Problem set 4: The EITC and diff-in-diff

Jamie Esmond

March 07, 2023

Contents

Introduction

In 1996, Nada Eissa and Jeffrey B. Liebman published a now-classic study on the effect of the Earned Income Tax Credit (EITC) on employment. The EITC is a special tax credit for low income workers that changes depending on (1) how much a family earns (the lowest earners and highest earners don't receive a huge credit, as the amount received phases in and out), and (2) the number of children a family has (more kids = higher credit). See this brief explanation for an interactive summary of how the EITC works.

Eissa and Liebman's study looked at the effects of the EITC on women's employment and wages after it was initially substantially expanded in 1986. The credit was expanded substantially again in 1993. For this problem set, you'll measure the causal effect of this 1993 expansion on the employment levels and annual income for women.

A family must have children in order to quality for the EITC, which means the presence of 1 or more kids in a family assigns low-income families to the EITC program (or "treatment"). We have annual data on earnings from 1991–1996, and because the expansion of EITC occurred in 1993, we also have data both before and after the expansion. This treatment/control before/after situation allows us to use a difference-in-differences approach to identify the causal effect of the EITC.

The dataset I've provided (`eitc.dta`) is a Stata data file containing more than 13,000 observations. This is non-experimental data—the data comes from the US Census's Current Population Survey (CPS) and includes all women in the CPS sample between the ages of 20–54 with less than a high school education between 1991–1996. There are 11 variables:

- `state`: The woman's state of residence. The numbers are Census/CPS state numbers: http://unionstats.gsu.edu/State_Code.htm
- `year`: The tax year
- `urate`: The unemployment rate in the woman's state of residence
- `children`: The number of children the woman has
- `nonwhite`: Binary variable indicating if the woman is not white (1 = Hispanic/Black)
- `finc`: The woman's family income in 1997 dollars
- `earn`: The woman's personal income in 1997 dollars
- `age`: The woman's age
- `ed`: The number of years of education the woman has
- `unearn`: The woman's family income minus her personal income, in *thousands* of 1997 dollars

```
library(tidyverse)  # For ggplot, %>%, mutate, filter, group_by, and friends
library(haven)      # For loading data from Stata
library(broom)      # For showing models as data frames
library(modelsummary)
library(kableExtra)

# This turns off this message that appears whenever you use summarize():
# `summarise()` ungrouping output (override with `.groups` argument)
options(dplyr.summarise.inform = FALSE)

# Load EITC data
eitc <- read_stata("data/eitc.dta") %>%
  # case_when() is a fancy version of ifelse() that takes multiple conditions
  # and outcomes. Here, we make a new variable named children_cat(egorical)
  # with three different levels: 0, 1, and 2+
  mutate(children_cat = case_when(
    children == 0 ~ "0",
    children == 1 ~ "1",
```

```
    children >= 2 ~ "2+"
  ))
```

1. Exploratory data analysis

Create a new variable that shows if women have 0 children, 1 child, or 2+ children (I did this for you already above).

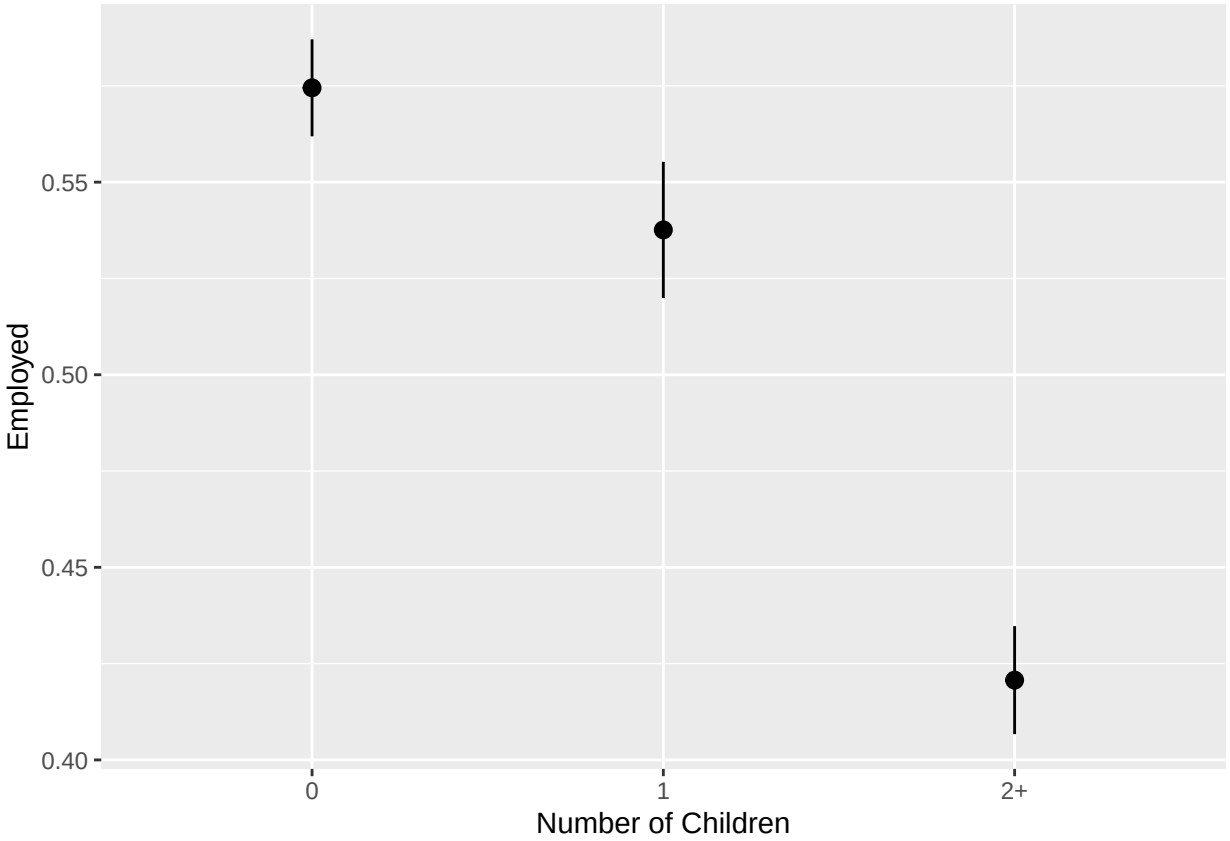
What is the average of `work`, `finc`, `earn`, `nonwhite`, `ed`, and `age` across each of these different levels of children? How are these groups different? Describe your findings in a paragraph.

