Return On Ad Spending [ROAS]

Raj Gopalan

Project done as part of UCLA MBA, Marketing Analytics course

An advertiser conducted a randomized experiment (A/B test) using cities as the unit of measure, half the cities got the ads and half didn't. Now the advertiser wants to measure the return on ad spend (ROAS). ROAS is calculated as:

ROAS = Total Incremental Sales / Total Advertising Spend

Example - An ROAS of 3 indicates that the advertiser made $2 in sales for every $1 in ad spend.

The A/B test lasted 30 days. Daily sales by city for 60 days prior to the test and for the 30 days during the test is provided. The ad spend for each day is provided as well (there was no advertising before the test started). 10,000 cities were randomly chosen and randomly assigned to a test group (5000) and control group (5000).

The question that needs to be answered: Does advertising impact sales? If so can you quantify this effect?

```
# Use install_bitbucket('perossichi/DataAnalytics') to download the
# DataAnalytics library
library(DataAnalytics)
library(ggplot2)
library(reshape)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following object is masked from 'package:reshape':
##
##     melt
```

Disclaimer:

I can't provide the data files as per my professor. But, the intent is to analyze the files and bring out the insights from the data, which can help answer the question: what is the ROAS? So, let's go!

```
adspend <- read.csv("data/adSpend.csv", header = TRUE)
assignments <- read.csv("data/gmaAssignment.csv", header = TRUE)
sales <- read.csv("data/salesData.csv", header = TRUE)
```

Merge the three sources of data using 'gma' as the primary key.

```
# merge_recurse() can intelligently guess the primary key, so there is no
# need to explicitly mention it in the function call.  This is because 'gma'
# is the only column that is common to all three data sets.
df1 <- merge_recurse(list(adspend, assignments, sales))
sum(is.na(df1))
```

