the Tarzan

[R] + applied economics.

About | ECNS 561 | Nuts'n Bolts | Resources

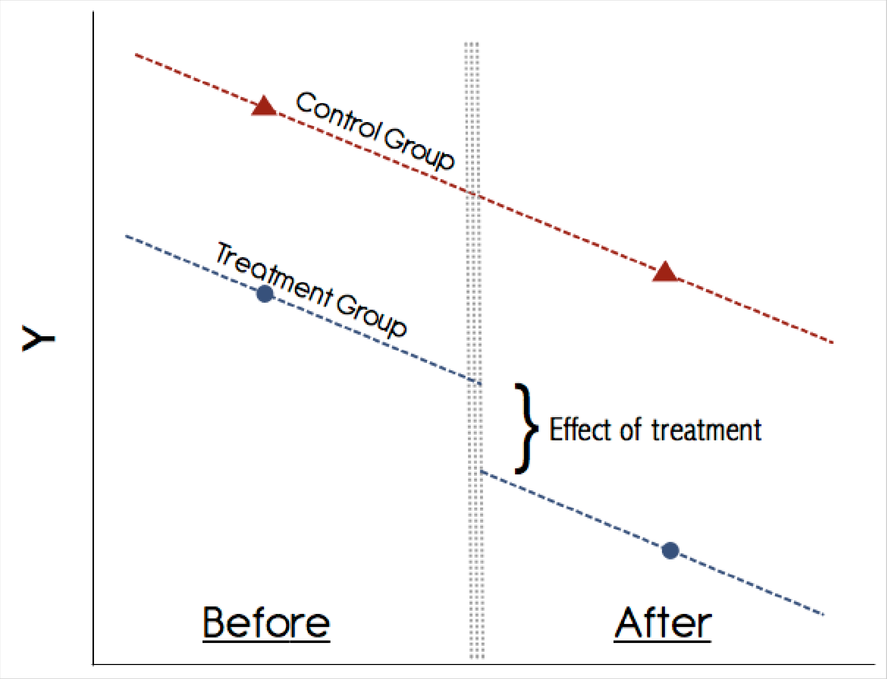
« TikZ for economists: A how-to guide with examples | TikZ diagrams with R: tikzDevice »

Differences-in-Differences estimation in R and Stata

{ a.k.a. Difference-in-Difference, Difference-in-Differences,DD, DID, D-I-D. }

DID estimation uses four data points to deduce the impact of a policy change or some other shock (a.k.a. treatment) on the treated population: *the effect of the treatment on the treated*. The structure of the experiment implies that the treatment group and control group have similar characteristics and are trending in the same way over time. This means that the counterfactual (unobserved scenario) is that had the treated group *not* received treatment, its mean value would be the same distance from the control group in the second period. See the diagram below; the four data points are the observed mean (average) of each group. These are the only data points necessary to calculate *the effect of the treatment on the treated*. The dotted lines represent the trend that is not observed by the researcher. Notice that although the means are different, they both have the same time trend (i.e. slope).

For a more thorough work through of the effect of the Earned Income Tax Credit on female employment, see an earlier post of mine:



Calculate the D-I-D Estimate of the Treatment Effect

We will now use R and Stata to calculate the unconditional difference-in-difference estimates of the effect of the 1993 EITC expansion on employment of single women.

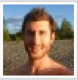
R:

```r
# Load the foreign package
require(foreign)

# Import data from web site

require(foreign)

# update: first download the file eitc.dta from this link:
# https://docs.google.com/open?id=0B0iAUHM7ljQ1cUZvRWxjUmpfVXM
# Then import from your hard drive:
eitc = read.dta("C:/link/to/my/download/folder/eitc.dta")

# Create two additional dummy variables to indicate before/after
# and treatment/control groups.

# the EITC went into effect in the year 1994
eitc$post93 = as.numeric(eitc$year >= 1994)

# The EITC only affects women with at least one child, so the
# treatment group will be all women with children.
eitc$anykids = as.numeric(eitc$children >= 1)

# Compute the four data points needed in the DID calculation:
a = sapply(subset(eitc, post93 == 0 & anykids == 0, select=work), mean)
b = sapply(subset(eitc, post93 == 0 & anykids == 1, select=work), mean)
c = sapply(subset(eitc, post93 == 1 & anykids == 0, select=work), mean)
d = sapply(subset(eitc, post93 == 1 & anykids == 1, select=work), mean)
```

Search this blog

Search...

