ECON 453
In-Class Exercise 8
November 2, 2023

**ANSWER KEY**

Please download the file "IC8.gdt", a gretl data file. This dataset comes from the 2000 United States Census. During this exercise, we will replicate the methodology used in Dynarski (2008). The dataset includes information for people born in Alabama, Florida, Georgia, South Carolina, and Texas that were between the ages of 22 and 34 when the 2000 Census was conducted. The study that we are replicating uses a difference-in-difference methodology and analyzes the impacts of the Georgia Hope Scholarship program, which began for students that graduated high school in 1993.

1. Let's begin by conducting a descriptive analysis.
    a. Make a dummy variable for those born in Georgia (call it **georgia**).
        i. **Add -> Define new variable**: georgia=birth_state==3
    b. Make a dummy variable for those *not* born in Georgia (call it **control**).
        i. **Add -> Define new variable**: control=georgia==0
    c. Make a dummy variable for those that graduated high school in 1993 or more recently (call it **after**). We assume everyone graduates high school at age 18.
        i. **Add -> Define new variable**: after=age<26
    d. Make a dummy variable for those that graduated high school before 1993 (call it **before**).
        i. **Add -> Define new variable**: before=age>25
    e. Make a dummy variable for those that have *at least an associate degree* (call it **college**).
    f. **Add -> Define new variable**: college=degree>2
    g. Use sample restrictions to fill out the simple difference-in-difference table below. For each of the cells, find the percentage that have at least an associate degree (using summary statistics for your **college** variable).

***To do this, I went to* Sample -> Restrict, based on criterion, *then entered the command for each cell that I was trying to estimate. For example, to fill in the top left, I restricted as*** before==1 && georgia==1. ***Then I viewed the summary statistics for the college variable and recorded the mean. Make sure you restore the range after each time.***

| | **18 before 1993** | **18 after 1993** | **Difference** |
|---|---|---|---|
| **Born in Georgia** | .2515 | .1815 | -0.070 |
| **Not born in Georgia** | .2699 | .1686 | -0.1013 |
| **Difference** | N/A | N/A | **0.0313** |

    h. Discuss the what the results indicate about the potential impact of the scholarship program on the level of college attainment in Georgia.

***The results tell us that the program is estimated to increase the likelihood a person born in Georgia has some form of college degree by about 3.13 percentage points. Since the people graduating in 1993 and after are younger, they are less likely to have a degree in all states. However, based on the control states, we would have expected a 10-percentage point drop, and we only observed a 7-percentage point drop in Georgia. This is evidence that the program may have been effective in increasing educational attainment.***

2. Next, we want to formally estimate the effect with the regression form of the difference-in-difference model.
    a. Create an interaction term between your **georgia** and **after** variables.
        i. **Add -> Define new variable**: inter = georgia*after
    b. Run a regression using **college** as the dependent variable and **georgia, after**, and the **interaction** as the regressors.

```
Q2: OLS, using observations 1-53312
Dependent variable: college

            coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       0.269856     0.00238046   113.4     0.0000    ***
  georgia    -0.0183393    0.00563869    -3.252   0.0011    ***
  after      -0.101248     0.00455172   -22.24    4.07e-109 ***
  inter       0.0312529    0.0109753      2.848   0.0044    ***

Mean dependent var   0.240584   S.D. dependent var   0.427442
Sum squared resid    9641.278   S.E. of regression   0.425276
R-squared            0.010163   Adjusted R-squared   0.010108
```

c. Discuss what each of the coefficients tells us:

**Georgia:** ***The coefficient is negative and significant. It tells us that people born in Georgia have lower rates of educational attainment than the control states (by about 1.83 percentage points) before the program began. Notice that the constant gives us the percentage with at least some college in the control states before 1993 (same as table above). If we subtract the Georgia coefficient, we get the value in the upper left (Georgia, before) cell.***

**After:** ***This tells is that individuals in the "after" period are much less likely to have some form of college degree. This makes sense, since they are younger when surveyed in the year 2000. If we look at the table from question 1, this is the difference in the before and after period for the control group. Those graduating in 1993 and after are 10.13 percentage points less likely to have a degree than those graduating in these states before 1993.***

