KAIST Summer Session 2018

Module 2. Causal Inference with STATA

Replication (2) Difference-in-Differences

KAIST College of Business

Jiyong Park

26 July, 2018

Let's Replicate the Work of Greenwood and Wattal (2017)

RESEARCH ARTICLE

SHOW ME THE WAY TO GO HOME: AN EMPIRICAL INVESTIGATION OF RIDE-SHARING AND ALCOHOL RELATED MOTOR VEHICLE FATALITIES[1]

Brad N. Greenwood and Sunil Wattal
Fox School of Business, Temple University, 1801 Liacouras Walk,
Philadelphia, PA 19122 U.S.A. {greenwood@temple.edu} {swattal@temple.edu}

In this work, we investigate how the entry of ride-sharing services influences the rate of alcohol related motor vehicle fatalities. While significant debate has surrounded ride-sharing, limited empirical work has been devoted to uncovering the societal benefits of such services (or the mechanisms which drive these benefits). Using a ***difference in difference*** *approach to exploit a natural experiment, the entry of Uber Black and Uber X into California markets between 2009 and 2014, we find a significant drop in the rate of fatalities after the introduction of Uber X. Further, results suggest that not all services have the same effect, insofar as the effect of the Uber Black car service is intermittent and manifests only in selective locations (i.e., large cities). These results underscore the importance of coupling increased availability with cost savings in order to exploit the public welfare gains offered by the sharing economy. Practical and theoretical implications are discussed.*

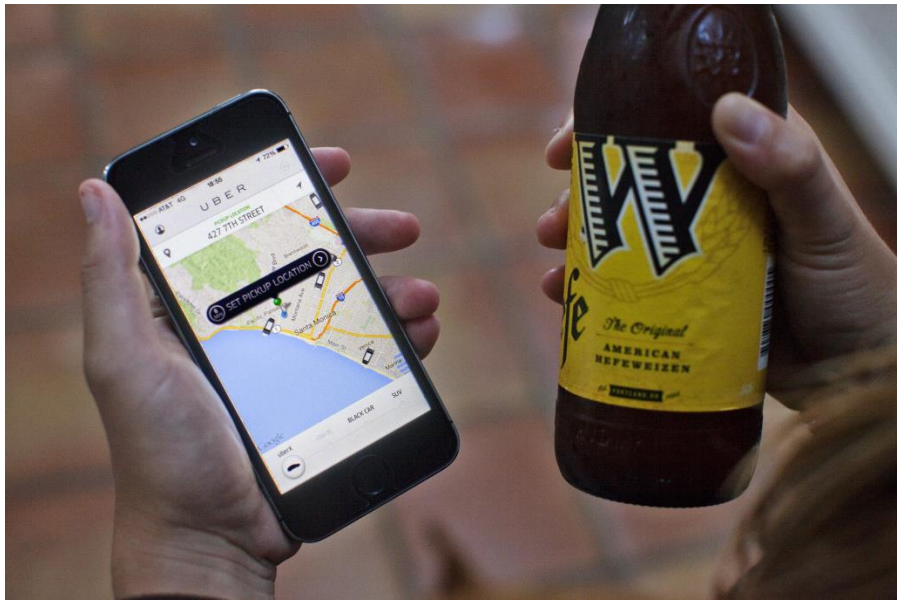
Keywords: Uber, sharing economy, ride-sharing, drunk driving, vehicular fatalities, difference in difference, natural experiment, platforms

Greenwood, B.N. and Wattal, S., 2017. Show Me the Way to Go Home: An Empirical Investigation of Ride-Sharing and Alcohol Related Motor Vehicle Fatalities. *MIS Quarterly*, 41(1), pp.163-187.

Let's Replicate the Work of Greenwood and Wattal (2017)

- Research question:

  Can ride-sharing services reduce alcohol-related motor vehicle fatalities?



Greenwood, B.N. and Wattal, S., 2017. Show Me the Way to Go Home: An Empirical Investigation of Ride-Sharing and Alcohol Related Motor Vehicle Fatalities. *MIS Quarterly*, 41(1), pp.163-187.