Work

```
# Work
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(work))
```

```
## # A tibble: 3 x 2
##   children_cat avg_work
##   <chr>           <dbl>
## 1 0               0.574
## 2 1               0.538
## 3 2+              0.421
```

```
# stat_summary() here is a little different from the geom_*() layers you've seen
# in the past. stat_summary() takes a function (here mean_se()) and runs it on
# each of the children_cat groups to get the average and standard error. It then
# plots those with geom_pointrange. The fun.args part of this lets us pass an
# argument to mean_se() so that we can multiply the standard error by 1.96,
# giving us the 95% confidence interval
ggplot(eitc, aes(x = children_cat, y = work)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of Children",
       y = "Employed")
```

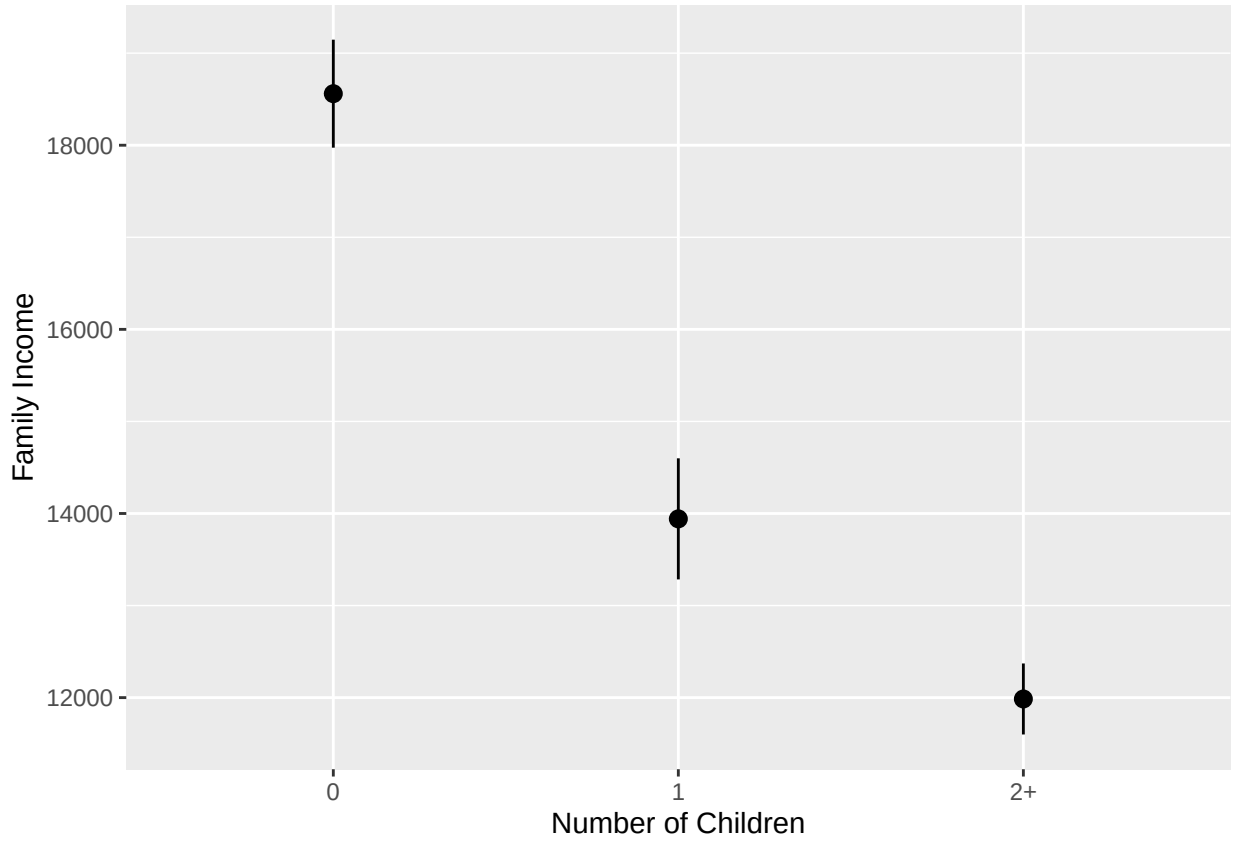

Employed
0.55
0.50
0.45
0.40
0
1
2+
Number of Children

Family income

```
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(finc))
```

```
## # A tibble: 3 x 2
##   children_cat avg_work
##   <chr>           <dbl>
## 1 0              18560.
## 2 1              13942.
## 3 2+             11985.
```

```
ggplot(eitc, aes(x = children_cat, y = finc)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of Children",
       y = "Family Income")
```

