Difference-in-Differences

Lucas Owen

There are different methods for identifying causality with observational data - difference-in-differences, regression discontinuity, and instrumental variables. Today we'll focus on diff-in-diff.

The data

This data was collected to evaluate the National Supported Work (NSW) Demonstration, a randomized employment program studied by LaLonde (1986). The NSW program was designed to help disadvantaged individuals—such as welfare recipients, high school dropouts, ex-offenders, and ex-drug addicts transition into the labor market. Participants were randomly assigned to either a treatment group, which received program services, or a control group, which did not. This random assignment helped ensure that differences in outcomes between groups could be attributed to the program itself rather than pre-existing characteristics.

Those in the treatment group were guaranteed 9 to 18 months of subsidized employment, worked in small crews, and received counseling and performance-based wage increases. Job placements varied by site and included roles in construction, service, and clerical work. Implemented in ten U.S. cities during the mid-1970s, the NSW program aimed to provide structured work experience and a pathway into regular employment.

The paper in question: *Lalonde, R. (1986) Evaluating the Econometric Evaluations of Training Programs, American Economic Review, 76, 604-620.*

Variable	Description
id	Matched pair id, 1, 1, 2, 2, . . . , 185, 185.
z	z=1 for treated, z=0 for control
age	Age in years
edu	Education in years
black	1=black, 0=other
hisp	1=Hispanic, 0=other
married	1=married, 0=other
nodegree	1=no High School degree, 0=other
re74	Earnings in 1974, a covariate
re75	Earnings in 1975, a covariate
re78	Earnings in 1978, an outcome

1. Descriptive statistics

First, let's load the data and make sure it is clean. Let's also assess the balance between treatment and control groups for the variables `black`, `married`, `nondegree`, and `age`.

```
#### LOAD DATA ####

nsw <- read.csv(file="data/321/nsw.csv")
head(nsw)
```

```
##   id z age edu black hisp married nodegree re74 re75      re78
## 1  1 1  37  11     1    0       1        1    0    0  9930.046
## 2  1 0  34  11     1    0       1        1    0    0  6040.335
## 3  2 1  22   9     0    1       0        1    0    0  3595.894
## 4  2 0  20   9     0    1       0        1    0    0  8329.823
## 5  3 1  30  12     1    0       0        0    0    0 24909.449
## 6  3 0  33  12     1    0       0        0    0    0  5970.257
```

```
#### CLEAN DATA ####
freq.na(nsw) # are there any NAs?
```

```
##          missing %
## id             0 0
## z              0 0
## age            0 0
## edu            0 0
## black          0 0
## hisp           0 0
## married        0 0
## nodegree       0 0
## re74           0 0
## re75           0 0
## re78           0 0
```

```
nsw$treat <- nsw$z
nsw$treat_f <- factor(nsw$z,
                      levels = c(0,1),
                      labels=c("control","treatment"))
nsw$black_f <- factor(nsw$black,
                      levels = c(0,1),
                      labels=c("non-black","black"))
nsw$married_f <- factor(nsw$married,
                      levels = c(0,1),
                      labels=c("non-married","married"))
nsw$nodegree_f <- factor(nsw$nodegree,
                      levels = c(0,1),
                      labels=c("degree","nodegree"))

#### CHECKING BALANCE ####

## proportion of blacks
mean(nsw$black)
```

```
## [1] 0.8486486
```

```
t.test(black ~ treat_f, data = nsw)
```

```
##
##  Welch Two Sample t-test
##
## data:  black by treat_f
## t = 0.28936, df = 367.68, p-value = 0.7725
```

```
## alternative hypothesis: true difference in means between group control and group treatment is not equ
## 95 percent confidence interval:
##  -0.06265644  0.08427806
## sample estimates:
##   mean in group control mean in group treatment
##               0.8540541               0.8432432
```

```r
## proportion of college educated
mean(nsw$nodegree)
```

```
## [1] 0.7378378
```

```r
t.test(nodegree ~ treat_f, data = nsw)
```

```
##
##  Welch Two Sample t-test
##
## data:  nodegree by treat_f
## t = 1.2997, df = 366.03, p-value = 0.1945
## alternative hypothesis: true difference in means between group control and group treatment is not equ
## 95 percent confidence interval:
##  -0.03050347  0.14942238
## sample estimates:
##   mean in group control mean in group treatment
##               0.7675676               0.7081081
```

```r
## proportion of married
mean(nsw$married)
```

```
## [1] 0.1945946
```

```r
t.test(married ~ treat_f, data = nsw)
```

```
##
##  Welch Two Sample t-test
##
## data:  married by treat_f
## t = 0.26195, df = 367.84, p-value = 0.7935
## alternative hypothesis: true difference in means between group control and group treatment is not equ
## 95 percent confidence interval:
##  -0.07034463  0.09196626
## sample estimates:
##   mean in group control mean in group treatment
##               0.2000000               0.1891892
```

```r
## balance on age
t.test(age ~ treat_f, data = nsw)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  age by treat_f
## t = -0.15139, df = 367.91, p-value = 0.8798
## alternative hypothesis: true difference in means between group control and group treatment is not equ
## 95 percent confidence interval:
##  -1.587988  1.360961
## sample estimates:
##   mean in group control mean in group treatment
##                25.70270                25.81622
```

