The EITC and diff-in-diff

Problem set 4 — PMAP 8521, Spring 2025

Tamara Mallory

March 13, 2025

Table of contents

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

```r
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

# 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).

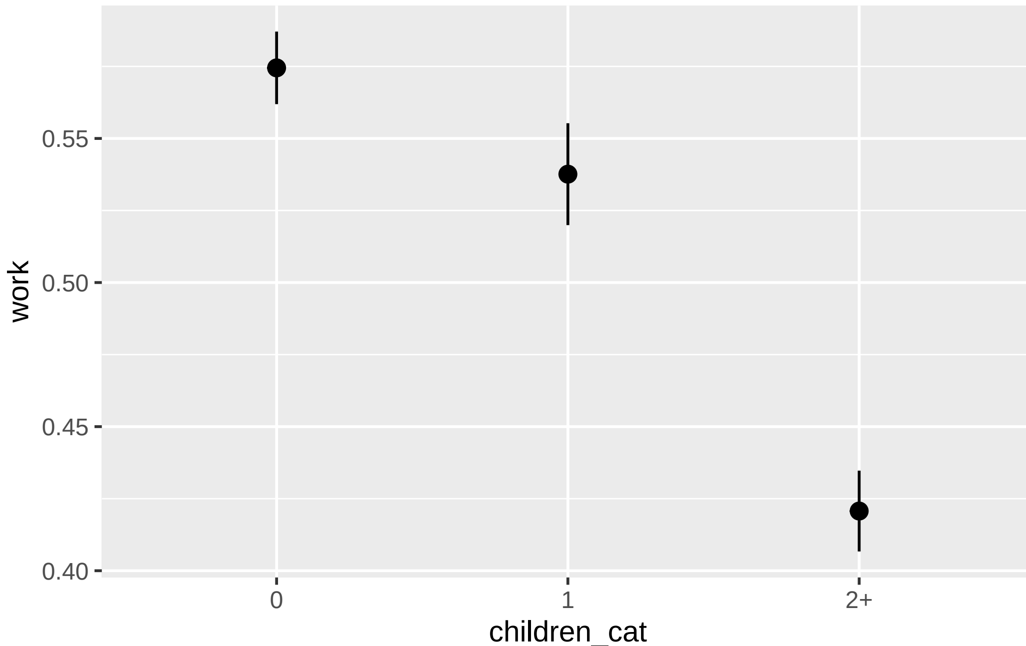
What is the average of `work`, `finc`, `earn`, `nonwhite`, `ed`, and `age` across each of these different levels of children? How are these groups different? Describe your findings in a paragraph.

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))
```

