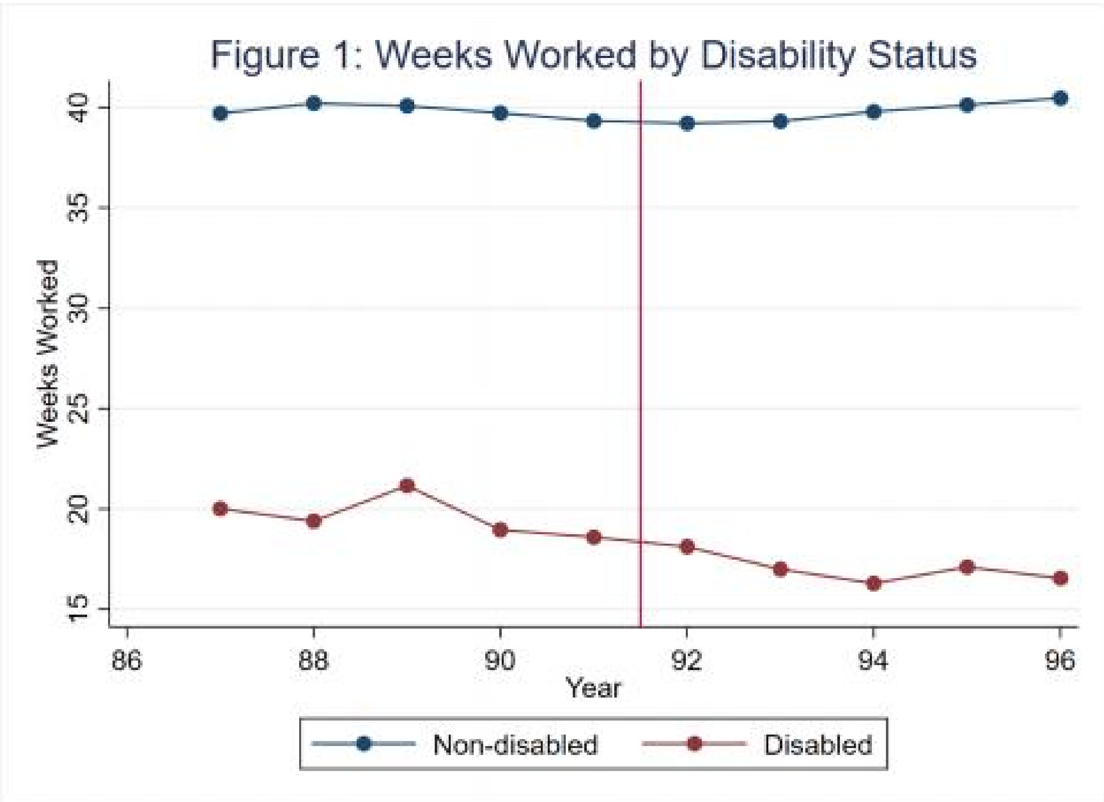
Economics 1016 Gregory A. Bruich, Ph.D.
Summer 2020 Harvard University

