STAT 511: HW #7 Multiple Linear Regression With Dummy Variables

Rumil Legaspi

4/26/2021

Contents

Multiple Regression & Brand Preference Dataset

Setting up workspace

```
library(tidyverse)

setwd("C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Analysis)/Week 10/Week 10/")

brand_data = read.table(file = "Brand.txt", header = FALSE, sep = "")

View(brand_data)

# #Adding headers
names(brand_data) <- c("Rating", "Moisture", "Sweetness")

# names(bank_data) <- c("", "")

#Defining dependent and independent vars
Rating = brand_data$Rating #Y
Moisture = brand_data$Moisture #X1
Sweetness = brand_data$Sweetness #X2
```

1a. Regress degree of brand liking on sweetness only. Write down the estimated regression model.

```
#Coding Sweetness as a dummy variable
cat_sweetness <- as.factor(Sweetness)
cat_sweetness
```

```
##  [1] 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4
## Levels: 2 4
```

```
#Regressing Rating on new Sweetness dummy variable
Sweetness_only <- lm(Rating ~ cat_sweetness, data = brand_data)

summary(Sweetness_only)
```

```
##
## Call:
## lm(formula = Rating ~ cat_sweetness, data = brand_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.375  -7.312  -0.125   8.688  16.625
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        77.375      3.851  20.094 1.01e-11 ***
```

```
## cat_sweetness4    8.750      5.446   1.607     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.89 on 14 degrees of freedom
## Multiple R-squared:  0.1557, Adjusted R-squared:  0.09539
## F-statistic: 2.582 on 1 and 14 DF,  p-value: 0.1304
```

Estimated Regression model:

$\hat{Y} = 77.375 + 8.750X$

```
#Checking how many levels there are in sweetness as a dummy
levels(cat_sweetness)
```

```
## [1] "2" "4"
```

```
typeof(cat_sweetness)
```

```
## [1] "integer"
```

```
attributes(cat_sweetness)
```

```
## $levels
## [1] "2" "4"
##
## $class
## [1] "factor"
```

So in this case since sweetness is a category with only 2 levels (sweetness level 2 & sweetness level 4) We will use sweetness level 2 as our reference group

```
#the "2" indicates the level marked as "2" given from our level() function and now setting sweetness le
cat_sweetness_new <- relevel(cat_sweetness, ref = "2")

#Regressing
Sweetness_only2 <- lm(Rating ~ cat_sweetness_new, data = brand_data)

summary(Sweetness_only2)
```

```
##
## Call:
## lm(formula = Rating ~ cat_sweetness_new, data = brand_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.375  -7.312  -0.125   8.688  16.625
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            77.375      3.851  20.094 1.01e-11 ***
```

```
## cat_sweetness_new4    8.750      5.446   1.607     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.89 on 14 degrees of freedom
## Multiple R-squared:  0.1557, Adjusted R-squared:  0.09539
## F-statistic: 2.582 on 1 and 14 DF,  p-value: 0.1304
```

1b. Compute the estimated mean degree of brand liking at each level of sweetness, i.e., what is

the estimated mean degree of brand liking at sweetness level 2? At sweetness level 4?

After reading the summary we can now say the estimated means when Sweetness level is 2 and 4 to be.

Table 1: Mean Rating at Level 2 or Level 4 Sweetness

$\hat{Y} = B_0 + B_1 X_1$	Sweetness = Level 2	Sweetness = Level 4
$\hat{Y} = 77.375 + 8.750X$	$\hat{Y} = 77.375 + 8.750(0)$ = 77.375 77.375	$\hat{Y} = 77.375 + 8.750(1)$ = 77.375 + 8.750 86.125

1c. Interpret the intercept coefficient.

$B_0 = 77.375$

The estimated mean Y-value when X = 0 (our reference/baseline group) is 77.375. When put in context, this represents the mean Rating when the brand's sweetness level is 2.

1d. Interpret the slope coefficient.

The change in mean Rating for Sweetness level 4 relative to sweetness level 2 is 86.125 (77.375 + 8.750(1)).

