ICA 2

Daniel Adiele

02.29.2024

Learning Objectives

In this assignment, you will practice your T-Test and R programming skills. You will use a fixed effects model and perform a difference-in-differences analysis. Please answer the questions in a R markdown file and "Knit" the file so that I can see your analysis.

Experiments

Reanalysis of Gerber, Green and Larimer (2008) 'Why do large numbers of people vote, despite the fact that, as Hegel once observed, "the casting of a single vote is of no significance where there is a multitude of electors"?'

This is the question that drives the experimental analysis of Gerber, Green and Larimer (2008). If it is irrational to vote because the costs of doings so (time spent informing oneself, time spent getting to the polling station, etc) are clearly greater than the gains to be made from voting (the probability that any individual voter will be decisive in an election is vanishingly small), then why do we observe millions of people voting in elections? One commonly proposed answer is that voters may have some sense of civic duty which drives them to the polls. Gerber, Green and Larimer investigate this idea empirically by priming voters to think about civic duty while also varying the amount of social pressure voters are subject to. In a field experiment in advance of the 2006 primary election in Michigan, nearly 350,000 voters were assigned at random to one of four treatment groups, where voters received mailouts which encouraged them to vote, or a control group where voters received no mailout. The treatment and control conditions were as follows:

```
Treatment 1 ("Civic duty"): Voters receive mailout reminding them that voting is a civic duty
Treatment 2 ("Hawthorne"): Voters receive mailout telling them that researchers would be studying their
Treatment 3 ("Self"): Voters receive mailout displaying the record of turnout for their household in pri
Treatment 4 ("Neighbors"): Voters receive mailout displaying the record of turnout for their household a
Control: Voters receive no mailout.
```

Load the replication data for Gerber, Green and Larimer (2008). This data is stored in a .Rdata format, which is the main way to save data in R. Therefore you will not be able to use read.csv but instead should use the function load.

```r
# You will need to change the file location for the code to work.
load("~/Downloads/ICA 2/gerber_green_larimer.Rdata")
```

Once you have loaded the data, familiarize yourself with the the gerber object which should be in your current environment. Use the str and summary functions to get an idea of what is in the data. There are 5 variables in this data.frame: Variable name Description

- voted: Indicator for whether the voter voted in the 2006 election (1) or did not vote (0)

	Control	Civic Duty	Hawthorne	Self	Neighbors
PCT Voted	0.30	0.31	0.32	0.35	0.38

- treatment: Factor variable indicating which treatment arm (or control group) the voter was allocated to
- sex: Sex of the respondent
- yob: Year of birth of the respondent
- p2004: Indicator for whether the voter voted in the 2004 election (Yes) or not (No)

1. Calculate the turnout rates, "voted", for each of the experimental groups (4 treatments, 1 control). Calculate the number of individuals allocated to each group. Recreate table 2 on p. 38 of the paper.

```
library(modelsummary)
```

```
## Version 2.0.0 of 'modelsummary', to be released soon, will introduce a
##   breaking change: The default table-drawing package will be 'tinytable'
##   instead of 'kableExtra'. All currently supported table-drawing packages
##   will continue to be supported for the foreseeable future, including
##   'kableExtra', 'gt', 'huxtable', 'flextable, and 'DT'.
##
##   You can always call the 'config_modelsummary()' function to change the
##   default table-drawing package in persistent fashion. To try 'tinytable'
##   now:
##
##   config_modelsummary(factory_default = 'tinytable')
##
##   To set the default back to 'kableExtra':
##
##   config_modelsummary(factory_default = 'kableExtra')
```

```
datasummary((`PCT Voted`=voted) ~ Mean*treatment, data=gerber)
```

2. Use the following code to create three new variables in the data.frame. First, a variable that is equal to 1 if a respondent is female, and 0 otherwise. Second, a variable that measures the age of each voter in years at the time of the experiment (which was conducted in 2006). Third, a variable that is equal to 1 if the voter voted in the November 2004 Midterm election.

```
## Female dummy variable
gerber$female <- ifelse(gerber$sex == "female", 1, 0)

## Age variable
gerber$age <- 2006 - gerber$yob

## 2004 variable
gerber$turnout04 <- ifelse(gerber$p2004 == "Yes", 1, 0)
```

3. Using these variables, conduct balance checks to establish whether there are potentially confounding differences between treatment and control groups. You do this by using the variables female, age, and turnout04 as response (dependent) variables. Use just the factor variable of treatment as your predictor (independent/explanatory) variable. Can you conclude from the results that randomization worked? How do you know?