Contributors



Kevin Goulding

Categories

- Econometrics
- Econometrics with R
- Numpy
- Python
- R tips & tricks
- Surviving Graduate Econometrics with R
- TikZ for Economists
- Visualizing Data with R
- White Papers

Twitter feed

RT @gappy3000: This post, apparently about #julialang and #pydata, explains why #rstats has become the standard of data analysis http:// ... 3 years ago

RT @justinwolfers: "if prediction markets are really as valuable as economists think, then..more experimentation could prove worthwhile. ... 3 years ago

RT @vsbuffalo: For me the biggest victory is for statistics and empiricism. Go Nate Silver and @fivethirtyeight for a brilliant forecast ... 3 years ago

Follow @baha_kev

Tag Cloud

cluster-robust Econometrics heteroskedasticity LaTeX Numpy Parallel Computing plots Python R STATA tex TikZ

```

# Compute the effect of the EITC on the employment of women with children:
(d-c)-(b-a)
```

The result is the width of the “shift” shown in the diagram above.

STATA:

```
cd "C:\DATA\Econ 562\homework"
use eitc, clear


gen anykids = (children >= 1)
gen post93 = (year >= 1994)


mean work if post93==0 & anykids==0     /* value 1 */
mean work if post93==0 & anykids==1     /* value 2 */
mean work if post93==1 & anykids==0     /* value 3 */
mean work if post93==1 & anykids==1     /* value 4 */
```

Then you must do the calculation by hand (shown on the last line of the R code).
(value 4 – value 3) – (value 2 – value 1)

Run a simple D-I-D Regression

Now we will run a regression to estimate the conditional difference-in-difference estimate of the effect of the Earned Income Tax Credit on “work”, using all women with children as the treatment group. This is exactly the same as what we did manually above, now using ordinary least squares. The regression equation is as follows:

$$work = \beta_0 + \delta_0 post93 + \beta_1 anykids + \delta_1 (anykids \times post93) + \varepsilon$$

Where ε is the white noise error term, and δ_1 is the effect of the treatment on the treated — the shift shown in the diagram. To be clear, the coefficient on $(anykids \times post93)$ is the value we are interested in (i.e., δ_1).

R:

```
eitc$p93kids.interaction = eitc$post93*eitc$anykids
reg1 = lm(work ~ post93 + anykids + p93kids.interaction, data = eitc)
summary(reg1)
```

The coefficient estimate on `p93kids.interaction` should match the value calculated manually above.

STATA:

```
gen interaction = post93*anykids
reg work post93 anykids interaction
```

Share this:

Share

★ Like

6 bloggers like this.

Related

Surviving Graduate Econometrics with R: Difference-in-Differences Estimation -- 2 of 8
In "Surviving Graduate Econometrics with R"

Surviving Graduate Econometrics with R: Fixed Effects Estimation -- 3 of 8
In "Surviving Graduate Econometrics with R"

Surviving Graduate Econometrics with R: Advanced Panel Data Methods -- 4 of 8
In "Surviving Graduate Econometrics with R"

Posted on June 20, 2011 at 2:21 pm in Econometrics with R | RSS feed | Reply | Trackback URL

Tags: R, STATA

46 Responses to “Differences-in-Differences estimation in R and Stata”

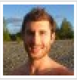
Claudio Gxp
November 6, 2011 at 11:10 am

Dear Kevin
Thanks for this example, but be aware of some mistakes:

in Stata:
“use eitc, clear” instead of “insheet using eitc.dta, clear”

in R:
eitc$post93 = as.numeric(eitc$year >= 1994)
eitc$anykids = as.numeric(eitc$children >= 1)
instead of
eitc$post93 = as.numeric(year >= 1994)
eitc$anykids = as.numeric(children >= 0)
Best
Claudio

Reply



Kevin Goulding
November 6, 2011 at 11:22 am

Hi Claudio,

Good catch; I’ve updated the or reading. -Kevin

Reply

Follow

Follow “the Tarzan”

Get every new post delivered to your Inbox.