1e. Is the slope coefficient significant? State the null, alternative, decision rule and conclusion.

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 Sweetness Level 4 ($\hat{\beta}_1 = 8.750$) p-value:

Because the p-value for sweetness level 4 is [1] 0.13 and is greater than our alpha (accepted error/significance level) of 0.05 we **fail to reject** our NULL hypothesis and conclude that our partial slope, **Sweetness Level 4** in reference to Sweetness Level of 2, does not show significance in our model.

2. Refer to the "Brand Preference" dataset. Code sweetness (X_2) as a dummy variable.

2a. Fit a multiple regression model with moisture content, sweetness, and their interaction.

```r
#Regressing Rating on Moisture and with Sweetness still as a dummy variable
full_lm <- lm(Rating ~ cat_sweetness + Moisture + cat_sweetness*Moisture, data = brand_data)

summary(full_lm)
```

```
##
## Call:
## lm(formula = Rating ~ cat_sweetness + Moisture + cat_sweetness *
##     Moisture, data = brand_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.150 -1.488  0.125  1.700  3.700
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              42.9000     2.8912  14.838 4.40e-09 ***
## cat_sweetness4           15.7500     4.0887   3.852   0.0023 **
## Moisture                  4.9250     0.3934  12.518 3.02e-08 ***
## cat_sweetness4:Moisture  -1.0000     0.5564  -1.797   0.0975 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.488 on 12 degrees of freedom
## Multiple R-squared:  0.9622, Adjusted R-squared:  0.9528
## F-statistic: 101.9 on 3 and 12 DF,  p-value: 8.379e-09
```

2b. Write down the estimated regression equation at each sweetness level

S = Sweetness

M = Moisture

$\hat{Y} = 42.90 + 15.75S + 4.925M - 1.0(S * M)$

Sweetness level 2: $\hat{Y} = 42.90 + 4.925M$

Sweetness level 4: $\hat{Y} = 58.68 + 3.925M$

2c. Interpret the slope coefficient in each estimated regression equation in Part (b).

$M = 4.925$: The mean Rating when sweetness is level 2 for a 1 unit change in Moisture content is an increases by 4.925.

$M = 3.925$: The increase in mean Rating when sweetness is 4 for a 1 unit change in Moisture content is an increase by 3.925.

2d. Is the interaction coefficient significant at = 5%? State the null, alternative, decision rule, conclusion.

Because the p-value for sweetness level 4 is [1] 0.0975 and is greater than our alpha (accepted error/significance level) of 0.05 we **fail to reject** our NULL hypothesis and conclude that our partial slope, **Sweetness Level 4** in reference to Sweetness Level of 2, does not show significance in our model.

2e. your answer is NO in Part (d), drop the interaction term and rerun the model. Write down the new estimated regression equation at each sweetness level.

```r
no_inter_lm <- lm(Rating ~ cat_sweetness + Moisture, data = brand_data)

summary(no_inter_lm)
```

```
##
## Call:
## lm(formula = Rating ~ cat_sweetness + Moisture, data = brand_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     46.4000     2.3129  20.061 3.66e-11 ***
## cat_sweetness4   8.7500     1.3466   6.498 2.01e-05 ***
## Moisture         4.4250     0.3011  14.695 1.78e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

Full estimated regression equation: $y = 46.4 + 8.75S + 4.425M$

Variables Assignment:

S = Sweetness Level

M = Moisture Content

Sweetness level 2: $y = 46.4 + 4.425M$

Sweetness level 4: $y = 55.15 + 4.425M$

3. Refer to the "Assessed Valuations" dataset (Value.txt)

```r
value_data = read.table(file = "C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Ana

View(value_data)

# #Adding headers
names(value_data) <- c("Price", "Valuation", "Corner_lot")

# names(bank_data) <- c("", "")

#Defining dependent and independent vars
Price = value_data$Price #Y
Valuation = value_data$Valuation #X1
Corner_lot = value_data$Corner_lot #X2
```

3a. Regress selling price on lot location only. Write down the estimated regression equation.