	Female	Age	Turnout
(Intercept)	0.499***	49.814***	0.400***
	(0.001)	(0.033)	(0.001)
Civic Duty	0.001	−0.155+	−0.001
	(0.003)	(0.081)	(0.003)
Hawthorne	0.000	−0.109	0.003
	(0.003)	(0.081)	(0.003)
Self	0.001	−0.021	0.002
	(0.003)	(0.081)	(0.003)
Neighbors	0.001	0.039	0.006*
	(0.003)	(0.081)	(0.003)
Num.Obs.	344 084	344 084	344 084
R2	0.000	0.000	0.000
R2 Adj.	0.000	0.000	0.000
AIC	499 477.3	2 814 316.6	485 852.4
BIC	499 541.8	2 814 381.1	485 916.9
Log.Lik.	−249 732.671	−1 407 152.317	−242 920.207
F	0.083	1.420	1.653
RMSE	0.50	14.45	0.49

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Based on the results gotten, we can conclude that randomization worked because there are no statistically significant differences between the treatment and control groups across the variables female, age, and turnout04. The coefficients for the treatment variable (Civic Duty, Hawthorne, Self, Neighbors) are not statistically significant ($p > 0.05$) for any of the response variables, indicating that there are no confounding differences between the groups. Additionally, the R-squared values equal zero, suggesting that the treatment variable does not explain a significant portion of the variation in the response variables. Therefore, randomization appears to have successfully balanced the groups, as there are no discernible differences between them based on the specified variables.Also based on the output gotten we also see that there is no difference in the means of the treatment group and controlled group across the variables female, age, and turnout04. These are some of the way we can tell that randomization worked.

```
exp1 <-lm(female ~ factor(treatment), data =gerber)
exp2 <-lm(age ~ factor(treatment), data =gerber)
exp3 <-lm(turnout04 ~ factor(treatment), data =gerber)
library(modelsummary)
modelsummary(list("Female"=exp1,"Age"=exp2,"Turnout"=exp3),coef_rename = coef_rename, star= TRUE)
```

4. Estimate the average treatment effects of the different treatment arms whilst controlling for the variables you created for the question above. How do these estimates differ from regression estimates of the treatment effects only (i.e. without controlling for other factors)? Why?

The estimates of average treatment effects (ATE) while controlling for other variables differ from regression estimates of treatment effects because they account for potential confounding factors. Controlling for variables such as female, age, and turnout04 helps isolate the specific impact of each treatment arm on the outcome variable by removing the influence of these confounding factors. Without controlling for other factors, regression estimates of treatment effects may be biased due to omitted variable bias, leading to less accurate and reliable estimates of the true treatment effects. Therefore, controlling for other factors provides a more nuanced and accurate understanding of the treatment effects by accounting for potential confounders and producing more robust estimates.

	(1)	(2)
(Intercept)	0.297***	0.044***
	(0.001)	(0.003)
Civic Duty	0.018***	0.019***
	(0.003)	(0.003)
Hawthorne	0.026***	0.026***
	(0.003)	(0.003)
Self	0.049***	0.048***
	(0.003)	(0.003)
Neighbors	0.081***	0.080***
	(0.003)	(0.003)
Female		−0.008***
		(0.002)
Age		0.004***
		(0.000)
Turnout04		0.148***
		(0.002)
Num.Obs.	344 084	344 084
R2	0.003	0.045
R2 Adj.	0.003	0.045
AIC	448 179.9	433 501.1
BIC	448 244.4	433 597.9
Log.Lik.	−224 083.935	−216 741.564
F	292.976	2317.892
RMSE	0.46	0.45

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Now based on the result gotten we can say they do not differ at all. Since the variables are truly independent of the error term, they will not affect the estimate, making them virtually the same values (meaning that randomization worked).

```
exp4 <-lm(voted ~ factor(treatment), data =gerber)
# Now use the same equation above but add female, age, and turnout04
exp5 <-lm(voted ~ factor (treatment)+female+age+turnout04, data=gerber)
modelsummary(list(exp4,exp5),coef_rename = coef_rename, stars = TRUE)
```

5. Estimate the treatment effects separately for men and women. Do you note any differences in the impact of the treatment amongst these subgroups?

