# Lab 3: Difference in Differences

Social Media and Web Analytics @ TiSEM

Last updated: 28 April, 2021

## Motivation

Last week, we reviewed linear regression - the workhorse model of a Marketing Analyst's toolkit. When linear regression is further combined with the ability to conduct experiments or find 'natural experiments', the analyst's toolkit is further strengthened through their ability to make **strong** causal claims about the effect of marketing interventions through the use of the Difference-in-Differences (DiD) methodology.

In this tutorial you will apply the DiD methodology to get first hand experience with these tools. We will focus both on how to implement them in `R` and how to correctly interpret the results. The empirical example demonstrates how to use the DiD toolkit to evaluate the effectiveness of search engine marketing on sales revenue of an online company.

## Learning Goals

By the end of this tutorial you will be able to:

1. Estimate treatment effects of a marketing intervention using difference in difference estimates from differences in group averages.
2. Estimate treatment effects of a marketing intervention using difference in difference estimates from linear regression.
3. Critically evaluate the assumptions required for difference in difference estimates to be valid.
4. Correctly interpret difference in difference regression estimates.
5. Display regression estimates in a regression table.

## Instructions to Students

These tutorials are **not graded**, but we encourage you to invest time and effort into working through them from start to finish. Add your solutions to the `lab-03_answer.Rmd` file as you work through the exercises so that you have a record of the work you have done.

Obtain a copy of both the question and answer files using Git. To clone a copy of this repository to your own PC, use the following command:

```
$ git clone https://github.com/tisem-digital-marketing/smwa-lab-03.git
```

Once you have your copy, open the answer document in RStudio as an RStudio project and work through the questions.

The goal of the tutorials is to explore how to "do" the technical side of social media analytics. Use this as an opportunity to push your limits and develop new skills. When you are uncertain or do not know what to do next - ask questions of your peers and the instructors on the class Slack channel `#lab-03-discussion`.

## Exercise 1: Difference in Differences

In 2014, Thomas Blake, Chris Nosko and Steve Tadelis published a study that examines the revenue impact of search engine marketing.[1] Essentially, they worked with eBay to run controlled experiments and turn off search engine marketing in certain parts of the USA and examine the effects on revenue in these regions compared to other regions where marketing was kept on. eBay (like many other companies) intensively used search engine marketing by bidding on different keywords on Google's AdWords platform.

For 8 weeks following May 22nd, 2012, eBay stopped using search engine marketing in a treatment group of 65 out of 210 Designated Market Areas in the USA.[2] eBay then tracked the revenues in each DMA (treatment and control) using the shipping address of customers. The question Blake, Nosko and Tadelis wanted to answer was whether turning off search engine marketing changed eBay's revenue. We want to replicate their analyses using a modified version of their original data.[3]

To gain access to the data, run the following code to download it and save it in the file `data/paid_search.csv`:

```r
url <- "https://raw.githubusercontent.com/TaddyLab/BDS/master/examples/paidsearch.csv"
# where to save data
out_file <- "data/paid_search.csv"
# download it!
download.file(url, destfile = out_file, mode = "wb")
```

You might need to use the following `R` libraries throughout this exercise:[4]

```r
library(readr)
library(dplyr)
library(tidyr)
library(lubridate)
library(fixest)
library(broom)
library(ggplot2)
library(modelsummary)
```

1. What is search engine advertising? Explain the mechanisms through which it might ultimately influence sales revenue.
2. Why would it be difficult to estimate the effectiveness of search engine advertising with purely observational data?
3. Explain why the experiment that Blake, Nosko and Tadelis runs allows them to avoid the issues in (2), and accurately measure the causal effect of search engine advertising on sales.

With some conceptual knowledge under our belt, let's get our hands dirty. We will start by cleaning up the data a little and then producing some summary plots to build up an understanding of the main patterns.

4. Load the data into `R` naming the data `paidsearch`.

[1] Read the paper here. It's a bit of a timeless classic in my opinion.

[2] A DMA is a region in the USA where the population receives the same (or similar) TV and radio offerings, and internet content.

[3] Matt Taddy makes this data available as part of his book Business Data Science. The data has been scaled and translated so that eBay's actual revenues remain unknown, but the transformed data give similar results to the analysis on real data. Any effect we find in our analysis will look very similar to the original paper.

