotalBsmtSF.** As the above ground living space increases by 1 square foot while the basement space is held steady, the home sale price increases by $89.19. When the basement space increases by 1 square foot and the above ground living space is held constant, the home sale price increases by $78.90. These parameter estimates are reasonable since a home with more space will likely sell for a higher price.

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 9: ANOVA Table for Model 3

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

Table 10: Overall Model Fit 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 11: Parameter Estimates Table for Model 3

An examination of the diagnostic plots in **Figure 10**, 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 variance since both plots show a funnel shape. The plots also reveal that the residuals do not have a mean of zero and the residuals increase with the predicted values. The residual-leverage plot in **Figure 10** suggests the presence of outliers and high-leverage points, and the Q-Q plot and residual histogram indicate an approximately normal distribution with heavy tails, possibly due to outliers. The plot of the response variable v. predicted value indicates that the model, with the exception of a few possible outliers, fits since the model's predictions and the observations show a mostly linear relationship. 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 and possible removal. Removing several outliers might improve the fit of this model. 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.

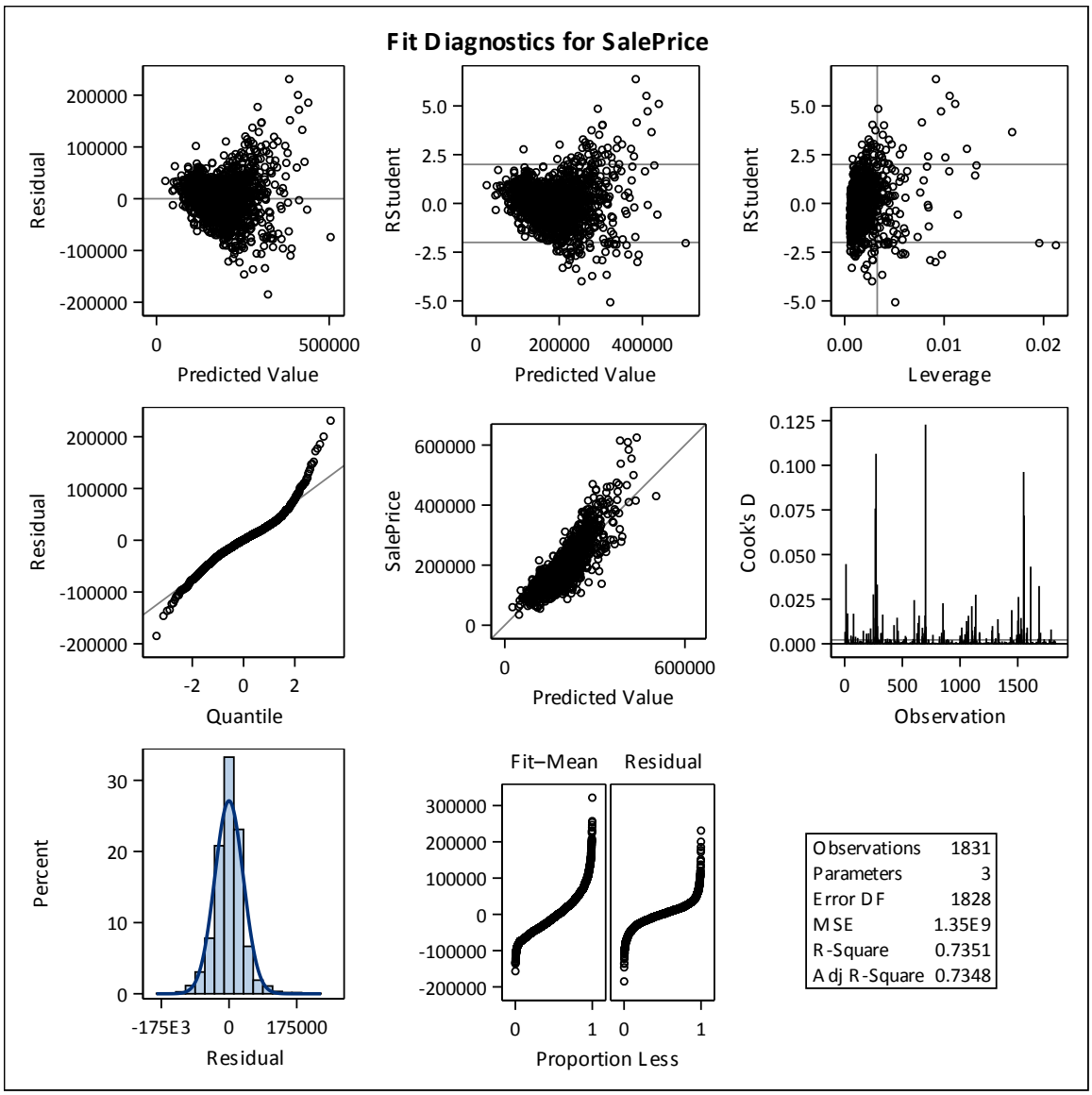


Figure 10: Fit Diagnostic Plots for Model 3

Multiple linear regression models do not produce a fit plot similar to **Figures 6 and 9**, but the results for Model 3 do include the residuals v. predictors plots. The plots in **Figure 11** reveal that the residuals violate the assumption of homoscedasticity since the data points form a funnel shape rather than a consistent band around mean zero.

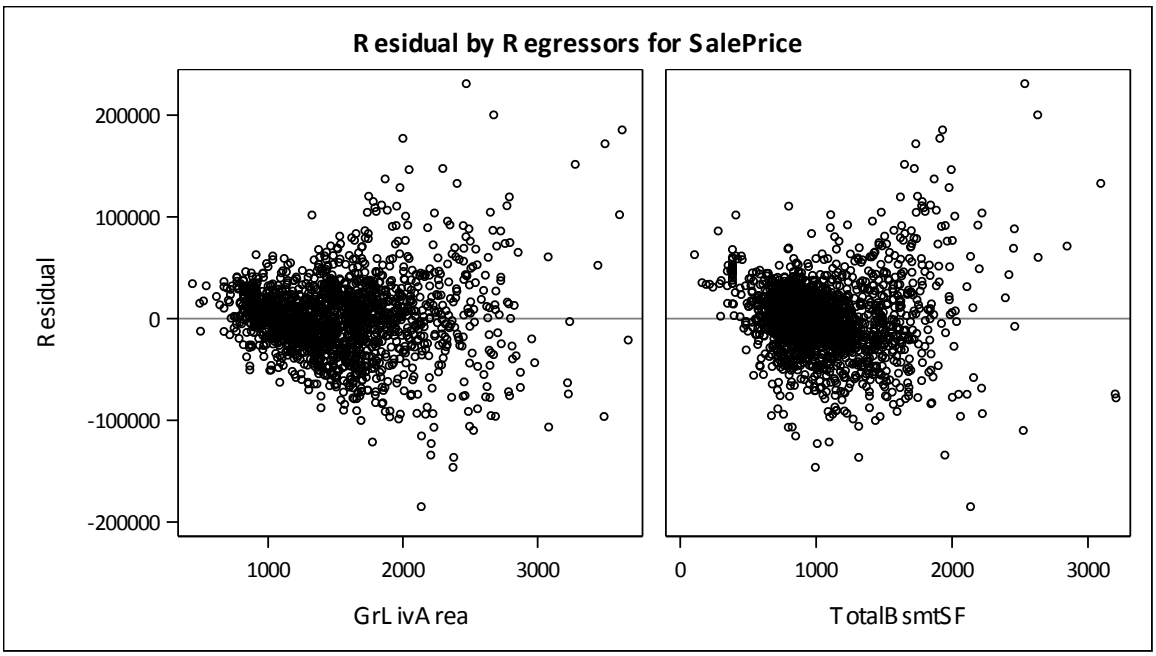


Figure 11: Residuals for Model 3

Although Model 3 has a more normal distribution that Models 1 and 2, it still violates the OLS assumptions of normality and homoscedasticity. We can, however, possibly improve the model fit by removing outliers or performing a transformation on one of the variables.

Outliers Identification

The models discussed above all indicate the presence of outliers in the data. It is important to note that what we are defining in this section as outliers are valid data points and are not necessarily true outliers in the context of the data set. We already accounted for any unusual data and outliers when defining our sample population, but here we are identifying certain values that are hampering the predictive potential of these models and causing the models to fit poorly. In order to try to produce a better fitting model, we decided to remove observations that appear to be outliers or observations that exert strong influence or leverage in the model.

After looking at the diagnostic plots for Model 3, we decided to remove observations that fell far from the median of each predictor variable since these are likely to have considerable leverage in the model. To do this, we found the interquartile range (IQR) and then calculated the IQR's upper and lower limits. We considered any observations that fell outside the limits an outlier.

Analysis Variable : GrLivArea		
Lower Quartile	Upper Quartile	Quartile Range
1130.00	1772.00	642.0000000

Table 12: IQR for GrLivArea

Analysis Variable : TotalBsmtSF		
Lower Quartile	Upper Quartile	Quartile Range
822.0000000	1248.00	426.0000000

Table 13: IQR for TotalBsmtSF

Based on the IQR in **Table 12**, we calculated the lower limit for GrLivArea as 167 and the upper limit as 2735. **Table 13** shows an IQR of 426 for TotalBsmtSF, which produced 183 as the lower limit and 1887 as the upper limit.

In addition to removing the observations that fell outside the IQR limits for each predictor variable, we removed values that were potentially high influence points based on the Cook's D output for Model 3. If $D_i > 4/n$, we dropped the observation from our sample. **Table 14** below shows how many observations were removed with each outlier definition.

Outliers Definition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1. GrLivArea Outliers	30	1.64	30	1.64
2. TotalBsmtSF Outliers	50	2.73	80	4.37
3. Cook D Outliers	76	4.15	156	8.52
4. Not An Outlier	1675	91.48	1831	100.00

Table 14: Outliers Definition for the Sample Population

From the 1831 observations in the sample population, 1675 observations remained after we removed the outliers.

Model 4: Multiple Linear Regression without the Outlier Observations

Removing the outliers identified in the section above greatly improved the fit of the multiple linear regression model. The low p-value and high F-statistic in **Table 15** suggest that the predictor variables have a significant effect on the response variable, just as the ANOVA table for Model 3 suggested. **Table 15** also shows the MSE as 7.27E8, which is considerably lower than the MSE for Model 3. The R-squared and adjusted R-squared values in **Table 16** show that the model explains approximately 73.7 percent of the variation in the response variable, which is just slightly more than the 73.5 percent for Model 3. Per **Table 17**, the fitted equation for this model is **SalePrice = -18153 + 86.87*GrLivArea + 63.01*TotalBsmtSF.** The parameter estimates for β_1 and β_2 both decreased slightly from Model 3, while the parameter estimate for β_0 changed considerably from -36306 in Model 3 to -18153 in Model 4.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	3.418983E12	1.709492E12	2350.11	<.0001
Error	1672	1.216228E12	727409293		
Corrected Total	1674	4.635212E12			

Table 15: ANOVA Table for Model 4

Root MSE	26971	**R-Square**	0.7376
Dependent Mean	170732	**Adj R-Sq**	0.7373
Coeff Var	15.79696		

Table 16: Overall Model Fit Table for Model 4

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-18153	2899.28035	-6.26	<.0001	-23839	-12466
GrLivArea	1	86.87080	1.64358	52.85	<.0001	83.64712	90.09448
TotalBsmtSF	1	63.00950	2.28467	27.58	<.0001	58.52839	67.49061

Table 17: Parameter Estimates Table for Model 4

Removing outliers from the sample population resulted in a more normal distribution of the residuals for Model 4. This is evident when we compare **Figures 12 and 13** below. The residuals in **Figure 13** create a nearly straight line and have very few points that deviate from the line, whereas **Figure 12** shows a slight curve and heavier tails. Removing the outliers also shortened the tails of the distribution, which is evident in both the Q-Q plots, as well as in the residuals histograms in **Figures 14 and 15**.

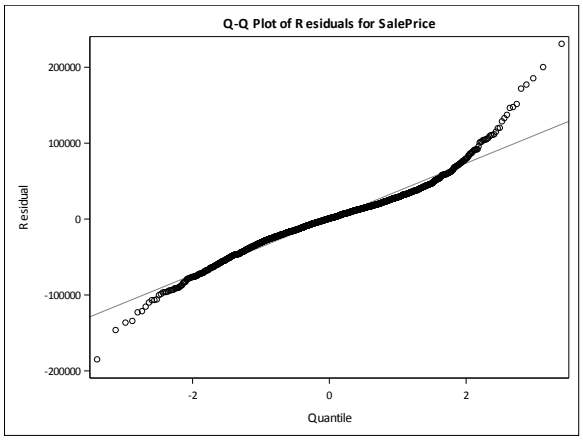


Figure 12: Q-Q Plot of Residuals for Model 3

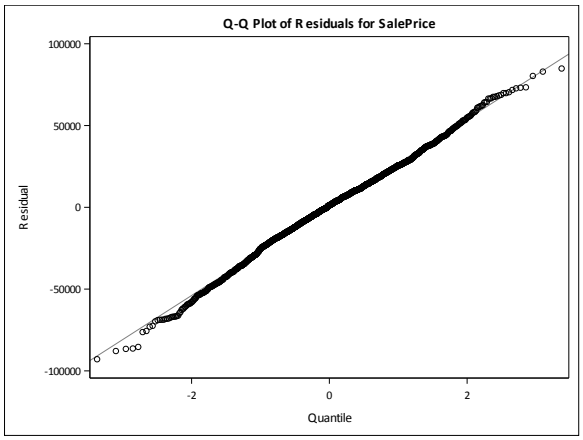


Figure 13: Q-Q Plot of Residuals for Model 4

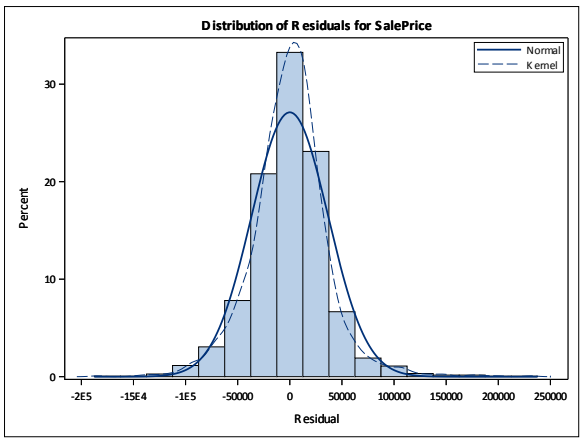


Figure 14: Residuals Histogram for Model 3

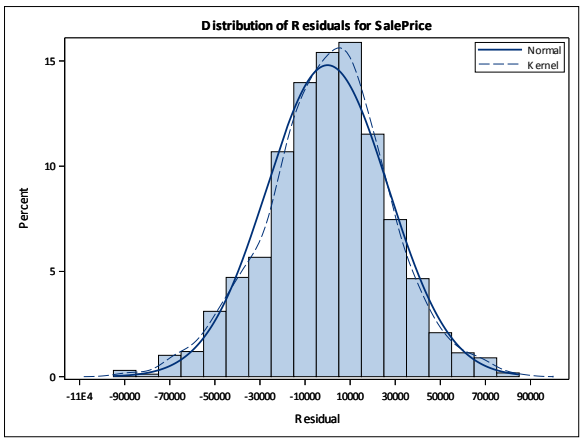


Figure 15: Residuals Histogram for Model 4

The observed v. predicted plots in **Figures 16 and 17** reveal that removing the outliers has greatly improved the fit of the model. Whereas the data points curve upward and begin to trail off in **Figure 16**, the data points in **Figure 17** do not exhibit this behavior and show instead a more linear relationship. **Figure 17** suggests that removing the outliers has created a model that can better predict the response variable.

