Predicting Residential Home Sales Prices Using Regression Analysis

Rumil Legaspi, [Rumil.legaspi@gmail.com](mailto:Rumil.legaspi@gmail.com)

Nancy Huerta, [Nhuerta20@apu.edu](mailto:Nhuerta20@apu.edu)

11 April 2021

# **Purpose**

We are conducting a **multiple linear regression** from the Real Estate Sales (APPENC07) dataset to analyze the relationship of the given features, ***bedrooms***, ***bathrooms***, and ***garage size***, with the outcome variable, ***house sales price*** in a midwestern city.

# **Our Data**

## Background on Dataset & Variables

Our dataset is comprised of *522 total transactions* from home sales during the year 2002.

|  |  |  |  |
| --- | --- | --- | --- |
| Response Variable (Y) | Explanatory Variable 1 () | Explanatory Variable 2 () | Explanatory Variable 3 () |
| “house\_price” | “beds” | “baths” | “garage\_size” |
| sales price of residence (in dollars) | Number of bedrooms | Number of bathrooms | Number of cars the garage can hold |

#Setting up our work environment  
setwd("C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Analysis)/Project 2")  
library(nortest)  
library(olsrr)  
library(car)  
library(lmtest)  
library(MASS)  
library(tidyverse)  
library(ggcorrplot)  
#Loading in the text data  
raw\_data = read.table(file = "APPENC07.txt", header = FALSE, sep = "")  
  
#Converting into tibble data frame for easier data analysis  
house\_data <- as\_tibble(raw\_data)

#Defining and renaming our Explanatory(X) and Response(Y) variables  
house\_data <- house\_data %>% select(house\_price = V2,  
 beds = V4,  
 baths = V5,  
 garage\_size = V7)  
  
#Setting explanatory and response variables  
house\_price <- house\_data %>% select(house\_price) #Y  
beds <- house\_data %>% select(beds) #X1  
baths <- house\_data %>% select(baths) #X2  
garage\_size <- house\_data %>% select(garage\_size) #X3

# Part 1 - Model Estimation and Interpretation

## 1a. Fitting a regression model estimating sales price using beds, baths, and garage size as predicting variables

#Using the lm function to fit a multiple regression model  
house\_lm <- lm(house\_price ~ beds + baths + garage\_size, data = house\_data)  
  
#Regression summary  
summary(house\_lm)

##   
## Call:  
## lm(formula = house\_price ~ beds + baths + garage\_size, data = house\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -249973 -55441 -16444 33862 423872   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -45886.3 17261.6 -2.658 0.0081 \*\*   
## beds 935.4 4966.4 0.188 0.8507   
## baths 67818.9 5150.4 13.168 <2e-16 \*\*\*  
## garage\_size 67332.3 7176.3 9.383 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 93300 on 518 degrees of freedom  
## Multiple R-squared: 0.5451, Adjusted R-squared: 0.5424   
## F-statistic: 206.9 on 3 and 518 DF, p-value: < 2.2e-16

## 1b. Interpretation of Coefficients

### Intercept & Partial Slopes

From summarizing our multiple linear regression model we can see:

|  |  |  |  |
| --- | --- | --- | --- |
| **Intercept** | Bedrooms | Bathrooms | Garage Size |
|  |  |  |  |
| -45886.3 | 935.4 | 67818.9 | 67332.3 |

and the estimated regression equation to be:

The partial slopes in our summary indicate that when any one of the partial slopes **Increase by 1 unit** and other explanatory variables held constant and unchanged we can expect:

* While holding our other explanatory variables Bathrooms and Garage Size constant and unchanged, when **Bedrooms** increase by 1 unit, we can expect our **house sales price** to increase by **roughly $935.4**.
* While holding our other explanatory variables Bedrooms and Garage Size constant and unchanged, when **Bathrooms** increases by 1 unit, we can expect our **house sales price** to increase by **roughly $67,818.9.**
* While holding our other explanatory variables Bedrooms and Bathrooms constant and unchanged, when **Garage size** increases by 1 unit, we can expect our **house sales price** to increase by **roughly $67,332.3**.

