Scott Jue
MSDS 410
1/29/2023

Modeling Assignment #4: Building Linear Regression Models

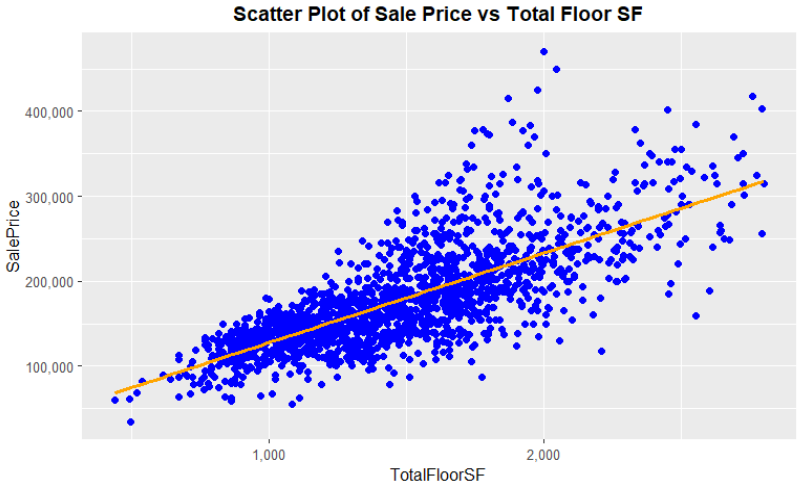
Introduction

For this analysis, we will use the Ames, IA housing dataset to build a simple linear regression model and various multiple linear regression models that predict the sale price of a typical home in Ames, IA between 2006 and 2010 using the continuous explanatory variables provided in the dataset. From previous exploratory data analysis work, we identified the sample population as observations that are in residential zones, of single-family homes type, have normal sale conditions, have a total square foot less than 2,800, have less than or equal to 9 total rooms above ground, have all public utilities, has a basement that is no bigger than 2,000 square feet, and have at least a 1 car garage, at least 1 full bath, at least 1 kitchen above ground, and at least 1 bedroom above ground. These filters and drop conditions have been applied to the dataset so that we can only select the observations that meet the criteria of the sample population of interest. Finally, after removing observations that do not meet the sample population criteria, I looked at the price per square feet variable. There is 1 extreme outlier that has been identified and removed from the data set. This observation is greater than 3 times the interquartile range and likely has some other unique features that increased the value of the sale price per square feet such that it does not resemble other observations in the sample population. This results in a sample population of 1,782 observations. Each of the regression models built will be evaluated for goodness of fit and the underlying assumptions will be assessed.

Results

1) For Model 1 I have selected TotalFloorSF as the explanatory variable because it shows a strong linear relationship to SalePrice as it has the highest Pearson Correlation coefficient out of the continuous explanatory variables with 0.7554.

Scatterplot of SalePrice vs TotalFloorSF



Model 1:

Linear Equation:

Y_hat = 22029.913 + 105.562*TotalFloorSF

The intercept can be interpreted as if TotalFloorSF is 0 then the sale price of a typical home, as defined by the drop conditions, in Ames, IA between 2006 and 2010 will have a predicted price of $22,029.91. However, the intercept value cannot be interpreted outside of this dataset since the value is below the minimum value of the dataset and therefore outside the range of the dataset. Additionally, is also not feasible for a home's total floor square feet to be 0 when interpreting home sale price. If there is 0 square feet, then the sale would be for a lot area and not considered a "typical home". In other words, outside of this modeling this dataset the SalePrice intercept cannot be interpreted. The coefficient for TotalFloorSF can be interpreted as for every 1 square foot increase in TotalFloorSF, the predicted sale price increases by $105.56.

The R-squared value for Model 1 is 0.5706. This means that the explanatory variable TotalFloorSF accounts for approximately 57% of the variability in the response variable SalePrice.

Model 1 Coefficient Summary Table:

```
Call:
lm(formula = SalePrice ~ TotalFloorSF, data = model_df)

Residuals:
    Min      1Q  Median      3Q     Max
-137823  -22473   -2012   18143  236846

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  22029.913   3302.368   6.671      0.0000000000338 ***
TotalFloorSF   105.562      2.171  48.632 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38990 on 1780 degrees of freedom
Multiple R-squared:  0.5706,    Adjusted R-squared:  0.5703
F-statistic:  2365 on 1 and 1780 DF,  p-value: < 0.00000000000000022
```

