Assignment #2
Crystal M. Mosley

Introduction:

The purpose of this task is to continue building upon the Exploratory Data Analysis as previously researched on the Ames, Iowa housing data set. We will begin building regression models for the home SalePrice by fitting specific models. After our EDA of the Ames housing data, we found the following continuous variables: GrLivArea, GarageArea, TotalBsmtSF, FirstFlrSF, MasVnrArea (although, an error in the data was found via code), BsmtFinSF1, and BsmtUnfSF. We'll be choosing to model the continuous variable which correlated near 0.5 with SalePrice since the variable that correlated approximately with 0.5 is displaying and error from the data (MasVrnArea), FirstFlrSF, as well as the variable which linearly correlated with SalePrice the strongest, GrLivArea.

Results:

Simple Linear Regression Models

FirstFlrSF to predict SalePrice

The model used:

$$SalePrice = \beta_0 + \beta_1 FirstFlrSF + \varepsilon$$

Utilize the "REG" procedure to show the estimates of each parameter:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33847	3611.67045	9.37	<.0001
FirstFlrSF	1	126.72825	2.95079	42.95	<.0001

The fitted model:

$$SalePrice = 33847 + 126.72825 \times FirstFlrSF$$

In reviewing the above variables, the model coefficients are indicative that if FirstFlrSF was 0, then the SalePrice of the house would be $33,847. Being that FirstFlrSF means "first floor square feet", it's unlikely that the observations for this variable will be null since all houses have at least one floor. The SalePrice of the house being $33,847 is reasonable since there's the possibility of the houses not having more than one floor level. The average change in the mean of SalePrice is about $127, and a one unit change in FirstFlrSF should be consistent since it's a continuous variable.

Since both t-values are > 0, they are significantly large; moreover, there p-values also show as significant. Because of this, we are able to reject the null hypothesis and conclude that these variables have slope and intercept that are greater than zero.

Goodness-of-fit information:

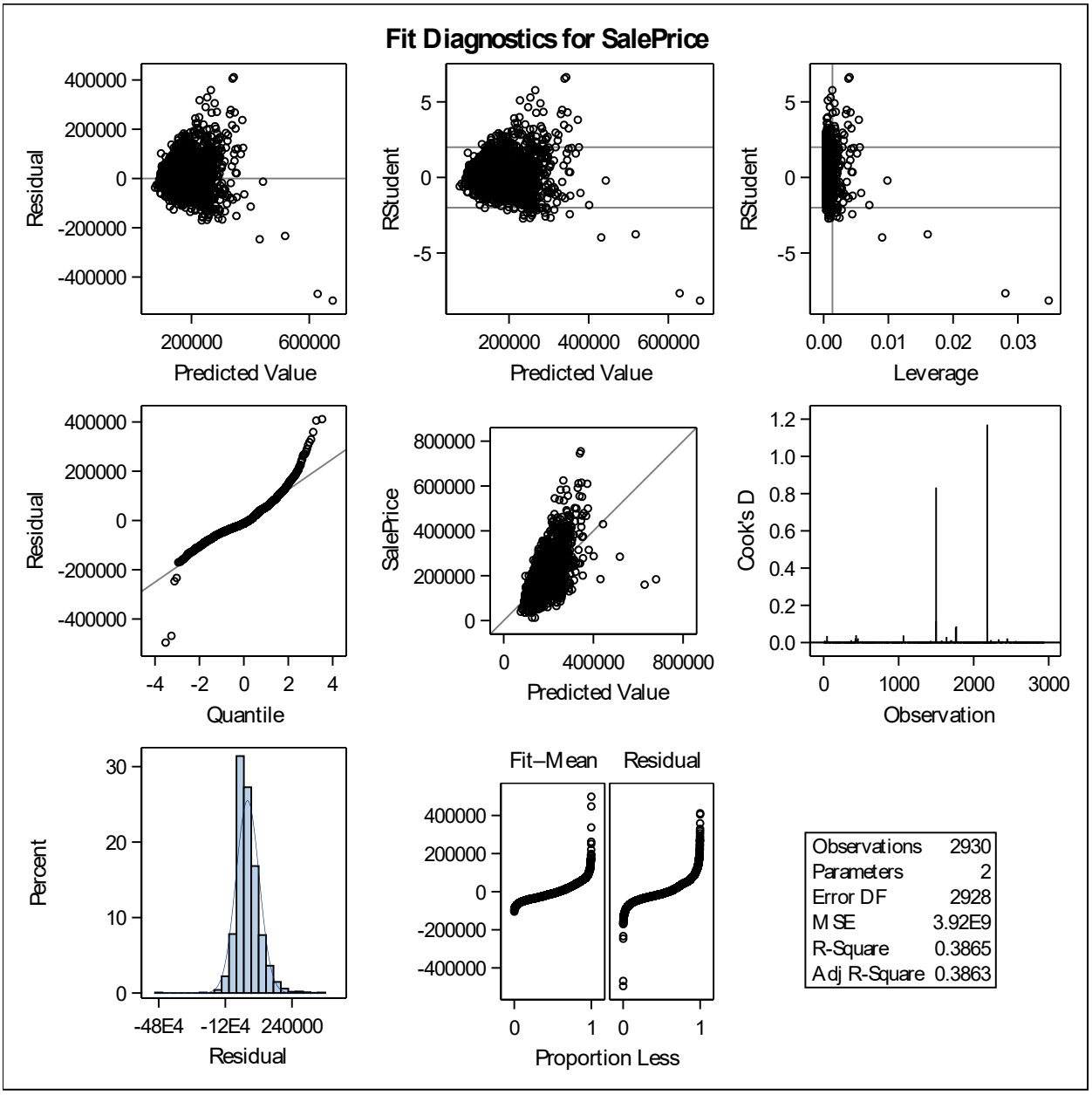
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7.224313E12	7.224313E12	1844.47	<.0001
Error	2928	1.146822E13	3916743268		
Corrected Total	2929	1.869254E13			

After reviewing the ANOVA table, we can see that the F-value is large since it is greater than one. This means the observations and regression will show contrast from the overall mean.

Root MSE	62584	R-Square	0.3865
Dependent Mean	180796	Adj R-Sq	0.3863
Coeff Var	34.61574		

After reviewing the source table, we see that the R-Square for this regression model shows about 38% of the variability in SalePrice while using FirstFlrSF. As the model building process continues, we will keep a close eye on the Adj. R-Square percentage to compare our models' performance in relation to the sample size and number of variables that are included in the model.