[4] If you haven't installed one or more of these packages, do so by entering `install.packages("PKG_NAME")` into the R console and pressing ENTER.

solution

```r
paidsearch <- read_csv('data/paid_search.csv')
```

```
##
```

```
## -- Column specification --------------------------------------------------------
## cols(
##   date = col_character(),
##   dma = col_double(),
##   treatment_period = col_double(),
##   `search stays on` = col_double(),
##   revenue = col_double()
## )
```

5. What are the column names in the data? If you find that one of the column names in the data has whitespace in it, you will want to replace the whitespace with underscores, "_"

**solution**

```r
print(colnames(paidsearch))
```

```
## [1] "date"             "dma"              "treatment_period" "search stays on"
## [5] "revenue"
```

```r
library(janitor) # students won't know this,
```

```
##
## Attaching package: 'janitor'
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
                 # but once they see it,
                 # they'll learn something
                 # There's many other ways to go

paidsearch <-
    paidsearch %>%
    clean_names() # by default lowercases all column names
                  # and replaces whitespace with underscores
```

6. Revenue is reported in USD. Modify the `revenue` variable so that the new values are in '000s of USD.

**solution**

```r
paidsearch <-
    paidsearch %>%
    mutate(revenue = revenue / 1000)
```

To draw some descriptive plots we will need the `date` variable to be formatted as a date (rather than as a character string). Run the following code to make this conversion:

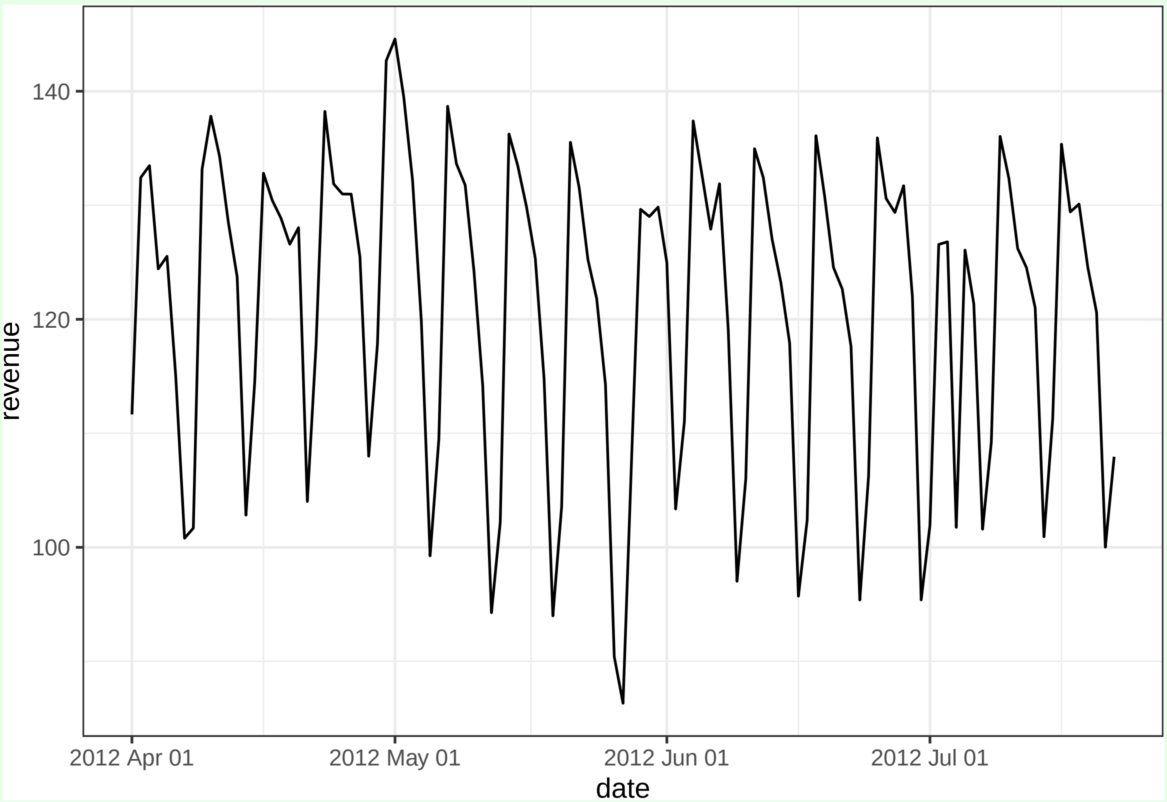
**solution**

```r
paidsearch <-
    paidsearch %>%
    mutate(date  = as_date(date, format = '%d-%b-%y'))
```