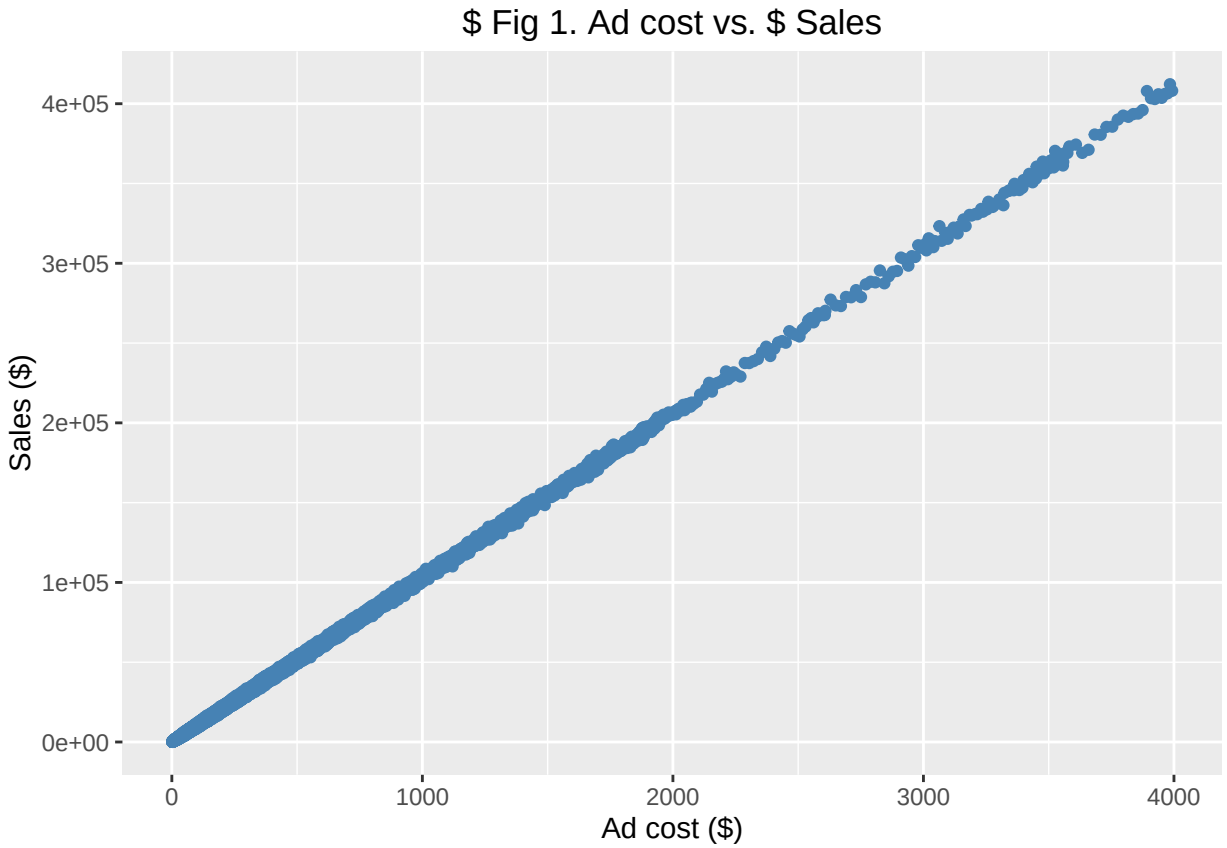
```
## [1] 0
```

Great, **no** NAs!

Time to visualize the data.

I will start by doing a quick **scatter plot** of *Ad cost vs. Sales* to see if there is a correlation between the two. I will be using data which contains ad cost information, ignoring the data when the ads were not shown. The expectation is that I should see a positive correlation.

```
fig1 <- ggplot(df1[df1$ad.clicks > 0, ], aes(cost.clicks, sales))
fig1 + geom_point(color = "steel blue") + ggtitle("$ Fig 1. Ad cost vs. $ Sales") +
    labs(x = "Ad cost ($)", y = "Sales ($)")
```



The **positive correlation** between sales and ads is clearly visible in *Fig 1.* above, i.e. as the cost of ads increases, the sales seems to be increasing as well. Now that I'm seeing a correlation between sales and advertisement spend, the approach I would like to take, is to first break down the problem that needs to be solved.

I know the A/B test lasted 30 days and that the daily sales by city for 60 days prior and 30 days during the test is provided. This means that there are three months worth of data. A quick summary of the *date* column should validate this.

```
summary(df1$date)
```

```
## 2014-01-01 2014-01-02 2014-01-03 2014-01-04 2014-01-05 2014-01-06
##      10000      10000      10000      10000      10000      10000
## 2014-01-07 2014-01-08 2014-01-09 2014-01-10 2014-01-11 2014-01-12
##      10000      10000      10000      10000      10000      10000
## 2014-01-13 2014-01-14 2014-01-15 2014-01-16 2014-01-17 2014-01-18
##      10000      10000      10000      10000      10000      10000
## 2014-01-19 2014-01-20 2014-01-21 2014-01-22 2014-01-23 2014-01-24
##      10000      10000      10000      10000      10000      10000
## 2014-01-25 2014-01-26 2014-01-27 2014-01-28 2014-01-29 2014-01-30
##      10000      10000      10000      10000      10000      10000
## 2014-01-31 2014-02-01 2014-02-02 2014-02-03 2014-02-04 2014-02-05
##      10000      10000      10000      10000      10000      10000
## 2014-02-06 2014-02-07 2014-02-08 2014-02-09 2014-02-10 2014-02-11
##      10000      10000      10000      10000      10000      10000
## 2014-02-12 2014-02-13 2014-02-14 2014-02-15 2014-02-16 2014-02-17
##      10000      10000      10000      10000      10000      10000
## 2014-02-18 2014-02-19 2014-02-20 2014-02-21 2014-02-22 2014-02-23
```

