STAT 592: Project 3

Rumil Legaspi, Rumil.legaspi@gmail.com   Efe Umukoro, email
Solange Ebobisse Mapenya, email

4/28/2021

Contents

Background & Objective

Given that a city tax assessor is interested in predicting residential home sales prices in a midwestern city with various characteristics, we will be conducting a **multiple linear regression analysis** from the Real Estate Sales (APPENC07) dataset from 2002. We aim to observe and predict the relationship using the given features, ***square feet***, the absence or presence of a ***swimming pool*** and ***air conditioning***, and our response variable as ***house sales price***.

```
#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)
```

```r
library(MASS)
library(tidyverse)
library(ggcorrplot)
library(knitr)
#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)
```

```r
#Defining and renaming our Explanatory(X) and Response(Y) variables
house_data <- house_data %>% select(sales_price = V2,
                                    square_feet = V3,
                                    swimming_pool = V8,
                                    air_conditioning = V6)

#Setting explanatory and response variables
sales_price <-  house_data %>% select(sales_price) #Y
square_feet <- house_data %>% select(square_feet) #X1
swimming_pool <- house_data %>% select(swimming_pool) #X2
air_conditioning <- house_data %>% select(air_conditioning) #X3

knitr::kable(house_data) %>%
  head(10)
```

```
##  [1] "| sales_price| square_feet| swimming_pool| air_conditioning|"
##  [2] "|-----------:|-----------:|-------------:|----------------:|"
##  [3] "|      360000|        3032|             0|                1|"
##  [4] "|      340000|        2058|             0|                1|"
##  [5] "|      250000|        1780|             0|                1|"
##  [6] "|      205500|        1638|             0|                1|"
##  [7] "|      275500|        2196|             0|                1|"
##  [8] "|      248000|        1966|             1|                1|"
##  [9] "|      229900|        2216|             0|                1|"
## [10] "|      150000|        1597|             0|                1|"
```

Part 1 - Regression using a Dummy Variable

1a. Estimated regression equation from regressing sales price on swimming pool only.

```r
#Regressing sales price only on swimming pool dummy variable
pool_only_lm <- lm(sales_price ~ swimming_pool, data = house_data)

#summarizing linear model
summary(pool_only_lm)
```

```
##
## Call:
```

```
## lm(formula = sales_price ~ swimming_pool, data = house_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -188396  -94396  -46896   52604  647604
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     272396       6195   43.97  < 2e-16 ***
## swimming_pool    79724      23589    3.38  0.00078 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 136600 on 520 degrees of freedom
## Multiple R-squared:  0.02149,    Adjusted R-squared:  0.01961
## F-statistic: 11.42 on 1 and 520 DF,  p-value: 0.0007799
```

Estimated Regression model:

$\hat{Y} = 272396 + 79724X$

1b. Interpretation of estimated intercept and slope.

Intercept: B_0 = 272396

The estimated mean Y-value when X = 0 (reference/baseline group) is 272396. When put in context, the mean sales price of a house when the property **does not** contain a swimming pool is estimated to be $272,396.

Slope: B_1 = 79724

The slope of 79724 in our model indicates the change for the sales price of a property **containing** a swimming pool, **relative** to a property **without** a swimming pool to be $352,120.

The calculations of these coefficients can be represented in this table.

Table 1: Property Sales Price With & Without Swimming Pool

$\hat{Y} = B_0 + B_1X_1$	Swimming Pool = No	Swimming Pool = Yes
$\hat{Y} = 77.375 + 8.750X$	$\hat{Y} = 272396 + 79724(0)$ $= 272396$	$\hat{Y} = 272396 + 79724(1)$ = 272396+ 79724
Estimated Mean Sales Price	**$272,396**	**$352,120**

1c. Hypothesis testing on the significance of the slope coefficient.

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.

Testing the significance of a property **with** a swimming pool ($\hat{\beta}_1 = 79724$)

Conclusion and Decision Rule using p-value:

Because the **p-value** for having a swimming pool is [1] 0.00078 and is significantly smaller than $\alpha = 0.05$, we reject our NULL hypothesis and conclude that our partial slope, that a property **containing** a swimming pool in reference to one **without a swimming pool**, shows statistical significance in our model.

Part 2 - Fitting a MLR model with the Interaction Term of a Dummy and Continuous Variable

2a. Regressing sales price on the (1)swimming pool dummy variable, (2)area of residence, and the (3)interaction between these two variables.

```r
pool_sqft_only_lm <- lm(sales_price ~ swimming_pool +
                                      square_feet +
                                      swimming_pool * square_feet,
                                      data = house_data)
summary(pool_sqft_only_lm)
```

```
##
## Call:
## lm(formula = sales_price ~ swimming_pool + square_feet + swimming_pool *
##     square_feet, data = house_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -247193  -40579   -7542   24476  384051
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -88538.996  12063.237  -7.340 8.34e-13 ***
## swimming_pool             105909.972  47262.735   2.241   0.0255 *
## square_feet                  161.910      5.168  31.331  < 2e-16 ***
## swimming_pool:square_feet    -37.213     17.102  -2.176   0.0300 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78890 on 518 degrees of freedom
## Multiple R-squared:  0.6747, Adjusted R-squared:  0.6728
## F-statistic: 358.1 on 3 and 518 DF,  p-value: < 2.2e-16
```