There is not much difference between the two variables. Upon reviewing the provided results, it appears that there are no substantial differences in the impact of the treatment among men and women. Both groups exhibit similar patterns of treatment effects across the different treatment arms, as indicated by the comparable coefficients and statistical significance levels. This suggests that the treatment has a consistent effect on both men and women, without significant variations between the two subgroups.

```
# modify the equation below for just men
exp6 <-lm(voted ~ factor(treatment), data =gerber[gerber$female==0, ])
# modify the equation below for just women
exp7 <-lm(voted ~ factor(treatment), data =gerber [gerber$female==1, ])
modelsummary(list("Men"=exp6,"Women"=exp7),coef_rename = coef_rename, stars = TRUE)
```

	Men	Women
(Intercept)	0.303***	0.290***
	(0.002)	(0.001)
Civic Duty	0.020***	0.016***
	(0.004)	(0.004)
Hawthorne	0.025***	0.027***
	(0.004)	(0.004)
Self	0.046***	0.051***
	(0.004)	(0.004)
Neighbors	0.082***	0.081***
	(0.004)	(0.004)
Num.Obs.	172 289	171 795
R2	0.003	0.004
R2 Adj.	0.003	0.003
AIC	226 155.0	221 958.4
BIC	226 215.3	222 018.8
Log.Lik.	−113 071.476	−110 973.214
F	142.616	151.163
RMSE	0.47	0.46

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Difference-in-Differences: Replication Exercise from the notes

The data are about the expansion of the Earned Income Tax Credit. The sample only contains single women. This legislation is 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.

Big Hint: Most of the code you need is in the notes

Variable Names and Definitions

state: Factor variable containing the state's FIPS code. year: Calendar Year urate: unemployment rate for the state and year children: number of children in the household nonwhite: the person identifies as non-White finc: Family household income earn: Earned income unearn: unearned income age: Age of the mother in years ed: Years of schooling work: Indicator variable equal to 1 if the person is currently working

The homework questions:

1. Provide Descriptive Statistics for the data. Format nicely, not just R printout. Here is an example below. I have already provided the code to read in the data below. You need to create the data summary table.

```
require(foreign)
eitc<-read.dta("https://github.com/CausalReinforcer/Stata/raw/master/eitc.dta")
library(modelsummary)
# the data mtcars is just an example. You need to replace it with eitc
datasummary((`Unemployment Rate` = urate)+(`Children` = children)+(`Non-White` = nonwhite)+(`Family Inc
```

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. **Hint: Use the tidyverse to make this table. You can either filter the data or use dplyr to construct groups.You can even use data summary to do this step. Below is one example**

	mean	SD	Min	Max
Unemployment Rate	6.76	1.46	2.60	11.40
Children	1.19	1.38	0.00	9.00
Non-White	0.60	0.49	0.00	1.00
Family Income	15 255.32	19 444.25	0.00	575 616.82
Earned Income	10 432.48	18 200.76	0.00	537 880.61
Unearned Income	4.82	7.12	0.00	134.06
Age	35.21	10.16	20.00	54.00
Years of Education	8.81	2.64	0.00	11.00
Work	0.51	0.50	0.00	1.00

	No Children	1 Child	2 or more children
Unemployment Rate	6.66	6.80	6.86
Children	0.00	1.00	2.80
Non-White	0.52	0.60	0.71
Family Income	18 559.86	13 941.57	11 985.30
Earned Income	13 760.26	9928.28	6613.55
Unearned Income	4.80	4.01	5.37
Age	38.50	33.76	32.05
Years of Education	8.55	8.99	9.01
Work	0.57	0.54	0.42

```
# Make the appropriate changes (i.e. data frame name and correct factor variable)
eitc$nochild <- eitc$children
eitc$nochild[eitc$nochild>2]<-2

eitc$nochild <- factor(eitc$nochild, labels = c ("No Children","1 Child", "2 or more children"))

datasummary((`Unemployment Rate` = urate)+(`Children` = children)+(`Non-White` = nonwhite)+(`Family Inc
```