ODS graphics for Fit Diagnostics – SalePrice



In assessing the normality of the residuals using the QQ Plot, we can see that the plot is heave-tailed. We can conclude that these observations do not follow the assumed distribution. There are two noticeable outliers at the bottom left corner of the graph that may be cause for an alarm. After observing the histogram of the residuals, the data appears to be normal.

It's clear that the graphs are showing heteroscedasticity, as opposed to homoscedasticity.

Next, we observe Cook's Distance. Cook's Distance (Cook's D.) represents values of "D" that are substantially larger than the rest. Utilizing the calculation for threshold of Cook's D., we'll be able to establish whether or not the spikes in the graph (data points) are influential.

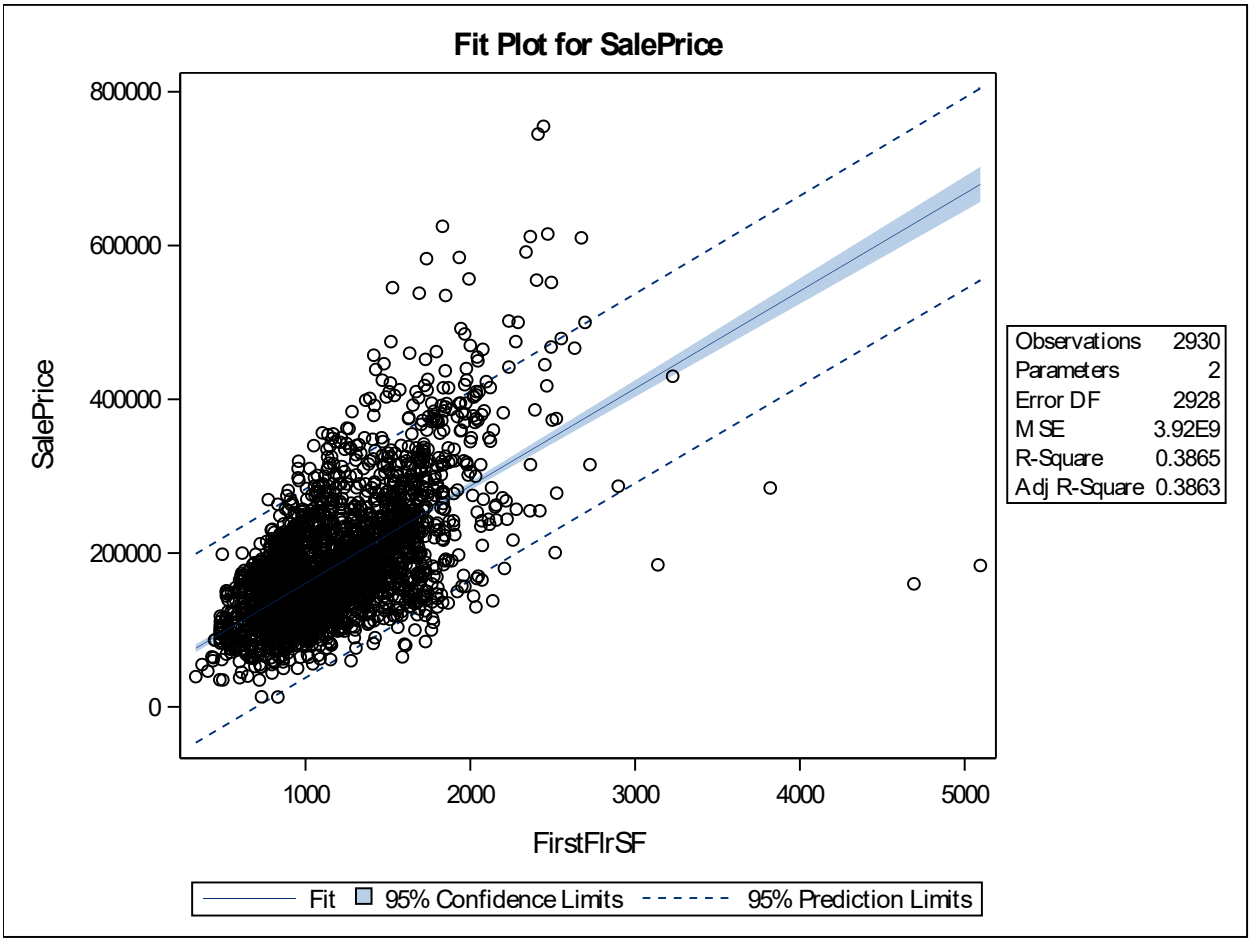
$$\frac{4}{(N-k-1)}$$

N = num. of corrected observations
k = num. of explanatory variables

$\frac{4}{(2929-1-1)}$ = 0.0013

Seeing that there are a couple of spikes over the 0.0013 threshold, this is indicative that there are some observations that *are* influential.

Fit Plot – SalePrice



For the most part, this fit plot is showing a positive, linear trend of heteroscedasticity.

Finding Best Continuous Variable Using R-Square Selection

Model used:

$$SalePrice = \beta_0 + \beta_{1'}Best' + \varepsilon$$

Utilize the Reg procedure with selections based on the *r-square* metric:

Model Comparison—

Number in Model	R-Square	Variables in Model
1	0.4995	GrLivArea
1	0.4100	GarageArea
1	0.3998	TotalBsmtSF
1	0.3864	FirstFlrSF
1	0.1875	BsmtFinSF1
1	0.0334	BsmtUnfSF

Next, we review the parameter estimates for GrLivArea:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13290	3269.70277	4.06	<.0001
GrLivArea	1	111.69400	2.06607	54.06	<.0001

The fitted model:

SalePrice = 13290 + 111.69400 x GrLivArea

In reviewing the above variables, the model coefficients are indicative that if GrLivArea was 0, then the SalePrice of the house would be $13,290. Being that GrLivArea means “above ground living area square feet”, it’s highly likely that this variable will not have any null observations. The average change in the mean of SalePrice is about $111.69, and a one unit change in GrLivArea should be consistent since it’s a continuous variable.

Since both t-values are > 0, they are significantly large; moreover, there p-values also show as significant. Because of this, we are able to reject the null hypothesis and conclude that these variables have slope and intercept that are greater than zero.

