RocketFuel Case Study - Heterogeneous Treatment Effects

Shih-Yuan Wang

```
# Load libraries
library(knitr)
library(psych)
library(dplyr)
library(skimr)
library(jtools)
library(rcompanion)
library(ggplot2)
library(car)
```

```
rm(list = ls()) # clear the workspace

x <- paste("C:/Users/User/Desktop/BUS 740 - Experiments and Causal Methods for Business Insights/",
           "Assignment/Homework 2_RocketFuel Case Study", sep="")
setwd(x)
```

1./2. Read in the file.

```
rocketfuel <- read.csv('rocketfuel_deciles.csv', header = TRUE)  # load the data file
dim(rocketfuel)
```

```
## [1] 588101      7
```

There are 588101 observations and 7 variables in the dataset.

3. Summarize the variables in the data.

```
# give a descriptive summary for all variables in rocketfuel
kable(summary(rocketfuel))
```

	user_id	test	converted	tot_impr	mode_impr_day	mode_impr_hour	tot_impr_decile
	Min. : 900000	Min. :0.00	Min. :0.00000	Min. : 1.00	Min. :1.000	Min. : 0.00	Min. : 1.00
	1st Qu.:1143190	1st Qu.:1.00	1st Qu.:0.00000	1st Qu.: 4.00	1st Qu.:2.000	1st Qu.:11.00	1st Qu.: 3.00
	Median :1313725	Median :1.00	Median :0.00000	Median : 13.00	Median :4.000	Median :14.00	Median : 5.00
	Mean :1310692	Mean :0.96	Mean :0.02524	Mean : 24.82	Mean :4.026	Mean :14.47	Mean : 5.38

user_id	test	converted	tot_impr	mode_impr_day	mode_impr_hour	tot_impr_decile
3rd Qu.:1484088 Max. :1654483	3rd Qu.:1.00 Max. :1.00	3rd Qu.:0.00000 Max. :1.00000	3rd Qu.: 27.00 Max. :2065.00	3rd Qu.:6.000 Max. :7.000	3rd Qu.:18.00 Max. :23.00	3rd Qu.: 8.00 Max. :10.00

```r
kable(describe(rocketfuel)[-1,1:9])
```

	vars	n	mean	sd	median	trimmed	mad	min	max
test	2	588101	0.9600001	0.1959592	1	1.000000	0.0000	0	1
converted	3	588101	0.0252389	0.1568499	0	0.000000	0.0000	0	1
tot_impr	4	588101	24.8208760	43.7151805	13	16.266608	14.8260	1	2065
mode_impr_day	5	588101	4.0255330	2.0040193	4	4.031916	2.9652	1	7
mode_impr_hour	6	588101	14.4690606	4.8346339	14	14.587541	4.4478	0	23
tot_impr_decile	7	588101	5.3803462	2.9535724	5	5.353332	2.9652	1	10

From the summary statistics, we can see that the size of the test (treatment) group in which the user was exposed to the real ad is 96%, while the size of the control group (was shown a PSA) is 4%. Around 2.5% of the users did buy the handbag during the campaign. The average of the total number of ad impressions the user encountered was around 25, but the median is just 13. The average of the day of the week on which the user encountered the most number of impressions is "Thursday", and the average of the hour of the day (0-23) in which the user encountered the most number of impressions is around 14:00.

4. Create a table to show the numbers and shares of individuals who were in the treatment vs. control group.

```r
attach(rocketfuel)
tb_treatment_full <- matrix(NA, nrow = 2, ncol = 2) # create a empty output matrix with 2 rows
# (for Frequency, i.e., count, and Proportion) and the 2 groups
tb_treatment_full[1,] <- format(table(test), digits = 0) # counts in treatment.
tb_treatment_full[2,] <- format(prop.table(table(test)), digits = 3) # proportion in treatments
rownames(tb_treatment_full) <- c("Frequency", "Proportion" ) # name the rows
colnames(tb_treatment_full) <- c("Control (PSA)", "Real Ad")  # name the columns
kable(tb_treatment_full, align = "rr") # output the table in a readable format
```

	Control (PSA)	Real Ad
Frequency	23524	564577
Proportion	0.04	0.96

```r
detach(rocketfuel)
```

The number of users who was shown a PSA in the control group is 23524 (4%), while the number of users who was exposed to the real ad in the treatment group is 564577 (96%).