**Interaction:** ***This is the difference-in-difference estimate. It tells us how to adjust the "after" coefficient for the treatment group (those born in Georgia). This tells us the estimated impact of the scholarship program on educational attainment in Georgia.***

d. Are the results different from the simple analysis above? Does the regression improve our study?
***What you should notice is that the results are the same in the two analyses. The benefit to running the regression is that we can see the statistical significance of our findings. Also, we can do it in one step, as opposed to restricting and restoring the sample each time. The final advantage to using the regression model is that we can then control for other factors, such as the gender, race, and age of the individual. I'd like to see you try that in the summary table format!***

3. Run the model again, but this time add the following regressors: **female, black, Hispanic,** and **age**.

```
Q3: OLS, using observations 1-53312
Dependent variable: college

            coefficient   std. error    t-ratio    p-value
  ----------------------------------------------------------
  const       0.136643     0.0243551      5.610   2.03e-08  ***
  georgia    -0.0257154    0.00559488    -4.596   4.31e-06  ***
  after      -0.0598563    0.00686489    -8.719   2.88e-018 ***
  inter       0.0329694    0.0107684      3.062   0.0022    ***
  female      0.0613601    0.00361558    16.97    1.99e-064 ***
  black      -0.160808     0.00443856   -36.23    5.95e-284 ***
  hispanic   -0.154347     0.00544511   -28.35    1.90e-175 ***
  age         0.00528341   0.000802599    6.583   4.66e-011 ***

Mean dependent var   0.240584   S.D. dependent var   0.427442
Sum squared resid    9276.864   S.E. of regression   0.417177
R-squared            0.047577   Adjusted R-squared   0.047452
```

a. Comment on what the coefficients on these new control variables tell us
***These coefficients tell us how gender, race/ethnicity, and age impact educational attainment in these states. We see females are more likely to have a college degree, black and Hispanic individuals are less likely, and that the probability of having a degree increases as a person ages.***

b. Comment on what happened to the estimated impact of the program once we added the additional control variables

***The coefficient on the interaction variable has changed slightly after adding these additional variables to the model. The impact of the scholarship program on degree attainment is now estimated to be 3.30 percentage points, instead of 3.13 (from questions 1 and 2).***

4. Create a new dummy variable (called **bach**) indicating whether an individual earned at least a bachelor's degree. Run the same regression as question 3, but use this new dummy variable as our dependent variable.

```
Q4: OLS, using observations 1-53312
Dependent variable: bach

            coefficient   std. error    t-ratio    p-value
  ------------------------------------------------------------
  const       0.119532     0.0220536       5.420   5.98e-08  ***
  georgia    -0.0142472    0.00506617     -2.812   0.0049    ***
  after      -0.0481959    0.00621616     -7.753   9.11e-015 ***
  inter       0.0256293    0.00975082      2.628   0.0086    ***
  female      0.0470725    0.00327391     14.38    8.70e-047 ***
  black      -0.138211     0.00401911    -34.39    2.44e-256 ***
  hispanic   -0.127007     0.00493055    -25.76    1.99e-145 ***
  age         0.00356827   0.000726754     4.910   9.14e-07  ***

Mean dependent var   0.181404   S.D. dependent var   0.385356
Sum squared resid    7606.390   S.E. of regression   0.377754
R-squared            0.039190   Adjusted R-squared   0.039064
```

a. What is the estimated effect of the program if we analyze this outcome? Is this a better outcome to look at?