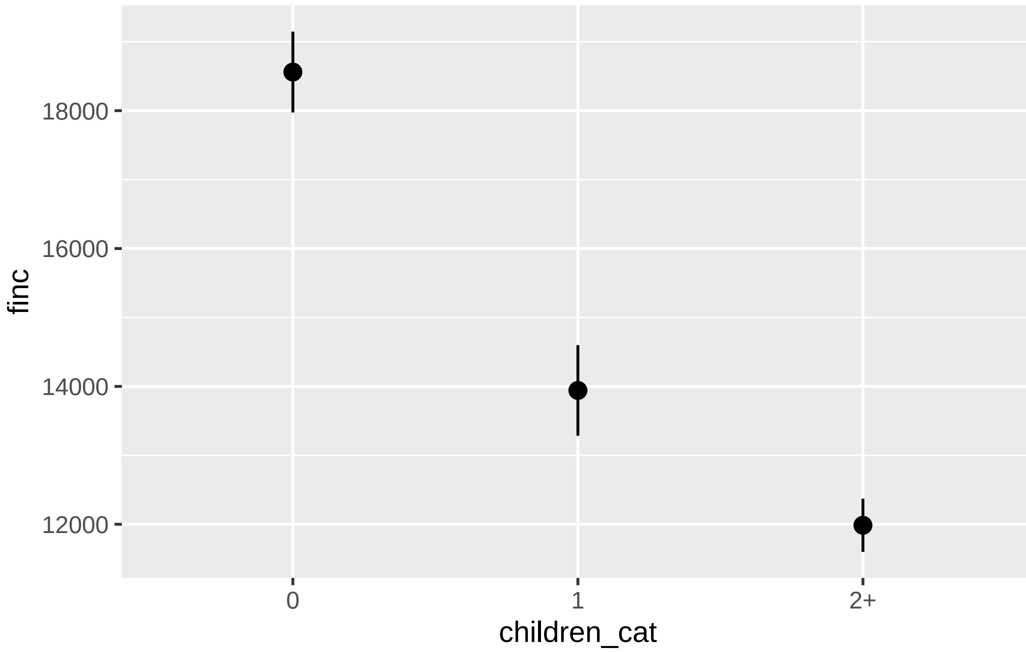


Family income

```
eitc |>
  group_by(children_cat) |>
  summarize(avg_finc = mean(finc))
```

```
# A tibble: 3 x 2
  children_cat avg_finc
  <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))
```

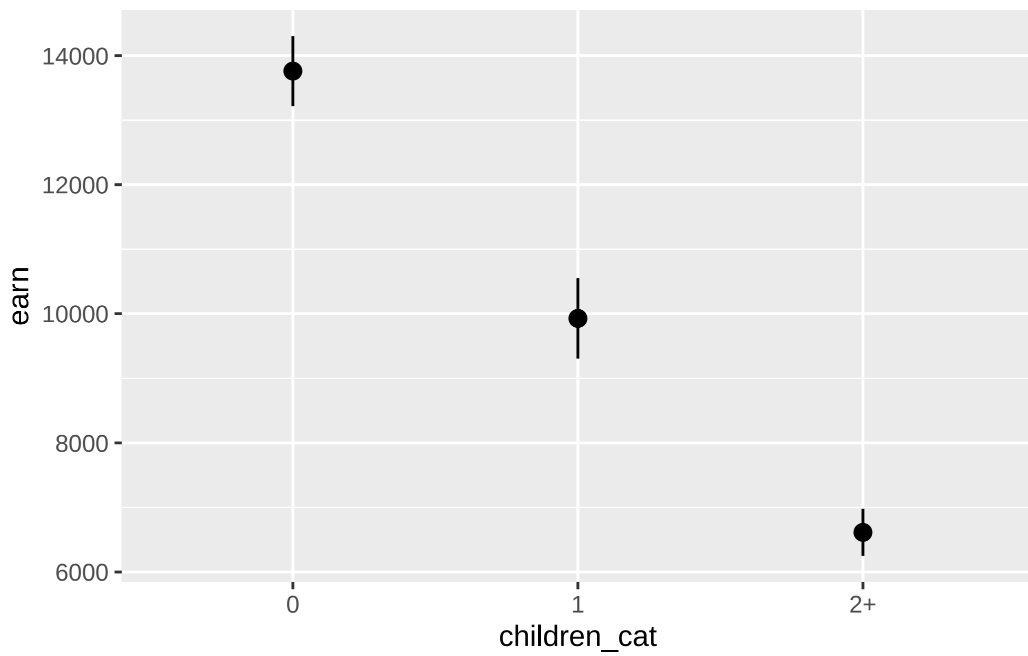


Earnings

```
eitc |>
  group_by(children_cat) |>
  summarize(avg_earn = mean(earn))
```

```
# A tibble: 3 x 2
  children_cat avg_earn
  <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))
```