Goodness-of-fit information:

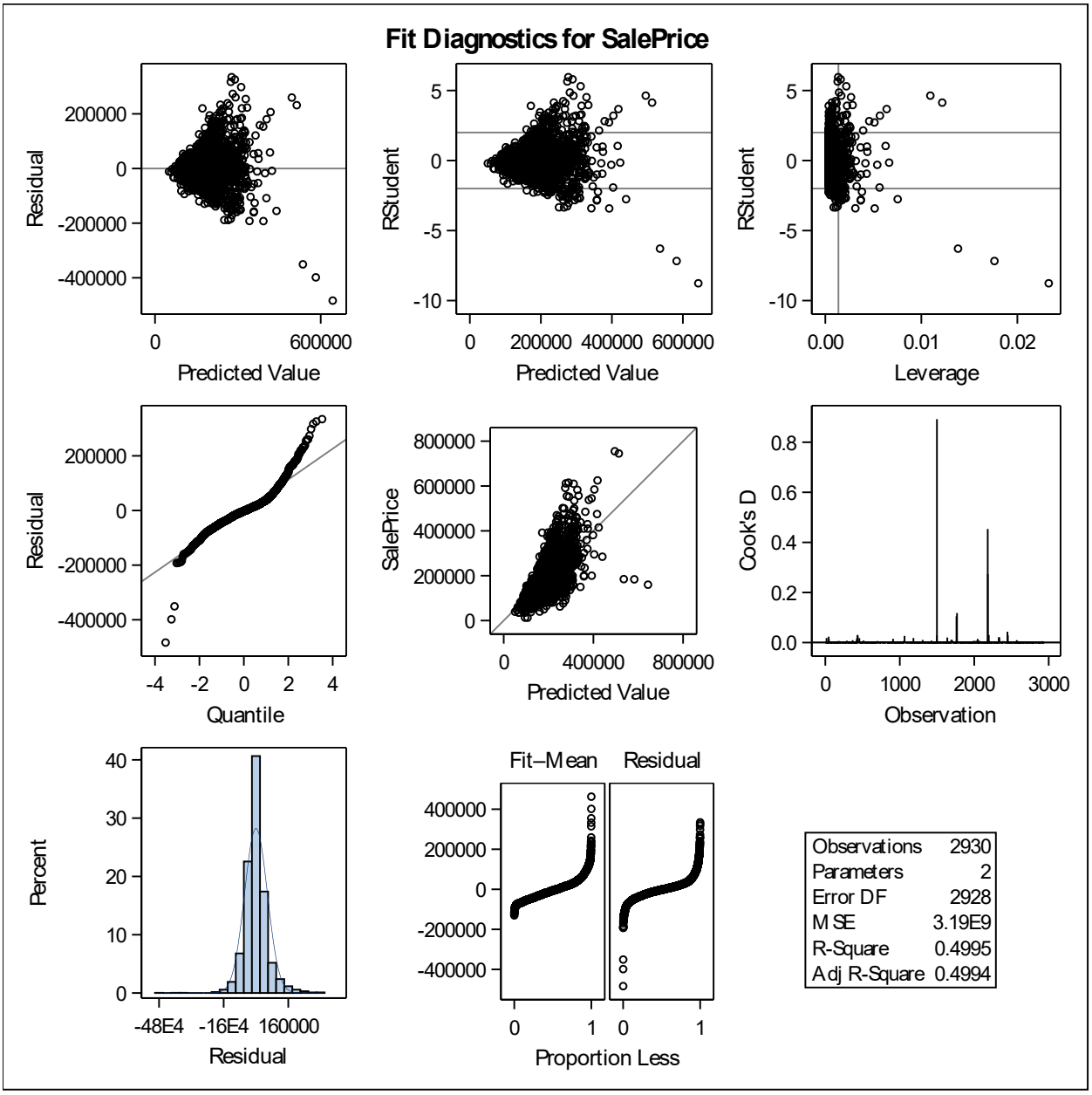
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9.33763E12	9.33763E12	2922.59	<.0001
Error	2928	9.354907E12	3194981962		
Corrected Total	2929	1.869254E13			

After reviewing the ANOVA table, we can see that the F-value is large since it is greater than one. This means the observations and regression will show contrast from the overall mean.

Root MSE	56524	R-Square	0.4995
Dependent Mean	180796	Adj R-Sq	0.4994
Coeff Var	31.26405		

After reviewing the source table, we see that the R-Square for this regression model shows about 50% of the variability in SalePrice while using GrLivArea. As the model building process continues, we will keep a close eye on the Adj. R-Square percentage to compare our models' performance in relation to the sample size and number of variables that are included in the model.

ODS graphics for Fit Diagnostics – SalePrice



In assessing the normality of the residuals using the QQ Plot, we can see that the plot is heave-tailed. We can conclude that these observations do not follow the assumed distribution. There are three noticeable outliers at the bottom left corner of the graph that may be cause for an alarm. After observing the histogram of the residuals, the data appears to be normal.

It's clear that the graphs are showing heteroscedasticity, as opposed to homoscedasticity.

Next, we observe Cook's Distance. Cook's Distance (Cook's D.) represents values of "D" that are substantially larger than the rest. Utilizing the calculation for threshold of Cook's D., we'll be able to establish whether or not the spikes in the graph (data points) are influential.

