Assignment 8.2 on Housing Price

Basak Atanu

05-09-2022

library("readxl")  
  
setwd("C:\\Users\\atanu\\Documents\\BellevueUniversity\_MSDS\\DSC520\\Repository\\dsc520")  
housing <- read\_excel("data\\week-6-housing.xlsx")  
housing['years\_old'] <- 2022 - housing['year\_built']  
housing['sales\_price'] <- housing["Sale Price"]

# Explain any transformations or modifications you made to the dataset

### I have created the year\_old variable using the year\_built variable.

# Create two variables; one that will contain the variables Sale Price and Square Foot of Lot (same variables used from previous assignment on simple regression) and one that will contain Sale Price and several additional predictors of your choice. Explain the basis for your additional predictor selections.

sale\_price\_simple <- lm(sales\_price~sq\_ft\_lot, data=housing)  
  
sale\_price\_multiple <- lm(sales\_price ~ bedrooms + year\_built + sq\_ft\_lot + bath\_full\_count + bath\_half\_count + building\_grade, data=housing)

# Execute a summary() function on two variables defined in the previous step to compare the model results. What are the R2 and Adjusted R2 statistics? Explain what these results tell you about the overall model. Did the inclusion of the additional predictors help explain any large variations found in Sale Price?

### Summary of simple regression

summary(sale\_price\_simple)

##   
## Call:  
## lm(formula = sales\_price ~ sq\_ft\_lot, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2016064 -194842 -63293 91565 3735109   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.418e+05 3.800e+03 168.90 <2e-16 \*\*\*  
## sq\_ft\_lot 8.510e-01 6.217e-02 13.69 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 401500 on 12863 degrees of freedom  
## Multiple R-squared: 0.01435, Adjusted R-squared: 0.01428   
## F-statistic: 187.3 on 1 and 12863 DF, p-value: < 2.2e-16

### Summary of multiple regression

summary(sale\_price\_multiple)

##   
## Call:  
## lm(formula = sales\_price ~ bedrooms + year\_built + sq\_ft\_lot +   
## bath\_full\_count + bath\_half\_count + building\_grade, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2181816 -132289 -46921 49526 3706800   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.179e+06 4.377e+05 -14.117 < 2e-16 \*\*\*  
## bedrooms 4.869e+04 4.061e+03 11.991 < 2e-16 \*\*\*  
## year\_built 2.900e+03 2.236e+02 12.974 < 2e-16 \*\*\*  
## sq\_ft\_lot 6.518e-01 5.823e-02 11.194 < 2e-16 \*\*\*  
## bath\_full\_count 4.148e+04 6.110e+03 6.788 1.19e-11 \*\*\*  
## bath\_half\_count 2.680e+04 6.425e+03 4.172 3.04e-05 \*\*\*  
## building\_grade 9.520e+04 3.610e+03 26.367 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 364000 on 12858 degrees of freedom  
## Multiple R-squared: 0.1902, Adjusted R-squared: 0.1898   
## F-statistic: 503.3 on 6 and 12858 DF, p-value: < 2.2e-16

### The inclusion of other predictor help explain large variations found in sales price thats why R squared has increased when we use multiple variables as predicter. Adjusted R squared is the modified version of R squared that depends on the number of attributes added.

# Considering the parameters of the multiple regression model you have created. What are the standardized betas for each parameter and what do the values indicate?

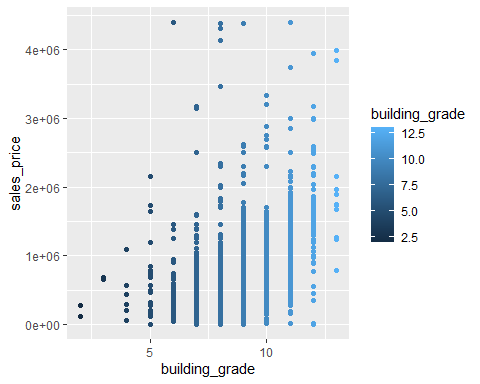
library(lm.beta)  
lm.beta(sale\_price\_multiple)

##   
## Call:  
## lm(formula = sales\_price ~ bedrooms + year\_built + sq\_ft\_lot +   
## bath\_full\_count + bath\_half\_count + building\_grade, data = housing)  
##   
## Standardized Coefficients::  
## (Intercept) bedrooms year\_built sq\_ft\_lot bath\_full\_count   
## NA 0.10548979 0.12350731 0.09177361 0.06675015   
## bath\_half\_count building\_grade   
## 0.03488338 0.25722651

### its gives the relative effect for the attributes on sales price, looking at the values of standardized beta the building grade has the highest impact on sale value, that also justified becase the houses build of stone and houses build of plastic siding must have completely diffent sales values. To do further analysis lets try to plot the building grade vs sales.

