Regression Model Building

Claire Boetticher, MSDS 410, SEC 56

September 27, 2020

Section 1: Introduction

The original Ames housing data set contains information from the Ames Assessor's Office for tax assessment purposes, specifically computing assessed values for single residential properties sold in Ames, Iowa from 2006 to 2010. The objective of this analysis is to explore the relationship between predictor variables related to home size, lot size, and build material quality and the response variable, sale price. Simple linear regression (SLR) models and multiple linear regression (MLR) models are fitted using the least squares method and evaluated for goodness-of-fit and potential to meaningfully capture the relationship between the predictor variables of interest and SalePrice. Predictor models are built to examine which predictors are more influential on their own and combined, the shape of the relationships between predictor and response variables, and whether the response variable warrants transformation with the MLR model.

Section 1.1: Data

The selected training data consists of 804 observations (of 1,135 original observations) on five predictor variables and one response variable, SalePrice, in the table below.

Table 1: Variable Definitions

Variable	Definition
SalePrice	Sale price in $
TotalSqftCalc	Total square footage of house
TotRmsAbvGrd	Total number of rooms above grade (does not include bathrooms)
LotArea	Lot size in square feet
GrLivArea	Above grade living area square feet
QualityIndex	Index value for quality: product of OverallQual (rating of overall material and finish of the house, 1=very poor to 10=very excellent) and OverallCond (rating of overall condition of the house, 1=very poor to 10=very excellent)

There are no missing observations from this subset of the original Ames housing data set. There are some select outliers, shown in Table 2, especially high-end ones for SalePrice and TotRmsAbvGrd, whose effect may be worth investigating further beyond the scope of this analysis.

Table 2: Summary Data Check

	No.Obs	Missing	Minimum	1st.Qtr	Median	3rd.Qtr	Maximum	Mean	Range	Lo.Outs	Hi.Outs
SalePrice	804	0	62,383	143,487.50	178,000	228,500	625,000	196,947.60	562,617	0	33
TotalSqftCalc	804	0	832	1,637	1,954.50	2,466	5,185	2,119.75	4,353	0	27
TotRmsAbvGrd	804	0	4	6	6	7	12	6.54	8	30	61
LotArea	804	0	3,182	8,640	9,938	11,900	215,245	11,060.44	212,063	1	28
GrLivArea	804	0	816	1,146	1,493	1,776	3,627	1,532.16	2,811	0	14
QualityIndex	804	0	15	30	35	40	72	34.38	57	0	6

Section 1.2: Exploratory Data Analysis and Correlations

Correlation values are calculated as an initial indicator of promising predictor variables of SalePrice. The correlation value of the predictor variables to the response variable, SalePrice, is calculated as an indicator of potential linear relationships and associations between a given predictor variable and the response variable. The correlation values range from -1, a strongly negative correlation, to 1, a strongly positive correlation. Results in Figure 1 show strongest correlations in GrLivArea (0.80), TotalSqftCalc (0.78), and TotRmsAbvGrd (0.64). The weakest correlation is in LotArea (0.29).

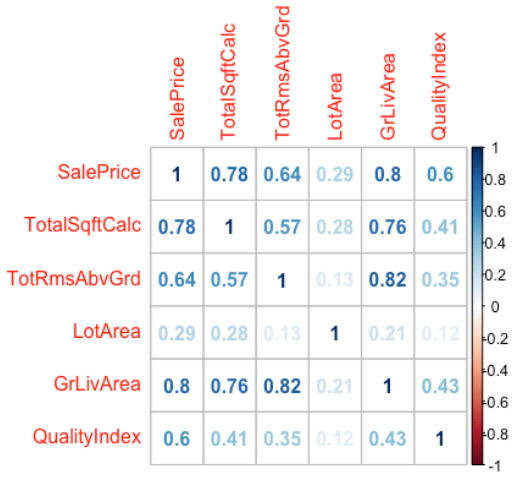

Figure 1: Correlation Matrix of Ames Housing Data

Section 2: Simple Linear Regression Models

Exploratory data analysis suggests two predictor variables, TotalSqftCalc and TotRmsAbvGrd, as valid candidates for building effective SLRs. The scatterplots in Figure 2 provide visual confirmation of their potential as predictor variables with SalePrice, with the locally estimated scatterplot smoothing (LOESS) line showing relative linearity in both TotalSqftCalc and TotRmsAbvGrd, falling along the regression line up to square footage values of 4000 for TotalSqftCalc and for all values taken for TotRmsAbvGrd.

