RMarkdown Assignment - Exercise 11 and 12

Pushkar Chougule

Oct 18th 2020

## R Markdown

**Reading the xlsx file into data frame. Using summary() and str() functions to understand data better and to understand the outliers**

library(readxl)  
  
housing\_df1 <- read\_excel("week-6-housing.xlsx")  
  
summary(housing\_df1)

## Sale Date Sale Price sale\_reason   
## Min. :2006-01-03 00:00:00 Min. : 698 Min. : 0.00   
## 1st Qu.:2008-07-07 00:00:00 1st Qu.: 460000 1st Qu.: 1.00   
## Median :2011-11-17 00:00:00 Median : 593000 Median : 1.00   
## Mean :2011-07-28 15:07:32 Mean : 660738 Mean : 1.55   
## 3rd Qu.:2014-06-05 00:00:00 3rd Qu.: 750000 3rd Qu.: 1.00   
## Max. :2016-12-16 00:00:00 Max. :4400000 Max. :19.00   
## sale\_instrument sale\_warning sitetype addr\_full   
## Min. : 0.000 Length:12865 Length:12865 Length:12865   
## 1st Qu.: 3.000 Class :character Class :character Class :character   
## Median : 3.000 Mode :character Mode :character Mode :character   
## Mean : 3.678   
## 3rd Qu.: 3.000   
## Max. :27.000   
## zip5 ctyname postalctyn lon   
## Min. :98052 Length:12865 Length:12865 Min. :-122.2   
## 1st Qu.:98052 Class :character Class :character 1st Qu.:-122.1   
## Median :98052 Mode :character Mode :character Median :-122.1   
## Mean :98053 Mean :-122.1   
## 3rd Qu.:98053 3rd Qu.:-122.0   
## Max. :98074 Max. :-121.9   
## lat building\_grade square\_feet\_total\_living bedrooms   
## Min. :47.46 Min. : 2.00 Min. : 240 Min. : 0.000   
## 1st Qu.:47.67 1st Qu.: 8.00 1st Qu.: 1820 1st Qu.: 3.000   
## Median :47.69 Median : 8.00 Median : 2420 Median : 4.000   
## Mean :47.68 Mean : 8.24 Mean : 2540 Mean : 3.479   
## 3rd Qu.:47.70 3rd Qu.: 9.00 3rd Qu.: 3110 3rd Qu.: 4.000   
## Max. :47.73 Max. :13.00 Max. :13540 Max. :11.000   
## bath\_full\_count bath\_half\_count bath\_3qtr\_count year\_built   
## Min. : 0.000 Min. :0.0000 Min. :0.000 Min. :1900   
## 1st Qu.: 1.000 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:1979   
## Median : 2.000 Median :1.0000 Median :0.000 Median :1998   
## Mean : 1.798 Mean :0.6134 Mean :0.494 Mean :1993   
## 3rd Qu.: 2.000 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:2007   
## Max. :23.000 Max. :8.0000 Max. :8.000 Max. :2016   
## year\_renovated current\_zoning sq\_ft\_lot prop\_type   
## Min. : 0.00 Length:12865 Min. : 785 Length:12865   
## 1st Qu.: 0.00 Class :character 1st Qu.: 5355 Class :character   
## Median : 0.00 Mode :character Median : 7965 Mode :character   
## Mean : 26.24 Mean : 22229   
## 3rd Qu.: 0.00 3rd Qu.: 12632   
## Max. :2016.00 Max. :1631322   
## present\_use   
## Min. : 0.000   
## 1st Qu.: 2.000   
## Median : 2.000   
## Mean : 6.598   
## 3rd Qu.: 2.000   
## Max. :300.000

str(housing\_df1)