Model 1 ANOVA Table:

```
Analysis of Variance Table

Response: SalePrice
              Df        Sum Sq       Mean Sq F value                Pr(>F)
TotalFloorSF    1 3595460685165 3595460685165    2365 < 0.00000000000000022 ***
Residuals    1780 2706045576724    1520250324
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Tests:

For the below hypothesis tests of the individual model coefficients, we will use a critical t-value of 1.9613 which is the value associated with a two tailed test at the 0.05 significance level and with degrees of freedom of 1,780.

Intercept:

Null hypothesis → H_0: $\beta_0 = 0$
Alternative hypotheses → H_A: $\beta_0 \neq 0$

t = 6.671

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_0 = 0$ and conclude that $\beta_0 \neq 0$. The intercept value provides significant useful information for predicting SalePrice.

TotalFloorSF:

Null hypothesis → H_0: $\beta_1 = 0$
Alternative hypotheses → H_A: $\beta_1 \neq 0$

t = 48.632

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_1 = 0$ and conclude that $\beta_1 \neq 0$. We can interpret this result as having the explanatory variable, TotalFloorSF, included in the model provides significant information for predicting the response variable (Y).

Omnibus Overall F-test:

Next, we will perform the Omnibus Overall F-test. For this hypothesis test of the overall model, we will use a critical F-value of 3.8467 which is the value associated with degrees of freedom of 1 and 1780, at the 0.05 significance level.

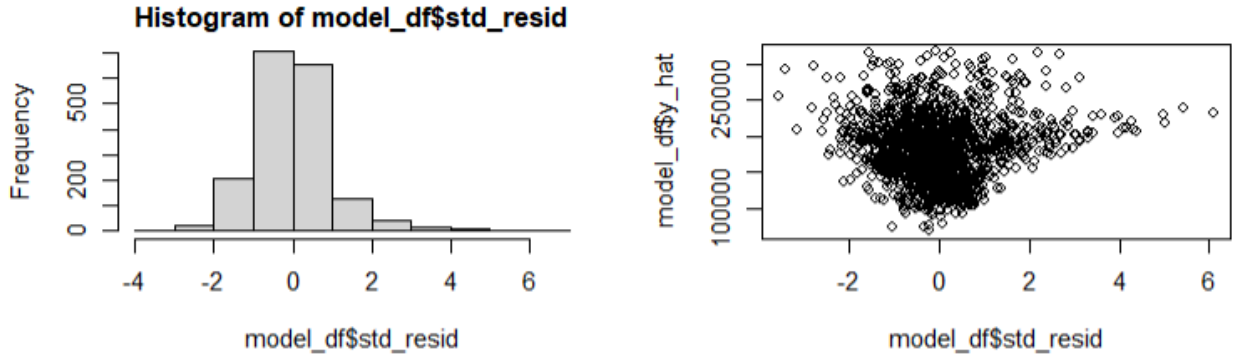
Null hypothesis → H_0: $\beta_i = 0$
Alternative hypotheses → H_A: at least one $\beta_i \neq 0$

From the Coefficient Summary Table, we can see that the F-statistic of 2365 is greater than the critical F-value of 2.2199. Therefore, we should reject the null hypothesis and conclude that at least one $\beta_i \neq 0$. This indicates that there is a significant relationship between the intercept and independent variable and the response variable.

Hypothesis Test Underlying Assumptions:

We can see in the below histogram that the distribution of the standardized residuals is slightly right-skewed. Additionally, the scatterplot shows that there is heteroscedasticity in the residuals as there the variance of the residuals increases as the predicted sales price increases. Therefore, the underlying assumptions have been violated and the hypothesis tests do not provide much value.

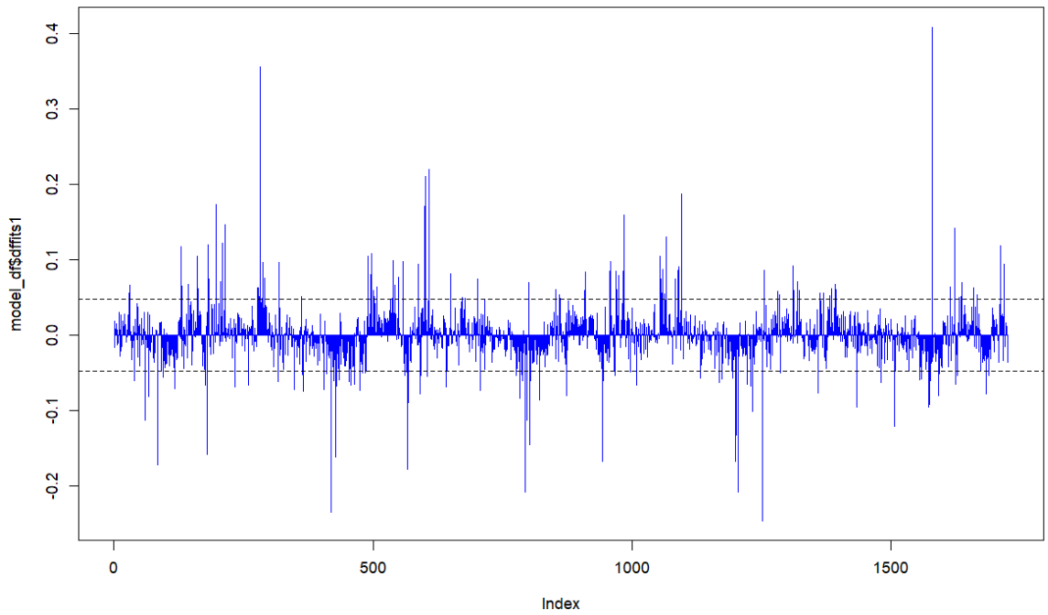
Histogram and Scatterplot of Standardized Residuals:



Influential Observations:

The below DFFITS plot shows that there are 201 observations that are greater than the DFFITS upper and lower limits indicating that these are influential observations. We should be concerned that these observations may pull the regression model too far in one direction, thereby decreasing the overall fit of the model. We could potentially remove these observations to increase the model performance and fit. However, these observations could also be valid representations of the sample population. So, further analysis should be done on these influential observations before removing them and refitting the model.

Model 1 DFFITS Plot:



2) For Model 2, we will select TotalFloorSF and OverallQual as the explanatory variables for a multiple linear regression model to predict SalePrice.

Model 2:

Linear Equation:

Y_hat = -69506.369 + 60.855* TotalFloorSF + 26103.013* OverallQual

The intercept for Model 2 can be interpreted as if TotalFloorSF is 0 and if OverallQual is 0, then the predicted sale price would be -$69,506.37. This value is only meaningful within this model using this specific dataset. It is outside the range of the SalePrice values in the dataset. Furthermore, the value is not realistic as that would imply a negative selling price if there is 0 TotalFloorSF and 0 OverallQual. There is some land value at the very least, so a negative intercept value is not feasible outside of this dataset. In other words, outside of this modeling this dataset the SalePrice intercept cannot be interpreted. The TotalFloorSF coefficient can be interpreted as for every 1 square foot increase in TotalFloorSF, the predict sale price increase by $60.86. The OverallQual coefficient can be interpreted as for every 1 unit increase in OverallQual, the predict sale price increase by $26,103.01. However, the OverallQual rating is ordinal so the max value is 10 and the minimum value is 1. So, this variable can only add a maximum of $261,030 a minimum of $26,103.01 to the predicted sale price. This interpretation is different than Model 1 because the intercept is negative and we have an ordinal variable so we know the minimum and maximum values that the OverallQual variable can add to the predicted sale price.

The R-squared value for Model 2 is 0.7488. This means that the explanatory variables TotalFloorSF and OverallQual account for approximately 75% of the variability in the response variable SalePrice. This is an increase of 0.1782 from Model 1, meaning that adding OverallQual to the model helped to explain an additional 17.82% of the variability in the response variable SalePrice.

Model 2 Coefficient Summary Table:

```
Call:
lm(formula = SalePrice ~ TotalFloorSF + OverallQual, data = model_df)

Residuals:
    Min      1Q  Median      3Q     Max
-109139  -18538    -290   16402  182869

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)  -69506.369   3608.371  -19.26 <0.0000000000000002 ***
TotalFloorSF     60.855      2.083   29.21 <0.0000000000000002 ***
OverallQual   26103.013    734.688   35.53 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29830 on 1779 degrees of freedom
Multiple R-squared:  0.7488,    Adjusted R-squared:  0.7485
F-statistic:  2652 on 2 and 1779 DF,  p-value: < 0.00000000000000022
```

Model 2 ANOVA Table:

```
Analysis of Variance Table

Response: SalePrice
              Df        Sum Sq       Mean Sq F value                Pr(>F)
TotalFloorSF    1 3595460685165 3595460685165  4041.0 < 0.00000000000000022 ***
OverallQual     1 1123170548085 1123170548085  1262.3 < 0.00000000000000022 ***
Residuals    1779 1582875028640     889755497
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hypothesis Tests:

For the below hypothesis tests of the individual model coefficients, we will use a critical t-value of 1.9613 which is the value associated with a two tailed test at the 0.05 significance level and with degrees of freedom of 1,779.

Intercept:

Null hypothesis → H_0: $\beta_0 = 0$
Alternative hypotheses → H_A: $\beta_0 \neq 0$

t = -19.26

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_0 = 0$ and conclude that $\beta_0 \neq 0$. The intercept value provides significant useful information for predicting SalePrice.

TotalFloorSF:

Null hypothesis → H_0: $\beta_1 = 0$
Alternative hypotheses → H_A: $\beta_1 \neq 0$

t = 29.21

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_1 = 0$ and conclude that $\beta_1 \neq 0$. We can interpret this result as having the explanatory variable, TotalFloorSF, included in the model provides significant information for predicting the response variable (Y).

OverallQual:

Null hypothesis → H_0: $\beta_2 = 0$
Alternative hypotheses → H_A: $\beta_2 \neq 0$

t = 35.53

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_2 = 0$ and conclude that $\beta_2 \neq 0$. We can interpret this result as having the explanatory variable, OverallQual, included in the model provides significant information for predicting the response variable (Y).

Omnibus Overall F-test:

Next, we will perform the Omnibus Overall F-test. For this hypothesis test of the overall model, we will use a critical F-value of 3.0008 which is the value associated with degrees of freedom of 2 and 1779, at the 0.05 significance level.

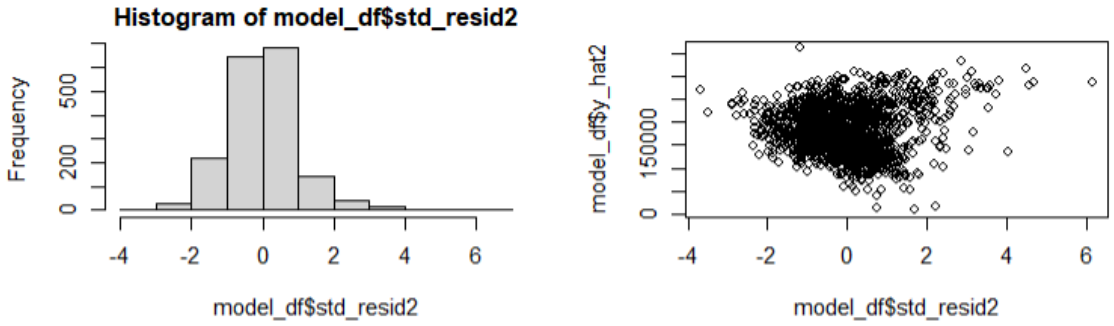
Null hypothesis → H_0: $\beta_1 = \beta_2 = 0$
Alternative hypotheses → H_A: at least one $\beta_i \neq 0$

From the Coefficient Summary Table, we can see that the F-statistic of 2652 is greater than the critical F-value of 3.0008. Therefore, we should reject the null hypothesis and conclude that at least one $\beta_i \neq 0$. This indicates that there is a significant relationship between the independent variables and the response variable.