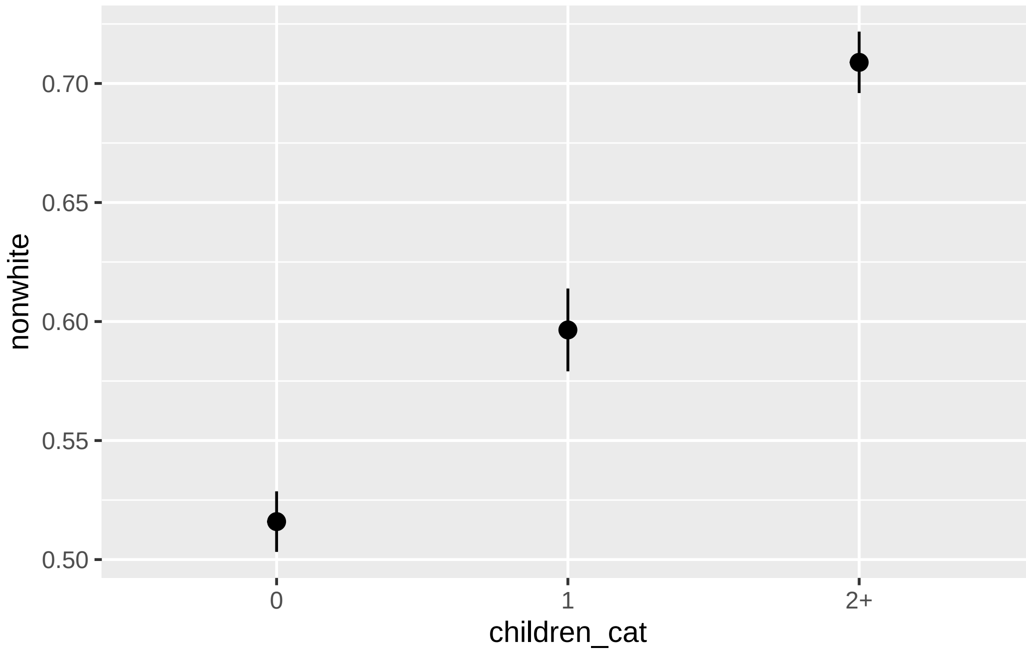


Race

```
eitc |>
  group_by(children_cat) |>
  summarize(avg_race = mean(nonwhite))
```

```
# A tibble: 3 x 2
  children_cat avg_race
  <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))
```

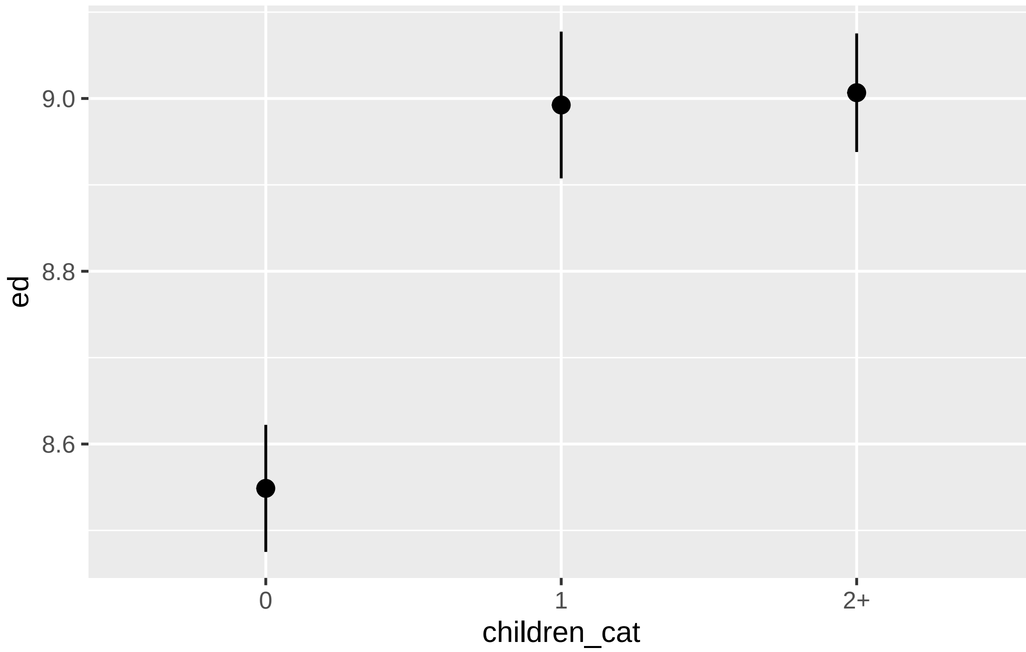


Education

```
eitc |>
  group_by(children_cat) |>
  summarize(avg_education = mean(ed))
```

```
# A tibble: 3 x 2
  children_cat avg_education
  <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))
```