7. Compute the average revenue per DMA each day. Plot the data so that you can see how average revenue evolves over time.

solution

```r
paidsearch %>%
    group_by(date) %>%
    summarise(revenue = mean(revenue)) %>%
    ggplot(aes(x = date,
               y = revenue
               )
           ) +
    geom_line() +
    scale_x_date(date_labels = "%Y %b %d") +
    theme_bw()
```




The plot in (7) likely shows a lot of cyclicality within each week. This makes it hard to visualize broader patterns or differences across groups. We will compute averages for each calendar week to smooth out the cyclicality within each week. To make this easier, run the following code to extract the calendar week from each date in the data:

solution

```r
paidsearch <-
    paidsearch %>%
    mutate(calweek = week(date))
```

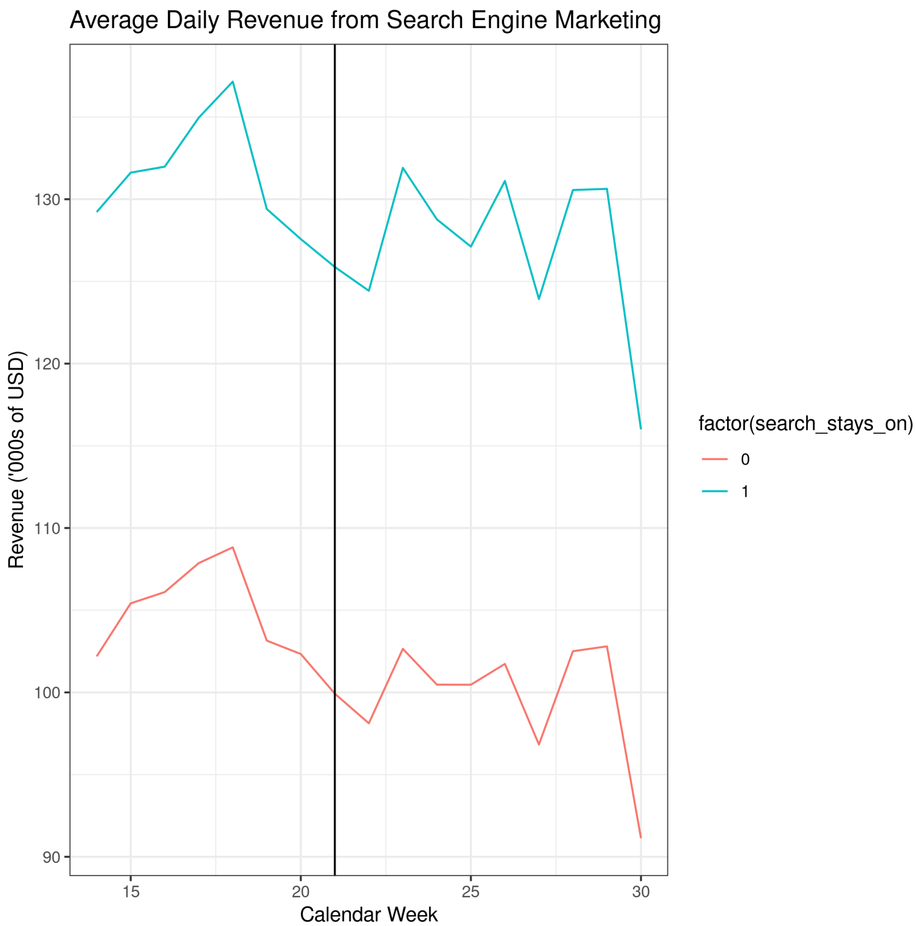
8. Compute the average daily revenue per calendar week for DMAs in which eBay turn off their search advertising and for DMAs that leave search advertising turned on. The resulting dataset should have 34 observations, 17 where `search_stays_on =1` and 17 where `search_stays_on =0`.

solution

```r
grp_avg <-
  paidsearch %>%
  group_by(search_stays_on, calweek) %>%
  summarise(revenue = mean(revenue)) %>%
  arrange(calweek, search_stays_on)
```

```
## `summarise()` has grouped output by 'search_stays_on'. You can override using the `.groups` argume
```

