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)
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)
```

```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
```

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(sales_price ~ swimming_pool, data = house_data)

#summarizing linear model
summary(pool_only)
```

```
##
## 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
```

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

$\hat{Y} = B_0 + B_1 X_1$	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**

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

Part 3 - MLR Only with the Interaction of Dummy Variables