MSDS 410 Assignment 2

Sameer Khan

Section 1: Introduction

In this week's assignment, the full Ames Housing dataset is examined to build different regression models. Simple and multiple regression models are constructed with the goal of finding the best predictors of SalePrice, the selling price of an individual home. There are a very large (>50) number of predictor variables, so it is important to select the most significant ones in determining price. The correlation coefficients of numeric variables with respect to SalePrice are calculated. Since we are only picking one variable each for simple linear regression and four total for multiple regression, it is ideal to select inputs with high correlation coefficient values as they are more likely to be good indicators of the response variable when applied to the specified model. Two variables, GrLivArea and TotalSqftCalc which respectively represent the above ground living area and total home square feet are fitted into a simple linear regression model. The models are compared using scatterplots, residual distribution, QQ plots, and a standard coefficient table. Two more variables are added to the initial ones to create a multiple regression model. Simple and multiple regression models are compared to see if adding additional variables increased the effectiveness of prediction. For a multiple regression comparison, the original MLR model's response variable undergoes a log transformation. Log transformation can often correct skew in data by spreading out consolidated points and bringing far datapoints closer to most observations.

Section 2: Simple Linear Regression Models

First, a SLR model is constructed using TotalSqftCalc as the independent variable. A scatterplot for TotalSqftCalc vs. SalePrice is produced in Figure 1, showing the fitted data mostly follows the regression line until very high values of TotalSqftCalc are reached.

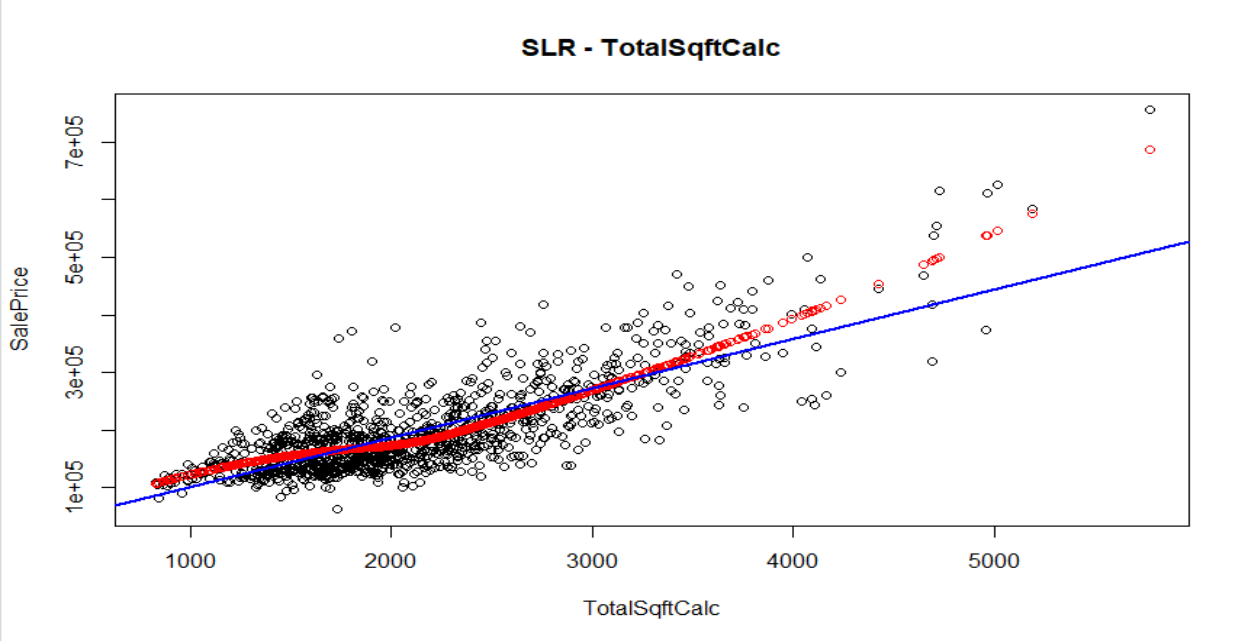

Figure 1: A scatterplot of TotalSftCalc against SalePrice

A similar second simple regression model and scatterplot is created using GrLivArea as the predictor, shown in Figure 2.