Join 78 other followers

Enter your email address

Sign me up

Build a website with WordPress.com

Dan M
January 6, 2012 at 1:34 am

Hi Kevin –

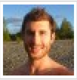
I have a question for you that to. I’m trying to run a DID model in stata, only I’m first differencing e time-variant control variables, but not all of them. In particular, my p erenced (because I assume it to have effects over multiple years, not plication is that not all program participants enter the program ticipation occurs, for some in 2003, some in 2004, some in 2005, have a pre and post dummy variable to interact with the treatment / control variable.

How do I carry this out? My initial impression, and after reading some math for the past 6 hours, is that when I first difference, I simply am left with the program participation variable that indicates a 1 only after the program is initiated, and that there is no more interaction term. Is this correct? This is as opposed to if I did not difference, then I would need the pre/post variable, the treatment/control variable, and the interaction of the two, which represents the affect of the treatment on the treated.

One last question / clarification. Assuming I’m correct that I simply have a variable in the regression that becomes a 1 for the treatment once treatment begins… and I don’t difference it, but I do difference the outcome variable… is there anything wrong with this?

Any help / opinion you can offer would be appreciated.

Reply



Kevin Goulding
July 14, 2012 at 9:06 am

Hi Dan, what is your motivation for first differencing your data? It might be less complicated to not first-difference your data but instead include a set of individual specific dummy variables (i.e. fixed effects) which will accomplish the same thing (control for between individual differences a.k.a. time invariant differences). From there you should be able to follow the model as above.

It is perfectly fine to have the treatment dummy switch on (=1) at different time for different individuals. Interpretation is the same. Hope this helps. -Kevin

Reply

Andre
May 28, 2015 at 12:53 pm

Awesome thanks for sharing this.

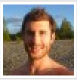
Karen
May 23, 2012 at 5:16 am

Hi Kevin,

thanks for this example, really useful resource. Could you please also give STATA command if outcome variable is binary.

thanks

Reply



Kevin Goulding
July 14, 2012 at 8:28 am

Hi Karen, If the outcome variable is binary, you do not want to use this OLS model because of the inherent heteroskedasticity in a Linear Probability Model (it’s inefficient). Also, interpretation of coefficients becomes difficult when predicted values lie outside [0,1].

If you can clearly state your research question I may be able to point you in the right direction — in the meantime check out log it and probit models. HTH. -Kevin

Reply

pythonscript
August 10, 2012 at 4:41 pm

Where can I find the data set you used? The link in your example is no longer functioning.

Reply

Kevin Goulding
August 10, 2012 at 5:14 pm

I've now updated the code to show how to find the eitc.dta data set. It is now available at this link:

https://docs.google.com/open?id=0B0iAUHM7ljQ1cUZvRWxjUmpfVXM

Let me know if this works. Cheers, Kevin

Reply

Raul
January 17, 2015 at 1:22 pm

Hello Kevin,
I am unable to obtain the EITC data set you used in your example. The link seems not to be functioning any longer.
Could you please provide another link
Thanks.
-Raul

John
August 10, 2012 at 9:18 pm

Works perfectly. Thank you for the quick response!

Reply

Raul
January 17, 2015 at 1:30 pm

I have just noticed that the link actually works. Sorry!