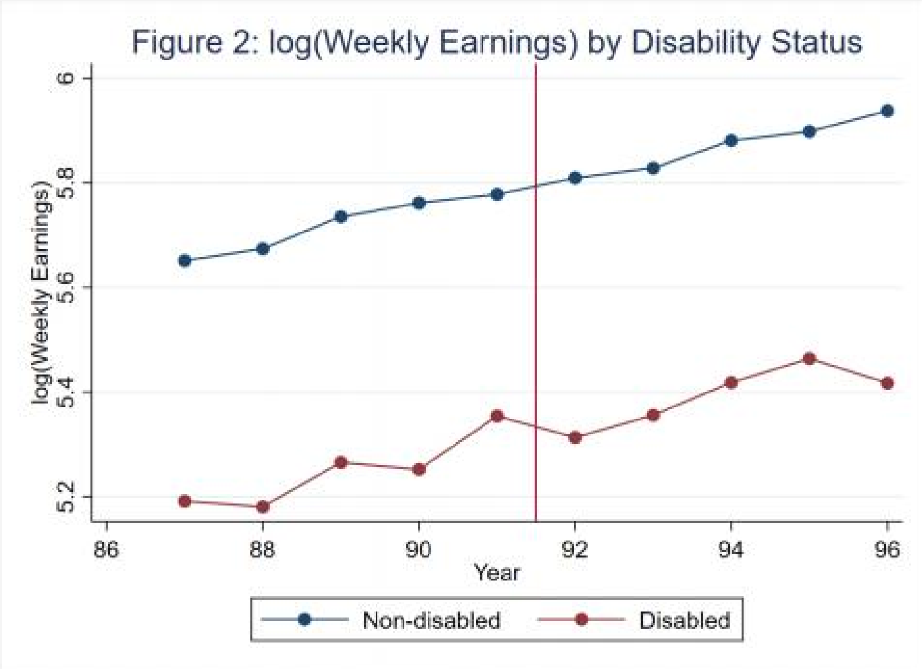
Empirical Exercise #1-- suggested solutions
Differences-in-Differences Empirical Exercise

1. *The parallel trends assumption.*
   a. State the parallel trends assumption in this empirical context, and explain how it could be violated.

   Answer: The parallel trends assumption states that absent the policy reform (implementation of the ADA), employment outcomes would have changed similarly across disabled and non-disabled individuals. It could be violated if other conditions were changing over time during this period (e.g., disability insurance benefits) that differentially affected disabled vs. non-disabled people.

   b. Replicate Figure 1 and Figure 2





c. Does it appear that the parallel trends assumption is violated for weeks worked? Does it appear that the parallel trends assumption is violated for log(earnings)? Explain.

Answer:

Weeks worked: There is some potential concern here that the parallel trends assumption is violated. The pre-treatment trend for disabled workers is noisier than that for non-disabled workers (although this is possibly because there are many fewer observations for disabled workers). The pre-treatment trend for disabled workers also seems to be on a downward trajectory whereas the pre-treatment trend for non-disabled workers appears to be basically flat.
Log weekly earnings: There is some potential concern here that the parallel trends assumption is violated – mostly because again the pre-treatment trend for disabled works in noisier than that for non-disabled workers (although this is possibly because we have fewer observations for disabled workers). However, it is reassuring that the earnings of both disabled and non-disabled workers appear to be on similar upward trajectories before the ADA was implemented.

2. *The basic difference-in-differences specification.*
   a. Run the following regression. Report and interpret the coefficient estimate for β_3 (the difference-in-differences estimator) and state whether it is statistically significant at the 5%-level. Specifically, what effect did the ADA have on weeks worked for the disabled?

$$wkswork_{it} = \beta_o + \beta_1 disabl1_i + \beta_2 post92_t + \beta_3 disabl1_i * post92_t + \varepsilon_{it} \qquad (1)$$

Answer:

$\widehat{\beta}_3$ = -2.49

Interpretation & statistical significance: Weeks worked by the disabled decreased by 2.49 weeks after the ADA was implemented relative to the non-disabled, and this estimate is significantly different from zero at the 5%-level since the ratio between the coefficient and its standard error is greater than 1.96. (A technical note is that one would probably like to cluster at some level, which could change the conclusion about the statistical significance).

b. Run the following regression. Report and interpret the coefficient estimate of (the difference-in-differences estimator) and state whether it is statistically significant at the 5%-level. Specifically, what effect did the ADA have on weekly earnings for the disabled?

$$lnwkwage_{it} = \beta_o + \beta_1 disabl1_i + \beta_2 post92_t + \beta_3 disabl1_i * post92_t + \varepsilon_{it} \qquad (2)$$

Answer:

$\widehat{\beta}_3$ = -0.0091

Interpretation & statistical significance: Weekly earnings by the disabled decreased by 0.9% after the ADA was implemented, relative to the non-disabled. With heteroskedasticity robust standard errors, this estimate is <u>not</u> significantly different from zero at the 5%-level. (A technical note is that one would probably like to cluster the standard errors at some level, which could change the conclusion about the statistical significance).

c. Assume that the regressions you ran in parts (a) and (b) have uncovered the causal effects of the ADA. Do your findings suggest that the ADA was helpful or harmful to employment outcomes for the disabled? Explain.

