Homework 1 for Marketing Aanlytics

Regression and Endogeneity

Harsh Tandon

Due on Wednesday, January 15, 2020

Feel free to conduct your analysis using this R notebook file. Please Knit to pdf file, print it, and submit your hard copy before class on Wednesday, Jan. 15.

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()`

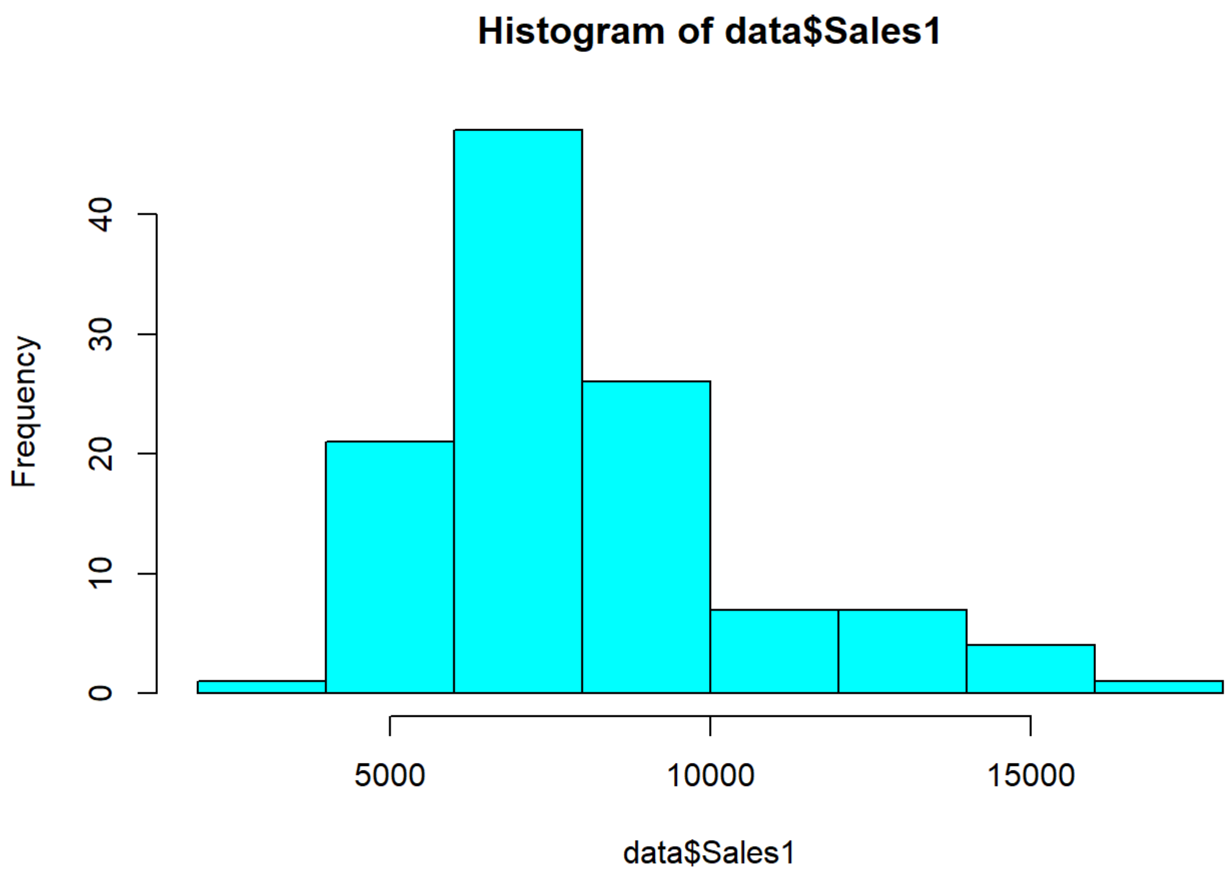
```
setwd("D:/2nd Qtr Study Material/Marketing Analytics/Lecture 2 Regression/Lecture 2 Regressio
n ")
#Install required packages
#install.packages("stargazer")
#install.packages("lmtest")
library(stargazer)
library(ggplot2)
library(graphics)
library(lmtest)
```

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 specification, 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.