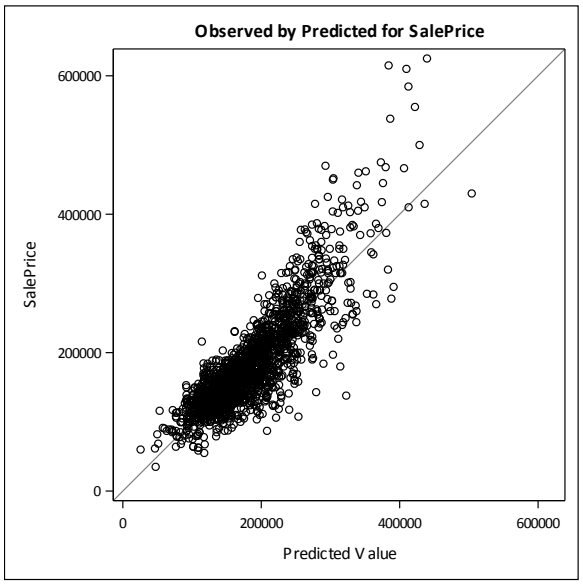


Figure 16: Observed by Predicted for Model 3

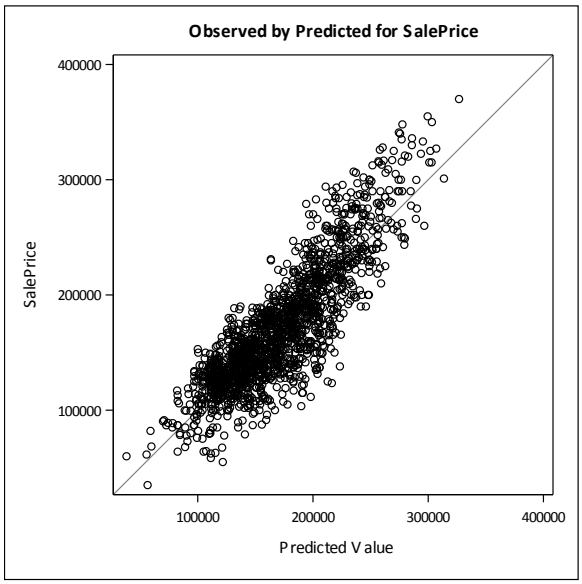


Figure 17: Observed by Predicted for Model 4

In the residual-fit plots below, the refitted model without outliers, as depicted in **Figure 19**, shows that the fit side is much taller than the residual side, indicating that the model mostly accounts for the variation in the data. Although **Figure 18** also shows a higher fit side than residual side, the fit side in **Figure 19** is much more pronounced and indicative that Model 4 better explains the variation than Model 3.

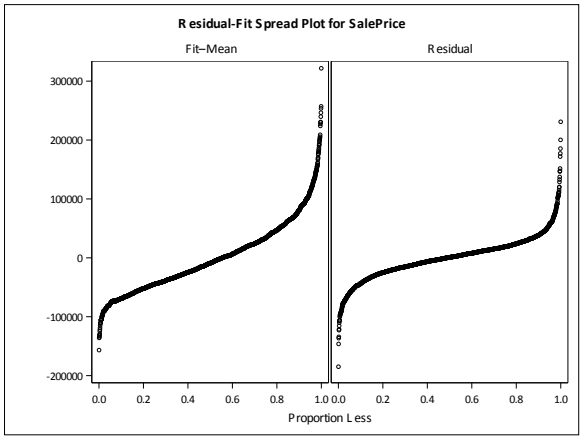


Figure 18: Residual-Fit Spread for Model 3

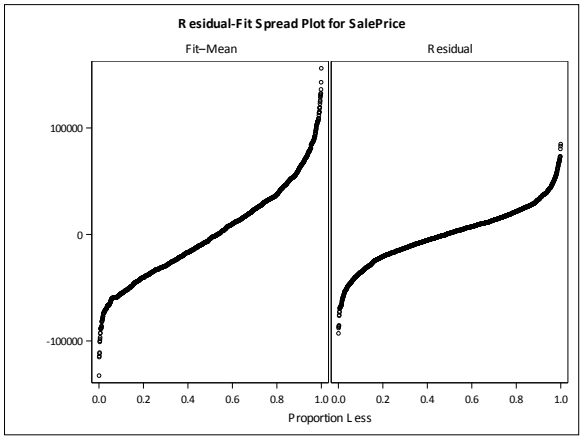


Figure 19: Residual-Fit Spread for Model 4

The studentized residuals v. predicted values plots, shown in **Figures 20 and 21**, reveal that removing the outliers has created more randomly distributed residuals. Whereas **Figure 20** has a clear funnel shape, **Figure 21** has no discernable pattern. **Figure 21** shows that Model 4 is more likely to validate the model assumptions that the residuals have equal variance, are not correlated, and have a mean of zero.

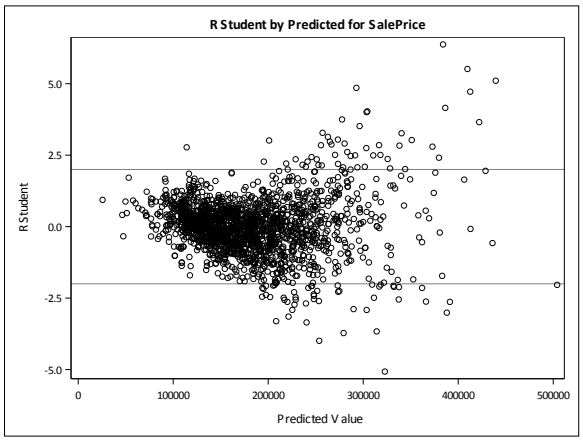


Figure 20: Studentized Residuals for Model 3

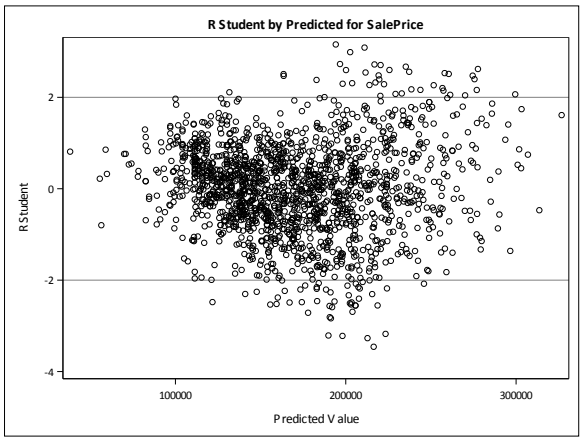


Figure 21: Studentized Residuals for Model 4

Although the studentized residual plots above suggest that removing the outliers has improved the fit of Model 4, the residual plots by the predictor variables in **Figures 22 and 23** suggest that Model 4 does not fully validate the OLS assumptions of homoscedasticity.

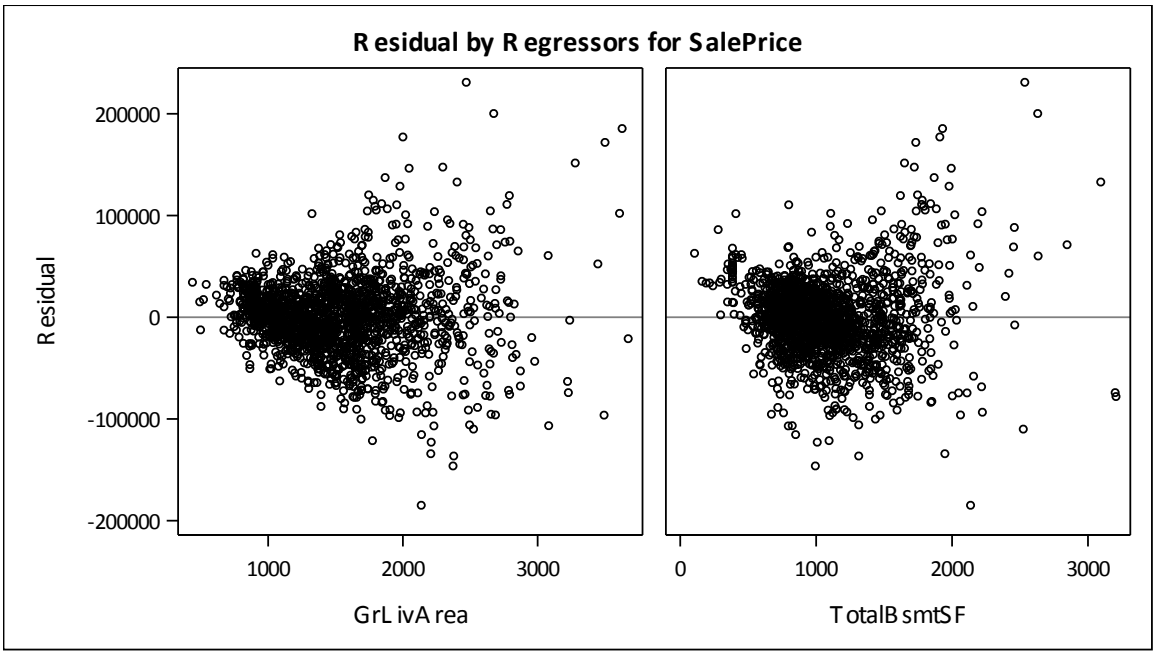


Figure 22: Residuals by Predictors for Model 3

The residuals by the predictor variables for Model 4 in **Figure 23** appear more random than the corresponding, funnel-shaped plots in **Figure 22** for Model 3, but the residuals for Model 4 still suggest non-constant variance. The Residuals v. TotalBsmtSF plot in **Figure 23** has no clear pattern, but the Residuals v. GrLivArea plot in **Figure 23** has a double bow shape. In order to improve the fit, a transformation of one of the predictor variables might be necessary.

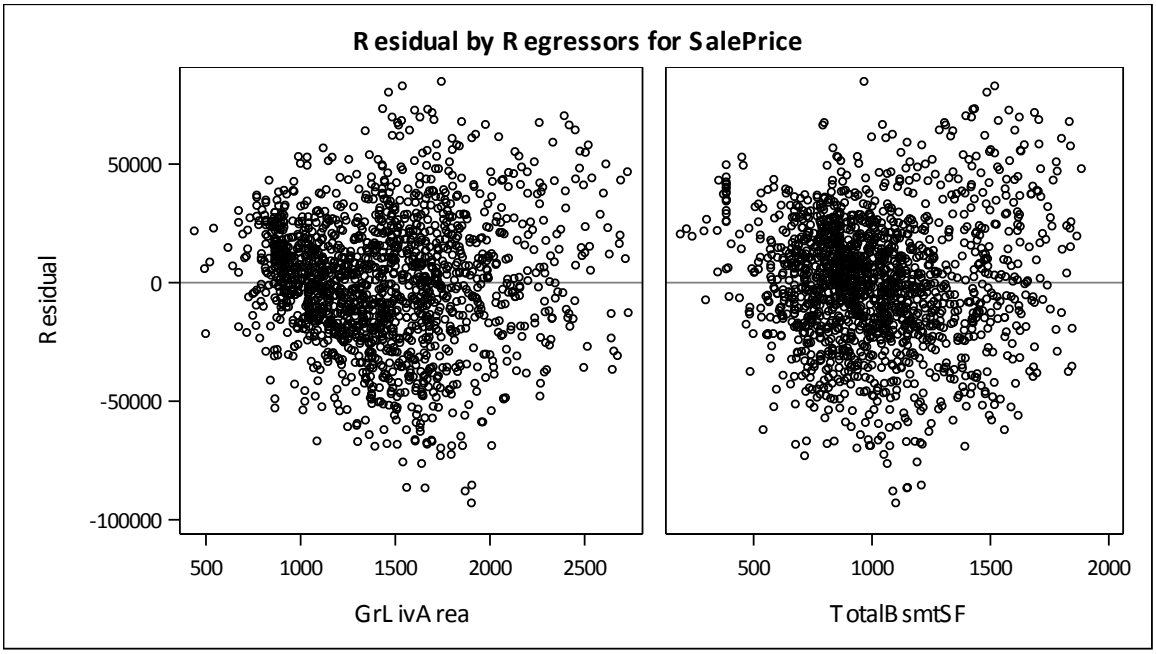


Figure 23: Residuals by Predictors for Model 4

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 a better fitting model. The log transformation of SalePrice, as shown in **Figure 25**, follows a much more normal distribution that the untransformed response variable shown in **Figure 24**.

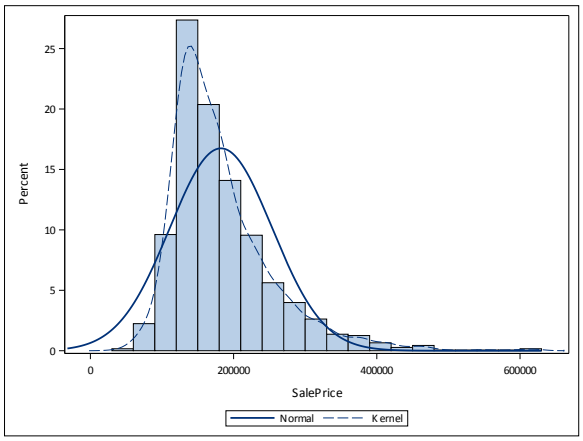


Figure 24: Histogram of SalePrice

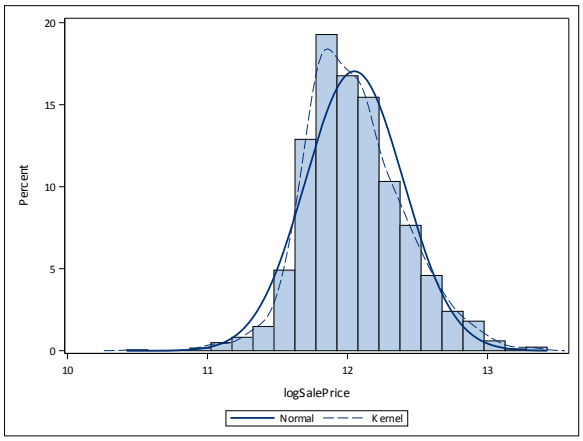


Figure 25: Histogram of log(SalePrice)

In the next section we will refit Model 3 so that it contains the same predictor variables but uses log(SalePrice) instead of 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 model with the original model in order to determine if the GOF has improved after the log transformation.

