Housing Data Analysis

Arindam Samanta

January 26 2020

### Source for the Housing data

Data for this assignment is focused on real estate transactions recorded from 1964 to 2016 and can be found in Week 7 Housing.xlsx. For statistical correlation, multiple regression and R programming, I am interested in some of the variables which have been included in the below code.

#### Loading the required libraries for our analysis

We have used the following libraries overall for our regression analysis. Regression analysis is used in stats to find trends in data. For example, we are trying to predict the sale price of the house based on some predictors and using regression analysis can help us quantify that.

library(dplyr)  
library(ggplot2)  
library(corrplot)  
library(readxl)

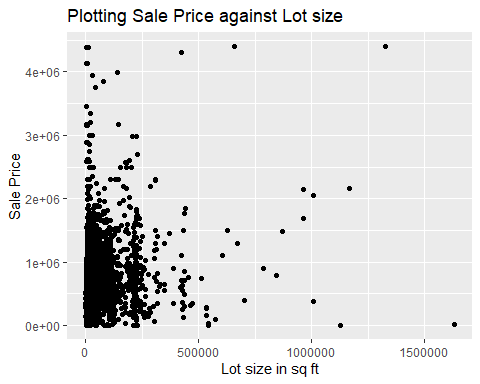
### a) Explain why you choose to remove data points from your ‘clean’ dataset.

Reading the excel sheet and loading the rows into a data frame. Then looking through the data using glimpse function. It shows that the sample has 12,865 observations and 24 variables. Each row represents a single house. That looks like a lot of variables so we would chose our variables that would be used for our prediction model and ignore the others. So essentially creating a smaller set of data with less variables. We are removing the other features from the data set so that we do not make our model overfit and so removing the unwanted data points.

#populating the housing\_data dataframe  
wd <- getwd()  
fname <- "week-7-housing.xlsx"  
path\_to\_file <- paste(wd,'/dataset/',fname, sep = "")  
path\_to\_file  
  
housing\_data <- read\_excel(path\_to\_file,col\_names=TRUE)  
glimpse(housing\_data)  
  
head(housing\_data)  
  
#Selecting only some variables from the dataframe into our analysis dataframe set 1  
housing\_data\_set1 <- housing\_data %>% dplyr::select(`Sale Price`,bedrooms,bath\_full\_count,sq\_ft\_lot)  
  
head(housing\_data\_set1)

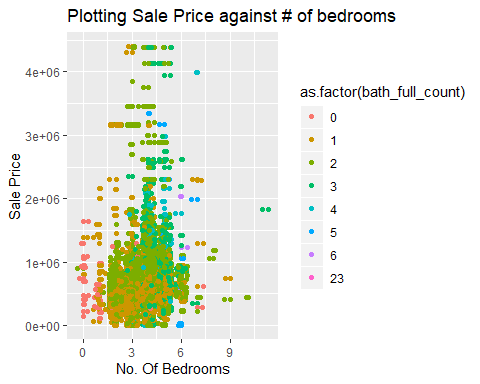
### b) Create two variables; one that will contain the variables Sale Price and Square Foot of Lot (same variables used from previous assignment on simple regression) and one that will contain Sale Price, Bedrooms, and Bath Full Count as predictors.

The first variable named var1 would be representing a simple linear regression model.Using ggplot we are trying to make a scatter plot for Sale Price as a function of sq foot of lot size. This is done in order to visualize the dataset.

 It shows that the data is fairly concentrated towards the lower end of the lot size.

Use lm() to fit a simple linear regression model for Price as a function of Lot size.

The second variable is using multiple linear regression using 2 predictors, bedrooms and and bath room full counts. Plotting them in a scatter plot looks like as below.

 Again using lm() to fit a multiple regression model var2 with 2 new predictors like bedrooms and bath\_full\_count. This is the second variable var2. We are trying to predict sale price as a function of both bath\_full\_count and Bedrooms.

### 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?

##   
## Call:  
## lm(formula = `Sale Price` ~ sq\_ft\_lot, data = housing\_data\_set1)  
##   
## 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

##   
## Call:  
## lm(formula = `Sale Price` ~ bedrooms + bath\_full\_count, data = housing\_data\_set1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3566590 -157368 -55794 67256 3891256   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 148987 14890 10.01 <2e-16 \*\*\*  
## bedrooms 70566 4048 17.43 <2e-16 \*\*\*  
## bath\_full\_count 148058 5450 27.17 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 383200 on 12862 degrees of freedom  
## Multiple R-squared: 0.1023, Adjusted R-squared: 0.1022   
## F-statistic: 733.2 on 2 and 12862 DF, p-value: < 2.2e-16

R-square measures the percentage of the variability in the response variable that is explained by the model. For the above two models Var1 has R-Square = 0.014 and Var2 has it 0.102. Increase in R-square is due to adding additional variable to my var2 model. For the above two models Var1 has Adj. R-Square = 0.014 and Var2 has it 0.102. But the increase in the Adj R-square shows that with the addition of the 2 new predictors they are better able to predict the sale price in the Var2 model.