3. Construct a variable for the "treatment" called ANYKIDS. This variable should equal 1 if they have any children and zero otherwise. Create a second variable to indicate after the expansion (called POST93-should be 1 for 1994 and later).

```
# 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)
```

4. Create a graph which plots mean annual employment rates by year (1991-1996) for single women with children (treatment) and without children (control). **Hint: you should have two lines on the same graph.** I would suggest to use ggplot to make this plot. Here is some sample code. The variable "work" is your dependent variable.

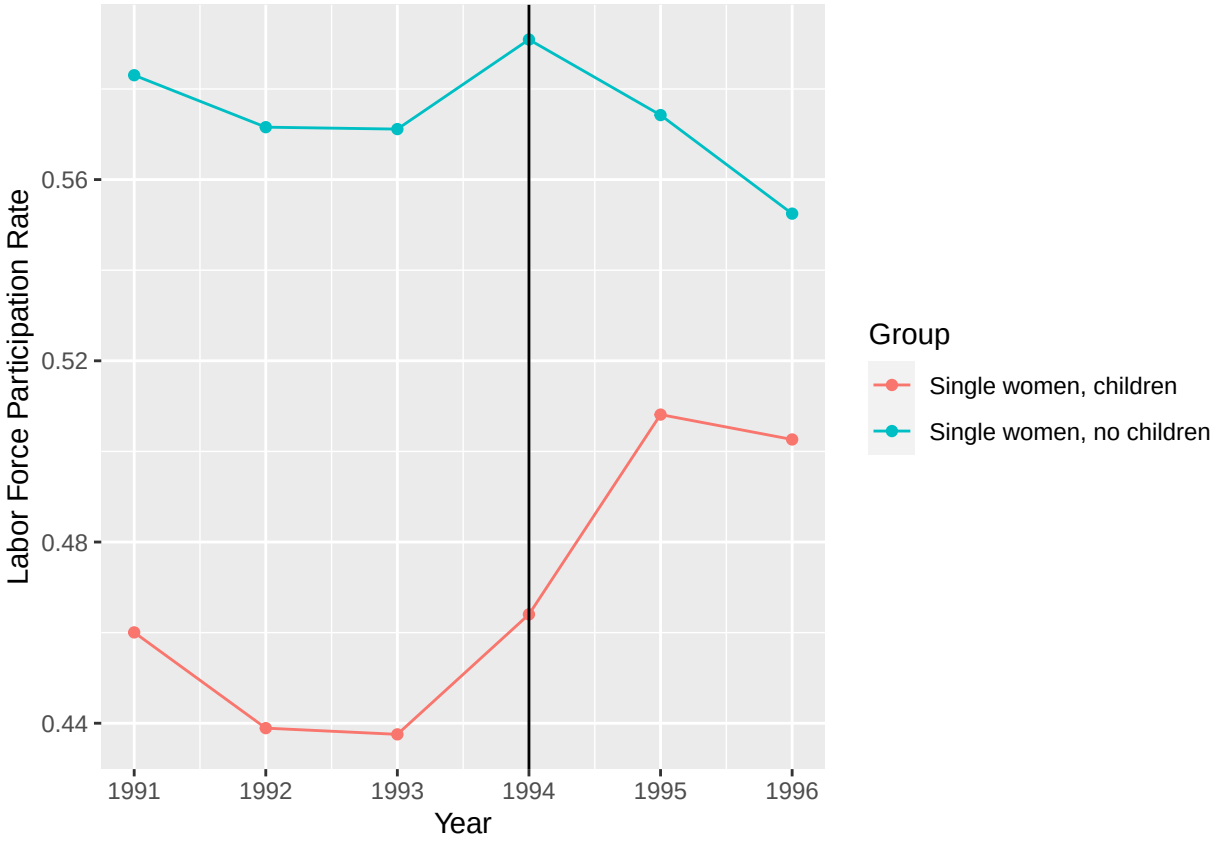
```
# Take average value of 'work' by year, conditional on anykids
minfo = aggregate(eitc$work, list(eitc$year,eitc$anykids == 1), mean)
# rename column headings (variables)
names(minfo) = c("YR","Treatment","LFPR")
```

```r
# 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
```

```
## Loading required package: ggplot2
```

```r
qplot(YR, LFPR, data=minfo, geom=c("point","line"), colour=Group,
xlab="Year", ylab="Labor Force Participation Rate")+geom_vline(xintercept = 1994)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



5. Do the trends between the two groups appear to be parallel? Why is this important?

Yes, because before 1994 the two trend lines (representing the two groups) are parallel and then afterwards that are shown not to be, this is important because we get to see the impact of the treatment on the treated group versus the controlled group. This is important because it supports/helps our argument that the change in trend of the treated group is influenced by the treatment given to the treated group else the two trends (the two groups) would have continued to be parallel. This parallel trend observation is very important because it verifies our difference in differences model has internal validity, which is important to our testing.

6. Calculate the unconditional difference-in-difference estimates of the effect of the 1993 EITC expansion on employment of single women. **Hint: This means calculate the DID treatment effect by just subtracting means (i.e. no regression)**

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