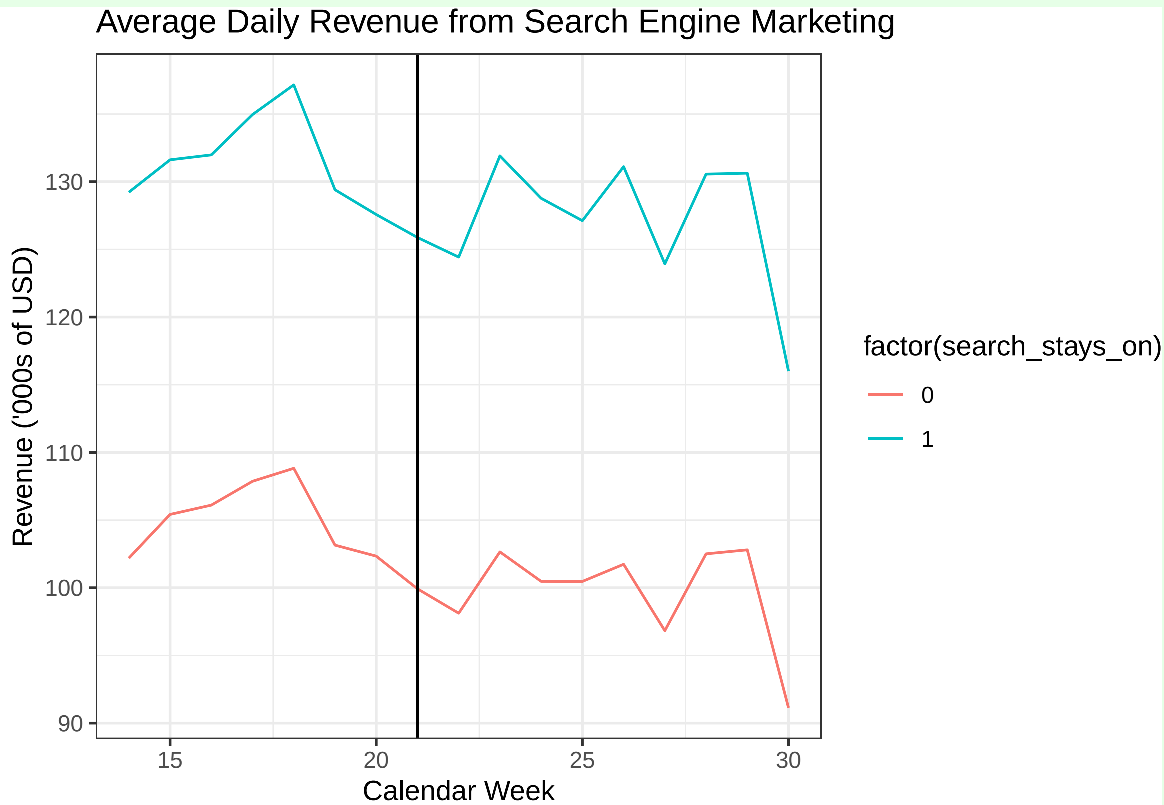
9. Plot the average daily revenue per week for`search_stays_on =1` and `search_stays_on =0`. Add a vertical line denoting where the experiment begins (22 May 2012). Your final plot should resemble this one:




solution

```r
exp_start <- as_date("2012-05-22")

grp_avg %>%
    ggplot(aes(x = calweek,
               y = revenue,
               color = factor(search_stays_on)
               )
           ) +
    geom_line() +
    geom_vline(xintercept = week(exp_start)) +
    ggtitle("Average Daily Revenue from Search Engine Marketing") +
    xlab("Calendar Week") +
    ylab("Revenue ('000s of USD)") +
    theme_bw()
```




10. Is average revenue different for DMAs that have turned off search advertising compared to those where it remains on? Why might this be the case?

11. From the graph above, can you "eye-conometrically" see any effect of turning off search engine ads?[5]

Now we will go back to working with the full dataset `paidsearch`.

12. Create a new variable `treatment` that takes the value 1 if search engine advertising is switched off, and 0 if search engine advertising stays on. Also rename the variable `treatment_period` to `after`.

