PREDICTING HOME SALES PRICES USING MULTIPLE LINEAR REGRESSION

STAT 511

Rumil Legaspi, Rumil.legaspi@gmail.com
Efe Umukoro, Eumukoro20@apu.edu
Solange Ebobisse Mapenya, bebobissemapenya20@apu.edu

4/28/2021

Table of 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 (MLR)** from the Real Estate Sales (APPENC07) dataset that was published in 2002. We aim to observe and predict the relationships 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***.

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

```
## # A tibble: 6 x 4
##   sales_price square_feet swimming_pool air_conditioning
##         <int>       <int>         <int>            <int>
## 1      360000        3032             0                1
## 2      340000        2058             0                1
## 3      250000        1780             0                1
## 4      205500        1638             0                1
## 5      275500        2196             0                1
## 6      248000        1966             1                1
```

Part 1 - Regression using a Dummy Variable

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

Estimated Regression model:

The linear regression model of sales price on the dummy variable, swimming pool is

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

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 that the estimated mean change for the sales price of a property **containing** a swimming pool, **relative** to a property **without** a swimming pool to be $79,724.

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

Property Sales Price With & Without Swimming Pool

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

1c. Hypothesis test 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 ($\widehat{\beta_1} = 79724$)

Conclusion and Decision Rule using p-value:

Looking at our model summary, we see that the **p-value** for owning a swimming pool is [1] 0.00078 which is less than the 0.05 significance level. As a result of this, we reject our NULL hypothesis and conclude with our alternative hypothesis. This means that there exists a significant difference between the mean change in the sales prices comparing properties **containing** a swimming pool in reference to one **without a swimming pool.**

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.

```
## 
## 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:

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

Variable Assignment:

X = Swimming pool

Y = Square feet

X * Y = Interaction of swimming pool and square feet

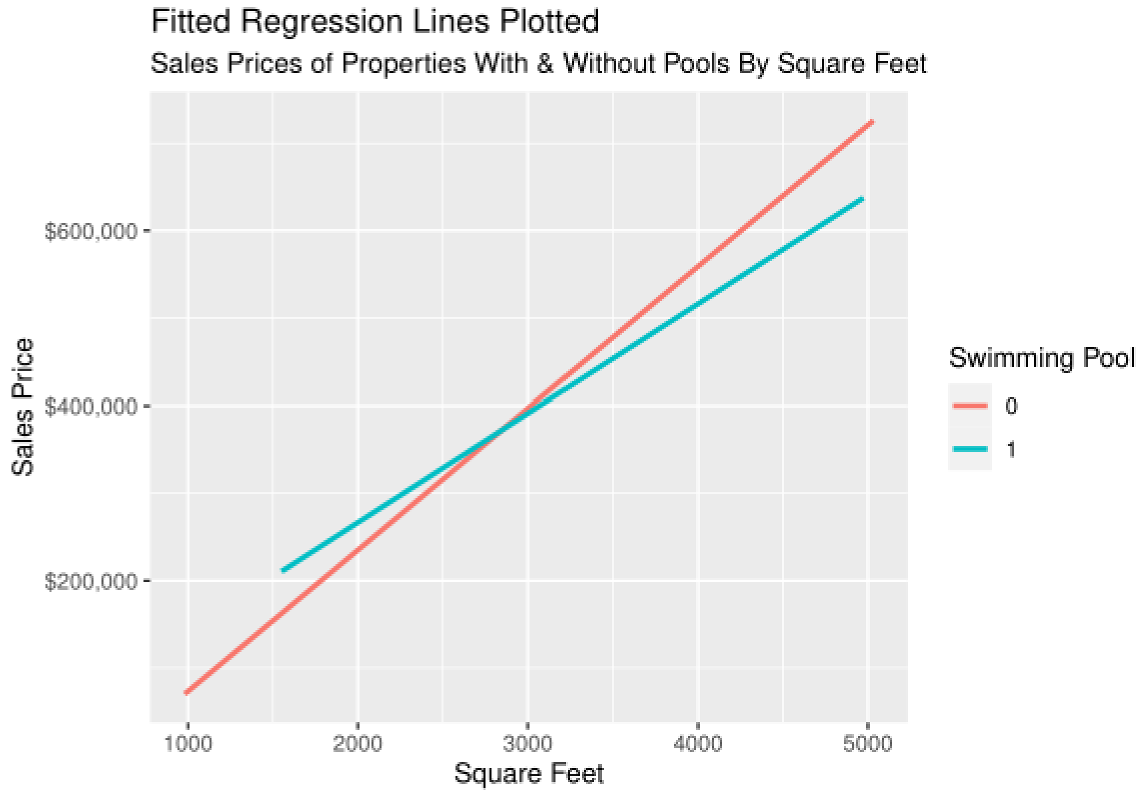
Calculating Estimated Regression Equations for Properties With and Without Pools

$\hat{Y} = B_0 + B_1X + B_2Y + B_3(X * Y)$	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

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



To find the value at which these two lines intersect algebraically we can set both equations equal to one another, solving for one variable, then plugging that back into the equation to get the other variable to obtain the coordinates.

The point of intersection of these two lines are when the values of: **Square feet is 2846.04 and sales price = 372264.58**.

2c. Testing if the two regression lines are parallel.

Using a significance level of $\alpha = 0.05$.

Null Hypothesis: H_0: $\beta_c = 0$ Partial slope of the interaction is 0.

Alternative Hypothesis: H_1: $\beta_c \neq 0$ Partial slope of the interaction is not 0.

Testing the significance of a property **with** a swimming pool ($\widehat{\beta_3} = -37.213$)

Conclusion and Decision Rule using p-value:

Based on the model summary, we see that the p-value of our interaction term is [1]0.0300. Since the p-value is less than the significance level, we **reject the** NULL hypothesis and conclude with our alternative hypothesis. This highlights that our regression lines are not parallel and a relationship exists between the two lines (because the interaction coefficient is not 0).

Part 3 - MLR Only With the Interaction of Dummy Variables

3a. Fitting a MLR on both swimming pool and AC dummy variables and find the estimated regression equation.