Let's Replicate the Work of Greenwood and Wattal (2017)

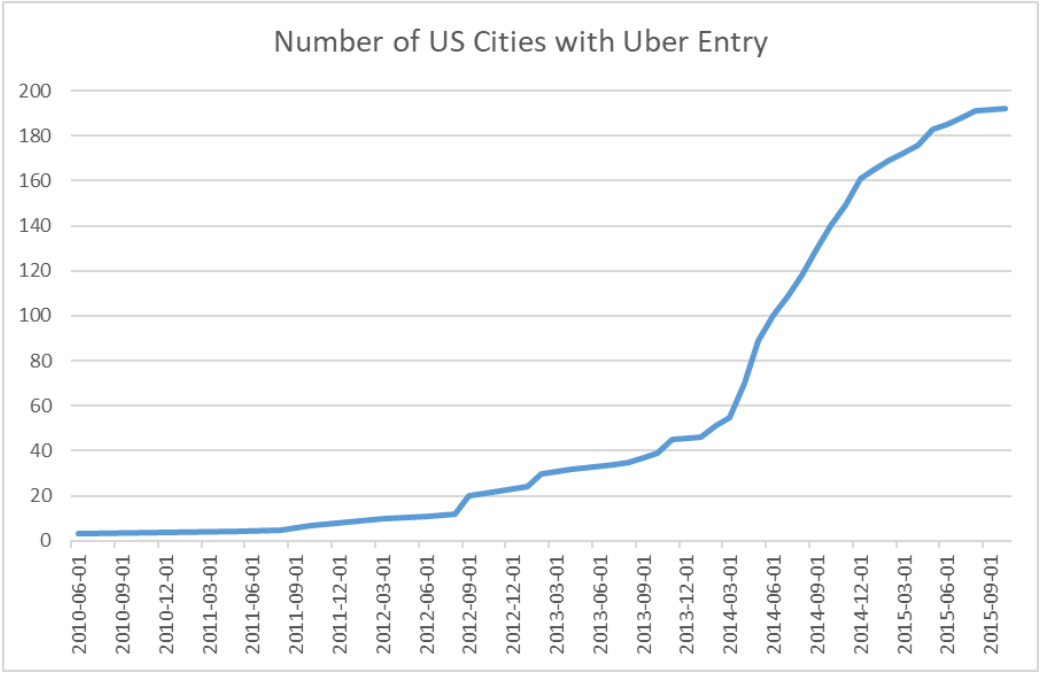
- To answer this research question, Greenwood and Wattal (2017) exploit a difference-in-differences design, given that the rollout of Uber is spatiotemporally staggered in the U.S.

	Before Treatment	**After Treatment**
Treatment (Cities with Uber entry)	T_B	T_A
Control (Cities without Uber entry)	C_B	C_A

Greenwood, B.N. and Wattal, S., 2017. Show Me the Way to Go Home: An Empirical Investigation of Ride-Sharing and Alcohol Related Motor Vehicle Fatalities. *MIS Quarterly*, 41(1), pp.163-187.

Let's Replicate the Work of Greenwood and Wattal (2017)

- To answer this research question, Greenwood and Wattal (2017) exploit a difference-in-differences design, given that the rollout of Uber is spatiotemporally staggered in the U.S.



Greenwood, B.N. and Wattal, S., 2017. Show Me the Way to Go Home: An Empirical Investigation of Ride-Sharing and Alcohol Related Motor Vehicle Fatalities. *MIS Quarterly*, 41(1), pp.163-187.

(1) Designing the Final Data Structure First

- Model specification
  - Difference-in-differences (DID)

$$y_{it} = \alpha_i + \beta_t + \gamma Uber\_Entry_{it} + CONTROLs + \varepsilon_{it}$$

where $Uber\_Entry_{it}$ indicates whether Uber services are available in the UA i at time t.

- Final panel structure (long panel)

Urbanized Area (i)	Year_Month (t)	Alcohol-related Fatalities	Uber_Entry	…
1	1	2	0	
1	2	1	0	
1	3	1	0	
1	4	0	1	
1	5	0	1	

(1) Designing the Final Data Structure First

- Where can you obtain the relevant data?
  - Fatality data: National Highway Traffic Safety Administration (NHTSA) – Fatality Analysis Reporting System (FARS, https://www-fars.nhtsa.dot.gov) (*Fatal Crash Data.csv*)

Two vehicles with single drivers are involved in this accident.

