**2023 Causal Inference Workshop: Stata, R, and Python Sessions**

**Day 3. Difference-in-differences**

Dataset: Causal Inference Workshops (All Years)\Stata and R Materials\Cheng-Hoekstra-castle-doctrine-simplified.dta

Underlying article: Cheng Cheng and Mark Hoekstra, *Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Expansions to Castle Doctrine*, **Journal of Human Resources** 48(3), 821-853 (2013).

**Answer Sheet**

**Question 1**. Estimate the average effect of castle doctrine law adoption on the 14 states that passed these laws in 2006 (ATT) using the treatment variable *treat* and the outcome variable log(homicides) in a difference-in-differences equation using years 2006-2010 as the treatment period (post = 1) and 2000-2005 as the pre-treatment period (post = 0). You will not be using covariates in this analysis.

**STATA Code for Step 1**:

**\*Generate log(homicide)**

gen homicide\_log=log(homicide)

**\*Drop all treated states except those that adopted castle doctrine laws in 2006**

drop if treatment\_date==2005 | treatment\_date==2007 | treatment\_date==2008 | treatment\_date==2009 | treatment\_date==2010

**\*Generate post treatment indicator**

gen post = 0

replace post = 1 if year>=2006

**\*Generate a dummy variable called treat = 0 if never-treated and 1 if treated in 2006.**

gen treat=0

replace treat=1 if treatment\_date==2006

**\*Estimate treatment effect, with and without population weights. Without weights:**

reg homicide\_log i.post i.treat i.post#i.treat, cluster(state)

**\*Alternative with population weights (more like Cheng and Hoekstra, but not part of the assignment)**

reg homicide\_log i.post i.treat i.post#i.treat [aweight=population], cluster(state) // more like Cheng and Hoekstra with population weights.

**\*Manual estimation of averages and differences**

preserve

**\*Generate averages for treated**

egen y11 = mean(homicide\_log) if post==1 & treat==1

egen y10 = mean(homicide\_log) if post==0 & treat==1

**\* Next steps remove the missing values of y11 and y10 (for the control states)**

egen ey11 = max(y11)

egen ey10 = max(y10)

**\*Generate averages for control**

egen y01 = mean(homicide\_log) if post==1 & treat==0

egen y00 = mean(homicide\_log) if post==0 & treat==0

**\* Next steps remove the missing values of y11 and y10 (for the treated states)**

egen ey01 = max(y01)

egen ey00 = max(y00)

**\*Generate Difference**

gen did = (ey11 - ey10) - (ey01 - ey00)

sum did

restore

|  |  |  |  |
| --- | --- | --- | --- |
|  | OLS estimate (with population weights) | OLS estimate (without weighting) | Manual estimation |
| Coefficient | .0436016 | .0893807 | .0893809 |
| Std. error | (.0341039) | (.0953924) |  |
| t-test | 1.2784 | 0.9370 |  |

**Question 2.**  What parallel trends assumption is needed for this DiD equation to estimate the ATT?

The core assumption for any DiD analysis, known as “parallel trends” is that the outcomes for the treated and control groups would have moved in parallel during the treatment period, but for the treatment. This assumption cannot be tested, but a core plausibility check is to verify whether trends are parallel during the *pretreatment* period. Good covariate balance between the two groups made the parallel trends assumption more plausible.

**Question 3**. Estimate the dynamic specification of the DiD design using OLS. One pre-treatment year must be omitted, as a baseline year; you can omit 2005. Otherwise, the dynamic model includes interactions of the treatment dummy with calendar year dummies. Estimate the event study model using OLS with standard errors clustered on state.

**STATA code for step 3**

**\*OLS, with year 2005 as omitted category**

reg homicide\_log treat##ib2005.year, cluster(state)

outreg2 using "$desktop\regression.xls", bdec(6) sdec(6) ctitle("w/out Analytical Weight") replace

**Question 3a.** Under what assumptions are the pre-treatment leads (e.g., 2000 to 2004) equal to zero?

No anticipation: the treatment effects should be zero prior to treatment.

**Question 3b**. Under what assumptions are the post-treatment lags (e.g., 2006 to 2010) estimates of the treatment effect?

1. SUTVA: this assumption rules out interference across units, spillover effects, and general equilibrium effects.

2. parallel trends

**Question 4**. Plot the event study coefficients you estimated, ideally with 95% confidence intervals.

**STATA code for step 4**

**\*Before starting step 4, save outreg2 output from Step 3 as “.xlsx”.**

**\*Rename**

rename (A B) (year estimate)

keep year estimate

**\*Drop omitted categories and vars we dont use**

drop in 1/47

drop in 11/12

drop in 21/27

**\*Cleaning**

replace year=subinstr(year, "1.treat#","",.)

replace year=subinstr(year, ".year","",.)

destring year, replace force

carryforward year if year==., replace

**\*Eliminate the parenthesis standard errors are reported with, and the stars**

replace estimate=subinstr(estimate, "(", "",.)

replace estimate=subinstr(estimate, ")", "",.)

replace estimate=subinstr(estimate, "\*\*\*", "",.)

replace estimate=subinstr(estimate, "\*\*", "",.)

replace estimate=subinstr(estimate, "\*", "",.)