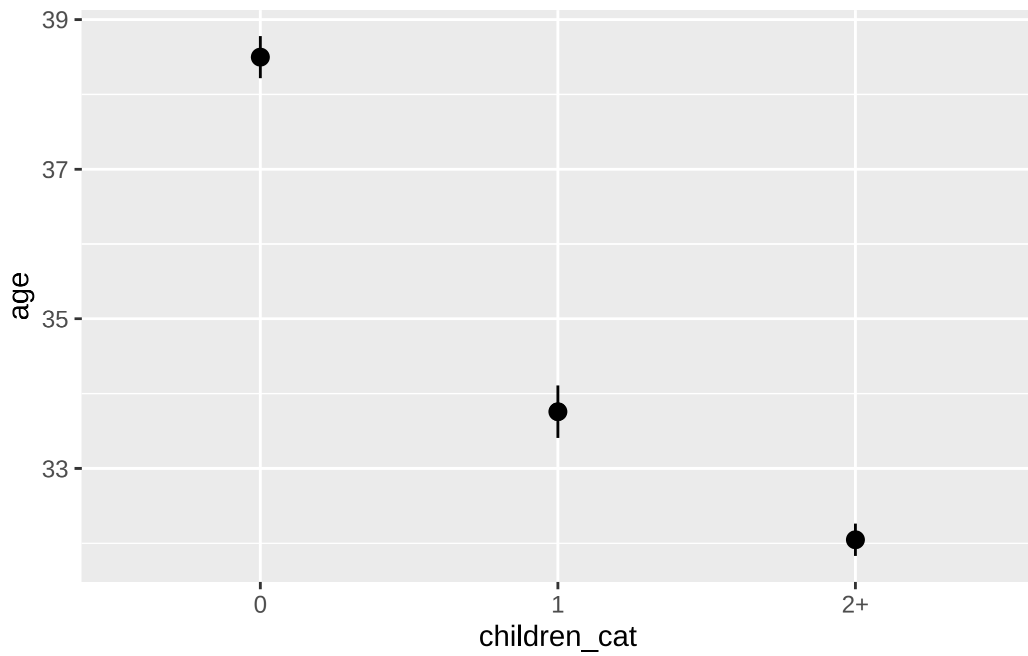


Age

```
eitc |>
  group_by(children_cat) |>
  summarize(avg_age = mean(age))
```

```
# A tibble: 3 x 2
  children_cat avg_age
  <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))
```



General summary

Describe your findings in a paragraph. How do these women differ depending on the number of kids they have? The women differ depending on the number of kids that they have. For average of women that were employed had 0 kids and as women had kids the less likely they were employed. Concerning family income, Women who had no kids had the highest family income and the women who had one kid had less family income and the income lessened more with 2+ kids. As the number of kids increased, the earned income for the women decreased. As the average of nonwhite women increase so did the number if kids the women had. As education increased, the number of kids woman had increased. As women aged the less amount of kids they had. At the age of 32 women had 2+ kids while at the age around 34 had 1 kid and the around the age of 39 had 0 kids.

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.

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. You can either do something like:
#
# eitc_new <- eitc |> whatever
#
# which would create a completely new data frame, or do something like:
#
# eitc <- eitc |> whatever
#
# which would overwrite the original eitc data frame with the modified one.
# Either approach is fine.

eitc_treatment_kids <- 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`.)