Answer: Our findings so far suggest that the ADA may have been harmful to employment outcomes for the disabled. There is evidence that disabled people were employed for fewer weeks per year as a result of the policy.

3. *Non-parametric difference-in-differences specification with separate interaction terms for each year.*

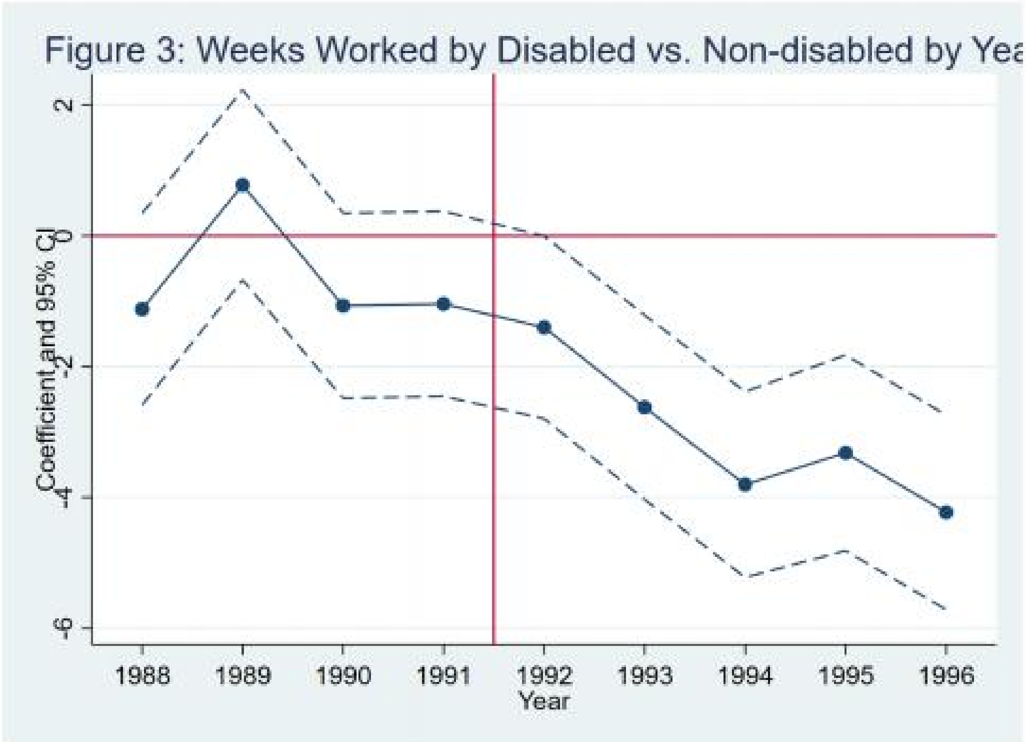
a. First, run a regression that takes equation (1) in question 2a and makes the following changes: 1) instead of including the variable *post92*, include a series of indicator variables (*y88*, *y89*, *y90* etc.) for the years 1988-1996, and 2) instead of including the variable $disabl1_i * post92_t$, include a series of interactions between disabled status and year (*disabl1_y88*, *disabl1_y89* etc.) for the years 1988-1996.

Note: 1987 is now the "excluded year," so all the interaction term coefficient estimates represent the change in weeks worked for disabled people relative to 1987.

See attached code.

b. Now, replicate Figure 3, which shows the coefficient estimates on the interaction terms (*disabl1_y88, disabl1_y89,* etc.) and their 95% confidence intervals.

See attached code.



c. Does there appear to be an immediate effect of the policy when it was implemented in 1992? What could explain what you observe?

Answer: No, there does not appear to be an immediate effect of the policy when it was implemented in 1992. Some potential reasons:

- It actually went into effect in July, halfway through the year.
- May have taken some time for employers to fully respond to the new regulation.

d. As a general matter, how can one use this graph to help assess the parallel trends assumption? What is the advantage of this type of graph over the raw averages plotted in Figure 1 and 2?

