---  
title: "ASSIGNMENT 6.1\_Housing\_Data"  
author: "Bhushan Suryawanshi"  
date: '2020-07-06'  
  
---

Work individually on this assignment. You are encouraged to collaborate on ideas and strategies pertinent to this assignment. Data for this assignment is focused on real estate transactions recorded from 1964 to 2016 and can be found in Week 6 Housing.xlsx. Using your skills in statistical correlation, multiple regression and R programming, you are interested in the following variables: Sale Price and several other possible predictors.  
  
Using your ‘clean’ data set from the previous week complete the following:

**a. Explain why you chose to remove data points from your ‘clean’ dataset.**

**Answer** - Ignored this question as its irrelevant here. (Ref Announcement for week.)

library("readxl")

## Warning: package 'readxl' was built under R version 4.0.2

housing\_data <- read\_excel("week-6-housing.xlsx")  
head(housing\_data)

## # A tibble: 6 x 24  
## `Sale Date` `Sale Price` sale\_reason sale\_instrument sale\_warning  
## <dttm> <dbl> <dbl> <dbl> <chr>   
## 1 2006-01-03 00:00:00 698000 1 3 <NA>   
## 2 2006-01-03 00:00:00 649990 1 3 <NA>   
## 3 2006-01-03 00:00:00 572500 1 3 <NA>   
## 4 2006-01-03 00:00:00 420000 1 3 <NA>   
## 5 2006-01-03 00:00:00 369900 1 3 15   
## 6 2006-01-03 00:00:00 184667 1 15 18 51   
## # ... with 19 more variables: sitetype <chr>, addr\_full <chr>, zip5 <dbl>,  
## # ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,  
## # building\_grade <dbl>, square\_feet\_total\_living <dbl>, bedrooms <dbl>,  
## # bath\_full\_count <dbl>, bath\_half\_count <dbl>, bath\_3qtr\_count <dbl>,  
## # year\_built <dbl>, year\_renovated <dbl>, current\_zoning <chr>,  
## # sq\_ft\_lot <dbl>, prop\_type <chr>, present\_use <dbl>

**b. 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.**

#sales\_price\_with\_sq\_ft\_lot <- housing\_data[,c(“Sale Price”,“sq\_ft\_lot”)]

cor(housing\_data$'Sale Price', housing\_data$square\_feet\_total\_living)^2 \* 100

## [1] 20.66499

cor.test(housing\_data$'Sale Price', housing\_data$square\_feet\_total\_living)

##   
## Pearson's product-moment correlation  
##   
## data: housing\_data$"Sale Price" and housing\_data$square\_feet\_total\_living  
## t = 57.884, df = 12863, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4407697 0.4681900  
## sample estimates:  
## cor   
## 0.4545876

cor(housing\_data$'Sale Price', housing\_data$building\_grade)^2 \* 100

## [1] 15.30602

cor.test(housing\_data$'Sale Price', housing\_data$building\_grade)

##   
## Pearson's product-moment correlation  
##   
## data: housing\_data$"Sale Price" and housing\_data$building\_grade  
## t = 48.214, df = 12863, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3764941 0.4057662  
## sample estimates:  
## cor   
## 0.3912291

cor(housing\_data$'Sale Price', housing\_data$year\_built)^2 \* 100

## [1] 5.888934

cor.test(housing\_data$'Sale Price', housing\_data$year\_built)

##   
## Pearson's product-moment correlation  
##   
## data: housing\_data$"Sale Price" and housing\_data$year\_built  
## t = 28.371, df = 12863, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2263401 0.2588660  
## sample estimates:  
## cor   
## 0.2426713

cor(housing\_data$'Sale Price', housing\_data$bedrooms)^2 \* 100

## [1] 5.083558

cor.test(housing\_data$'Sale Price', housing\_data$bedrooms)

##   
## Pearson's product-moment correlation  
##   
## data: housing\_data$"Sale Price" and housing\_data$bedrooms  
## t = 26.247, df = 12863, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2090015 0.2418056  
## sample estimates:  
## cor   
## 0.2254675

