Regression and Endogeneity

Jaskaran Singh Kohli

Part I Regression Basics

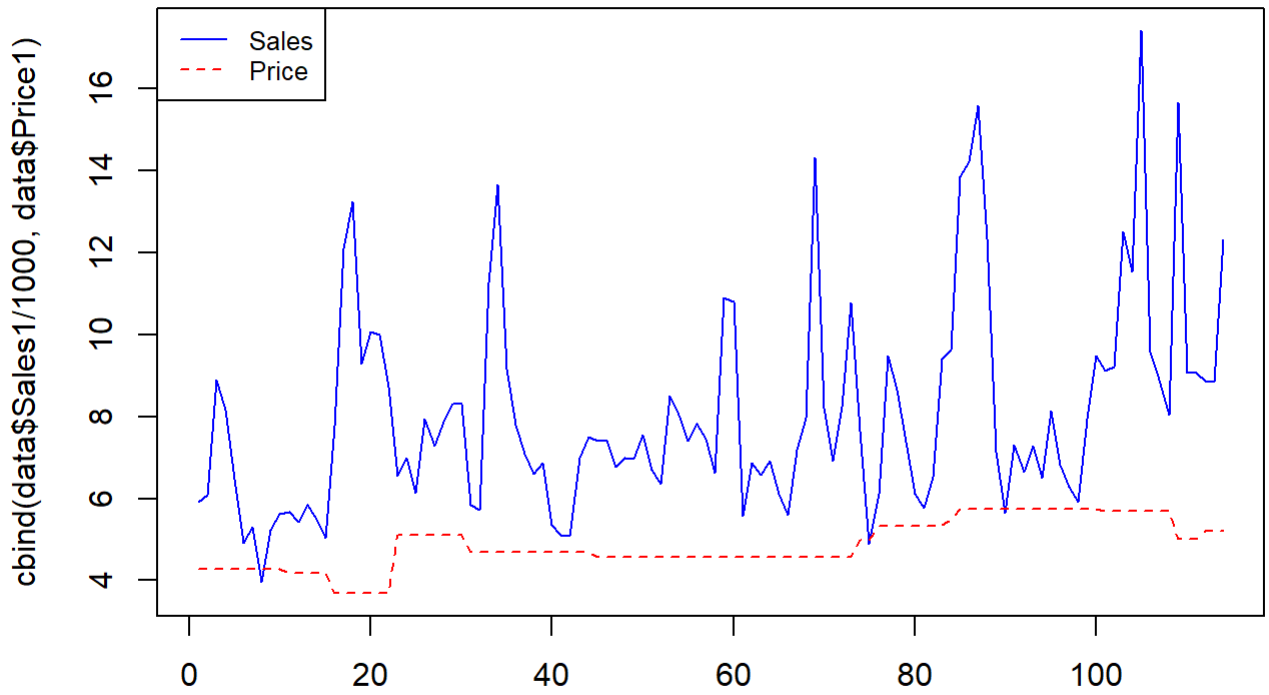
Follow the steps below:

1. Put the data and this file in a folder, and set it as your working folder through `setwd()`

```
#setwd("~Documents/Coffee_inClass.csv")
```

