**Category**

* [Advanced Modeling](https://datascienceplus.com/category/advanced-modeling/)

**Tags**

* [Linear Regression](https://datascienceplus.com/tag/linear-regression/)
* [Logistic Regression](https://datascienceplus.com/tag/logistic-regression/)
* [R Programming](https://datascienceplus.com/tag/rstats/)

Regression analysis is one of the most demanding machine learning methods in 2019. One group of regression analysis for measuring effects and to evaluate a policy program is Difference-in-Difference. This method is well suited for benchmarking and finding improvements for optimization in organizations. It can, therefore, be used to design organizations so they generate more value for employees and customers. In this article, you learn how to do difference-in-difference in R.

**Methodology**

The difference in differences (DiD) method is a statistical technique or quasi-experimental design method, and it is used primarily in the social sciences and econometrics. In social science, it is sometimes called a “controlled before-and-after” study.

The DiD method involves comparing results from two groups, with data from each group being recorded over two time periods. One group (the control group) is not exposed to any treatment or intervention whatsoever; the other (treatment group) is exposed to a treatment or intervention before or during one of the two time periods. The same observations are made in both groups over each time period.

The data is analyzed by first calculating the difference in first and second time periods, and then subtracting the average gain (or difference) in the control group from the average gain (or difference) in the treatment group.

**Difference-in-Difference Estimator**

In order to make the DiD estimator lets consider the main assumption of DID:  
Counterfactual Levels for treated and non-treated can be different, but their time variation is similar.

Let us explain this with statistics:  
$$ E(Y0(t1) – Y0(t0) | D=1)= E( Y0(t1) – Y0(t0) | D=0) $$  
$$ [E(Y0(t1)|D=1) is counterf] $$

In the absence of treatment, change in treated outcome would have been as a change in non-treated outcome, i.e. changes in the economy or life-cycle, etc (unrelated to treatment) affect the two groups in a similar way.

This implies an relaxing assumption:  
$$ E(Y0 | X, D=0) = E( Y0 | X,D=1) $$  
$$ E( Y1 | X,D=0)= E( Y1 | X,D=1) $$

Selectivity bias is allowed even conditional on X, but only through an individual fixed effect (i.e. time constant).

Let us again consider the main assumption of DID:  
$$ E(Y0(t1) – Y0(t0) | D=1) = E( Y0(t1) – Y0(t0) | D=0) $$

It is possible to show how this assumption is used to generate a “control group” that can be substituted in for the missing counterfactual:

$$  
TTE = E( Y1(t1) – Y0(t1) | D=1)  
$$  
$$  
= E( Y1(t1) – Y0(t0) + Y0(t0) -Y0(t1) | D=1)  
$$  
$$  
= E( Y1(t1) – Y0(t0) | D=1) – E( Y0(t1) – Y0(t0) | D=1) [2. term unobs]  
$$  
$$  
= E( Y1(t1) – Y0(t0) | D=1) – E( Y0(t1) – Y0(t0) | D=0) [2. term obs]  
$$

Formally let us show that this can be written in a regression framework with individual fixed effects and time fixed effects:  
$$ Y(it) = a(t)+b\*D(it)+m(i)+u(it) $$  
Where a(t)=time effect, m(i)=individual effect

Then we can write the model as:  
$$ Y(i1) -Y(i0) = [a(1)-a(0)]+ b\*[D(i1)-D(i0)] +[u(i1)-u(i0)] $$

And the fixed-effect estimator reduces to:  
$$ b = E [Y(i1)-Y(i0)|D=1] – E [Y(i1)-Y(i0)|D=0] $$

Sample version of this is the DiD estimator, where assumption in this model is:  
$$ E [u(i1)-u(i0)|D=1] = E [u(i1)-u(i0)|D=0] $$

**Methodological Assumptions for Difference in Differences**

When doing a DiD method analysis we assume that the composition of the groups being studied are stable over the time period we are concerned about. We also assume there are no spillover effects, the amount of treatment or intervention given is not determined by the outcome, and that both groups being studied have parallel trends in their outcome—i.e., if no treatment was given, the difference between the data from the two groups would have a consistent difference over time.

The most important assumption in DiD is the parallel trends assumption. It is, therefore, necessary to be cautious about studies that do not graphically show these trends. If there is no convincing graph that shows the parallel trends in the pre-treatment outcomes for the treatment and control groups, be cautious. If the parallel trends assumption holds and we can credibly rule out any other time-variant changes that may confound the treatment, then DiD is a trustworthy method.

The difference in difference method is intuitive and fairly flexible; it will show a causal effect from observational data if the basic assumptions are met. Since it focuses on change, rather than the absolute levels, the groups being compared can start at different levels. Another key strong point to the DiD method is that it accounts for change due to factors other than the treatment or intervention being studied.