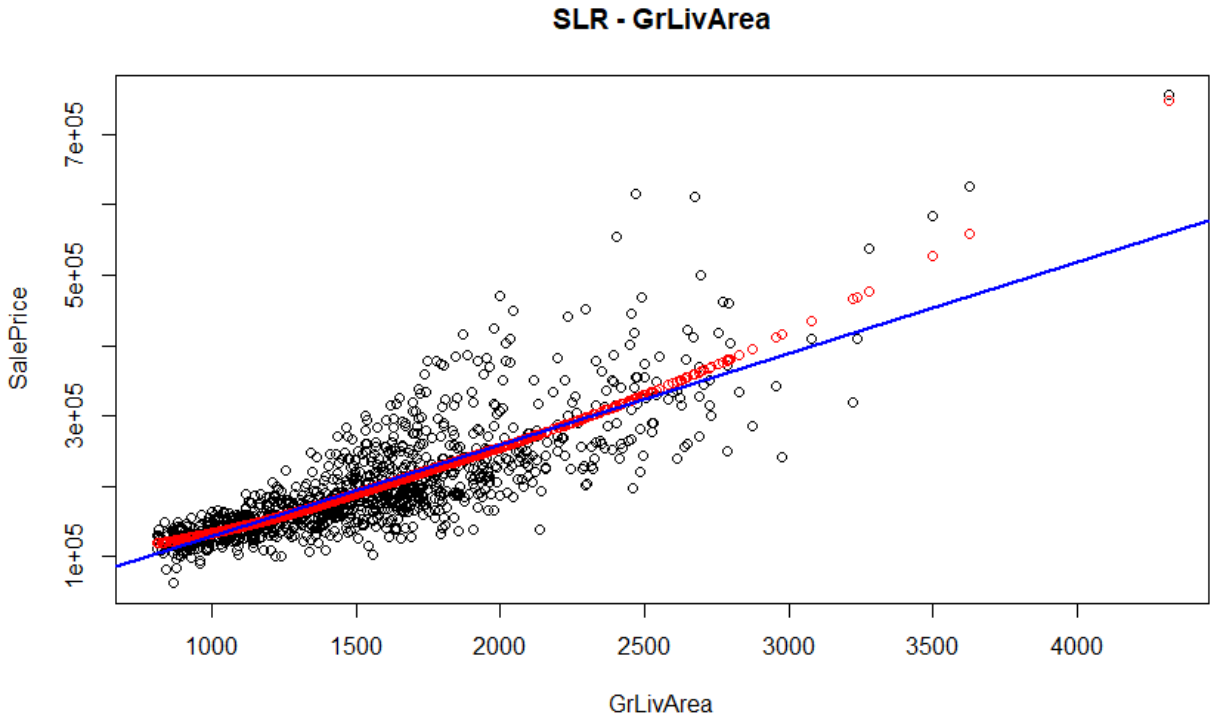

Figure 2: Scatterplot of GrLivArea against SalePrice

The fitted values on this plot are much closer to the regression line than the plot using TotalSqftCalc, although both indicate a good fit of the model. After a basic visualizing of the models, we look the regression coefficient output tables for each model, shown in Figure 3 and Figure 4.

SLR-TotalSqftCalc

	Dependent variable:
	SalePrice
TotalSqftCalc	85.661***
	(2.004)
Constant	15,455.700***
	(4,482.657)
Observations	1,135
R2	0.617
Adjusted R2	0.617
Residual Std. Error	47,856.170 (df = 1133)
F Statistic	1,827.529*** (df = 1; 1133)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 3: Regression output for SLR (TotalSqftCalc)

SLR-GrLivArea

	Dependent variable:
	SalePrice
GrLivArea	129.762***
	(2.788)
Constant	-1,149.016
	(4,469.090)
Observations	1,135
R2	0.657
Adjusted R2	0.656
Residual Std. Error	45,331.920 (df = 1133)
F Statistic	2,166.416*** (df = 1; 1133)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 4: Regression output for SLR (GrLivArea)

We see that in both models, the regression coefficient has a p-value less than 0.01, making both variables statistically significant. The R-Square values for each model are relatively close, but the SLR using GrLivArea accounts for about 4% more of the variation in SalePrice. Given that the dependent variable is the price of a home with a mean value of roughly 200,000 dollars based on 1,135 observations even a small increase in R-Squared can make our model more efficient. The residual plots for each SLR are produced in Figures 5 and 6.