[5] All "clean" (well-executed) Difference in Difference papers should produce a plot where you can visually assess what is going on. If you don't see one, its could a warning sign that whatever comes next is an artifact of the statistical model, rather than the experiment (or, if published in a journal *maybe* an editor asked to suppress the figure to save space).

solution

```r
paidsearch <-
    paidsearch %>%
    mutate(treatment = case_when(
        search_stays_on == 0 ~ 1,
        TRUE ~ 0
        )
    ) %>%
    rename(after = treatment_period)
```

13. We can compute a Difference in Difference estimate from a set of group means:

$$\hat{\beta}_{DiD} = (\bar{y}_{after=1,treat=1} - \bar{y}_{after=0,treat=1}) - (\bar{y}_{after=1,treat=0} - \bar{y}_{after=0,treat=0})$$

where $\bar{y}$ is average revenue, `after=1` denotes dates after the treatment starts, `after=0` denotes dates before the treatment starts, `treat=1` denotes DMAs that where part of the treatment group and `treat=0` denotes DMAS that where part of the control group. Thus $\bar{y}_{after=1,treat=1}$ is average revenue for DMAs in the treatment group for the period after the treatment has begun.

Compute each of these four $\bar{y}$'s.

solution

```r
treat_table <-
    paidsearch %>%
    group_by(after, treatment) %>%
    summarise(revenue = mean(revenue)) %>%
    ungroup()
```

```
## `summarise()` has grouped output by 'after'. You can override using the `.groups` argument.
```

14. Use the averages you computed in (13) to the treatment effect of turning off search engine advertising, i.e. $\hat{\beta}_{DiD}$.

solution

```r
# many ways to do this, here one way
treat_table %>%
    # makes treat and control group each on column
    pivot_wider(names_from = treatment,
                values_from = revenue
                ) %>%
    mutate_all(funs(. - lag(.))) %>%
    mutate(did_simple = `1` - `0`) %>%
    na.omit() %>%
    select(did_simple)
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
```

```
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))

## # A tibble: 1 x 1
##   did_simple
##        <dbl>
## 1      -1.09
```

15. Interpret the effect you computed from the point of view of a marketing analyst working at eBay. Is it large from a marketing viewpoint?
16. We rarely see analytical work use this difference in averages approach. Can you explain why that might be the case?
17. Estimate a linear regression that computes the equivalent Difference-in-Difference estimate as (13).

**solution**

```r
# with y --> levels
reg_did <- feols(revenue ~ after +
                       treatment +
                       after:treatment,
                   data = paidsearch)

tidy(reg_did,  conf.int = TRUE)
```

```
## # A tibble: 4 x 7
##   term            estimate std.error statistic      p.value conf.low conf.high
##   <chr>              <dbl>     <dbl>     <dbl>        <dbl>    <dbl>     <dbl>
## 1 (Intercept)       132.        2.69    48.9   0                126.     137.
## 2 after              -3.49      3.63    -0.959 0.337            -10.6      3.64
## 3 treatment         -26.6       4.73    -5.62  0.0000000199     -35.8    -17.3
## 4 after:treatment    -1.09      6.39    -0.171 0.864            -13.6     11.4
```

18. Is $\hat{\beta}_{DiD}$ statistically significant? How can you tell?
19. We can repeat the estimation in (17) using $\log(revenue)$ rather than *revenue*. Why might we want to do that? Run this regression and interpret the magnitude of the treatment effect.

solution

```
# with log y --> elasticity
reg_did_log <- feols(log(revenue) ~ after +
                       treatment +
                       after:treatment,
                     data = paidsearch)

tidy(reg_did_log, conf.int = TRUE)
```