## 1c. Interpretation of Adjusted R-Squared = 0.54

A adjusted R squared value, similar to the R square value, tells us how much of the variability in our model is explained by our predictor variables, while also penalizing redundant or otherwise useless predictor variables helping us to resist urges of adding too many variables into our model.

In this case our adjusted $R^2$ of 0.54 tells us that about 54% of the variation in our response variable is explained by our 3 explanatory variables.

# Part 2 - Prediction

## 2a. Predicting the house sales price for a house with 3 bedrooms, 3 bathrooms, and a 2-car garage

#We create an artifical observation where a given house has  
#3 Bedrooms, 3 Bathrooms, and a 2 car garage  
new\_house\_data <- data.frame(beds = 3, baths = 3, garage\_size = 2)

## 2b. Calculating the 95% confidence interval

#confidence interval  
ci\_house <- predict(house\_lm, new\_house\_data, interval = "confidence", level = 0.95)  
ci\_house

## fit lwr upr  
## 1 295041.2 284025.7 306056.6

### Interpretation

This 95% confidence interval, when Bedrooms = 3, Bathrooms = 3, and Garage Size = 2, is from **74.84094 to 79.70906.**

When Bedrooms = 3, Bathrooms = 3, and Garage Size = 2, with 95% confidence we can expect our confidence interval to capture the average of house sales prices (response variable).

## 2c. Calculating the 95% prediction interval

#prediction interval  
pi\_house <- predict(house\_lm, new\_house\_data, interval = "prediction", level = 0.95)  
pi\_house

## fit lwr upr  
## 1 295041.2 111422.3 478660

### Interpretation

From the results we can predict with 95% confidence that when their are 3 bedrooms, 3 bathrooms, and a garage that can hold 2 cars, the predicted house sales price will fall somewhere between **111,422 to 478,660 dollars.**

# Part 3 - Hypothesis Testing

## 3a. Checking the significance for each individual partial slope (independent variable)

### Using a significance level of

**Null Hypothesis**: : (slopes are showing no change), **is not** linearly associated with Y, therefore the partial slope **is not significant.**

**Alternative Hypothesis**: : (slopes are showing change), **is** linearly associated with Y, therefore the partial slope **is significant.**

Table Representation of Hypothesis Testing

|  |  |  |
| --- | --- | --- |
| Bedrooms () | Bathrooms () | Garage Size () |
| 0.8507 **>** | <2e-16 **<** | <2e-16 **<** |
| Fail to reject | Reject | Reject |
| Not Significant | Significant | Significant |
|  |  |  |

**Bedroom variable:**

The p-value of Bedroom is 0.8507 and is greater than our (accepted error) of 0.05, so we **fail to reject** our NULL hypothesis and must conclude with our NULL hypothesis. Stating that our partial slope, **Bedrooms**, does not show overall significance in our model.

**Bathroom & Garage Size variables:**

On the other hand because the p-value of Bathroom and Garage size are both <2e-16 and are incredibly smaller than our (accepted error) of 0.05, so we **reject** our NULL hypothesis and conclude with our alternative hypothesis. Our alternative hypothesis states that our partial slopes, **Bathroom and Garage Size**, shows overall significance in our model.

## 3b. Conducting an F-test to check overall model significance

### Using a significance level of

**Null Hypothesis**: : (**No** partial slopes are significant). Shows no change, therefore **does not** show overall model significance.

**Alternative Hypothesis**: : (**At least one** partial slope is significant). Shows change, therefore **showing** overall model significance

#We can use the qt() to find our critical value and compare with our t-value (test statistic)  
# We use 0.95 Because of our 95% confidence interval and 518 for our degrees of freedom  
qt(0.975, 518)

## [1] 1.964554

#Checking for our f-value   
anova(house\_lm)

## Analysis of Variance Table  
##   
## Response: house\_price  
## Df Sum Sq Mean Sq F value Pr(>F)   
## beds 1 1.6931e+12 1.6931e+12 194.515 < 2.2e-16 \*\*\*  
## baths 1 2.9426e+12 2.9426e+12 338.057 < 2.2e-16 \*\*\*  
## garage\_size 1 7.6627e+11 7.6627e+11 88.032 < 2.2e-16 \*\*\*  
## Residuals 518 4.5089e+12 8.7044e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Table Representation of F-test Hypothesis Testing

|  |  |  |  |
| --- | --- | --- | --- |
| Test Statistic Type & Result | Bedrooms () | Bathrooms () | Garage Size () |
| F-value | 194.515 > 1.96 | 338.057 > 1.96 | 88.032 > 1.96 |
| P-value | 2.2e-16 < 0.05 | 2.2e-16 < 0.05 | 2.2e-16 < 0.05 |
| Result | Significant | Significant | Significant |

Our p-value of < 2.2e-16 being less than our alpha and our F values being larger then our critical value tells us we can **reject** our NULL hypothesis and conclude with our alternative hypothesis, that **at least one** of our predictor variables **shows** overall model significance.