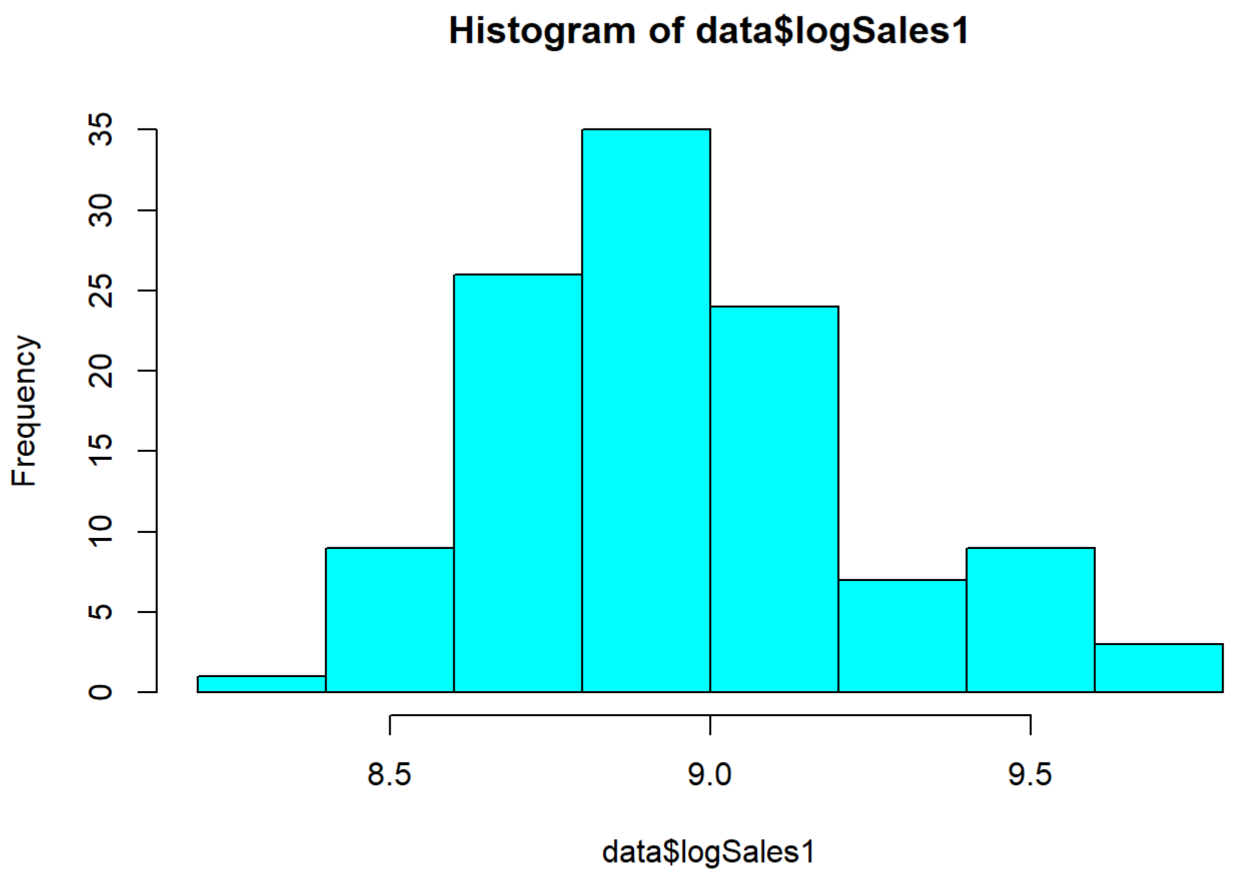
```
data = read.csv("Coffee_inClass.csv", header = TRUE) #read data

stargazer(data, type="text", median=TRUE, iqr=TRUE, digits=2, title="Descriptive Statistics")
#print summary statistics
```

```
##
## Descriptive Statistics
## ========================================================================
## Statistic  N    Mean   St. Dev.  Min  Pctl(25) Median Pctl(75)  Max
## ------------------------------------------------------------------------
## day       114  57.50    33.05     1     29.2    57.5    85.8    114
## dayofweek 113   3.98     2.01   1.00    2.00    4.00    6.00    7.00
## Sales1    114 8,078.75 2,565.97 3,968  6,388   7,416  9,072.2  17,418
## Price1    114   4.88     0.59   3.69    4.59    4.70    5.34    5.75
## feat1     114  18.14    24.41     0       0       0     33.8     98
## disp1     114   5.44     7.16     0       0      2.6     7.5      31
## ------------------------------------------------------------------------
```