## Warning: package 'QuantPsyc' was built under R version 3.6.2

## Loading required package: boot

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 3.6.2

##   
## Attaching package: 'MASS'

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

##   
## Attaching package: 'QuantPsyc'

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

## bedrooms bath\_full\_count   
## 0.1528878 0.2382795

### 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?

From the output of the Var2 we get the parameter values for the multiple regression model. So 70,566 is the paramter for the # of bedrooms and 148058 is for the bath\_full\_count. These values indicate the individual contribution of each predictor to the model.It also tells us that both the predictors are positively related to the sale price. paramter for bedrooms = 70566 indicates that if the number of bedrooms increases by 1 unit the sale price of the house increases by $70566. Similarly increase of full bath by 1 unit increases the sale price by $148058. The Standardized beta values (bedroom = 0.153, bath\_full\_count = 0.238) provide a better insight into the importance of a predictor in the model.

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

### 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.

## Analysis of Variance Table  
##   
## Model 1: `Sale Price` ~ sq\_ft\_lot  
## Model 2: `Sale Price` ~ bedrooms + bath\_full\_count  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 12863 2.0734e+15   
## 2 12862 1.8883e+15 1 1.8509e+14 1260.8 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We see from the above ANOVA table that the second model Var2 has F value of 1260.8 with a very small value of Pr(>F) is 2.2e -16 (i.e., 2.2 with decimal place moved 16 places to the left.) Thus we can say that Var2 significantly improved the fit of the model to the data compared to the simple regression model Var1.

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

In order to answer the question of whether our multiple regression model(Var2) fits the observed data well, or if it is influenced by a small number of cases, we can look for outliers and influential cases. We are using the augment function on the var2 model from the broom package to generate a new data frame housing\_data\_set1.augment storing the variables for the casewise diagnostics. To identify Outliers we would need Residuals and standardized residuals. These are stored in our dataframe as .resid and .std.resid To identify Influential cases: Cooks distance, hat values for leverage are used.These values are stored in .hat and .cooksd columns.

## Warning: package 'broom' was built under R version 3.6.2

## # A tibble: 6 x 11  
## Sale.Price bedrooms bath\_full\_count .fitted .se.fit .resid .hat .sigma  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 698000 4 2 727368. 3957. -2.94e4 1.07e-4 3.83e5  
## 2 649990 4 2 727368. 3957. -7.74e4 1.07e-4 3.83e5  
## 3 572500 4 1 579310. 6356. -6.81e3 2.75e-4 3.83e5  
## 4 420000 3 1 508744. 5382. -8.87e4 1.97e-4 3.83e5  
## 5 369900 3 1 508744. 5382. -1.39e5 1.97e-4 3.83e5  
## 6 184667 4 2 727368. 3957. -5.43e5 1.07e-4 3.83e5  
## # … with 3 more variables: .cooksd <dbl>, .std.resid <dbl>, cov.ratios <dbl>

### 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.

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

It is expected that in an ordinary sample we would expect 95% of cases to have standardized residuals (.std.resid) within about +2 and -2. Here we have a sample of 12,865 so it is reasonable to expect about (643) to have standardized residuals outside these limits.

## [1] 334

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

creating a new data frame to store those cases which are outside our 95% range.

## # A tibble: 334 x 5  
## Sale.Price bedrooms bath\_full\_count .std.resid large.residual  
## <dbl> <dbl> <dbl> <dbl> <lgl>   
## 1 1900000 4 3 2.67 TRUE   
## 2 1390000 0 1 2.85 TRUE   
## 3 1588359 2 2 2.62 TRUE   
## 4 1450000 3 2 2.07 TRUE   
## 5 1450000 2 1 2.64 TRUE   
## 6 270000 4 23 -9.76 TRUE   
## 7 2500000 4 2 4.63 TRUE   
## 8 2169000 4 3 3.38 TRUE   
## 9 1534000 4 1 2.49 TRUE   
## 10 555000 6 6 -2.37 TRUE   
## # … with 324 more rows

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

Let us now look at the leverage (hat value), cook’s distance and covariance ratio for these 334 for these large residual cases