Second, let's check the correlation matrix of all our quantitative variables.

```
round(
  cor(nsw[,-c(1,13:16)]), digits = 2
)
```

```
##              z   age   edu black  hisp married nodegree  re74  re75  re78 treat
## z         1.00  0.01  0.04 -0.02  0.01   -0.01    -0.07  0.01  0.01  0.15  1.00
## age       0.01  1.00  0.00  0.08 -0.06    0.20    -0.09  0.01  0.05  0.07  0.01
## edu       0.04  0.00  1.00 -0.01 -0.09    0.07    -0.67  0.09  0.01  0.13  0.04
## black    -0.02  0.08 -0.01  1.00 -0.58    0.06     0.07 -0.03 -0.08 -0.14 -0.02
## hisp      0.01 -0.06 -0.09 -0.58  1.00    0.00     0.04  0.01  0.09  0.08  0.01
## married  -0.01  0.20  0.07  0.06  0.00    1.00     0.00  0.18  0.28  0.04 -0.01
## nodegree -0.07 -0.09 -0.67  0.07  0.04    0.00     1.00 -0.09  0.05 -0.12 -0.07
## re74      0.01  0.01  0.09 -0.03  0.01    0.18    -0.09  1.00  0.70  0.09  0.01
## re75      0.01  0.05  0.01 -0.08  0.09    0.28     0.05  0.70  1.00  0.08  0.01
## re78      0.15  0.07  0.13 -0.14  0.08    0.04    -0.12  0.09  0.08  1.00  0.15
## treat     1.00  0.01  0.04 -0.02  0.01   -0.01    -0.07  0.01  0.01  0.15  1.00
```

Next, let's make a table of descriptive statistics for all variables except `id`.

```
# Note: to display LaTeX tables from stargazer in your knitted PDF you must set your code chunk with th
stargazer(nsw[,-1],
          header = FALSE,
          type = "latex") # change to type = "text" to display in the console
```