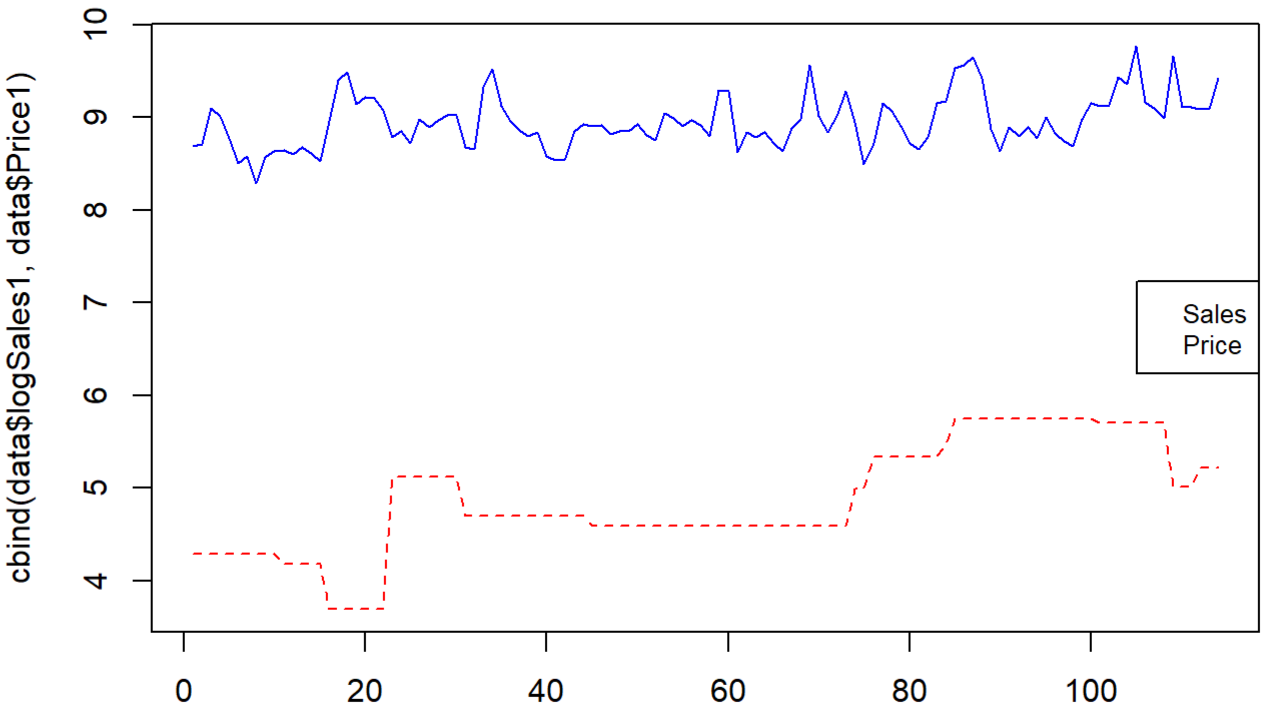
```
data$dayofweek[is.na(data$dayofweek) == TRUE] = 4 #accounting for missing values

#Lets visualize the relationship between Sales and Price through a scatter plot
hist(data$Sales1, col = "cyan")
```



```
data$logSales1=log(data$Sales1)    #log transforming sales
hist(data$logSales1, col = "cyan")
```



```
#plotting the relationship between logSales and Price
{matplot(cbind(data$logSales1,data$Price1),type="l",col=c("blue","red"))
legend("right",c("Sales","Price"),col=c("blue","red"),cex=0.8)}
```



```
#creating dummy variables
#convert the variable to a dummy (factor) variable
data$dayofweek = factor(data$dayofweek)
levels(data$dayofweek) #printing initial levels
```