Hypothesis Test Underlying Assumptions:

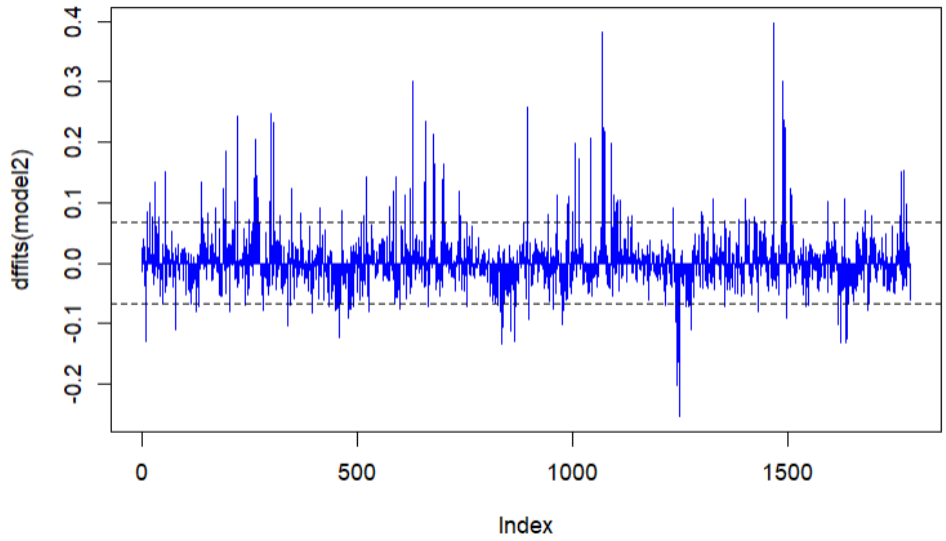
We can see in the below histogram that the distribution of the standardized residuals is slightly right-skewed. Additionally, the scatterplot shows that there is heteroscedasticity in the residuals as there the variance of the residuals increases as the predicted sales price increases. Therefore, the underlying assumptions have been violated and the hypothesis tests do not provide much value.

Histogram and Scatterplot of Standardized Residuals:



Influential Observations:

The below DFFITS plot shows that there are 171 observations that are greater than the DFFITS upper and lower limits indicating that these are influential observations. We should be concerned that these observations may pull the regression model too far in one direction, thereby decreasing the overall fitness of the model. We could potentially remove these observations to increase the model performance and fit. However, these observations could also be valid representations of the sample population. So, further analysis should be done on these influential observations before removing them and refitting the model.



Model Comparison:

We should retain both variables as predict variables of SalePrice since the OverallQual variable provides significant information that is useful to predicting SalePrice. The addition of OverallQual increases the R-squared value of the model by .1782 meaning that OverallQual accounts for approximately 18% of the variability in the SalePrice when added to the regression model with TotalFloorSF.

3) For Model 3, we will add GarageArea to the multiple regression model that was created for Model 2. I selected GarageArea because it had the next highest Pearson Correlation coefficient of 0.6228. Additionally, GarageArea accounts for additional usable house space that is not captured in the

TotalFloorSF variable. So, if we are assuming a linear relationship between house size and sales price, then adding GarageArea should help in predicting SalesPrice.

Model 3:

Linear Equation:
Y_hat = -71201.392 + 54.811*TotalFloorSF + 21472.912*OverallQual + 80.338*GarageArea

The intercept for Model 3 can be interpreted as if the explanatory variables, TotalFloorSF. OverallQual, and GarageArea, are equal to 0, then the predicted sale price would be -$71,201.39. This value is only meaningful within this model using this specific dataset. It is outside the range of the SalePrice values in the dataset. Furthermore, the value is not realistic as that would imply a negative selling price if TotalFloorSF, OverallQual, and GarageArea are all 0. There is some land value at the very least, so a negative intercept value is not feasible outside of this dataset. In other words, outside of this modeling this dataset the SalePrice intercept cannot be interpreted. The TotalFloorSF coefficient can be interpreted as for every 1 square foot increase in TotalFloorSF, the predict sale price increase by $54.81. The OverallQual coefficient can be interpreted as for every 1 unit increase in OverallQual, the predict sale price increase by $21,472.91. However, the OverallQual rating is ordinal so the max value is 10 and the minimum value is 1. So, this variable can only add a maximum of $214,729.12 a minimum of $21,472.91 to the predicted sale price. The coefficient for GarageArea can be interpreted as for every 1 square foot increase in GarageArea, the predicted SalePrice increases by $80.33. This interpretation of these coefficients is different than Model 1 and 2. Because of the addition of GarageArea to the model, the coefficients for TotalFloorSF and OverallQual decreased slightly to account for the additional dollars added from the GarageArea parameter.

The R-squared value for Model 3 is 0.7897. This means that the explanatory variables TotalFloorSF, OverallQual, and GarageArea account for approximately 79% of the variability in the response variable SalePrice. This is an increase of 0.04 from Model 2, meaning that adding GarageArea to the model helped to explain an additional 4% of the variability in the response variable SalePrice. Therefore, adding GarageArea as an explanatory variable helps improve the model's explanatory ability.

Model 3 Coefficient Summary Table:

```
Call:
lm(formula = SalePrice ~ TotalFloorSF + OverallQual + GarageArea,
    data = model_df)

Residuals:
    Min      1Q  Median      3Q     Max
-116155  -16586   -1804   15345  180318

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)  -71201.392   3303.475  -21.55 <0.0000000000000002 ***
TotalFloorSF     54.811      1.934   28.34 <0.0000000000000002 ***
OverallQual   21472.912    716.931   29.95 <0.0000000000000002 ***
GarageArea       80.338      4.318   18.61 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27300 on 1778 degrees of freedom
Multiple R-squared:  0.7897,    Adjusted R-squared:  0.7894
F-statistic:  2226 on 3 and 1778 DF,  p-value: < 0.00000000000000022
```