Model 5: Multiple Linear Regression of GrLivArea and TotalBsmtSF v. log(SalePrice)

Transforming the response variable did not significantly improve the fit of the model. The low p-value and high F-statistic in **Table 18** indicates that the predictor variables have a significant effect on the transformed response variable. **Table 18** also shows that the MSE is very small, which suggests the model is likely to make accurate predictions. The R-squared and adjusted R-squared values in **Table 19** show that the model explains approximately 72.6 percent of the variation in the response variable, which is just slightly less than the 73.5 percent explained in Model 3. The fitted equation for this model is **SalePrice = 10.98 + 0.0005*GrLivArea + 0.0004*TotalBsmtSF** per **Table 20.** This is significantly different from the fitted equation for Model 3, which was **SalePrice = -36306 + 89.19*GrLivArea + 78.90*TotalBsmtSF.** The drastic change in parameter estimates for β_0, β_1, and β_2 are not unusual, however, given that the model contains a log transformation.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	163.89965	81.94982	2422.49	<.0001
Error	1828	61.83895	0.03383		
Corrected Total	1830	225.73860			

Table 18: ANOVA Table for Model 5

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

Table 19: Overall Model Fit Table for Model 5

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	10.98472	0.01596	688.08	<.0001	10.95341	11.01603
GrLivArea	1	0.00045523	0.00000960	47.43	<.0001	0.00043641	0.00047406
TotalBsmtSF	1	0.00035418	0.00001256	28.21	<.0001	0.00032956	0.00037881

Table 20: Parameter Estimates Table for Model 5

Performing a log transformation of SalePrice in Model 5 changed the skew from Model 3, but it did not create a more normal distribution. The residuals distribution shifted from a right skew to a left skew, as indicated by the curvature of the Q-Q plots in **Figures 26 and 27**.

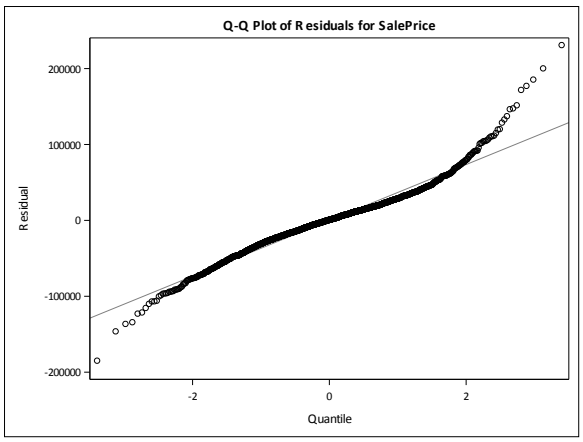


Figure 26: Residuals for Model 3

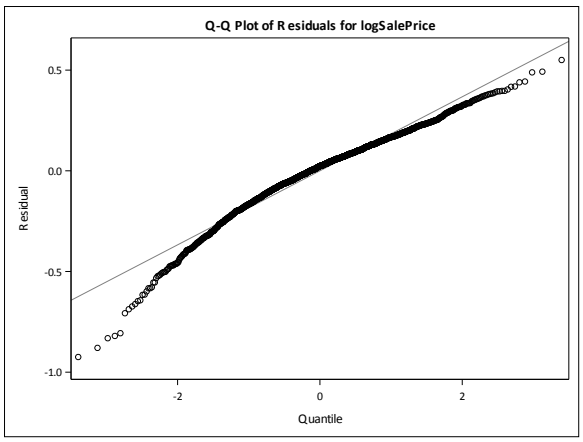


Figure 27: Residuals for Model 5

The observed v. predicted plots in **Figures 28 and 29** reveal only a slight improvement in model fit for the transformed model. There are fewer data points that stray from the line in **Figure 29** than in **Figure 28**, indicating that Model 5 better predicts the response variable than Model 3.

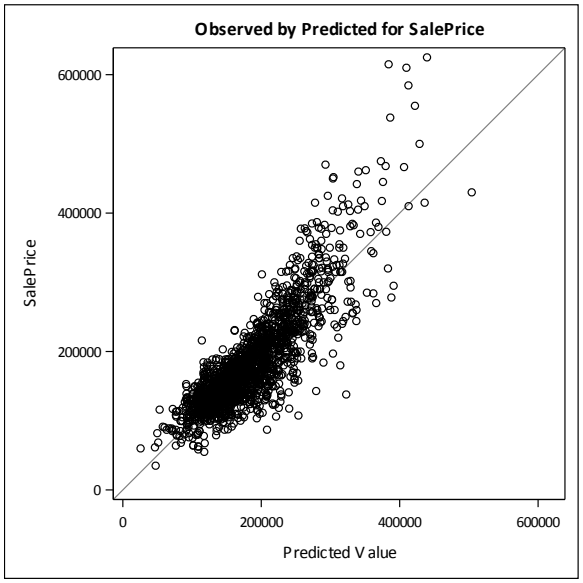


Figure 28: Observed by Predicted for Model 3

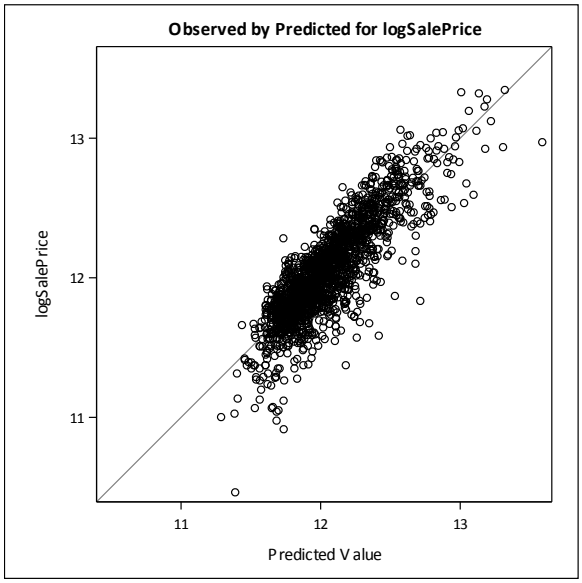


Figure 29: Observed by Predicted for Model 5

The studentized residuals v. predicted values plots, shown in **Figures 30 and 31**, reveal that performing a log transformation of the response variable has produced more randomly distributed residuals. **Figure 30** has a clear funnel shape, but **Figure 31** is more scattered and the residuals more likely have a mean of zero than in **Figure 30**. Both plots, however, indicate the presence of possible outliers.

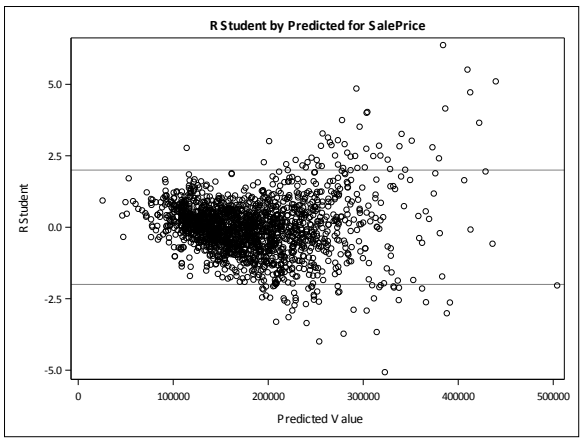


Figure 30: Studentized Residuals for Model 3

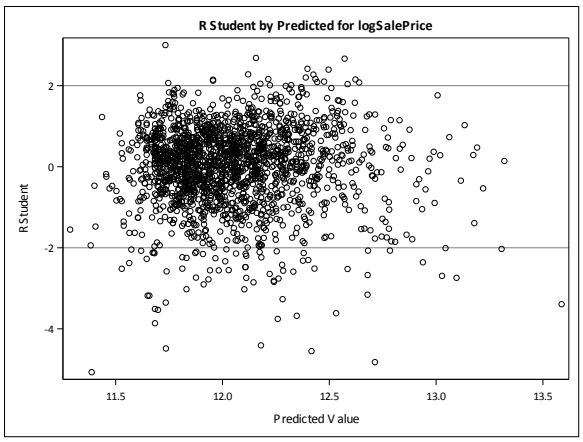


Figure 31: Studentized Residuals for Model 5

Although the studentized residual plots above suggest that the log transformation of the response variable has improved the fit of Model 5 over Model 3, the residual plots by the predictor variables in **Figures 32 and 33** reveal that Model 5 still violates the OLS assumptions of homoscedasticity.

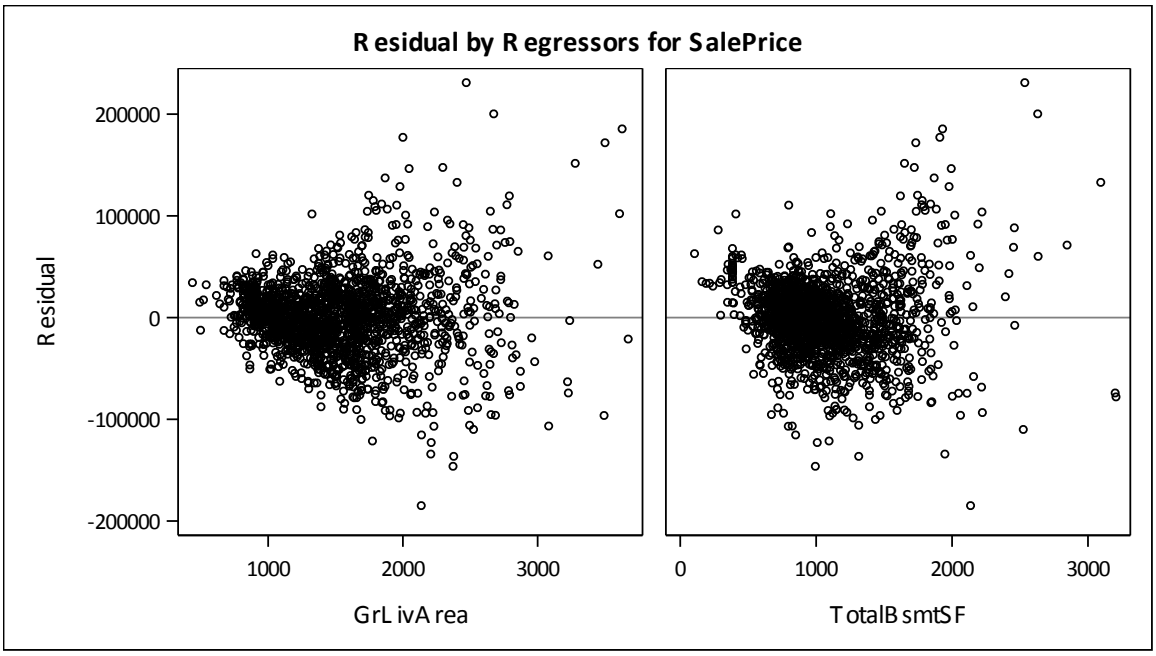


Figure 32: Residuals by Predictors for Model 3

The residuals by the predictor variables for Model 5 in **Figure 33** appear more random than the corresponding, funnel-shaped plots in **Figure 32** for Model 3, but the residuals for Model 5 still suggest non-constant variance. The Residuals v. GrLivArea plot in **Figure 33** has a slight double bow shape while the Residuals v. TotalBsmtSF has funnel shape, indicating that a transformation of one or both of these predictor variables might be necessary. Both models also show a stack of points near the 300 mark for TotalBsmtSF, which warrants further investigation as well. In general, the predictor variables require further analysis and modifications in order to make the model support the OLS assumptions.

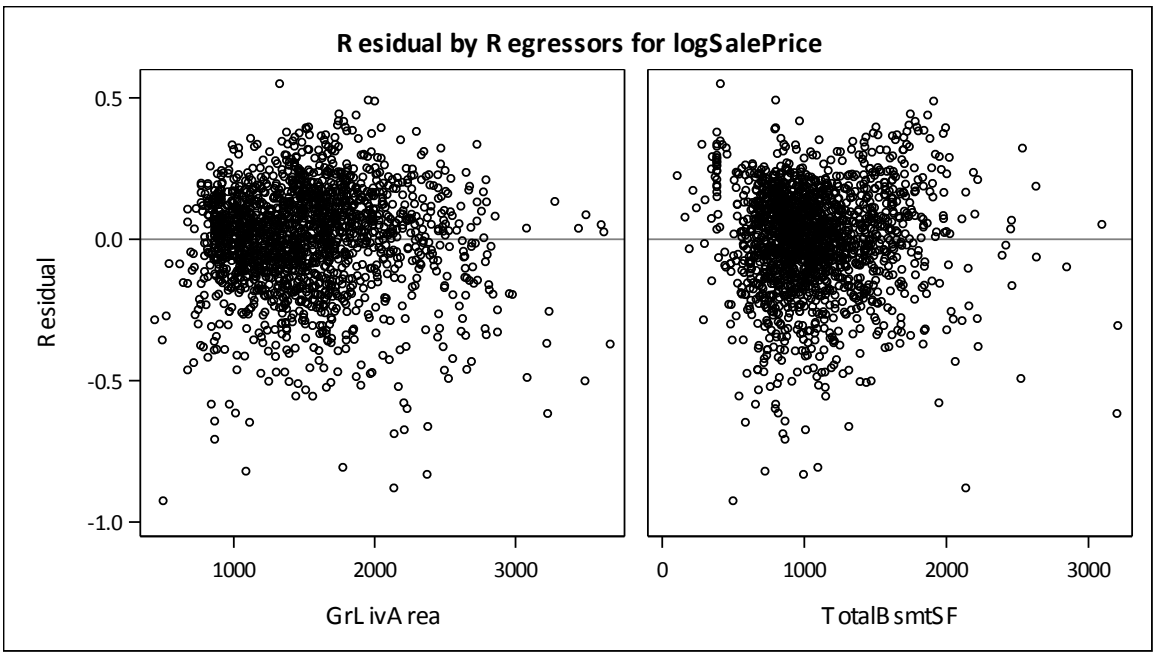


Figure 33: Residuals by Predictors for Model 5

In both of the residual-fit plots in **Figures 34 and 35**, the fit side is taller than the residual side, suggesting that the models mostly account for the variation in the data. The difference in height in **Figure 35** is greater than in **Figure 34**, which suggests that Model 5 fits better than Model 3.