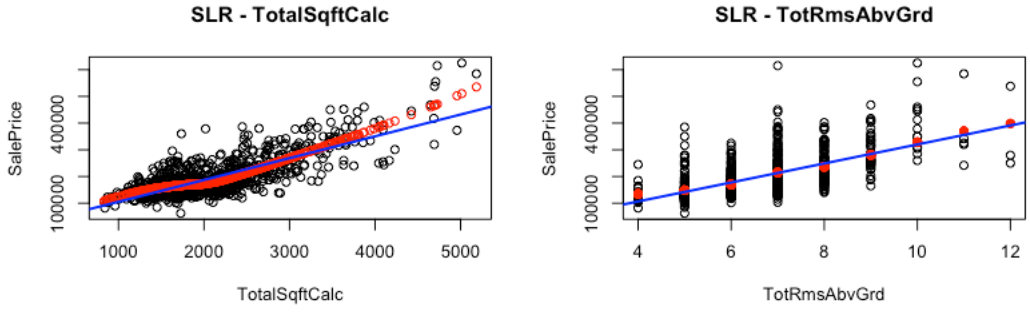

Figure 2: Scatterplots of Select Predictor Variables and SalePrice

Section 2.1: Model #1 (TotalSqftCalc)

Table 4 shows the Ordinary Least Squares (OLS) regression results for the single variable TotalSqftCalc.

Table 3: Linear Regression Results - TotalSqftCalc

	Dependent variable:
	SalePrice
TotalSqftCalc	81.68***
	(2.35)
Constant	23,812.96***
	(5,258.48)
Observations	804
R^2	0.60
Adjusted R^2	0.60
Residual Std. Error	47,516.56 (df = 802)
F Statistic	1,206.58*** (df = 1; 802)
Note:	*p**p***p<0.01

The simple regression model is

$$SalePrice = 23812.96 + 81.68 * TotalSqftCalc + \varepsilon$$

For each rise in 1 unit of TotalSqftCalc, the SalePrice of the home increases by $81.68. This corresponds with the positive correlation calculated in Figure 1 (0.78) and observed in Figure 2. The R-squared value, a number between zero and one that quantifies the variance explained in a statistical model, of 0.60 suggests relatively decent explanation of variance by the single predictor variable. The diagnostic plots to assess goodness-of-fit for the model in Figure 3 support the statistical results. The residuals vs. predictor plot is relatively shapeless without much of a clear pattern in the data up until TotalSqftCalc values of 4000, suggesting homoscedasticity in the errors. The Quantile-Quantile (Q-Q) plot is positively skewed, though, seen in the points' deviation from the line at the upper end of the plot, suggesting some potential model shortcomings.

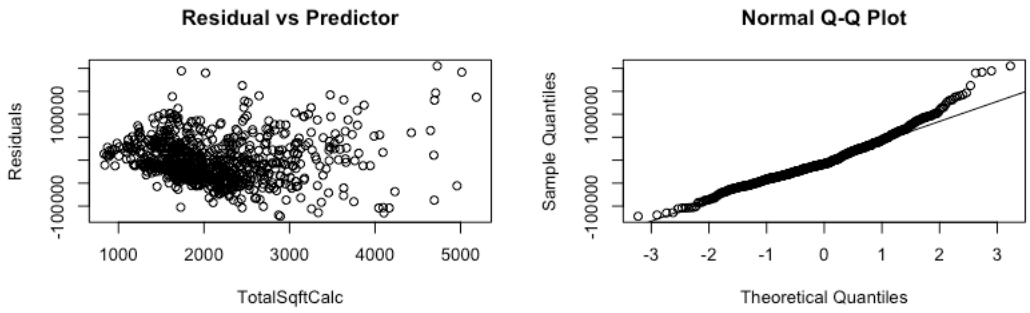

Figure 3: Simple Regression Diagnostic Plots

Section 2.2: Model #2 (TotRmsAbvGrd)

Table 5 shows the OLS regression results for the single variable TotRmsAbvGrd.

Table 5: Linear Regression Results - TotRmsAbvGrd

	Dependent variable:
	SalePrice
TotRmsAbvGrd	35,576.23***
	(1,518.51)
Constant	-35,625.35***
	(10,135.08)
Observations	804
R^2	0.41
Adjusted R^2	0.41
Residual Std. Error	57,940.17 (df = 802)
F Statistic	548.89*** (df = 1; 802)
Note:	*p**p***p<0.01

The simple regression model is

$$SalePrice = -35625.35.23 + 35576.23 * TotRmsAbvGrd + \varepsilon$$

For each rise in 1 unit of TotRmsAbvGrd, the SalePrice of the home increases by $35576.23. This corresponds to the positive correlation value calculated in EDA (0.64, Figure 1). The R-squared value of 0.41 suggests that TotRmsAbvGrd, though positively correlated with SalePrice, explains less than half the variance in SalePrice.