Model 3 ANOVA Table:

```
Analysis of Variance Table

Response: SalePrice
              Df        Sum Sq       Mean Sq F value                Pr(>F)
TotalFloorSF    1 3595460685165 3595460685165 4824.97 < 0.00000000000000022 ***
OverallQual     1 1123170548085 1123170548085 1507.25 < 0.00000000000000022 ***
GarageArea      1  257948420530  257948420530  346.16 < 0.00000000000000022 ***
Residuals    1778 1324926608110     745178070
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Tests:

For the below hypothesis tests of the individual model coefficients, we will use a critical t-value of 1.9613 which is the value associated with a two tailed test at the 0.05 significance level and with degrees of freedom of 1,778.

Intercept:

Null hypothesis → H_0: $\beta_0 = 0$
Alternative hypotheses → H_A: $\beta_0 \neq 0$

t = -21.55

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_0 = 0$ and conclude that $\beta_0 \neq 0$. The intercept value provides significant useful information for predicting SalePrice.

TotalFloorSF:

Null hypothesis → H_0: $\beta_1 = 0$
Alternative hypotheses → H_A: $\beta_1 \neq 0$

t = 28.34

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_1 = 0$ and conclude that $\beta_1 \neq 0$. We can interpret this result as having the explanatory variable, TotalFloorSF, included in the model provides significant information for predicting the response variable (Y).

OverallQual:

Null hypothesis → H_0: $\beta_2 = 0$
Alternative hypotheses → H_A: $\beta_2 \neq 0$

t = 29.95

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_2 = 0$ and conclude that $\beta_2 \neq$

0. We can interpret this result as having the explanatory variable, OverallQual, included in the model provides significant information for predicting the response variable (Y).

GarageArea:

Null hypothesis → H_0: $\beta_3 = 0$
Alternative hypotheses → H_A: $\beta_3 \neq 0$

t = 18.61

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_3 = 0$ and conclude that $\beta_3 \neq 0$. We can interpret this result as having the explanatory variable, GarageArea, included in the model provides significant information for predicting the response variable (Y).

Omnibus Overall F-test:

Next, we will perform the Omnibus Overall F-test. For this hypothesis test of the overall model, we will use a critical F-value of 2.6099 which is the value associated with degrees of freedom of 3 and 1778, at the 0.05 significance level.

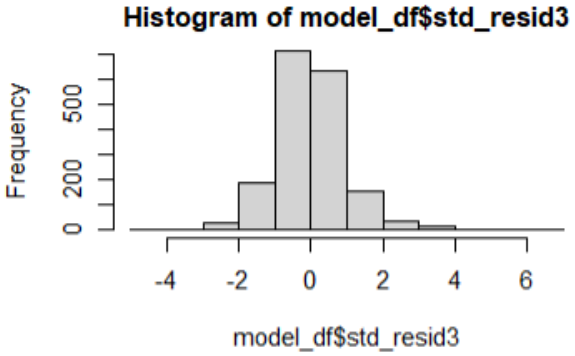
Null hypothesis → H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
Alternative hypotheses → H_A: at least one $\beta_i \neq 0$

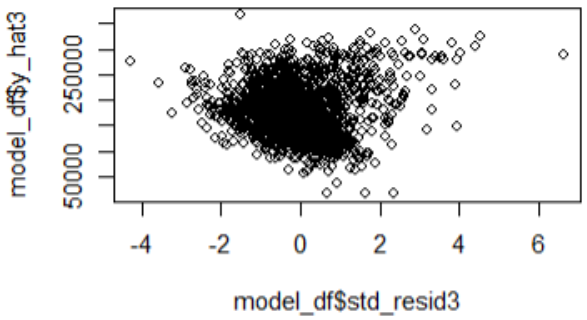
From the Coefficient Summary Table, we can see that the F-statistic of 2226 is greater than the critical F-value of 2.6099. Therefore, we should reject the null hypothesis and conclude that at least one $\beta_i \neq 0$. This indicates that there is a significant relationship between the independent variables and the response variable.

Hypothesis Test Underlying Assumptions:

We can see in the below histogram that the distribution of the standardized residuals is slightly right-skewed. Additionally, the scatterplot shows that there is heteroscedasticity in the residuals as there the variance of the residuals increases as the predicted sales price increases. Therefore, the underlying assumptions have been violated and the hypothesis tests do not provide much value.

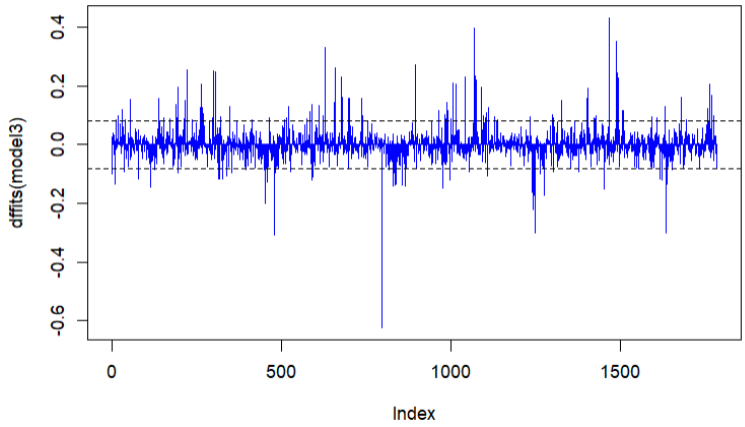
Histogram and Scatterplot of Standardized Residuals:





Influential Observations:

The below DFFITS plot shows that there are 165 observations that are greater than the DFFITS upper and lower limits indicating that these are influential observations. We should be concerned that these observations may pull the regression model too far in one direction, thereby decreasing the overall fitness of the model. We could potentially remove these observations to increase the model performance and fit. However, these observations could also be valid representations of the sample population. So, further analysis should be done on these influential observations before removing them and refitting the model.



Model Comparison:

Based on the above information, we should retain all three variables as predictor variables as the addition of GarageArea increases the R-squared value by .04 and provides significant information that is useful to predicting SalePrice. The addition of GarageArea accounts for approximately 4% of the additional variability in SalePrice when added to the multiple regression model with TotalFloorSF and OverallQual. Including all three variables results in an R-squared value of approximately .79 which is the highest of all three models.

4) For Model 4, we will use the same explanatory variables as Model 3 (TotalFloorSF, OverallQual, and GarageArea; however, we will transform the response variable SalePrice using a log base transformation. Initial exploratory analysis shows that the distribution of SalePrice was right-skewed, so performing a log transformation provides a more normal distribution shape of the response variable.

Model 4 Coefficient Summary Table:

```
Call:
lm(formula = logSalePrice ~ TotalFloorSF + OverallQual + GarageArea,
    data = model_df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69128 -0.07656  0.00572  0.09776  0.57907

Coefficients:
               Estimate  Std. Error t value            Pr(>|t|)
(Intercept)  10.67989850  0.01776688  601.11 <0.0000000000000002 ***
TotalFloorSF  0.00031010  0.00001040   29.81 <0.0000000000000002 ***
OverallQual   0.11754934  0.00385583   30.49 <0.0000000000000002 ***
GarageArea    0.00039333  0.00002322   16.94 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1468 on 1778 degrees of freedom
Multiple R-squared:  0.7933,    Adjusted R-squared:  0.7929
F-statistic:  2274 on 3 and 1778 DF,  p-value: < 0.00000000000000022
```

Model 4 ANOVA Table:

```
Analysis of Variance Table

Response: logSalePrice
              Df  Sum Sq Mean Sq F value               Pr(>F)
TotalFloorSF   1 108.482 108.482 5032.90 < 0.00000000000000022 ***
OverallQual    1  32.410  32.410 1503.60 < 0.00000000000000022 ***
GarageArea     1   6.183   6.183  286.86 < 0.00000000000000022 ***
Residuals   1778  38.324   0.022
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hypothesis Tests:

For the below hypothesis tests of the individual model coefficients, we will use a critical t-value of 1.9613 which is the value associated with a two tailed test at the 0.05 significance level and with degrees of freedom of 1,778.

Intercept:

Null hypothesis → H_0: $\beta_0 = 0$
Alternative hypotheses → H_A: $\beta_0 \neq 0$

t = 601.11

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_0 = 0$ and conclude that $\beta_0 \neq 0$. The intercept value provides significant useful information for predicting SalePrice.

TotalFloorSF:

Null hypothesis → H_0: $\beta_1 = 0$
Alternative hypotheses → H_A: $\beta_1 \neq 0$

t = 29.81

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_1 = 0$ and conclude that $\beta_1 \neq 0$. We can interpret this result as having the explanatory variable, TotalFloorSF, included in the model provides significant information for predicting the response variable (Y).

OverallQual:

Null hypothesis → H_0: $\beta_2 = 0$
Alternative hypotheses → H_A: $\beta_2 \neq 0$

t = 30.49

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_2 = 0$ and conclude that $\beta_2 \neq 0$. We can interpret this result as having the explanatory variable, OverallQual, included in the model provides significant information for predicting the response variable (Y).

GarageArea:

Null hypothesis → H_0: $\beta_3 = 0$
Alternative hypotheses → H_A: $\beta_3 \neq 0$

t = 16.94

From the Coefficient Summary, we can see that the absolute value of t is greater than the critical t-value of 1.9613. Therefore, we can reject the null hypothesis that $\beta_3 = 0$ and conclude that $\beta_3 \neq 0$. We can interpret this result as having the explanatory variable, GarageArea, included in the model provides significant information for predicting the response variable (Y).

Omnibus Overall F-test:

Next, we will perform the Omnibus Overall F-test. For this hypothesis test of the overall model, we will use a critical F-value of 2.6099 which is the value associated with degrees of freedom of 3 and 1778, at the 0.05 significance level.

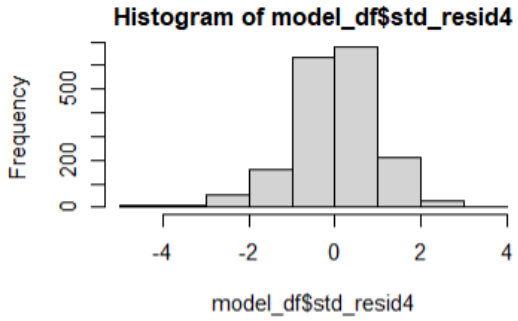
Null hypothesis → H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
Alternative hypotheses → H_A: at least one $\beta_i \neq 0$

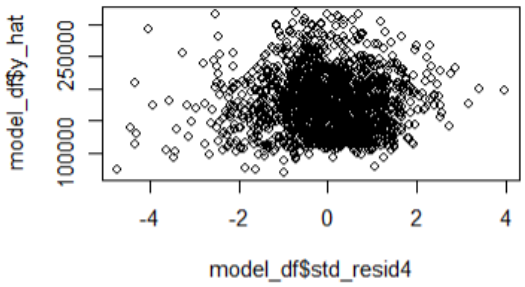
From the Coefficient Summary Table, we can see that the F-statistic of 2774 is greater than the critical F-value of 2.6099. Therefore, we should reject the null hypothesis and conclude that at least one $\beta_i \neq 0$. This indicates that there is a significant relationship between the independent variables and the response variable.

Hypothesis Test Underlying Assumptions:

We can see in the below histogram that the shows more of a normal distribution than the prior models. More importantly, the scatterplot shows that there is homoscedasticity in the residuals as there the variance of the residuals is more equally spread across the predicted SalePrice value. Therefore, the underlying assumptions have been met and the hypothesis tests performed on Model 4 can provide value in our assessment of the model.

