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

The following replication exercise closely follows the homework assignment #2 in ECNS 562. The data for this exercise can be found here.

The data is about the expansion of the Earned Income Tax Credit. This is a legislation aimed at providing a tax break for low income individuals. For some background on the subject, see

> Eissa, Nada, and Jeffrey B. Liebman. 1996. Labor Supply Responses to the Earned Income Tax Credit. Quarterly Journal of Economics. **111**(2): 605-637.

The homework questions (abbreviated):

1. Describe and summarize data.
2. Calculate the sample means of all variables for (a) single women with no children, (b) single women with 1 child, and (c) single women with 2+ children.
3. Create a new variable with earnings conditional on working (missing for non-employed) and calculate the means of this by group as well.
4. Construct a variable for the "treatment" called ANYKIDS and a variable for after the expansion (called POST93—should be 1 for 1994 and later).
5. Create a graph which plots mean annual employment rates by year (1991-1996) for single women with children (treatment) and without children (control).
6. Calculate the unconditional difference-in-difference estimates of the effect of the 1993 EITC expansion on employment of single women.
7. Now run a regression to estimate the conditional difference-in-difference estimate of the effect of the EITC. Use all women with children as the treatment group.
8. Reestimate this model including demographic characteristics.
9. Add the state unemployment rate and allow its effect to vary by the presence of children.
10. Allow the treatment effect to vary by those with 1 or 2+ children.
11. Estimate a "placebo" treatment model. Take data from only the pre-reform period. Use the same treatment and control groups. Introduce a placebo policy that begins in 1992 (so 1992 and 1993 both have this fake policy).

A review: Loading your data

Recall the code for importing your data:

STATA:

```
/*Last modified 1/11/2011 */

*********************************************************************
*The following block of commands go at the start of nearly all do files*/
*Bracket comments with /* */ or just use an asterisk at line beginning

clear                                /*Clears memory*/
set mem 50m                          /*Adjust this for your particular dataset*/
cd "C:\DATA\Econ 562\homework"        /*Change this for your file structure*/
log using stata_assign2.log, replace  /*Log file records all commands & results*/
display "$S_DATE $S_TIME"
set more off
insheet using eitc.dta, clear
*********************************************************************
```

R:

```
# Kevin Goulding
# ECNS 562 - Assignment 2

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

# Import data from web site
# 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")</pre>
Note that any comments can be embedded into R code, simply by putting a <code> # </code> to the left of your co

1eitc = read.dta("C:\DATA\Courses\Econ 562\homework\eitc.dta")
```

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

Describe and summarize your data

Recall from part 1 of this series, the following code to describe and summarize your data:

STATA:

```
des
sum
```

R:

In R, each column of your data is assigned a class which will determine how your data is treated in various functions. To see what class R has interpreted for all your variables, run the following code:

```
sapply(eitc,class)
summary(eitc)
source('sumstats.r')
sumstats(eitc)
```

To output the summary statistics table to LaTeX, use the following code:

```
require(xtable)                  # xtable package helps create LaTeX code from R.
xtable(sumstats(eitc))
```

Note: You will need to re-run the code for `sumstats()` which you can find in an earlier post.

Calculate Conditional Sample Means

STATA:

```
summarize if children==0
summarize if children == 1
summarize if children >=1
summarize if children >=1 & year == 1994


mean work if post93 == 0 & anykids == 1
```

R:

```
# The following code utilizes the sumstats function (you will need to re-run this code)
sumstats(eitc[eitc$children == 0, ])
sumstats(eitc[eitc$children == 1, ])
sumstats(eitc[eitc$children >= 1, ])
sumstats(eitc[eitc$children >= 1 & eitc$year == 1994, ])

# Alternately, you can use the built-in summary function
summary(eitc[eitc$children == 0, ])
summary(eitc[eitc$children == 1, ])
summary(eitc[eitc$children >= 1, ])
summary(eitc[eitc$children >= 1 & eitc$year == 1994, ])

# Another example: Summarize variable 'work' for women with one child from 1993 onwards.
summary(subset(eitc, year >= 1993 & children == 1, select=work))
```

The code above includes all summary statistics – but say you are only interested in the mean. You could then be more specific in your coding, like this:

```
mean(eitc[eitc$children == 0, 'work'])
mean(eitc[eitc$children == 1, 'work'])
mean(eitc[eitc$children >= 1, 'work'])
```

Try out any of the other headings within the summary output, they should also work: `min()` for minimum value, `max()` for maximum value, `stdev()` for standard deviation, and others.