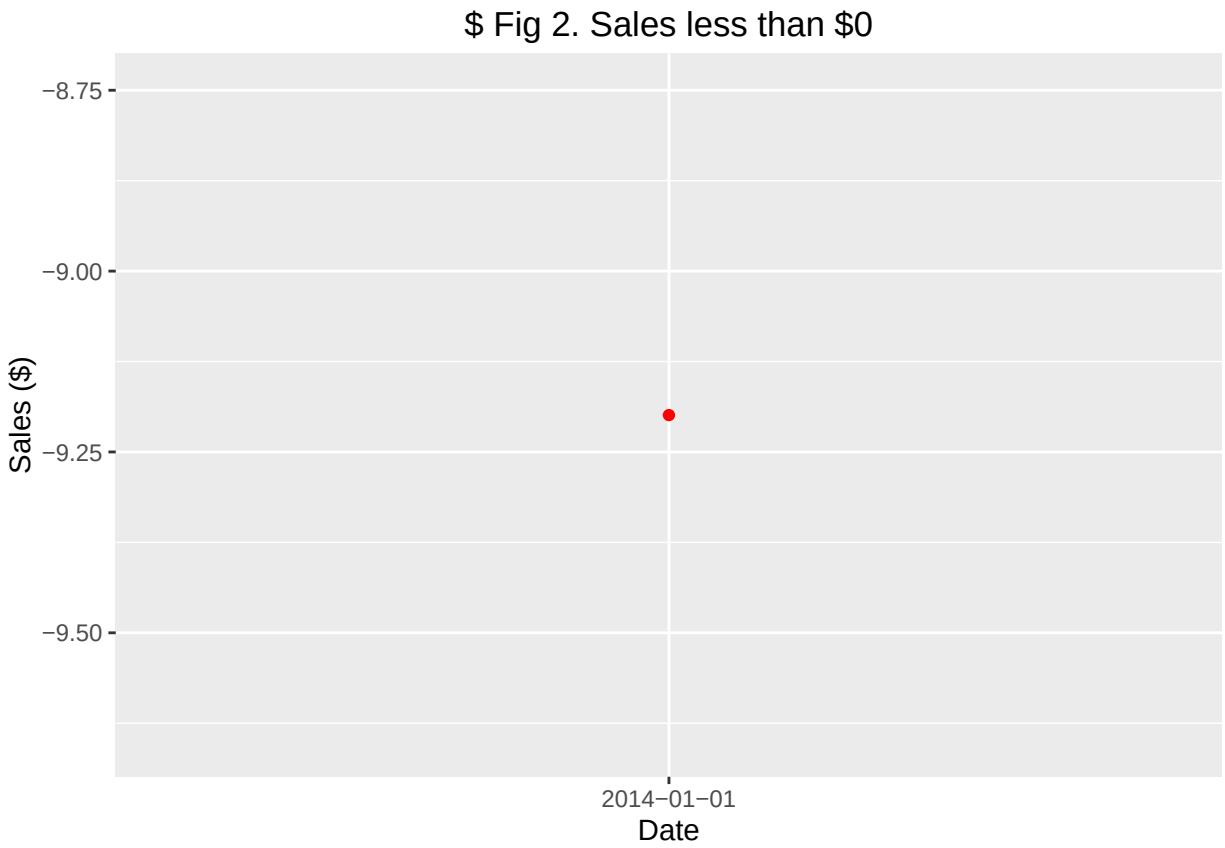
```
##       10000       10000       10000       10000       10000       10000
## 2014-02-24 2014-02-25 2014-02-26 2014-02-27 2014-02-28 2014-03-01
##       10000       10000       10000       10000       10000       10000
## 2014-03-02 2014-03-03 2014-03-04 2014-03-05 2014-03-06 2014-03-07
##       10000       10000       10000       10000       10000       10000
## 2014-03-08 2014-03-09 2014-03-10 2014-03-11 2014-03-12 2014-03-13
##       10000       10000       10000       10000       10000       10000
## 2014-03-14 2014-03-15 2014-03-16 2014-03-17 2014-03-18 2014-03-19
##       10000       10000       10000       10000       10000       10000
## 2014-03-20 2014-03-21 2014-03-22 2014-03-23 2014-03-24 2014-03-25
##       10000       10000       10000       10000       10000       10000
## 2014-03-26 2014-03-27 2014-03-28 2014-03-29 2014-03-30 2014-03-31
##       10000       10000       10000       10000       10000       10000
```

As expected, the summary identifies the start and end dates. There is information on January, February and March. I am converting the dates into months, with the intent of identifying seasonal effects, if any. I would like to ensure that the months are added as categorical variables so that running a regression becomes straighforward.

```
df1$January <- ifelse((months(as.Date(df1$date, "%Y-%m-%d")) == "January"),
    1, 0)
df1$February <- ifelse((months(as.Date(df1$date, "%Y-%m-%d")) == "February"),
    1, 0)
df1$March <- ifelse((months(as.Date(df1$date, "%Y-%m-%d")) == "March"), 1, 0)
```

I first need to ensure that all sales values are greater than 0. A quick plot of sales values less than 0 would help check this:

```
fig2 <- ggplot(df1[df1$sales < 0, ], aes(date, sales))
fig2 + geom_point(color = "red") + ggtitle("$ Fig 2. Sales less than $0") +
    labs(x = "Date", y = "Sales ($)")
```



Observations: It looks like there is one sales value less than $0. A logical interpretation could be that it was a return that happened. So, I will simply let it be there.