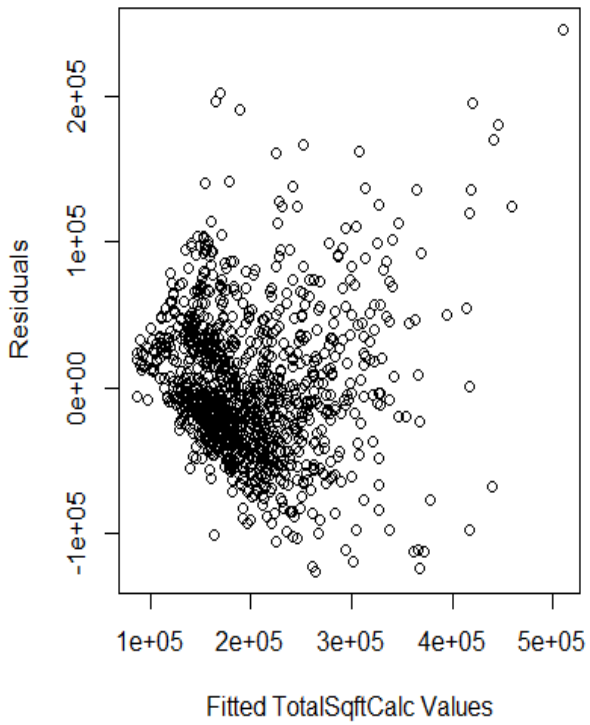

Figure 5: TotalSqftCalc Residual Plot

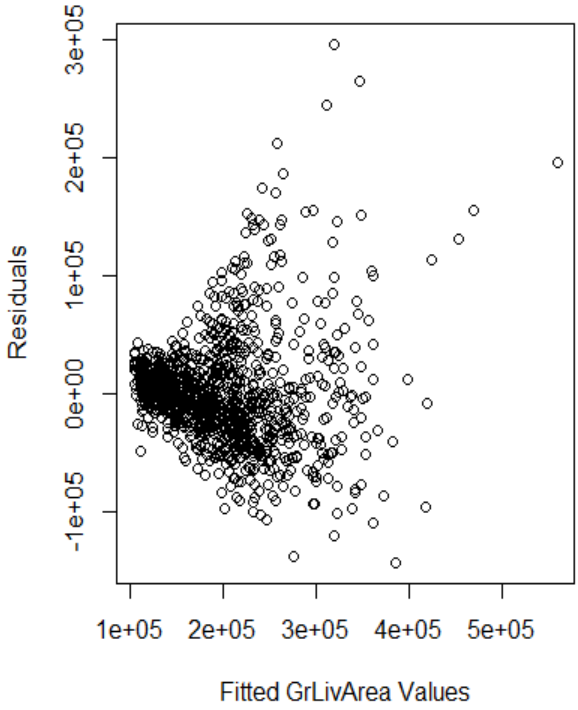

Figure 6: GrLivArea Residual Plot

The residuals for both plots appear to be randomly distributed, but the TotalSqftCalc residuals follow the principle of homoscedasticity closer than GrLivArea. The variance of GrLivArea starts off small for small values, but greatly increases as the independent variable increases. QQ plots are generated to check the normality of the probability distribution of sample and theoretical quantities:

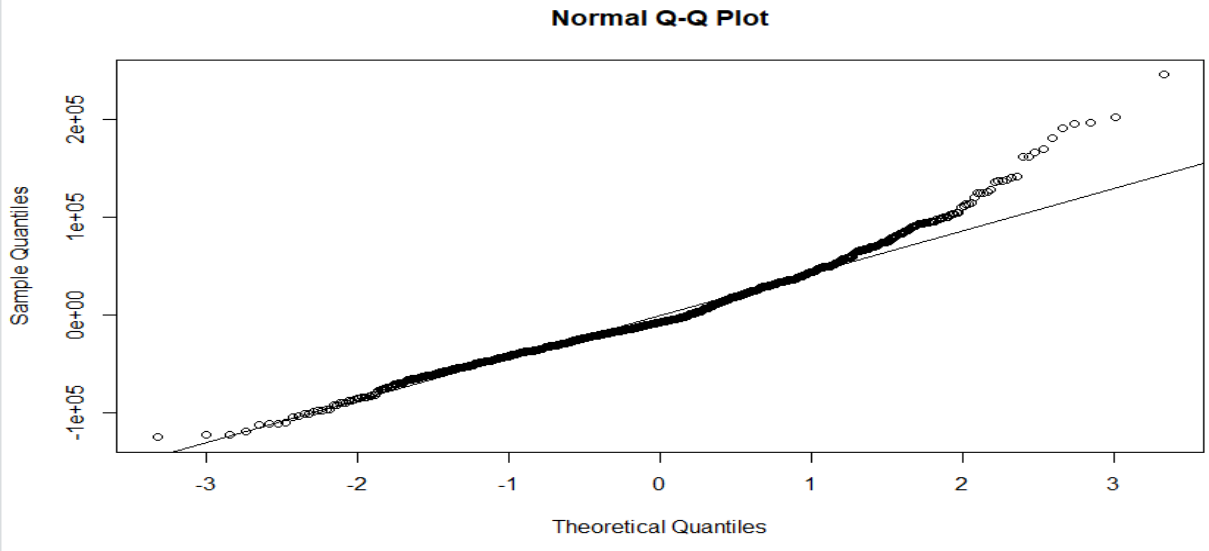

Figure 7: QQ Plot for TotalSqftCalc

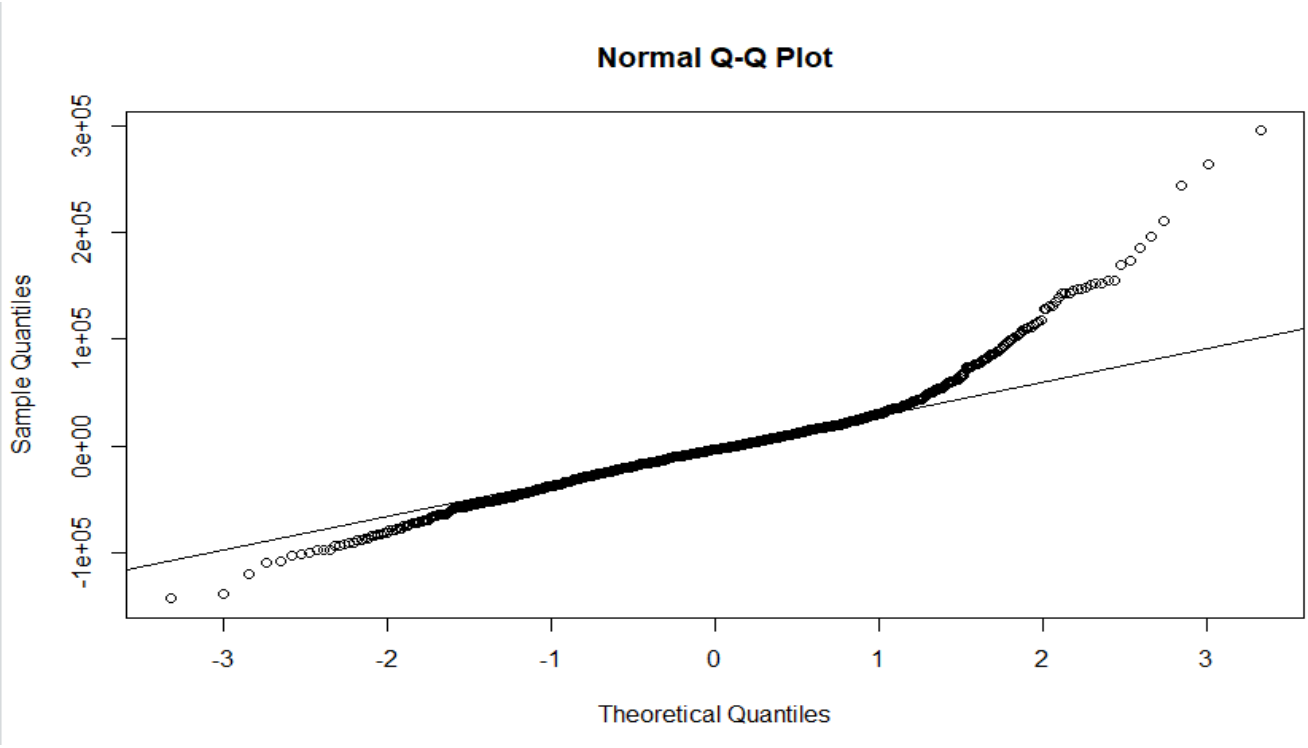

Figure 8: QQ Plot for GrLivArea

The plots indicated that the data is close to normal, but not exact. For TotalSqftCalc, the distribution indicates a slight right skew. In the context of house prices, this could be represented by a normal distribution of typical homes and a few extremely large houses with high selling prices. For GrLivArea, the distribution deviates from the straight line at both ends of the QQ plot. This indicates a normal distribution of values with large tails.