Reply

René Zacker
December 4, 2012 at 8:32 am

Hi Kevin, I write to you to consult you about a question I have. I estimated a differences-in-differences model as follows: (t = a + b1treatment b2dt + b3 + (treatment * dt) + e) plus state x year fixed effects. This specification use it with different dependent variables such as employment, productivity and sales

The results I get to b3 are not significant. However, if I include firm fixed effects coefficients are all significant. I have doubts if this latter specification picking another effect different from the original idea of DID.

I appreciate the help that I can provide.

René.

Reply

ahgecker
December 28, 2012 at 1:59 pm

THANK YOU. It's 9 months since I turned in my thesis, and I couldn't for the life of me remember how to do this. I'm pretty sure it's a stress-induced mental block haha. Anyway, this was super clear and helpful — much appreciated!!

Reply

estatistics
January 16, 2013 at 3:20 am

I will like to run a regression diff-in-diff. Variables: a household has a child(=0) or not (1) Household must be predicted from predictors. two years 2008 and 2010. education (7 levels) and treatment (received social support=1, they didnt receive social support=0). So, i would like to run two regressions, Household(i,1)=b0+b2*educ(i)+a(i)+m(i,1), the second: Household(i,1)=b0+b1*treatment(i,2)+b2*educ(i)+d(treatment)(i,2)*educ(i)+a(i)+m(i,2), then i must substract first model from second one…. DHousehold(i)=a+b1(D)treatment(i)+dDtreatment(i)*educ(i)+Dm(i). . I have problem because Household and treatment are 0/1 vars, and educ is ordinal (none to high educ). I must use reg in STATA? or logit? And how can i calc effects? what effects i have? fixed? the treatment group? Can you tell me some calcs on stata in order to understand better my problem?

Reply

estatistics
January 16, 2013 at 3:28 am

edit: Household(i,2)=b0+b1*treatment(i,2)+b2*educ(i)+d(treatment)(i,2)*educ(i)+a(i)+m(i,2)
Household(i,1)=2008, Household(i,2)=2010

Reply

Dexter
February 12, 2013 at 10:28 am

What if your dependent variable is continuous? What if all of your variables are continuous? Does that mean in the 'manual' / sapply computation (lines 23 – 27) a threshold needs to be chosen? In certain cases that threshold would be obvious, like zero, but in other instances it is unclear. I imagine thresholds could come from the literature, from theory, or in a data driven approach like mean, median, or breaking the data up by quantile…. can you please clarify what to do in these instances?

Reply

Kevin Goulding
February 13, 2013 at 10:16 am

Dexter — in the example above, the dependent variable is continuous. It was also designed to be simple for illustration. If you are wanting to look at *incremental* impacts of continuous variables on the outcome of interest, you will have to think hard about precisely what you are trying to measure so that you can set up your model properly. I think a good place to start would be to search on "regression and causality" or "modeling incremental impacts". HTH-

Reply

Khamphong
December 11, 2013 at 3:08 am

Dear Kevin,
I do clearly understand how to do DID by hand as you mentioned:

```
gen anykids = (children >= 1)
gen post93 = (year >= 1994)

mean work if post93==0 & anykids==0 /* value 1 */
mean work if post93==0 & anykids==1 /* value 2 */
mean work if post93==1 & anykids==0 /* value 3 */
mean work if post93==1 & anykids==1 /* value 4 */
```

However, what should I do if need Standard Error of each mean or Standard Error value 1-4?
Thank you so much
Khamphong

Rumayya
March 4, 2013 at 6:48 pm

Hi Kevin, I have a question. Could i use perceived change of Y variable as proxy for DID method? I have two set of survey data in different year (2000 and 2007) but unfortunately my main variable of interest, the dependent variable is only asked in later survey. The variable come from these questions: 1)how is the level of Y in this year (2007); 2)how has the Y changed compared with 2000 (increasing, the same, worse). Could I use the second as proxy for delta Y in DID model? what are the statistical implication (bias?) and how i could overcome it? thanks in advance. -Rumayya-

Reply

Kevin Goulding
March 4, 2013 at 6:54 pm

Rumayya — thanks for reading my blog. I think you have an insurmountable problem with your data set because of the missing outcome in year=2000. This fundamentally leaves a gaping hole in your experimental design. It might be better to simply use a cross sectional approach, although it will be difficult to show causality.