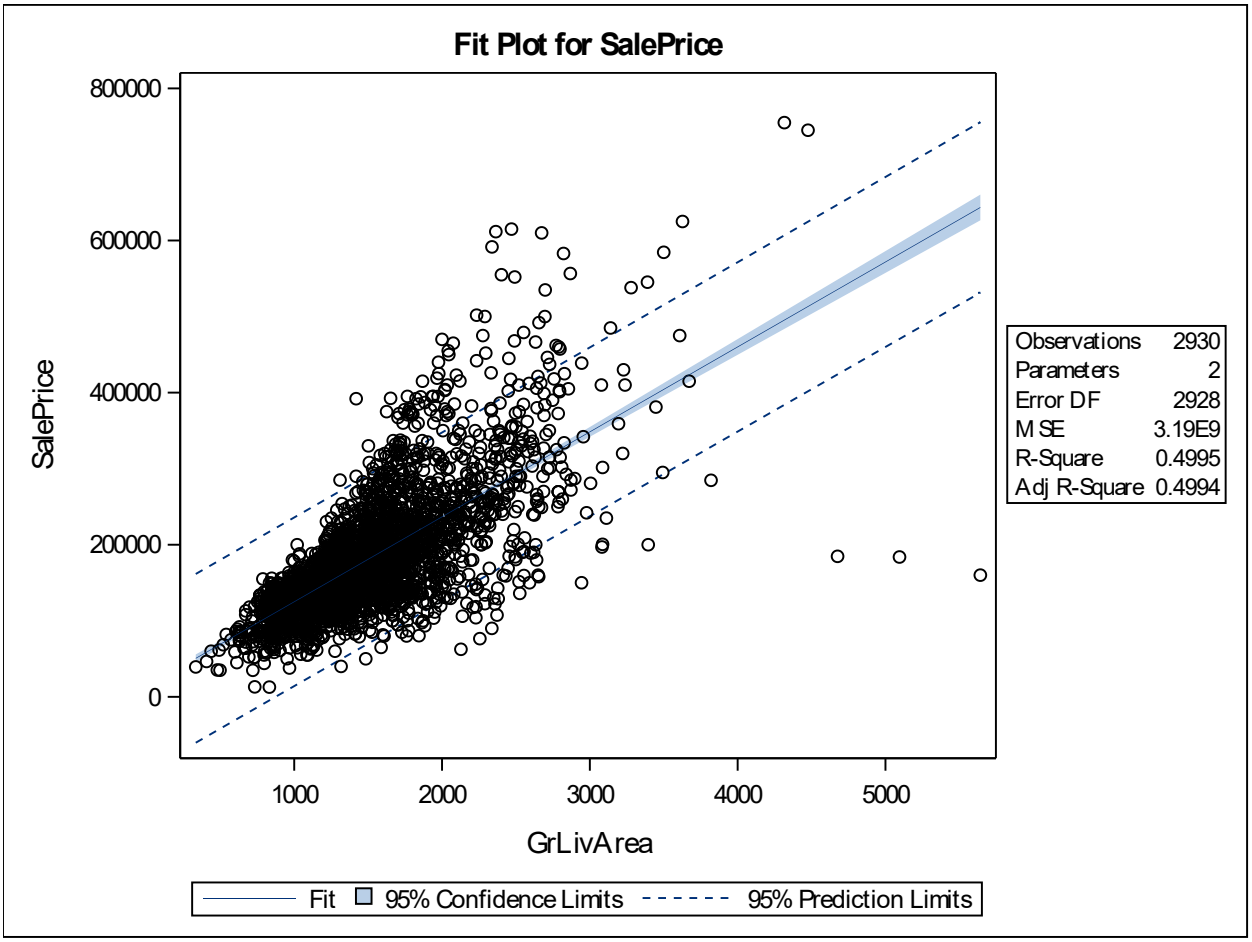
$$\frac{4}{(N-k-1)}$$

N = num. of corrected observations
k = num. of explanatory variables

$\frac{4}{(2929-1-1)}$ = 0.0013

Seeing that there are a couple of spikes over the 0.0013 threshold, this is indicative that there are some observations that *are* influential.

Fit Plot – SalePrice



For the most part, this fit plot is showing a positive, linear trend of heteroscedasticity. Being that the Adj. R-Square can only explain about 50% of variability in our predictor, it may be of use to utilize a categorical variable or a model of more complexity to explain the variability.

Fireplaces to predict SalePrice

The model used:

$$SalePrice = \beta_0 + \beta_1 Fireplaces + \varepsilon$$

Utilize the "REG" procedure to show the estimates of each parameter:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	145729	1770.04460	82.33	<.0001
Fireplaces	1	58512	2005.67334	29.17	<.0001

The fitted model:

$$SalePrice = 145729 + 58512 \times Fireplaces$$

In reviewing the above variables, the model coefficients are indicative that if Fireplaces was 0, then the SalePrice of the house would be $145,729. Being that Fireplaces means "number of fireplaces", it's likely that some observations about be null since not all houses are required to have fireplaces; moreover, these houses may only have one fireplace rather than multiple. If there is a one unit change, this model tells us that the average change in the mean of SalePrice is about $58512. This is a large slope for this categorical variable.

Since both t-values are > 0, they are significantly large; moreover, there p-values also show as significant. Because of this, we are able to reject the null hypothesis and conclude that these variables have slope and intercept that are greater than zero.

Goodness-of-fit information:

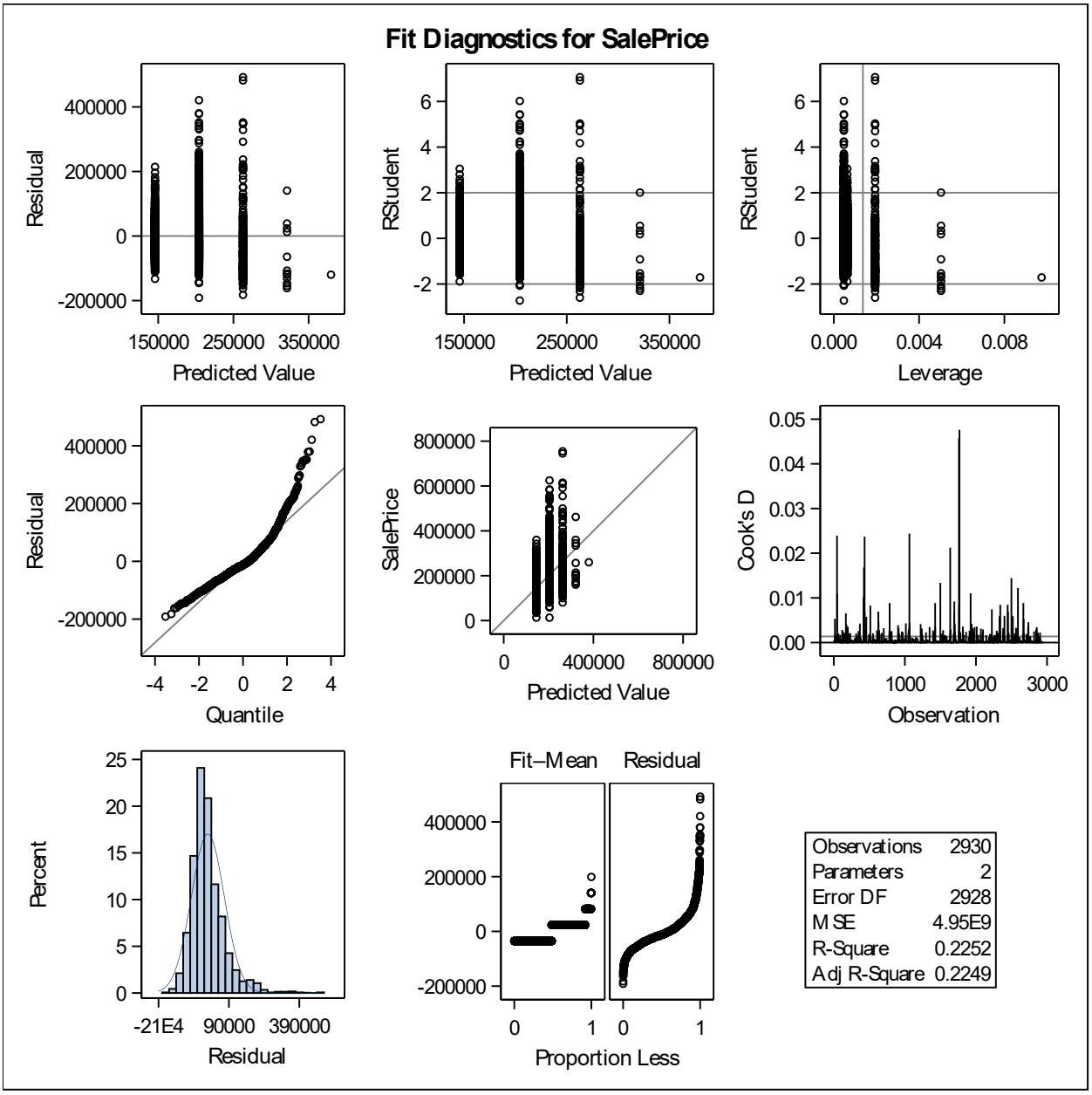
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.20966E12	4.20966E12	851.07	<.0001
Error	2928	1.448288E13	4946337816		
Corrected Total	2929	1.869254E13			