### Plot the sales price against the building grade

library(ggplot2)  
ggplot(data = housing, aes(y = sales\_price, x = building\_grade, color = building\_grade)) + geom\_point()

 ### the pattern on the plot also says the same thing the higher the building grade having higher sale prices.

# Calculate the confidence intervals for the parameters in your model and explain what the results indicate.

confint(sale\_price\_multiple)

## 2.5 % 97.5 %  
## (Intercept) -7.036955e+06 -5.321031e+06  
## bedrooms 4.073003e+04 5.664869e+04  
## year\_built 2.462127e+03 3.338515e+03  
## sq\_ft\_lot 5.376966e-01 7.659874e-01  
## bath\_full\_count 2.949929e+04 5.345290e+04  
## bath\_half\_count 1.420964e+04 3.939786e+04  
## building\_grade 8.812267e+04 1.022769e+05

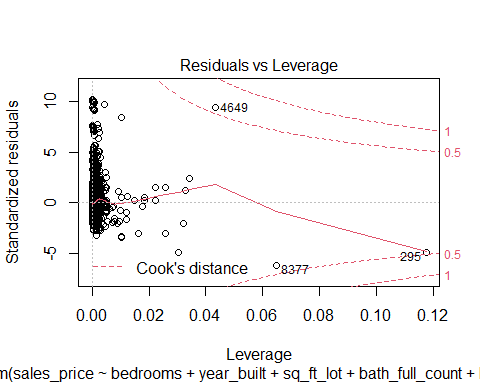
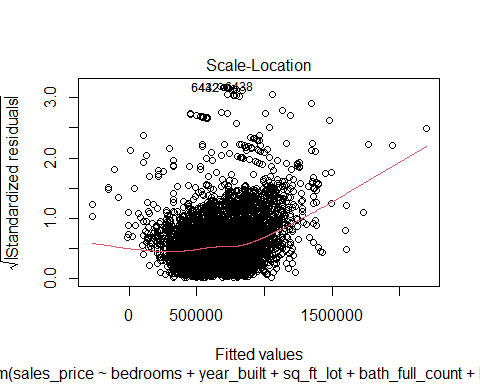
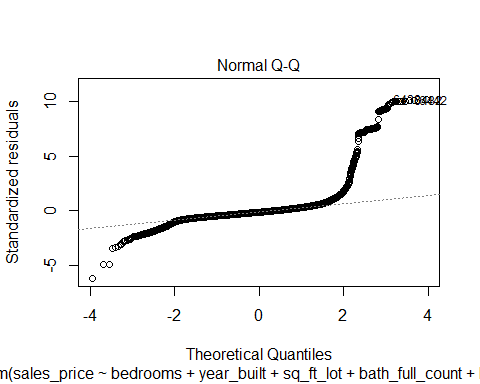
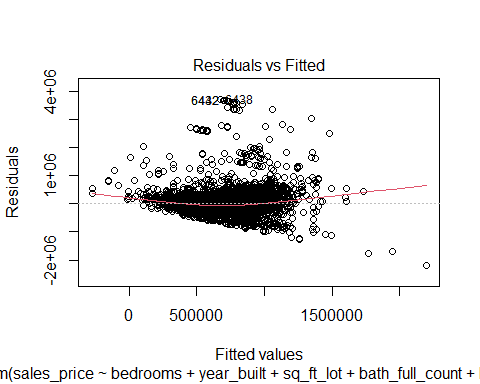
### If the confidence interval for the parameter contains 0, then it can be concluded that there no signifiance evedence of liner relation between predictor and response. Here all are non-zero so the relationship is linear.