5. Check for balance in the pre-experiment variables across treatment and control.

For a successful experiment, the pre-treatment variables should be balanced across treatment and control groups. The variables of interest here are the total number of ad impressions the user encountered (tot_impr), the day of the week on which the user encountered the most number of impressions (mode_impr_day), and the hour of the day in which the user encountered the most number of impressions(mode_impr_hour). We are hoping to have these look similar between the groups, since they can't be affected by the treatment.

a. Create a table comparing the means between treatment and controls.

```
attach(rocketfuel)
# create a data frame with the treatment variables and the pre-treatment variables
preexp <- rocketfuel %>%
  select(tot_impr, mode_impr_day, mode_impr_hour) # pre-experiment variables

# Summarize the means of those variables by treatment
tb_preexp <- matrix(NA, nrow = 3, ncol = 2) # define the empty output matrix
colnames(tb_preexp) <- c("Mean Control", "Mean Real Ad") # name the columns
rownames(tb_preexp) <- colnames(preexp) # name the rows
m <- as.matrix(round(aggregate(.~test, preexp, mean), 2))
# summarize all variables by treatment and store the outputs as a matrix

tb_preexp[,1:2] <-t(m)[2:4,] # transpose the matrix and delete the treatment(test) row
kable(tb_preexp) # output the table in a readable format
```

	Mean Control	Mean Real Ad
tot_impr	24.76	24.82
mode_impr_day	3.95	4.03
mode_impr_hour	14.30	14.48

```
detach(rocketfuel)
```

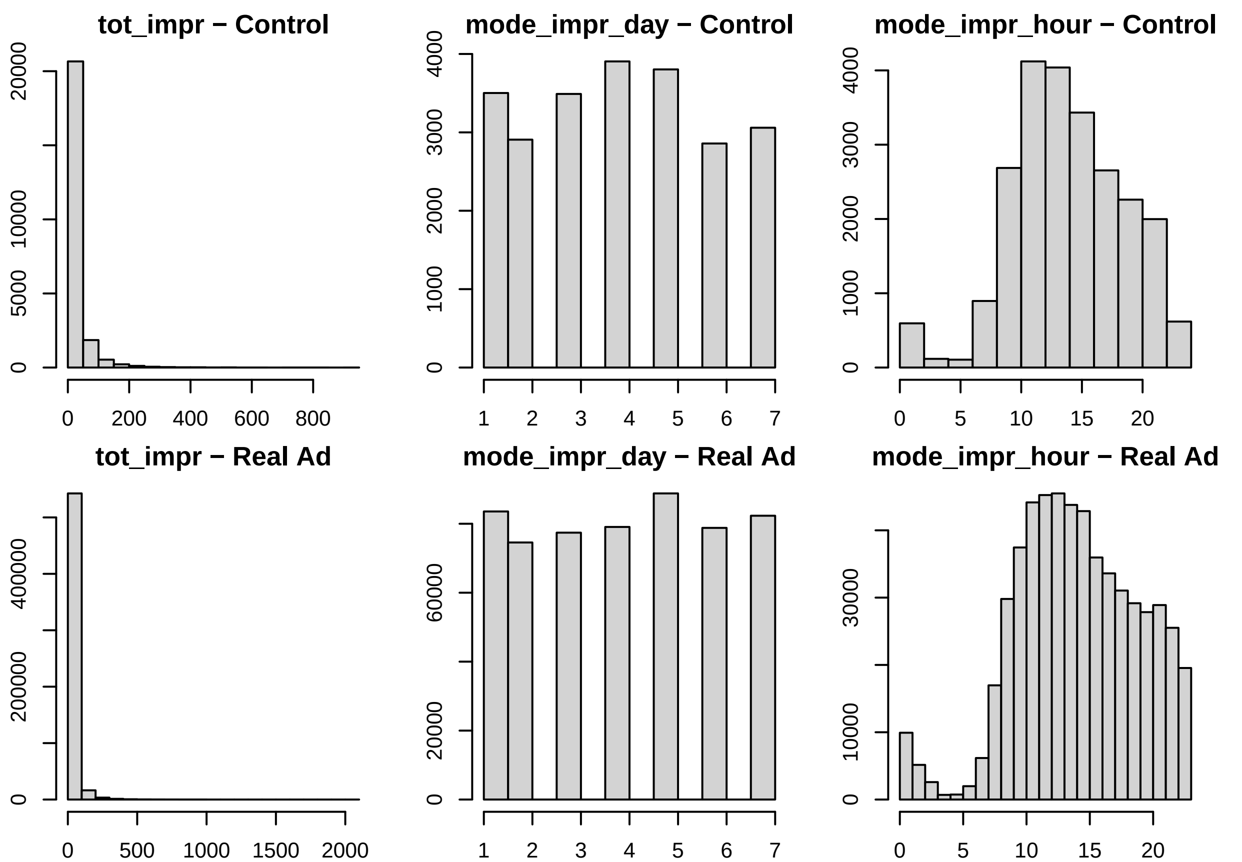
We can see that everything looks balanced here. The averages for each of these variables look similar across control and treatment groups.