After reviewing the ANOVA table, we can see that the F-value is large since it is greater than one. This means the observations and regression will show contrast from the overall mean.

Root MSE	70330	R-Square	0.2252
Dependent Mean	180796	Adj R-Sq	0.2249
Coeff Var	38.90030		

After reviewing the source table, we see that the R-Square for this regression model shows about 23% of the variability in SalePrice while using Fireplaces. As the model building process continues, we will keep a close eye on the Adj. R-Square percentage to compare our models' performance in relation to the sample size and number of variables that are included in the model.

ODS graphics for Fit Diagnostics – SalePrice



In assessing the normality of the residuals using the QQ Plot, we can see that the plot is heave-tailed. We can conclude that these observations do not follow the assumed distribution. There are two noticeable outliers at the top right corner of the graph that may be cause for an alarm. After observing the histogram of the residuals, the data appears to be normal.

It's clear that the graphs are showing heteroscedasticity, as opposed to homoscedasticity.

Next, we observe Cook's Distance. Cook's Distance (Cook's D.) represents values of "D" that are substantially larger than the rest. Utilizing the calculation for threshold of Cook's D., we'll be able to establish whether or not the spikes in the graph (data points) are influential.

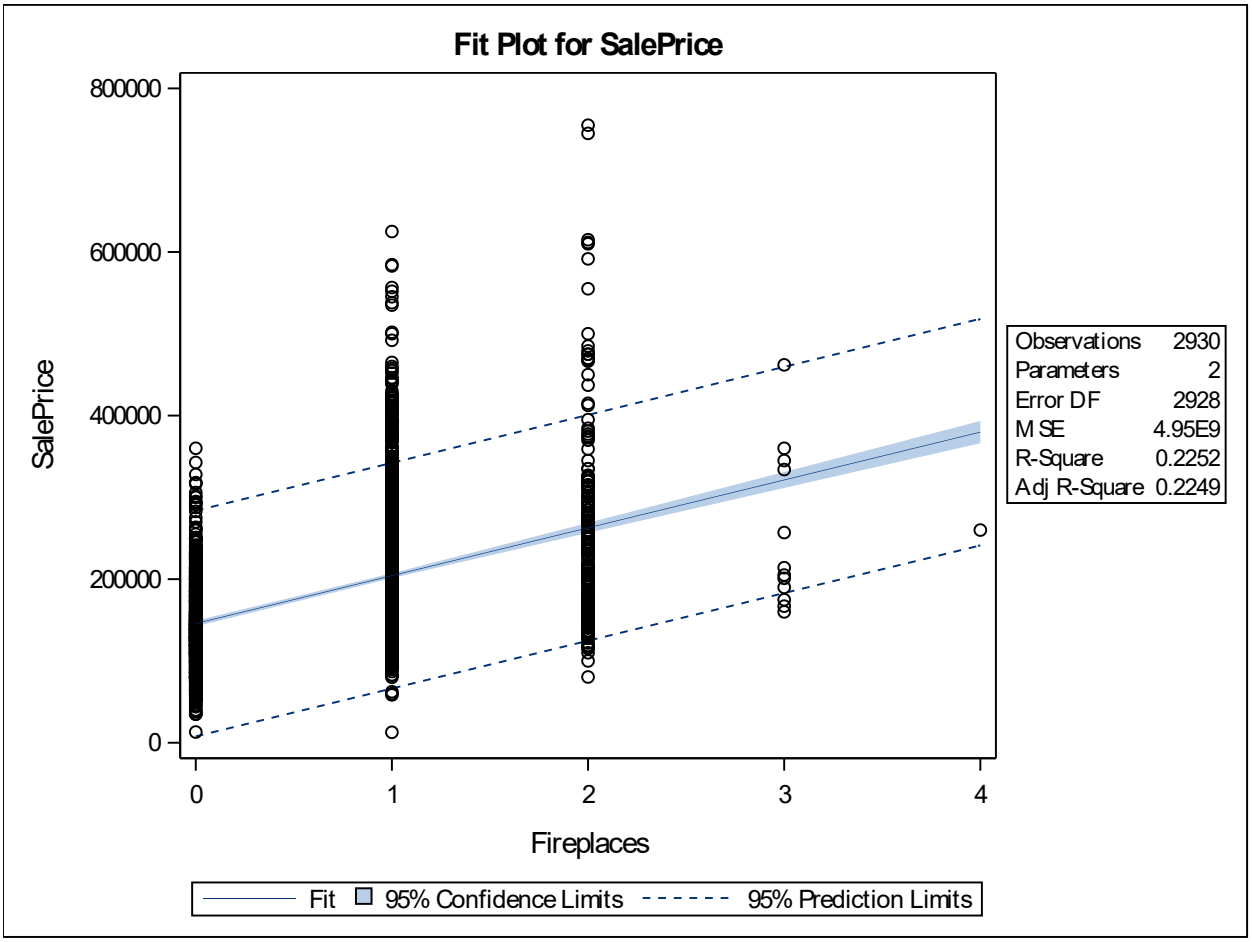
$$\frac{4}{(N-k-1)}$$

N = num. of corrected observations
k = num. of explanatory variables

$\frac{4}{(2929-1-1)}$ = 0.0013

Seeing that there are a couple of spikes over the 0.0013 threshold, this is indicative that there are some observations that *are* influential.