Histogram and Scatterplot of Standardized Residuals:





Model Comparison:

Model	R-Squared	Adjusted R-Squared
3	0.7897445	0.7893897
4	0.7932886	0.7929399

Compared to Model 3, using the Natural Log of SalePrice as the response variable provides a better fit to the data. We can see the R-squared and Adjusted R-Squared of Model 4 is higher than Model 3. Using a transformed response variable may cause confusion for a non-technical audience and thus, making it difficult to interpret the model. It is more difficult to understand what a 1 square foot change in TotalFloorSF does to sales price when using the Natural Log of SalePrice instead of just SalePrice. Therefore, it is important to only use transformations when justified and necessary. In terms of R-squared value, the improvement is marginal and I would not use log transformation to obtain this increase in R-Squared with the trade-off of a less interpretable model. However, for the underlying assumptions of hypothesis testing to be met, the SalePrice variable should be transformed using the Natural Log transformation so that we can obtain residuals that display a homoscedasticity pattern. So, for this reason, there is justification for use of a transformed model in addition to improved model fit.

5) Next, we will refit Model 4 after removing the influential observations and compare the results. Using DFFITS to identify the influential observations, 147 observations were removed from the dataset. The refitted model produced improved R-Squared and Adjusted R-Squared values as can be seen in the table below. Both values increase by approximately 0.05.

Model	Observations Removed	R-Squared	Adjusted R-Squared
4	0	0.7932886	0.7929399
4_dffits	147	0.8442296	0.8439431

With influential observations, we should be concerned that these observations could be outliers that pull the regression model too far in one direction, thereby decreasing the overall fitness of the model. By removing these observations, we can increase the model performance and fit. However, these observations could also be valid outliers that still represent the sample population and removing them we could introduce bias. However, some of these influential observations could be outliers because they are not representative of the sample population. For example, HouseAge is negatively correlated with SalesPrice; however, a few of the oldest houses in the dataset are associated with higher sales prices. So, there could be an element of historical significance of these houses that do not qualify them as a “typical” home. So, in this case, they could potentially be removed since they do not represent the desired sample population of “typical” homes. Therefore, further analysis should be done on each of these influential observations before removing them and refitting the model, so that we do not remove legitimate data points. But I do not think it is justifiable to blindly remove influential observations for the sake of improved fit of the model, because we would introduce modeler bias.

6) For Model 5, I will use the explanatory variables TotalFloorSf, OverallQual, GarageArea, TotalBsmtSF, LotArea, and HouseAge to predict the response variable SalePrice. The first 3 variables are the same as Model 3 which showed to explain approximately 79% of the variability in SalePrice, so I thought that would be a good model to build off of. Adding TotalBsmtSF and LotArea should help capture more of the overall size of the house and property. Both also show a linear relationship to SalePrice. Additionally, HouseAge shows a negative linear relationship to SalePrice. Logically, this makes sense as newer house are generally valued higher due to having less wear and tear and also may have new features or materials.

Model 5:

Linear Equation:

Y_hat = -48836.21 + 52.9681*TotalFloorSF + 15803.3495*OverallQual + 41.9937*GarageArea + 35.0784*TotalBsmtSF + 0.8671*LotArea - 328.7413*HouseAge,

The intercept for Model 5 can be interpreted as if all the explanatory variables are equal to 0, then the predicted sale price would be -$48,836.21. This value is only meaningful within this model using this specific dataset. It is outside the range of the SalePrice values in the dataset. Furthermore, the value is not realistic as that would imply a negative selling price if the explanatory variables are all equal 0. Again, a negative intercept value is not feasible outside of modeling this dataset, so the SalePrice intercept cannot be interpreted. The TotalFloorSF coefficient can be interpreted as for every 1 square foot increase in TotalFloorSF, the predict sale price increase by $52.97. The OverallQual coefficient can be interpreted as for every 1 unit increase in OverallQual, the predict sale price increase by $15,803.35. The GarageArea coefficient can be interpreted as for every 1 square foot increase in GarageArea, the predict sale price increase by $41.99. The TotalBsmtSF coefficient can be interpreted as for every 1 square foot increase in TotalBsmtSF the predict sale price increase by $35.08. The LotArea coefficient can be interpreted as for every 1 square foot increase in LotArea, the predict sale price increase by $0.87. The HouseAge coefficient can be interpreted as for every additional year increase in HouseAge, the predict sale price decreases by $328.74.

Model 5 Coefficient Summary Table:

```
Call:
lm(formula = SalePrice ~ TotalFloorSF + OverallQual + GarageArea +
    TotalBsmtSF + LotArea + HouseAge, data = model_df)

Residuals:
   Min     1Q Median     3Q    Max
-95279 -14428  -1044  11849 157846

Coefficients:
               Estimate  Std. Error t value            Pr(>|t|)
(Intercept)  -48836.2117   4093.8443 -11.929 <0.0000000000000002 ***
TotalFloorSF     52.9681      1.6441  32.217 <0.0000000000000002 ***
OverallQual   15803.3495    660.0877  23.941 <0.0000000000000002 ***
GarageArea       41.9937      3.8673  10.859 <0.0000000000000002 ***
TotalBsmtSF      35.0784      1.9568  17.927 <0.0000000000000002 ***
LotArea           0.8671      0.0928   9.343 <0.0000000000000002 ***
HouseAge       -328.7413     23.8381 -13.791 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22760 on 1775 degrees of freedom
Multiple R-squared:  0.8541,    Adjusted R-squared:  0.8536
F-statistic:  1732 on 6 and 1775 DF,  p-value: < 0.00000000000000022
```