Then, we plot the histogram for these three variables to check whether the pre-treatment variables are really balanced across control and treatment groups.

b. Graph the histograms of each variable separately for treatment and control groups.

```
attach(rocketfuel)
par(mar = c(2,2,2,2))
par(mfrow=c(2,3)) # output multiple subfigures into one figure,
# with 3 subfigures each row and 2 rows (one for each treatment) in total

options(scipen=999) # adjust the scientific notation setting

# plot the histogram of tot_impr for control group
```

```r
hist(tot_impr[test==0], main = paste("tot_impr - Control"), xlab = "Control")
# plot the histogram of mode_impr_day for control group
hist(mode_impr_day[test==0], main = paste("mode_impr_day - Control"), xlab = "Control")
# plot the histogram of mode_impr_hour for control group
hist(mode_impr_hour[test==0], main = paste("mode_impr_hour - Control"), xlab = "Control")

# plot the histogram of tot_impr for treatment group
hist(tot_impr[test==1], main = paste("tot_impr - Real Ad"), xlab = "Real Ad")
# plot the histogram of mode_impr_day for treatment group
hist(mode_impr_day[test==1], main = paste("mode_impr_day - Real Ad"), xlab = "Real Ad")
# plot the histogram of mode_impr_hour for treatment group
hist(mode_impr_hour[test==1], main = paste("mode_impr_hour - Real Ad"), xlab = "Real Ad")
```




```r
detach(rocketfuel)
```

By looking down each of these columns of histograms, we can see that not only are the averages similar but the distributions are quite similar too. As this looks well balanced, their outcome differences will be due to the "treatment".