**Load dataset in R**

First we will load the dataset in R and do descriptive statistics:

data\_file = '../\_data/did\_data.dta'

if (!file.exists(data\_file)) {

download.file(url = 'https://drive.google.com/uc?authuser=0&id=0B0iAUHM7ljQ1cUZvRWxjUmpfVXM&export=download',

destfile = data\_file)

}

dataf = haven::read\_dta(data\_file)

skim(dataf, work, year, children) %>%

skimr::kable(digits = 0)

*Skim summary statistics*

*n obs: 13746*

*n variables: 11*

*Variable type: numeric*

*variable missing complete n mean sd p0 p25 p50 p75 p100 hist*

*---------- --------- ---------- ------- --------- ------ ------ ------ ------ ------ ------ ----------*

*## children 0 13746 13746 1.19 1.38 0 0 1 2 9 ▇▂▁▁▁▁▁▁*

*## work 0 13746 13746 0.51 0.5 0 0 1 1 1 ▇▁▁▁▁▁▁▇*

*## year 0 13746 13746 1993.35 1.7 1991 1992 1993 1995 1996 ▇▇▁▇▇▁▆▆*

Now let us construct dichotom variables for (a) before-and-after the EITC takes effect in 1994 and (b) the treatment group (1 or more children). We’ll keep these as logicals:

dataf = dataf %>%

mutate(post93 = year >= 1994, anykids = children >= 1)

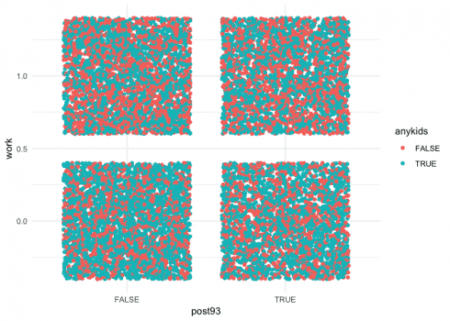
**Difference-in-difference regression in R**

The regression equation is:

work=β0+δ0post93+β1anykids+δ1(anykids×post93)+ε

The “difference-in-differences coefficient” is δ1, which indicates how the effect of kids changed after the EITC went into effect. Let’s take a second to plot this:

ggplot(dataf, aes(post93, work, color = anykids)) + geom\_jitter() + theme\_minimal()

The above coding gives us the following plot:  
[](https://i0.wp.com/datascienceplus.com/wp-content/uploads/2019/07/p1.png?ssl=1)

Let us also plot the variable means:

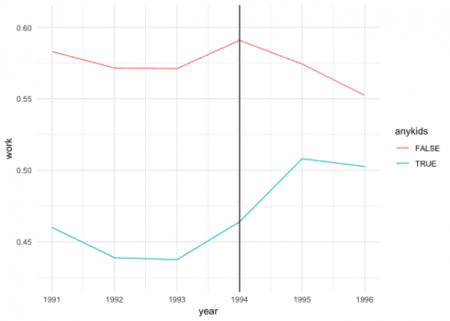
ggplot(dataf, aes(year, work, color = anykids)) +

stat\_summary(geom = 'line') +

geom\_vline(xintercept = 1994) +

theme\_minimal()

*No summary function supplied, defaulting to `mean\_se()*

The above coding gives us the following plot:  
[](https://i0.wp.com/datascienceplus.com/wp-content/uploads/2019/07/p2.png?ssl=1)

The parallel trends assumption looks good. However, parallel trends are nonlinear.  
Let us make the regression, which is created by using a linear probability model, where the DiD coefficient is the interaction coefficient δ1(anykids×post93):

model = lm(work ~ anykids\*post93, data = dataf)

summary(model)

*Call:*

*lm(formula = work ~ anykids \* post93, data = dataf)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-0.5755 -0.4908 0.4245 0.5092 0.5540*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 0.575460 0.008845 65.060 < 2e-16 \*\*\**

*anykidsTRUE -0.129498 0.011676 -11.091 < 2e-16 \*\*\**

*post93TRUE -0.002074 0.012931 -0.160 0.87261*

*anykidsTRUE:post93TRUE 0.046873 0.017158 2.732 0.00631 \*\**

*---*

*Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1*

*Residual standard error: 0.4967 on 13742 degrees of freedom*

*Multiple R-squared: 0.0126, Adjusted R-squared: 0.01238*

*F-statistic: 58.45 on 3 and 13742 DF, p-value: < 2.2e-16*

The above model shows that the EITC increased work by 5% among families with at least 1 child. In the second plot, the blue line goes from about 45% prior to 1994 to about 50%, afterward, due to the 5% increase in the model.