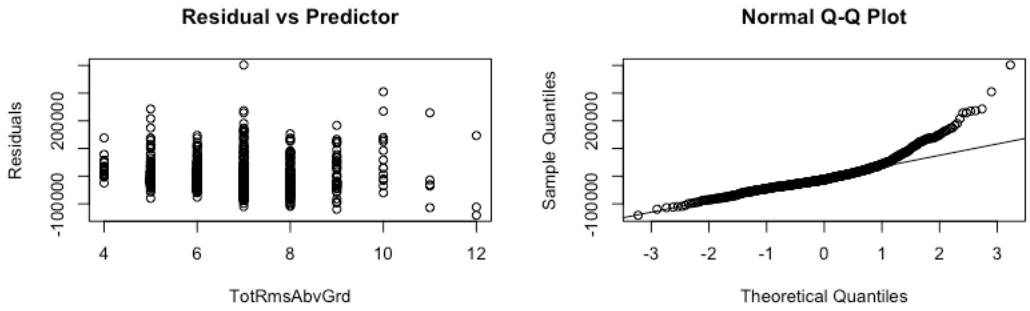

Figure 4: Simple Regression Diagnostic Plots

The diagnostic plots in Figure 4 show some potential issues with this model. In the residual vs. predictor plot, the plot is not evenly distributed vertically and residuals show relative uniformity at values of four to nine TotRmsAbvGrd. The Q-Q plot shows sizeable positive skew, suggesting the distribution assumption of normality may not be met.

Section 2.3: Model Comparison

SLR model 1, using TotalSqftCalc as a single predictor variable, generally meets more criteria for goodness-of-fit than SLR model 2 using TotRmsAbvGrd. These criteria include the visual diagnostics provided in Figure 2, where the data adheres more closely to the regression line and to the LOESS line as well, versus the greater deviance from that line with model 2. The residual plot and Q-Q plot for each model is even more telling. The residual vs. predictor plot for model 1 suggests homoscedasticity of errors with a relative lack of pattern in the plot and even spread above and below the zero line for residuals up to values of 4000, whereas this same plot for model 2 shows somewhat of a visual trend for certain values of TotRmsAbvGrd, showing possible issues of fit. The Q-Q plot for model 1 and 2 both show positive skew, but it is more pronounced in model 2, again suggesting some issues of model adequacy. The R-squared values for each are somewhat less revealing of model adequacy; while model 1's higher value could suggest a generally stronger explanatory relationship between TotalSqftCalc and SalePrice, this statistic serves more as a supporting signal for each model's particular goodness-of-fit than a reliable point of comparison between the two.

Section 3: Multiple Linear Regression Model (Model #3)

As a comparison to the two SLRs from Section 2, a multiple linear regression model (MLR) with the same response variable, SalePrice, and multiple predictor variables (TotalSqftCalc, TotRmsAbvGrd, LotArea, and QualityIndex) is fitted and evaluated. LotArea, though calculated with a low correlation value in Figure 1 (0.29), is a variable worth exploring for total outside area's effect and QualityIndex (correlation value of 0.6) merits exploration given its measurement of a home's actual building materials and general condition, not just its size. GrLivArea is dropped. SLRs are constructed for each of these two added predictor variables as a baseline to evaluate individual explanatory potential for SalePrice and the shape of their relationship with that response variable.

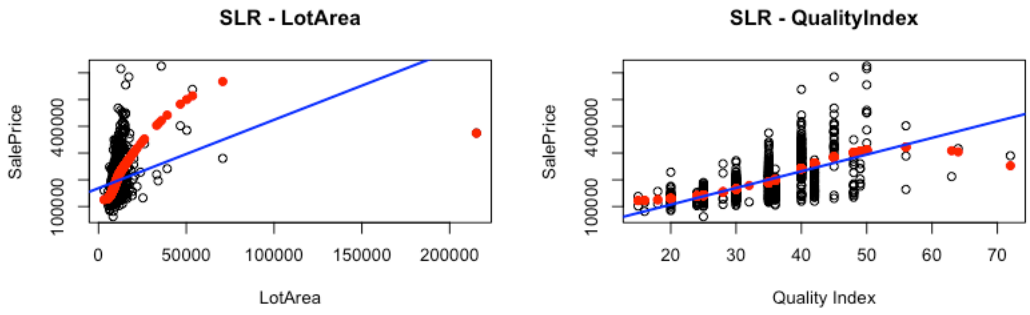

Figure 5: Scatterplots of Select Predictor Variables and SalePrice (LotArea and QualityIndex)

The shape of LotArea's relationship to SalePrice appears influenced by certain outlier points; gauged as a single predictor, its correlation is not as strong as other selected variables. However, correlation value alone may not reflect its influence when combined with other variables so it is included in the MLR; moreover, outlier treatment may reflect a clearer relationship in future analyses. The LOESS line for QualityIndex falls neatly along the regression line up until values of 55. Residual plots and quantile-quantile plots for each individual added predictor variable provide a further baseline for assessing their respective model's goodness-of-fit, as a point of comparison for the MLR of combined predictor variables.