## tibble [12,865 x 24] (S3: tbl\_df/tbl/data.frame)  
## $ Sale Date : POSIXct[1:12865], format: "2006-01-03" "2006-01-03" ...  
## $ Sale Price : num [1:12865] 698000 649990 572500 420000 369900 ...  
## $ sale\_reason : num [1:12865] 1 1 1 1 1 1 1 1 1 1 ...  
## $ sale\_instrument : num [1:12865] 3 3 3 3 3 15 3 3 3 3 ...  
## $ sale\_warning : chr [1:12865] NA NA NA NA ...  
## $ sitetype : chr [1:12865] "R1" "R1" "R1" "R1" ...  
## $ addr\_full : chr [1:12865] "17021 NE 113TH CT" "11927 178TH PL NE" "13315 174TH AVE NE" "3303 178TH AVE NE" ...  
## $ zip5 : num [1:12865] 98052 98052 98052 98052 98052 ...  
## $ ctyname : chr [1:12865] "REDMOND" "REDMOND" NA "REDMOND" ...  
## $ postalctyn : chr [1:12865] "REDMOND" "REDMOND" "REDMOND" "REDMOND" ...  
## $ lon : num [1:12865] -122 -122 -122 -122 -122 ...  
## $ lat : num [1:12865] 47.7 47.7 47.7 47.6 47.7 ...  
## $ building\_grade : num [1:12865] 9 9 8 8 7 7 10 10 9 8 ...  
## $ square\_feet\_total\_living: num [1:12865] 2810 2880 2770 1620 1440 4160 3960 3720 4160 2760 ...  
## $ bedrooms : num [1:12865] 4 4 4 3 3 4 5 4 4 4 ...  
## $ bath\_full\_count : num [1:12865] 2 2 1 1 1 2 3 2 2 1 ...  
## $ bath\_half\_count : num [1:12865] 1 0 1 0 0 1 0 1 1 0 ...  
## $ bath\_3qtr\_count : num [1:12865] 0 1 1 1 1 1 1 0 1 1 ...  
## $ year\_built : num [1:12865] 2003 2006 1987 1968 1980 ...  
## $ year\_renovated : num [1:12865] 0 0 0 0 0 0 0 0 0 0 ...  
## $ current\_zoning : chr [1:12865] "R4" "R4" "R6" "R4" ...  
## $ sq\_ft\_lot : num [1:12865] 6635 5570 8444 9600 7526 ...  
## $ prop\_type : chr [1:12865] "R" "R" "R" "R" ...  
## $ present\_use : num [1:12865] 2 2 2 2 2 2 2 2 2 2 ...

**a. Data clean up task.**

**Answer a.**

understanding above data and manually looking at excel data, determined possible outliers values and dropped them from dataframe

housing\_df <- housing\_df1[housing\_df1$`Sale Price` > 50000 & housing\_df1$`Sale Price` < 3000000 & housing\_df1$square\_feet\_total\_living < 8000, ]  
  
summary(housing\_df)

## Sale Date Sale Price sale\_reason   
## Min. :2006-01-03 00:00:00 Min. : 53502 Min. : 0.000   
## 1st Qu.:2008-07-10 00:00:00 1st Qu.: 460000 1st Qu.: 1.000   
## Median :2011-11-28 00:00:00 Median : 590000 Median : 1.000   
## Mean :2011-08-04 06:20:00 Mean : 633875 Mean : 1.527   
## 3rd Qu.:2014-06-16 00:00:00 3rd Qu.: 745000 3rd Qu.: 1.000   
## Max. :2016-12-16 00:00:00 Max. :2988000 Max. :19.000   
## sale\_instrument sale\_warning sitetype addr\_full   
## Min. : 0.000 Length:12672 Length:12672 Length:12672   
## 1st Qu.: 3.000 Class :character Class :character Class :character   
## Median : 3.000 Mode :character Mode :character Mode :character   
## Mean : 3.611   
## 3rd Qu.: 3.000   
## Max. :27.000   
## zip5 ctyname postalctyn lon   
## Min. :98052 Length:12672 Length:12672 Min. :-122.2   
## 1st Qu.:98052 Class :character Class :character 1st Qu.:-122.1   
## Median :98052 Mode :character Mode :character Median :-122.1   
## Mean :98053 Mean :-122.1   
## 3rd Qu.:98053 3rd Qu.:-122.0   
## Max. :98074 Max. :-121.9   
## lat building\_grade square\_feet\_total\_living bedrooms   
## Min. :47.46 Min. : 2.000 Min. : 240 Min. : 0.000   
## 1st Qu.:47.67 1st Qu.: 8.000 1st Qu.:1830 1st Qu.: 3.000   
## Median :47.69 Median : 8.000 Median :2420 Median : 4.000   
## Mean :47.68 Mean : 8.244 Mean :2532 Mean : 3.478   
## 3rd Qu.:47.70 3rd Qu.: 9.000 3rd Qu.:3110 3rd Qu.: 4.000   
## Max. :47.73 Max. :13.000 Max. :7980 Max. :11.000   
## bath\_full\_count bath\_half\_count bath\_3qtr\_count year\_built   
## Min. : 0.000 Min. :0.0000 Min. :0.0000 Min. :1900   
## 1st Qu.: 1.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:1979   
## Median : 2.000 Median :1.0000 Median :0.0000 Median :1998   
## Mean : 1.796 Mean :0.6117 Mean :0.4942 Mean :1993   
## 3rd Qu.: 2.000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:2007   
## Max. :23.000 Max. :6.0000 Max. :7.0000 Max. :2016   
## year\_renovated current\_zoning sq\_ft\_lot prop\_type   
## Min. : 0.0 Length:12672 Min. : 785 Length:12672   
## 1st Qu.: 0.0 Class :character 1st Qu.: 5400 Class :character   
## Median : 0.0 Mode :character Median : 7998 Mode :character   
## Mean : 25.7 Mean : 21533   
## 3rd Qu.: 0.0 3rd Qu.: 12600   
## Max. :2016.0 Max. :1166246   
## present\_use   
## Min. : 0.000   
## 1st Qu.: 2.000   
## Median : 2.000   
## Mean : 6.501   
## 3rd Qu.: 2.000   
## Max. :300.000