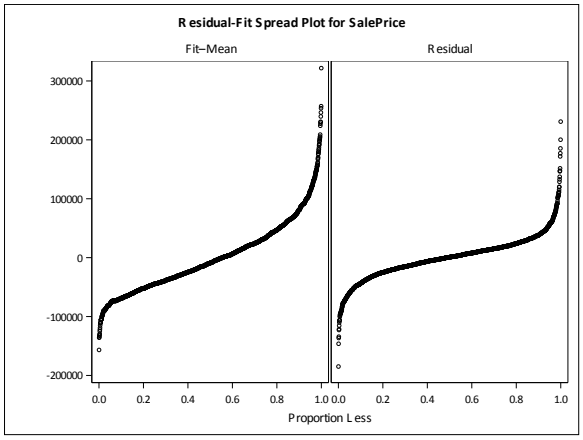


Figure 34: Residual-Fit Spread for Model 3

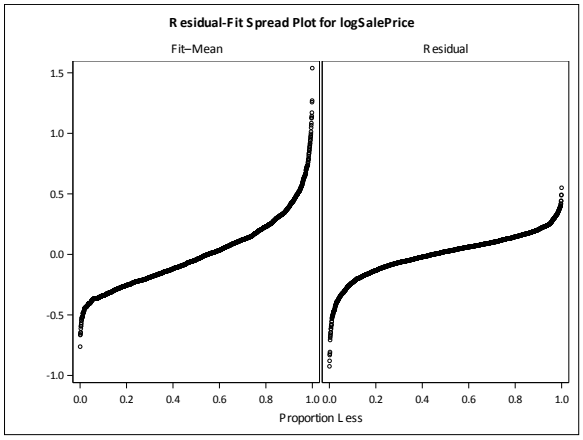


Figure 35: Residual-Fit Spread for Model 5

Based on the statistical tables and diagnostic plots, Model 5 fits better than Model 3, but it still violates the OLS assumptions of normality and homoscedasticity and is therefore not a reliable model for predicting the sale price of homes in Ames, Iowa.

Because the log transformation of the response variable did not fully improve the model, we decided that transforming the predictor variables might improve the fit of the model. We kept the log transformation of the response variable as part of the model but refitted the model by performing either a square root or inverse transformation on either one or both predictors. Refitting the model with these transformations did not greatly improve the fit and made it worse in some cases. The results of the transformations of the predictor variables are in Appendix A.

Conclusions

In this report we assessed the GOF of several models and determined how modifying the model, either through eliminating potential outliers or performing a logarithmic transformation on the response variable, might improve the model's fit. Although more predictor variables do not necessarily mean a better fitting model, the multiple linear regression models explored in this report fit better than either of the simple linear regression models we created. All of the models we examined violated at least one OLS assumption, which indicates that we still have more work to do in creating a more reliable model. Of the models we examined, Model 4, in which we removed potential outliers, offered the most promising predictive potential. We would like to conduct an even more thorough outliers assessment in the future to improve the model's fit. We also believe that the combination of creating a model with the outliers removed and performing a log transformation on the response variable may enable us to create a better model for predicting single-family home sale prices in Ames, Iowa.

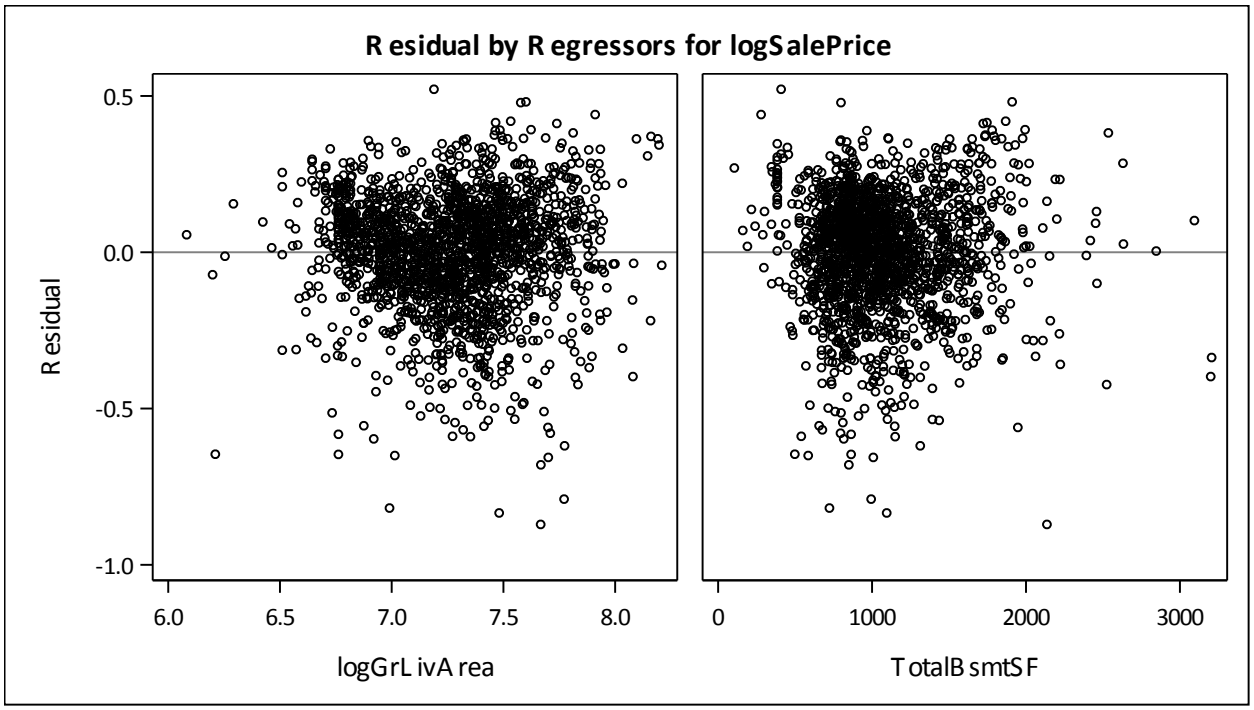
Appendix A

Results from log(SalePrice) v. log(GrLivArea) and TotalBsmtSF

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	164.32683	82.16342	2445.70	<.0001
Error	1828	61.41177	0.03360		
Corrected Total	1830	225.73860			

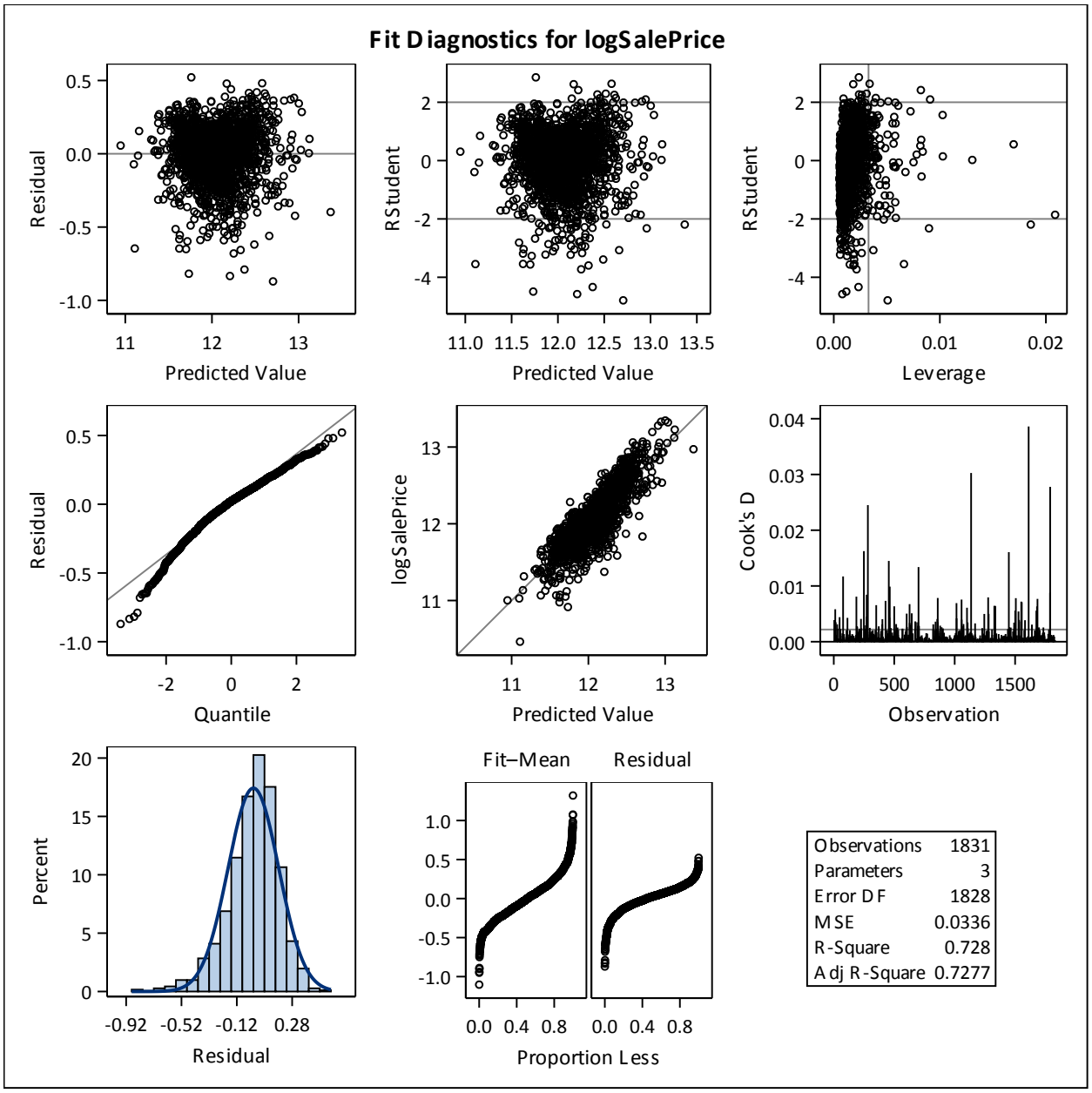
Root MSE	0.18329	R-Square	0.7280
Dependent Mean	12.04898	Adj R-Sq	0.7277
Coeff Var	1.52120		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	6.59378	0.10200	64.64	<.0001	6.39373	6.79383
logGrLivArea	1	0.69875	0.01464	47.73	<.0001	0.67003	0.72746
TotalBsmtSF	1	0.00035309	0.00001252	28.21	<.0001	0.00032855	0.00037764



Appendix A

Results from log(SalePrice) v. log(GrLivArea) and TotalBsmtSF [continued]



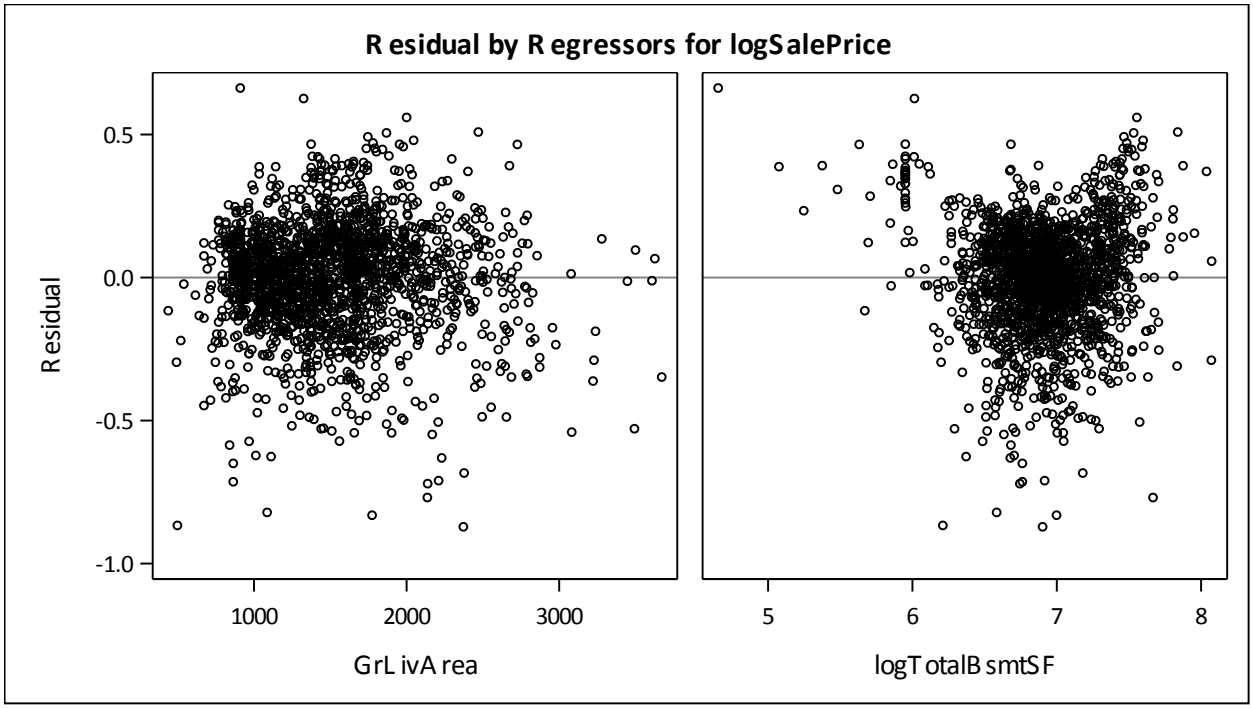
Appendix A

Results from log(SalePrice) v. GrLivArea and log(TotalBsmtSF)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	159.68420	79.84210	2209.56	<.0001
Error	1828	66.05440	0.03613		
Corrected Total	1830	225.73860			

