# AB Testing and Regression Analysis

Matteo Montrucchio

2024-04-08

## AB Testing

In the following file, I will evaluate the effectiveness of featuring actions. I will analyze a dataset characterized by weekly sales and prices of orange juices across different stores for three brands. To evaluate the effectiveness of featuring actions, I compare the average sales by feat category.

```
library(readr)
oj.data <- read_csv("/Users/matteomontrucchio/Desktop/oj_data.csv", show_col_types = FALSE)
```

```
# show the first rows of the dataset
head(oj.data)
```

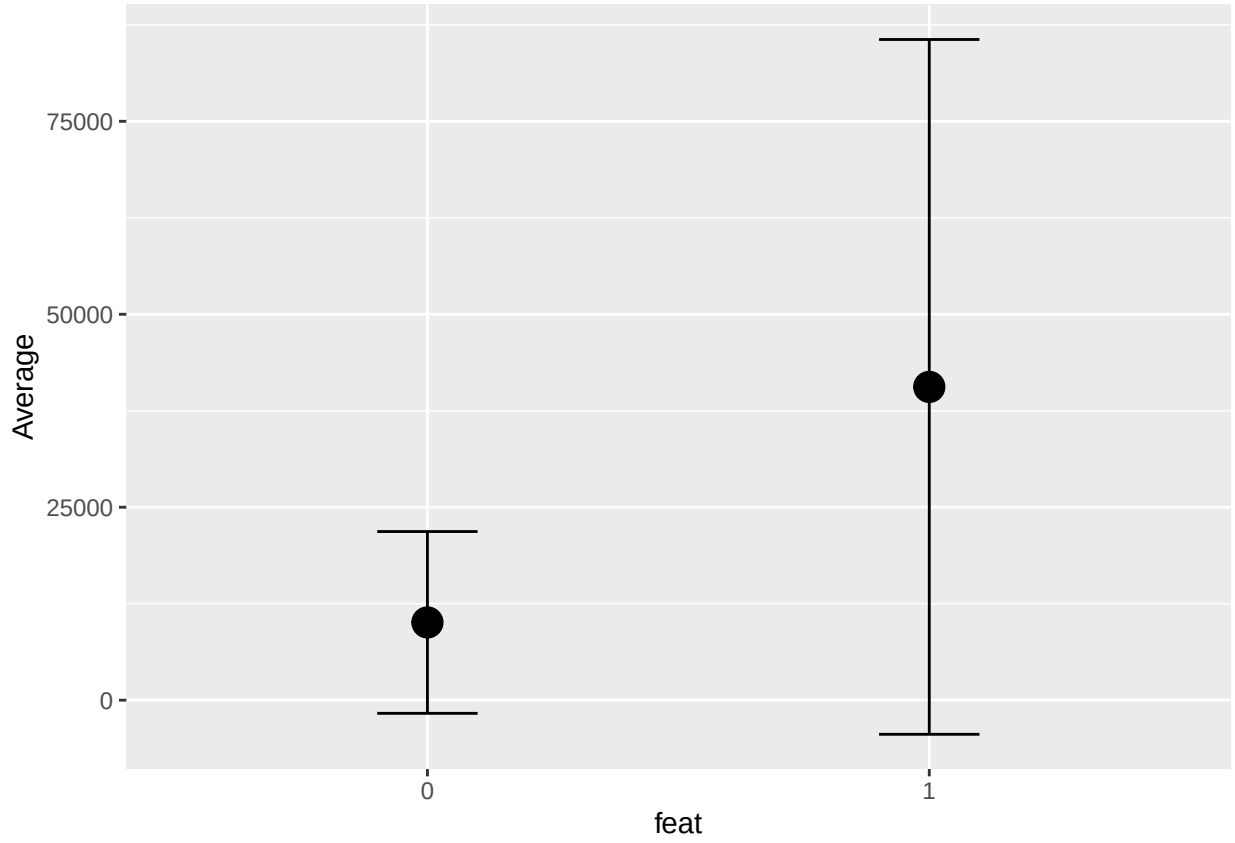
```
## # A tibble: 6 x 4
##   sales price brand      feat
##   <dbl> <dbl> <chr>     <dbl>
## 1 8256.  3.87 tropicana     0
## 2 6144.  3.87 tropicana     0
## 3 3840.  3.87 tropicana     0
## 4 8000.  3.87 tropicana     0
## 5 8896.  3.87 tropicana     0
## 6 7168.  3.87 tropicana     0
```

```
# summarize the information contained in the dataset
summary(oj.data)
```

```
##      sales            price          brand                feat
##  Min.   :    64   Min.   :0.520   Length:28947       Min.   :0.0000
##  1st Qu.:  4864   1st Qu.:1.790   Class :character   1st Qu.:0.0000
##  Median :  8384   Median :2.170   Mode  :character   Median :0.0000
##  Mean   : 17312   Mean   :2.282                      Mean   :0.2373
##  3rd Qu.: 17408   3rd Qu.:2.730                      3rd Qu.:0.0000
##  Max.   :716416   Max.   :3.870                      Max.   :1.0000
```

```
# generate the plot comparing average sales by feat category and their SD
oj.data %>%
  mutate(feat= as.factor(feat)) %>%     # from the summary(), feat is stored as numeric variable
  group_by(feat) %>%
  summarise(Average = mean(sales),
            SD = sd(sales)) %>%
  ggplot(aes(feat,Average)) +
```

```
  geom_point(size=5) +
  geom_errorbar(aes(ymin=Average-SD, ymax=Average+SD), width=.2)
```




```
t.table <- oj.data %>%
  mutate(feat= as.factor(feat)) %>%
  group_by(feat) %>%
  summarise(Average = mean(sales),
            Var = var(sales),
            N = n())
t.table
```

```
## # A tibble: 2 x 4
##   feat  Average         Var     N
##   <fct>   <dbl>       <dbl> <int>
## 1 0      10071.  138704271. 22079
## 2 1      40590. 2026106369.  6868
```

```
t.test(oj.data[oj.data$feat==1, 1], oj.data[oj.data$feat==0, 1], var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj.data[oj.data$feat == 1, 1] and oj.data[oj.data$feat == 0, 1]
## t = 55.601, df = 7161.6, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  29443.30 31595.31
## sample estimates:
## mean of x mean of y
##  40590.47  10071.17
```