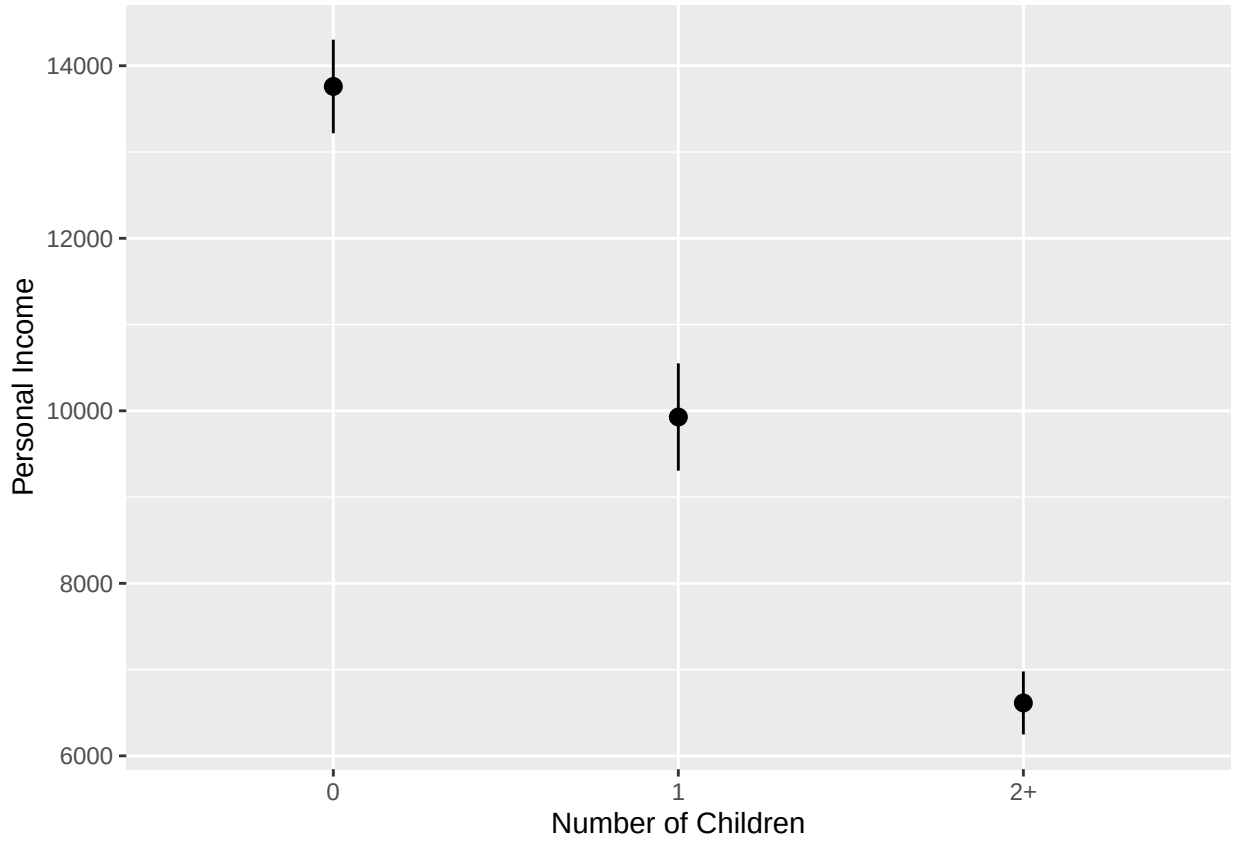


Earnings

```
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(earn))
```

```
## # A tibble: 3 x 2
##   children_cat avg_work
##   <chr>           <dbl>
## 1 0              13760.
## 2 1               9928.
## 3 2+              6614.
```

```
ggplot(eitc, aes(x = children_cat, y = earn)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of Children",
       y = "Personal Income")
```