str(housing\_df)

## tibble [12,672 x 24] (S3: tbl\_df/tbl/data.frame)  
## $ Sale Date : POSIXct[1:12672], format: "2006-01-03" "2006-01-03" ...  
## $ Sale Price : num [1:12672] 698000 649990 572500 420000 369900 ...  
## $ sale\_reason : num [1:12672] 1 1 1 1 1 1 1 1 1 1 ...  
## $ sale\_instrument : num [1:12672] 3 3 3 3 3 15 3 3 3 3 ...  
## $ sale\_warning : chr [1:12672] NA NA NA NA ...  
## $ sitetype : chr [1:12672] "R1" "R1" "R1" "R1" ...  
## $ addr\_full : chr [1:12672] "17021 NE 113TH CT" "11927 178TH PL NE" "13315 174TH AVE NE" "3303 178TH AVE NE" ...  
## $ zip5 : num [1:12672] 98052 98052 98052 98052 98052 ...  
## $ ctyname : chr [1:12672] "REDMOND" "REDMOND" NA "REDMOND" ...  
## $ postalctyn : chr [1:12672] "REDMOND" "REDMOND" "REDMOND" "REDMOND" ...  
## $ lon : num [1:12672] -122 -122 -122 -122 -122 ...  
## $ lat : num [1:12672] 47.7 47.7 47.7 47.6 47.7 ...  
## $ building\_grade : num [1:12672] 9 9 8 8 7 7 10 10 9 8 ...  
## $ square\_feet\_total\_living: num [1:12672] 2810 2880 2770 1620 1440 4160 3960 3720 4160 2760 ...  
## $ bedrooms : num [1:12672] 4 4 4 3 3 4 5 4 4 4 ...  
## $ bath\_full\_count : num [1:12672] 2 2 1 1 1 2 3 2 2 1 ...  
## $ bath\_half\_count : num [1:12672] 1 0 1 0 0 1 0 1 1 0 ...  
## $ bath\_3qtr\_count : num [1:12672] 0 1 1 1 1 1 1 0 1 1 ...  
## $ year\_built : num [1:12672] 2003 2006 1987 1968 1980 ...  
## $ year\_renovated : num [1:12672] 0 0 0 0 0 0 0 0 0 0 ...  
## $ current\_zoning : chr [1:12672] "R4" "R4" "R6" "R4" ...  
## $ sq\_ft\_lot : num [1:12672] 6635 5570 8444 9600 7526 ...  
## $ prop\_type : chr [1:12672] "R" "R" "R" "R" ...  
## $ present\_use : num [1:12672] 2 2 2 2 2 2 2 2 2 2 ...