6. Plot the means and confidence intervals of "converted" by control and treatment.

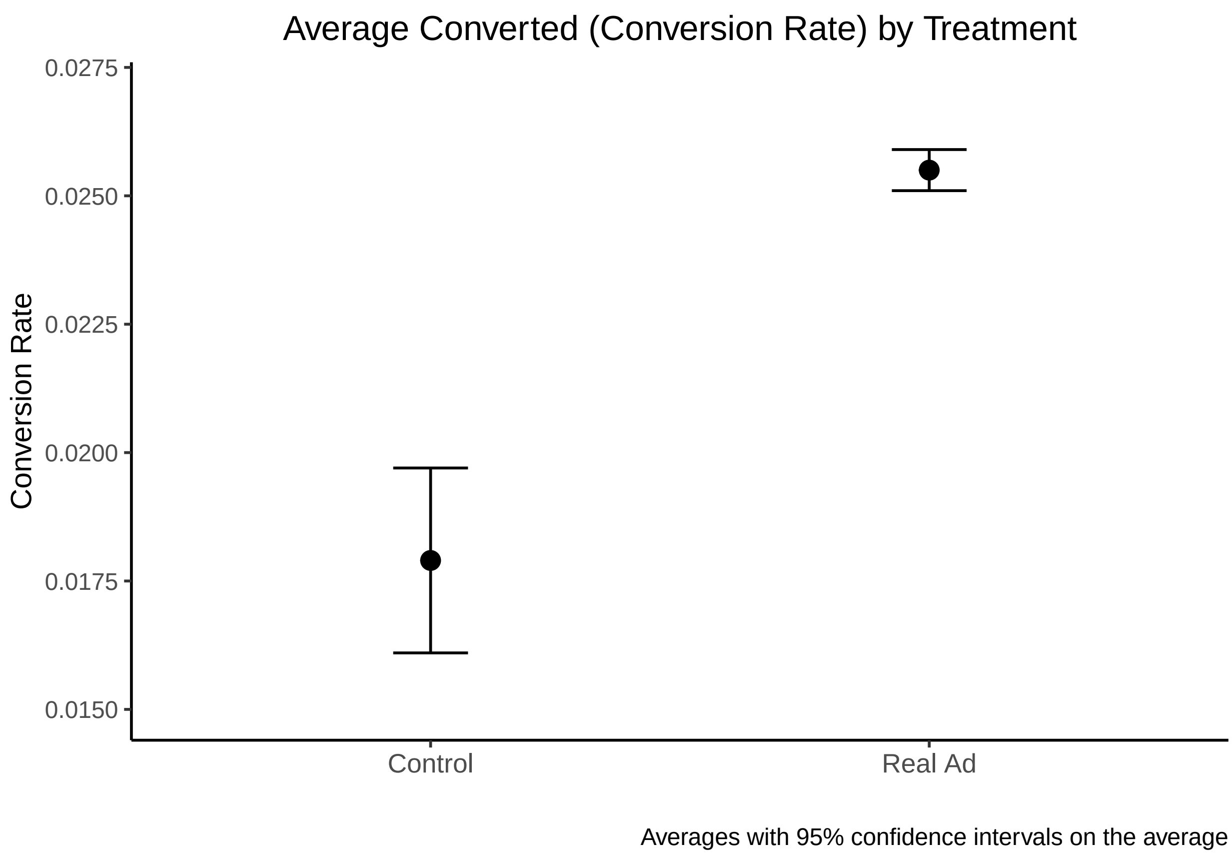
```r
# Create a summary table
summary <- rocketfuel %>%  # create a table called summary that will hold the info
```

```r
  mutate(test = as.factor(test)) %>% # test is a factor variable taking discrete levels
  group_by(test) %>%                 # create groups by test
  summarise(n = length(user_id),     # create a new table with summary measures
            mean.converted = round(mean(converted), 4), # get the mean for each group
            # calculate the standard error on the mean using standard formula
            error.converted = round(sd(converted)/sqrt(n), 4),
            # calculate 95% CI boundaries
            LCI.converted = round(mean.converted - 1.96*error.converted, 4),
            UCI.converted = round(mean.converted + 1.96*error.converted, 4))

kable(summary, caption = "**Average Converted (Conversion Rate)**")
```

Table 5: **Average Converted (Conversion Rate)**

test	n	mean.converted	error.converted	LCI.converted	UCI.converted
0	23524	0.0179	0.0009	0.0161	0.0197
1	564577	0.0255	0.0002	0.0251	0.0259

```r
# Plot the average converted along with 95% CI from that summary table
summary %>%
  ggplot(aes(x=test)) +
  geom_point(aes(y = mean.converted), size = 3) +
  ylim(0.015, 0.027) +
  scale_shape_manual(values=c(15, 16)) +
  labs(
    title = "Average Converted (Conversion Rate) by Treatment",
    caption = "Averages with 95% confidence intervals on the average"
  ) +
  ylab("Conversion Rate") +
  scale_x_discrete(labels=c("0" = "Control", "1" = "Real Ad")) +
  # label the value of treatment on x-axis
  xlab("")+ # eliminate the title of x-axis
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"),
        axis.text.x= element_text(size = 10), legend.position=c(.5,.5),
        plot.title=element_text(hjust=.5))+
  geom_errorbar(aes(ymin = LCI.converted,
                    ymax = UCI.converted), width = .15)+
  scale_color_manual(values=c("darkgrey","black"))
```