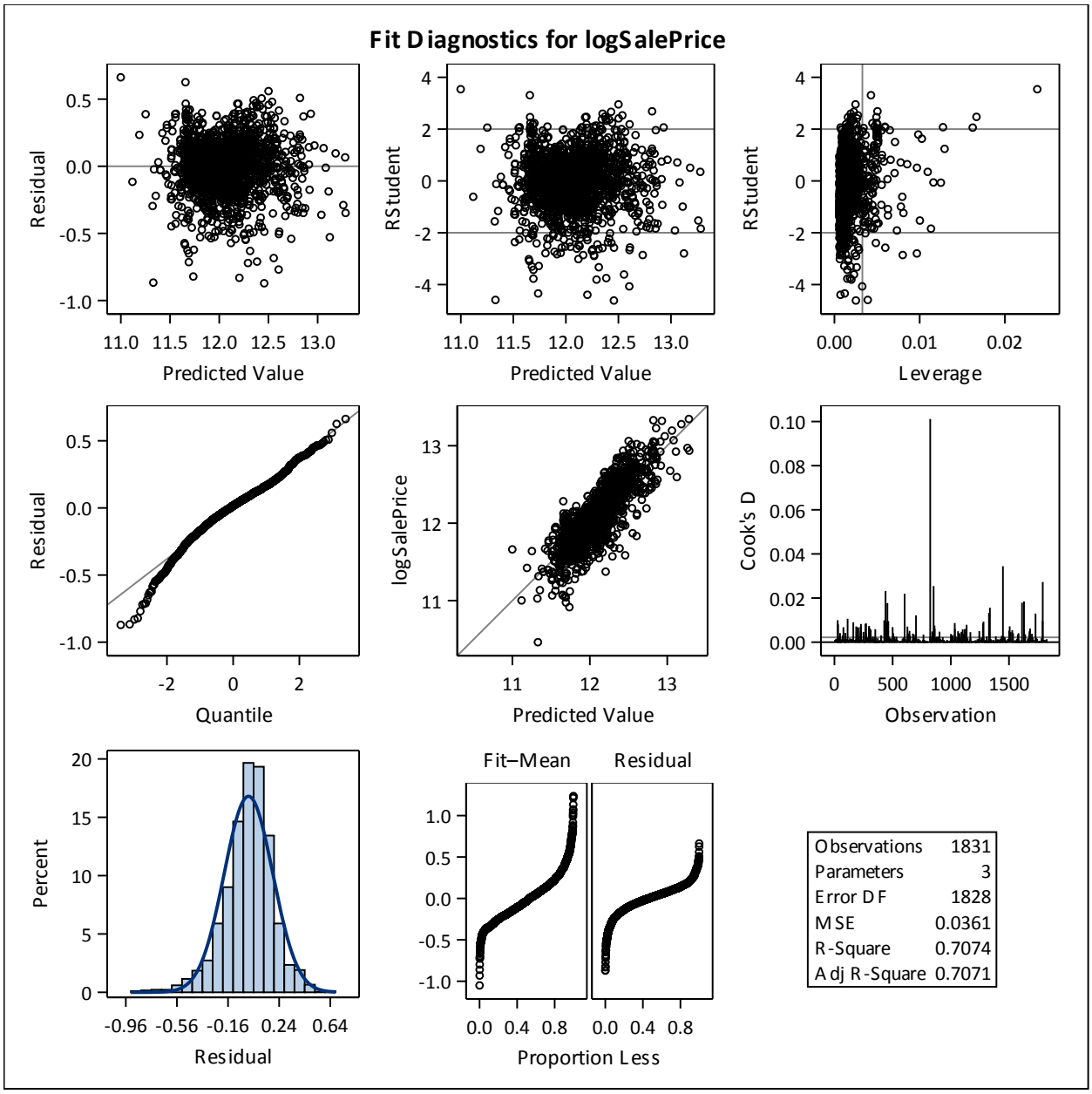
Root MSE	0.19009	R-Square	0.7074
Dependent Mean	12.04898	Adj R-Sq	0.7071
Coeff Var	1.57766		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	8.98845	0.08945	100.48	<.0001	8.81301	9.16389
GrLivArea	1	0.00047639	0.00000973	48.94	<.0001	0.00045730	0.00049548
logTotalBsmtSF	1	0.33879	0.01352	25.07	<.0001	0.31228	0.36530



Appendix A

Results from log(SalePrice) v. GrLivArea and log(TotalBsmtSF) [continued]



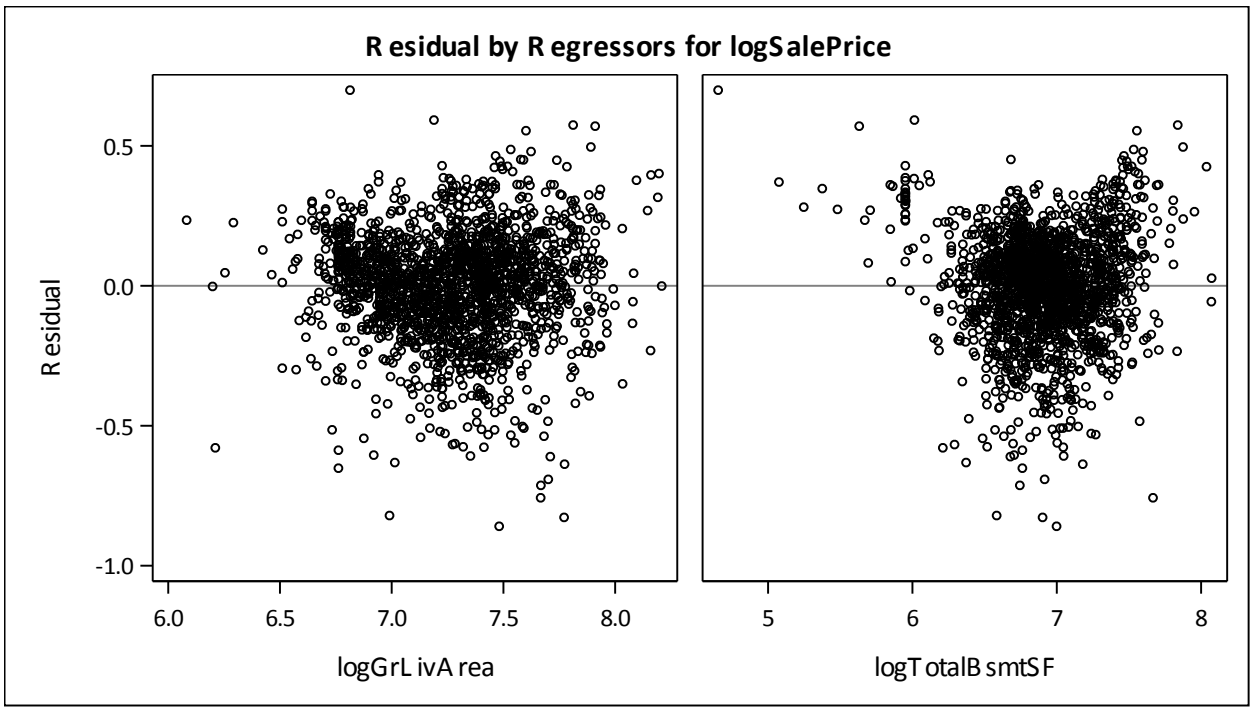
Appendix A

Results from log(SalePrice) v. log(GrLivArea) and log(TotalBsmtSF)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	159.54013	79.77007	2202.77	<.0001
Error	1828	66.19847	0.03621		
Corrected Total	1830	225.73860			

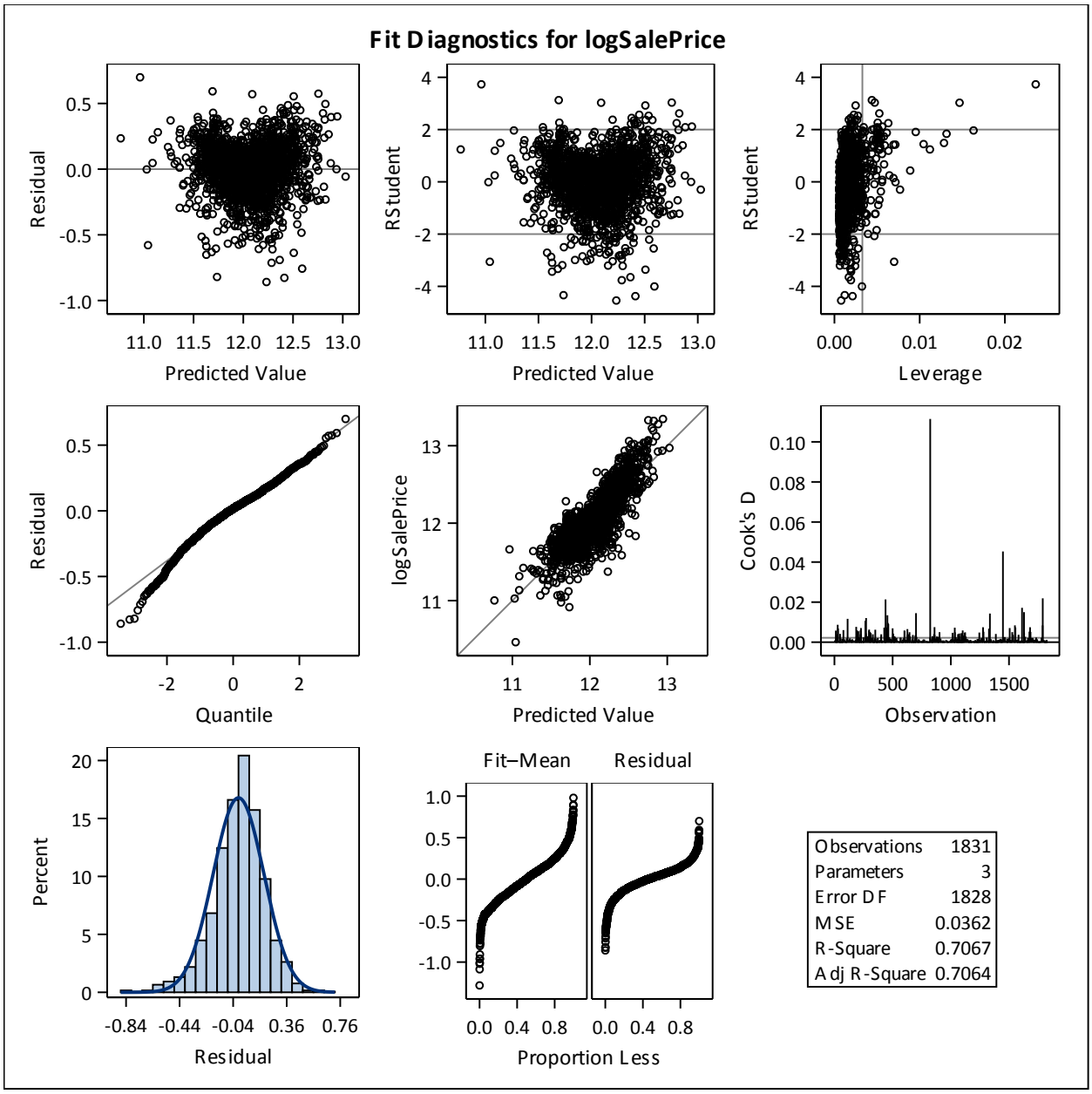
Root MSE	0.19030	R-Square	0.7067
Dependent Mean	12.04898	Adj R-Sq	0.7064
Coeff Var	1.57938		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	4.43201	0.11563	38.33	<.0001	4.20523	4.65880
logGrLivArea	1	0.73028	0.01495	48.84	<.0001	0.70096	0.75961
logTotalBsmtSF	1	0.33402	0.01357	24.62	<.0001	0.30741	0.36063



Appendix A

Results from log(SalePrice) v. log(GrLivArea) and log(TotalBsmtSF) [continued]



Appendix A

Results from log(SalePrice) v. sqrt(GrLivArea) and TotalBsmtSF

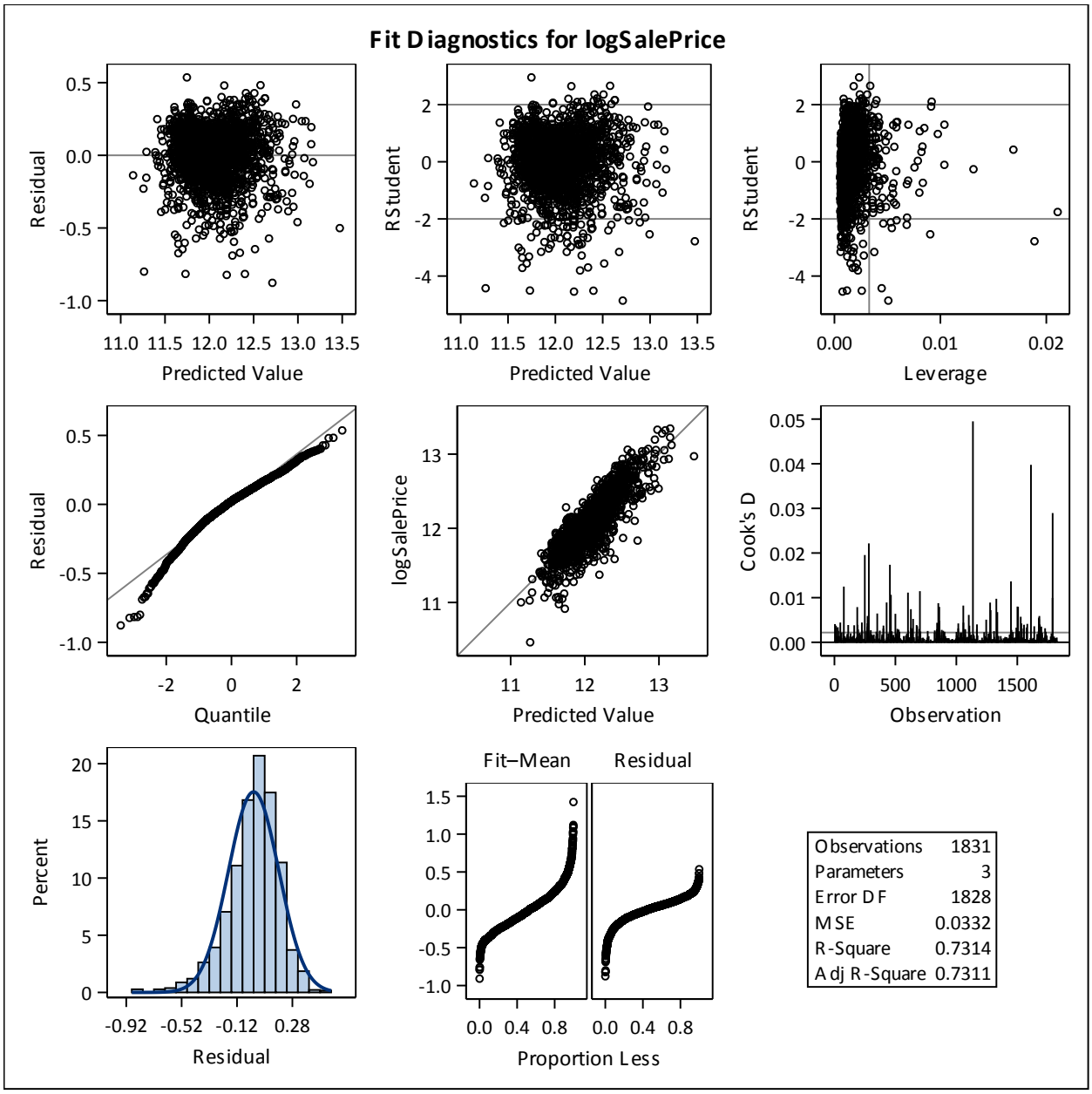
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	165.10393	82.55197	2488.76	<.0001
Error	1828	60.63467	0.03317		
Corrected Total	1830	225.73860			