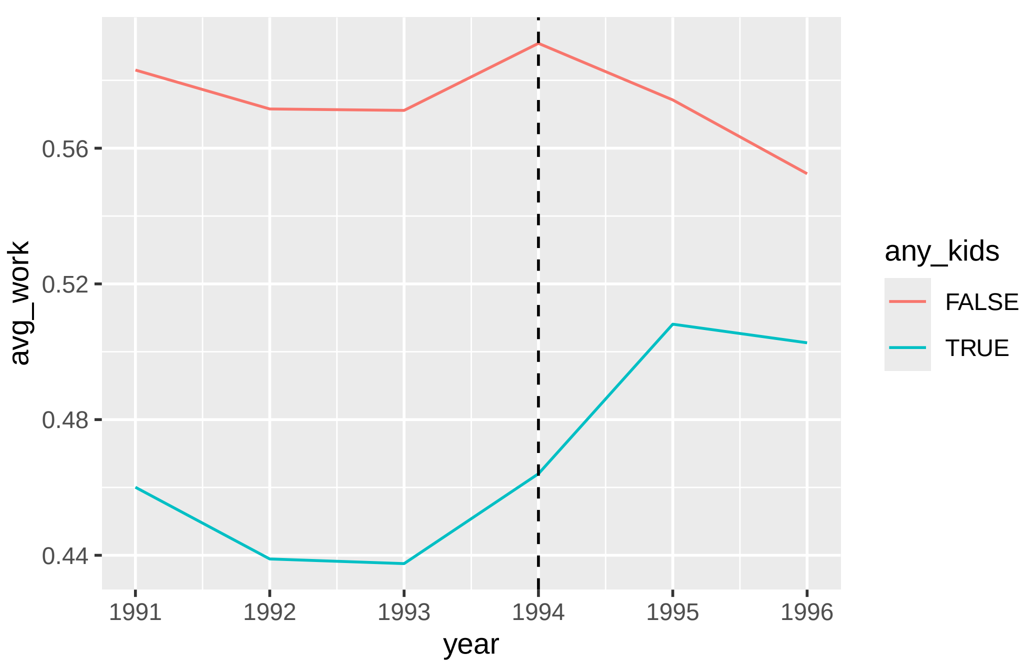
```
# Find average of work across year and any_kids
# Store this as a new object and then print it, like so:
#
# eitc_by_year_kids <- eitc |> whatever
# print(eitc_by_year_kids)

eitc_by_year_kids <- eitc_treatment_kids |>
  group_by(year,any_kids) |>
  summarize(avg_work = mean(work))
print(eitc_by_year_kids)
```

```
# A tibble: 12 x 3
# Groups:   year [6]
```

```
    year any_kids avg_work
   <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)`.

```r
# Add plot here, with x = year, y = average employment, and color = any_kids.
# Add a vertical line too.
ggplot(eitc_by_year_kids, aes(x = year, y =avg_work, color = any_kids ) )+
         geom_line()+
         geom_vline(xintercept = 1994, linetype = "dashed", color = "black" )
```



Do the pre-treatment trends appear to be similar? The pre-treatment trends appear to be similar according to the parallel trend assumption. Both displayed similar trends of decreasing then stabilizing then increasing up until 1994. After 1994 the treatment group continued increasing while the control group decreased.

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
eitc_treatment_kids |>
  group_by(after_1993,any_kids) |>
  summarize(avg_work = mean(work, na.rm = T))
```

```
# A tibble: 4 x 3
# Groups:   after_1993 [2]
  after_1993 any_kids avg_work
  <lgl>      <lgl>       <dbl>
1 FALSE      FALSE       0.575
2 FALSE      TRUE        0.446
3 TRUE       FALSE       0.573
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), an d 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.)

cell_A_before_control <- eitc_treatment_kids |>
  filter(after_1993 == 0, any_kids == 0) |>
  summarize(avg_work = mean(work))
print(cell_A_before_control)
```

```
# A tibble: 1 x 1
  avg_work
```

```
     <dbl>
1    0.575
```

```r
cell_B_before_treatment <- eitc_treatment_kids |>
  filter(after_1993 == 0, any_kids == 1) |>
  summarize(avg_work = mean(work))
print(cell_B_before_treatment)
```

```
# A tibble: 1 x 1
  avg_work
     <dbl>
1    0.446
```

```r
cell_C_after_control <- eitc_treatment_kids |>
  filter(after_1993 == 1 , any_kids == 0) |>
  summarize(avg_work = mean(work))
print(cell_C_after_control)
```

```
# A tibble: 1 x 1
  avg_work
     <dbl>
1    0.573
```

```r
cell_D_after_treatment <- eitc_treatment_kids |>
  filter(after_1993 == 1, any_kids == 1) |>
    summarize(avg_work = mean(work))
print(cell_D_after_treatment)
```

```
# A tibble: 1 x 1
  avg_work
     <dbl>
1    0.491
```

```r
diff_treatment_before_after <- cell_D_after_treatment - cell_B_before_treatment
diff_treatment_before_after
```

```
    avg_work
1 0.04479962
```

```r
diff_control_before_after <- cell_C_after_control - cell_A_before_control
diff_control_before_after
```

```
      avg_work
1 -0.002073509
```

```r
diff_diff <- diff_treatment_before_after - diff_control_before_after
diff_diff
```

```
    avg_work
1 0.04687313
```