We can see that the conversion rate (average converted) for control group is 1.79%, while the conversion rate for real ad (treatment) group is 2.55%, which implies that advertising did help increase conversion rate. Also, the 95% confidence intervals (0.0251, 0.0259) on the average of the treatment group is higher and very narrow (larger sample size in the treatment group), and do not overlap with those of the control group. It suggests that we are quite confident about the estimates of the conversion rate for advertising and know that the ads do make a difference. The result makes sense since we expect that the advertising campaign was effective.

7. Calculate the estimate of the Average Treatment Effect (ATE) of the ads for each treatment relative to control and provide a 95% confidence intervals on the Average Treatment Effects.

a. Calculate the estimate of ATE in a table.

```
attach(rocketfuel)
ATE <- matrix(NA, ncol = 3, nrow = 2)  # create a matrix to store the results
colnames(ATE) <- c("Treatment Effect", "Lower 95% CI", "Upper 95% CI" )
rownames(ATE) <- c("Real Ad", "Control Mean")

# calculate the average treatment effect
effect <- c(summary$mean.converted[2]-summary$mean.converted[1],
            summary$mean.converted[1])
error_ate <- c(sqrt(summary$error.converted[1]^2+summary$error.converted[2]^2), NA)

# calculate the standard error of ATE
LCI <- effect - 1.96*error_ate
```

```r
UCI <- effect + 1.96*error_ate

ATE[,1] <- round(effect,4)
ATE[,2] <- round(LCI,4)
ATE[,3] <- round(UCI,4)

kable(ATE, caption = "**Average Treatment Effect on Conversion Rate**" )
```

Table 6: **Average Treatment Effect on Conversion Rate**

	Treatment Effect	Lower 95% CI	Upper 95% CI
Real Ad	0.0076	0.0058	0.0094
Control Mean	0.0179	NA	NA

```r
detach(rocketfuel)
```

b. Using the regression approach to estimate ATE.

```r
attach(rocketfuel)
# create "dummy variable" for the treatment
treatment <- as.numeric(test == 1)
# estimate standard errors that allow for heteroskedasticty
# (i.e., different standard errors for each treatment)
fit.converted <- lm(converted~treatment, data = rocketfuel)  # simple linear regression
summ(fit.converted, robust = "HC1", confint = TRUE, digits = 4)
```

```
## MODEL INFO:
## Observations: 588101
## Dependent Variable: converted
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,588099) = 54.3229, p = 0.0000
## R² = 0.0001
## Adj. R² = 0.0001
##
## Standard errors: Robust, type = HC1
## ---------------------------------------------------------------
##                       Est.     2.5%    97.5%    t val.        p
## ------------------ -------- -------- -------- --------- --------
## (Intercept)         0.0179   0.0162   0.0195   20.6793   0.0000
## treatment           0.0077   0.0060   0.0094    8.6573   0.0000
## ---------------------------------------------------------------
```

```r
detach(rocketfuel)
```