Model 5 ANOVA Table:

```
Analysis of Variance Table

Response: SalePrice
              Df        Sum Sq       Mean Sq  F value                Pr(>F)
TotalFloorSF    1 3595460685165 3595460685165 6940.775 < 0.00000000000000022 ***
OverallQual     1 1123170548085 1123170548085 2168.199 < 0.00000000000000022 ***
GarageArea      1  257948420530  257948420530  497.951 < 0.00000000000000022 ***
TotalBsmtSF     1  260324498651  260324498651  502.538 < 0.00000000000000022 ***
LotArea         1   46599574813   46599574813   89.957 < 0.00000000000000022 ***
HouseAge        1   98516898919   98516898919  190.180 < 0.00000000000000022 ***
Residuals    1775  919485635726     518020076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

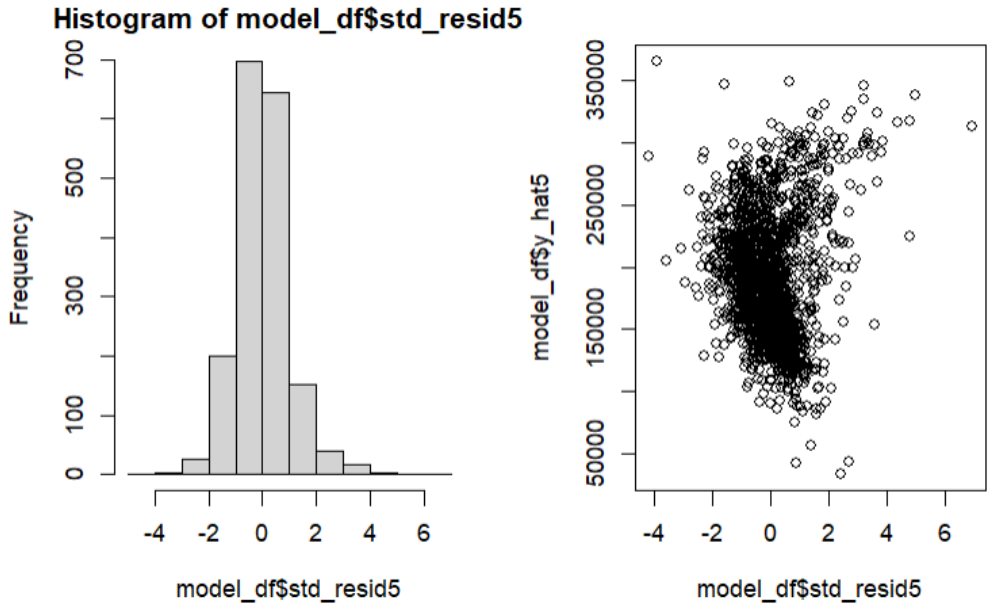
The R-squared value for Model 5 is 0.8541. This means that the explanatory variables account for approximately 85% of the variability in the response variable SalePrice. This is an increase of 0.0644 from Model 3 which is nested in Model 5, meaning that adding TotalBsmtSF, LotArea, and HouseAge to

the model helped to explain an additional 6.4% of the variability in the response variable SalePrice. Therefore, adding these explanatory variable helps improve the model's explanatory ability.

Hypothesis Test and Underlying Assumptions:

From the Coefficient Summary Table, we can see that the absolute values of the t-values for each coefficient is greater than the critical t-value of 1.9613, so we can reject the null hypothesis for each coefficient that they are equal to 0 and conclude that $\beta_i \neq 0$. Additionally, the F-value of 1732 is greater than the critical F-value of 2.1037, so we can reject the null hypothesis of the omnibus overall test and conclude that at least one $\beta_i \neq 0$. However, these hypothesis tests do not provide much value as the underly assumptions of homoscedasticity has been violated. We can see in the below histogram that the distribution of the standardized residuals appears to be normally distributed. However, the scatterplot shows that there is heteroscedasticity in the residuals as there the variance of the residuals increases as the predicted sales price increases. Therefore, the underlying assumptions have been violated and the results of hypothesis tests cannot be relied on.

Histogram and Scatterplot of Standardized Residuals:



Conclusion

This analysis shows that variable transformation and outlier detection/removal can alter the modeling process and impact the model results. Outlier deletion can introduce modeler bias, especially if these are legitimate and valid data points. This is similar to how the definition of the sample population can be subjective to the modeler. We can see that removing outliers can improve model performance; however, at the same time could reduce how well the model performs when applied to different datasets If other datasets have observations that are similar to the outliers that were removed, then this could lead to inaccurate predictions on these observations. However, it could also be appropriate to remove outliers that do not represent or share similar qualities as the sample population as they can negatively influence a model's fit. Transformations could also be a valid modeling process if there is justification, such as creating a homoscedasticity pattern in the residuals to satisfy underlying assumptions of hypothesis tests. However, with transformations we could potentially lose the

interpretability of a model especially when dealing with a non-technical audience. So, there needs to be caution on when to use transformation and to only use it when it is necessary and justifiable. Futhermore, I think that hypothesis tests can be trusted only to a certain extent. There is some bias in subjectivity when selecting and defining the sample population and also if and which observations should be removed from the dataset. Additionally, it is difficult to trust statistical hypothesis tests if we are not sure if the underlying assumptions have been met or have been violated. If they have been violated, then we cannot trust the results of these hypothesis tests.

For the next steps in the modeling process and future work, we could look at interactions between variables and check for any collinearity issues. In my initial exploratory analysis, I noticed that some of the independent variables have linear relationships with each other. So, these relationships and interactions would need to be explored further and/or accounted for in subsequent modeling efforts. Lastly, due to the heteroscedasticity pattern displayed in the non-transformed model residuals, I would recommend using confidence intervals instead of hypothesis tests for further assessment of these models to determine if statistical inference is appropriate.