Year	statenum	casenum	vnumber	pnumber	city	accmon	latitude	longitude	numfatal	alcres	dui
2005	1	1	1	1	0	1	30.50191	-88.1576	1	96	0
2005	1	1	2	1	0	1	30.50191	-88.1576	1	21	1
2005	1	2	1	1	0	1	33.99218	-86.9706	1	22	1
2005	1	3	1	1	0	1	33.60046	-86.3226	1	99	0
2005	1	4	1	1	0	1	34.63493	-86.4338	1	0	0
2005	1	5	1	1	0	1	.	.	1	99	0
2005	1	6	1	1	0	1	31.46446	-87.384	1	96	0
2005	1	6	1	2	0	1	31.46446	-87.384	1	96	0
2005	1	6	1	3	0	1	31.46446	-87.384	1	96	0
2005	1	7	1	1	0	1	33.09427	-86.5356	1	15	1
2005	1	8	2	1	0	1	33.39353	-86.6655	1	96	0
2005	1	8	1	1	0	1	33.39353	-86.6655	1	33	1

One vehicle with three passengers is involved in this accident.

(1) Designing the Final Data Structure First

- Where can you obtain the relevant data?
  - Uber entry data: Uber blogs and media announcements (hand-collected; *Uber Entry.dta*)
  - For each urbanized area, you can obtain annual statistics about demographics and socio-economic status (*Demographics_panel.xlsx*).

Name	Uber_entrydate
Antioch, CA	2010-06-15
Concord, CA	2010-06-15
San Francisco-Oakland, CA	2010-06-15
New York-Newark, NY-NJ-CT	2011-05-15
Seattle, WA	2011-08-15
Kankakee, IL	2011-10-15
Chicago, IL-IN	2011-10-15
Santa Clarita, CA	2012-03-15
Lancaster-Palmdale, CA	2012-03-15
Los Angeles-Long Beach-Anaheim, CA	2012-03-15
Philadelphia, PA-NJ-DE-MD	2012-06-15
San Diego, CA	2012-08-15
Athens-Clarke County, GA	2012-09-15
Gainesville, GA	2012-09-15
Macon, GA	2012-09-15
Albany, GA	2012-09-15
Atlanta, GA	2012-09-15
Boston, MA-NH-RI	2012-09-15

(2) Pre-processing Datasets

- **[STATA Practice]** Collapsing data

Year	statenum	casenum	DUI
2005	1	1	1
2005	1	1	1
2005	1	1	2
2005	1	2	1
2005	1	3	1
2005	1	3	3
2005	1	4	1
2005	1	4	2

Year	statenum	casenum	DUI
2005	1	1	14
2005	1	2	1
2005	1	3	4
2005	1	4	3

You need to set the directory where your files are located.

```
cd "E:\Desktop\"
import delimited "STATA_Lab2\Fatal Crash Data.csv", case(preserve) clear

drop if latitude ==.
drop if longitude ==.
collapse (first) latitude longitude (mean) numfatal (sum) dui , by( Year Month statenum casenum)
replace dui=1 if dui>0

save "STATA_Lab2\Fatal Crash Data_crash level.dta"
```

(2) Pre-processing Datasets

- **[STATA Practice]** Spatial mapping
  - The best way to deal with spatial data (e.g., longitude/latitude) is to use geographic information systems (e.g., ArcGIS). Here, we use STATA packages for spatial mapping in a easier, but less precise way.
  - In my opinion, spatial econometrics is a powerful tool for social big data (e.g., Uber transactions, Airbnb listings, mobile usage).

```
use "STATA_Lab2\Fatal Crash Data_crash level.dta", clear
append using "STATA_Lab2\City Location.dta", generate(city)

gen _id=_n
gen crash=1
replace crash=0 if city==1
distmatch, id(_id) lat( latitude) long( longitude) don(city) rec(crash) near(1) max(10)

replace ID=0 if city==0
vlookup __id1 , gen(_included) key( _id ) value(ID)
collapse (sum) numfatal dui, by(_included Year Month)
rename _included ID
drop if ID==.

save "STATA_Lab2\Fatal Crash Data_city level.dta"
```

(2) Pre-processing Datasets

- **[STATA Practice]** Extracting substring

Name	Uber_entrydate
Antioch, CA	2010-06-15
Concord, CA	2010-06-15
San Francisco-Oakland, CA	2010-06-15
New York-Newark, NY-NJ-CT	2011-05-15
Seattle, WA	2011-08-15
Kankakee, IL	2011-10-15
Chicago, IL-IN	2011-10-15
Santa Clarita, CA	2012-03-15
Lancaster-Palmdale, CA	2012-03-15
Los Angeles-Long Beach-Anaheim, CA	2012-03-15