2. Read in the data file `Coffee_inClass.csv` , and run a regression analysis try to answer the question “how price influence sales”? You can try different model specificatoin, but only leave the final version of your code here. Make sure you include some dummy variables, and interactions between some dummy with other variables.

```
# Importing the dataset
data = read.csv("Coffee_inClass.csv",header=TRUE)
#Plotting the Data and Scaling down sales by 1000
matplot(cbind(data$Sales1/1000,data$Price1),type="l",col=c("blue","red"))
legend('topleft',c("Sales","Price"),lty=1:2,col=c("blue","red"),cex=0.8)
```



```
# Encoding categorical data
data$dayofweek = factor(data$dayofweek)
levels(data$dayofweek)
```

```
## [1] "1" "2" "3" "4" "5" "6" "7"
```

```
levels(data$dayofweek) =levels=c('Monday','Tuesday','Wednesday','Thursday','Friday','Saturday',
'Sunday')
# Fitting Multiple Linear Regression to the Training set
regressor = lm(formula = Sales1 ~Price1 + feat1  + dayofweek + dayofweek*Price1 ,data = data)
summary(regressor)
```

```
##
## Call:
## lm(formula = Sales1 ~ Price1 + feat1 + dayofweek + dayofweek *
##     Price1, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4231.5 -1236.0  -277.6   830.5  6203.4
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1711.007   4587.887  -0.373   0.7100
## Price1                     1757.298    929.668   1.890   0.0617 .
## feat1                        55.773      8.979   6.211 1.24e-08 ***
## dayofweekTuesday           1154.119   6287.828   0.184   0.8547
## dayofweekWednesday        -1584.989   6435.437  -0.246   0.8060
## dayofweekThursday          8489.655   6414.101   1.324   0.1887
## dayofweekFriday           11627.587   6602.723   1.761   0.0813 .
## dayofweekSaturday          5889.904   6585.982   0.894   0.3733
## dayofweekSunday            4972.143   6567.069   0.757   0.4508
## Price1:dayofweekTuesday    -158.840   1276.967  -0.124   0.9013
## Price1:dayofweekWednesday   367.518   1304.106   0.282   0.7787
## Price1:dayofweekThursday  -1692.079   1302.097  -1.300   0.1968
## Price1:dayofweekFriday    -2215.400   1346.161  -1.646   0.1030
## Price1:dayofweekSaturday  -1244.752   1343.193  -0.927   0.3563
## Price1:dayofweekSunday    -1040.077   1336.058  -0.778   0.4382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2199 on 99 degrees of freedom
## Multiple R-squared:  0.3567, Adjusted R-squared:  0.2658
## F-statistic: 3.922 on 14 and 99 DF,  p-value: 2.604e-05
```

```
# Predicting the values on the same model
y_Pred = predict(regressor, newdata =data)
```

3. List what are the control variables (including dummy variables, and interactions) included in the model? Explain for each control variable, why it needs to be included?

Control Variables

feat1

As we know that feat1 data shows the percentage of particular coffee being featured .

dayofweek

This is the dummy variable showing the sales on specific day on compared to reference day in our case which is Monday .

daysofweek*price

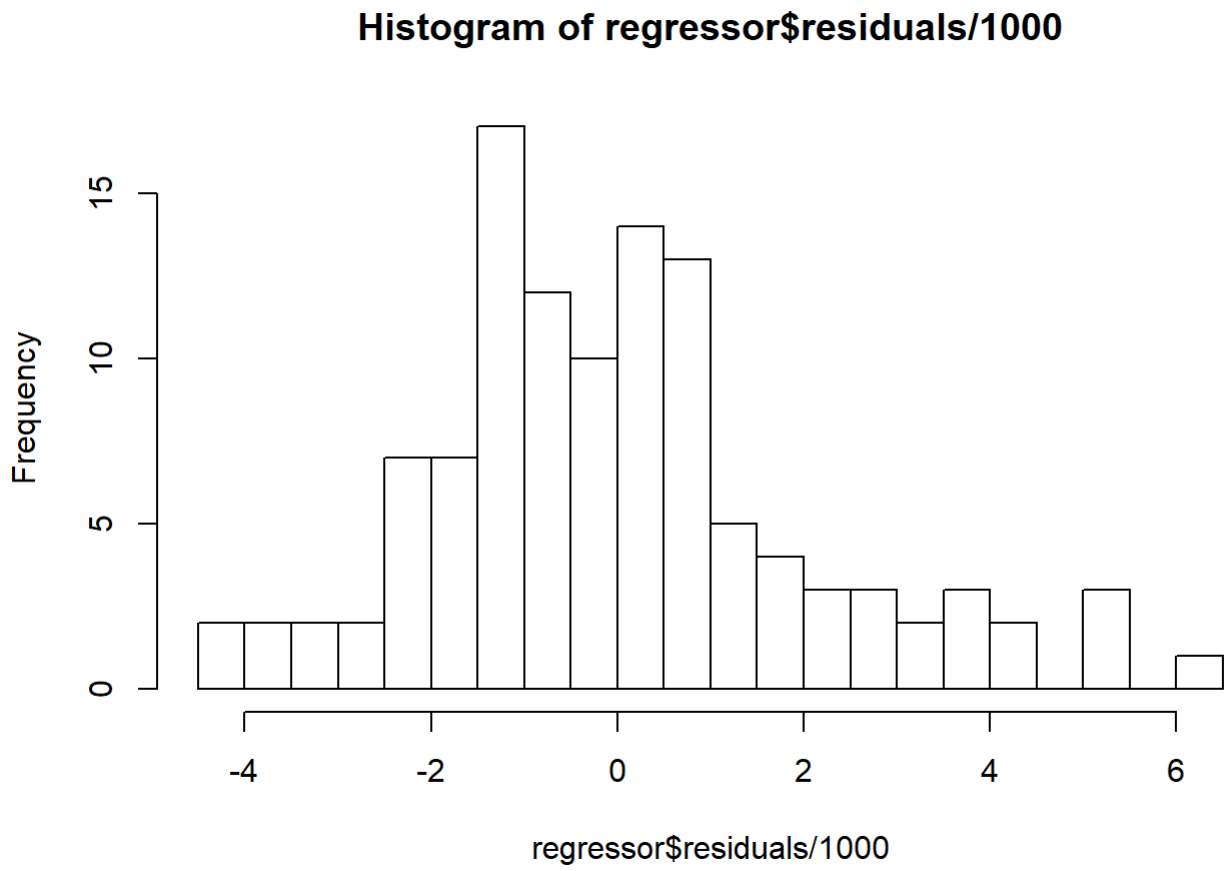
Days of the Week (Categorical Variable/Dummy Variable) shows the sales of coffee on a specific day. We need this control variable to check the impact of price on Sales on a Specifc day . So I am using interaction variables to sove this problem .