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

```
levels(data$dayofweek) = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"
, "Sunday")
levels(data$dayofweek)  #printing levels after changes
```

```
## [1] "Monday"    "Tuesday"   "Wednesday" "Thursday"  "Friday"    "Saturday"
## [7] "Sunday"
```

```
#running our lm model
model1 = lm(logSales1~ Price1 + feat1 + Price1*feat1 + dayofweek, data = data)

#print regression results
stargazer(model1,
          title="Regression Results", type="text",
          column.labels=c("Model-1"),
          df=FALSE, digits=3, star.cutoffs = c(0.05,0.01,0.001))
```

```
##
## Regression Results
## =================================================
##                           Dependent variable:
##                      -----------------------------
##                                logSales1
##                                 Model-1
## -------------------------------------------------
## Price1                         0.201***
##                                 (0.055)
##
## feat1                           0.023**
##                                 (0.007)
##
## dayofweekTuesday                 0.055
##                                 (0.086)
##
## dayofweekWednesday               0.034
##                                 (0.087)
##
## dayofweekThursday                0.037
##                                 (0.085)
##
## dayofweekFriday                  0.111
##                                 (0.086)
##
## dayofweekSaturday               -0.012
##                                 (0.086)
##
## dayofweekSunday                 -0.038
##                                 (0.086)
##
## Price1:feat1                    -0.004*
##                                 (0.002)
##
## Constant                       7.838***
##                                 (0.278)
##
## -------------------------------------------------
## Observations                      114
## R2                               0.343
## Adjusted R2                      0.286
## Residual Std. Error              0.243
## F Statistic                    6.036***
## =================================================
## Note:               *p<0.05; **p<0.01; ***p<0.001
```

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?

Dependent Variable: **Sales1** is a dependent variable which depends on values of other independent variables. Key Independent Variable: **Price1** is the key independent variable whose value we will change to measure how it impacts dependent variable (Sales), keeping other control variables constant.

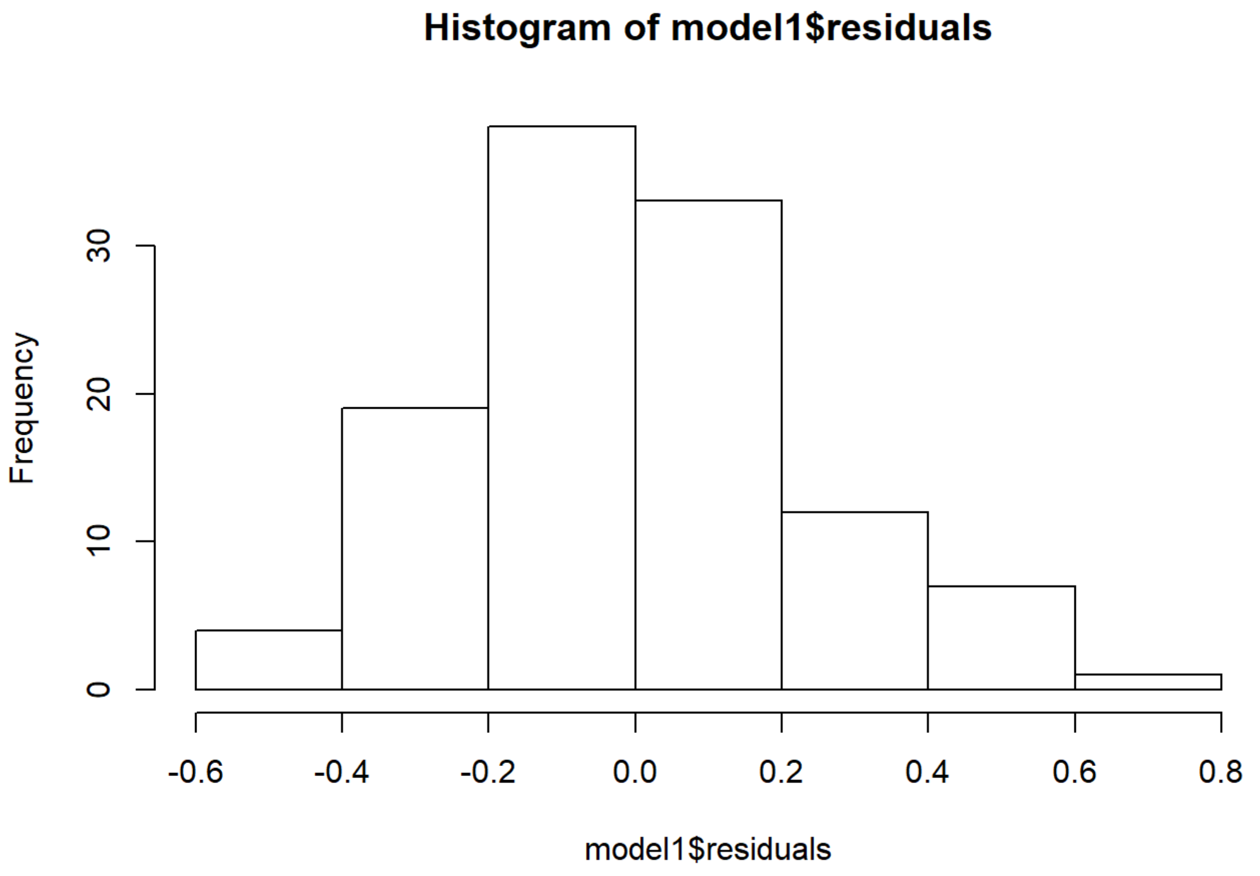
Control Variables: The following variables are control variables-