Root MSE	0.18213	R-Square	0.7314
Dependent Mean	12.04898	Adj R-Sq	0.7311
Coeff Var	1.51155		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	10.27314	0.02708	379.36	<.0001	10.22003	10.32625
GrLivAreaSqRt	1	0.03655	0.00075697	48.28	<.0001	0.03506	0.03803
TotalBsmtSF	1	0.00035097	0.00001244	28.20	<.0001	0.00032656	0.00037537

Appendix A

Results from log(SalePrice) v. sqrt(GrLivArea) and TotalBsmtSF [continued]



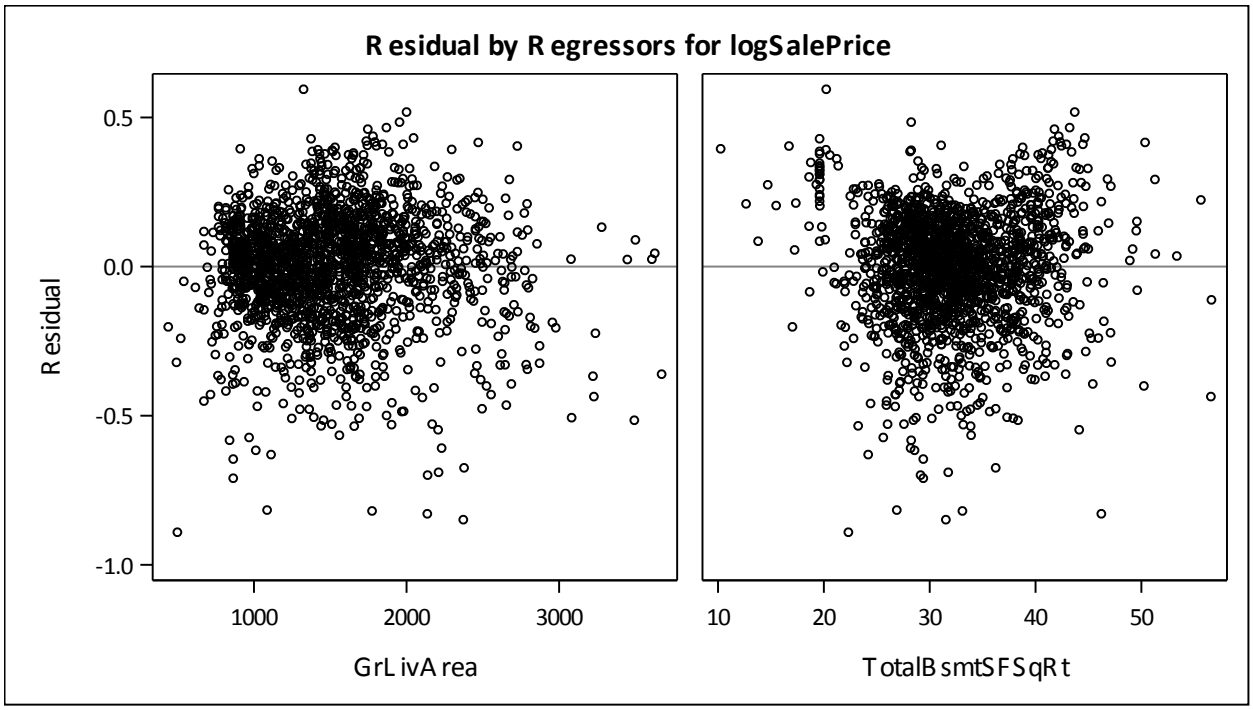
Appendix A

Results from log(SalePrice) v. GrLivArea and sqrt(TotalBsmtSF)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	162.62314	81.31157	2355.01	<.0001
Error	1828	63.11546	0.03453		
Corrected Total	1830	225.73860			

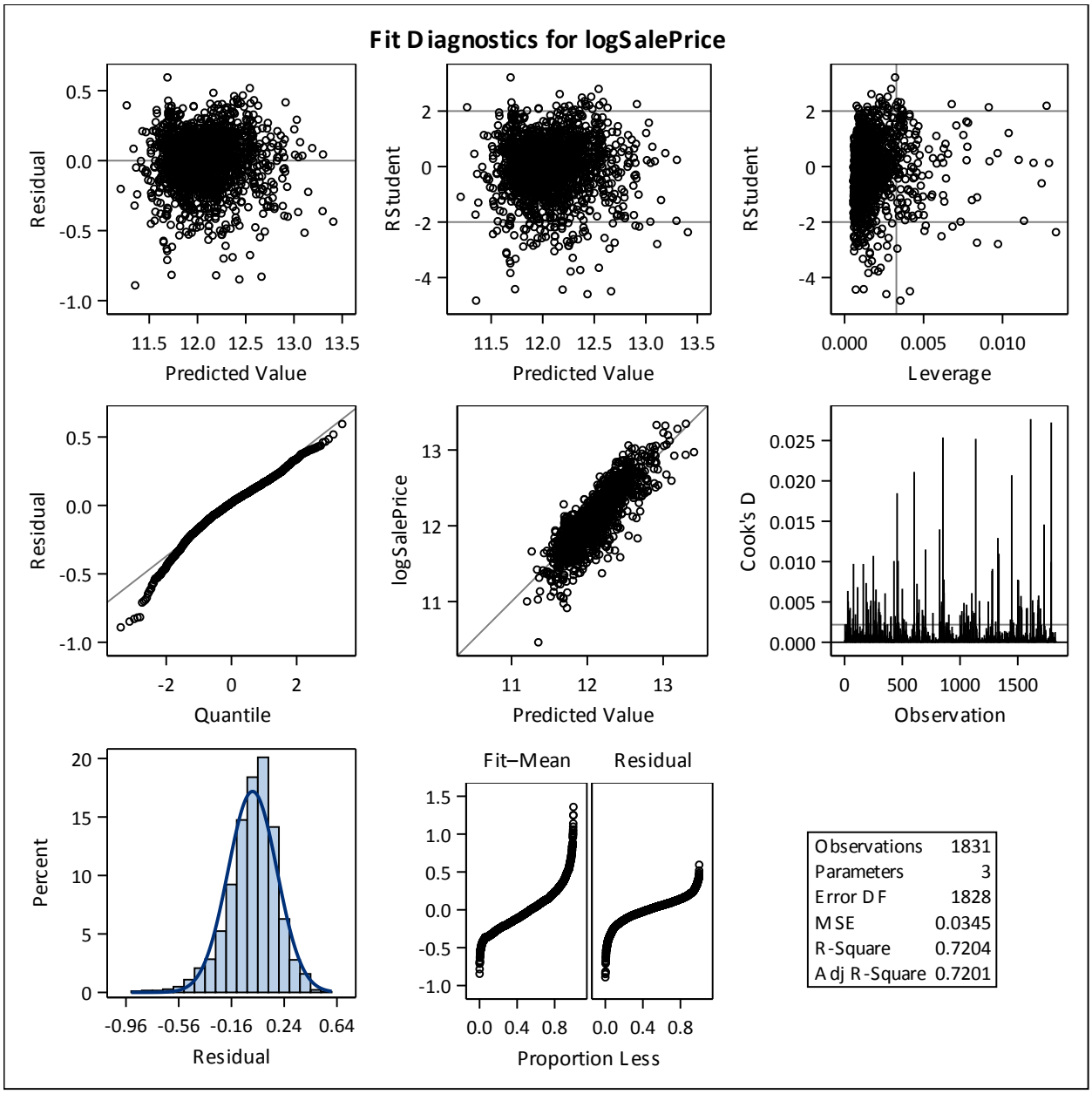
Root MSE	0.18581	R-Square	0.7204
Dependent Mean	12.04898	Adj R-Sq	0.7201
Coeff Var	1.54216		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	10.60883	0.02593	409.11	<.0001	10.55797	10.65968
GrLivArea	1	0.00046303	0.00000962	48.12	<.0001	0.00044416	0.00048190
TotalBsmtSFSqRt	1	0.02305	0.00084585	27.25	<.0001	0.02139	0.02471



Appendix A

Results from log(SalePrice) v. GrLivArea and sqrt(TotalBsmtSF) [continued]



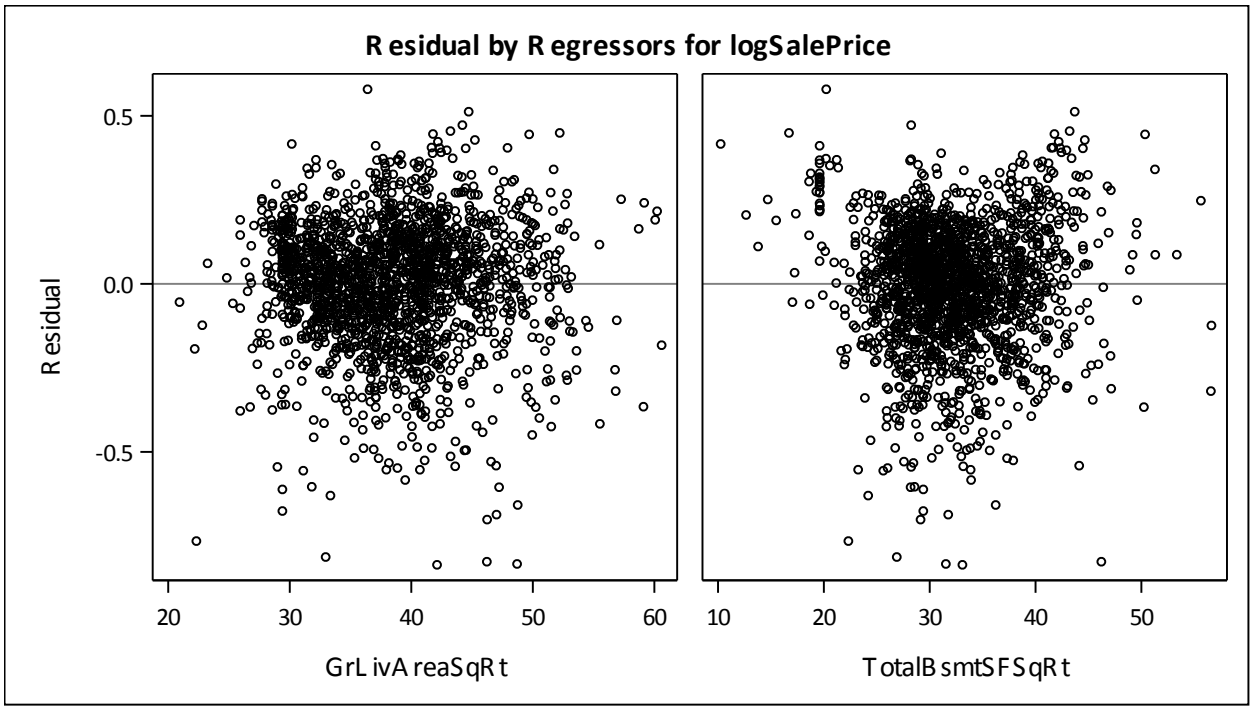
Appendix A

Results from log(SalePrice) v. sqrt(GrLivArea) and sqrt(TotalBsmtSF)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	163.70511	81.85255	2412.03	<.0001
Error	1828	62.03349	0.03394		
Corrected Total	1830	225.73860			

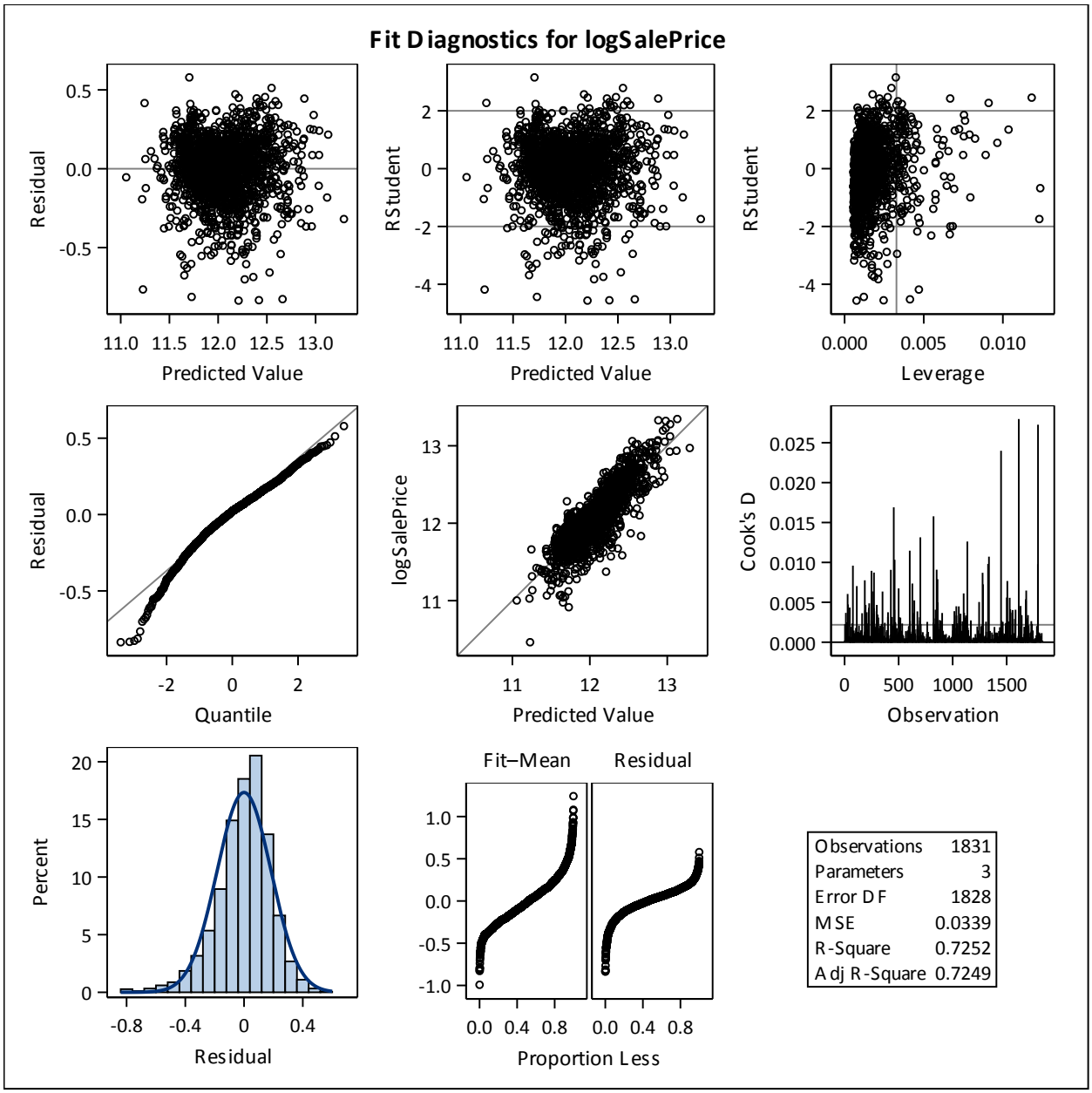
Root MSE	0.18422	R-Square	0.7252
Dependent Mean	12.04898	Adj R-Sq	0.7249
Coeff Var	1.52888		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	9.89084	0.03164	312.61	<.0001	9.82879	9.95289
GrLivAreaSqRt	1	0.03715	0.00076025	48.87	<.0001	0.03566	0.03864
TotalBsmtSFSqRt	1	0.02279	0.00083980	27.13	<.0001	0.02114	0.02443