## # A tibble: 1 x 7  
## Sale.Price bedrooms bath\_full\_count .hat .cooksd cov.ratios large.residual  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <lgl>   
## 1 270000 4 23 0.0900 3.14 1.07 TRUE

## # A tibble: 44 x 7  
## Sale.Price bedrooms bath\_full\_count .hat .cooksd cov.ratios large.residual  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <lgl>   
## 1 1390000 0 1 1.30e-3 0.00354 1.000 TRUE   
## 2 270000 4 23 9.00e-2 3.14 1.07 TRUE   
## 3 555000 6 6 3.39e-3 0.00636 1.00 TRUE   
## 4 450000 5 5 1.96e-3 0.00281 1.00 TRUE   
## 5 32000 6 5 2.12e-3 0.00794 1.000 TRUE   
## 6 1085000 0 0 1.51e-3 0.00301 1.00 TRUE   
## 7 2988000 5 4 1.01e-3 0.00824 0.996 TRUE   
## 8 1600000 1 1 7.11e-4 0.00246 0.999 TRUE   
## 9 2625000 6 3 8.02e-4 0.00472 0.997 TRUE   
## 10 1384950 1 1 7.11e-4 0.00167 0.999 TRUE   
## # … with 34 more rows

## # A tibble: 287 x 7  
## Sale.Price bedrooms bath\_full\_count .hat .cooksd cov.ratios large.residual  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <lgl>   
## 1 1900000 4 3 3.43e-4 8.18e-4 0.999 TRUE   
## 2 1588359 2 2 3.57e-4 8.15e-4 0.999 TRUE   
## 3 1450000 3 2 1.20e-4 1.72e-4 0.999 TRUE   
## 4 1450000 2 1 3.43e-4 7.97e-4 0.999 TRUE   
## 5 270000 4 23 9.00e-2 3.14e+0 1.07 TRUE   
## 6 2500000 4 2 1.07e-4 7.61e-4 0.995 TRUE   
## 7 2169000 4 3 3.43e-4 1.30e-3 0.998 TRUE   
## 8 1534000 4 1 2.75e-4 5.70e-4 0.999 TRUE   
## 9 555000 6 6 3.39e-3 6.36e-3 1.00 TRUE   
## 10 1968000 4 3 3.43e-4 9.30e-4 0.999 TRUE   
## # … with 277 more rows

Executing the command prints the variables but only for the cases for which large.residual = TRUE. The above output shows that the observation for the house with sale price of $270,000 and having a bath count of 23 only has a cook’s distance > 1 so it would have an undue influence on the model. The average leverage is (2+1)/12865 = .0002. Now looking for values Greater than (2\*.0002 = .0004) or .0006 comes back with 44 rows The CVR value upper limit is 1.0006 and the lower limit is .9994. Now taking data from the third table shows that there are 287 rows that could be problematic as they have the leverage outside of the limit.(> 1.0006 or < 0.9994)

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

We can test the assumption of independent errors using the Durbin-Watson test.

## Warning: package 'lmtest' was built under R version 3.6.2

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Durbin-Watson test  
##   
## data: var2  
## DW = 0.71677, p-value < 2.2e-16  
## alternative hypothesis: true autocorrelation is greater than 0

Here the test shows that DW is < 1 with a p value of very small. So this model is a cause for concern.