The data provided shows how average sales are higher when the product is featured.

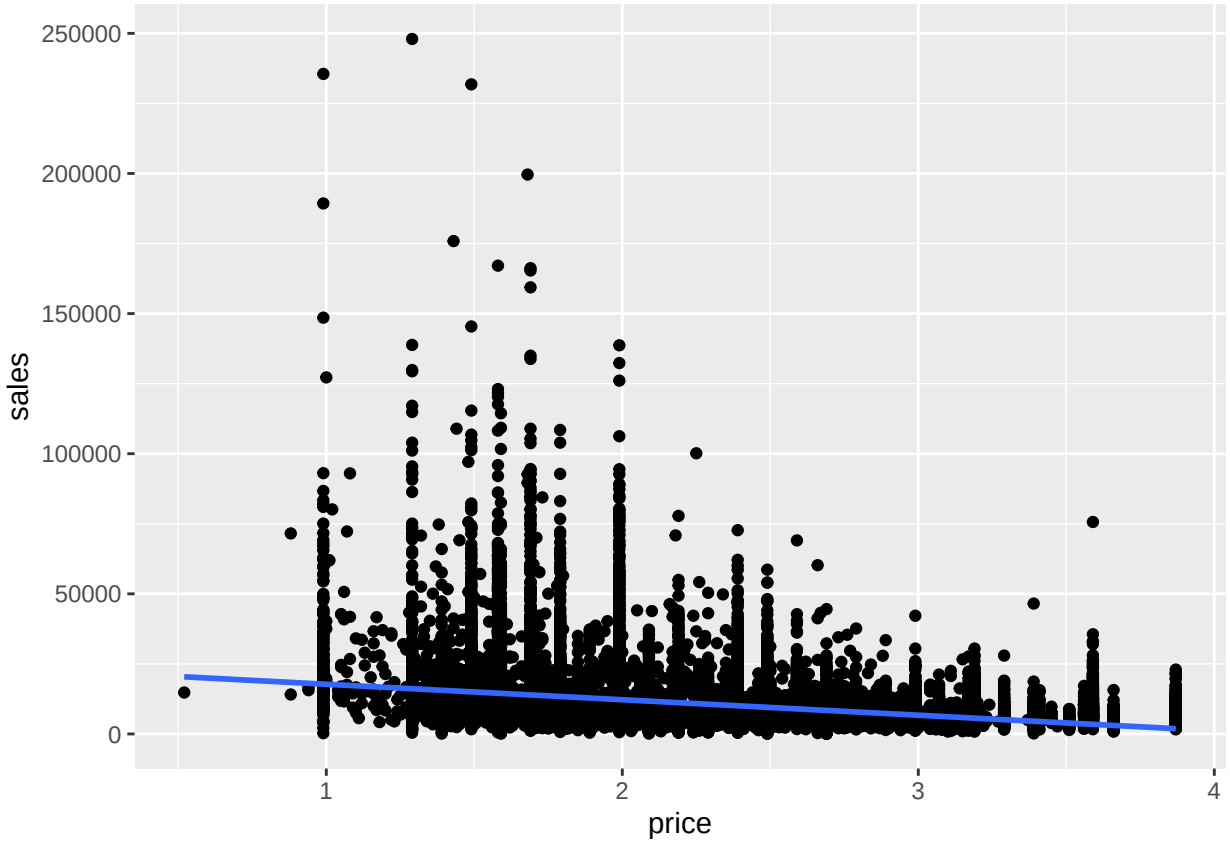
On average, 40590 quantities are sold when product is featured against an average of 10071 quantities sold when not featured, resulting in an average difference of 30519 quantities (with a 95% probability, a store will sell between 29443.16 and 31594.84 more quantities).

## Regression

The goal is now to evaluate the effect that a price change has on the sales.

I will at first perform a simple linear regression analysis not considering the featuring.

```
# plot sales VS price for feat equal to 0
ggplot(oj.data%>% filter(feat==0),aes(price,sales))+geom_point()+
  geom_smooth(method='lm',formula=y~x)
```




```
# summary of the model
summary(lm(sales~price, oj.data %>% filter(feat==0) ))
```

```
##
## Call:
## lm(formula = sales ~ price, data = oj.data %>% filter(feat ==
##     0))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17535  -5707  -2100   2079 231866
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23258.2      284.1   81.86   <2e-16 ***
## price        -5522.5      114.7  -48.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11200 on 22077 degrees of freedom
## Multiple R-squared:  0.09501,    Adjusted R-squared:  0.09497
## F-statistic:  2318 on 1 and 22077 DF,  p-value: < 2.2e-16
```

This highlights how an increase in price of one unit has a negative effect on the average weekly sales that will reduce of 5522.5 units (more precisely, with a 95% level of confidence, between 5747.312 and 5297.688)

I will now run a second model in which I include the featuring effect in order to see if in the first model there was omitting variable bias.

```
no.ovb <- lm(sales~price+feat, oj.data)
summary(no.ovb)
```

```
##
## Call:
## lm(formula = sales ~ price + feat, data = oj.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42604  -9192  -2545   4702 665908
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34902.6      551.4   63.29   <2e-16 ***
## price       -10399.0      221.4  -46.98   <2e-16 ***
## feat         25900.2      337.2   76.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23340 on 28944 degrees of freedom
## Multiple R-squared:  0.2783, Adjusted R-squared:  0.2782
## F-statistic:  5580 on 2 and 28944 DF,  p-value: < 2.2e-16
```

Considering the featuring effect, the estimated decrease in sales is of 10399 (95% C.I.: between 9965.256 and 10833.144).