**Answer:**  
Looking at the correlation we can see all the variables shown above are positively correlated. However, building grade and square feet of living room share 20% and 15% variation in determining sales price. Hence building grade and square feet of living are chosen as predictors for the model over year\_built and bedrooms.

sales\_price\_with\_sq\_ft\_lot <- lm(housing\_data$'Sale Price' ~ housing\_data$sq\_ft\_lot, data = housing\_data)  
sales\_price\_with\_others <- lm(housing\_data$'Sale Price' ~ housing\_data$sq\_ft\_lot + housing\_data$square\_feet\_total\_living + housing\_data$building\_grade, data = housing\_data)

**c. 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(sales\_price\_with\_sq\_ft\_lot)

##   
## Call:  
## lm(formula = housing\_data$"Sale Price" ~ housing\_data$sq\_ft\_lot,   
## data = housing\_data)  
##   
## 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 \*\*\*  
## housing\_data$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(sales\_price\_with\_others)

##   
## Call:  
## lm(formula = housing\_data$"Sale Price" ~ housing\_data$sq\_ft\_lot +   
## housing\_data$square\_feet\_total\_living + housing\_data$building\_grade,   
## data = housing\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1943318 -115934 -42965 39043 3747512   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -8.142e+04 2.810e+04 -2.898 0.00377  
## housing\_data$sq\_ft\_lot 1.330e-01 5.723e-02 2.324 0.02016  
## housing\_data$square\_feet\_total\_living 1.475e+02 4.889e+00 30.178 < 2e-16  
## housing\_data$building\_grade 4.424e+04 4.348e+03 10.175 < 2e-16  
##   
## (Intercept) \*\*   
## housing\_data$sq\_ft\_lot \*   
## housing\_data$square\_feet\_total\_living \*\*\*  
## housing\_data$building\_grade \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 358700 on 12861 degrees of freedom  
## Multiple R-squared: 0.2132, Adjusted R-squared: 0.213   
## F-statistic: 1161 on 3 and 12861 DF, p-value: < 2.2e-16

**Answer:**

For the first model R-Squared value is 0.014, which means that sq\_ft\_lot accounts for 1.4% of the variation in sale price. However, when the other two predictors are included as well, this value increases to 0.213, or 21.3% of the variance in sale price. Therefore, if sq\_ft\_lot accounts for 1.4%, we can tell that square\_fee\_total\_living and building\_grade account for an additional 20.0%. So the inclusion of the two new predictors has explained quite a large amount of the variation in sale price. R-squared and adjusted r-squared for the model with additional predictors shows difference of 0% (0.213 - 0.213). Which means if the model were derived from the population rather than a sample it would account for 0% variance in the outcome.

**d. 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("QuantPsyc")

## Warning: package 'QuantPsyc' was built under R version 4.0.2

## Loading required package: boot

## Loading required package: MASS

##   
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':  
##   
## norm

lm.beta(sales\_price\_with\_others)

## housing\_data$sq\_ft\_lot housing\_data$square\_feet\_total\_living   
## 0.01872256 0.36113656   
## housing\_data$building\_grade   
## 0.11952538

**Answer:**

The standardized beta estimates tell us the number of standard deviations by which the outcome will change as a result of one standard deviation change in the predictor.

In this case, 1 standard deviation of change in Sq\_ft\_lot causes sales price to change by 0.019 standard deviation. 1 standard deviation change in square\_feet\_total\_living causes sales price to change by 0.361 standard deviation and 1 standard deviation change in building\_grade can cause 0.120 standard deviation change in sale price.

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

confint(sales\_price\_with\_others, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) -1.365003e+05 -2.634307e+04  
## housing\_data$sq\_ft\_lot 2.079828e-02 2.451639e-01  
## housing\_data$square\_feet\_total\_living 1.379562e+02 1.571221e+02  
## housing\_data$building\_grade 3.571437e+04 5.275855e+04

**Answer:**

The confidence interval shows that there is positive relation between all predictors and outcome. Also the 95% confidence interval range is not very big which means this sample if close to the beta of population.