```r
#Coding Corner Location as a dummy variable
cat_corner_lot <- as.factor(Corner_lot)

#For good measure I set 0 as base reference group
cat_corner_lot <- relevel(cat_corner_lot, ref = "0")

location_onlylm <- lm(Price ~ Corner_lot, data = value_data)
summary(location_onlylm)
```

```
##
## Call:
## lm(formula = Price ~ Corner_lot, data = value_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.854  -5.637   1.157   6.196  16.446
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   81.154      1.198  67.715  < 2e-16 ***
## Corner_lot    -8.523      2.397  -3.556 0.000728 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.303 on 62 degrees of freedom
## Multiple R-squared:  0.1694, Adjusted R-squared:  0.156
## F-statistic: 12.64 on 1 and 62 DF,  p-value: 0.0007283
```

y = 81.154 - 8.5C

3b. Based on your regression result in Part (a), what is the estimated mean selling price for corner lots? For non-corner lots?

Estimated mean selling price for non-corner lots:

The estimated mean selling price for non-corner lots is roughly $81K

For corner lots:

The estimated mean selling price for corner lots is roughly $72.5k, which is taken by 81k - 8.5k.

3c. Based on your regression result in Part (a), what is the estimated difference in selling price between corner and non-corner lots? Is this difference statistically significant?

The estimated mean selling price between corner and non-corner lots is $8.5k. Looking at the p-value of 0.000728 we can see that the difference is statistically significant.

3d. Regress selling price on assessed valuation, lot location, and their interaction. Write down the estimated regression equation for corner lots, and for non-corner lots respectively

```r
#regressing on valuation, lot location, plus their interaction
val_lot_interlm <- lm(Price ~ Corner_lot + Valuation + Corner_lot*Valuation, data = value_data)
summary(val_lot_interlm)
```