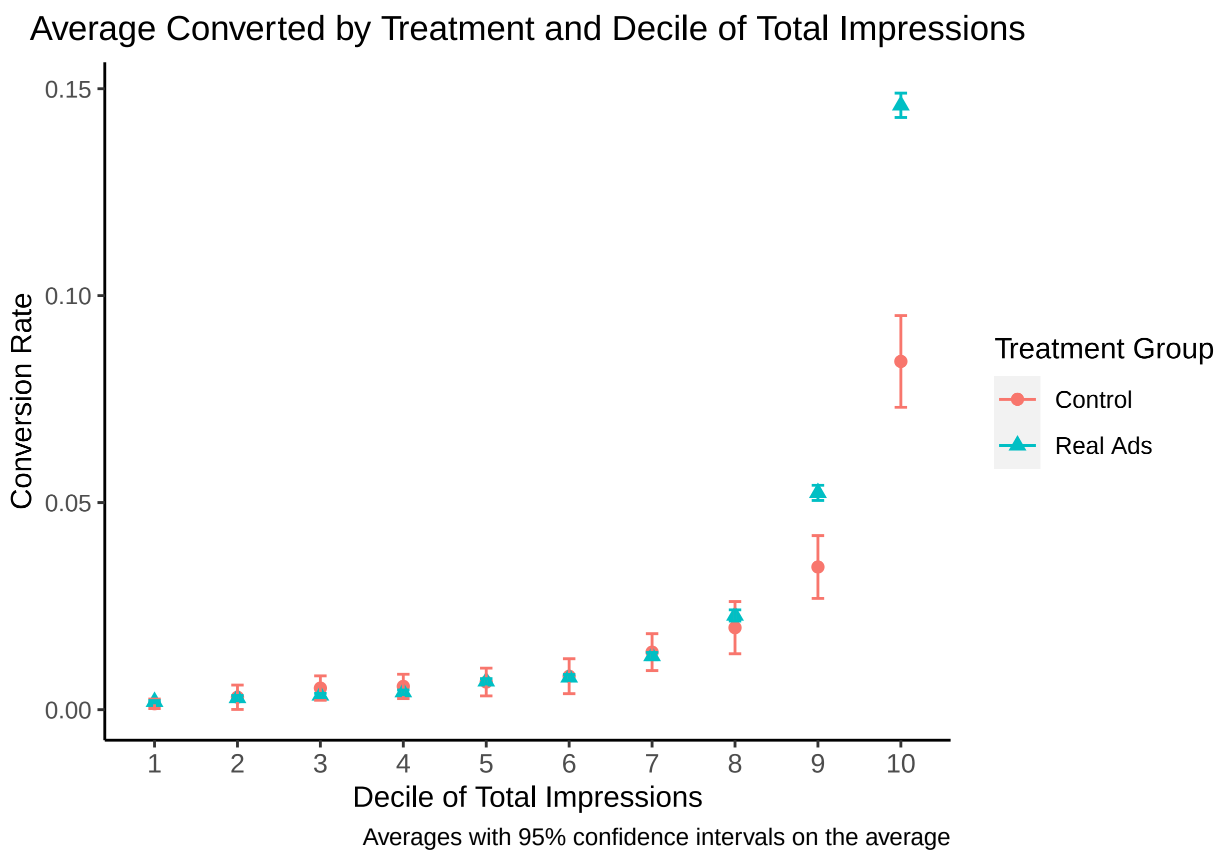
The results of regression approach match the ATE table above.

We can find that relative to the control group (was shown a PSA), "real ad" treatment increases conversion rate by 0.0077. For the treatment group, the interval (0.0060, 0.0094) has 95% chance of containing the true difference in the expected conversion rate. Thus, advertising has a statistically positive significant effect on conversion of users at the 95% confidence level.

8. Next create a graph that shows the mean and 95% CI on "converted" separately for treatment and control plotted over the 10 deciles of total impressions.

```
# create a summary data frame by treatment and decile for graph
summary2 = rocketfuel %>%
  mutate(test = as.factor(test)) %>%  # denote test is a factor variable
  # create decile as a factor variable for tot_impr_decile
  mutate(decile = as.factor(tot_impr_decile)) %>%
  group_by(test, decile) %>%      # create groups by treatment and decile
  # create a new table with summary measures by these groups
  summarise(n = length(user_id),
            mean = mean(converted),   # get the mean for each group
            # calculate the standard error on the mean using standard formula
            error = sd(converted)/sqrt(n),
            # calculate 95% CI boundaries
            LCI = mean - 1.96*error,
            UCI = mean + 1.96*error)
```

```
# Plot the data from summary2
summary2 %>%
  ggplot(aes(decile)) +  # plot over deciles of total impressions
  # plot the averages and give it different shapes and colors by treatment
  geom_point(aes(y = mean, shape = test, color = test), size = 2) +
  # give it error bars with the same coloring by treatment
  geom_errorbar(aes(ymin = LCI, ymax = UCI, color=test), width = .15) +
  labs(
    title = "Average Converted by Treatment and Decile of Total Impressions",
    caption = "Averages with 95% confidence intervals on the average"
  ) +
  # Label the axes
  ylab("Conversion Rate") +
  xlab("Decile of Total Impressions") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
      panel.background = element_blank(), axis.line = element_line(colour = "black"),
      axis.text.x = element_text(size = 10),
      plot.title = element_text(hjust=.5) ) +
  # remake the legend by both color and shape
  scale_shape_discrete(name = "Treatment Group", labels = c("Control", "Real Ads")) +
  scale_color_discrete(name = "Treatment Group", labels = c("Control", "Real Ads"))
```