Create a New Variable

To create a new variable called "c.earn" equal to earnings conditional on working (if "work" = 1), "NA" otherwise ("work" = 0) – use the following code:

STATA:

```
gen cearn = earn if work == 1
```

R:

```
eitc$c.earn=eitc$earn*eitc$work
z = names(eitc)
X = as.data.frame(eitc$c.earn)
X[] = lapply(X, function(x){replace(x, x == 0, NA)})
eitc = cbind(eitc,X)
eitc$c.earn = NULL
names(eitc) = z
```

Construct a Treatment Variable

Construct a variable for the treatment called "anykids" = 1 for treated individual (has at least one child); and a variable for after the expansion called "post93" = 1 for 1994 and later.

STATA:

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

R:

```
eitc$post93 = as.numeric(eitc$year >= 1994)
eitc$anykids = as.numeric(eitc$children > 0)
```

Create a plot

Create a graph which plots mean annual employment rates by year (1991-1996) for single women with children (treatment) and without children (control).

STATA:

```
preserve
collapse work, by(year anykids)
gen work0 = work if anykids==0
label var work0 "Single women, no children"
gen work1 = work if anykids==1
label var work1 "Single women, children"
twoway (line work0 year, sort) (line work1 year, sort), ytitle(Labor Force Participation Rates)
graph save Graph "homework\eitc1.gph", replace
```

R:

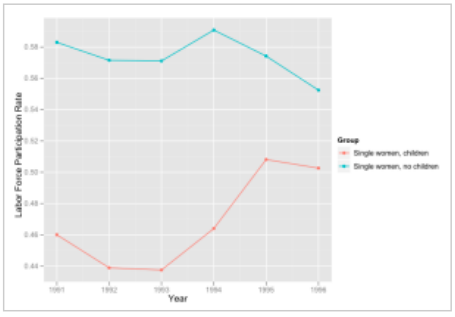
```
# Take average value of 'work' by year, conditional on anykids
minfo = aggregate(eitc$work, list(eitc$year,eitc$anykids == 1), mean)

# rename column headings (variables)
names(minfo) = c("YR","Treatment","LFPR")

# Attach a new column with labels
minfo$Group[1:6] = "Single women, no children"
minfo$Group[7:12] = "Single women, children"
minfo

require(ggplot2)    #package for creating nice plots

qplot(YR, LFPR, data=minfo, geom=c("point","line"), colour=Group,
        xlab="Year", ylab="Labor Force Participation Rate")
```



The ggplot2 package produces some nice looking charts.

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

Calculate the unconditional difference-in-difference estimates of the effect of the 1993 EITC expansion on employment of single women.

STATA:

```
mean work if post93==0 & anykids==0
mean work if post93==0 & anykids==1
mean work if post93==1 & anykids==0
mean work if post93==1 & anykids==1
```

R:

```
a = colMeans(subset(eitc, post93 == 0 & anykids == 0, select=work))
b = colMeans(subset(eitc, post93 == 0 & anykids == 1, select=work))
c = colMeans(subset(eitc, post93 == 1 & anykids == 0, select=work))
d = colMeans(subset(eitc, post93 == 1 & anykids == 1, select=work))
(d-c)-(b-a)
```

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. 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.

STATA:

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

R:

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

Include Relevant Demographics in Regression

Adding additional variables is a matter of including them in your coded regression equation, as follows:

STATA:

```
gen age2 = age^2          /*Create age-squared variable*/
gen nonlaborinc = finc - earn     /*Non-labor income*/


reg work post93 anykids interaction nonwhite age age2 ed finc nonlaborinc
```

R:

```
reg2 = lm(work ~ anykids + post93 + post93*anykids + nonwhite
                 + age + I(age^2) + ed + finc + I(finc-earn), data = eitc)
summary(reg2)
```

Create some new variables

We will create two new interaction variables:

1. The state unemployment rate interacted with number of children.
2. The treatment term interacted with individuals with one child, or more than one child.