The first intuition I would like to check, is if there are any time series trends affecting the sales. Running a linear regression of sales on February and March, could help identify monthly trends, if any. I am splitting the data based on the test and control groups, to check if the effects are the same across the groups. The expectation is that, since the test group vs control group assignment was random, the effects should be the same in both data sets.

```
df1.cg <- df1[df1$gma.group == 0, ]
df1.tg <- df1[df1$gma.group == 1, ]
df1.cg$gma.group = NULL
df1.tg$gma.group = NULL
lm.1.Results <- lm(data = df1.cg, sales ~ February + March)
lmSumm(lm.1.Results)
```

```
## Multiple Regression Analysis:
##     3 regressors(including intercept) and 450000 observations
##
## lm(formula = sales ~ February + March, data = df1.cg)
##
## Coefficients:
##             Estimate Std Error t value p value
## (Intercept)    10710     39.59  270.43       0
## February        3279     57.46   57.06       0
## March           2679     55.98   47.85       0
## ---
## Standard Error of the Regression:  15580
```

```
## Multiple R-squared:  0.008  Adjusted R-squared:  0.008
## Overall F stat: 1891.4 on 2 and 449997 DF, pvalue= 0
```

```
lm.2.Results <- lm(data = df1.tg, sales ~ February + March)
lmSumm(lm.2.Results)
```

```
## Multiple Regression Analysis:
##     3 regressors(including intercept) and 450000 observations
##
## lm(formula = sales ~ February + March, data = df1.tg)
##
## Coefficients:
##             Estimate Std Error t value p value
## (Intercept)    11330     50.18  225.84       0
## February        3471     72.84   47.65       0
## March           3246     70.96   45.74       0
## ---
## Standard Error of the Regression:  19760
## Multiple R-squared:  0.006  Adjusted R-squared:  0.006
## Overall F stat: 1468.22 on 2 and 449997 DF, pvalue= 0
```

Observations:

1. Both February and March are statistically significant at the 1% level.
2. In both groups, average January sales (intercept value) were lower than in February or March.
3. In the control group, average March sales were $3279 - $2679 = $600 less than average February sales. Whereas, in the test group, average March sales were $3471 - $3246 = $225 less than average February sales. So, the time series effect seems to be consistent across the randomly chosen groups.
4. In March, when the A/B testing was in progress, **the average sales in the test group were *higher* on than those of the control group by $3246 - $2679 = $567**.

Next hypothesis that I would like to test is, the day of the week made a difference in sales numbers. For this purpose, I am creating a set of categorical variables for the weekday corresponding to the date of each observation. I will then run a regression of the test group sales in March against days of the week, to see if there is a statistical significance.

```
df1.tg$Monday <- ifelse((weekdays(as.Date(df1.tg$date, "%Y-%m-%d")) == "Monday"),
    1, 0)
df1.tg$Tuesday <- ifelse((weekdays(as.Date(df1.tg$date, "%Y-%m-%d")) == "Tuesday"),
    1, 0)
df1.tg$Wednesday <- ifelse((weekdays(as.Date(df1.tg$date, "%Y-%m-%d")) == "Wednesday"),
    1, 0)
df1.tg$Thursday <- ifelse((weekdays(as.Date(df1.tg$date, "%Y-%m-%d")) == "Thursday"),
    1, 0)
df1.tg$Friday <- ifelse((weekdays(as.Date(df1.tg$date, "%Y-%m-%d")) == "Friday"),
    1, 0)
df1.tg$Saturday <- ifelse((weekdays(as.Date(df1.tg$date, "%Y-%m-%d")) == "Saturday"),
    1, 0)
df1.tg$Sunday <- ifelse((weekdays(as.Date(df1.tg$date, "%Y-%m-%d")) == "Sunday"),
    1, 0)

df1.tg.March <- df1.tg[df1.tg$March == 1, ]
```

```
lm.4.Results <- lm(data = df1.tg.March, sales ~ Monday + Tuesday + Wednesday +
    Thursday + Friday + Saturday)
lmSumm(lm.4.Results)
```

```
## Multiple Regression Analysis:
##     7 regressors(including intercept) and 155000 observations
##
## lm(formula = sales ~ Monday + Tuesday + Wednesday + Thursday +
##     Friday + Saturday, data = df1.tg.March)
##
## Coefficients:
##             Estimate Std Error t value p value
## (Intercept) 14570.00     133.4  109.24   0.000
## Monday        -79.91     188.6   -0.42   0.672
## Tuesday       168.30     200.0    0.84   0.400
## Wednesday      89.20     200.0    0.45   0.656
## Thursday       10.18     200.0    0.05   0.959
## Friday        -78.43     200.0   -0.39   0.695
## Saturday      -14.24     188.6   -0.08   0.940
## ---
## Standard Error of the Regression:  21090
## Multiple R-squared:  0  Adjusted R-squared:  0
## Overall F stat: 0.38 on 6 and 154993 DF, pvalue= 0.895
```

Observations:

Unfortunately, **none** of the days of the week are statistically significant. So, I will ignore days of the week in my model.