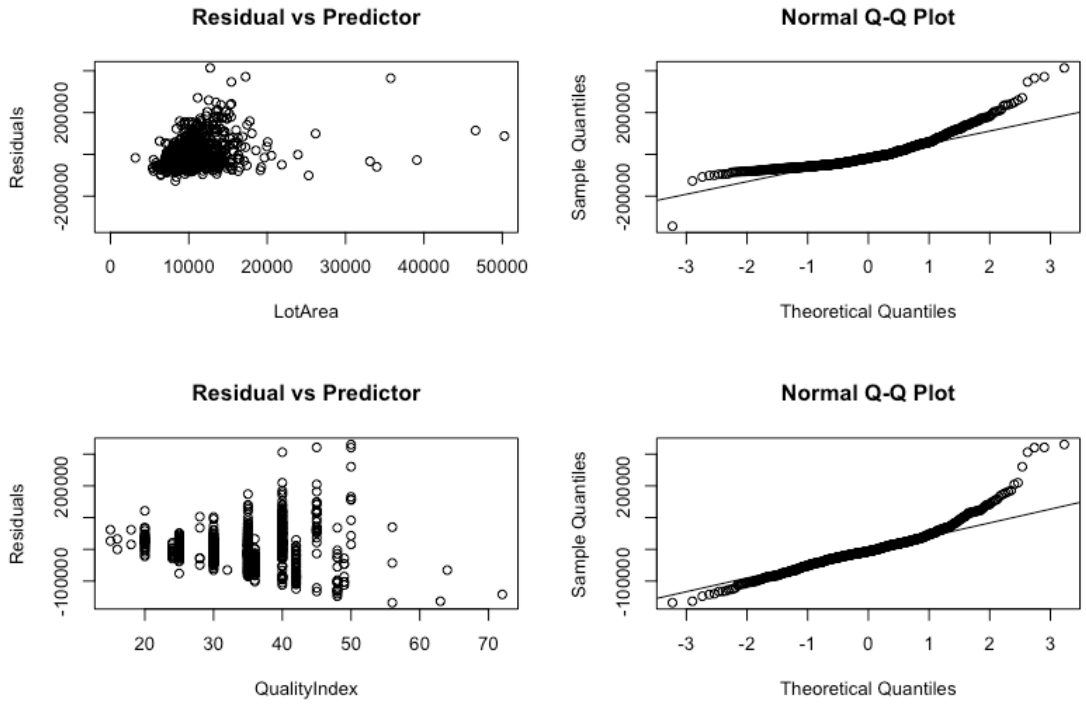

Figure 6: Simple Regression Diagnostic Plots (LotArea and QualityIndex)

LotArea's residuals vs. predictor plot does seem to reflect homoscedasticity, with positive residual values increasing as LotArea values increase up to values of 20,000. The Q-Q plot shows positive skew as well, though much of the plot does fall along the line. For QualityIndex, there is a trend of residuals increasing as QualityIndex values increase, reflecting a possible issue with the linear model. The Q-Q plot shows strong positive skew as well. Though these two predictor variables, assessed in SLR models in the context of their individual relationship to SalePrice, reflect potential issues in model adequacy, they are worth exploring as potential complementary factors in predicting SalePrice that address potential influential factors outside of a home's size.

Table 6 shows the OLS regression results for the multiple variables, TotalSqftCalc, TotRmsAbvGrd, LotArea, and QualityIndex.

Table 6: Multiple Linear Regression Results

	Dependent variable:
	SalePrice
TotalSqftCalc	51.83***
	(2.49)
TotRmsAbvGrd	13,177.94***
	(1,252.47)
LotArea	0.78***
	(0.17)
QualityIndex	3,166.75***
	(210.50)
Constant	-116,608.30***
	(8,190.73)
Observations	804
R^2	0.74
Adjusted R^2	0.74
Residual Std. Error	38,580.72 (df = 799)
F Statistic	561.94*** (df = 4; 799)
Note:	*p**p***p<0.01

The multiple variable regression model is

$$\textit{SalePrice} = -116608.30 + (51.83 * \textit{TotalSqftCalc}) + (13177.94 * \textit{TotRmsAbvGrd}) + (0.78 * \textit{LotArea}) + (3166.75 * \textit{QualityIndex}) + \varepsilon$$

The regression results show an improvement in R-squared from model 1 (0.60) and model 2 (0.41) to 0.74. This suggests an improved explanation of the variance, with these variables together, over either SLR model.

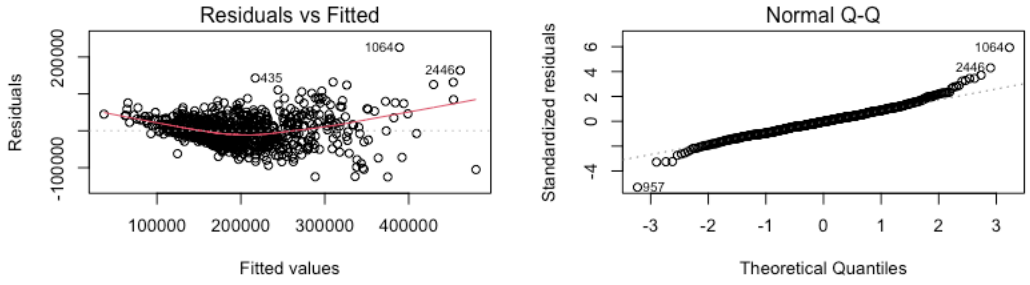

Figure 7: Multiple Linear Regression Diagnostic Plots

The general residual plot in Figure 7 suggests potential shortcomings with this model's goodness-of-fit, showing a slightly curved line and the residuals appearing to increase at certain points as the fitted values increase, suggesting heteroscedasticity. The Q-Q plot shows a noticeable right skew, a possible sign of not meeting the distribution assumption

of normality. Individual diagnostic plots are constructed for each of the added two predictor variables to investigate their relationship to SalePrice and potential explanatory influence for variance.