Key independent Variables

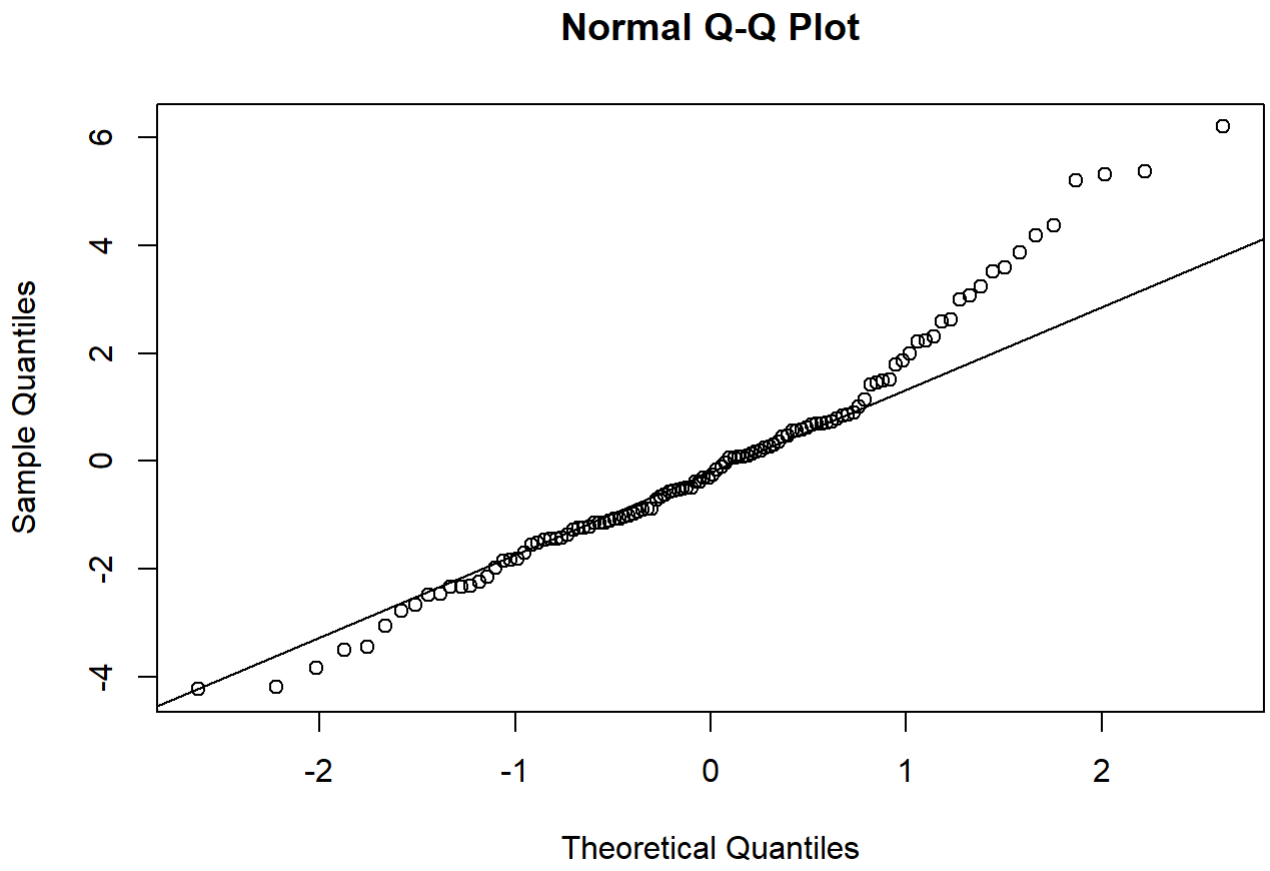
This is our most important control variable as we need to find out how our price is effecting sales.When we run a regression , we can see that price has a positive impact (Directly Proportional)on sales which is ideally not possible.