We can see that when the total number of ad impressions the user encountered is relatively high, especially for 9th-10th decile, the conversion rate of both control and treatment groups increases. It indicates that users are more likely to buy the new handbag when the total number of ad impressions they encountered is large enough. Besides, while the ads are generating enough impressions, advertising will increase conversion rate more significantly if the total number of ad impressions is higher. However, when the total number of ad impressions is at lower level (1st-8th decile), advertising has almost no effect on the conversion rate.

9. Optional: create a graph of the treatment effect with 95% confidence interval on the treatment effect by decile of total impressions.

```
# Re-calculate the treatment effect by decile of total impressions.
attach(rocketfuel)
HTE <- matrix(NA, ncol = 5, nrow = 10)  # create a matrix to store the results
colnames(HTE) <- c("Decile", "ControlMean", "TreatmentEffect", "LowerCI", "UpperCI" )
HTE[,1] <- summary2$decile[1:10]

# call the means from summary2 table
mean.control <- summary2$mean[1:10]
mean.treat <- summary2$mean[11:20]
HTE[,2] <- mean.control

# calculate treatment effect
HTE[,3] <- effect.treat <- mean.treat-mean.control

# calculate sd to construct CI on TE
```

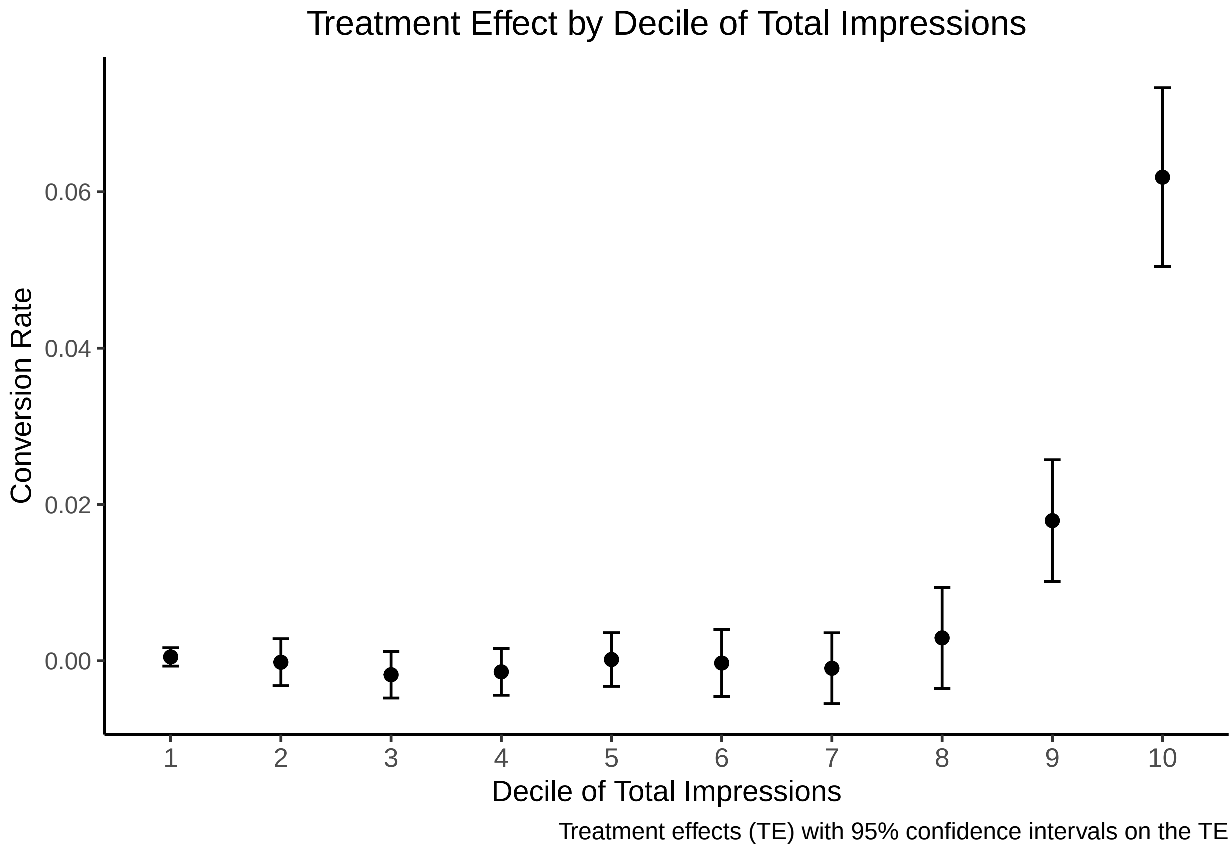
```r
sd.control <- summary2$error[1:10]
sd.treat <- summary2$error[11:20]

# construct the s.d. for computing CI based on the s.d. vector
error.treat <- sqrt(sd.control^2+sd.treat^2)

# computing 95% CI on TE
HTE[,4] <- LCI.treat <- effect.treat - 1.96*error.treat
HTE[,5] <- UCI.treat <- effect.treat + 1.96*error.treat

kable(HTE, caption = "**Treatment Effect by Decile of Total Impressions**")
```