	Before 1993	After 1993	Difference
Women with no kids			
Women with kids			
Difference			

What is the difference-in-difference estimate? Discuss the result. (Hint, these numbers are percents, so you can multiply them by 100 to make it easier to interpret. For instance, if the diff-in-diff number is 0.15 (it's not), you could say that the EITC caused the the proportion of mothers in the workplace to increase 15 percentage points.)

The EITC caused the proportion of mothers in the workplace to increase by 4.7%.

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).

```r
# Regression model here

model_eitc_program <- lm(work ~ any_kids + after_1993
                         + any_kids *after_1993, data = eitc_treatment_kids)

tidy(model_eitc_program)
```

```
# 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 value is the same compared to what was found earlier in part 4.The advantage of doing a regression than a 2x2 table is because a regression can include control variables to help isolate the effect and also can tell if a variable is statistically significantly.

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).)

```r
# Make new dataset with columns for age squared and education squared
# Regression model with demographic controls here
#
# R tends to put interaction terms last in regression tables, so you might not
# see the any_kids * after_1993 coefficient on the first page of the table here

eitc_squared <- eitc_treatment_kids |>
  mutate(age_squared = age^2,
         education_squared = unearn^2)
print(eitc_squared)
```

```
# A tibble: 13,746 x 16
   state  year urate children nonwhite   finc   earn   age    ed  work unearn
   <dbl> <dbl> <dbl>    <dbl>    <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
 1    11  1991  7.60        0        1 18714. 18714.    26    10     1  0
 2    12  1991  7.20        1        0  4839.   471.    22     9     1  4.37
```

```
 3    13  1991  6.40        2         0  8178.      0     33    11     0  8.18
 4    14  1991  9.10        0         1  9370.      0     43    11     0  9.37
 5    15  1991  8.60        3         1 14707. 14707.    23     7     1  0
 6    16  1991  6.80        1         0 21605. 18855.    53     7     1  2.75
 7    21  1991  7.30        0         1 19147. 14141.    52    11     1  5.01
 8    22  1991  6.70        0         1 64312. 63803.    51    11     1  0.509
 9    23  1991  7           1         1 17676. 17676.    20    11     1  0
10    31  1991  6.40        2         1 12214.  2358.    32    11     1  9.86
# i 13,736 more rows
# i 5 more variables: children_cat <chr>, any_kids <lgl>, after_1993 <lgl>,
#   age_squared <dbl>, education_squared <dbl>
```

```
model_squared <- lm(work ~ unearn + any_kids*after_1993 +
                      nonwhite + age + age_squared + ed + education_squared, data = eitc_squa
tidy(model_squared)
```

```
# A tibble: 10 x 5
   term                         estimate std.error statistic   p.value
   <chr>                           <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)                  0.105    0.0572         1.83 6.75e-  2
 2 unearn                      -0.0348   0.000880     -39.6  0
 3 any_kidsTRUE                -0.108    0.0115        -9.45 3.97e- 21
 4 after_1993TRUE              -0.0131   0.0121        -1.08 2.81e-  1
 5 nonwhite                    -0.0803   0.00829       -9.69 4.04e- 22
 6 age                          0.0276   0.00319        8.66 5.31e- 18
 7 age_squared                 -0.000332 0.0000438     -7.58 3.72e- 14
 8 ed                           0.0139   0.00155        9.01 2.28e- 19
 9 education_squared            0.000363 0.0000149     24.4  6.85e-129
10 any_kidsTRUE:after_1993TRUE  0.0611   0.0161         3.79 1.49e-  4
```

Does the treatment effect change? Interpret these findings. The treatment does effect change. The treatment causes a 6.10 percentage point increase in employment for women with children after 1993. The p-value is statistically significant which means that we can reject the null hypothesis that the treat did not have any effect.

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)`.