**calculate R2 values for select few variables, to understand the potential predictor variables, which can have most impacts**

cor(housing\_df$"Sale Price", housing\_df$sq\_ft\_lot)^2

## [1] 0.03459774

cor(housing\_df$"Sale Price", housing\_df$square\_feet\_total\_living)^2

## [1] 0.444824

cor(housing\_df$"Sale Price", housing\_df$building\_grade)^2

## [1] 0.3697124

cor(housing\_df$"Sale Price", housing\_df$bedrooms)^2

## [1] 0.1096689

cor(housing\_df$"Sale Price", housing\_df$bath\_full\_count)^2

## [1] 0.147803

cor(housing\_df$"Sale Price", housing\_df$year\_built)^2

## [1] 0.0686501

#cor(housing\_df$"Sale Price", housing\_df$current\_zoning)^2  
#cor(housing\_df$"Sale Price", housing\_df$sitetype)^2

**b. Create two models : simple linear regression and multiple regression**

housing\_mod1 <- lm(`Sale Price` ~ sq\_ft\_lot, data = housing\_df)  
  
housing\_mod2 <- lm(`Sale Price` ~ sq\_ft\_lot + square\_feet\_total\_living + building\_grade, data = housing\_df)

**Answer b.**  > For Multiple regression model, considered only few numeric variables, after looking at above R-squared (R2) values. square feet total living has 44.48% impact, building grade has 36.97% impact and used in addition to square feet lot variable.

**c. execute summary() function. explain R2 and adjusted R2 values**

summary(housing\_mod1)

##   
## Call:  
## lm(formula = `Sale Price` ~ sq\_ft\_lot, data = housing\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1257168 -166914 -38020 113408 2267970   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.120e+05 2.647e+03 231.16 <2e-16 \*\*\*  
## sq\_ft\_lot 1.017e+00 4.771e-02 21.31 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 274700 on 12670 degrees of freedom  
## Multiple R-squared: 0.0346, Adjusted R-squared: 0.03452   
## F-statistic: 454.1 on 1 and 12670 DF, p-value: < 2.2e-16

summary(housing\_mod2)

##   
## Call:  
## lm(formula = `Sale Price` ~ sq\_ft\_lot + square\_feet\_total\_living +   
## building\_grade, data = housing\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1219650 -87492 -16387 64705 2068722   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.550e+05 1.610e+04 -15.841 < 2e-16 \*\*\*  
## sq\_ft\_lot 2.561e-01 3.616e-02 7.083 1.48e-12 \*\*\*  
## square\_feet\_total\_living 1.390e+02 2.929e+00 47.450 < 2e-16 \*\*\*  
## building\_grade 6.448e+04 2.515e+03 25.642 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 202800 on 12668 degrees of freedom  
## Multiple R-squared: 0.4737, Adjusted R-squared: 0.4735   
## F-statistic: 3800 on 3 and 12668 DF, p-value: < 2.2e-16

**Answer c.**

Simple linear regression model has R2 value of 0.0346 and adjusted R2 value of 0.03452, which is roughly about 3.46% of an impact on Sale Price. The difference between these two values is 0.00008 i.e. very minimal. SO, if the model were derived from the population rather than a sample and it would account for 0.0080% of variance, which is very small. Multiple linear regression model has R2 value of 0.4737 and adjusted R2 value of 0.4735, which is roughly about 47.37% of an impact on Sale Price. The difference between these two values is 0.0002 i.e. very minimal. SO, if the model were derived from the population rather than a sample and it would account for 0.02% of variance, which is very small. We can infer that the second model increased R2 value from 3.46% to 47.37%, which is very significant increase

**d. standardized betas function.**

library(QuantPsyc)

## Loading required package: boot

## Loading required package: MASS

##   
## Attaching package: 'QuantPsyc'

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

lm.beta(housing\_mod2)

## sq\_ft\_lot square\_feet\_total\_living building\_grade   
## 0.04686682 0.46939216 0.24978199

**Answer d.**

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.

With our cleaned up sample data 1 standard deviation of change in Square feet lot causes sales price to change by 0.04686682 standard deviation. 1 standard deviation change in square feet total living causes sales price to change by 0.46939216 standard deviation and 1 standard deviation change in building\_grade can cause 0.24978199 standard deviation change in sale price.

**e. confidence interval**

confint(housing\_mod2, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) -2.865651e+05 -2.234570e+05  
## sq\_ft\_lot 1.852629e-01 3.270240e-01  
## square\_feet\_total\_living 1.332166e+02 1.446973e+02  
## building\_grade 5.954845e+04 6.940630e+04

**Answer e.**

The confidence interval shows that there is positive relation between all the 3 predictors and outcome variable. Also the 95% confidence interval range for the 3 predictor variables is not very large / wide which indicates this cleaned up sample is fairly close representation of the beta values of the population. So, our model is closer to the real data.