- **dayofweek**: dayofweek is a dummy variable with multiple levels (called a factor variable). It is mapped as following: Monday-1, Tuesday-2, Wednesday-3 and so on till Sunday-7. We include this factor variable to assess if Sales differ with different days of a week. Theoretical assumption is that coffee Sales on weekend(Saturday and Sunday) would be lower than Sales on weekdays, because schools and offices, which account for most coffee customers, remain closed on weekends.
- **feat1**: feat1 is a continuous variable. We include this variable to assess the impact of featuring the product on Sales. Theoretically we assume that if a product is featured more (featuring increases), Sales should increase.
- **Price1:feat1**: This is an interaction term that assesses the impact of Price on Sales, when product is featured more versus when product is featured less. Theoretically we assume that as featuring increases, the impact of Price on sales will be greater.

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

- Histogram: Let's look at the histogram of residuals. We notice that the plot is almost normally distributed, but it is not symmetrical, it is slightly right-skewed.
- Mean: The mean of residuals show that mean is approximately 0.
- Q-Q Plot: The Q-Q plot shows that most part of our sample quantiles are close to theoretical quantiles. However, at the top right area, sample quantiles start to drift away from theoretical quantiles.
  The residuals looks good, however they are NOT ideal. Ideally we would want residuals that are normally distributed with a mean 0.