```
##
## Call:
## lm(formula = sales_price ~ swimming_pool + (air_conditioning) +
##     swimming_pool * air_conditioning, data = house_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -181752  -92704  -35504   44546  629546
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     189578.2    14087.8  13.457  < 2e-16 ***
## swimming_pool                      421.8   132154.9   0.003    0.997
## air_conditioning                100875.8    15548.0   6.488 2.03e-10 ***
## swimming_pool:air_conditioning   65876.5   134169.7   0.491    0.624
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 131400 on 518 degrees of freedom
## Multiple R-squared:  0.09756,    Adjusted R-squared:  0.09233
## F-statistic: 18.67 on 3 and 518 DF,  p-value: 1.642e-11
```

Estimated regression equation for each kind of property:

$$\hat{Y} = 189578.2 + 421.8X + 100875.8Z - 65876.5(X * Z)$$

Variable Assignment:

X = Swimming pool

Z = Air conditioning

X * Z = Interaction of swimming pool and air conditioning

3c. Calculating estimated mean sales prices for 4 types of properties using estimated regression equation:

1. No swimming pool and no AC

```
## [1] 189578.2
```

The estimated mean sales price of a property without a swimming pool and AC is $189,578.2.

2. No swimming pool and has AC

```
## [1] 290454
```

The estimated mean sales price of a property without a swimming pool but has AC is $290,454.

3. Has swimming pool and no AC

```
## [1] 190000
```

The estimated mean sales price of a property with a swimming pool and no AC is $190,000.

4. Has swimming pool and has AC

```
## [1] 356752.3
```

The estimated mean sales price of a property with a swimming pool and AC is $356,752.3.

Conclusion & Section Summary

In analyzing home sale prices our goal was to establish whether the variables coupled with the interaction variables impacted the average sales price. The variables studied, which were included in our model, include swimming pool, AC and swimming pool*AC. In order to establish whether these variables had an impact on the average sale price we studied the regression utilizing the dummy variable, created a Multi-Linear model that contains the interaction term of a dummy variable and a continuous variable, and exploited the Multi-Linear regression model with only the interaction of dummy variables.

Part 1

In the first part of our analytical study, a regression model that utilized the dummy variable was created. Our goal was to establish whether the slope coefficient is significant. The slope of $79724 in our model indicates that the estimated mean change for the sales price of a property containing a swimming pool, relative to a property without a swimming pool to be $79,724. Based on our two-sided hypothesis test, we concluded that the slope coefficient was in fact significant. This tells us that a significant difference between the mean change in the sales prices comparing properties containing a swimming pool in reference to one without a swimming pool does in fact exist.

Part 2

In the second part of our analytical study, a multi-linear regression model that contains the interaction term of the dummy variable and the continuous variable was created. The goal of this was to establish whether the different linear regression models that resulted when the home contained a pool or did not contain a pool. We also looked at the effect that this would have dependent on the square footage of the property. The model revealed that our regression lines are not parallel, and a relationship exists between the two lines.

Part 3

Our last analytical study included a multi-linear regression model that contained the dummy variable only. The goal of this model was to establish the mean price of the property given that it may have contained or lacked a pool and an AC. Based on our regression analysis it is clear that a property with a swimming pool and air conditioning ($356,752.3), cost significantly more than a property without ($189,578.2.).

Looking Forward

For future reference we understand that our dataset is unbalanced with only about 7%, or 36 out of 522 observations owning swimming pools and 16%, or 88 out of 522 observations having air conditioning. Moving forward one way to correct this would be to collect more data from houses containing these features. Also, since the Interaction term between owning a swimming pool and having air conditioning is not significant and therefore remove this from the model.