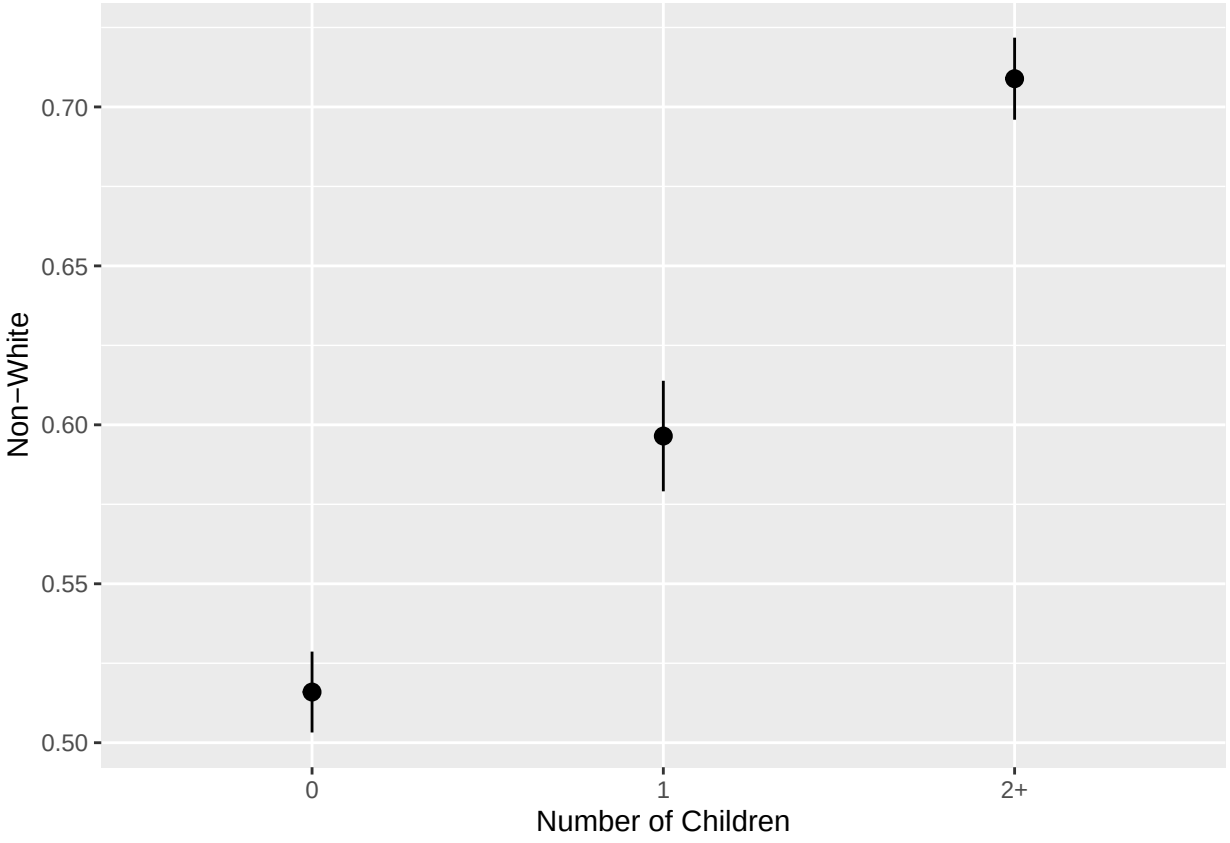


Race

```
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(nonwhite))
```

```
## # A tibble: 3 x 2
##   children_cat avg_work
##   <chr>           <dbl>
## 1 0               0.516
## 2 1               0.596
## 3 2+              0.709
```

```
ggplot(eitc, aes(x = children_cat, y = nonwhite)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of Children",
       y = "Non-White")
```

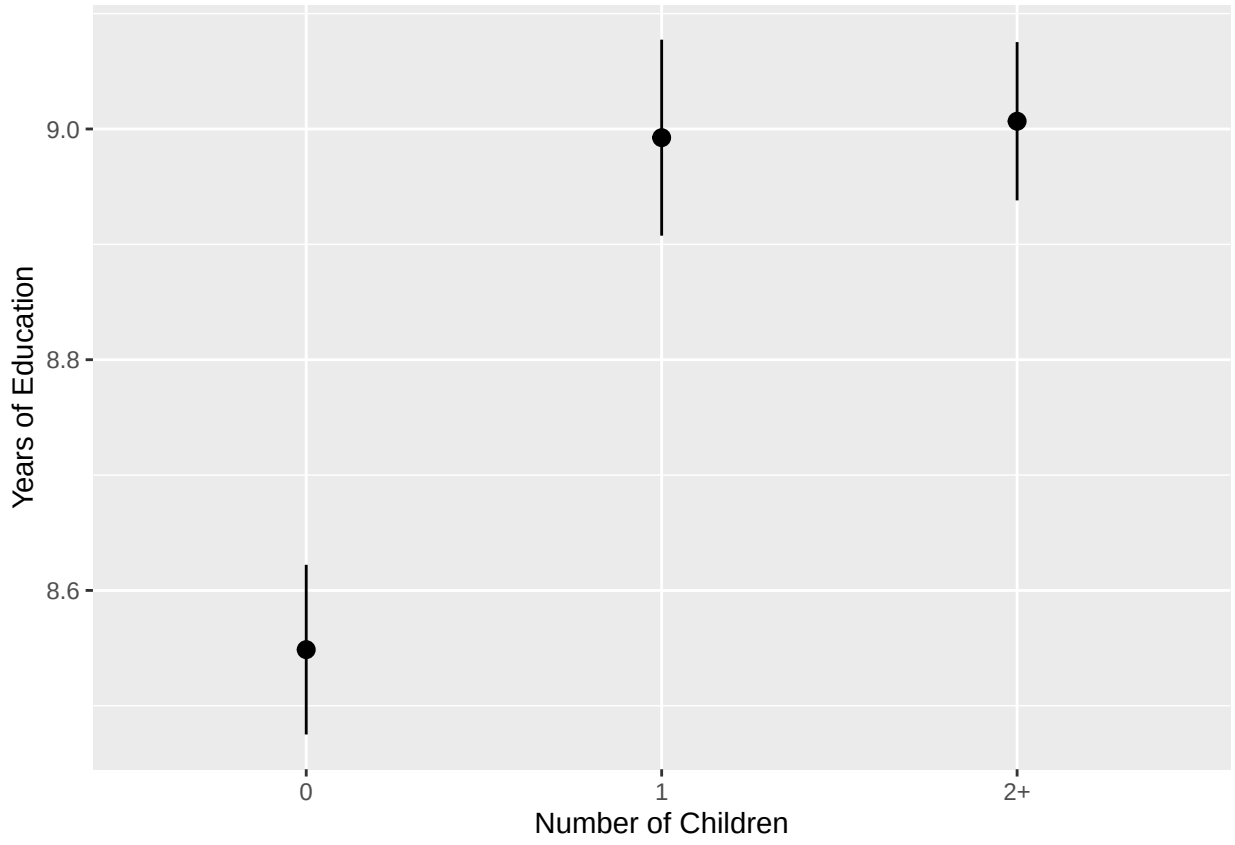


Education

```
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(ed))
```

```
## # A tibble: 3 x 2
##   children_cat avg_work
##   <chr>           <dbl>
## 1 0                8.55
## 2 1                8.99
## 3 2+               9.01
```

```
ggplot(eitc, aes(x = children_cat, y = ed)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of Children",
       y = "Years of Education")
```