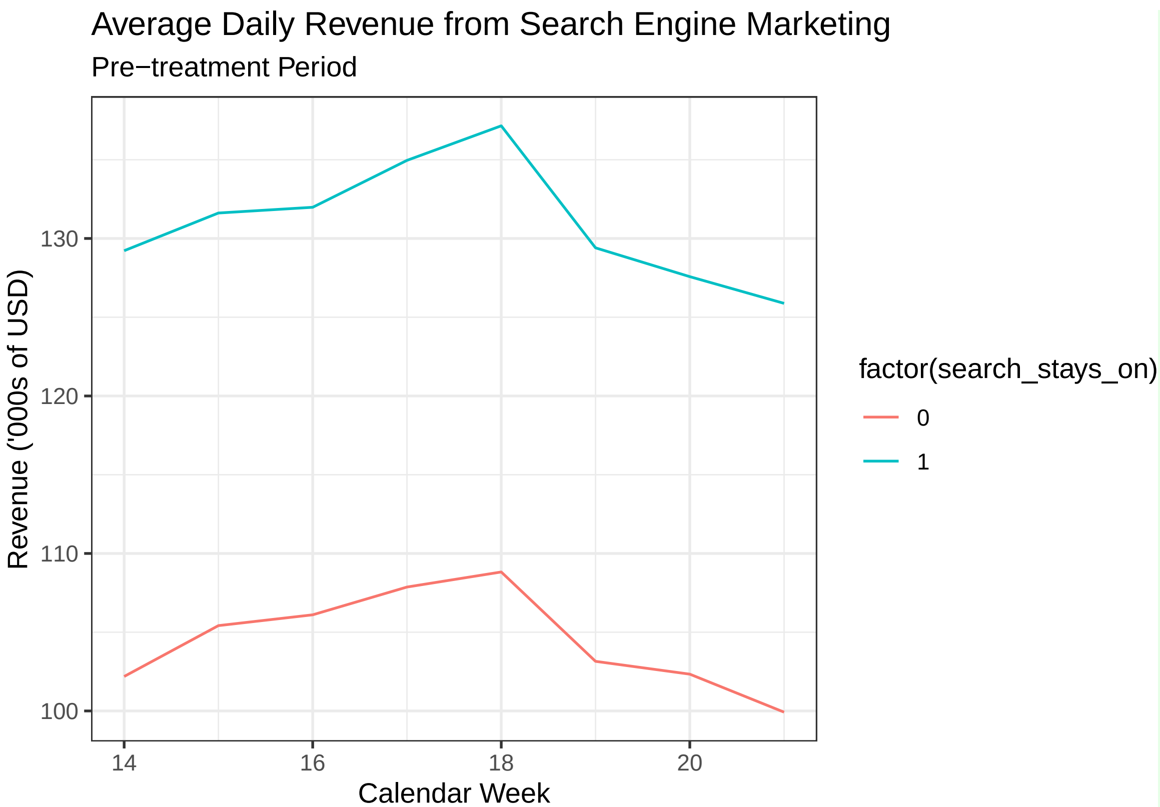
```
## # A tibble: 4 x 7
##   term            estimate std.error statistic p.value conf.low conf.high
##   <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)       4.04      0.0142   286.     0         4.01     4.07
## 2 after            -0.0394    0.0191    -2.06   0.0392  -0.0769  -0.00195
```

```
## 3 treatment        0.0141     0.0249     0.566  0.571   -0.0347   0.0628
## 4 after:treatment -0.00659    0.0336    -0.196  0.844   -0.0724   0.0592
```

20. What standard errors were computed with your regression estimates so far? Are these appropriate? If not, adjust them to compute what you believe to be more conservative (and unbiased) standard error estimates.
21. Do you think fixed effects should be added into the regression? Why? Add them to your regression and report the results.
22. Construct a regression table that presents three or more of the regressions you have run. Make it look presentable, such that you could use it in a presentation to eBay stakeholders if you were in the shoes of Blake, Nosko and Tadelis.
23. Construct a coefficient plot that presents the difference in difference estimates of three or more of the regressions you have run. Make it look presentable, such that you could use it in as an alternative to the regression table in presentation to eBay stakeholders if you were in the shoes of Blake, Nosko and Tadelis.
24. A crucial assumption for the Difference-in-Differences estimate to accurately measure the effect of turning off search engine ads is the presence of parallel trends. What is the parallel trends assumption and why do we need it? Does it appear satisfied in our data?

**solution**

```
grp_avg %>%
    filter(calweek <= week(exp_start)) %>%
    ggplot(aes(x = calweek,
               y = revenue,
               color = factor(search_stays_on)
               )
           ) +
    geom_line() +
    ggtitle("Average Daily Revenue from Search Engine Marketing",
            subtitle = "Pre-treatment Period") +
    xlab("Calendar Week") +
    ylab("Revenue ('000s of USD)") +
    theme_bw()
```




25. Can we definitively conclude from the results above that search engine marketing does not pay off? Explain your answer.