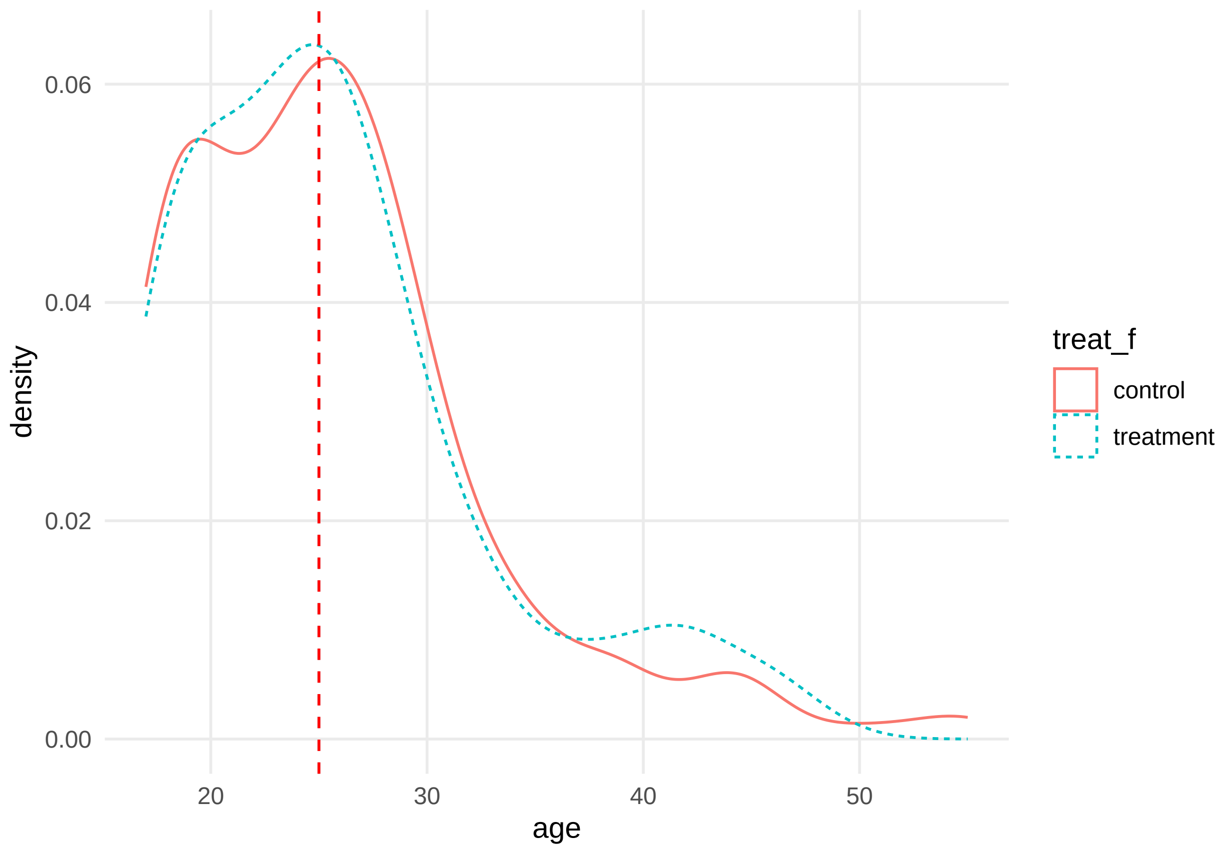
Table 2:

Statistic	N	Mean	St. Dev.	Min	Max
z	370	0.500	0.501	0	1
age	370	25.759	7.202	17	55
edu	370	10.270	1.865	3	16
black	370	0.849	0.359	0	1
hisp	370	0.057	0.232	0	1
married	370	0.195	0.396	0	1
nodegree	370	0.738	0.440	0	1
re74	370	2,052.511	4,945.299	0.000	35,040.070
re75	370	1,508.557	3,308.116	0.000	25,142.240
re78	370	5,328.255	6,643.759	0.000	60,307.930
treat	370	0.500	0.501	0	1