Reply

eda
March 7, 2013 at 7:22 am

Hi Kevin,
Can u help me understand this expression 'standard errors were bootstrapped using 5000 simulations'?
thanks in advance

Reply

Kevin Goulding
March 7, 2013 at 8:52 am

Hi eda — bootstrapping is one method used to correct bias in standard errors and/or an estimator by resampling the original data many times — in your example above, 5000 times. A search on 'bootstapped standard errors' or the R package on CRAN called 'boot' is a good place to start investigating this method. HTH

Reply

eda
March 31, 2013 at 6:15 am

Hi again Kevin, thanks for the reply. I have another question….i see that together with DID, a study uses also propensity score matching…do u have any idea how is that done?

Reply

Kevin Goulding
March 31, 2013 at 10:23 am

Hi Eda — propensity score matching is a quasi-experimental approach that attempts to control for selection bias into the treatment group. Typically, this approach is used when the researcher cannot randomly assign treatment herself and therefore relies on observational data. While PSM is out of scope for the blog post above, it is a related topic and perhaps I'll add it to my list for a future post. In the meantime, check out the CRAN package `Matching`. HTH, Kevin

Reply

Dan Matisoff
March 31, 2013 at 6:59 pm

Reply

Donald
April 8, 2013 at 11:37 am

Hello,

Am having trouble with the syntax for DID, i have baseline data and follow up data and i want to assess the impact of a project. how do i go about it?

Reply

Tierney Leddy
May 7, 2013 at 6:53 pm

Neither of the links for the data are working- wondering if you could post a link for the stata data again?

Thanks, I'm trying to figure out where the coefficient estimate on 'p93kids.internaction' would be in the stata regression- thanks!!

Reply

Kevin Goulding
May 7, 2013 at 7:03 pm

Very sorry about the broken links. I won't be able to change them in the near term. Please try googling, you might just get lucky.

Reply

Luciana Chalela
June 1, 2013 at 6:19 pm

I am running diff-in-diff in stata. Do you know how can I produce good graphs to illustrate the net treatment effect? Something like the first graph you have in this post…Thank you

Reply

Yuri
October 8, 2013 at 9:35 am

Hi Kevin,
Great post!
I'm running diff-in-diff with propensityscore in stata and I'd like to know if the casual effect of my model will be exactly the interaction coefficient. Do you know anything about it?

Reply

swangila
November 16, 2013 at 5:40 pm

so in terms of interpretation what does it imply if the coefficient of the DiD term is negative…does it imply that the effect of the program on the treated was negative or is it that it was positive but was decelerating over time?

Reply

Kevin Goulding
November 16, 2013 at 5:43 pm

Swangila — assuming the model is properly specified, a negative and statistically significant coefficient on the DID term implies a negative effect.

Reply

reajul
November 20, 2013 at 5:32 pm

hi Kevin
It has been a great help!
Thanks a lot…

Reply

sweetwilliam
February 20, 2014 at 6:33 pm

hi Kenvin
Could u please show me how to Test whether the coefficient (delta1) is
statistically significant at a given alpha

Reply

Chaima Ben Abderrahmen
April 6, 2014 at 2:21 am

Hi Kevin, please i need to know how to use to run a generalized diff-in-diff approach under stata

Reply

GJ
April 9, 2014 at 10:43 pm

```
# Compute the four data points needed in the DID calculation:
a = sapply(subset(eitc, post93 == 0 & anykids == 0, select=work), mean)
b = sapply(subset(eitc, post93 == 0 & anykids == 1, select=work), mean)
c = sapply(subset(eitc, post93 == 1 & anykids == 0, select=work), mean)
d = sapply(subset(eitc, post93 == 1 & anykids == 1, select=work), mean)
```

For this part of the code, what does "select=work" do?