Appendix A

Results from log(SalePrice) v. sqrt(GrLivArea) and sqrt(TotalBsmtSF) [continued]



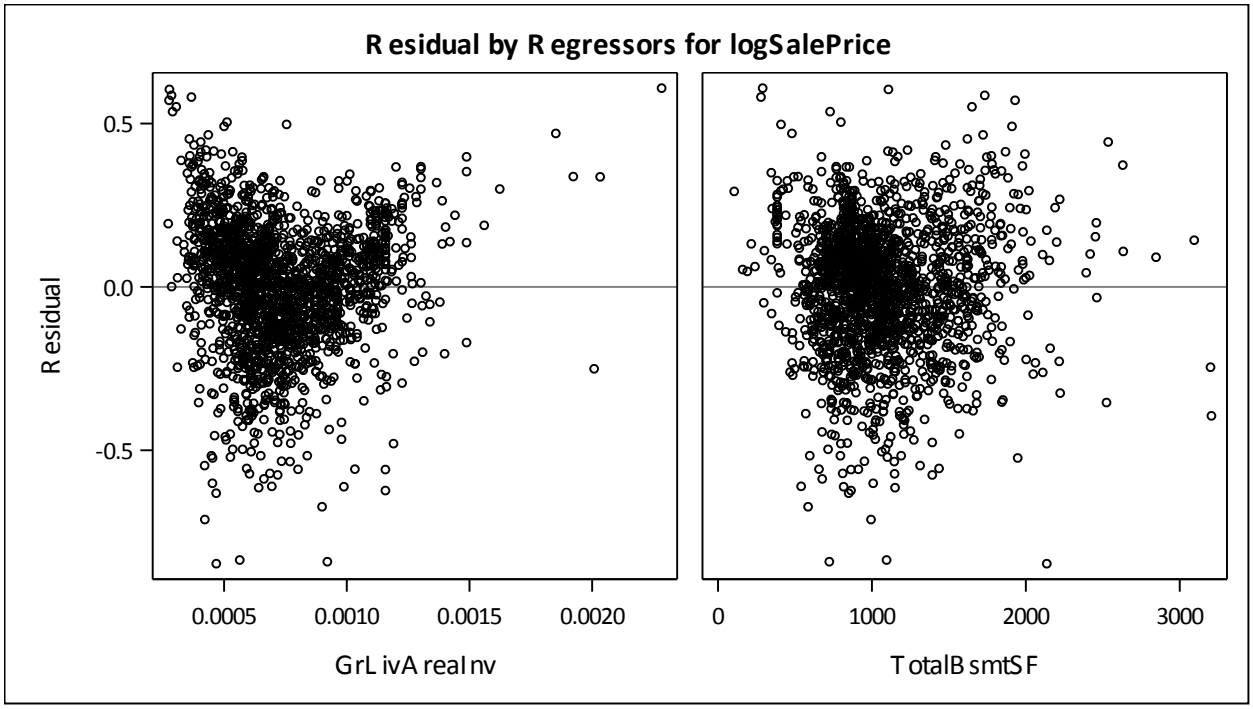
Appendix A

Results from log(SalePrice) v. inv(GrLivArea) and TotalBsmtSF

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	157.23236	78.61618	2097.77	<.0001
Error	1828	68.50624	0.03748		
Corrected Total	1830	225.73860			

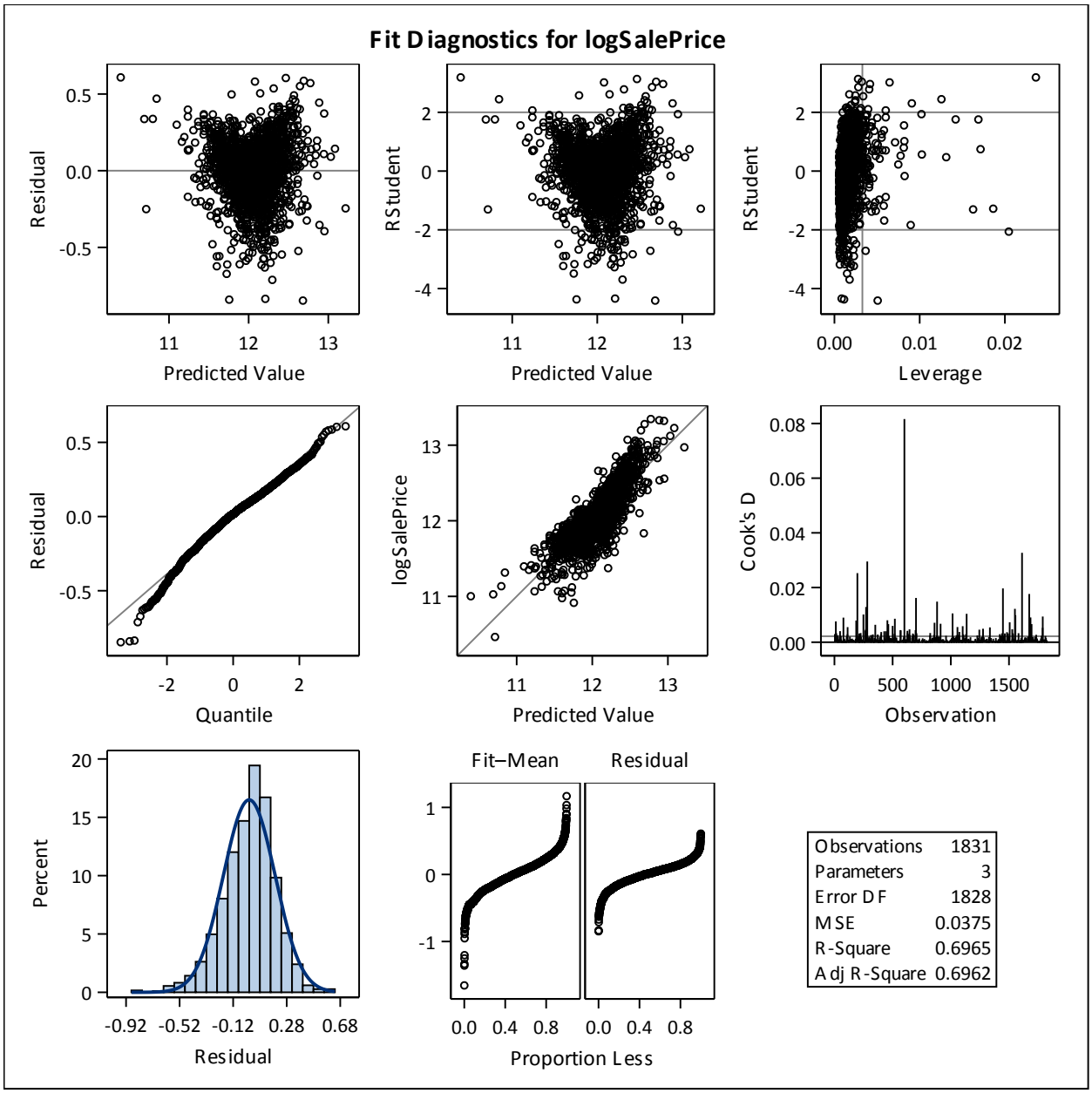
Root MSE	0.19359	R-Square	0.6965
Dependent Mean	12.04898	Adj R-Sq	0.6962
Coeff Var	1.60667		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	12.29881	0.02453	501.33	<.0001	12.25070	12.34693
GrLivAreaInv	1	-881.72234	20.48274	-43.05	<.0001	-921.89437	-841.55032
TotalBsmtSF	1	0.00037236	0.00001315	28.32	<.0001	0.00034657	0.00039814



Appendix A

Results from log(SalePrice) v. inv(GrLivArea) and TotalBsmtSF [continued]



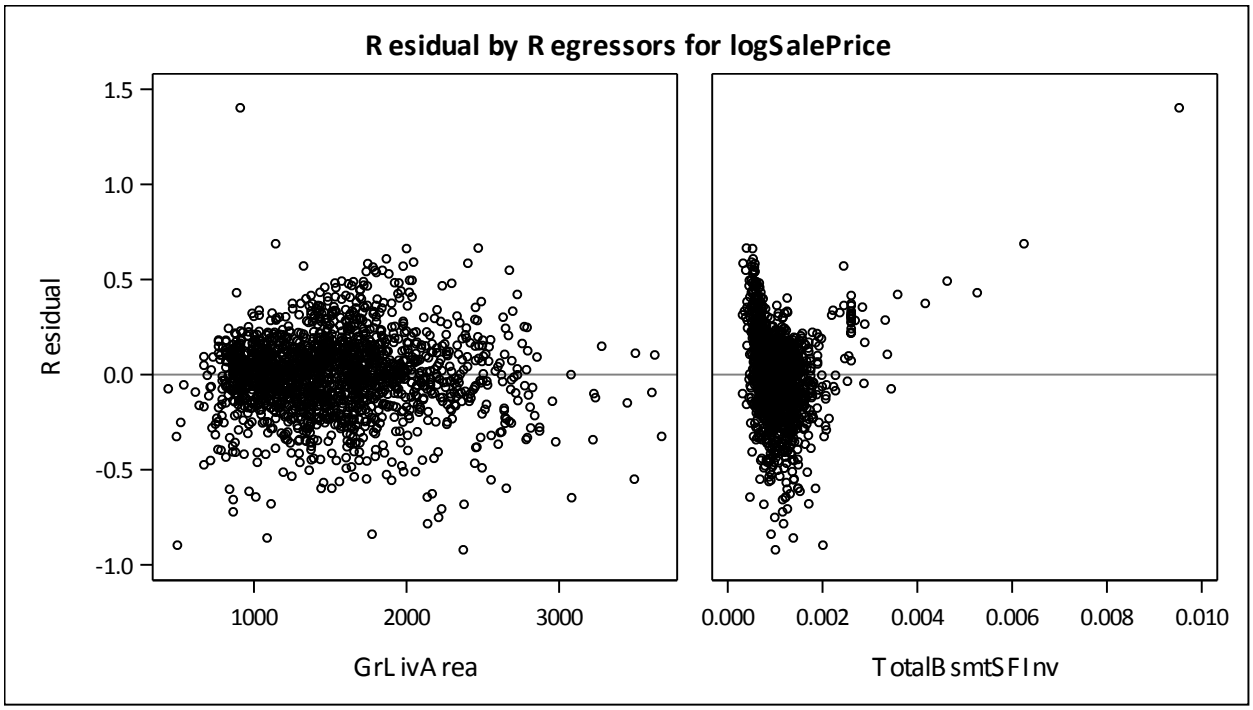
Appendix A

Results from log(SalePrice) v. GrLivArea and inv(TotalBsmtSF)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	148.73874	74.36937	1765.55	<.0001
Error	1828	76.99986	0.04212		
Corrected Total	1830	225.73860			

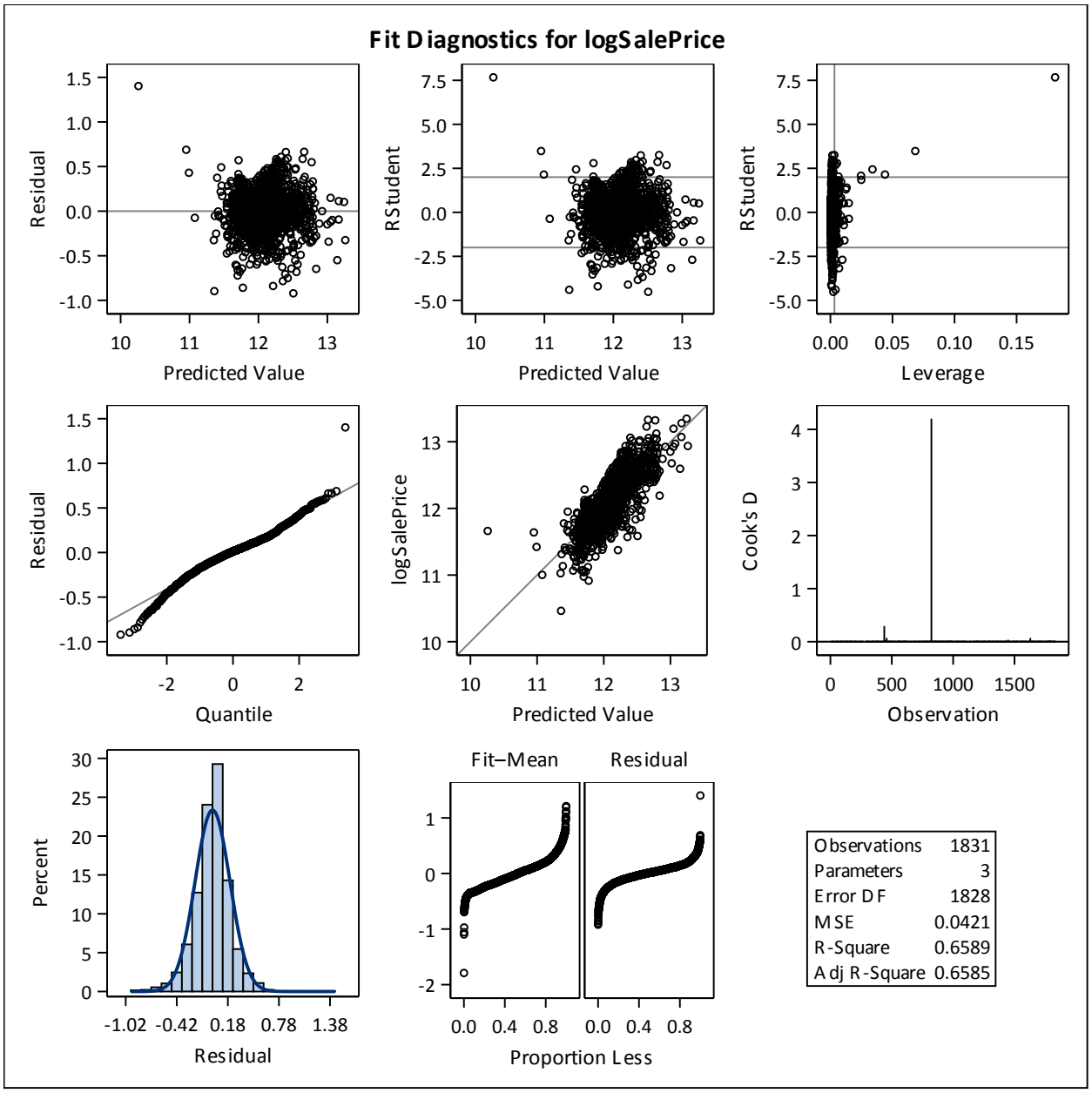
Root MSE	0.20524	R-Square	0.6589
Dependent Mean	12.04898	Adj R-Sq	0.6585
Coeff Var	1.70336		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	11.45284	0.02170	527.73	<.0001	11.41027	11.49540
GrLivArea	1	0.00051833	0.00001018	50.90	<.0001	0.00049836	0.00053830
TotalBsmtSFInv	1	-174.90016	10.46842	-16.71	<.0001	-195.43149	-154.36883