Name			
Antioch, CA	2010	6	15
Concord, CA	2010	6	15
San Francisco-Oakland, CA	2010	6	15
New York-Newark, NY-NJ-CT	2011	5	15
Seattle, WA	2011	8	15
Kankakee, IL	2011	10	15
Chicago, IL-IN	2011	10	15
Santa Clarita, CA	2012	3	15
Lancaster-Palmdale, CA	2012	3	15
Los Angeles-Long Beach-Anaheim, CA	2012	3	15

```
use "STATA_Lab2\Uber Entry.dta", clear

gen Year = substr( Uber_entrydate,1,4)
gen Month = substr( Uber_entrydate,6,2)
destring Year, replace
destring Month, replace

save "STATA_Lab2\Uber Entry_modified.dta"
```

(3) Matching Treatment and Control Groups

- **[STATA Practice]** Matching (1) – Propensity score matching
  - ➢ You can match the cities with Uber entry (treatment) to other cities (control), based on average demographic statistics before 2010.

```
import excel "STATA_Lab2\Demographics_panel.xlsx", sheet("Sheet1") firstrow clear
keep if Year<2010
collapse (mean) Population Percentbachelorsdegree Income Unemployment, by( ID UA_name)
merge 1:1 UA_name using "STATA_Lab2\Uber Entry_modified.dta"
drop if _merge==2

gen lnIncome = ln(Income)
gen lnPopulation=ln(Population)
gen treatment=0
replace treatment=1 if _merge==3

ssc install psmatch2
psmatch2 treatment lnPopulation Percentbachelorsdegree lnIncome Unemployment, n(1)
pstest lnPopulation Percentbachelorsdegree lnIncome Unemployment, both

vlookup _id, gen(_included) key(_n1) value(_treated)
gen matched = 0
replace matched=1 if _included==1
replace matched=1 if _treated==1

save "STATA_Lab2\Matching.dta"
```

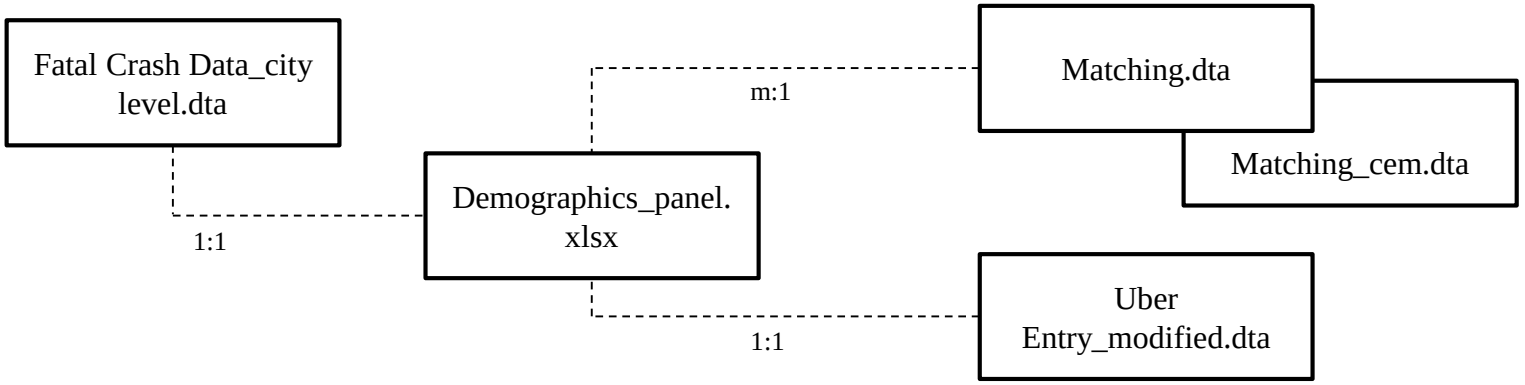
(3) Matching Treatment and Control Groups