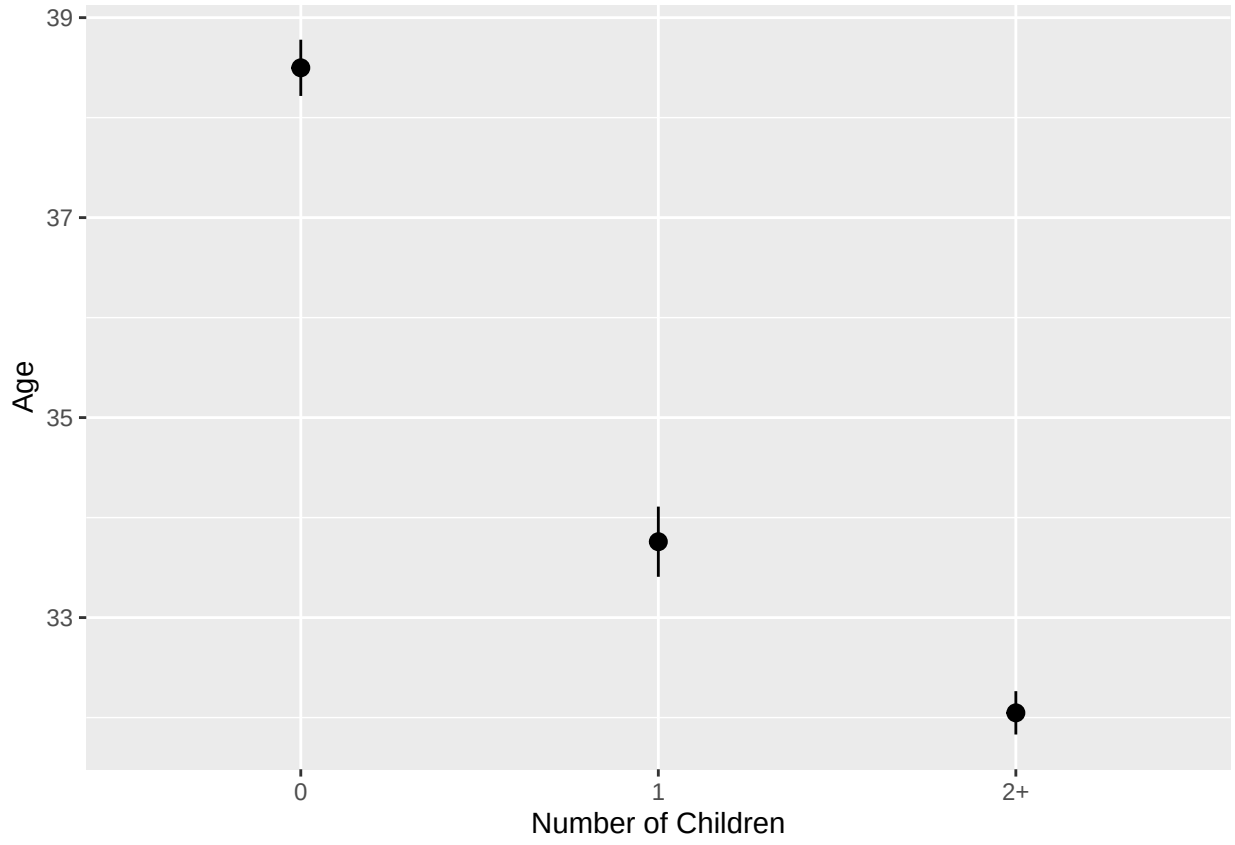


Age

```
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(age))
```

```
## # A tibble: 3 x 2
##   children_cat avg_work
##   <chr>           <dbl>
## 1 0                38.5
## 2 1                33.8
## 3 2+               32.0
```

```
ggplot(eitc, aes(x = children_cat, y = age)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of Children",
       y = "Age")
```



General summary

Describe your findings in a paragraph. How do these women differ depending on the number of kids they have?

Women in the sample are less likely to be employed if they have children, the difference is even greater if she has more than one child. These women differ economically depending on the number children they have. Both family and person income is much higher for women without children, and increasingly lower for more than one child. Women with more than one child are much more likely to be non-white. There is almost a 10 percentage point difference for white and non-white women between those with no children, with one child, and with more than one child. A woman having children is correlated with more years of education, but the number of children does not influence the level of education. Women with no children tend to be older than women with children; women with more than one child are even younger on average.

2. Create treatment variables

Create a new variable for treatment named `any_kids` (should be TRUE or 1 if `children` > 0) and a variable for the timing named `after_1993` (should be TRUE or 1 if `year` > 1993).