STATA:

```
gen interu = urate*anykids


gen onekid = (children==1)
gen twokid = (children>=2)
gen postXone = post93*onekid
gen postXtwo = post93*twokid
```

R:

```
# The state unemployment rate interacted with number of children
eitc$urate.int = eitc$urate*eitc$anykids

##
# Creating a new treatment term:

# First, we'll create a new dummy variable to distinguish between one child and 2+.
eitc$manykids = as.numeric(eitc$children >= 2)

# Next, we'll create a new variable by interacting the new dummy
# variable with the original interaction term.
eitc$tr2 = eitc$p93kids.interaction*eitc$manykids
```

Estimate a Placebo Model

Testing a placebo model is when you arbitrarily choose a treatment time before your actual treatment time, and test to see if you get a significant treatment effect.

STATA:

```
gen placebo = (year >= 1992)
gen placeboXany = anykids*placebo


reg work anykids placebo placeboXany if year<1994
```

In R, first we'll subset the data to exclude the time period after the real treatment (1993 and later). Next, we'll create a new treatment dummy variable, and run a regression as before on our data subset.

R:

```
# sub set the data, including only years before 1994.
eitc.sub = eitc[eitc$year <= 1993,]

# Create a new "after treatment" dummy variable
# and interaction term
eitc.sub$post91 = as.numeric(eitc.sub$year >= 1992)

# Run a placebo regression where placebo treatment = post91*anykids
reg3 <- lm(work ~ anykids + post91 + post91*anykids, data = eitc.sub)
summary(reg3)
```

The entire code for this post is available here (File –> Save As). If you have any questions or find problems with my code, you can e-mail me directly at **kevingoulding {at} gmail [dot] com**.

To continue on to Part 3 of our series, Fixed Effects estimation, click here.

Related

Differences-in-Differences estimation in R and Stata
In "Econometrics with R"

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

Calculate OLS regression manually using matrix algebra in R
In "Econometrics with R"

Posted on May 24, 2011 at 6:00 pm in Surviving Graduate Econometrics with R | RSS feed | Reply | Trackback URL

Tags: R, STATA

19 Comments to "Surviving Graduate Econometrics with R: Difference-in-Differences Estimation — 2 of 8"

Tony
March 11, 2012 at 3:01 pm

I have one suggestion on doing the diff-in-diff regression in R. You could use the built-in functionality of R for interactions instead of making your own.

reg1 = lm(work~anykids*post93, data = eitc)

is enough to estimate exactly what your reg1 does. Thanks for posting the code.

Reply

Kevin Goulding
March 11, 2012 at 4:25 pm

Thanks Tony, for the suggestion. It's now updated. Thanks for reading! –Kevin

Reply

Jonathan
March 19, 2012 at 5:01 pm

Any good ways of exporting the dif-in-diff into excel? It would be very helpful to know how (save me hours of hand calculations)

Reply

Kevin Goulding
March 19, 2012 at 5:17 pm

Jonathan – I'm not sure exactly what you mean by 'export'. You can export any data.frame from R using the command "write.csv(df3, 'df3.csv')". Hope this helps-

Reply

Jonathan
March 19, 2012 at 5:38 pm

Hey Kevin you might be confused because I was being incredibly vague. Essentially I am trying to find a package in STATA or R that exports the marginal output from a difference in difference estimation into Latex or excel. So that the output would be a table of the means for the four periods, and the differences. I can write my own program to do this in STATA (not as proficient in R). However if you know a program that already does this it would be extremely helpful. I know there are several individuals who have asked this before on other forums without any luck (for R and STATA). I can provide an example if I am still being vague.

Kevin Goulding
March 19, 2012 at 6:35 pm

Thanks for clarifying. I do not know of a package to do this, but it shouldn't be too hard to code it in R. Try this (using the data / example from this post):

create a table:

```
agg = aggregate(eitc$work,list(Time = eitc[,"post93"] > 0,
Treatment = eitc[,"anykids"] > 0),mean)
require(reshape)
tb1 = data.frame(cast(agg, Treatment ~ Time))
names(tb1) = c('Treatment','Before','After')
tb1$diff = tb1[,2]-tb1[,3]
tb1$diff.in.diff = c(NA,tb1[1,4]-tb1[2,4])
tb1$Treatment = as.character(tb1$Treatment)
tb1
```

print to LaTex:

```
require(xtable)
print(xtable(tb1),include.rownames=FALSE)
```

print to csv for use in Excel:

```
write.csv(tb1,'tb1.csv')
```

Jonathan
March 20, 2012 at 9:50 am

Thanks Kevin I will try this. If it works you will save me hours of work!