Model fit given a comparison of regression results with attention to R-squared values and these diagnostic plots aid in deciphering strength of relationships amongst variables and their explanatory potential with respect to SalePrice, however a formal declaration of goodness-of-fit is not possible with these criteria alone. Moreover, the selected predictor variables' interaction amongst themselves may be as much an influence on goodness-of-fit. More experimentation with both single predictors and other combinations of multiple predictors would be warranted for further exploration.

Section 4: Transformed MLR Model (Model #4)

The heteroscedasticity seen in Figure 7 indicates that the response variable may merit transformation. SalePrice is transformed to the natural logarithm of SalePrice to determine effect on model fit using the same MLR from model 3.

Section 4.1: Model #4 Log SalePrice Model

Table 7 shows the OLS regression results for the multiple variables, TotalSqftCalc, TotRmsAbvGrd, LotArea, and QualityIndex.

Table 7: Multiple Linear Regression Results - log(SalePrice)

	Dependent variable:
	log(SalePrice)
TotalSqftCalc	0.0002***
	(0.0000)
TotRmsAbvGrd	0.07***
	(0.01)
LotArea	0.0000***
	(0.0000)
QualityIndex	0.02***
	(0.001)
Constant	10.67***
	(0.04)
Observations	804
R^2	0.75
Adjusted R^2	0.75
Residual Std. Error	0.17 (df = 799)
F Statistic	589.11*** (df = 4; 799)
Note:	*p**p***p<0.01

The multiple variable regression model is

$$\log(SalePrice) = 10.67 + (0.0002 * TotalSqftCalc) - (0.071 * TotRmsAbvGrd) + (0.000 * LotArea) + (0.02 * QualityIndex) + \varepsilon$$

The regression results show a minor improvement in R-squared from model 3 (0.74) to model 4 (0.75).

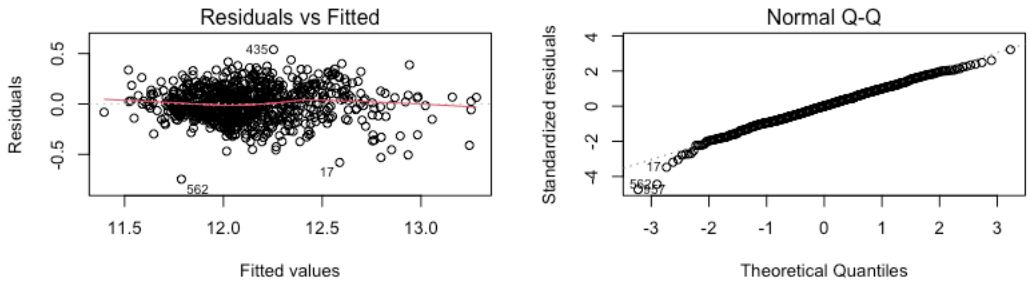

Figure 8: Multiple Linear Regression Diagnostic Plots

From a visual inspection, the general residuals vs. fitted plot in Figure 8 shows slightly less concentration and more even spacing around the zero residual axis, showing stronger evidence of homoscedasticity. There is less deviation from the line in the Q-Q plot with a slight left skew, indicating greater likelihood of meeting the distribution assumption of normality.

Section 4.2: Comparison of MLR Models

The minimal difference in R-squared values between model 3 using the raw response variable and model 4, with SalePrice transformed to the natural logarithm of SalePrice, does not provide foundation alone for an assessment of goodness-of-fit. The diagnostic plots in Figure 8, however, indicate a possibly-improved explanation of the variance using the transformed response variable over model 3. As an additional measure of comparison, Mean Squared Error (MSE) and Mean Absolute Error (MAE) are calculated for each model on the SalePrice scale to enable an appropriate comparison.

Table 8: Mean Squared Error and Mean Absolute Error for MLRs

Model	Mean Squared Error	Mean Absolute Error
Model 3: SalePrice	1479215462	29079.58
Model 4: log(SalePrice)	1394482721	26962.78

Results from Table 8 show that model 4 predicts with a smaller error, both in terms of MSE and MAE. Overall, it seems to be a reasonable improvement to the statistical model based on these specific statistical and graphical results.