Remember you can use the following syntax for creating a new binary variable based on a test:

```
new_dataset <- original_dataset %>%
  mutate(new_variable = some_column > some_number)
```

```
# Make new dataset here.

eitc1 <- eitc %>%
  mutate(any_kids = children > 0,
         after_1993 = year > 1993)
```

3. Check pre- and post-treatment trends

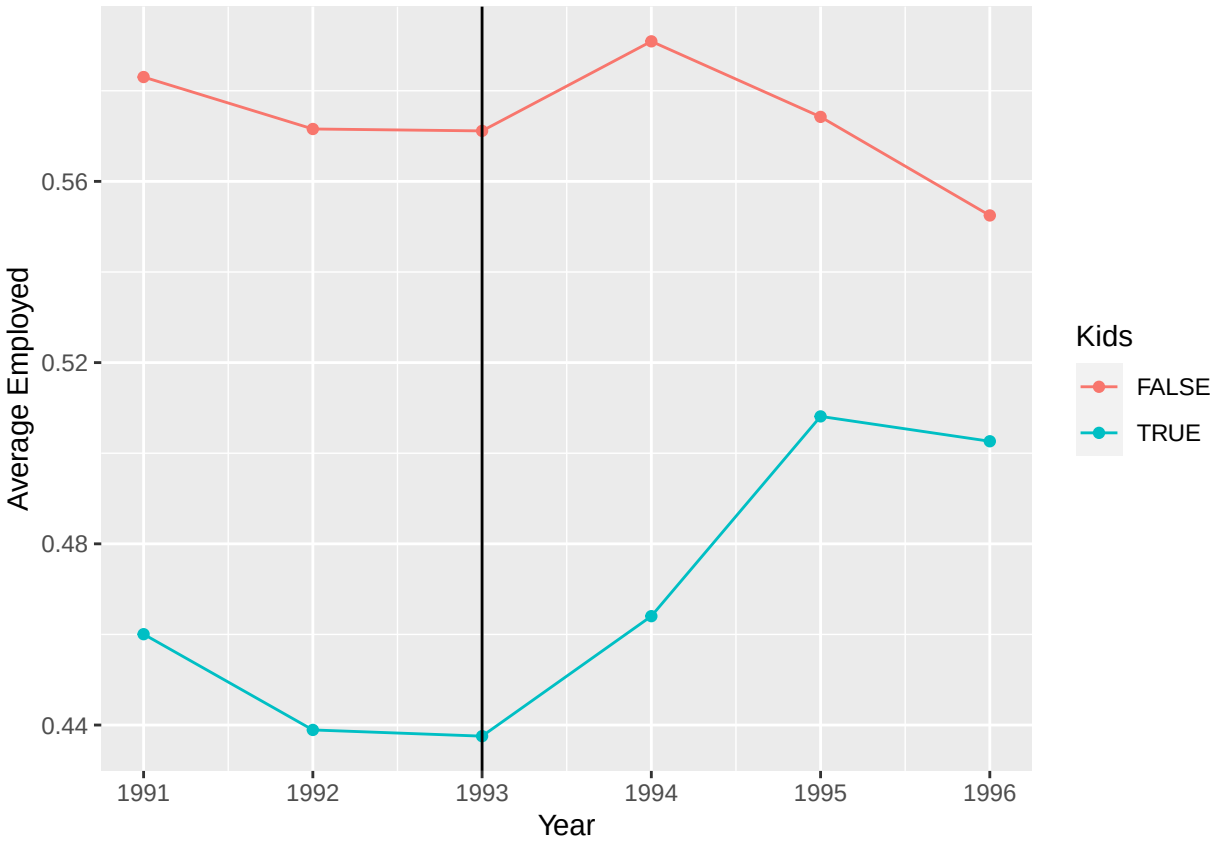
Create a new dataset that shows the average proportion of employed women (`work`) for every year in both the treatment and control groups (i.e. both with and without kids). (Hint: use `group_by()` and `summarize()`, and group by both `year` and `any_kids`.)

```
# Find average of work across year and any_kids
# Store this as a new object and then print it, like so:
#
eitc_by_year_kids <- eitc1 %>%
  group_by(year, any_kids) %>%
  summarise(workmean = mean(work))

print(eitc_by_year_kids)
```

```
## # A tibble: 12 x 3
## # Groups:   year [6]
##     year any_kids workmean
##    <dbl> <lgl>       <dbl>
##  1  1991 FALSE       0.583
##  2  1991 TRUE        0.460
##  3  1992 FALSE       0.572
##  4  1992 TRUE        0.439
##  5  1993 FALSE       0.571
##  6  1993 TRUE        0.438
##  7  1994 FALSE       0.591
##  8  1994 TRUE        0.464
##  9  1995 FALSE       0.574
## 10  1995 TRUE        0.508
## 11  1996 FALSE       0.552
## 12  1996 TRUE        0.503
```

Plot these trends using colored lines and points, with year on the x-axis, average employment on the y-axis. Add a vertical line at 1994 (hint: use `geom_vline(xintercept = SOMETHING)`.

```
# Add plot here, with x = year, y = average employment, and color = any_kids.
# Add a vertical line too.

eitc_by_year_kids %>%
  group_by(any_kids) %>%
  ggplot(aes(year, workmean, color = any_kids)) +
    geom_line() +
    geom_point() +
    geom_vline(xintercept = 1993) +
    labs(x = "Year",
         y = "Average Employed",
         color = "Kids")
```



Do the pre-treatment trends appear to be similar?