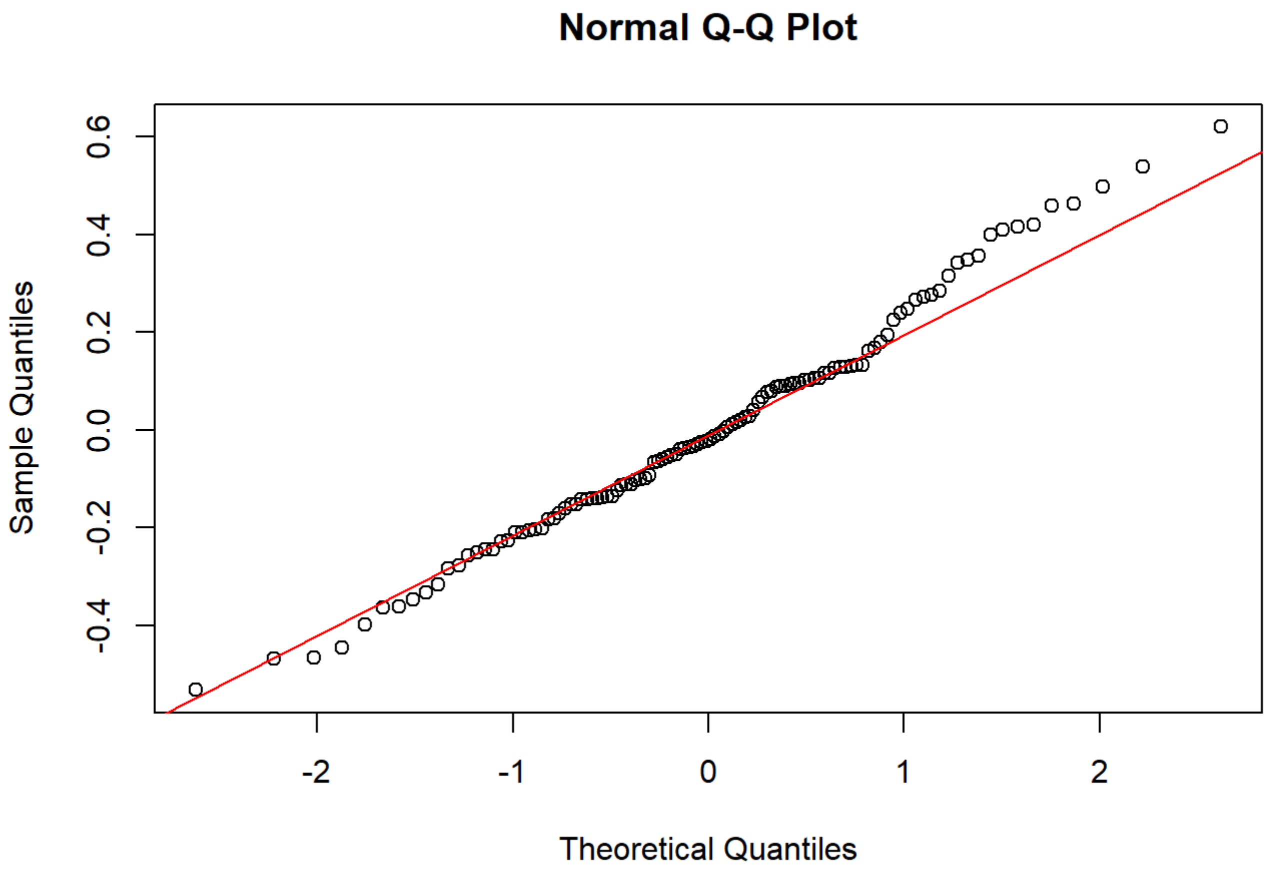
```
hist(model1$residuals) # The histogram of residuals looks normally distributed.
```



```
mean(model1$residuals) # mean of residuals
```

```
## [1] 9.130124e-19
```

```
{qqnorm(model1$residuals) #plotting the qqplot
qqline(model1$residuals,col=2)}
```



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?

The coefficient estimates, standard errors, t-value and p-values for each parameter is given below in a summary table. Interpretations:

- **Price1:feat1**: As featuring of the product increases by one unit, the impact of Price on logSales decreases by a 0.003553 units. In simpler words, as the product is featured more, the impact of Price on logSales decreases. This is in contradiction to our theoretical assumption. The standard error is 0.001515.
- **Price1**: One dollar increase in price is associated with 0.200511 units increase in logSales. If we exponentiate these units we can say that, a one dollar increase in price is associated with approximately 20.05% increase in Sales. The standard error is 0.055076.
- **feat1**: One unit increase in featuring(feat1) is associated with 0.023411 units increase in logSales. In other words, one unit increase in featuring is associated with approximately 2.34% increase in Sales. The standard error is 0.007498.

Notice the p-value of the aforementioned estimates. All the p-values are less than 0.05, thus making these estimates statistically significant. t-value of these estimates are also greater than 2, which shows statistical significance.
On the other hand, notice the p-value and t-values of the estimates given below. All p-values are greater than 0.05 and t-values are less than 2, thus showing that these estimates are statistically non-significant.