**f. 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.**

anova(sales\_price\_with\_sq\_ft\_lot, sales\_price\_with\_others)

## Analysis of Variance Table  
##   
## Model 1: housing\_data$"Sale Price" ~ housing\_data$sq\_ft\_lot  
## Model 2: housing\_data$"Sale Price" ~ housing\_data$sq\_ft\_lot + housing\_data$square\_feet\_total\_living +   
## housing\_data$building\_grade  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 12863 2.0734e+15   
## 2 12861 1.6551e+15 2 4.1823e+14 1624.9 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Answer:** The value in column labelled Pr(>F) is 2.2e−16 (i.e., 2.2 with the decimal place moved 16 places to the left, or a very small value indeed); we can say that sales\_price\_with\_others significantly improved the fit of the model to the data compared to sales\_price\_with\_sq\_ft\_lot.

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

housing\_data$standardized\_residuals<- rstandard(sales\_price\_with\_others)  
head(housing\_data$standardized\_residuals)

## 1 2 3 4 5 6   
## -0.09526631 -0.25749989 -0.30602206 -0.25859457 -0.20013735 -1.83622727

housing\_data$studentized\_residuals<-rstudent(sales\_price\_with\_others)  
head(housing\_data$studentized\_residuals)

## 1 2 3 4 5 6   
## -0.09526264 -0.25749054 -0.30601128 -0.25858519 -0.20012988 -1.83639662

housing\_data$cooks\_distance<-cooks.distance(sales\_price\_with\_others)  
head(housing\_data$cooks\_distance)

## 1 2 3 4 5 6   
## 3.020345e-07 2.155168e-06 2.761224e-06 3.074759e-06 1.905215e-06 1.108648e-03

housing\_data$dfbeta<-dfbeta(sales\_price\_with\_others)  
head(housing\_data$dfbeta)

## (Intercept) housing\_data$sq\_ft\_lot housing\_data$square\_feet\_total\_living  
## 1 15.53348 1.476123e-05 0.001227376  
## 2 38.42380 4.519048e-05 0.002074116  
## 3 -46.69060 5.226207e-05 -0.008579373  
## 4 18.77844 -4.558569e-06 0.012611383  
## 5 -43.25320 2.403663e-06 0.003317644  
## 6 -1375.07195 7.990754e-04 -0.302595677  
## housing\_data$building\_grade  
## 1 -2.625498  
## 2 -6.295360  
## 3 7.133413  
## 4 -7.028207  
## 5 3.542687  
## 6 251.748864

housing\_data$dffit<-dffits(sales\_price\_with\_others)  
head(housing\_data$dffit)

## 1 2 3 4 5 6   
## -0.001099111 -0.002935992 -0.003323269 -0.003506871 -0.002760488 -0.066598878

housing\_data$leverage<-hatvalues(sales\_price\_with\_others)  
head(housing\_data$leverage)

## 1 2 3 4 5 6   
## 0.0001331006 0.0001299962 0.0001179247 0.0001838876 0.0001902239 0.0013134999

housing\_data$covariance\_ratios<-covratio(sales\_price\_with\_others)  
head(housing\_data$covariance\_ratios)

## 1 2 3 4 5 6   
## 1.000441 1.000421 1.000400 1.000474 1.000489 1.000577

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

housing\_data$large\_residual <- housing\_data$studentized\_residuals > 2 | housing\_data$studentized\_residuals < -2

**i. Use the appropriate function to show the sum of large residuals.**

sum(housing\_data$large\_residual)

## [1] 317

**j. Which specific variables have large residuals (only cases that evaluate as TRUE)?**

housing\_data[housing\_data$large\_residual, c('Sale Price', 'sq\_ft\_lot', 'square\_feet\_total\_living', 'building\_grade', "standardized\_residuals")]