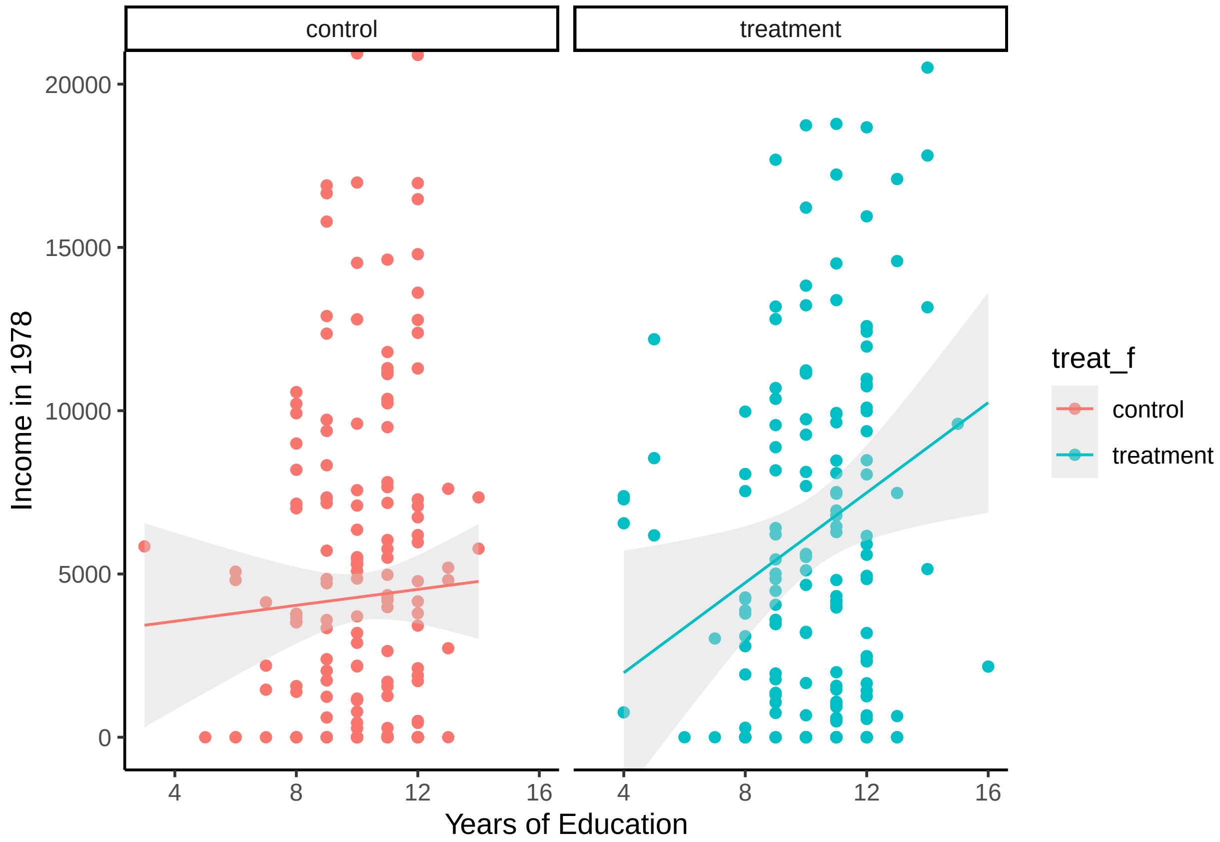
2. Exploratory data analysis

Let's check the distribution of our data more visually.

```
## visualize the age distribution by treatment group
ggplot(data=nsw, aes(x=age, group=treat_f)) +
  geom_density(aes(linetype=treat_f, color=treat_f)) +
  geom_vline(xintercept = median(nsw$age[nsw$treat==1]),
             linetype="dashed", color="lightblue") +
  geom_vline(xintercept = median(nsw$age[nsw$treat==0]),
             linetype="dashed", color="red")
```




```
# scatter plot
ggplot(data=nsw,
       mapping = aes(x=edu,
                     y=re78,
                     color=treat_f)) +
  geom_point() +
  geom_smooth(method = "lm", fill="lightgrey", linewidth=0.5) +
  coord_cartesian(ylim=c(0,20000)) +
  facet_wrap(~treat_f) +
  theme_classic()+
  xlab("Years of Education")+
  ylab("Income in 1978")
```



3. Estimation of policy effects

We want to evaluate the treatment effect on the outcome. Before estimating this quantity, let's double check whether the outcome `re78` is correlated or associated with the unbalanced variables. Then, let's estimate the following conditional expectations:

$$E[\ re78 \mid z = treated] \quad - \quad E[\ re78 \mid z = control]$$

Can we interpret this estimator as a causal effect? Why or why not?

```
t.test(re78 ~ black_f, data = nsw)
```

```
##
##  Welch Two Sample t-test
##
## data:  re78 by black_f
## t = 3.1077, df = 86.058, p-value = 0.002555
## alternative hypothesis: true difference in means between group non-black and group black is not equal
## 95 percent confidence interval:
##   940.8257 4281.1545
## sample estimates:
## mean in group non-black     mean in group black
##                7544.068                4933.078
```

```r
t.test(re78 ~ nodegree_f, data = nsw)
```

```
##
##  Welch Two Sample t-test
##
## data:  re78 by nodegree_f
## t = 2.0654, df = 145.09, p-value = 0.04066
## alternative hypothesis: true difference in means between group degree and group nodegree is not equal
## 95 percent confidence interval:
##    76.09038 3456.50640
## sample estimates:
##   mean in group degree mean in group nodegree
##               6631.497               4865.198
```

```r
t.test(re78 ~ married, data = nsw)
```

```
##
##  Welch Two Sample t-test
##
## data:  re78 by married
## t = -0.74489, df = 104.73, p-value = 0.458
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2444.982  1109.640
## sample estimates:
## mean in group 0 mean in group 1
##         5198.33         5866.00
```

```r
t.test(re78 ~ treat_f, data = nsw)
```

```
##
##  Welch Two Sample t-test
##
## data:  re78 by treat_f
## t = -2.9873, df = 310.03, p-value = 0.003039
## alternative hypothesis: true difference in means between group control and group treatment is not equ
## 95 percent confidence interval:
##  -3386.6340  -696.9204
## sample estimates:
##   mean in group control mean in group treatment
##                4307.366                6349.144
```

Let's re-estimate the quantity using linear regression.

```r
lm1 <- lm(re78 ~ treat_f, data=nsw)
summary(lm1)
```

```
##
## Call:
## lm(formula = re78 ~ treat_f, data = nsw)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6349  -4307  -1973   2989  53959
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4307.4      483.3   8.912   <2e-16 ***
## treat_ftreatment    2041.8      683.5   2.987    0.003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6574 on 368 degrees of freedom
## Multiple R-squared:  0.02368,    Adjusted R-squared:  0.02102
## F-statistic: 8.924 on 1 and 368 DF,  p-value: 0.003003
```

```r
lm2 <- lm(re78 ~ treat_f + edu + nodegree + black_f, data=nsw)
summary(lm2)
```

```
##
## Call:
## lm(formula = re78 ~ treat_f + edu + nodegree + black_f, data = nsw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9761  -4253  -1611   3124  54208
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2789.8     3151.5   0.885  0.37661
## treat_ftreatment    1933.8      675.4   2.863  0.00444 **
## edu                  387.5      242.8   1.596  0.11139
## nodegree            -373.5     1032.3  -0.362  0.71770
## black_fblack       -2513.0      943.9  -2.662  0.00810 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6480 on 365 degrees of freedom
## Multiple R-squared:  0.0589, Adjusted R-squared:  0.04859
## F-statistic: 5.711 on 4 and 365 DF,  p-value: 0.0001812
```

```r
stargazer(lm1,lm2,
          header = FALSE,
          type="text")  # change to type = "text" to display in the console
```

```
##
## ==================================================================
##                                  Dependent variable:
##                      ---------------------------------------------
##                                         re78
##                               (1)                    (2)
## ------------------------------------------------------------------
## treat_ftreatment        2,041.777***           1,933.751***
##                          (683.485)              (675.382)
```

```
##
## edu                                              387.511
##                                                 (242.825)
##
## nodegree                                         -373.506
##                                                (1,032.313)
##
## black_fblack                                  -2,512.997***
##                                                 (943.872)
##
## Constant                   4,307.366***          2,789.775
##                              (483.297)          (3,151.452)
##
## -----------------------------------------------------------------
## Observations                   370                  370
## R2                            0.024                0.059
## Adjusted R2                   0.021                0.049
## Residual Std. Error  6,573.553 (df = 368)   6,480.335 (df = 365)
## F Statistic         8.924*** (df = 1; 368) 5.711*** (df = 4; 365)
## =================================================================
## Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

Given the temporal dimension of this experiment, we can instead do a diff-in-diff! let's estimate a differences-in-differences estimator using `re75` to represent participants' earnings before the intervention and `re78` to represent earnings after the intervention. Under what assumptions does this estimator identify a causal effect? Recall:

$$\text{DiD} = [\bar{Y}(1)_{after} - \bar{Y}(0)_{after}] - [\bar{Y}(1)_{before} - \bar{Y}(0)_{before}]$$