Appendix A

Results from log(SalePrice) v. GrLivArea and inv(TotalBsmtSF) [continued]



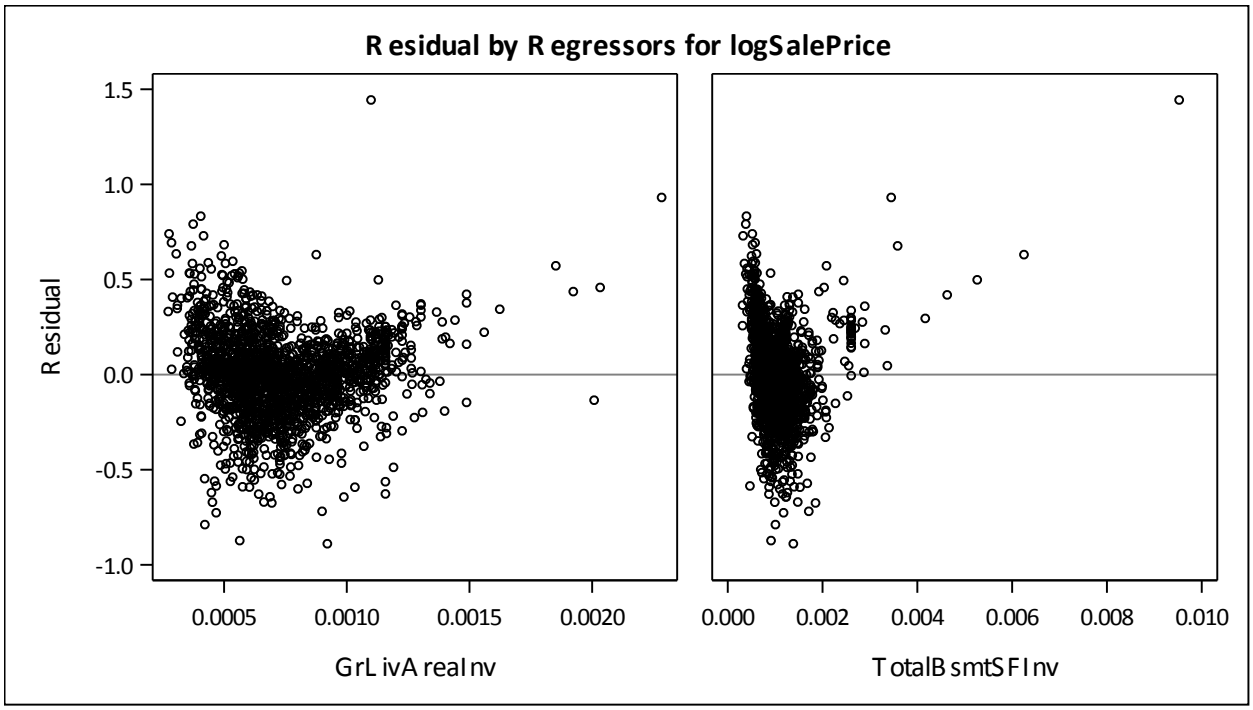
Appendix A

Results from log(SalePrice) v. inv(GrLivArea) and inv(TotalBsmtSF)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	138.54202	69.27101	1452.21	<.0001
Error	1828	87.19658	0.04770		
Corrected Total	1830	225.73860			

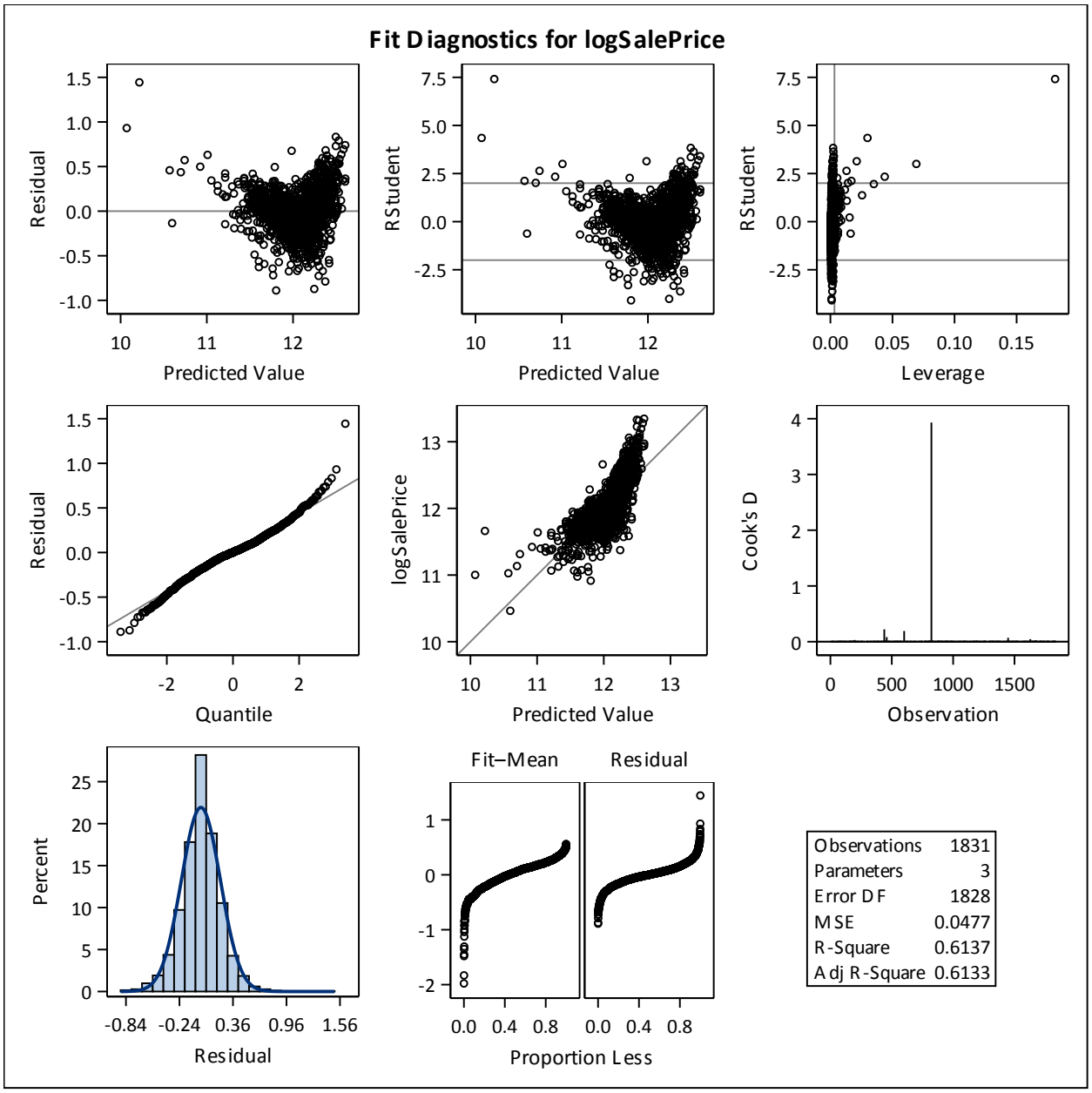
Root MSE	0.21840	R-Square	0.6137
Dependent Mean	12.04898	Adj R-Sq	0.6133
Coeff Var	1.81264		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	12.97364	0.01811	716.48	<.0001	12.93813	13.00916
GrLivAreaInv	1	-1010.45288	22.18638	-45.54	<.0001	-1053.96620	-966.93957
TotalBsmtSFInv	1	-172.80021	11.18945	-15.44	<.0001	-194.74567	-150.85475



Appendix A

Results from log(SalePrice) v. inv(GrLivArea) and inv(TotalBsmtSF) [continued]



Code

```
*************************************************************************;
* 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;

* Make 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 transformation variables;
data sample;
set temp;
if (drop_condition='09: Sample Population');
logSalePrice = log(SalePrice);
logGrLivArea = log(GrLivArea);
logTotalBsmtSF = log(TotalBsmtSF);
GrLivAreaSqRt = sqrt(GrLivArea);
TotalBsmtSFSqRt = sqrt(TotalBsmtSF);
GrLivAreaInv = 1/(GrLivArea);
TotalBsmtSFInv = 1/(TotalBsmtSF);
* Would be interesting to use for outliers,
but outliers cannot be defined by the response variable;
* PricePerSqFtGrLiv=(SalePrice/GrLivArea);
* PricePerSqFtBsmt=(SalePrice/TotalBsmtSF);
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;

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

* See which variables have strongest correlation with SalePrice;
ods graphics on;
proc corr data=sample plots=matrix;
var SalePrice GrLivArea TotalBsmtSF LotFrontage LotArea MasVnrArea BsmtFinSF1
BsmtFinSF2 BsmtUnfSF FirstFlrSF SecondFlrSF LowQualFinSF GarageArea WoodDeckSF
OpenPorchSF EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea MiscVal;
title;
```

```
run;
ods graphics off;

* Create plots of top 3 predictors + response variable;
proc sgplot data=sample;
histogram SalePrice;
density SalePrice / type=normal;
density SalePrice / type=kernel;
run; quit;

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

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

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

* Create matrix of 2 most promising predictors;
ods graphics on;
proc corr data=sample plots(maxpoints=none)=matrix;
var SalePrice GrLivArea TotalBsmtSF;
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;

*************************************************************************;
* 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;
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 Models;
*************************************************************************;

* 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;

* Create multiple linear regression model;
proc reg data=sample alpha=0.05;
model SalePrice=GrLivArea TotalBsmtSF / clb;
output out=sample_out cookd=cookd;
run;
```

```
*************************************************************************;
* Remove Outliers;
*************************************************************************;

* EDA for predictor variables;
proc means data=sample min q1 mean median q3 max qrange;
var GrLivArea TotalBsmtSF;
run; quit;

proc options option=macro;
run;

* Create IQR Macro;
%macro PredictorsIQR(indata,x,outdata);
proc means data=&indata q1 q3 qrange;
var &x;
output out=&outdata q1=q1 q3=q3 qrange=qrange;
run;
quit;
%mend PredictorsIQR;

* Calculate IQR for GrLivArea;
%PredictorsIQR(indata=%str(sample),
x=%str(GrLivArea),
outdata=%str(GrLivArea_IQR));

* Calculate upper and lower fence for GrLivArea;
data GrLivArea_IQR;
set GrLivArea_IQR;
GrLiv_LL = q1 - 1.5*qrange;
GrLiv_UL = q3 + 1.5*qrange;
run; quit;

proc print data=GrLivArea_IQR;
run; quit;

* Lower Limit: 167;
* Upper Limit: 2735;

* Calculate IQR for TotalBsmtSF;
%PredictorsIQR(indata=%str(sample),
x=%str(TotalBsmtSF),
outdata=%str(TotalBsmtSF_IQR));

* Calculate upper and lower fence for TotalBsmtSF;
data TotalBsmtSF_IQR;
set TotalBsmtSF_IQR;
Bsmt_LL = q1 - 1.5*qrange;
Bsmt_UL = q3 + 1.5*qrange;
run; quit;
```

```
proc print data=TotalBsmtSF_IQR;
run; quit;

* Lower Limit: 183;
* Upper Limit: 1887;

* Create outliers waterfall;

data temp2;
	set sample_out;
	format outlier_def $40.;
	if (GrLivArea < 167 or GrLivArea > 2735) then do;
		outlier_def = '1. GrLivArea Outliers';
		outlier_code = 1;
	end;
	else if (TotalBsmtSF < 183 or TotalBsmtSF > 1887) then do;
		outlier_def = '2. TotalBsmtSF Outliers';
		outlier_code = 2;
	end;
	else if (cookd > 4/1831) then do;
		outlier_def = '3. Cook D Outliers';
		outlier_code = 3;
	end;
	else do;
		outlier_def = '4. Not An Outlier';
		outlier_code = 0;
	end;
run;

* View frequency of outliers;
proc freq data=temp2;
tables outlier_def;
title 'Outliers';
run; quit;

* Subset data for just sample without outliers;
data outliers;
set temp2;
if (outlier_code > 0) then delete;
run;

* Create outliers multiple linear regression model;
proc reg data=outliers alpha=0.05;
model SalePrice=GrLivArea TotalBsmtSF / clb;
run;

* Create unpacked outliers model;
ods graphics on;
proc reg data=outliers alpha=0.05 plots(unpack);
model SalePrice=GrLivArea TotalBsmtSF / clb;
```

```
run;
ods graphics off;

************************************************************************;
* Create Transformed Models;
************************************************************************;

* Create log response 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 response multiple linear regression model;
proc reg data=sample alpha=0.05;
model logSalePrice=GrLivArea TotalBsmtSF / clb;
run;

* Create log predictors and log response simple linear regression models;
proc reg data=sample alpha=0.05;
model logSalePrice=logGrLivArea / clb;
run;

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

* Create log predictors and log response multiple linear regression models;
proc reg data=sample alpha=0.05;
model logSalePrice=logGrLivArea TotalBsmtSF / clb;
run;

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

proc reg data=sample alpha=0.05;
model logSalePrice=logGrLivArea logTotalBsmtSF / clb;
run;

* Create square root predictors and log response multiple linear regression models;
proc reg data=sample alpha=0.05;
model logSalePrice=GrLivAreaSqRt TotalBsmtSF / clb;
run;

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

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

* Create inverse predictors and log response multiple linear regression models;
proc reg data=sample alpha=0.05;
model logSalePrice=GrLivAreaInv TotalBsmtSF / clb;
run;

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

proc reg data=sample alpha=0.05;
model logSalePrice=GrLivAreaInv TotalBsmtSFInv / clb;
run;

**************************************************************************;
* Create Unpacked Models;
**************************************************************************;

* 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;
**************************************************************************;
```