# Assess the improvement of the new model compared to your original model (simple regression model) by testing whether this change is significant by performing an analysis of variance.

### for simple regression r-squared was 0.01435 where as for multiple regression it increased to 0.1902, so adding multiple predictor to the model explains more variability.

# Perform casewise diagnostics to identify outliers and/or influential cases, storing each function’s output in a dataframe assigned to a unique variable name.

plot(sale\_price\_multiple)

 ### This plot is used to identify influential observations. If any points in this plot fall outside of Cook’s distance (the dashed lines) then it is an influential observation. the observation are obs 4649, 8377 and 295 there are the influential observation for the model. ### Normal Q-Q plot is used to determine if residuals are normally distributed. ### Residuals vs Fitted plotf is used to determine if the residuals exhibit non-linear patterns. If the red line across the center of the plot is roughly horizontal then we can assume that the residuals follow a linear pattern.

# Calculate the standardized residuals using the appropriate command, specifying those that are +-2, storing the results of large residuals in a variable you create.

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

standard\_res <- rstandard(sale\_price\_multiple)  
data\_with\_res <- cbind(housing, standard\_res)  
  
data\_with\_res\_ = data\_with\_res[,c('sales\_price','bedrooms','year\_built','sq\_ft\_lot','bath\_full\_count','bath\_half\_count','building\_grade','standard\_res')]  
data\_with\_res\_2 = data\_with\_res\_ %>% filter(between(abs(standard\_res),2,3))

# Use the appropriate function to show the sum of large residuals.

sum( data\_with\_res\_2[,c('standard\_res')])

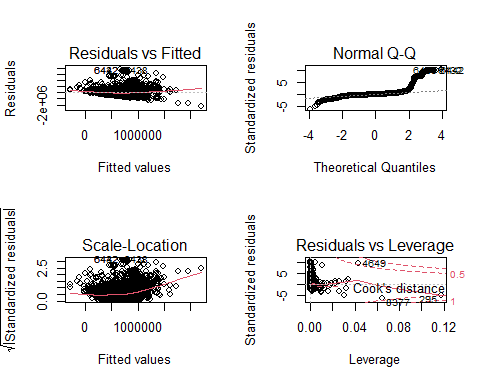
## [1] 26.52065

# Investigate further by calculating the leverage, cooks distance, and covariance rations. Comment on all cases that are problematics.

x <- influence.measures(sale\_price\_multiple)

# Perform the necessary calculations to assess the assumption of independence and state if the condition is met or not.

par(mfrow = c(2, 2))  
plot(sale\_price\_multiple)

 ### as per the residual vs fitted plot it looks like assumption of independence has met.

# Perform the necessary calculations to assess the assumption of no multicollinearity and state if the condition is met or not.

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

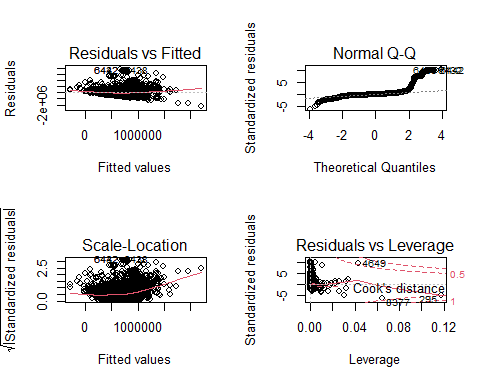
car::vif(sale\_price\_multiple)

## bedrooms year\_built sq\_ft\_lot bath\_full\_count bath\_half\_count   
## 1.228897 1.438916 1.067277 1.535325 1.110172   
## building\_grade   
## 1.511059

### as per the VIF values its not close to 10 so there are not high so there is no multicolinearity in the model.

# Visually check the assumptions related to the residuals using the plot() and hist() functions. Summarize what each graph is informing you of and if any anomalies are present.

par(mfrow = c(2, 2))  
plot(sale\_price\_multiple)

 ### Normal Q-Q plot is used to determine if residuals are normally distributed. ### Residuals vs Fitted plotf is used to determine if the residuals exhibit non-linear patterns. If the red line across the center of the plot is roughly horizontal then we can assume that the residuals follow a linear pattern