4. Plot the residuals, and comment on the residules, are they ideal? Any concerns?

```
hist(regressor$residuals/1000,20)
```




```
{qqnorm(regressor$residuals/1000)
 qqline(regressor$residuals/1000)}
```



When I analyzed regressor$residual histogram, I could see that our model needs concern regarding the prediction of values.The regression is not showing Normal Distribution(NOT SYMMETRIC) and then I thought that we can use a plot of standardized residuals (QQ Plot) versus predicted values can show whether points are equally distributed across all values of the independent variables. It is not following a straight line hence our model is not accurate .

5. How do you interpret each of the parameter estimates? Make sure your interpretation of each estimates include the values of the estimates, the standard error, the t-statistics and the p-value. Be careful with the dummy variables and the interaction variables?

```
a = summary(regressor)
Parameter_Estimates = a$coefficients
Parameter_Estimates
```

```
##                              Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)               -1711.00683 4587.886628 -0.3729401 7.099903e-01
## Price1                     1757.29789  929.668277  1.8902419 6.165048e-02
## feat1                        55.77254    8.979422  6.2111508 1.242835e-08
## dayofweekTuesday           1154.11922 6287.828338  0.1835481 8.547434e-01
## dayofweekWednesday        -1584.98933 6435.437228 -0.2462909 8.059670e-01
## dayofweekThursday          8489.65461 6414.100753  1.3235923 1.886869e-01
## dayofweekFriday           11627.58670 6602.722625  1.7610291 8.132117e-02
## dayofweekSaturday          5889.90431 6585.982073  0.8943092 3.733255e-01
## dayofweekSunday            4972.14276 6567.068825  0.7571327 4.507686e-01
## Price1:dayofweekTuesday    -158.84048 1276.967198 -0.1243889 9.012596e-01
## Price1:dayofweekWednesday   367.51753 1304.106499  0.2818156 7.786732e-01
## Price1:dayofweekThursday  -1692.07909 1302.097308 -1.2995028 1.967894e-01
## Price1:dayofweekFriday    -2215.39981 1346.161165 -1.6457166 1.029934e-01
## Price1:dayofweekSaturday  -1244.75241 1343.193196 -0.9267114 3.563301e-01
## Price1:dayofweekSunday    -1040.07736 1336.057958 -0.7784672 4.381503e-01
```

Price:

Parameter Estimates - According to our model the increase in price will increase sales which is practically not possible unless the product is a luxury good . Ideally , Sales should decrease when the price increase . Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 1.948,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

feat1 :

Parameter Estimates - According to our model feat1 has a positive impact on sales there for an increase in feature will lead to an increase in sales . Standard Error - Standard Error is low , indicating the sample is more representative to the actual population . t-statistics - Here the t value is 6.640,So, when t-stat >2 .The estimate is said to be positive and significant p-value - p values for price is less than 0.05 , So this indicate that the estimate is statistically significant.

dayofweekTuesday :