Section 3: Multiple Linear Regression Model

A multiple regression model is created using the two variables from Section 2 as well as two new ones, OverallQual and GarageCars. OverallQual represents the overall material and finish quality of a home, GarageCars is the size of the garage in car capacity. These two variables had relatively high correlation coefficients with respect to SalesPrice, 0.82 for OverallQual and 0.69 for GarageCars. The SLR models for variables with similar coefficients yielded good fitting models, so we should expect a better fit if we add more strongly correlated predictors to the models. The multiple regression results confirm this hypothesis in Table 1:

```
MLR Results
================================================
                        Dependent variable:
                    ----------------------------
                              SalePrice
------------------------------------------------
TotalSqftCalc                  39.831***
                                (1.904)

GrLivArea                      25.721***
                                (3.238)

OverallQual                  25,079.550***
                              (1,022.922)

GarageCars                   15,753.810***
                              (1,793.684)

Constant                    -114,684.900***
                              (4,323.249)

------------------------------------------------
Observations                      1,135
R2                                0.860
Adjusted R2                       0.860
Residual Std. Error    28,955.810 (df = 1130)
F Statistic         1,739.186*** (df = 4; 1130)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
> |
```

Table 1: Multiple Linear Regression output using four predictor variables

The regression results show that each predictor variable is significant with a low p-value. The R-Squared value (0.86) for the model is much better than either simple linear regression model, accounting for 20% more of the variation in SalePrice. Adding more variables will always increase the R-Square value, but only indicates a better fit if the added predictors are statistically significant. Since our two added variables increased the R-Square value by 0.2, they should be deemed as good additions to the model. The QQ-Plots for multiple regression show a heavily right skewed distribution in Figure 9:

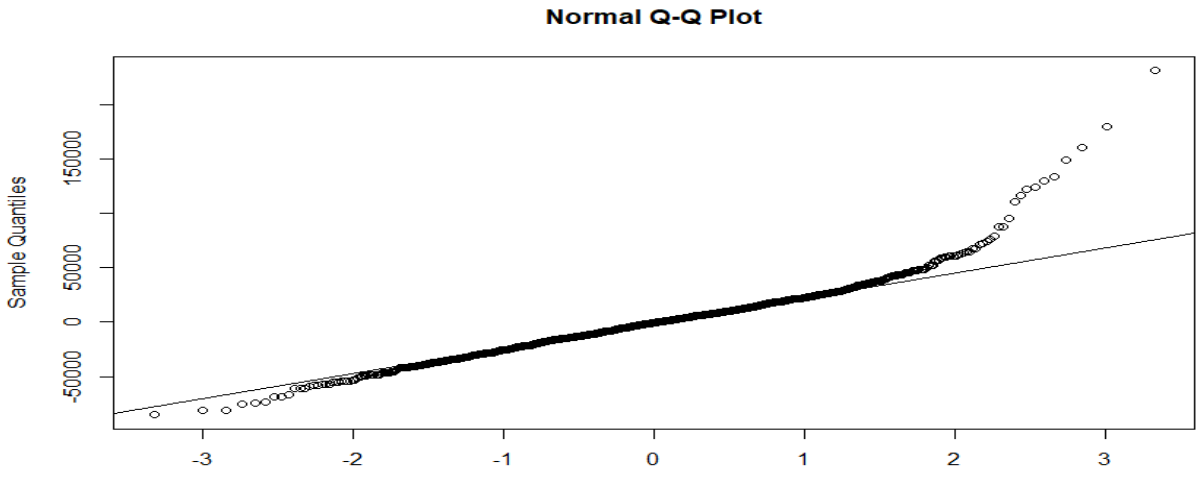

Figure 9: QQ Plot for MLR with SalePrice as the response

Skewness has the potential to decrease the predictive power of out model. The distribution should be approximately normal, and the presence of outlier points might be responsible for the deviation from the 45-degree theoretical/sample line. A log transformation is conducted in Section 4 to try and normalize the data.