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

## Warning: package 'car' was built under R version 3.6.2

## Loading required package: carData

##   
## Attaching package: 'car'

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

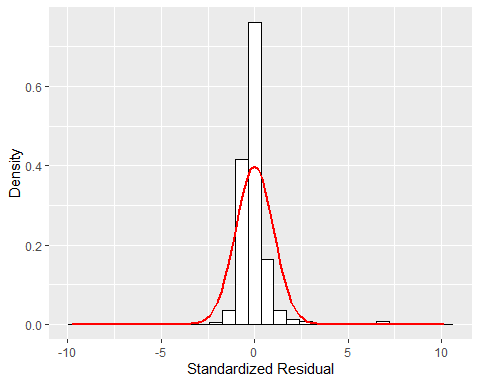
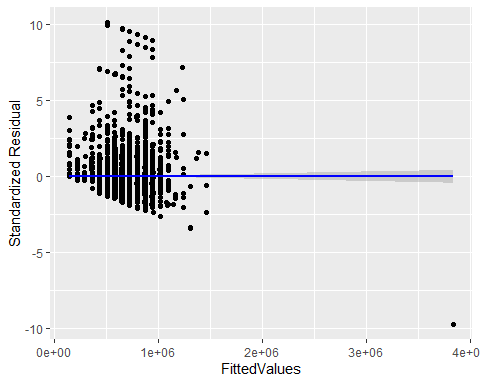
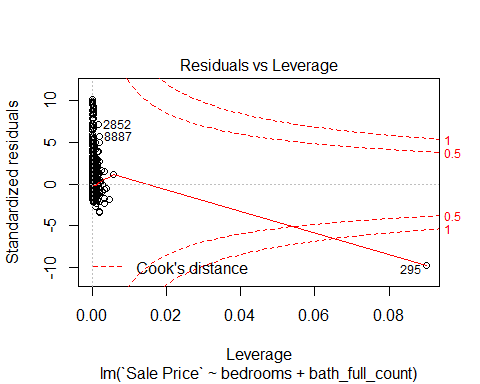
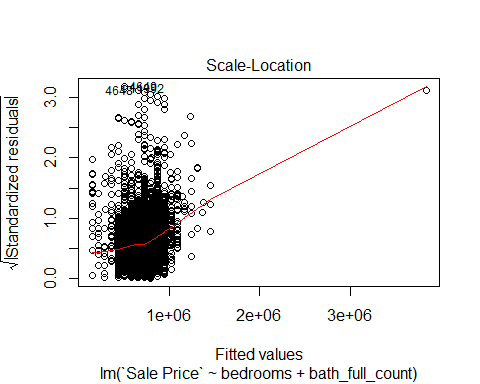
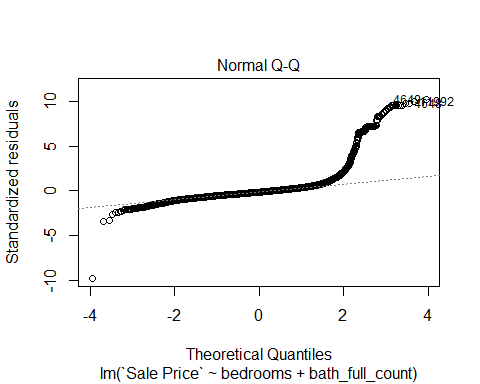
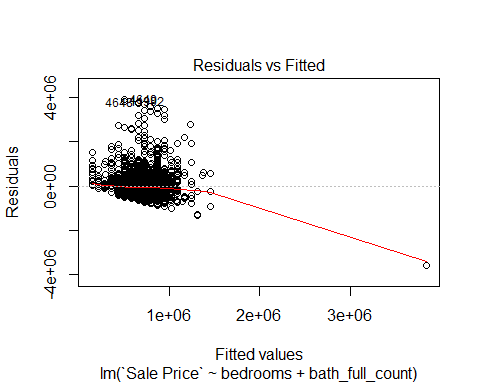
## The following object is masked from 'package:dplyr':  
##   
## recode

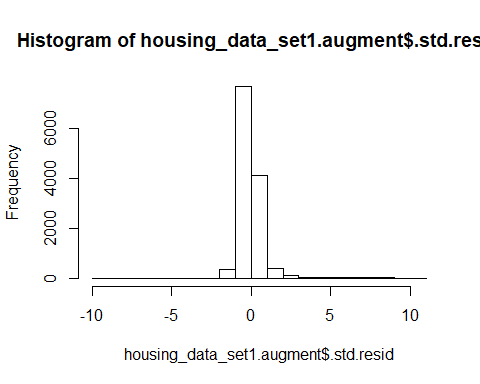
## bedrooms bath\_full\_count   
## 1.102269 1.102269

## bedrooms bath\_full\_count   
## 0.9072195 0.9072195

The VIF and tolerance statistics are useful statistics to assess collinarity. For our current model the VIF values for both the variables are all well below 10 and the tolerance statistics all well above 0.2. Based on these measures we can safely conclude that there is no collinearity within our data. ### 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.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 The above plot shows the histogram of standardized residuals. Let us now draw the Scatter plot of standardized residuals against predicted values.   From the above Residuals vs Fitted plot it shows that the scatter points are not random and rather concentrated towards the positive values for the residuals. Hence the data shows a non linear relationship. The QQ plot shows deviation from linearity. The normal distribution shown by the stariaght line do not have all the points over it so it shows skewness of the data towards the higher end of the model.

 The histogram for the data is roughly skewed to the right as it does not have a bell-curve shape. ### o) Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us about the sample vs. the entire population model? So overall this regression model is not unbiased. It is not a perfectly good model for the data as it has outliers,influential cases, etc.and so we cannot use that model to draw conclusions about our sample. We cannot generalize your findings beyond your sample to the entire population.