```r
# differences in means before the intervention:
before <- mean(nsw$re75[nsw$z==1]) - mean(nsw$re75[nsw$z==0])

# differences in means after the intervention:
after <- mean(nsw$re78[nsw$z==1]) - mean(nsw$re78[nsw$z==0])

# differences in differences:
(DiD <- after - before)
```

```
## [1] 1994.78
```

What is the internal and external validity of this experiment results?

4. DID & Parallel trends

A Difference-in-Differences (DiD) design is a statistical method used to estimate the causal effect of a treatment or intervention when we have data over time for both a treatment group (those exposed to the intervention) and a control group (those not exposed). Rather than just comparing outcomes after the intervention, DiD uses changes in outcomes over time—differences in differences—to control for underlying trends that might affect both groups.

The key assumption behind DiD is called the parallel trends assumption. This means that, in the absence of the treatment, the average outcomes for the treatment and control groups would have followed the same trend over time. If this assumption holds, any divergence in outcomes after the intervention can be attributed to the treatment.

The power of DiD comes from using within-unit changes over time to control for unobserved fixed differences between treated and control units. In other words, DiD differences out time-invariant confounders even if treatment was not randomly assigned. In this case, it measures the average effect of treatment on the treated, not the average treatment effect! These are different because the treatment effect size may differ between treatment and control groups. Random assignment strengthens internal validity but is not required.

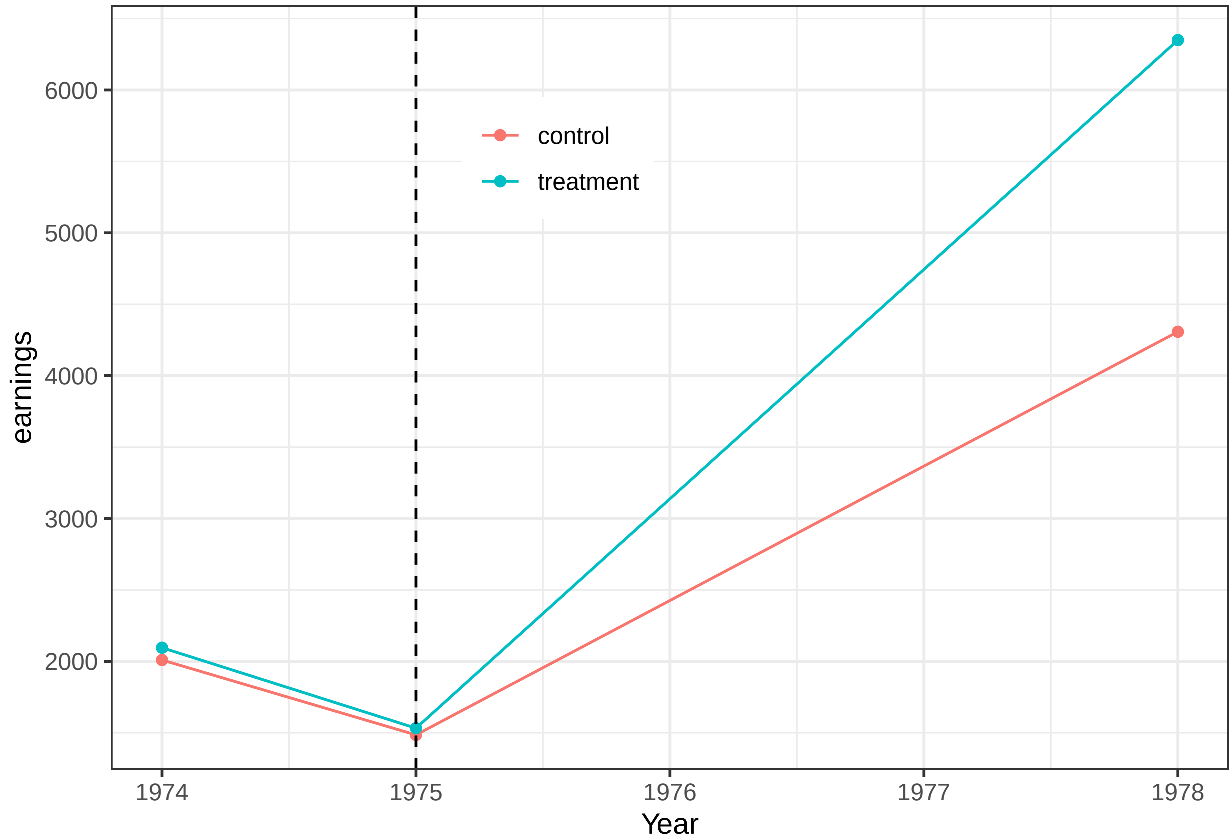
When treatment is not random, you must be extra careful to justify the parallel trends assumption. Time variant confounders should main concern!

Although we can't directly test this assumption (since we never observe the counterfactual), we can provide visual evidence to support it by comparing trends in the outcome before the intervention.

To do this, we'll create a line plot that shows average earnings for the treatment and control groups across the years 1974, 1975, and 1978. This will help us assess whether the two groups followed similar trends before the intervention occurred (in 1978).

Let's start by extracting the variables `re74`, `re75`, and `re78`, and convert the dataset from wide to long format using `pivot_longer()`. This will give us one column for the year, one for earnings, and one indicating whether the individual was in the treatment or control group. Then, we'll use the `aggregate()` function to compute average earnings by year and group. Finally, we plot the trends using `ggplot2`.