Section 4: Transformed MLR Model

For the transformed MLR model we keep all of the predictor variables the same and change the response from SalePrice to log(SalePrice). Log transformation can reduce skewness by reducing the distance of large outliers from the rest of the dataset. This transformation also changes the regression coefficients as shown in Table 2:

```
MLR Results
================================================
                        Dependent variable:
                    ---------------------------
                             log(SalePrice)
------------------------------------------------
TotalSqftCalc                  0.0001***
                               (0.00001)

GrLivArea                      0.0002***
                               (0.00001)

OverallQual                    0.115***
                                (0.004)

GarageCars                     0.091***
                                (0.007)

Constant                       10.723***
                                (0.016)

------------------------------------------------
Observations                      1,135
R2                                0.902
Adjusted R2                       0.901
Residual Std. Error         0.107 (df = 1130)
F Statistic           2,592.548*** (df = 4; 1130)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
> |
```

Table 2: Multiple Regression output with log(SalePrice) as response

We notice that the p-values of the predictors do not change, and they are still relevant to the model. The R-Square value has also slightly increased, now accounting for over 90% of the variation in the dependent variable. To interpret the coefficients, we would have to exponentiate the coefficient and subtract 1 to find the fixed percent change in SalePrice resulting from a single unit increase in the predictor variable. The QQ plot displayed in Figure 10 follows normality much more closely than the untransformed model.

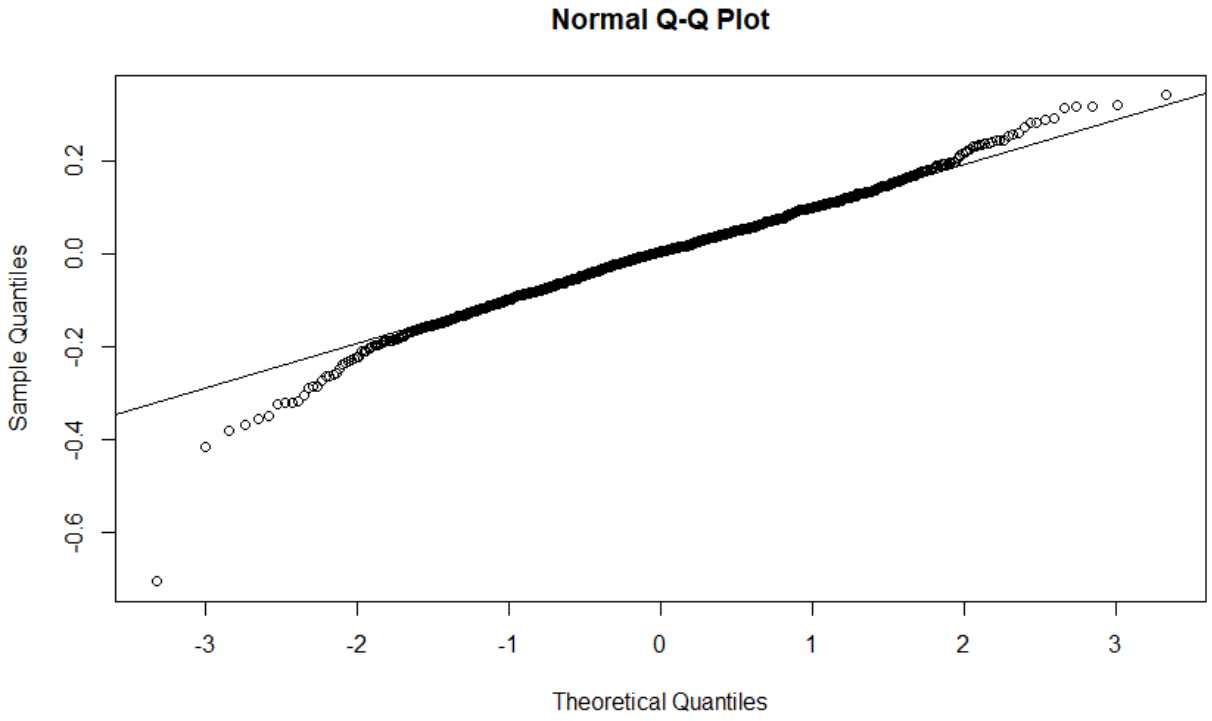

Figure 10: QQ Plot of MLR using log(SalePrice) as response

The theoretical values and sample values are much closer to each other once the transformation has been performed.