Table 7: **Treatment Effect by Decile of Total Impressions**

Decile	ControlMean	TreatmentEffect	LowerCI	UpperCI
1	0.0014205	0.0004991	-0.0006714	0.0016696
2	0.0030008	-0.0001831	-0.0031870	0.0028208
3	0.0052083	-0.0017775	-0.0047613	0.0012063
4	0.0056225	-0.0014120	-0.0043981	0.0015740
5	0.0066667	0.0001662	-0.0032601	0.0035926
6	0.0080738	-0.0002857	-0.0045612	0.0039898
7	0.0138993	-0.0009535	-0.0054900	0.0035831
8	0.0198073	0.0029344	-0.0035275	0.0093964
9	0.0344673	0.0179333	0.0101494	0.0257171
10	0.0841237	0.0618751	0.0504378	0.0733124

```r
detach(rocketfuel)
```

```r
# Plot the data from HTE
HTE_df <- as.data.frame(HTE)
HTE_df %>%
  ggplot(aes(x=as.factor(Decile))) +  # plot over deciles of total impressions
  # plot the estimates of TE
  geom_point(aes(y = TreatmentEffect), size = 2) +
  scale_shape_manual(values=c(15, 16)) +
  geom_errorbar(aes(ymin = LowerCI, ymax = UpperCI), width = .15) +
  labs(
    title = "Treatment Effect by Decile of Total Impressions",
    caption = "Treatment effects (TE) with 95% confidence intervals on the TE"
  ) +
  # Label the axes
  ylab("Conversion Rate") +
  xlab("Decile of Total Impressions") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"),
        axis.text.x = element_text(size = 10),
        plot.title = element_text(hjust=.5) ) +
  scale_color_manual(values=c("darkgrey","black"))
```



As we found, the treatment effect for 1st-8th decile of total impressions is around zero, which implies that advertising has almost no effect on the conversion rate for low number of impressions. However, the confidence intervals (0.01015, 0.02572) and (0.05044, 0.07331) on the treatment effect for the 9th and 10th decile of total impressions has 95% chance of containing the true difference in the expected conversion rate. Thus, advertising has a statistically positive significant effect on conversion of users at the 95% confidence level if the total number of ad impressions they encountered is large enough. And the treatment effect on conversion rate for the 10th decile of total impressions is the most significant.