Kevin Goulding
July 14, 2012 at 8:46 am

Hi Jonathon, check out the 'xtable' package. This will print the latex code of any R table (including regression results). I've used this approach extensively to drop R results into my thesis.

lilian
July 11, 2012 at 7:46 am

was not able to assess the data link. need it to understand. could it be reloaded

Reply

Kevin Goulding
August 11, 2012 at 9:09 am

Hi Lilian, the data is now available for download at the following link:

https://docs.google.com/open?id=0B0iAUHM7IjQ1cUZvRWxjUmpfVXM

I've also updated the code above. Cheers, Kevin

Reply

Manuel S Gonzalez Canche
November 29, 2012 at 1:02 pm

Great post Kevin!!!
I have a small suggestion to create a new variable cearn conditional on working you can do have a double subset function as follows:
#Create a place holder for cearn
eitc$cearn<-NA
#Add the double condition
eitc[eitc$work==1,]$cearn <- eitc[eitc$work==1,]$earn

Thank you for sharing your amazing work!!!

Reply

Juno
March 8, 2013 at 5:30 pm

Thank you, Kevin!
Your code is very helpful and nice.
By the way, I have a questions about the model, especially for the dependent variable "work."
Can we run just a simple linear regression even if the dependent variable is a 0 and 1 binary variable?
If it is fine, then I am okay.
But, if it is not okay, should we use a logistic regression?
Then, another question will come out:
can we put an interaction term (post93*anykids) in the logistic regression even if the interpretation about the estimated coefficient of the interaction term depends on other control variables (covariates)?

I hope to hear your answers to my questions.
Thank you in advance!!

– Juno

Reply

Kevin Goulding
March 8, 2013 at 8:08 pm

The downside to using a linear model when the dependent variable is binary is that (1) there is inherent heteroskedasticity, and (2) predictions can fall outside the range [0,1]. Search on "linear probability model". The upside is that the coefficients are easy to interpret. I suggest using a logistic model. This would take care of the heteroskedasticity issue; however, you then need to be careful to interpret the coefficients properly. A logistic model can handle dummy variables and interaction variables just like a linear model can.
HTH, Kevin

Reply

James
April 22, 2014 at 4:19 pm

I don't know if you still check this site but, i had to use DID estimation on Stata for my dissertation – and while i study advanced econometrics this was painful to run through on STATA – this and following pages truly helped. thanks for everything!

Reply

Kevin Goulding
April 23, 2014 at 12:32 pm

Hi James — I don't do much on here lately, but your response put a smile on my face. You are very welcome and glad it helped! The idea was to hopefully allow others to avoid some of the pain I went through to get up and running in R. Cheers-

Reply

Xiomara
October 3, 2014 at 9:42 pm

Hi! I could have sworn I've been to this website before but after checking through some of the post I realized it's new to me. Nonetheless, I'm definitely happy I found it and I'll be bookmarking and checking back often!

Reply

Praveen rawat
June 17, 2015 at 11:11 pm

Hi Kevein,
I did exactly what you have suggested for estimating diff-in-diff, but it is not giving me the co-efficents, std error, sig. etc for interaction term (like post93*anykids in your case)

It shows this line
Coefficients: (1 not defined because of singularities)

Please help me in tackling this. Since this is the diff-in-diff estimator. the whole motive behind doing diff-in-diff.

Reply

Kevin Goulding
June 19, 2015 at 4:02 pm

Hi Praveen — Thanks for reading my blog. Normally when you get an error referring to singularities, it is due to the data set itself and means that one or more of your variables are collinear. Recall that in order for OLS to work (see: Gauss-Markov assumptions), the matrix has to be full rank. Look at the data set, and ensure that this is correct; hopefully, you'll identify the problem there. Hope this helps-

Kevin

Reply

Leah
July 28, 2015 at 4:57 pm

Thanks so much, this page was extremely helpful!!

Reply

Leave a Reply

Enter your comment here...

Tags

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

Calendar

May 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	31					

Jun »

Archives

October 2012
February 2012
July 2011
June 2011
May 2011

Blogroll

Documentation
Plugins
Suggest Ideas
Support Forum
Themes
WordPress Blog
WordPress Planet