The mean square error, which measures the average value of the squares of each model's error terms is used to determine the quality of the model. The MSE should be minimized to maximize the effectiveness of the regression model. The MSE for each of the four models created are calculated below:

SLR with TotalSqftCalc as predictor: 2,286,177,444

SLR with GrLivArea as predictor: 2,051,361,824

MLR with SalePrice as response: 834,745,622

MLR with log(SalePrice) as response: 502,160,344

With each consecutive model, the MSE decreases with the log transformed model giving the lowest result. The values of the MSE are very large which could be due to the squaring of the residuals that are taking place. Since home prices are in the hundreds of thousands, even a residual of a few thousand will create a large MSE value when squared.

Appendix

```r
#Sameer Khan
#MSDS 410 Assignment 2

ames.df <- readRDS('C:\\Users\\samee/Downloads\\ames_sample.RData');
str(ames.df)
num_col <- unlist(lapply(ames.df, is.numeric))
num_col
cor(ames.df[,num_col],ames.df$SalePrice)

#For SLR, the two chosen variables are GrLivArea and TotalSqftCalc

loess1 <- loess(SalePrice ~ TotalSqftCalc,data=ames.df);
model1 <- lm(SalePrice ~ TotalSqftCalc,data=ames.df);

plot(ames.df$TotalSqftCalc, ames.df$SalePrice,xlab='TotalSqftCalc',ylab='SalePrice')
points(loess1$x,loess1$fitted,type='p',col='red')
abline(coef=model1$coef,col='blue',lwd=2)
title('SLR - TotalSqftCalc')

loess2 <- loess(SalePrice ~ GrLivArea,data=ames.df);
model2 <- lm(SalePrice ~ GrLivArea,data=ames.df);

plot(ames.df$GrLivArea, ames.df$SalePrice,xlab='GrLivArea',ylab='SalePrice')
points(loess2$x,loess2$fitted,type='p',col='red')
abline(coef=model2$coef,col='blue',lwd=2)
title('SLR - GrLivArea')

install.packages("stargazer")
library(stargazer)

#Coeff table for each model
stargazer(model1,title="SLR-TotalSqftCalc", align=TRUE, type="text")
stargazer(model2,title="SLR-GrLivArea", align=TRUE, type="text")

#Residual Plots
model1_res <- resid(model1)
model2_res <- resid(model2)

plot(fitted(model1),model1_res,xlab='Fitted TotalSqftCalc Values',ylab='Residuals')
plot(fitted(model2),model2_res,xlab='Fitted GrLivArea Values',ylab='Residuals')

#QQ Plots
qqnorm(model1_res)
qqline(model1_res)

qqnorm(model2_res)
qqline(model2_res)


#Multiple Linear Regression
mlr <- lm(SalePrice ~ TotalSqftCalc + GrLivArea + OverallQual + GarageCars, data=ames.df)
stargazer(mlr, title='MLR Results',align=TRUE,type='text')

plot(fitted(mlr),mlr_res)
mlr_res <- resid(mlr)

qqnorm(mlr_res)
qqline(mlr_res)


#Multiple Linear Regression with log(SalePrice) as the response
mlr2 <- lm(log(SalePrice) ~TotalSqftCalc + GrLivArea + OverallQual + GarageCars, data=ames.df )
stargazer(mlr2, title='MLR Results',align=TRUE,type='text')

mlr2_res <- resid(mlr2)
plot(fitted(mlr2),mlr2_res)
qqnorm(mlr2_res)
qqline(mlr2_res)

#MSE Values for Each Model

#SLR with TotalSqftCalc as predictor
x1 <- mean(model1$residuals^2)

#SLR with GrLivArea as predictor
x2 <- mean(model2$residuals^2)

#MLR with SalePrice as response
x3 <- mean(mlr$residuals^2)

#MLR with log(SalePrice) as response
x4 <- mean((ames.df$SalePrice-exp(predict(mlr2)))^2)
```