- **dayofweekTuesday**: logSales on Tuesday is 0.054631 units higher than logSales on Monday. In other words, Sales on Tuesday is approximately 5.46% higher than Sales on Monday. The standard error is 0.085951.
- **dayofweekWednesday**: logSales on Wednesday is 0.034125 units higher than logSales on Monday. In other words, Sales on Wednesday is approximately 3.41% higher than Sales on Monday. The standard error is 0.086825.
- **dayofweekThursday**: logSales on Thursday is 0.037064 units higher than logSales on Monday. In other words, Sales on Thursday is approximately 3.70% higher than Sales on Monday. The standard error is 0.084771.
- **dayofweekFriday**: logSales on Friday is 0.111311 units higher than logSales on Monday. In other words, Sales on Friday is approximately 11.13% higher than Sales on Monday. The standard error is 0.086318.
- **dayofweekSaturday**: logSales on Saturday is 0.012236 units lower than logSales on Monday. In other words, Sales on Saturday is approximately 1.22% lower than Sales on Monday. This is in accordance with our theoretical assumption that sales would dip on weekends. The standard error is 0.086231.
- **dayofweekSunday**: logSales on Sunday is 0.038149 units lower than logSales on Monday. In other words, Sales on Sunday is approximately 3.81% lower than Sales on Monday. This is in accordance with our theoretical assumption that sales would dip on weekends. The standard error is 0.086292.

```
summary(model1)
```

```
##
## Call:
## lm(formula = logSales1 ~ Price1 + feat1 + Price1 * feat1 + dayofweek,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53179 -0.14947 -0.02088  0.12722  0.61973
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.838486   0.278245  28.171  < 2e-16 ***
## Price1              0.200511   0.055076   3.641 0.000426 ***
## feat1               0.023411   0.007498   3.122 0.002324 **
## dayofweekTuesday    0.054631   0.085951   0.636 0.526432
## dayofweekWednesday  0.034125   0.086825   0.393 0.695100
## dayofweekThursday   0.037064   0.084771   0.437 0.662853
## dayofweekFriday     0.111311   0.086318   1.290 0.200069
## dayofweekSaturday  -0.012236   0.086231  -0.142 0.887439
## dayofweekSunday    -0.038149   0.086292  -0.442 0.659343
## Price1:feat1       -0.003553   0.001515  -2.346 0.020858 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2429 on 104 degrees of freedom
## Multiple R-squared:  0.3431, Adjusted R-squared:  0.2863
## F-statistic: 6.036 on 9 and 104 DF,  p-value: 8.792e-07
```

6. Based on the above estimation results, what’s your answer to the question “how does price influence sales”?

Based on the estimation results, an increase in price is associated with an increase in sales. However, this impact of price on sales decreases with an increase in featuring of the product.

7. Comment on your model fit: R-squared, adjusted R-squared, F-statistics.

F-statistic is 6.036 and its p-value is less than 0.001, which makes F-stat statistically significant. This shows that our model fits the data significantly better than a null-model.
R-squared is 0.3431, which shows that our model fits 34.31% of our data. This value is above 0.10, thus we consider this value as significant. However, this value will go up as we introduce more control variable in our model. Thus we look at a more reliable statistic, that is adjusted R-square.
Adjusted R-squared is 0.2863, which shows that our model fits 28.63% of our data. Since this value is greater than 0.10, we consider this significant.

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.

