Predicting Residential Home Sales Prices Using Regression Analysis

Rumil Legaspi, Rumil.legaspi@gmail.com   Nancy Huerta, Email

9 April 2021

Contents

Purpose

We are conducting a multiple linear regression model 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 and variables

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

Response Variable (Y)	Explanatory Variable 1 (X_1)	Explanatory Variable 2 (X_2)	Explanatory Variable 3 (X_3)
"house_price" sales price of residence (in dollars)	"beds" Number of bedrooms	"baths" Number of bathrooms	"garage_size" 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

Fitting a regression model estimating sales price using predictors

```
#test
test_lm <- lm(house_price ~ baths + beds + garage_size, data = house_data)
test2_lm <- lm(house_price ~ garage_size +baths + beds, data = house_data)
summary(test_lm)
```

```
##
## Call:
## lm(formula = house_price ~ baths + beds + 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 **
## baths        67818.9     5150.4  13.168   <2e-16 ***
## beds           935.4     4966.4   0.188   0.8507
## 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
```

```r
summary(test2_lm)
```

```
##
## Call:
## lm(formula = house_price ~ garage_size + baths + beds, 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 **
## garage_size  67332.3     7176.3   9.383   <2e-16 ***
## baths        67818.9     5150.4  13.168   <2e-16 ***
## beds           935.4     4966.4   0.188   0.8507
## ---
## 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
```

```r
#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
```

Interpretation of coefficients

Intercept & Partial Slopes

From summarizing our multiple linear regression model we can see:

Intercept	Bedrooms	Bathrooms	Garage Size
β_0 -45886.3	β_1 935.4	β_2 67818.9	β_3 67332.3

and the estimated regression equation to be:

$\hat{Y} = -45886.3 + 935.4X + 67818.9X + 67332.3X$

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:

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

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, but also penalizes 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

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

```
#Creating a 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)
```

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
```

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
```

Part 3 - Hypothesis Testing

3a. Conducting a hypothesis test for each individual partial slope (independent variable) to check for significance using a significance level of $\alpha = 0.05$

Null Hypothesis: H_0: $\beta_j = 0$ (slopes are showing no change), X_j **is not** linearly associated with Y, therefore the partial slope **is not significant.**

Alternative Hypothesis: H_1: $\beta_j \neq 0$ (slopes are showing change), X_j **is** linearly associated with Y, therefore the partial slope **is significant.**

Table 3: Table Representation of Hypothesis Testing

Bedrooms (X_1)	Bathrooms (X_2)	Garage Size (X_3)
$0.8507 > \alpha = 0.05$	<2e-16 $< \alpha = 0.05$	<2e-16 $< \alpha = 0.05$
Fail to reject H_0	Reject H_0	Reject H_0
Not Significant	Significant	Significant

Bedroom variable:

The p-value of Bedroom is 0.8507 and is greater than our α (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 α (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 with a significance level of $\alpha = 0.05$

Null Hypothesis: H_0: $\beta_1 = \beta_2 = 0$ (**No** partial slopes are significant). Shows no change, therefore **does not** show overall model significance.

Alternative Hypothesis: H_1: $\beta_1 = \beta_2 \neq 0$ (**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 4: Table Representation of F-test Hypothesis Testing

Test Statistic Type & Result	Bedrooms (X_1)	Bathrooms (X_2)	Garage Size (X_3)
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.

Partial F tests

**Are the number of bathrooms (2) and garage size (3) jointly significant in the MLR model? Conduct a partial F-test and conclude. Choose your own .

We will be conducting partial F tests to see if the number of bathrooms (X2) and garage size(X3) are jointly significant using a significance level of 0.05.

** In other words is there significance in adding bathrooms and garage size into our model?

** is it comparing full model (y=b0+b1X1+b2X2+b3X3) to (y=b0+b1X1)

THIS ONE ^^^

```
#full model
#house_lm

#reduced model bathrooms and garage size (w/o bath a)
#bed_lm(house_lm ~ beds, data = house_data)
```

or comparing full model (y=b0+b1X1+b2X2+b3X3) to (y=b0+b1x1+b2X2) and (y=b0+b1x1+b3X3)?

```
#full model
#house_lm

#model without garage size only
#bed_bath_lm <- lm(house_lm ~ beds, baths, data = house_data)
#model without bath only
#bed_garage_lm <- lm(house_lm ~ beds, garage_size, data = house_data)
```