***When we focus on attaining a bachelor's degree or higher as the outcome, we see a smaller estimated impact of the program. The effect has dropped from 3.30 percentage points (in question 3), to 2.56 percentage points. This makes sense, this is a more restrictive outcome to measure. The earlier estimate included those with associates, bachelors, or graduate degrees, and now we only include the latter two.***

***In terms of which dependent variable is better to use, that is open for debate. The goal of the program is to increase educational attainment in the state of Georgia. In my opinion, moving people from high school graduates to having an associate degree would be achieving that goal. On the other hand, from a revenue-generation standpoint, the earnings of those with a bachelor's degree are much higher than those with associate's, so the state likely wants to examine this category from a cost/benefit perspective.***

5. Let's investigate whether the program had heterogeneous impacts on different groups. For each of the following groups, restrict the sample and run the simplified version of the model from question 2 (no control variables). Report the coefficient on the interaction term in each case. Use **college** as your dependent variable.

*Female, Male, Black, Hispanic, White*

***OMG, an opportunity to create a model table in gretl. Don't pass these opportunities up when you have them. To do this, I restricted the sample (e.g.*** female==1***), then ran the regression, then restored the sample, then did the same thing for the next group.***

OLS estimates
Dependent variable: college

| | Q5 females | Q5 males | Q5 black | Q5 hispanic | Q5 whites |
|---|---|---|---|---|---|
| const | 0.2953*** | 0.2438*** | 0.1546*** | 0.1626*** | 0.3151*** |
| | (0.003480) | (0.003228) | (0.004250) | (0.004881) | (0.002971) |
| georgia | -0.01484* | -0.02163*** | -0.01150 | 0.1204** | -0.01497** |
| | (0.008260) | (0.007630) | (0.008617) | (0.04757) | (0.007159) |
| after | -0.09380*** | -0.1076*** | -0.06146*** | -0.07858*** | -0.1137*** |
| | (0.006702) | [illegible] | [illegible] | [illegible] | (0.005772) |
| inter | 0.04476*** | 0.01817 | 0.01972 | 0.05872 | 0.03704*** |
| | (0.01619) | (0.01475) | (0.01599) | (0.09253) | (0.01432) |
| n | 26793 | 26519 | 11952 | 7292 | 37323 |
| Adj. R**2 | 0.0076 | 0.0132 | 0.0055 | 0.0121 | 0.0111 |

Comment on what we learn about the effects of the program on different groups:

***The effects of the program are estimated by the interaction coefficient, so this is where we should focus our energy. We see, interestingly, that the scholarship program increased the attainment of females by a significant amount, but there was not a significant effect for males. On the race/ethnicity side, we see that the effect is significant for whites, but not blacks or Hispanics. The coefficients for the Hispanic subsample is the largest, but the sample is the smallest, so there is not enough evidence for significance. This is one of the arguments against this type of program – that it helps those that don't traditionally need as much help (higher income, white students) as opposed to scholarship programs more based on financial need.***

6. The control group in this dataset includes people born in Alabama, Florida, South Carolina, and Texas. Choose which of these you believe might be the best comparison to Georgia and restrict the sample to compare just the two states. Run the same specification as in question 2.
    a. What is the estimated impact of the program, and how does this compare to the effect we found using the broader group of control states?

***What I want you to think about for this one is that our decision of what to include as a control state matters. I created a model table with each of the 4 control states. We see that the effect of the Georgia program is significant if we use Alabama, Florida, or Texas as the only control state. However, if we use South Carolina, we would conclude there is no significant effect. From an ethical standpoint, we should decide which state(s) are the best control group BEFORE we see the results.***

| | AL | FL | SC | TX |
|---|---|---|---|---|
| const | 0.2648*** | 0.2806*** | 0.2711*** | 0.2661*** |
| | (0.006223) | (0.004947) | (0.007115) | (0.003309) |
| georgia | -0.01327* | -0.02903*** | -0.01956** | -0.01460** |
| | (0.008034) | (0.007134) | (0.008760) | (0.006055) |
| after | -0.09873*** | -0.09722*** | -0.07776*** | -0.1092*** |
| | (0.01214) | (0.009366) | (0.01360) | (0.006321) |
| inter | 0.02874* | 0.02723** | 0.007769 | 0.03918*** |
| | (0.01568) | (0.01373) | (0.01687) | (0.01175) |
| n | 15638 | 19741 | 14295 | 31775 |
| Adj. R**2 | 0.0072 | 0.0083 | 0.0058 | 0.0108 |