Section 5: Conclusion

On review of the Ames housing data set, the data check found 804 records adequate for modeling as a training set from 1,135 original records. EDA on single predictors – TotalSqftCalc and TotRmsAbvGrd – and the shape of their relationship with the response variable – SalePrice – identified each as suitable for OLS regression modeling to predict SalePrice. The SLR models developed and evaluated for each identified model 1, using TotalSqftCalc as a predictor, as fitting this data slightly better based on regression results and diagnostic plots. Regression results from MLR model 3 using the original two predictor variables plus LotArea and QualityIndex (identified as candidates via EDA and potential for complementing fit) and diagnostic plots showed improvement from the single variable models. Upon refitting model 3 using the natural logarithm of SalePrice, model 4 (otherwise replicating parameters of model 3) showed further

improvement in terms of regression results, diagnostic plots, and from compared MSE and MAE values. This is not necessarily evidence of added variables always enabling better regression results. Additionally, the interaction amongst the four predictor variables selected could have been a factor in results, not simply a sign of each variable's explanatory effect for SalePrice. For this analysis using this data as defined, though, the inclusion of added variables and transformation may result in additional model improvements.

R Script

```
###################################
# Regression Model Building
#
# Claire Boetticher
# MSDS 410 - Data Modeling for Supervised Learning, Sec 56
# Northwestern University
#
# 09/27/20
#############################################################################
# Data Import and Initial Examination of Data
#############################################################################

# Set working directory
path <- "/Users/clb/Documents/MSDS410/2-Regression_Modeling/"
# set directory to desired working location
setwd(path)

# Load packages
library(corrplot)
library(stargazer)

# Import and examine data
ames.df <- readRDS("data/ames_sample.Rdata")
str(ames.df)      # compact display of structure of data set
head(ames.df)     # returns the first parts of data frame
summary(ames.df)  # summary statistics for data set

# Create output tables for both data sets
out.path <- 'report_outputs/' # define path for html tables

# Perform EDA on training data set
train.df <- subset(ames.df,train==1)
str(train.df)      # compact display of structure of data set
head(train.df)     # returns the first parts of data frame
summary(train.df)  # summary statistics for data set

# Subset of predictor variables
small.df <- train.df[,c('SalePrice','TotalSqftCalc','TotRmsAbvGrd','LotArea','GrLivArea',
                        'QualityIndex')]

#####################################################################
# Summary Data Check function
# Function to output descriptive statistics for data set; credit/gratitude to Scott Forrey
my.summary <- function (data, varname)
{
  g.var <- paste(varname, "sum", sep = ".")
```

```
  outcomes <- c("outcome")
  vars <- setdiff(colnames(data), c(outcomes, "rgroup"))
  numvars <- vars[sapply(data[, vars], class) %in% c("numeric", "integer")]
  k <- length(numvars)
  sum.df <- data.frame(matrix(NA, nrow = 0, ncol = 11))
  colnames(sum.df) <- c("No.Obs", "Missing", "Minimum", "1st.Qtr", "Median", "3rd.Qtr",
                        "Maximum", "Mean", "Range", "Lo.Outs", "Hi.Outs")
  for (i in numvars)
  {
    lo.out <- quantile(data[, i], probs = .25, na.rm = T) - 1.5*(quantile(data[, i], probs = .75,
na.rm = T)-quantile(data[, i], probs = .25, na.rm = T))
    hi.out <- quantile(data[, i], probs = .75, na.rm = T) + 1.5*(quantile(data[, i], probs = .75,
na.rm = T)-quantile(data[, i], probs = .25, na.rm = T))
    sum <- cbind(No.Obs = round(NROW(data[, i]), 0),
                 Missing = round(sum(is.na(data[, i])) + sum(is.null(data[, i])), 0),
                 Minimum = quantile(data[, i], probs = 0, na.rm = T),
                 "1st.Qtr" = quantile(data[, i], probs = .25, na.rm = T),
                 Median = quantile(data[, i], probs = .50, na.rm = T),
                 "3rd.Qtr" = quantile(data[, i], probs = .75, na.rm = T),
                 Maximum = quantile(data[, i], probs = 1, na.rm = T),
                 Mean = round(mean(data[, i], na.rm = T), 4),
                 # Std.Dev = round(sd(data[, i], na.rm = T), 4),
                 # Variance = round(var(data[, i], na.rm = T), 4),
                 Range = quantile(data[, i], probs = 1, na.rm = T)-quantile(data[, i], probs = 0,