Yes, they both decrease in 1992 and held steady in 1993.

4. Difference-in-difference by hand-ish

Calculate the average proportion of employed women in the treatment and control groups before and after the EITC expansion. (Hint: group by `any_kids` and `after_1993` and find the average of `work`.)

```
# Calculate average of work across any_kids and after_1993

eitc1treatwork <- eitc1 %>%
  group_by(any_kids, after_1993) %>%
  summarise(meanwork = mean(work))

print(eitc1treatwork)
```

```
## # A tibble: 4 x 3
## # Groups:   any_kids [2]
##   any_kids after_1993 meanwork
##   <lgl>    <lgl>         <dbl>
## 1 FALSE    FALSE         0.575
## 2 FALSE    TRUE          0.573
## 3 TRUE     FALSE         0.446
## 4 TRUE     TRUE          0.491
```

Calculate the difference-in-difference estimate given these numbers. (Recall from class that each cell has a letter (A, B, C, and D), and that the diff-in-diff estimate represents a special combination of these cells.)

```
# It might be helpful to pull these different cells out with filter() and pull()
# like in the in-class examples from 8. Store these as objects like cell_A,
# cell_B, etc. and do the math here (like cell_B - cell_A, etc.)

A <- eitc1treatwork %>%
  filter(any_kids == FALSE, after_1993 == FALSE) %>%
  pull(meanwork)
A * 100
```

```
## [1] 57.54597
```

```
B <- eitc1treatwork %>%
  filter(any_kids == FALSE, after_1993 == TRUE) %>%
  pull(meanwork)
B * 100
```

```
## [1] 57.33862
```

```
C <- eitc1treatwork %>%
  filter(any_kids == TRUE, after_1993 == FALSE) %>%
  pull(meanwork)
C * 100
```

```
## [1] 44.59619
```

```r
D <- eitc1treatwork %>%
  filter(any_kids == TRUE, after_1993 == TRUE) %>%
  pull(meanwork)
D * 100
```

```
## [1] 49.07615
```

```r
nokids_diff <- B - A
kids_diff <- D - C

diff <- kids_diff - nokids_diff

diff * 100
```

```
## [1] 4.687313
```

	Before 1993	After 1993	Difference
Women with no kids	57.54597	57.33862	-2.0735
Women with kids	44.59619	49.07615	4.4799
Difference			**4.6873**

What is the difference-in-difference estimate? Discuss the result.

The difference-in-difference estimate is a 4.7 percentage point increase in employment for women with children. The employment rate of women without children decreased slightly, and the employment rate for women with children rose by 4.5 percentage points creating a difference of 4.7 percentage points between the treatment (women with children) and control (women without children) group from before and after the program was implemented.

5. Difference-in-difference with regression

Run a regression model to find the diff-in-diff estimate of the effect of the EITC on employment (`work`) (hint: remember that you'll be using an interaction term).

```
# Regression model here

model <- lm(work ~ any_kids * after_1993,
            data = eitc1)

tidy(model)
```

```
## # A tibble: 4 x 5
##   term                          estimate std.error statistic  p.value
##   <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                     0.575     0.00885    65.1   0
## 2 any_kidsTRUE                   -0.129     0.0117    -11.1   1.84e-28
## 3 after_1993TRUE                 -0.00207   0.0129     -0.160 8.73e- 1
## 4 any_kidsTRUE:after_1993TRUE     0.0469    0.0172      2.73  6.31e- 3
```

How does this value compare with what you found in part 4 earlier? What is the advantage of doing this instead of making a 2x2 table?

The regression method is much simpler than making the 2x2 table. The regression does all the math for you and gives you the same result.

6. Difference-in-difference with regression and controls

Run a new regression model with demographic controls. Eissa and Liebman used the following in their original study: non-labor income (family income minus personal earnings, or the `unearn` column), number of children, race, age, age squared, education, and education squared. You'll need to make new variables for age squared and education squared. (These are squared because higher values of age and education might have a greater effect: someone with 4 years of education would have 16 squared years, while someone with 8 years (twice as much) would have 64 squared years (way more than twice as much).)

```
# Make new dataset with columns for age squared and education squared

eitc2 <- eitc1 %>%
  mutate(agesq = age^2,
         edsq = ed^2)

# Regression model with demographic controls here

model1 <- lm(work ~ unearn + children + nonwhite + age +
               agesq + ed + edsq + any_kids * after_1993,
             data = eitc2)

tidy(model1)
```

```
## # A tibble: 11 x 5
##    term                          estimate std.error statistic   p.value
##    <chr>                            <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                    0.0611   0.0600        1.02  3.08e-  1
##  2 unearn                        -0.0178   0.000576    -30.8   5.26e-202
##  3 children                      -0.0518   0.00452     -11.5   2.93e- 30
##  4 nonwhite                      -0.0628   0.00848      -7.40  1.40e- 13
##  5 age                            0.0301   0.00327       9.22  3.49e- 20
##  6 agesq                         -0.000377 0.0000449    -8.39  5.23e- 17
##  7 ed                            -0.00398  0.00595      -0.669 5.03e-  1
##  8 edsq                           0.00142  0.000434      3.26  1.11e-  3
##  9 any_kidsTRUE                  -0.0210   0.0147       -1.43  1.52e-  1
## 10 after_1993TRUE                -0.00868  0.0123       -0.704 4.82e-  1
## 11 any_kidsTRUE:after_1993TRUE    0.0581   0.0164        3.55  3.81e-  4
```

Does the treatment effect change? Interpret these findings.