Answer: To assess whether the parallel trends assumption is plausible, we examine the coefficients on interaction terms for 1988-1991. These coefficients are "placebo" difference in difference estimates: if we pretend that the policy change happened in each of these years, would we have found a "zero" estimated effect? If the parallel trends assumption is plausible, then the answer should be yes.

The advantage of this type of graph compared with plotting raw averages is that we can assess the parallel trends assumption after adding additional controls. For example, one could control for age, education, gender, and race. We can be very flexible in the way that we control for these variables by defining indicator variables for each category. We can also interact these indicators with the year fixed effects, allowing the impact of each of the controls to be different in each year.

Stata Code

```
clear all
capture log close
version 16.1

/*
*** Install binscatter
ssc install binscatter

*** Install coefplot
ssc install coefplot
*/

*** Open the dataset
use "marcps_w.dta", clear

*** Replace year of survey with year for labor supply variable
replace year = year -1

*** Keep observations between ages 21 and 39
keep if age>=21 & age<=39

*** Create log(weekly earnings) variable
generate lnwkwage=log(wsal_val/wkswork)

*** QUESTION 1:

*** QUESTION 1b:

binscatter wkswork year, by(disabl1) line(connect) xline(91.5) ///
legend(label(1 "Non-disabled") label(2 "Disabled" )) ///
title("Figure 1: Weeks Worked by Disability Status") ///
ytitle("Weeks Worked") xtitle("Year")

graph export "Figure1.png", replace

binscatter lnwkwage year, by(disabl1) line(connect) xline(91.5) ///
legend(label(1 "Non-disabled") label(2 "Disabled" )) ///
title("Figure 2: log(Weekly Earnings) by Disability Status") ///
ytitle("log(Weekly Earnings)") xtitle("Year")

graph export "Figure2.png", replace


*** QUESTION 2:

*** QUESTION 2a:

generate post92 = (year >= 92)
generate disabl1_post92 = disabl1*post92
```

```
regress wkswork disabl1 post92 disabl1_post92, robust

*** QUESTION 2b:

regress lnwkwage disabl1 post92 disabl1_post92, robust


*** QUESTION 3:

*** generate year indicator variables (can also do this more efficiently with a for loop)
generate y88 = (year == 88)
generate y89 = (year == 89)
generate y90 = (year == 90)
generate y91 = (year == 91)
generate y92 = (year == 92)
generate y93 = (year == 93)
generate y94 = (year == 94)
generate y95 = (year == 95)
generate y96 = (year == 96)

*** generate year-diabl1 interaction terms (can also do this more efficiently with a for loop)
generate disabl1_y88 = disabl1*y88
generate disabl1_y89 = disabl1*y89
generate disabl1_y90 = disabl1*y90
generate disabl1_y91 = disabl1*y91
generate disabl1_y92 = disabl1*y92
generate disabl1_y93 = disabl1*y93
generate disabl1_y94 = disabl1*y94
generate disabl1_y95 = disabl1*y95
generate disabl1_y96 = disabl1*y96

*** run the regression
regress wkswork disabl1  y8* y9* disabl1_y8* disabl1_y9* , robust

*** create the coefficient plot
coefplot, recast(connected) ciopts(recast(rline) lpattern(dash)) vertical ///
keep(disabl1_y*) xline(4.5) yline(0) coeflabel(disabl1_y88  = "1988" ///
disabl1_y89  = "1989" ///
disabl1_y90  = "1990" disabl1_y91  = "1991" disabl1_y92  = "1992" ///
disabl1_y93  = "1993" disabl1_y94  = "1994" disabl1_y95  = "1995" ///
disabl1_y96  = "1996" disabl1_y97  = "1997") xtitle(Year) ///
ytitle("Coefficient and 95% CI") ///
title("Figure 3: Weeks Worked by Disabled vs. Non-disabled by Year")
graph export "Figure3.png", replace
```