Fit Plot – SalePrice



For the most part, this fit plot is showing a positive, linear trend of heteroscedasticity. In comparison to the previous two continuous variables, the largest amount of variability in observation of

SalePrice comes from when a house has at least two fireplaces. We do make note that there are noticeable outliers at 1 and 2.

Comparing all model Adj. R-Square, this model explains the least amount of variability, 23%; whereas, the GrLivArea continuous variable explains the most variability of all three, 50%.

Simple Linear Model Comparisons:

Model	Adj. R-Square	F Value
SalePrice = 33847 + 126.72825 x FirstFlrSF	0.3863	1844.47
SalePrice = 13290 + 111.69400 x GrLivArea	0.4994	2922.59
SalePrice = 145729 + 58512 x Fireplaces	0.2249	851.07

In reviewing the above criteria about each of the models created, the model that was 'best', based on explanation of variability in SalePrice would be the model utilizing GrLivArea as the explanatory variable.

Multiple Linear Regression Models

FirstFlrSF and GrLivArea to predict SalePrice

The model used:

$$SalePrice = \beta_0 + \beta_1 FirstFlrSF + \beta_2 GrLivArea + \varepsilon$$

Utilize the "REG" procedure to show the estimates of each parameter:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-20542	3374.96975	-6.09	<.0001
FirstFlrSF	1	66.86452	2.97666	22.46	<.0001
GrLivArea	1	82.55353	2.30763	35.77	<.0001

The fitted model:

SalePrice = 66.86452 x FirstFlrSF + 82.55353 x GrLivArea – 20542

In reviewing the above variables, the model coefficients are indicative that if FirstFlrSF were 0 and GrLivArea were 0, then the SalePrice of the house would be $-20,542. It's unlikely that either of these variables could be 0 due to most all houses being above ground, and all houses having at least a first floor. The average change in the mean of SalePrice is about $149.41, and a one unit change in FirstFlrSF and GrLivArea should be consistent since they're continuous variables.

Since all t-values are > 0, they are significantly large; moreover, there p-values also show as significant. Because of this, we are able to reject the null hypothesis and conclude that these variables have slope and intercept that are greater than zero.

Goodness-of-fit information:

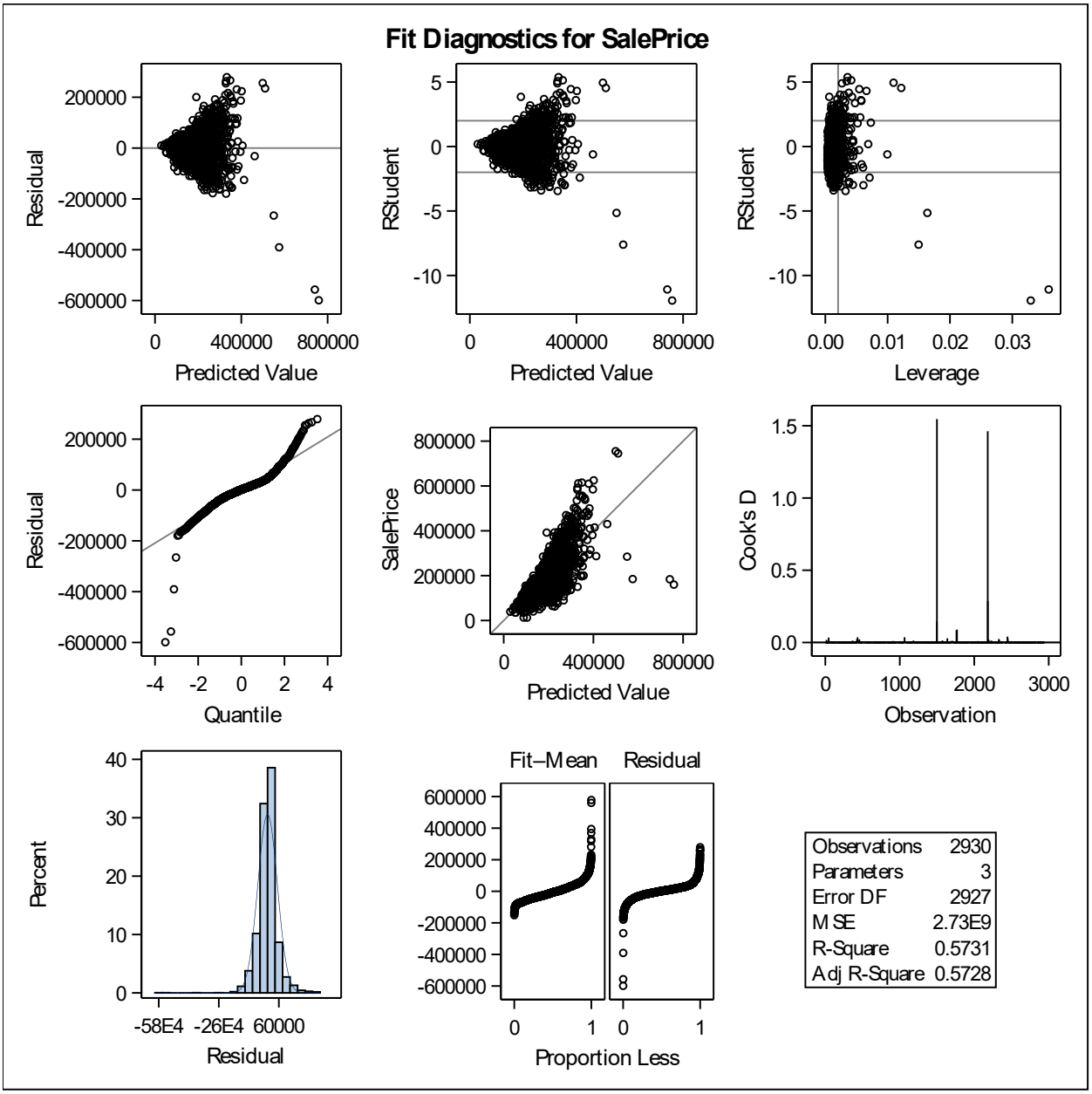
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.071318E13	5.356592E12	1964.91	<.0001
Error	2927	7.979352E12	2726119658		
Corrected Total	2929	1.869254E13			

After reviewing the ANOVA table, we can see that the F-value is large since it is greater than one. This means the observations and regression will show contrast from the overall mean.