**f. analysis of variance**

anova(housing\_mod1, housing\_mod2)

## Analysis of Variance Table  
##   
## Model 1: `Sale Price` ~ sq\_ft\_lot  
## Model 2: `Sale Price` ~ sq\_ft\_lot + square\_feet\_total\_living + building\_grade  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 12670 9.5589e+14   
## 2 12668 5.2116e+14 2 4.3473e+14 5283.6 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Answer f.**

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 really very small value); we can say that housing\_mod2 has significantly improved the fit of the model to the data compared to housing\_mod1.

**g. casewise diagnostics**

**Answer g.**

casewise diagnostics performed using the formuale given in (Field, Miles, and Field 2012) book, page 289

housing\_df$standardized.residuals <- rstandard(housing\_mod2)  
head(housing\_df$standardized.residuals)

## 1 2 3 4 5 6   
## -0.09591938 -0.37925188 -0.37167447 -0.33715111 -0.14031129 -2.91879468

housing\_df$studentized.residuals <- rstudent(housing\_mod2)  
head(housing\_df$studentized.residuals)

## 1 2 3 4 5 6   
## -0.09591563 -0.37923906 -0.37166183 -0.33713932 -0.14030586 -2.91966139

housing\_df$cooks.distance <- cooks.distance(housing\_mod2)  
head(housing\_df$cooks.distance)

## 1 2 3 4 5 6   
## 3.121938e-07 4.781441e-06 4.291748e-06 5.604655e-06 9.790214e-07 3.070981e-03

housing\_df$dfbeta <- dfbeta(housing\_mod2)  
head(housing\_df$dfbeta)

## (Intercept) sq\_ft\_lot square\_feet\_total\_living building\_grade  
## 1 8.956452 1.037009e-05 0.0006971453 -1.513854  
## 2 32.268057 4.608331e-05 0.0015961445 -5.261074  
## 3 -33.766324 4.129145e-05 -0.0066463149 5.307528  
## 4 14.996795 -1.238558e-06 0.0104027413 -5.665371  
## 5 -17.327262 1.582812e-06 0.0015801317 1.339939  
## 6 -1295.144267 7.789333e-04 -0.3031042669 242.483553

housing\_df$dffit <- dffits(housing\_mod2)  
head(housing\_df$dffit)

## 1 2 3 4 5 6   
## -0.001117442 -0.004373154 -0.004143166 -0.004734665 -0.001978834 -0.110865779

housing\_df$leverage <- hatvalues(housing\_mod2)  
head(housing\_df$leverage)

## 1 2 3 4 5 6   
## 0.0001357102 0.0001329553 0.0001242553 0.0001971851 0.0001988752 0.0014398071

housing\_df$covariance.ratios <- covratio(housing\_mod2)  
head(housing\_df$covariance.ratios)

## 1 2 3 4 5 6   
## 1.0004487 1.0004034 1.0003965 1.0004772 1.0005086 0.9990661

**h. large standardized residuals**

**Answer h.**

large standardized residuals calculated using the formula given in (Field, Miles, and Field 2012) book

housing\_df$large\_residual <- housing\_df$standardized.residuals > 2 | housing\_df$standardized.residuals < -2

**i. sum of large standardized residuals**

**Answer i.**

sum function to calculate sum of large standardized residuals

sum(housing\_df$large\_residual)

## [1] 514

**j. specific variables with large standardized residuals**

**Answer j.**

where large standardized residuals is TRUE

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

## # A tibble: 514 x 5  
## `Sale Price` sq\_ft\_lot square\_feet\_total\_l~ building\_grade standardized.resi~  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 184667 7280 4160 7 -2.92  
## 2 165000 278891 1850 9 -2.41  
## 3 265000 112650 4920 10 -4.13  
## 4 1392000 17291 3740 9 2.68  
## 5 148000 3430 1930 9 -2.20  
## 6 1900000 37017 6610 11 2.55  
## 7 1520000 19173 4640 9 2.69  
## 8 1390000 225640 660 6 5.47  
## 9 1390000 225640 3280 10 2.40  
## 10 229000 236966 3840 10 -3.73  
## # ... with 504 more rows