**\*Destring estimates**

destring estimate, replace force

**\*Generate separate columns for standard errors and coefficients**

gen obs=\_n

gen odd=mod(obs,2)

gen coef=estimate if odd==1

carryforward coef if coef==., replace

keep if odd==0

rename (estimate) (se)

drop obs odd

order year

**\*Generate confidence intervals**

gen ci\_up=coef+1.96\*se

gen ci\_low=coef-1.96\*se

**\*Generate 2005 null value**

expand 2 if year==2004

sort year

**\*Generate year identifier**

gen id=\_n

replace se=. if id==6

replace coef=0 if id==6

replace ci\_low=0 if id==6

replace ci\_up=0 if id==6

replace year=2005 if id==6

**\*Labels**

label var ci\_up "CI"

label var ci\_low "CI"

label var coef "Year FE \* Treatment"

label var id "Year"

**\*Plot**

twoway scatter coef id, connect(l) msymbol(triangle) lcol(purple) lwidth(medthick) /\*

\*/ || scatter ci\_low id, connect(l) lpattern(shortdash\_dot) lcol(red) lwidth(medthick) msymbol(none) /\*

\*/ || scatter ci\_up id, connect(l) lpatter(shortdash\_dot) lcol(red) lwidth(medthick) msymbol(none) /\*

\*/ yline(0, lcolor(black) lwidth(thin) lpattern(shortdash\_dot)) /\*

\*/ xline(6.5, lcolor(black) lwidth(thin) lpattern(line)) /\*

\*/ xtitle("Year", size(medium)) /\*

\*/ xlabel(1 "2000" 2 "2001" 3 "2002" 4 "2003" 5 "2004" 6 "2005" 7 "2006" 8 "2007" 9 "2008" 10 "2009" 11 "2010") /\*

\*/ legend(order(1 2) col(2) size(medium)) /\*

\*/ ymtick(0.1(0.025)-0.2) /\*

\*/ ylabel(0.1 "0.1" 0 "0" -0.1 "-0.1" -0.2 "-0.2", grid) /\*

\*/ ytitle("Marginal effect", size(medium) height(5)) /\*

\*/ title("Year FE \* Treatment on Log(Homicide)", size(medium)) subtitle("")

graph export "$desktop\graph\_year\_fe\_treatment.png", as(png) replace width(3050) height(1350)

A graph with red and blue lines

Description automatically generated

**Question 4a**. What is your estimate of the ATT (without population weights)?

The ATT is 0.0894 (averaged over 2006-2010, relative to a base of the average over 2000-2005) but, as the leads-and-lags graphs suggests, is not statistically significant at the 5% level.

**Question 4b**. Using the event study plots to assess the plausibility of the parallel trends assumption during the pre-treatment period, does it seem that the two groups are trending different during the pre-treatment period?

Using an “eyeball test,” the two groups have similar trends during the pre-treatment period.

[Beyond what the assignment asked for. It can also be informative to plot trends separately for the treated and control groups. This is done below.]

\***STATA Code for step 4b**.

**\*Generate means over years 2000-2010.**

forval yearx=2000/2010{

**\*Treated group**

egen y1\_`yearx' = mean(homicide\_log) if treat==1 & year==`yearx'

egen ey1\_`yearx' = max(y1\_`yearx')

**\*Control group**

egen y0\_`yearx' = mean(homicide\_log) if treat==0 & year==`yearx'

egen ey0\_`yearx' = max(y0\_`yearx')

}

**\*Keep only one line for each year\*treatment group**

keep state ey1\* ey0\* //keep state for the reshape command

duplicates drop state ey1\* ey0\*, force

**\*Rename and reshape.**

rename (ey1\_\* ey0\_\*) (ey1\* ey0\*)

reshape long ey1 ey0, i(state) j(year)

duplicates drop ey1 ey0, force //to have just one row for each group\*year

drop state

**\*Labels**

lab var ey1 "Treated"

lab var ey0 "Control"

**\*Plot**

twoway scatter ey1 year, connect(l) msymbol(triangle) lcol(purple) lwidth(medthick) /\*

\*/ || scatter ey0 year, connect(l) msymbol(diamond) lcol(red) lwidth(medthick) /\*

\*/ yline(0, lcolor(black) lwidth(thin) lpattern(shortdash\_dot)) /\*

\*/ xline(2005.5, lcolor(black) lwidth(thin) lpattern(line)) /\*

\*/ xtitle("Year", size(medium)) /\*

\*/ xlabel(2000 "2000" 2001 "2001" 2002 "2002" 2003 "2003" 2004 "2004" 2005 "2005" 2006 "2006" 2007 "2007" 2008 "2008" 2009 "2009" 2010 "2010") /\*

\*/ legend(order(1 2) col(2) size(medium)) /\*

\*/ ytitle("Average", size(medium) height(5)) /\*

\*/ title("Log(Homicide) average for treated and control groups overtime", size(medium)) subtitle("")

graph export "$desktop\graph\_trend\_for\_loghomicide\_by\_group\_year.png", as(png) replace width(3050) height(1350)

A graph with a line and a line

Description automatically generated with medium confidence

The major observation from this plot, and one reason why it is worthwhile to plot the two groups separately: The two groups are *not similar*. Treated states had higher homicide rates than control states in the pre-treatment period. This calls into question the parallel trends assumption, since whatever factors caused these differences might change during the treatment period, and lead to an apparent treatment effect, which was not a true outcome of the treatment. Note the change in the control states during 2009, with no similar change in the treated states.

**Question 4c.** Treatment effects often emerge over time. One can interpret the event study coefficients during the treatment period as ATT estimates by year. Does it appear that the effects change over time during the treatment period, or does it appear as if there was a one-time treatment effect, at the time of treatment, that thereafter is constant over time?

For this particular project, there does not seem to be gradual emergence of a treatment effect.