Root MSE	52212	R-Square	0.5731
Dependent Mean	180796	Adj R-Sq	0.5728
Coeff Var	28.87909		

After reviewing the source table, we see that the R-Square for this regression model shows about 57% of the variability in SalePrice while using FirstFlrSF and GrLivArea. As the model building process continues, we will keep a close eye on the Adj. R-Square percentage to compare our models' performance in relation to the sample size and number of variables that are included in the model.

ODS graphics for Fit Diagnostics – SalePrice



In assessing the normality of the residuals using the QQ Plot, we can see that the plot is heave-tailed. We can conclude that these observations do not follow the assumed distribution. There are noticeable outliers at the bottom left corner of the graph that may be cause for an alarm. After observing the histogram of the residuals, the data appears to be normal.

It's clear that the graphs are showing heteroscedasticity, as opposed to homoscedasticity.

Next, we observe Cook's Distance. Cook's Distance (Cook's D.) represents values of "D" that are substantially larger than the rest. Utilizing the calculation for threshold of Cook's D., we'll be able to establish whether or not the spikes in the graph (data points) are influential.

$$\frac{4}{(N-k-1)}$$

N = num. of corrected observations
k = num. of explanatory variables

$\frac{4}{(2929-1-1)}$ = 0.0013

Seeing that there are a couple of spikes over the 0.0013 threshold, this is indicative that there are some observations that *are* influential.

Comparing this multiple regression model against the simple regression models as previously created, based only off of the Adj. R-Square value, this models' 57% variability is greater than all the models, and will perform better than the rest.

Finding Worst Continuous Variable

The continuous variable that correlated the worst with SalePrice is BsmtUnfSF that showed a correlation of 0.18286.

Model used:

$$SalePrice = \beta_0 + \beta_1 FirstFlrSF + \beta_2 GrLivArea + \beta_3 BsmtUnfSF + \varepsilon$$

Utilize the Reg procedure:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-19619	3381.46139	-5.80	<.0001
FirstFlrSF	1	68.86756	3.03566	22.69	<.0001
GrLivArea	1	83.18210	2.31336	35.96	<.0001
BsmtUnfSF	1	-7.46924	2.30380	-3.24	0.0012

The fitted model:

SalePrice = 68.86756 x FirstFlrSF + 83.18210 x GrLivArea - 7.46924 x BsmtUnfSF – 19619

In reviewing the above variables, the model coefficients are indicative that if FirstFlrSF were 0, GrLivArea were 0, and BsmtUnfSF were 0, then the SalePrice of the house would be $-19,619. It's likely that BsmtUnfSF could have null values because not all houses will have a basement. The average change in the mean of SalePrice is about $144.58, and a one unit change in FirstFlrSF, GrLivArea, and BsmtUnfSF should be consistent since they're continuous variables. The Intercept and t-value of the intercept has changed a little, as well as the estimates for FirstFlrSF and GrLivArea after we added BsmtUnfSF to the model.

Since all t-values are > 0, they are significantly large. The p-values of both FirstFlrSF and GrLivArea show significance, but the p-value of BsmtUnfSF is much higher. From this, we can conclude

that both FirstFlrSF and GrLivArea have slopes that are greater than zero; however, the null hypothesis must be accepted for BsmtUnfSF—this means the slope is zero.
Goodness-of-fit information:

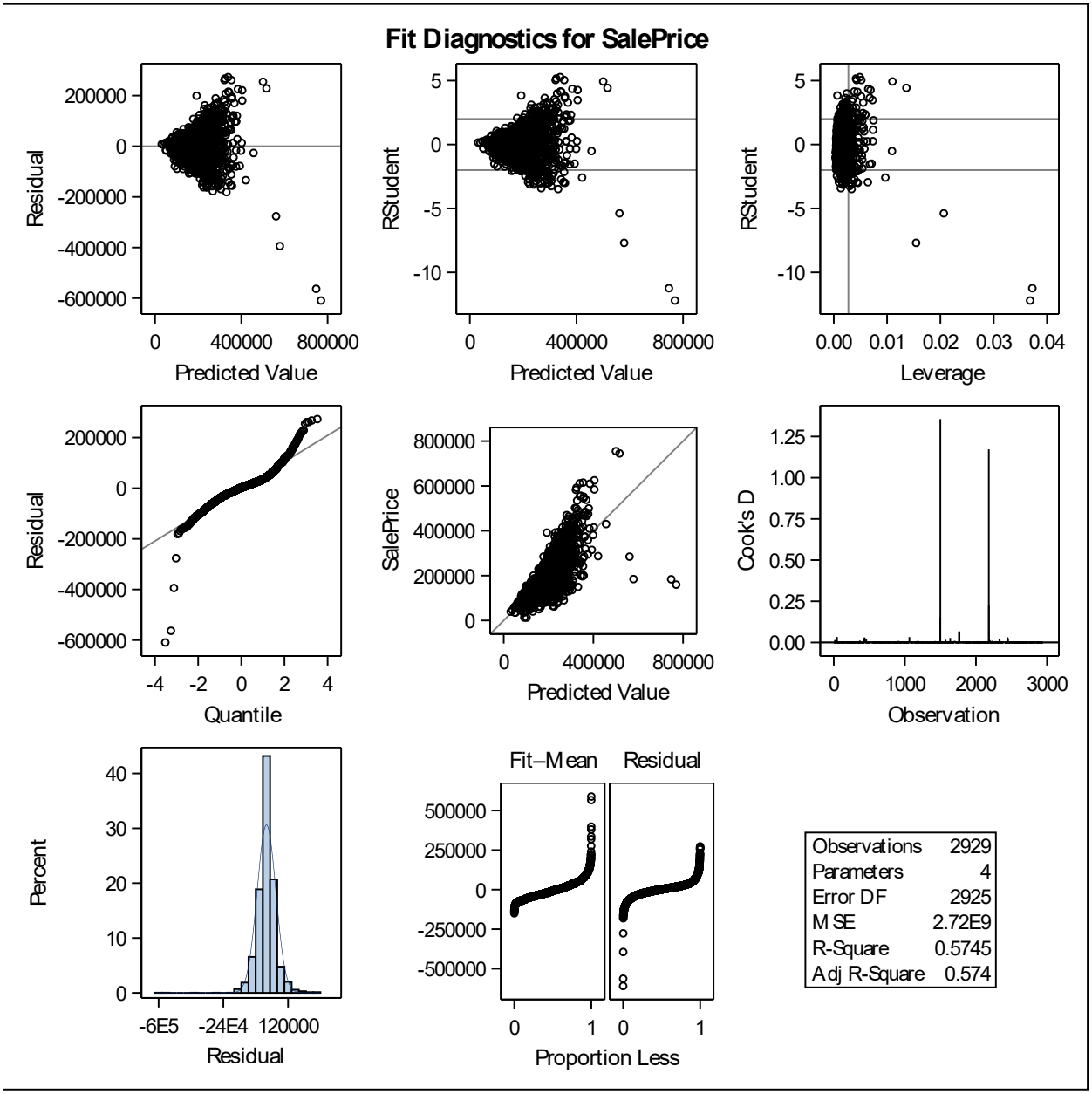
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.073257E13	3.577522E12	1316.32	<.0001
Error	2925	7.949604E12	2717813328		
Corrected Total	2928	1.868217E13			

