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

```
#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	β_1	β_2	β_3
-45886.3	935.4	67818.9	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 $67332.3**.

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

Testing if **each of the individual predictors** (partial slopes) in this regression to see if they hold significance.

Testing for Individual Parameter Significance Using p-value significance level ($\alpha = 0.05$)

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.**

Bedrooms variable:

The p-value of Bedroom is 0.8507 and is greater than our α (accepted error) of 0.05 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.

Bathrooms & 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 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.

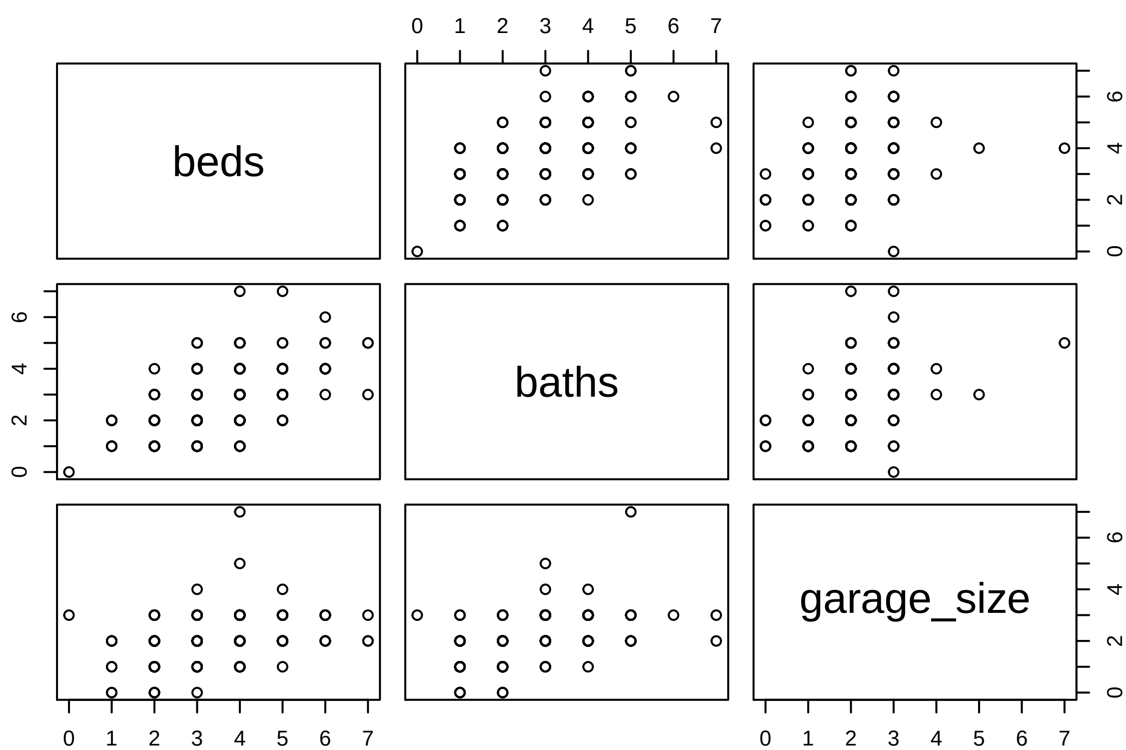
Table 3: Table Representation of Hypothesis Testing

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

Part 4 - Multicolinearity

Creating scatterplot and correlation matrices

```
#Plotting a scatterplot matrix **(why does it look symmterical?)
scat_matrix <- c(beds, baths, garage_size) %>% data.frame() %>%
  plot()
```



```r
scat_matrix
```

```
## NULL
```

```r
#Correlation Matrix
corr_matrix <- c(beds, baths, garage_size) %>% data.frame() %>%
  cor()
corr_matrix
```

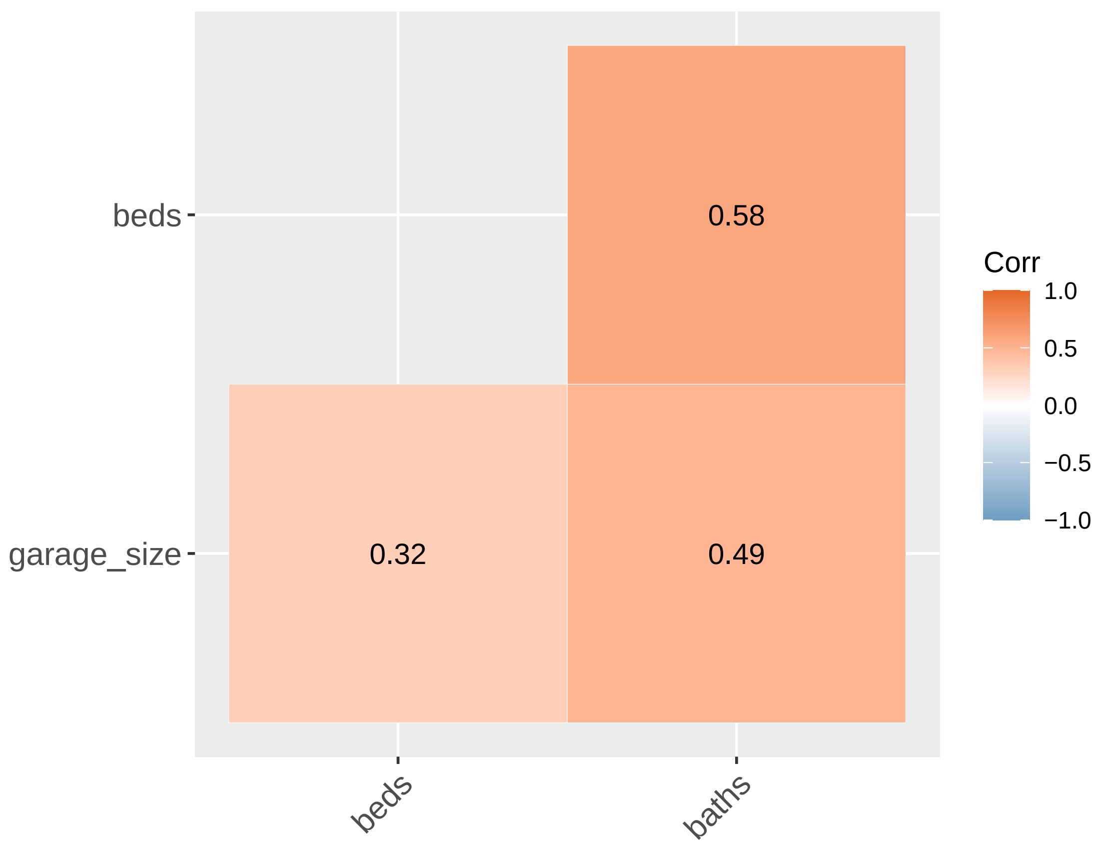
```
##                  beds     baths garage_size
## beds        1.0000000 0.5834469   0.3168137
## baths       0.5834469 1.0000000   0.4898981
## garage_size 0.3168137 0.4898981   1.0000000
```

```r
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
```

```r
ggcorr_matrix
```



The correlation coefficient shows a moderately positive relationship between our predictor variables except for______