**k. calculate leverage, cooks distance and covariance ratios and problematic cases**

large\_residuals <- housing\_df[housing\_df$large\_residual, c("cooks.distance", "leverage", "covariance.ratios")]  
head(large\_residuals)

## # A tibble: 6 x 3  
## cooks.distance leverage covariance.ratios  
## <dbl> <dbl> <dbl>  
## 1 0.00307 0.00144 0.999  
## 2 0.00395 0.00270 1.00   
## 3 0.00306 0.000718 0.996  
## 4 0.000416 0.000232 0.998  
## 5 0.000438 0.000361 0.999  
## 6 0.00274 0.00168 1.00

#Check if any problematic cases exist, with cooks.distance greater than 1  
cooks.distance.gt1.df <- housing\_df[housing\_df$cooks.distance > 1, ]  
  
head(cooks.distance.gt1.df)

## # A tibble: 0 x 32  
## # ... with 35 variables: `Sale Date` <dttm>, `Sale Price` <dbl>,  
## # sale\_reason <dbl>, sale\_instrument <dbl>, sale\_warning <chr>,  
## # 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>, standardized.residuals <dbl>,  
## # studentized.residuals <dbl>, cooks.distance <dbl>,  
## # dfbeta[,"(Intercept)"] <dbl>, [,"sq\_ft\_lot"] <dbl>,  
## # [,"square\_feet\_total\_living"] <dbl>, [,"building\_grade"] <dbl>,  
## # dffit <dbl>, leverage <dbl>, covariance.ratios <dbl>, large\_residual <lgl>

**Answer k.**

Looking at above, amongst 514 large residual instances, none of them has cooks.distance greater than 1. So, none of the cases us having any undue influence on the model.

#average leverage calculations = (3 + 1) / 12672  
  
avg\_leverage <- (3 + 1) / 12672  
avg\_leverage

## [1] 0.0003156566

# calculate twice the average leverage and thrice the average leverage values  
avg\_leverage\_2 <- 2 \* avg\_leverage  
avg\_leverage\_3 <- 3 \* avg\_leverage  
  
nrow(large\_residuals[large\_residuals$leverage > avg\_leverage\_2,])

## [1] 216

nrow(large\_residuals[large\_residuals$leverage > avg\_leverage\_3,])

## [1] 149

we see that there are 216 large residuals cases with leverage greater than twice the average leverage and 149 large residuals cases with leverage greater than thrice the average leverage. however, none of the case has cooks distance greater than 1, as previously seen. So, we may not need to worry on this.

cvr\_upper <- 1 + avg\_leverage\_3  
cvr\_lower <- 1 - avg\_leverage\_3  
  
cvr\_lower

## [1] 0.999053

cvr\_upper

## [1] 1.000947

nrow(large\_residuals[large\_residuals$covariance.ratios > cvr\_upper | large\_residuals$covariance.ratios < cvr\_lower,])

## [1] 393

nrow(large\_residuals[large\_residuals$covariance.ratios > cvr\_upper,])

## [1] 20

nrow(large\_residuals[large\_residuals$covariance.ratios < cvr\_lower,])

## [1] 373

we see that there are total of 393 large residuals cases with covariance ratio outside of lower limit of 0.999053 and upper limit of 1.000947. However, none of the case has cooks distance greater than 1, as previously seen. So, we may not need to worry on this.

**l. assumptions of independence**

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

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

dwt(housing\_mod2)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.4252643 1.149453 0  
## Alternative hypothesis: rho != 0

**Answer l.**

The Durbin Watson test reports a test statistic indicates if the values fall outside the range of 1 to 3, then we may run into problems of autocorrelation between predicor variables. If closer to 2, indicated no autocorrelation between predictor variables. Durbin Watson statistic value above calculated is approx 1.15 and seems to be in the range of 1 to 3. But we can notice that there a dependence existing, nevertheless, between square feet lot and square feet total living variables and shown by Autocorrelation value of 0.4252643

**m. assumptions of no multicolinearity**

vif(housing\_mod2)

## sq\_ft\_lot square\_feet\_total\_living building\_grade   
## 1.053607 2.355268 2.283872

1/vif(housing\_mod2)

## sq\_ft\_lot square\_feet\_total\_living building\_grade   
## 0.9491204 0.4245801 0.4378528

mean(vif(housing\_mod2))