- **[STATA Practice]** Matching (2) – Coarsened exact matching
  - ➢ You can match the cities with Uber entry (treatment) to other cities (control), based on average demographic statistics before 2010.

```
import excel "STATA_Lab2\Demographics_panel.xlsx", sheet("Sheet1") firstrow clear
keep if Year<2010
collapse (mean) Population Percentbachelorsdegree Income Unemployment, by( ID UA_name)
merge 1:1 UA_name using "STATA_Lab2\Uber Entry_modified.dta"
drop if _merge==2

gen lnIncome = ln(Income)
gen lnPopulation=ln(Population)
gen treatment=0
replace treatment=1 if _merge==3

ssc install cem
cem Population (100000 250000 500000 1000000) Percentbachelorsdegree (#5) Income (#5) Unemployment
(#5) , treatment(treatment)

save "STATA_Lab2\Matching_cem.dta"
```

(4) Merging Datasets

- **[STATA Practice]** Constructing the final panel dataset




```
import excel "STATA_Lab2\Demographics_panel.xlsx", sheet("Sheet1") firstrow clear

merge 1:1 ID Year Month using "STATA_Lab2\Fatal Crash Data_city level.dta", gen(crash_merge)

replace numfatal=0 if crash_merge==1
replace dui=0 if crash_merge==1
drop crash_merge

merge m:1 ID using "STATA_Lab2\Matching.dta", gen(match_merge)
drop Uber_entrydate _merge lnIncome lnPopulation treatment _pscore _treated _support _weight _id _n1 _nn
_pdif _included match_merge

merge m:1 ID using "STATA_Lab2\Matching_cem.dta", gen(cem_match_merge)
drop Uber_entrydate _merge lnIncome lnPopulation treatment cem_match_merge

merge 1:1 ID Year Month using "STATA_Lab2\Uber Entry_modified.dta", gen(Uber_merge)
drop if Uber_merge==2
```

(5) Difference-in-Differences Estimations

- **[STATA Practice]** Econometric analysis
  - Model specification

$$y_{it} = \alpha_i + \beta_t + \gamma Uber\_Entry_{it} + CONTROLs + \varepsilon_{it}$$

where $Uber\_Entry_{it}$ indicates whether Uber services are available in the UA i at time t.

```
egen time = group(Year Month)
xtset ID time
gen Uber_entry=0
replace Uber_entry=1 if Uber_entrydate~=""
replace Uber_entry=L.Uber_entry if Uber_entry==0
replace Uber_entry=0 if Uber_entry==.
gen lnDUI = ln(dui+1)
gen lnIncome = ln(Income)
gen lnPopulation = ln(Population)

xtreg lnDUI Uber_entry  , fe cluster(ID)
reg lnDUI Uber_entry  i.time , cluster(ID)

xtreg lnDUI Uber_entry  i.time , fe cluster(ID)
xtreg lnDUI Uber_entry  i.time if matched==1 , fe cluster(ID)
xtreg lnDUI Uber_entry  i.time if cem_matched==1 , fe cluster(ID)
xtreg lnDUI Uber_entry lnPopulation lnIncome Percentbachelorsdegree Unemployment i.time if matched==1 , fe cluster(ID)
xtreg lnDUI Uber_entry lnPopulation lnIncome Percentbachelorsdegree Unemployment i.time if cem_matched==1 , fe cluster(ID)
```

(6) Reporting Estimation Results

- **[STATA Practice]** Displaying regression tables

```
ssc install estout

xtreg lnDUI Uber_entry  , fe cluster(ID)
eststo results1
reg lnDUI Uber_entry  i.time , cluster(ID)
eststo results2
xtreg lnDUI Uber_entry  i.time , fe cluster(ID)
eststo results3
xtreg lnDUI Uber_entry  i.time if matched==1 , fe cluster(ID)
eststo results4
xtreg lnDUI Uber_entry  i.time if cem_matched==1 , fe cluster(ID)
eststo results5
xtreg lnDUI Uber_entry lnPopulation lnIncome Percentbachelorsdegree Unemployment i.time if matched==1 , fe cluster(ID)
eststo results6
xtreg lnDUI Uber_entry lnPopulation lnIncome Percentbachelorsdegree Unemployment i.time if cem_matched==1 , fe cluster(ID)
eststo results7

esttab results1 results2 results3 results4 results5 results6 results7
esttab results1 results2 results3 results4 results5 results6 results7, keep(Uber_entry lnPopulation lnIncome
Percentbachelorsdegree Unemployment) se star(* 0.10 ** 0.05 *** 0.01)
esttab results1 results2 results3 results4 results5 results6 results7 using DID_estimations.csv, keep(Uber_entry lnPopulation
lnIncome Percentbachelorsdegree Unemployment) se star(* 0.10 ** 0.05 *** 0.01)
```

End of Document