## 3c. Conducting Partial F tests

### Which variable is actually contributing?

Conducting partial F tests is important to see if the number of bathrooms (X2) and garage size(X3) are **jointly significant.**

### Using a significance level of 0.05

**Null Hypothesis**: : ***There is no*** change when adding certain predictors to the significance of our model

**Alternative Hypothesis**: : ***There is*** change when adding certain predictors towards the significance of our model

#full model  
house\_lm

##   
## Call:  
## lm(formula = house\_price ~ beds + baths + garage\_size, data = house\_data)  
##   
## Coefficients:  
## (Intercept) beds baths garage\_size   
## -45886.3 935.4 67818.9 67332.3

#reduced model without bathrooms and garage size  
bed\_lm <- lm(house\_price ~ beds, data = house\_data)

We now ***compare*** our reduced model with our complete model

anova(bed\_lm, house\_lm)

## Analysis of Variance Table  
##   
## Model 1: house\_price ~ beds  
## Model 2: house\_price ~ beds + baths + garage\_size  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 520 8.2178e+12   
## 2 518 4.5089e+12 2 3.7089e+12 213.04 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since the p-value is **2.2e-16** is less than our significance level of 0.05 we see that bathroom and garage size are both jointly significant and therefore we can reject the null hypothesis, indicating there is significance in keeping both bathroom and garage size in our model.

In effect, we are concluding that bathroom and garage size are predictors that do contribute information in the prediction of house sales price and therefore should be retained in the model.

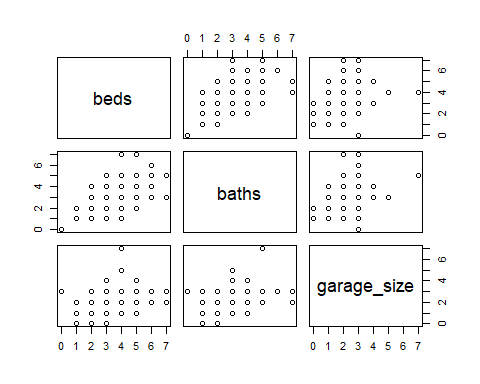
# Part 4 - Multicolinearity

### Why bother with multicolinearity?

Having multicolinearity is problematic because by having multiple correlated predictor variables, it becomes harder for our model to attribute significance to our predictor variables. It creates redundant and duplicate information, thereby negatively affecting the results of our regression model.

## 4a. Creating scatterplots and correlation matrices

#Plotting a scatterplot matrix \*\*(why does it look symmterical?)  
scat\_matrix <- c(beds, baths, garage\_size) %>%   
 data.frame() %>%  
 plot()



scat\_matrix

## NULL

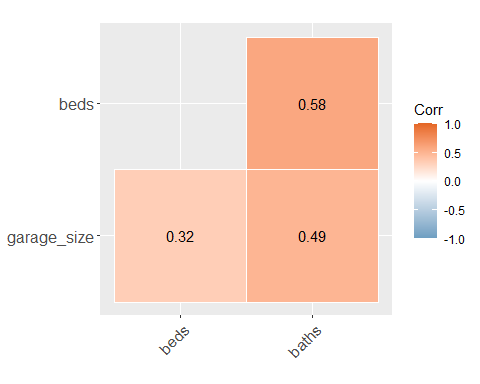
#Correlation Matrix  
corr\_matrix <- c(beds, baths, garage\_size) %>%   
 data.frame() %>%   
 cor()  
  
corr\_matrix

## beds baths garage\_size  
## beds 1.0000000 0.5834469 0.3168137  
## baths 0.5834469 1.0000000 0.4898981  
## garage\_size 0.3168137 0.4898981 1.0000000

ggcorr\_matrix <- ggcorrplot(corr\_matrix, hc.order = TRUE, type = "lower", lab = TRUE,  
 outline.col = "white",  
 ggtheme = ggplot2::theme\_gray,  
 colors = c("#6D9EC1", "white", "#E46726"))  
  
