Housing Data Analysis

Arindam Samanta

January 19 2020

Using your ‘clean’ data set from the previous week complete the following:

Explain why you choose to remove data points from your ‘clean’ dataset.  
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.   
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?  
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?  
Calculate the confidence intervals for the parameters in your model and explain what the results indicate.  
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.  
Perform casewise diagnostics to identify outliers and/or influential cases, storing each functions output in a dataframe assigned to a unique variable name.  
Calculate the standardized residuals using the appropriate command, specifying those that are +-2, storing the results of large residuals in a variable you create.  
Use the appropriate function to show the sum of large residuals.  
Which specific variables have large residuals (only cases that evaluate as TRUE)?  
Investigate further by calculating the leverage, cooks distance, and covariance rations. Comment on all cases that are problematics.  
Perform the necessary calculations to assess the assumption of independence and state if the condition is met or not.  
Perform the necessary calculations to assess the assumption of no multicollinearity and state if the condition is met or not.  
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.  
Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us about the sample vs. the entire population model?

## Part

### 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. 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.

## Data Preparations

We have used the following libraries 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)

## 

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.

#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 %>%  
 select (`Sale Price`,`Sale Date`,bath\_full\_count, sq\_ft\_lot,bedrooms)

Now we start looking at the dataset

## # A tibble: 6 x 5  
## `Sale Price` `Sale Date` bath\_full\_count sq\_ft\_lot bedrooms  
## <dbl> <dttm> <dbl> <dbl> <dbl>  
## 1 698000 2006-01-03 00:00:00 2 6635 4  
## 2 649990 2006-01-03 00:00:00 2 5570 4  
## 3 572500 2006-01-03 00:00:00 1 8444 4  
## 4 420000 2006-01-03 00:00:00 1 9600 3  
## 5 369900 2006-01-03 00:00:00 1 7526 3  
## 6 184667 2006-01-03 00:00:00 2 7280 4

Summaize the data based on each of the columns to see how the distribution looks

## # A tibble: 8 x 2  
## bath\_full\_count `# of houses`  
## <dbl> <int>  
## 1 0 51  
## 2 1 3772  
## 3 2 7905  
## 4 3 1032  
## 5 4 87  
## 6 5 14  
## 7 6 3  
## 8 23 1

We can consider bath count to be a categorical variable. There are 8 distinct bath counts in the data set. Now lets look at the other categorical variable bedrooms.

## # A tibble: 12 x 2  
## bedrooms `# of houses`  
## <dbl> <int>  
## 1 0 19  
## 2 1 33  
## 3 2 1658  
## 4 3 4493  
## 5 4 5515  
## 6 5 1047  
## 7 6 83  
## 8 7 11  
## 9 8 2  
## 10 9 2  
## 11 10 1  
## 12 11 1

The summarized data shows that there are 12 distinct bedroom counts for the dataset.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 698 460000 593000 660738 750000 4400000

## # A tibble: 2 x 5  
## `Sale Price` `Sale Date` bath\_full\_count sq\_ft\_lot bedrooms  
## <dbl> <dttm> <dbl> <dbl> <dbl>  
## 1 4400000 2010-03-02 00:00:00 2 657816 3  
## 2 4400000 2010-03-02 00:00:00 1 1327090 3

Now lets convert the bath\_full\_count and bedrooms to factors and create a .mod dataset

## Observations: 12,865  
## Variables: 5  
## $ `Sale Price` <dbl> 698000, 649990, 572500, 420000, 369900, 184667, 10500…  
## $ `Sale Date` <date> 2006-01-03, 2006-01-03, 2006-01-03, 2006-01-03, 2006…  
## $ bath\_full\_count <fct> 2, 2, 1, 1, 1, 2, 3, 2, 2, 1, 2, 2, 1, 2, 2, 2, 2, 2,…  
## $ sq\_ft\_lot <dbl> 6635, 5570, 8444, 9600, 7526, 7280, 97574, 30649, 426…  
## $ bedrooms <fct> 4, 4, 4, 3, 3, 4, 5, 4, 4, 4, 3, 3, 4, 3, 3, 3, 4, 4,…