After reviewing the ANOVA table, we can see that the F-value is large since it is greater than one. This means the observations and regression will show contrast from the overall mean. There are some linear associations with the observations. From Pr > |t| we reject the null hypothesis. There is a linear relationship between FirstFlrSF, GrLivArea, BsmtUnfSF, and SalePrice.

Root MSE	52133	R-Square	0.5745
Dependent Mean	180831	Adj R-Sq	0.5740
Coeff Var	28.82952		

After reviewing the source table, we see that the R-Square for this regression model shows about 57% of the variability in SalePrice while using FirstFlrSF, GrLivArea, and BsmtUnfSF. As the model building process continues, we will keep a close eye on the Adj. R-Square percentage to compare our models' performance in relation to the sample size and number of variables that are included in the model.

ODS graphics for Fit Diagnostics – SalePrice



In assessing the normality of the residuals using the QQ Plot, we can see that the plot is heave-tailed. We can conclude that these observations do not follow the assumed distribution. There are noticeable outliers at the bottom left corner of the graph that may be cause for an alarm. After observing the histogram of the residuals, the data appears to be normal.

It's clear that the graphs are showing heteroscedasticity, as opposed to homoscedasticity.

Next, we observe Cook's Distance. Cook's Distance (Cook's D.) represents values of "D" that are substantially larger than the rest. Utilizing the calculation for threshold of Cook's D., we'll be able to establish whether or not the spikes in the graph (data points) are influential.

$$\frac{4}{(N-k-1)}$$

N = num. of corrected observations
k = num. of explanatory variables

$\frac{4}{(2928-1-1)}$ = 0.0013

Seeing that there are a couple of spikes over the 0.0013 threshold, this is indicative that there are some observations that *are* influential.

Comparing this against the previous model, by adding a poor explanatory variable into our model, we can see that the R-Squared value increased by 0.0014, and the Adj. R-Squared value increased by 0.0012; however, the F-value decreased significantly: -648.59

From what we have gathered throughout the model building process, it seems that if we add more predictor variables, even the worst correlation to the response variable, the R-Squared results are better. This does not indicate a better *fit* because it also seems like you can put any variable into the multivariate model to get a better explanation of variability in the predictor variable.

Conclusions:

Throughout this model building process, we were able to build comparisons between simple linear regression models and multiple linear regression models by utilizing the R-Squared, Adj. R-Squared, and F-Value estimates. Although MasVrnArea was not used during this model building process due to data error, this type of variable is one that should be avoided in the beginning because of its potential of having mass quantities of null values.

The next in the modeling building process would be to present our initial findings to our stakeholders so that we can pinpoint our model building process based off of their feedback before we being validation (Do we use continuous or categorical variables? Do we use simple linear or multiple linear models?). We would show that during our initial assessment, we have found that continuous variables perform the best for explaining variability of SalePrice.

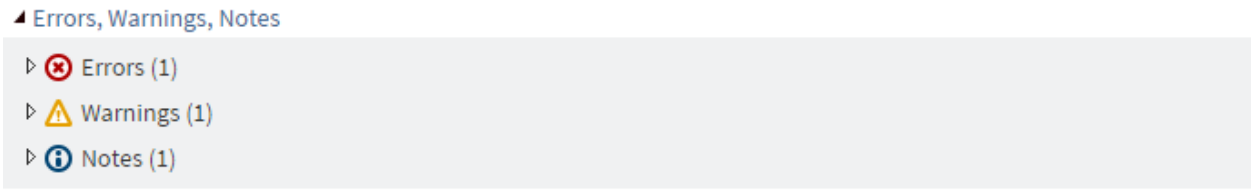
Code:

```
libname mydata '/scs/wtm926/' access=readonly;

proc datasets library=mydata;
run;

Data temp1;
set mydata.ames_housing_data;

proc contents data=temp1;
run;
```

```
proc corr data=temp1;
var saleprice;
with _numeric_;
run;

ods graphics on;

proc reg;
model saleprice = firstflrsf;
run;

proc reg;
model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF BsmtFinSF1 BsmtUnfSF/
selection=rsquare start=1 stop=1;
run;

proc reg;
model saleprice = grlivarea;
run;

proc reg;
model saleprice = fireplaces;
run;

proc means data=temp1;
var FirstFlrSF GrLivArea Fireplaces;
run;

proc reg;
model saleprice = firstflrsf grlivarea;
run;

proc reg;
model saleprice = firstflrsf grlivarea bsmtunfsf;
run;
```

Please note the variable MasVrnArea error that I continue to receive when trying to build the models utilizing this variable:

```
proc reg;
model saleprice = MasVrnArea;
run;
```




```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
55
56         proc reg;
57         model saleprice = MasVrnArea;
ERROR: Variable MASVRNAREA not found.
NOTE: The previous statement has been deleted.
58         run;

WARNING: No variables specified for an SSCP matrix. Execution terminating.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.00 seconds
      cpu time            0.01 seconds


59
60         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
72
```