Estimated regression equation for each kind of property:

$\hat{Y} = -88538.996 + 105909.972X + 161.910Y - 37.213(X * Y)$

note variables:

X = Swimming pool

Y = Square feet

X * Y = Interaction of swimming pool and square feet

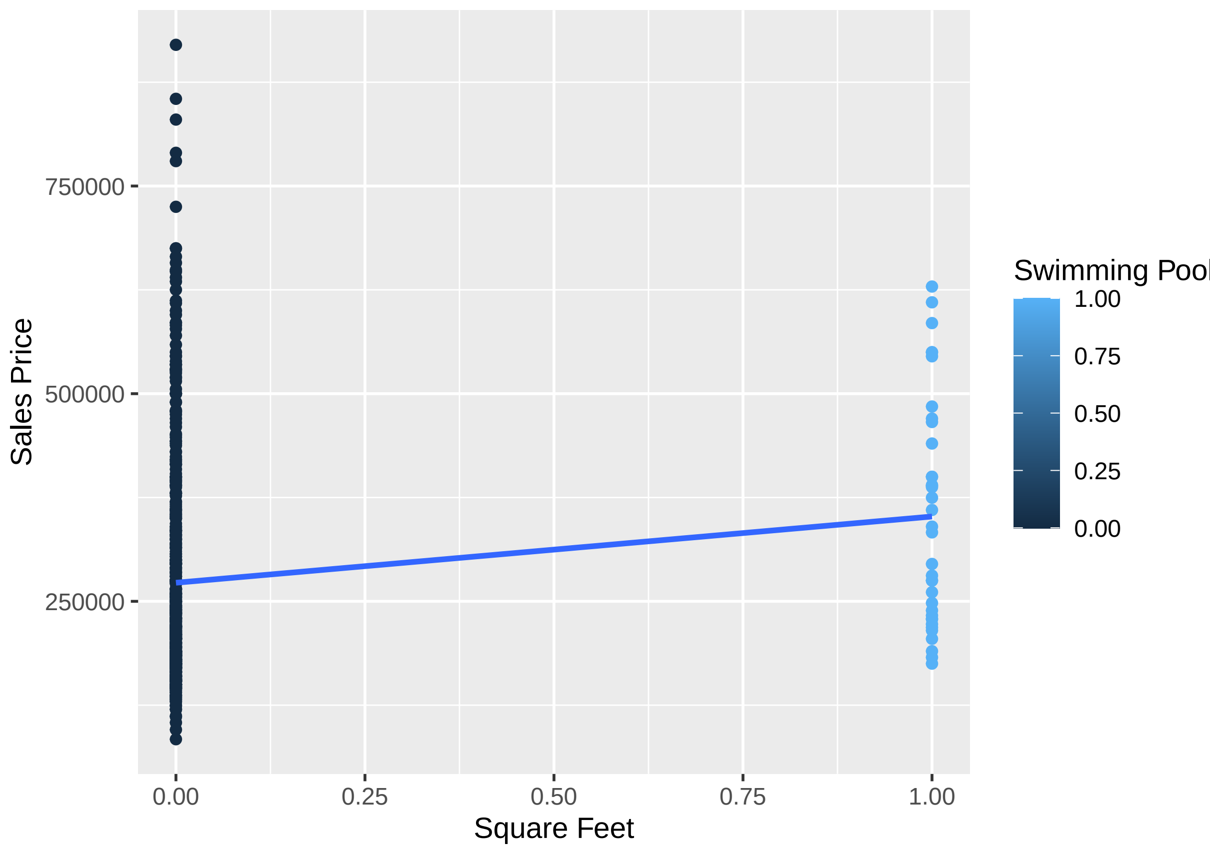
Table 2: Calculating Estimated Regression Equations for Properties With and Without Pools

$\hat{Y} = B_0 + B_1X + B_2Y + B_3(X*Y)$	Swimming Pool = No	Swimming Pool = Yes
$\hat{Y} = -88538.996 + 105909.972X + 161.910Y - 37.213(X*Y)$	$\hat{Y} = -88538.996 + 105909.972(0) + 161.910Y - 37.213(0*Y)$	$\hat{Y} = -88538.996 + 105909.972(1) + 161.910Y - 37.213(1*Y)$
	$= -88538.996 + 161.910Y$	$= -88538.996 + 105909.972 + 161.910Y - 37.213(Y)$ $= 17370.976 + 124.697Y$
Estimated Regression Equations	$= -88538.996 + 161.910Y$	$= 17370.976 + 124.697Y$

2b. Plotting Fitted regression lines

```
#no_pool <- house_data %>%
           #select(swimming_pool == 0)
# Code to plot regression equations model:
ggplot(pool_sqft_only_lm, aes(x = swimming_pool, y = sales_price, color = swimming_pool)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Square Feet", y = "Sales Price", color = "Swimming Pool")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Part 3 - MLR Only with the Interaction of Dummy Variables