na.rm = T),
                 Lo.Outs = round(sum(data[, i] < lo.out[[1]]), 0),
                 Hi.Outs = round(sum(data[, i] > hi.out[[1]]), 0))
    sum.df[i, ] <- sum
  }
  sum.df[, 2:6] <- sapply(sum.df[, 2:6], as.numeric)
  assign(g.var, sum.df, envir = .GlobalEnv)
  return(sum.df)
}

table_sum <- my.summary(small.df, 'Ames_small.datacheck')

# Convert table to a data frame and create output table
summary.table <- as.data.frame(table_sum)

file.name <- '2_datacheck_small.html'
stargazer(summary.table, type=c('html'),out=paste(out.path,file.name,sep=''),
          title=c('Table XX: Summary Data Check'),
          align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE,
          summary=FALSE)

# Check correlation values for subset of predictors from training set
cor(small.df)
```

```
# Create output table of correlation values
file.name <- '3_correlation_table.html'
stargazer(cor(small.df), type=c('html'),out=paste(out.path,file.name,sep=''),
          title=c('Table XX: Predictor Variable Correlations to Response Variable: SalePrice'),
          align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE, median=TRUE)

# Correlation plot
corrplot(cor(small.df),method='number')

########################################
# Section 1: Simple Linear Regression Models
########################################

# Predictor variables: TotalSqftCalc and TotRmsAbvGrd
# Response Variable: SalePrice

# LM 1 (TotalSqftCalc)

# Panel plots
par(mfrow=c(2,2))

# Plot and SLR - SalePrice and TotalSqftCalc
loess.1 <- loess(SalePrice ~ TotalSqftCalc,data=small.df)
lm.1 <- lm(SalePrice ~ TotalSqftCalc,data=small.df)

# Save table output of model
summary(lm.1)
file.name <- 'lm1_totalsqftcal.html'
stargazer(lm.1, type=c('html'),out=paste(out.path,file.name,sep=''),
          title=c('Table 4: Linear Regression Results - TotalSqftCalc'),
          align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE)

# Scatterplot
plot(small.df$TotalSqftCalc, small.df$SalePrice,xlab='TotalSqftCalc',ylab='SalePrice')
points(loess.1$x,loess.1$fitted,type='p',col='red')
abline(coef=lm.1$coef,col='blue',lwd=2)
title('SLR - TotalSqftCalc')

# LM 2 (TotRmsAbvGrd)

# Plot and SLR - SalePrice and TotRmsAbvGrd
loess.2 <- loess(SalePrice ~ TotRmsAbvGrd,data=small.df)
lm.2 <- lm(SalePrice ~ TotRmsAbvGrd,data=small.df)

# Save table output of model
summary(lm.2)
```

```
file.name <- 'lm2_totaltmsabvgrd.html'
stargazer(lm.2, type=c('html'),out=paste(out.path,file.name,sep=''),
          title=c('Table 5: Linear Regression Results - TotRmsAbvGrd'),
          align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE)

# Scatterplot
plot(small.df$SalePrice ~ small.df$TotRmsAbvGrd,xlab='TotRmsAbvGrd',ylab='SalePrice')
points(loess.2$x,loess.2$fitted,type='p',col='red',pch=19)
abline(coef=lm.2$coef,col='blue',lwd=2)
title('SLR - TotRmsAbvGrd')

# Save table output of models together
file.name <- 'SLR_together.html';
stargazer(lm.1, lm.2, type=c('html'),out=paste(out.path,file.name,sep=''),
          title=c('Table XX: Comparison of Model #1 and Model #2'),
          align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE,
          column.labels=c('Model #1','Model #2'), intercept.bottom=FALSE )

# Diagnostic plots for model goodness-of-fit - LM 1 (TotalSqftCalc)

# Panel the plots
par(mfrow = c(2, 2))

# Scatterplot of residuals against predictor - check for homoscedasticity
plot(small.df$TotalSqftCalc,lm.1$residuals,xlab='TotalSqftCalc',ylab='Residuals')
title('Residual vs Predictor')

# Quantile-quantile plot - check for distribution assumption of normality
qqnorm(lm.1$residuals)
qqline(lm.1$residuals)

# ANOVA
anova(lm.1)

# Review coefficients
lm.1$coefficients

# Access residuals from lm object to compute Mean Square Error (MSE) and Mean Absolute Error (MAE)
mse.1 <- mean(lm.1$residuals^2)
mae.1 <- mean(abs(lm.1$residuals))

mse.1
mae.1

# Diagnostic plots for model goodness-of-fit - LM 2 (TotRmsAbvGrd)

# Panel the plots
```

```
par(mfrow = c(2, 2))

# Scatterplot of residuals against predictor - check for homoscedasticity
plot(small.df$TotRmsAbvGrd,lm.2$residuals,xlab='TotRmsAbvGrd',ylab='Residuals')
title('Residual vs Predictor')

# Quantile-quantile plot - check for distribution assumption of normality
qqnorm(lm.2$residuals)
qqline(lm.2$residuals)

# ANOVA
anova(lm.2)

# Review coefficients
lm.2$coefficients

# Access residuals from lm object to compute Mean Square Error (MSE) and Mean Absolute Error (MAE)
mse.2 <- mean(lm.2$residuals^2)
mae.2 <- mean(abs(lm.2$residuals))

mse.2
mae.2