```r
# 1. Add individual ID
nsw$id <- 1:nrow(nsw)

# 2. Select and reshape data to long format
df_long <- nsw[, c("id", "treat_f", "re74", "re75", "re78")]

df_long <- pivot_longer(df_long,
                        cols = c("re74", "re75", "re78"),
                        names_to = "year",
                        values_to = "earnings")

# 3. Recode year as numeric
df_long$year <- as.numeric(gsub("re", "19", df_long$year))

# 4. Create post-treatment indicator (1978 is post)
df_long$post <- ifelse(df_long$year == 1978, 1, 0)

#5. Pre-trends
vis <- aggregate(earnings ~ year + treat_f, data=df_long, FUN=mean)
ggplot(data=vis,
       mapping=aes(x=year,
                   y=earnings,
                   group=treat_f,
                   color=treat_f)) +
  geom_line() +
  geom_point() +
  theme_bw() +
  labs(x="Year") +
  theme(legend.position = c(0.4, 0.8),
        legend.title = element_blank()) +
  geom_vline(xintercept=1975, linetype=2)
```




```r
# 6. Run DiD regression with clustered SEs by individual
did_model <- lm_robust(earnings ~ treat_f * post,
                       clusters = id,
                       data = df_long)

# 7. View results
summary(did_model)
```

```
##
## Call:
## lm_robust(formula = earnings ~ treat_f * post, data = df_long,
##     clusters = id)
##
## Standard error type:  CR2
##
## Coefficients:
##                        Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper
## (Intercept)             1747.25      291.2  6.0000 1.026e-08   1172.7   2321.8
## treat_ftreatment          66.56      397.5  0.1675 8.671e-01   -715.1    848.2
## post                    2560.11      441.2  5.8028 2.802e-08   1689.7   3430.5
## treat_ftreatment:post   1975.22      758.8  2.6031 9.612e-03    483.1   3467.3
##                         DF
## (Intercept)            184
## treat_ftreatment       368
## post                   184
## treat_ftreatment:post  368
```

```
##
## Multiple R-squared:  0.1074 ,    Adjusted R-squared:  0.105
## F-statistic: 29.31 on 3 and 369 DF,  p-value: < 2.2e-16
```

```r
# effect size!
reference <- did_model$coefficients[1]+did_model$coefficients[2]+did_model$coefficients[3]
did_model$coefficients[4] / reference
```

```
## treat_ftreatment:post
##             0.4515886
```

```r
#45% higher relative to the counterfactual!
```