We now compare our reduced model with our complete model

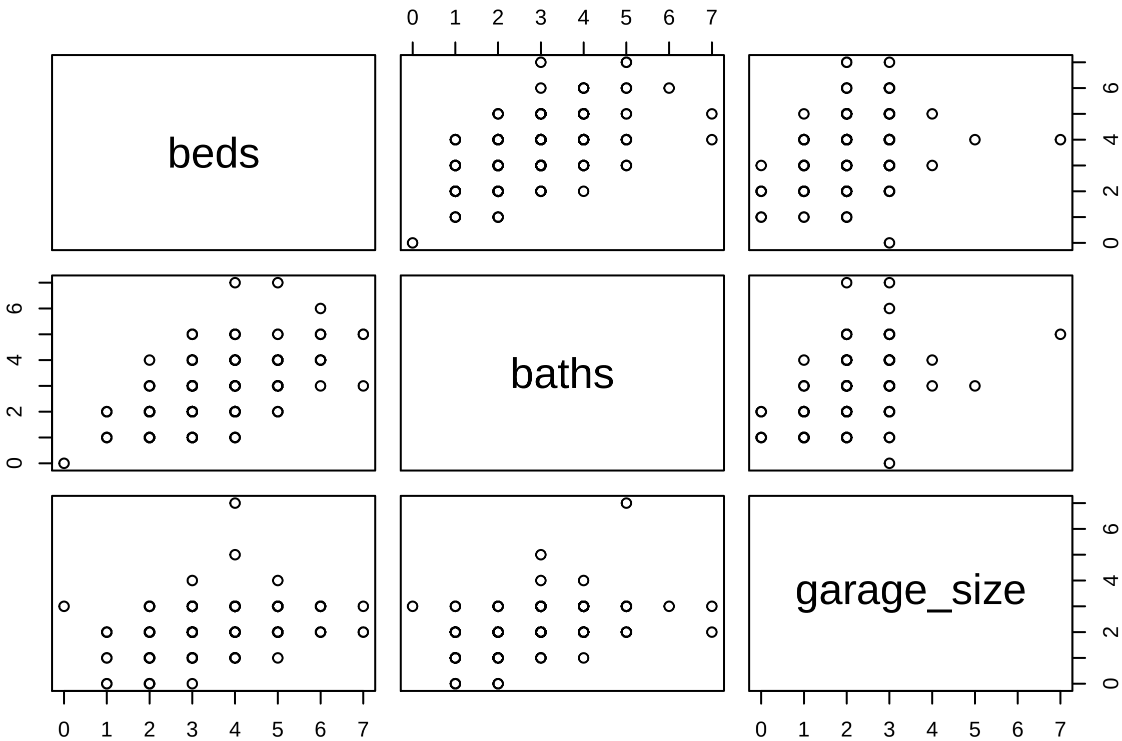
```
#anova(house_lm,)
```

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

**Part 4 - Multicolinearity

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
```

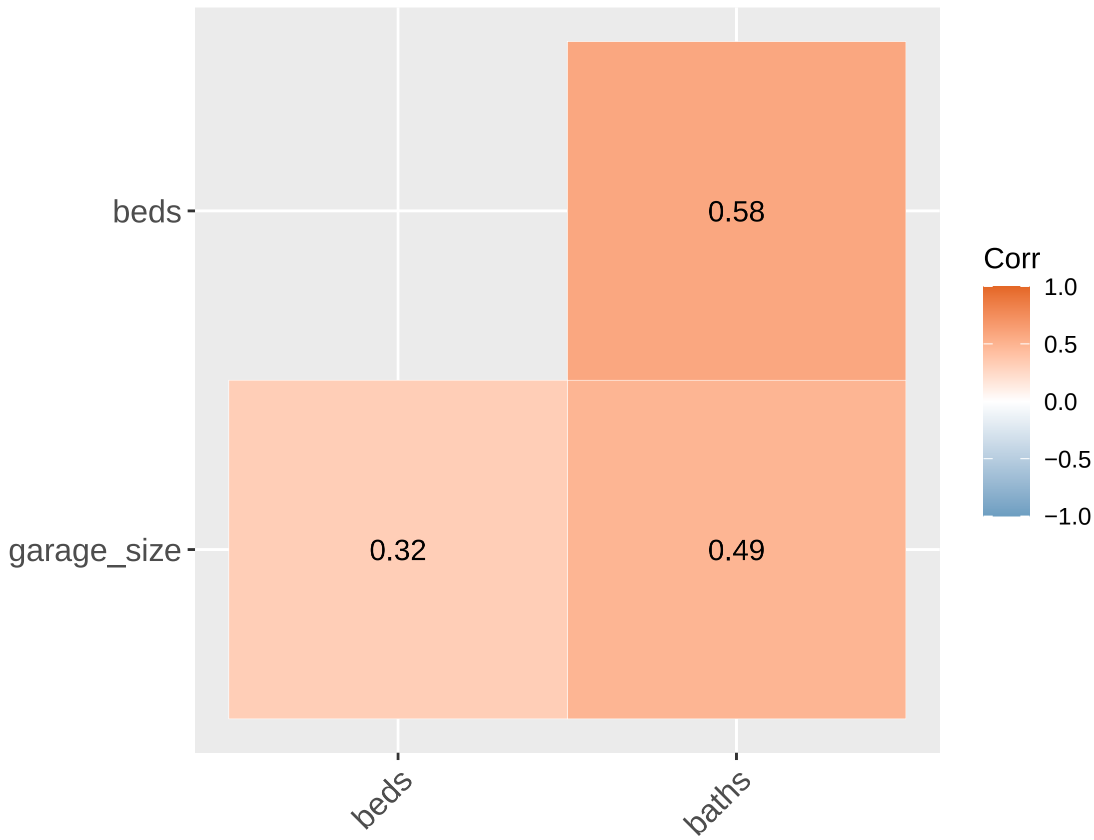
```
##                  beds     baths garage_size
## beds        1.0000000 0.5834469   0.3168137
## baths       0.5834469 1.0000000   0.4898981
## garage_size 0.3168137 0.4898981   1.0000000
```

```
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
```



The correlation coefficient shows a moderately positive relationship between most of our predictor variables. Moreover, it does not necessarily look like are

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

A way to combat this is by removing a highly correlated predictor.

```
vif(house_lm)
```

```
##        beds       baths garage_size
##    1.519003    1.798078    1.318314
```

```
#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
```

```r
#removing baths
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
```

```r
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
```