Parameter Estimates - Comparing it to the refernce point(Monday), Tuesday shows a negative impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain and notepresentative to the actual population . t-statistics - Here the t value is 0.209,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

dayofweekWednesday :

Parameter Estimates - Comparing it to the refernce point(Monday), Wednesday shows a negative impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is -0.252,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

dayofweekThursday :

Parameter Estimates - Comparing it to the refernce point(Monday), Thursday shows a postive impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 1.57,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

dayofweekFriday :

Parameter Estimates - Comparing it to the refernce point(Monday), Friday shows a postive impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 1.83,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

dayofweekSaturday :

Parameter Estimates - Comparing it to the refernce point(Monday), Saturday shows a postive impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 0.92,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

dayofweekSunday :

Parameter Estimates - Comparing it to the refernce point(Monday), Sunday shows a postive impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 0.74,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekTuesday :

Parameter Estimates - Comparing it to the refernce point(Monday), Price on Tuesday shows a negative impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain and notepresentative to the actual population . t-statistics - Here the t value is -0.141,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekWednesday :

Parameter Estimates - Comparing it to the refernce point(Monday),Price on Wednesday shows a positive impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 0.29,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekThursday :

Parameter Estimates - Comparing it to the refernce point(Monday), Price on Thursday shows a negative impact on sales
Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is -1.59,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekFriday :

Parameter Estimates - Comparing it to the refernce point(Monday), Price on Friday shows a negative impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is -1.71,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekSaturday :

Parameter Estimates - Comparing it to the refernce point(Monday), Price on Saturday shows a negative impact on sales . Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is -0.95,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for

price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekSunday :

Parameter Estimates - Comparing it to the refernce point(Monday), Price on Sunday shows a negative impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is -0.76,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

6. Based on the above estimation results, what's your answer to the question "how does price influence sales"? Price doesn't influence sales because the model show's that as we increase the price the sales should increase which is not the case and practically not possible . So when we combine the price with dayofweek , we are actually checking that change in price/day causing change in sales . We can use that for checking the exact influence of price on a specific day on sales . We see that at some days it is increasing and vice versa .
7. Comment on your model fit: R-squared, adjusted R-squared, F-statistics.

```
a$r.squared
```

```
## [1] 0.3567409
```

```
a$adj.r.squared
```

```
## [1] 0.2657749
```

```
a$fstatistic
```

```
##     value     numdf     dendf
##  3.921697 14.000000 99.000000
```

The *R-squared* value of 0.381 means the model explains about 38% of the variability in the response. The *adjusted R-squared* is a modified version of R-squared for the number of predictors in a model.Generally adjusted R-Square is slightly smaller than R square but its not true for our case as R adjusted for my model is 0.293 / 29.3% which is approximately 9% less than my R Square .So our model is not good . Lastly when we look for *fstatistics* it gives us a vague idea that our model is significant , but it wont help us predicting the significance of each variable .