```
##       work
## 0.04687313
```

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. **Hint: your answers for 6 and 7 should match.**

```
# Estimate a difference in difference regression. You should be using ANYKIDS and POST93 in your regres

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

8. Re-estimate this model including demographic characteristics as well as state and year fixed effect. Use the variable nonwhite, age, ed, and unearn as demographics.

```
library(fixest)
reg2 <- feols(work~post93*anykids+nonwhite+age+ed+unearn|state+year, data=eitc)
```

```
## The variable 'post93' has been removed because of collinearity (see $collin.var).
```

9. Explain why can't you use finc, earn, and uearn in the same regression.

Household income is the sum of earned income and unearned income. We can't use all the three variables in the same regression model because it would produce multicollinearity within the model, this is because household income is gotten from the sum of 'earn - earned income' and 'uearn - unearned income' and would lead to multicollinearity if all three were put together in the same regression model.

10. 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 instead of 1994 (so 1992 and 1993 both have this fake policy).

```
eitc$post91 = as.numeric(eitc$year >= 1992)
reg3 <- feols(work~post91*anykids+nonwhite+age+ed+unearn|state+year,
data=eitc[eitc$year<1994, ])
```

```
## The variable 'post91' has been removed because of collinearity (see $collin.var).
```

```
label <- c(work = "Work", post93 = "Post 1993", post91 = "Post 1991", anykids = "Any Kids", nonwhite =

modelsummary(list(reg1,reg2,reg3), vcov = c("robust","robust","robust"), stars = TRUE, coef_rename = lab
```

	(1)	(2)	(3)
(Intercept)	0.575***		
	(0.009)		
Post 1993	−0.002		
	(0.013)		
Any Kids	−0.129***	−0.106***	−0.094***
	(0.012)	(0.012)	(0.019)
Post 1993:Any Kids	0.047**	0.053**	
	(0.017)	(0.016)	
Non-White		−0.081***	−0.076***
		(0.010)	(0.013)
Age		0.003***	0.003***
		(0.000)	(0.001)
Education		0.016***	0.014***
		(0.002)	(0.002)
Unearned Income		−0.017***	−0.018***
		(0.001)	(0.001)
Post 1991:Any Kids			−0.017
			(0.023)
Num.Obs.	13 746	13 746	7401
R2	0.013	0.116	0.122
R2 Adj.	0.012	0.112	0.115
R2 Within		0.086	0.091
R2 Within Adj.		0.086	0.090
AIC	19 779.8	18 380.1	9896.7
BIC	19 817.5	18 846.9	10 304.3
Log.Lik.	−9884.917		
F	58.637		
RMSE	0.50	0.47	0.47
Std.Errors	HC3	HC1	HC1
FE: year		X	X
FE: state		X	X

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001