```
##
## Call:
## lm(formula = Price ~ Corner_lot + Valuation + Corner_lot * Valuation,
##     data = value_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8470  -2.1639   0.0913   1.9348   9.9836
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -126.9052    14.7225  -8.620 4.33e-12 ***
## Corner_lot            76.0215    30.1314   2.523  0.01430 *
## Valuation              2.7759     0.1963  14.142  < 2e-16 ***
## Corner_lot:Valuation  -1.1075     0.4055  -2.731  0.00828 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.893 on 60 degrees of freedom
## Multiple R-squared:  0.8233, Adjusted R-squared:  0.8145
## F-statistic: 93.21 on 3 and 60 DF,  p-value: < 2.2e-16
```

C - Corner or non corner lots

V - Valuation

Estimated regression equation for **non corner lots**:

$\hat{Y} = -126.9052 + 2.7759V$

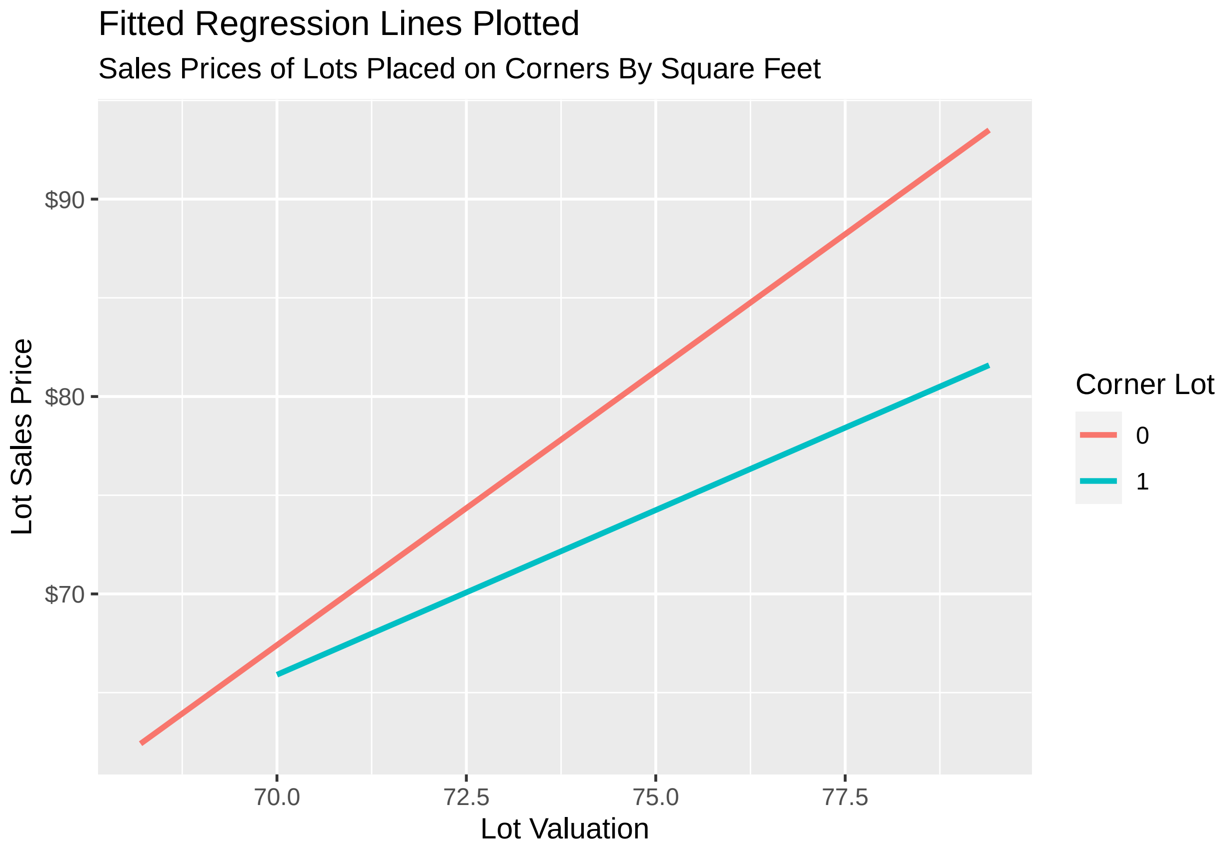
and for **corner lots** respectively:

$\hat{Y} = -50.8837 + 1.67V$

3e. Plot the estimated regression lines for the two groups and describe their differences

Plotting Fitted Regression Lines

```r
#plot fitted regression lines
#Save estimated coefficients
#Coef = val_lot_interlm$coefficients
#Coef

#Non corner lots: When Corner lots takes value 0
#price1 = Coef[1] + Coef[3] * Valuation

#corner lots: When Corner lots takes value 1
#price2 = Coef[1] + Coef[3] * Valuation + Coef[2] +Coef[4] * Valuation

#plotting plot()
#plot(Valuation, price1, type = 'l', col = "blue",
    # xlab = "Valuation" , ylab = "Price")
#lines(Valuation, price2, type = 'l', col = "red")

#legend
#legend("topleft", legend = c("Line 1", "Line 2"),
       #col = c("blue", "red"), lty = 1, cex = 0.5)
#--------------------------------------------------------------
#Plotting ggplot way with similar results
plot_coef <- value_data %>%

  ggplot(aes(x = Valuation, y = Price, color = as.factor(Corner_lot))) +

  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Lot Valuation", y = "Lot Sales Price", color = "Corner Lot",
        title = "Fitted Regression Lines Plotted",
       subtitle = "Sales Prices of Lots Placed on Corners By Square Feet")+
  scale_y_continuous(labels = scales::dollar)

plot_coef
```



3f. 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 the interaction term ($\hat{\beta}_3 = -1.1075$)

Conclusion and Decision Rule using p-value:

Looking at our model summary, we see that the p-value of our interaction term is [1]0.00828 which means we reject NULL hypothesis and conclude with our alternative hypothesis and that our regression lines are not parallel and a relationship exists between the two lines (because the interaction coefficient is not 0).