Difference-in-Difference Estimation

1. (From 13.6) In 1985, neither Florida nor Georgia had laws banning open alcohol containers in vehicle compartments. By 1990, Florida had passed such a law, but Georgia had not.
	1. Suppose you can collect random samples of the driving age population in both states, for 1985 and 1990. Let arrest be a binary variable equal to unity if a person was arrested for drunk driving during the year. Without controlling for any factors, write down a linear probability model that allows you to test whether the open container law reduced the probability of being arrested for drunk driving. Which coefficient in your model measures the effect of the law?
	2. Why might you want to control for other factors in the model? What other factors might you want to include?
	3. Now, suppose you can only collect data for 1985 and 1990 and the county level for the two states. The dependent variable would be the fraction of licensed drivers arrested for drunk driving during the year. How does the data structure differ from the individual level data described in part (a)? What econometric method would you use?
2. A master’s student a few years ago worked on a paper examining the effect of tribal colleges on various outcomes for American Indians. Tribal colleges were formed in different states over the past 30 years, beginning with Dine College in 1968. The last tribal college formed was in 1997. Suppose he wanted to estimate the effects of tribal colleges on American Indian health outcomes.
3. What would be the “ideal” dataset for examining this question? Describe (1) who would be included in the dataset, (2) who the treatment and control groups would be, and (3) how the treatment would be assigned.
4. What would your estimating equation be using this data? Interpret your coefficients of interest.
5. Of course, we do not have ideal data. Census data provide a large nationally, representative survey of adults in the United States. The data are repeated cross sections—individuals are not followed over time. Data include the county in which an individual lives, the individual’s race and ethnicity, the individual’s education level, and information about the individual’s health. We also can match this will colleges in that county and when they open. Give at **least two** examples of potential treatment and control groups. Hint: this is the hard part of the question.
6. Give the estimating equations that correspond to the treatment and control groups you described. The notation in your equations (subscripts, etc) needs to be precise. Interpret your coefficients of interest.
7. Do Computer exercises in Wooldridge 13.6 and 13.12

One thing to note: the Wooldridge data sets for these problems have the “change” variables already constructed, usually labeled dXXX. You may find it useful to replicate the construction of these variables so that you understand how the models work.

**Replication Exercise—Differences in Differences**

Eissa and Liebman “Labor Supply Responses to the Earned Income Tax Credit” (QJE 1996) is a classic difference-in-difference papers. The Earned Income Tax Credit is a refundable tax credit for low income wage workers. Families have to be employed to receive the credit, which is essentially a wage subsidy. The credit also depends on the number of children (none, one, 2+). Economists have advocated using this credit in lieu of cash welfare because it causes fewer labor supply distortions for the participation decision.

This paper examines the 1986 expansion of the credit. It’s a great paper—I encourage you to read it if you are interested in this area. <http://www.jstor.org/stable/pdfplus/2946689.pdf>

This credit was substantially expanded again in 1993. This problem set will use Eissa and Liebman’s strategy to estimate the effects of the 1993 expansion. Essentially, you will compare labor supply for single women before and after 1993 by whether or not they had children: the EITC largely applied to women with children.

You will need to download the dataset EITC.dta to start. This dataset contains CPS data for single women 20-54 with less than a high school education, as this group is most likely to be affected by the EITC.

1. Describe and summarize your data using the describe and summarize commands in STATA. (I expect you to do that with every dataset you work with. Get in the habit at looking at means, max values, and min values to get a sense of the data and to look for outliers/coding errors.)
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. Earning are reported as zero for women who are not employed. Create a new variable with earnings conditional on working (missing for non-employed) and calculate the means of this by group as well.

(Use summarize if \*\*\*\*) Your results here parallel those reported in EL Table I (but will be different because you have more years.)

1. How do these three samples differ?
2. Construct a variable for the “treatment” called ANYKIDS and a variable for after the expansion (called POST93—should be 1 for 1994 and later).

STATA hint:

 gen dummyx = (x>=0)

This will create a variable, dummyx, with values of 1 if the statement in the parentheses is true, and 0 otherwise

1. Create a graph which plots mean annual employment rates by year (1991-1996) for single women with children (treatment) and without children (control). Use this graph to discuss the validity of using single women without children as a control group. Given the other expansions prior to 1993, are there difficulties in testing for differences in “pre-treatment” trends?

 Some useful STATA commands for doing this

preserve;

/\*allows you to make changes and then later revert back to original data set\*/;

/\*Note that you can’t run this interactively—once the do file is complete, the dataset reverts back to the original\*/;

collapse work, by (year anykids);

/\*creates means by year & kids—read up on the collapse command—one of my favs\*/;

 restore /\*restores dataset to what it was before collapse command\*/;

gen work0 = work if anykids==0;

 gen work1=work if anykids==1;

/\*You can also normalize these, making the scales comparable. Lots of ways to do that—try making both start with a value of “1” in the first year\*/;

Use the graph commands to plot the series work0 and work1—mess around with the options to learn how to make pretty graphs.

1. Calculate the unconditional difference-in-difference estimates of the effect of the 1993 EITC expansion on employment of single women. For comparison, here are the EL results, with standard errors in parentheses. Your table should be parallel this format. Calculate estimates with all women with children as the treatment (single women with no children as the control), women with children as the control. Discuss these results. Are they consistent or inconsistent with the theoretical predictions?

From Table II, Eissa and Liebman (1996)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Before 1986 | After 1986 | Difference | D-D |
| Less that High school, with children | .479(.010) | .497(.010) | .018(.014) | .041(.019) |
| Less than High school, without children | .784(.010) | .761(.009) | -.023(.013) |

 Some useful STATA commands:

 sort anykids

 by anykids: summarize \*\*\*\*\*\*\*\*\* if post93==1 etc.

will give you means and standard deviations by category

 mean work will give you means and standard errors

Reminder: To calculate standard errors for the means, divide standard deviations of the variable by the square root of n

The standard error for a difference in means is

 $se\left(\overbar{Y\_{1}}-\overbar{Y}\_{2}\right)=\sqrt{\frac{σ\_{1}^{2}}{n\_{1}}+\frac{σ\_{2}^{2}}{n\_{2}}}$

1. 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. Do not include any variables other that the few needed to calculate that effect. How do these results compare with what you found in (6)? What is the interpretation of the coefficient on the variables you included?
2. Reestimate this model including demographic characteristics. The ones in EL include non-labor income, number of preschool children, nonwhite dummy, age, age squared, education, and education squared. You will need to create some of these variables. How does the treatment effect change? Is that consistent with what you would expect? Interpret the parameter estimates on nonwhite and non-labor income variables. (This compared to EL Table III.)
3. Add the state unemployment rate and allow its effect to vary by the presence of children. (You will need to make another variable.) Discuss your results. (This compared to EL Table III.)
4. Allow the treatment effect to vary by those with 1 or 2+ children. Discuss your results and how they compare to your expectations.
5. 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 (so 1992 and 1993 both have this fake policy). What do you find? Discuss the implications of this result.