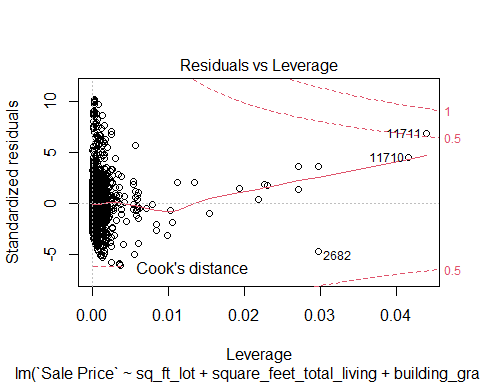
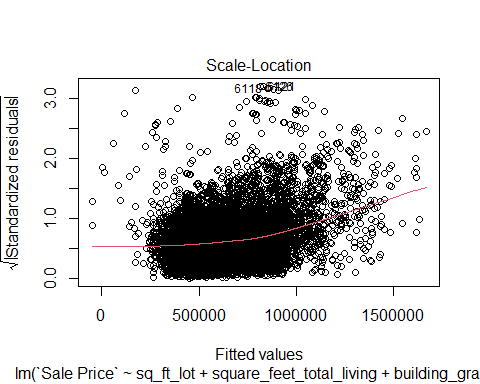
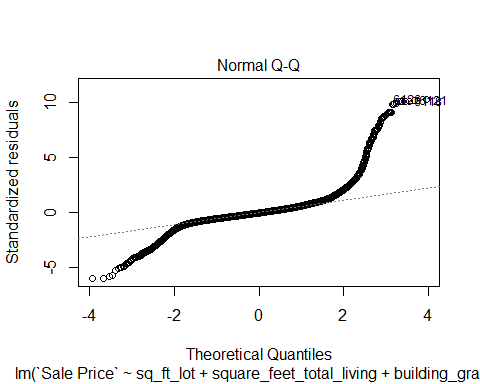
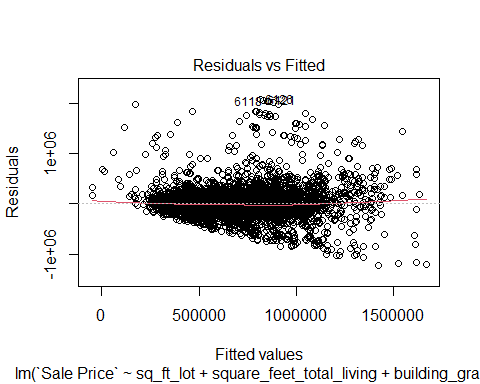
## [1] 1.897583

**Answer m.**

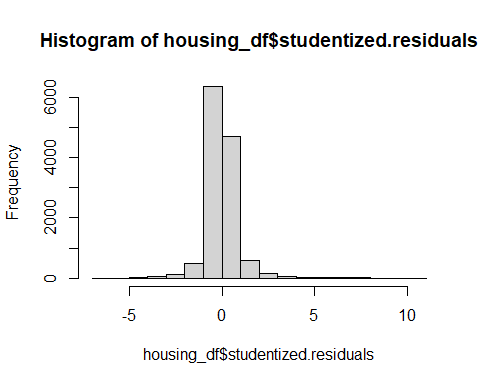
Based on (Field, Miles, and Field 2012) book All of the VIF values are well below 10. SO, there is no cause of concerns. All of the tolerance values are well above 0.1 Mean of VIF values is little above 1, indicates our model might be slightly biased and may be needs to consider additional variables or more cleaning of data needed or change predictor variables is needed somewhat

**n. plot() and hist() functions**

plot(housing\_mod2)



hist(housing\_df$studentized.residuals)



**Answer n.**

looking at the “Residuals vs Fitted Values” plot, residuals in our model shows a fairly random pattern, which is indicative of situation in which the assumptions of linearity, randomness and homoscedasticity have been met. Q-Q plot shows there is deviation from normality on either ends of the data, indicating we might still be having some skewed data / outliers in the sample and hence deviation from the normal straight line Looking at the histogram, we can see that the middle portion of the plot, which has most data points present, is near normal. But we have outliers present on either side (plot extends towards 10 on positive side). Suggesting we might need additional data cleaning to get a little better.

**o. unbiased regression model?**

**Answer o.**

Looking at the model, it is fairly close representation of the sample and a generalizable model to the larger poulation. Can be improved further by deleting outliers for model building purpose.

## References

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