I had noted earlier that when the A/B testing was in progress in March, the sales in test group were higher on average than those of control group. Now, I would like to run a linear regresion of the average sales in the test period (grouped by gma) against the cost.clicks and the sales in the pre-test period.
The hypothesis is that:
1. If ad spending (cost.clicks) improved sales, I should see a statistical significance of the ad spending.
2. By using sales in the pre-test period as an independent variable, it would take away that component of sales, which is related to organic search, leaving behing the sales numbers related to the paid search, i.e. the incremental sales lift. The end result is that the regression coefficient of the ad spending variable (cost.clicks) would then directly provide us with the ROAS.

Now, its time to first setup the data. I'm splitting the data set into pre-test and test periods.

```
df1.pre <- data.table(df1[as.Date(df1$date, "%Y-%m-%d") < "2014-03-02", ])
df1.test <- data.table(df1[as.Date(df1$date, "%Y-%m-%d") >= "2014-03-02", ])
```

Then I'm calculating the aggregrate of 'sales' (response metric) and aggregate of 'cost.clicks' (ad spend) for every gma. This will generate two data sets, each with 10000 observations.

```
agg.df1.pre <- aggregate(cbind(sales, cost.clicks) ~ gma, data = df1.pre, FUN = mean)
agg.df1.test <- aggregate(cbind(sales, cost.clicks) ~ gma, data = df1.test,
    FUN = mean)
```

I don't need the cost.clicks column in the pre-test phase, since there was no advertising before the stest started. So, I will drop cost.clicks from the pre-test phase and rename the relevant sales columns as 'sales.pre' and 'sales.test', before merging the above two data sets using 'gma' as the primary key.

```r
agg.df1.pre$cost.clicks <- NULL
colnames(agg.df1.pre)[colnames(agg.df1.pre) == "sales"] <- "sales.pre"
colnames(agg.df1.test)[colnames(agg.df1.test) == "sales"] <- "sales.test"
agg.df1.combined <- merge_recurse(list(agg.df1.pre, agg.df1.test))
```

The understanding here is that the aggregate of cost.clicks provides the aggregate of the *actual ad spend in the test phase minus the ad spend in the pre-test phase*. Since there was no advertising in the pre-test phase, the regression coefficient of the aggregate of ad spend would directly give us the incremental lift in ad spend, i.e. ROAS. Now, its time to run the regression:

```r
lm.4.Results <- lm(data = agg.df1.combined, sales.test ~ sales.pre + cost.clicks)
lmSumm(lm.4.Results)
```

```
## Multiple Regression Analysis:
##     3 regressors(including intercept) and 10000 observations
##
## lm(formula = sales.test ~ sales.pre + cost.clicks, data = agg.df1.combined)
##
## Coefficients:
##             Estimate Std Error  t value p value
## (Intercept)   -2.561 1.464e+00    -1.75    0.08
## sales.pre      1.086 9.784e-05 11104.53    0.00
## cost.clicks    2.986 1.035e-02   288.61    0.00
## ---
## Standard Error of the Regression:  116.8
## Multiple R-squared:  1  Adjusted R-squared:  1
## Overall F stat: 129554159 on 2 and 9997 DF, pvalue= 0
```

Observations:

1. As expected cost.clicks and pre-test sales are both statistically significant (p-value < 0.01) at the 1% level.
2. \$1 increase in ad spend lifts sales by \$2.96. So, the incremental lift in sales for \$1 increase in ad spend = \$2.96, all else remaining the same. In industry terms, the ROAS is approximately 299. This means that the advertiser got back **almost 2x the ad spend**. The industry benchmark for ROAS is close to 260, so this adveriser has a comparitively good ROAS.

Now let's look at the confidence interval of the ROAS:

```r
confint(lm.4.Results, level = 0.99)
```

```
##                 0.5 %   99.5 %
## (Intercept) -6.333539 1.210712
## sales.pre    1.086181 1.086685
## cost.clicks  2.959108 3.012414
```

So, 99% confidence interval of the ROAS varies between 296 and 301, a width of 5.

This analysis involved ad spend of \$0 in the pre-test phase. If there were non-zero ad spend in the pre-test phase, then the treatment of the data and the corresponding regression would be slightly different, but doable.