# Examine ratio for comparison of error values

mse.1/mse.2
mae.1/mae.2

########################################
# Section 2: Multiple Linear Regression Models
########################################

# Addition of 2 predictor variables: LotArea and QualityIndex
# Predictor variables: TotalSqftCalc, TotRmsAbvGrd, LotArea, QualityIndex
# Response Variables: SalePrice and log(SalePrice)

# Scatterplots and diagnostic plots for added variables - LotArea and QualityIndex
par(mfrow = c(2,2))

# Scatterplot with LOESS line - LotArea
loess.lotarea <- loess(SalePrice ~ LotArea,data=small.df)
lm.lotarea<- lm(SalePrice ~ LotArea,data=small.df)
plot(small.df$SalePrice ~ small.df$LotArea,xlab='LotArea',ylab='SalePrice')
points(loess.lotarea$x,loess.lotarea$fitted,type='p',col='red',pch=19)
abline(coef=lm.lotarea$coef,col='blue',lwd=2)
title('SLR - LotArea')
```

```
# Scatterplot with LOESS line - QualityIndex
loess.qual <- loess(SalePrice ~ QualityIndex,data=small.df)
lm.qual<- lm(SalePrice ~ QualityIndex,data=small.df)
plot(small.df$SalePrice ~ small.df$QualityIndex,xlab='Quality Index',ylab='SalePrice')
points(loess.qual$x,loess.qual$fitted,type='p',col='red',pch=19)
abline(coef=lm.qual$coef,col='blue',lwd=2)
title('SLR - QualityIndex')

par(mfrow=c(2,2))

# Scatterplot of residuals against predictor - check for homoscedasticity
plot(small.df$LotArea,lm.lotarea$residuals,xlab='LotArea',ylab='Residuals',xlim=c(0,50000))
title('Residual vs Predictor')

# # Quantile-quantile plot - check for distribution assumption of normality
qqnorm(lm.lotarea$residuals)
qqline(lm.lotarea$residuals)

# Scatterplot of residuals against predictor - check for homoscedasticity
plot(small.df$QualityIndex,lm.qual$residuals,xlab='QualityIndex',ylab='Residuals')
title('Residual vs Predictor')

# Quantile-quantile plot - check for distribution assumption of normality
qqnorm(lm.qual$residuals)
qqline(lm.qual$residuals)

par(mfrow=c(1,1))

# Plot and MLR - SalePrice and predictor variables
loess.3 <- loess(SalePrice ~ TotalSqftCalc+TotRmsAbvGrd+LotArea+QualityIndex,data=small.df)
lm.3<- lm(SalePrice ~ TotalSqftCalc+TotRmsAbvGrd+LotArea+QualityIndex,data=small.df)

# Save table output of model
summary(lm.3)
file.name <- 'lm3_mlr.html'
stargazer(lm.3, type=c('html'),out=paste(out.path,file.name,sep=''),
          title=c('Table 6: Multiple Linear Regression Results'),
          align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE)

# Review coefficients
lm.3$coefficients

# General residual plots for MLR model
par(mfrow = c(2, 2))
plot(lm.3)

# ANOVA
```

```
anova(lm.3)

file.name <- 'lm3_anova.html'
stargazer(anova(lm.3), type=c('html'),out=paste(out.path,file.name,sep=''),
          title=c('Table 7: ANOVA Results for MLR'),
          align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE,
          summary=FALSE)

########################################
# Section 3: Regression Models for the Transformed Response
########################################
# Predictor variables: TotalSqftCalc, TotRmsAbvGrd, LotArea, QualityIndex
# Response Variable: log(SalePrice)

# Fit model using log(SalePrice)
lm.4 <- lm(log(SalePrice) ~ TotalSqftCalc+TotRmsAbvGrd+LotArea+QualityIndex, data=small.df)

# Save table output of model

file.name <- 'lm4_logsaleprice.html'
stargazer(lm.4, type=c('html'),out=paste(out.path,file.name,sep=''),
          title=c('Table 8: Multiple Linear Regression Results - log(SalePrice)'),
          align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE)

# Review coefficients
lm.4$coefficients

# Scatterplot of residuals against predictors and Quantile-quantile plot
par(mfrow = c(2, 2))
plot(lm.4)

# ANOVA
anova(lm.4)

file.name <- 'lm3_anova.html'
stargazer(anova(lm.4), type=c('html'),out=paste(out.path,file.name,sep=''),
          title=c('Table 9: ANOVA Results for MLR'),
          align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE,
          summary=FALSE)

# Calculate MSE and MAE values for models 3 and 4
# Model 3
# Access residuals from lm object to compute Mean Square Error (MSE) and Mean Absolute Error (MAE)
mse.3 <- mean(lm.3$residuals^2)
mae.3 <- mean(abs(lm.3$residuals))

# Model 4
```

```
# Take exp of fitted value for mse and mae to get back to original SalePrice
mse.4 <- mean((small.df$SalePrice-exp(lm.4$fitted.values))^2)
mae.4 <- mean(abs(small.df$SalePrice-exp(lm.4$fitted.values)))

# Model 4 predicts with a smaller error (MSE for 3 higher and for MAE as well)
# transformation makes that possible, can compare by converting back to same scale

mse.3
mae.3

mse.4
mae.4

# Calculate ratios
mse.3/mse.4
mae.3/mae.4
```