The treatment effect changed from 0.047 to 0.058, or from a 4.7 percentage point change to a 5.8 percentage point change. When controlling for these characteristics, the impact may be increased.

```
my_gof <- tribble(
  ~raw, ~clean, ~fmt,
  "nobs", "N", 0,
  "adj.r.squared", "R2", 2)

modelsummary(list("Simple" = model, "Full" = model1),
             gof_map = my_gof) %>%
    row_spec(7, background = "#f7fabe")
```

	Simple	Full
(Intercept)	0.575	0.061
	(0.009)	(0.060)
any_kidsTRUE	−0.129	−0.021
	(0.012)	(0.015)
after_1993TRUE	−0.002	−0.009
	(0.013)	(0.012)
any_kidsTRUE × after_1993TRUE	0.047	0.058
	(0.017)	(0.016)
unearn		−0.018
		(0.001)
children		−0.052
		(0.005)
nonwhite		−0.063
		(0.008)
age		0.030
		(0.003)
agesq		0.000
		(0.000)
ed		−0.004
		(0.006)
edsq		0.001
		(0.000)
N	13 746	13 746
R2	0.01	0.10

7. Varying treatment effects

Make two new binary indicator variables showing if the woman has one child or not and two children or not. Name them `one_kid` and `two_plus_kids` (hint: use `mutate(BLAH = children == SOMETHING)`.

```
# Make new dataset with one_kid and two_plus_kids indicator variables

eitc3 <- eitc2 %>%
  mutate(one_kid = children == 1,
         two_plus_kids = children >= 2)
```

Rerun the regression model from part 6 (i.e. with all the demographic controls), but remove the `any_kids` and `any_kids * after_1993` terms and replace them with two new interaction terms: `one_kid * after_1993` and `two_plus_kids * after_1993`.

```
# Run regression with both of the new interaction terms instead of
# any_kids * after_1993

model2 <- lm(work ~ unearn + children + nonwhite + age +
                    agesq + ed + edsq +
                    (one_kid * after_1993) +
                    (two_plus_kids * after_1993),
             data = eitc3)

tidy(model2)
```

```
## # A tibble: 13 x 5
##    term                              estimate std.error statistic   p.value
##    <chr>                                <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                        0.0610   0.0600        1.02  3.09e-  1
##  2 unearn                            -0.0178   0.000576    -30.8   5.88e-202
##  3 children                          -0.0526   0.00649      -8.12  5.14e- 16
##  4 nonwhite                          -0.0627   0.00848      -7.39  1.57e- 13
##  5 age                                0.0301   0.00327       9.21  3.62e- 20
##  6 agesq                             -0.000377 0.0000449    -8.39  5.51e- 17
##  7 ed                                -0.00405  0.00595      -0.681 4.96e-  1
##  8 edsq                               0.00142  0.000434      3.28  1.05e-  3
##  9 one_kidTRUE                       -0.0144   0.0159       -0.904 3.66e-  1
## 10 after_1993TRUE                    -0.00868  0.0123       -0.704 4.81e-  1
## 11 two_plus_kidsTRUE                 -0.0222   0.0222       -1.00  3.16e-  1
## 12 one_kidTRUE:after_1993TRUE         0.0438   0.0211        2.07  3.84e-  2
## 13 after_1993TRUE:two_plus_kidsTRUE   0.0674   0.0185        3.64  2.70e-  4
```

For which group of women is the EITC treatment the strongest for (i.e. which group sees the greatest change in employment)? Why do you think that is?

The change is stronger for women with 2 or more children. Women with only one child saw employment increase by 4.4 percentage points, and women with more than one children saw an increase of 6.7 percentage points.

Women with more than one child may have a harder time finding and retaining employment because child care is more expensive and arguably exponentially more demanding than only one child. The relief offered to these mothers is more impactful because they are more in need of assistance. This group was the least likely to be employed before the treatment, therefore then say the most gains.

8. Check parallel trends with fake treatment

To make sure this effect isn't driven by any pre-treatment trends, we can pretend that the EITC was expanded in 1991 (starting in 1992) instead of 1993.

Create a new dataset that only includes data from 1991–1993 (hint: use `filter()`). Create a new binary before/after indicator named `after_1991` (hint: `year >= 1992`). Use regression to find the diff-in-diff estimate of the EITC on `work` (don't worry about adding demographic controls).

```
# Make new dataset that only includes rows less than 1994 (with filter), and add
# a new binary indicator variable for after_1991

eitc4 <- eitc3 %>%
  filter(year < 1994) %>%
  mutate(after_1991 = year >= 1992)

# Run simple regression with interaction term any_kids * after_1991

model3 <- lm(work ~ any_kids * after_1991,
             data = eitc4)

tidy(model3)
```

```
## # A tibble: 4 x 5
##   term                          estimate std.error statistic   p.value
##   <chr>                            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                     0.583     0.0149    39.1   1.34e-304
## 2 any_kidsTRUE                   -0.123     0.0196    -6.26  4.02e- 10
## 3 after_1991TRUE                 -0.0117    0.0185    -0.631 5.28e-  1
## 4 any_kidsTRUE:after_1991TRUE    -0.0101    0.0244    -0.415 6.78e-  1
```

Is there a significant diff-in-diff effect? What does this mean for pre-treatment trends?

There is not a significant diff-in-diff effect. This means that the treatment and control group had similar trends before the treatment began, and were likely to continue that trend without intervention.