#Printing both matrices  
scat\_matrix

## NULL

ggcorr\_matrix



summary(house\_lm)

##   
## Call:  
## lm(formula = house\_price ~ beds + baths + garage\_size, data = house\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -249973 -55441 -16444 33862 423872   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -45886.3 17261.6 -2.658 0.0081 \*\*   
## beds 935.4 4966.4 0.188 0.8507   
## baths 67818.9 5150.4 13.168 <2e-16 \*\*\*  
## garage\_size 67332.3 7176.3 9.383 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 93300 on 518 degrees of freedom  
## Multiple R-squared: 0.5451, Adjusted R-squared: 0.5424   
## F-statistic: 206.9 on 3 and 518 DF, p-value: < 2.2e-16

corr\_matrix

## beds baths garage\_size  
## beds 1.0000000 0.5834469 0.3168137  
## baths 0.5834469 1.0000000 0.4898981  
## garage\_size 0.3168137 0.4898981 1.0000000

## 4b. Removing Two Strongly Correlated Variables

A way to combat this is by removing a highly correlated predictor. From the correlation matrix and by looking at our correlation coefficient, we can see moderately positive relationship between bedrooms and bathrooms which might be worth further investigating.

We know there is some kind of multicolinearity issue with bed and baths, I am more interested in beds than baths. So I can remove baths from our model thereby correcting our multicolinearity issue.

#summary of original predictor  
summary(house\_lm)

##   
## Call:  
## lm(formula = house\_price ~ beds + baths + garage\_size, data = house\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -249973 -55441 -16444 33862 423872   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -45886.3 17261.6 -2.658 0.0081 \*\*   
## beds 935.4 4966.4 0.188 0.8507   
## baths 67818.9 5150.4 13.168 <2e-16 \*\*\*  
## garage\_size 67332.3 7176.3 9.383 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 93300 on 518 degrees of freedom  
## Multiple R-squared: 0.5451, Adjusted R-squared: 0.5424   
## F-statistic: 206.9 on 3 and 518 DF, p-value: < 2.2e-16

We notice bedrooms is not a significant variable from looking at the p-value, when in fact, the reality is it **should** be significant. Knowing this, we can check to see how well our model performs when removing bathroom since there is a multicolinearity issue.

#removing beds\*\*  
nobeds\_lm <- lm(house\_price ~ baths + garage\_size, data = house\_data)  
summary(nobeds\_lm)

##   
## Call:  
## lm(formula = house\_price ~ baths + garage\_size, data = house\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -249830 -55576 -15656 33933 423631   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -44091 14377 -3.067 0.00228 \*\*   
## baths 68321 4402 15.521 < 2e-16 \*\*\*  
## garage\_size 67391 7163 9.409 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 93210 on 519 degrees of freedom  
## Multiple R-squared: 0.545, Adjusted R-squared: 0.5433   
## F-statistic: 310.9 on 2 and 519 DF, p-value: < 2.2e-16

#removing baths  
nobaths\_lm <- lm(house\_price ~ garage\_size + beds, data = house\_data)  
summary(nobaths\_lm)

##   
## Call:  
## lm(formula = house\_price ~ garage\_size + beds, data = house\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -352121 -66704 -28488 42621 529386   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -62861 19868 -3.164 0.00165 \*\*   
## garage\_size 104753 7606 13.773 < 2e-16 \*\*\*  
## beds 34804 4904 7.098 4.19e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 107700 on 519 degrees of freedom  
## Multiple R-squared: 0.3928, Adjusted R-squared: 0.3904   
## F-statistic: 167.9 on 2 and 519 DF, p-value: < 2.2e-16

In both cases we can see that by either removing bedroom or bathroom in our model, the predictors still remain significant but more importantly we can now see that bedrooms is in fact a significant predictor when removing the bathroom variable in our model concluding that we have addressed our issue of multicolinearity.