Thanks for your time.

GJ

Reply

Dimi
July 8, 2014 at 10:47 am

How can I estimate the differences-in-differences using SPSS?

Reply

Ica
August 14, 2014 at 1:57 pm

Hey Kevin, I am facing with my first time coding in STATA, and I am completely confused. I have to interact a lot of variables and I do not have a clue how to fo it. I am examining the influence of smoking on wages, and I wanted to use DID method. But the problem is that I have to examine separately for female and male. Therefore I have to write the codes for dummies for wages of females who smoke year 2001- wages of females who smoke year 2000, and wages of females who do not smoke year 2001-wages of females who do not smoke year 2000. Please HELP .. :)))

Reply

Nicholas Zaleski
August 21, 2014 at 6:59 am

This post is old, but hopefully I get a response. In the beginning you did a manual version of the D-D estimator. I am no expert, but I didn't think that the mean at the beginning and end of a sample, was statistically the same as the beginning and end of a regression line. Kevin, if you still receive notifications could you please answer, as to this question pertains to a project I am currently working on

Reply

n. a.
November 20, 2014 at 8:30 am

I'm not Kevin, but:

@GJ — "select = work" is used to select the "work" variable from the dataset. Try it out yourself by loading the dataset and trying "subset(eitc, post93 == 0 & anykids == 0, select=nonwhite), or any other variable. Kevin does this because he wants to take the average of the "work" variable for the 4 different groups.

@Ica – I don't use stata, but how about you use stata (or excel) to seperate your dataset into two datasets: female and male? Then the code Kevin wrote should apply.

@Nicholas — the dependant variables, "post" and "anykids" only take the values 0 and 1. So, there is no regression "line". Try writing out the formula for a regression and you'll see that in the case of only 0 and 1 values, it simplifies to kevin's statement. (Also verified by the fact that using a "regression" returns the same values!

Reply

Sumit
February 19, 2015 at 2:30 am

Dear Kevin/n.a.,

Thank you very much for your blog. I have a question regarding stata command "diff" (please "help diff" in stata) for the pooled/repeated cross section data (whether arguments for this command change for this data). I will write my code here and please confirm if I get it right.

I have data for two countries (lets say, countryA and countryB) for 5 years from 2001 until 2005. The treatment happens to countryB in the year 2002. So the stata commands I am using for the difference-in-difference estimator are:

*code starts here
period=1 if year>=2002
period=0 if year<2002

treated=1 if country=="countryB"
treated=0 if country=="countryA"

diff outcomevar, period(period) treated(treated) cov(a,b,c,d)

I am aware that this coding is pretty much a standard for this command. However, I would like to confirm if it prevails for the repeated/pooled cross-sectional data, and also, I would like to confirm if I got everything right.

Thanks and regards,
Sumit

Reply

Genet
March 24, 2015 at 11:15 am

How can I estimate the differences-in-differences using R?

Reply

agusjay55
August 22, 2015 at 11:08 pm

If I use simple regression for DID, what criteria do I have to reject null hypothesis? The p-value of "interaction"?

Reply

jashem
September 8, 2015 at 1:48 am

dear kevin,
thanks for your blog, i have a question. for you. How to set up data for running the diff in diff?can you give an example? i am running a simple DID Estimation with two periods, the before 1993 and after 2003. Is this acceptable estimation?

Reply

Trackbacks

Yes, they're young and inexperienced. But Teach First participants have the right stuff | IOE London blog
September 5, 2013 at 3:06 am

Leave a Reply

Enter your comment here...

Tags

cluster-robust econometrics heteroskedasticity latex numpy parallel computing plots python r stata tex tikz

Calendar

June 2011

M	T	W	T	F	S	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			

« May Jul »

Archives

- October 2012
- February 2012
- July 2011
- June 2011
- May 2011

Blogroll

- Documentation
- Plugins
- Suggest Ideas
- Support Forum
- Themes
- WordPress Blog
- WordPress Planet