The chunk of code below shows the results of changing the reference variable from Monday to Wednesday.

```
#Changing the reference from Monday to Wednesday
data$dayofweek = relevel(data$dayofweek, ref = "Wednesday")
#running our lm model
model2 = lm(logSales1~ Price1 + feat1 + Price1*feat1 + dayofweek, data = data)
#print regression results
stargazer(model1,model2,
          title="Regression Results", type="text",
          column.labels=c("Ref-Monday","Ref-Wednesday"),
          df=FALSE, digits=6, star.cutoffs = c(0.05,0.01,0.001))
```

```
##
## Regression Results
## ==================================================
##                          Dependent variable:
##                     ------------------------------
##                               logSales1
##                       Ref-Monday    Ref-Wednesday
##                           (1)            (2)
## --------------------------------------------------
## Price1                0.200511***    0.200511***
##                       (0.055076)      (0.055076)
##
## feat1                 0.023411**      0.023411**
##                       (0.007498)      (0.007498)
##
## dayofweekMonday                       -0.034125
##                                       (0.086825)
##
## dayofweekTuesday       0.054631        0.020506
##                       (0.085951)      (0.084846)
##
## dayofweekWednesday     0.034125
##                       (0.086825)
##
## dayofweekThursday      0.037064        0.002939
##                       (0.084771)      (0.086069)
##
## dayofweekFriday        0.111311        0.077186
##                       (0.086318)      (0.086469)
##
## dayofweekSaturday      -0.012236      -0.046361
##                       (0.086231)      (0.086173)
##
## dayofweekSunday        -0.038149      -0.072274
##                       (0.086292)      (0.087337)
##
## Price1:feat1          -0.003553*      -0.003553*
##                       (0.001515)      (0.001515)
##
## Constant              7.838486***    7.872611***
##                       (0.278245)      (0.278895)
##
## --------------------------------------------------
## Observations              114            114
## R2                     0.343120        0.343120
## Adjusted R2            0.286275        0.286275
## Residual Std. Error    0.242916        0.242916
## F Statistic           6.036028***    6.036028***
## ==================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = logSales1 ~ Price1 + feat1 + Price1 * feat1 + dayofweek,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53179 -0.14947 -0.02088  0.12722  0.61973
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.872611   0.278895  28.228  < 2e-16 ***
## Price1             0.200511   0.055076   3.641 0.000426 ***
## feat1              0.023411   0.007498   3.122 0.002324 **
## dayofweekMonday   -0.034125   0.086825  -0.393 0.695100
## dayofweekTuesday   0.020506   0.084846   0.242 0.809504
## dayofweekThursday  0.002939   0.086069   0.034 0.972823
## dayofweekFriday    0.077186   0.086469   0.893 0.374104
## dayofweekSaturday -0.046361   0.086173  -0.538 0.591731
## dayofweekSunday   -0.072274   0.087337  -0.828 0.409834
## Price1:feat1      -0.003553   0.001515  -2.346 0.020858 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2429 on 104 degrees of freedom
## Multiple R-squared:  0.3431, Adjusted R-squared:  0.2863
## F-statistic: 6.036 on 9 and 104 DF,  p-value: 8.792e-07
```

These are the following changes in our estimates:

Standard Errors: We notice that the standard errors of the coeffients do not vary much.

- **dayofweekWednesday**: This parameter is NOT present in our model-2, because Wednesday is the reference. This parameter, however, appeared in model-1 and had the following interpretation: logSales on Wednesday are 0.034125 units more than logSales on Monday.
- **dayofweekMonday**: This parameter was not present in our model-1, because Monday was the reference. This parameter appears in model-2 and has the following interpretation: logSales on Monday are 0.034125 units lesser than logSales on Wednesday.
- **dayofweekTuesday**: logSales on Tuesday is 0.020506 units higher than logSales on Wednesday. In other words, Sales on Tuesday is approximately 2.05% higher than Sales on Wednesday.
- **dayofweekThursday**: logSales on Thursday is 0.002939 units higher than logSales on Wednesday. In other words, Sales on Thursday is approximately 0.29% higher than Sales on Wednesday.
- **dayofweekFriday**: logSales on Friday is 0.077186 units higher than logSales on Wednesday. In other words, Sales on Friday is approximately 7.72% higher than Sales on Wednesday.
- **dayofweekSaturday**: logSales on Saturday is 0.046361 units lower than logSales on Wednesday. In other words, Sales on Saturday is approximately 4.64% lower than Sales on Wednesday.
- **dayofweekSunday**: logSales on Sunday is 0.072274 units lower than logSales on Wednesday. In other words, Sales on Sunday is approximately 7.23% lower than Sales on Wednesday.

Notice that p-value of all these estimates are still greater than 0.05, thus all the above mentioned estimates are not significant.
Other estimates like, Price1*feat1, Price1, feat1 are not affected by changing reference of dayofweek.