R Code

```
#Gregory Bruich, Ph.D.
#Harvard University
#Send corrections and suggestions to gbruich@fas.harvard.edu

rm(list=ls()) # removes all objects from the environment
cat('\014') # clears the console

# Let's install a number of useful packages.
# To make things easy, the following snippet of code will download
# and install everything you'll need.
# But for future reference, remember that to install a package
# you only have to type
# > install.packages("<packagename>")
# And then you can load it with
# > library(lib)

packages <- c("haven"
              ,"ggplot2"
              ,"sandwich"
              ,"lmtest")

not_installed <- !packages %in% installed.packages()
if (any(not_installed)) install.packages(packages[not_installed])
lapply(packages,require,character.only=TRUE)
```

```
# Now all packages should be installed and loaded!

#Set working directory to be location of data
#Session -> set working directory -> choose working directory

#read in data
#File import data set -> from stata -> browse
marcps_w <- read_dta("marcps_w.dta")

#Data prep
marcps_w$year <- marcps_w$year - 1

#Subset data by age
df <- subset(marcps_w, age>=21 & age<=39)

#Define log weekly wage income worked
df$lnwkwage <- log(df$wsal_val/df$wkswork)
#eliminate log(0) entries which are -Infinity
df$lnwkwage[which(df$lnwkwage==-Inf)] <- NA

summary(df$disabl1)

#Replicate figure 1
ggplot(df, aes(x=year,y=wkswork, shape= factor(disabl1, labels = c("Non-disabled", "Disabled"))))
+
  geom_vline(xintercept=91.5) +
  stat_summary(fun.y = "mean",geom="point") +
  stat_summary(fun.y = "mean",geom="line")  +
  labs(x = "Year", y = "Weeks Worked", shape = "") +
  theme(legend.position="bottom")
ggsave("plot1.png")

#Replicate figure 2
ggplot(df, aes(x=year,y=lnwkwage, shape= factor(disabl1, labels = c("Non-disabled",
"Disabled")))) +
  geom_vline(xintercept=91.5) +
  stat_summary(fun.y = "mean",geom="point") +
  stat_summary(fun.y = "mean",geom="line")  +
  labs(x = "Year", y = "Log weekly wage", shape = "") +
  theme(legend.position="bottom")
ggsave("plot2.png")


#Generate post 19926 indicator variable
df$post <- 0
df$post[which(df$year >= 92)] <- 1

#Generate interation term
df$post_disabl1 <- df$disabl1*df$post

#Estimate regression 1
reg1 <- lm(wkswork ~ post + disabl1 + post_disabl1, data=df)
#Report homoskedastic standard errors
summary(reg1)
#Report heteroskedasticity robust standard errors
coeftest(reg1, vcov = vcovHC(reg1, type="HC1"))

#Estimate regression 2
reg2 <- lm(lnwkwage ~ post + disabl1 + post_disabl1, data=df)
coeftest(reg2, vcov = vcovHC(reg2, type="HC1"))

#Estimate regression 3
reg3 <- lm(wkswork ~ factor(year):disabl1 + factor(year) + disabl1 , data=df)
coeftest(reg3, vcov = vcovHC(reg3, type="HC1"))

years <- 1988:1996
beta <- reg3$coef[12:20]
se1 <-  sqrt(diag(vcovHC(reg3, type="HC1")))
se <- se1[12:20]
dfgraph = data.frame(years, beta, se)
```

```
years <- 1987
beta <- 0
se <- 0
df1987 = data.frame(years, beta, se)

forgraph <- rbind(dfgraph,df1987)
forgraph$ub <- forgraph$beta + (1.96*forgraph$se)
forgraph$lb <- forgraph$beta - (1.96*forgraph$se)

##Draw graph
ggplot(data=forgraph, aes(x=years)) +
  geom_errorbar(aes(ymin=lb, ymax=ub), width=.1, color="red") +
  geom_point(aes(y=beta)) +
  geom_line(aes(y=beta)) +
  labs(title = "",
       y = "Coefficient and 95% CI",
       x = "Year") +
  geom_hline(yintercept=0) +
  geom_vline(xintercept=1991.5)

#Save graph
ggsave("plot3.png")
```