## # A tibble: 317 x 5  
## `Sale Price` sq\_ft\_lot square\_feet\_total\_l~ building\_grade standardized\_resi~  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 265000 112650 4920 10 -2.33  
## 2 1390000 225640 660 6 3.01  
## 3 229000 236966 3840 10 -2.04  
## 4 390000 63162 5800 11 -2.45  
## 5 1588359 8752 3360 9 2.16  
## 6 1450000 14043 900 6 3.15  
## 7 163000 18498 4710 9 -2.37  
## 8 270000 89734 5060 11 -2.49  
## 9 200000 288367 6880 10 -3.39  
## 10 300000 55303 4490 11 -2.16  
## # ... with 307 more rows

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

housing\_data[housing\_data$large\_residual, c("cooks\_distance", "leverage", "covariance\_ratios")]

## # A tibble: 317 x 3  
## cooks\_distance leverage covariance\_ratios  
## <dbl> <dbl> <dbl>  
## 1 0.000841 0.000617 0.999  
## 2 0.00391 0.00173 0.999  
## 3 0.00135 0.00130 1.00   
## 4 0.00139 0.000922 0.999  
## 5 0.000172 0.000147 0.999  
## 6 0.00102 0.000410 0.998  
## 7 0.000903 0.000641 0.999  
## 8 0.00106 0.000685 0.999  
## 9 0.00885 0.00307 1.00   
## 10 0.000676 0.000578 0.999  
## # ... with 307 more rows

cook\_dist <- housing\_data$cooks\_distance > 1  
  
sum(cook\_dist)

## [1] 1

**Answer:**

Out of 317 large residuals only 1 value has cook distance greater than 1. Hence only 1 influential observation in the model. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 424) )

k <- 3  
sample\_size <- 12865  
leverage <- k/sample\_size  
leverage

## [1] 0.0002331908

leverage\_data <- housing\_data$cooks\_distance > 0.0006 | housing\_data$cooks\_distance < 0.0009  
sum(leverage\_data)

## [1] 12865

Leverage we have is 0.0003 which means we have to look for values between 0.0006 (twice of leverage) and 0.0009 (thrice of leverage). We can see that all the observations are within the required range and hence there is no problem with the observation used in the model. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 424) )

#Covariance ratio boundary  
  
#CVRi > 1 + [3(k + 1)/n] = 1 + [3(3 + 1)/sample\_size] = 1.0009;  
  
#CVRi < 1 – [3(k + 1)/n] = 1 – [3(3 + 1)/sample\_size] = 0.9991  
  
cvr\_min <- 1 + 12/12865  
  
cvr\_min

## [1] 1.000933

cvr\_max <- 1 - 12/12865  
  
cvr\_max

## [1] 0.9990672

cvr\_data <- housing\_data$cooks\_distance > cvr\_min | housing\_data$cooks\_distance < cvr\_max  
  
sum(cvr\_data)

## [1] 12865

**Answer:**

From the covariance data above we can see that all the observation in the dataset lies between cvr\_min, cvr\_mx and hence there is no outlier as per the covarince ratio calculation. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 425) )

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

#Need to import car library as Durbin Watson Test was not working.  
  
library("car")

## Warning: package 'car' was built under R version 4.0.2

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:boot':  
##   
## logit

dwt(sales\_price\_with\_others)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.7402678 0.5194617 0  
## Alternative hypothesis: rho != 0

**Answer:** The Durbin Watson test reports a test statistic, with a value from 0 to 4, where:

* 2 is no autocorrelation.
* ‘0 to <2’ is positive autocorrelation (common in time series data).
* ‘>2 to 4’ is negative autocorrelation (less common in time series data).

Based on the Durbin watson test we can see the DWT-statistic value for the model is 0.52. Which means there is positive autocorrelation and the condition is met. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 426) )

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

vif(sales\_price\_with\_others)

## housing\_data$sq\_ft\_lot housing\_data$square\_feet\_total\_living   
## 1.061258 2.340696   
## housing\_data$building\_grade   
## 2.255639

mean(vif(sales\_price\_with\_others))