8. In utilizing the dummy variables indicating the day of week, the above model has left one of the day-of-week dummy variable out. Now change the specification by leaving out a different day-of-week dummy variable (for example instead of leaving out the Monday dummy, now include the Monday dummy but leave out the Tuesday (or any other day) dummy). Please explain the changes in the estimates, standard errors of all the estimate.

```
data$dayofweek= relevel(data$dayofweek,ref="Tuesday")
data$dayofweek
```

```
##   [1] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##   [8] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [15] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [22] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [29] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [36] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [43] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [50] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [57] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [64] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [71] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [78] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [85] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [92] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
##  [99] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
## [106] Tuesday   Wednesday Thursday  Friday    Saturday  Sunday    Monday
## [113] Tuesday   Thursday
## Levels: Tuesday Monday Wednesday Thursday Friday Saturday Sunday
```

```
regressor = lm(formula = Sales1 ~Price1 + feat1  + dayofweek + dayofweek*Price1 ,data = data)
summary(regressor)
```

```
##
## Call:
## lm(formula = Sales1 ~ Price1 + feat1 + dayofweek + dayofweek *
##     Price1, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4231.5 -1236.0  -277.6   830.5  6203.4
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -556.888   4291.191  -0.130   0.8970
## Price1                     1598.457    874.631   1.828   0.0706 .
## feat1                        55.773      8.979   6.211 1.24e-08 ***
## dayofweekMonday           -1154.119   6287.828  -0.184   0.8547
## dayofweekWednesday        -2739.109   6246.084  -0.439   0.6620
## dayofweekThursday          7335.535   6198.330   1.183   0.2395
## dayofweekFriday           10473.467   6385.914   1.640   0.1042
## dayofweekSaturday          4735.785   6389.120   0.741   0.4603
## dayofweekSunday            3818.024   6429.535   0.594   0.5540
## Price1:dayofweekMonday      158.840   1276.967   0.124   0.9013
## Price1:dayofweekWednesday   526.358   1267.189   0.415   0.6788
## Price1:dayofweekThursday  -1533.239   1260.239  -1.217   0.2266
## Price1:dayofweekFriday    -2056.559   1305.321  -1.576   0.1183
## Price1:dayofweekSaturday  -1085.912   1305.243  -0.832   0.4074
## Price1:dayofweekSunday     -881.237   1305.607  -0.675   0.5013
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2199 on 99 degrees of freedom
## Multiple R-squared:  0.3567, Adjusted R-squared:  0.2658
## F-statistic: 3.922 on 14 and 99 DF,  p-value: 2.604e-05
```

Price:

Parameter Estimates - According to our model the increase in price will increase sales which is practically not possible unless the product is a luxury good . Ideally , Sales should decrease when the price increase . Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 1.863,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

feat1 :

Parameter Estimates - According to our model feat1 has a positive impact on sales there for an increase in feature will lead to an increase in sales . Standard Error - Standard Error is low , indicating the sample is more representative to the actual population . t-statistics - Here the t value is 6.640,So, when t-stat >2 .The estimate is said to be positive and significant p-value - p values for price is less than 0.05 , So this indicate that the estimate is statistically significant.

dayofweekMonday :

Parameter Estimates - Comparing it to the refernce point(Tuesday), Monday shows a negative impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain and notepresentative to the actual population . t-statistics - Here the t value is -0.209,So, when t stat values lies between -2 to 2 .The estimate is said

to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

dayofweekWednesday :

Parameter Estimates - Comparing it to the refernce point(Tuesday), Wednesday shows a negative impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is -0.470,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

dayofweekThursday :

Parameter Estimates - Comparing it to the refernce point(Tuesday), Thursday shows a postive impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 1.414,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

dayofweekFriday :

Parameter Estimates - Comparing it to the refernce point(Tuesday), Friday shows a postive impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 1.689,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

dayofweekSaturday :

Parameter Estimates - Comparing it to the refernce point(Tuesday), Saturday shows a postive impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 0.751,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

dayofweekSunday :

Parameter Estimates - Comparing it to the refernce point(Tuesday), Sunday shows a postive impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 0.556,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekMonday :

Parameter Estimates - Comparing it to the refernce point(Tuesday), Price on Monday shows a negative impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain and notepresentative to the actual population . t-statistics - Here the t value is 0.142,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekWednesday :

Parameter Estimates - Comparing it to the refernce point(Tuesday),Price on Wednesday shows a positive impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is 0.446,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekThursday :

Parameter Estimates - Comparing it to the refernce point(Monday), Price on Thursday shows a negative impact on sales
Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is -1.509,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekFriday :

Parameter Estimates - Comparing it to the refernce point(Tuesday), Price on Friday shows a negative impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is -1.627,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

Price1:dayofweekSaturday :

Parameter Estimates - Comparing it to the refernce point(Tuesday), Price on Saturday shows a negative impact on sales . Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is -0.847,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

####Price1:dayofweekSunday :

Parameter Estimates - Comparing it to the refernce point(Tuesday), Price on Sunday shows a negative impact on sales Standard Error - Standard Error is high , indicating the price data is uncertain . t-statistics - Here the t value is -0.646 ,So, when t stat values lies between -2 to 2 .The estimate is said to be statistically 0 p-value - p values for price is more than 0.05 , So this indicate that the estimate is statistically insignificant.

The new adjusted R square is similar to the old one , hence we’ll reject this model . Other variables are also same in comparison to the first model .