```r
# Make new dataset with one_kid and two_plus_kids indicator variables
eitc_counting_kids <- eitc_squared |>
mutate(
  one_kid = children == 1,
  two_plus_kids = children >= 2
)
print(eitc_counting_kids)
```

```
# A tibble: 13,746 x 18
   state  year urate children nonwhite   finc   earn   age    ed  work unearn
   <dbl> <dbl> <dbl>    <dbl>    <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
 1    11  1991  7.60        0        1 18714. 18714.    26    10     1  0
 2    12  1991  7.20        1        0  4839.   471.    22     9     1  4.37
 3    13  1991  6.40        2        0  8178.      0    33    11     0  8.18
 4    14  1991  9.10        0        1  9370.      0    43    11     0  9.37
 5    15  1991  8.60        3        1 14707. 14707.    23     7     1  0
 6    16  1991  6.80        1        0 21605. 18855.    53     7     1  2.75
 7    21  1991  7.30        0        1 19147. 14141.    52    11     1  5.01
 8    22  1991  6.70        0        1 64312. 63803.    51    11     1  0.509
 9    23  1991  7           1        1 17676. 17676.    20    11     1  0
10    31  1991  6.40        2        1 12214.  2358.    32    11     1  9.86
# i 13,736 more rows
# i 7 more variables: children_cat <chr>, any_kids <lgl>, after_1993 <lgl>,
#   age_squared <dbl>, education_squared <dbl>, one_kid <lgl>,
#   two_plus_kids <lgl>
```

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`.

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

model_new_terms <- lm(work ~ unearn + one_kid*after_1993 +
                      two_plus_kids*after_1993 +
                    nonwhite + age + age_squared + ed + education_squared, data = eitc_counti
tidy(model_new_terms)
```

```
# A tibble: 12 x 5
   term                              estimate std.error statistic   p.value
   <chr>                                <dbl>     <dbl>     <dbl>     <dbl>
```

```
 1 (Intercept)                        0.0677    0.0574          1.18 2.39e-  1
 2 unearn                            -0.0342    0.000884      -38.7  1.50e-311
 3 one_kidTRUE                       -0.0619    0.0143         -4.32 1.55e-  5
 4 after_1993TRUE                    -0.0131    0.0121         -1.08 2.80e-  1
 5 two_plus_kidsTRUE                 -0.144     0.0130        -11.1  2.69e- 28
 6 nonwhite                          -0.0750    0.00832        -9.02 2.11e- 19
 7 age                                0.0298    0.00320         9.30 1.56e- 20
 8 age_squared                       -0.000365  0.0000440      -8.29 1.23e- 16
 9 ed                                 0.0141    0.00154         9.13 7.84e- 20
10 education_squared                  0.000356  0.0000149      23.9  5.47e-124
11 one_kidTRUE:after_1993TRUE         0.0472    0.0208          2.27 2.32e-  2
12 after_1993TRUE:two_plus_kidsTRUE   0.0694    0.0182          3.82 1.35e-  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 women with 2+ kids had the greatest change in employment. I believe that this group of women had the greatest change in employment because mothers had more financial incentives to seek employment and were already employed because of financial responsibilities for the kids.

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).

```r
eitc_fake_treatment <- eitc_treatment_kids |>
  filter(year < 1994) |>
  mutate(after_1991 = year >= 1992)
# Make new dataset that only includes rows less than 1994 (with filter), and add
# a new binary indicator variable for after_1991

# Run simple regression with interaction term any_kids * after_1991
model_simple <- lm(work ~ any_kids*after_1991,
                   data = eitc_fake_treatment)
tidy(model_simple)
```

```
# 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 in this regression model. This means that the pre-treatment trends were not parallel or that the policy had no true effect.