## [1] 1.885864

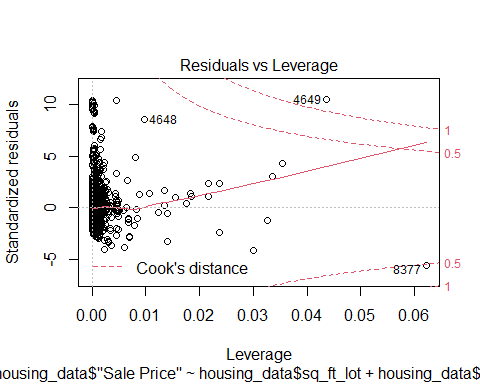
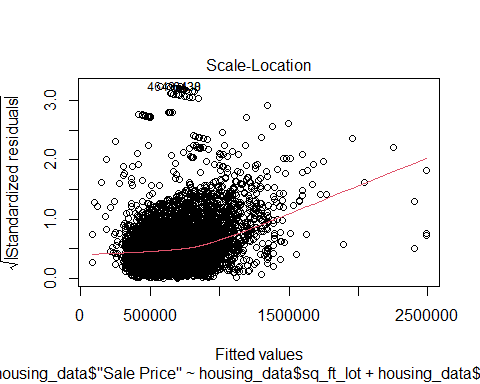
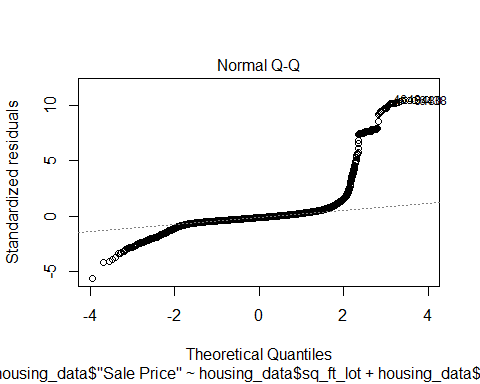
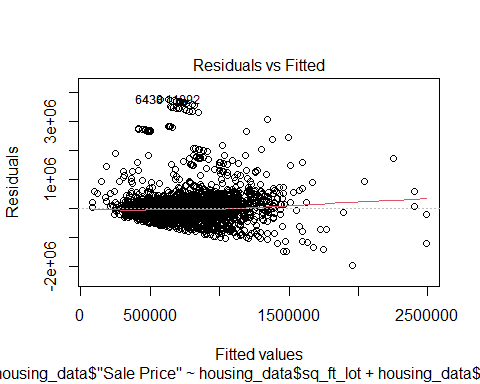
#Tolerence Statistics  
1/vif(sales\_price\_with\_others)

## housing\_data$sq\_ft\_lot housing\_data$square\_feet\_total\_living   
## 0.9422776 0.4272234   
## housing\_data$building\_grade   
## 0.4433334

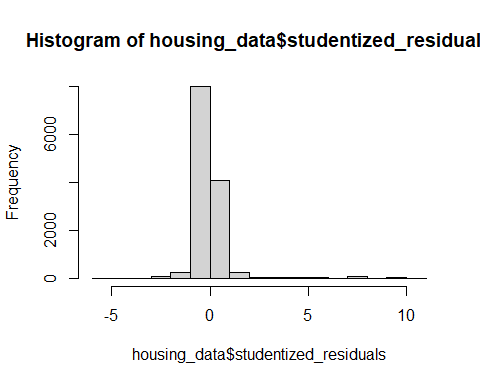
**Answer:** VIF values are all well below 10 and the tolerance statistics all well above 0.2. Also, the average VIF is very close to 1. Based on these measures we can safely conclude that there is no collinearity within our data.(Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 428) )

**n. 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.**

plot(sales\_price\_with\_others)



hist(housing\_data$studentized\_residuals)



Residuals vs fitted plot shows a fairly random pattern, which means the assumptions of linearity, randomness and homoscedasticity have been met. The Q-Q plot shows less normality but this can happen due to smaller sample than the population.

(Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 428) )

**0. Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us about the sample vs. the entire population model?**

**Answer:** Based on the above histogram which looks like a normal distribution, we can say that the model is both accurate and generalizable to the population.

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using R*. SAGE Publications. <https://books.google.com/books?id=wd2K2zC3swIC>.