Empirical Applications to Labor Supply: Lecture 4

Empirical Application: Differences in Differences
Paper: Eissa and Liebman (QJE 1996)

Difference-in differences strategies are simple panel data methods applied to sets of group means in cases when certain groups are exposed to the causing variable of interest and others are not. The approach is well suited to estimating the effect of sharp changes in the economic environment or changes in government policy when a suitable control group can be found. We'll consider the Diff-in-diff approach with the application of examining the impact of welfare reform in the U.S.

Eissa and Liebman want to examine the impact of the U.S. Tax Reform Act of 1986. In particular, part of the act increased an existing earned income tax-credit for single women with children. The earned income tax credit (EITC) began in 1975. In general, a taxpaying family is eligible for the subsidy if 1) earned income is below a particular amount (about $28,000 in 1996), and 2) parents have a child under 19 years old.

The credit works roughly the following way: the full credit is phased in at a 11% rate over the first $5000 of income (the subsidy increases with hours worked, unlike the SSP where it decreases). The maximum credit is 550 holds until earnings are $6,500, and is then phased out at 12.22 percent, until earnings is $11,000.

<write down figure>

The 1987 change increased the subsidy phase-in rate from 11% to 14%, and the maximum credit increased to $851. The phaseout rate was also lowered, from 12.33% to 10%. Taxpayers with incomes between $11,000 and $15,432 became eligible for the credit for the first time in 1987.

The tax reform act also raised non-work related deductions for households with and without children, fairly substantially relative to the change in the EITC. This shifts the budget line up – which should decrease hours of work for eligible taxpayers who are already in the workforce and view leisure as a normal good.

In contrast, the EITC unambiguously predicts a positive impact on labor force participation, because income effects from the program for those not participating are zero.

Estimation strategy:

Let Y_{igt} be the observed outcome (hours worked in this example), for individual i, from group g, at time t. The effect of average interest is $E(Y_{1igt} - Y_{0igt})$, Y_{1igt} is the outcome if the policy is implemented, and Y_{0igt} is the outcome if not. Group 1 is at some time exposed to the policy and group 2 is not.

The underlying assumption in the diff-in-diff framework is that

$$E(Y_{0igt} \mid g,t) = e_g + e_t$$

that is, in the absence of the new policy, average hours worked can be decomposed into a time effect that is common to groups and a group effect that is fixed over time.

<draw trending pattern on board>

Suppose that the average effect of the program is simply a constant, so that:

$$E(Y_{1igt} \mid g,t) = E(Y_{0igt} \mid g,t) + \beta$$

If D_i is an indicator for whether group g is exposed to the policy, then we can write:

$$Y_i = \beta D_i + e_g + e_t + e_i,$$

where $E(e_i \mid g,t) = 0$.

This is just a regression equation, with fixed effects for time and group.

If we take the difference in outcomes across time, we identify the average effect. Suppose when t = 1, no reform has taken place, and in between t=1 and t=2 the policy changes and affects group 1. Then:

$$\begin{aligned}&E(Y_i \mid g=1,t=1) - E(Y_i \mid g=2,t=1)\\&+E(Y_i \mid g=1,t=2) - E(Y_i \mid g=2,t=2) = \beta\end{aligned}$$

In this example, Eissa and Liebman want to examine how the EITC reforms impacted single women parents' labor supply. Where Y_{igt} is labor supply, t=1 for the year=1985 (before the reform) and t=2 if the year = 1987 (after the reform). Note we are note examining a specific change in tax liability, but the overall effect in the entire shift in the budget constraint. Individuals will be affected in different ways. The only thing we can predict is that the reform should raise overall employment, among single eligible parents.

Key here is, what is the counterfactual?? We'd like to know the treatment effect relative to a similar group that was not eligible. Picking the control group is crucial, since we assume that both groups are affected by time identically.

Eissa and Liebman suggest using as the control group the population of single women without children. This group was not eligible for the EITC,

The difference in differences strategy makes 2 crucial assumptions: 1) the interaction terms are zero in the absence of the intervention. The outcomes are trending exactly the same way without the change. If hours worked evolves differently across single women with and without children, we have a problem. We can often test whether trends in outcomes are the same before the policy break. A related assumption is the composition of both groups remains stable before and after the policy change. We're assuming the population background characteristics within groups (not correlated to the policy change) remain stable.

The smaller the time range examined, the less likely trends will deviate.

Note, essentially the way I've described this analysis, there are only 4 observations: the mean labor supply for the 2 groups, before and after the change. If we can observe other factors for individuals that could affect labor supply (that could change between periods), we may be able to get more efficient estimates by controlling for these observables and working at a smaller level of data than means.

$$Y_i = X_i^{'}\delta + \beta D_i + e_g + e_t + e_i$$

Controlling for other individual characteristics, X_i, the estimate of β only if X_i and D_i are correlated, conditional on group and time main effects.

Also in practice, we can sometimes allow for the effect to vary with time.

A quick note on regression discontinuity: why do we even need a control group in the first place? Why not measure the change in hours supplied the year before and the year after the change? Under a different set of assumptions, we can identify the causal effect. I'll present a good example of regression discontinuity later. A discontinuity approach doesn't work well here because we have few observations before and after the reform change, and the policy is likely to take some time to have an effect and so there is less likely a discontinuity in outcomes right at the year the policy changes.

Results:

Table 1 shows that the average characteristics, such as income, hours worked, age, are quite different between single women with and without children. It should make us nervous about the identification assumptions – are these groups really likely to experience the same time shocks?

Eissa and Liebman try to address this by focussing only on single women with less than high school. Table II shows main result: labor force participation rises for the treatment group by 4 percentage points after the reform.

Figure II, tries to convince us that there are no underlying time trends. We can squint and see main results here: labor force participation going up for females with children, and perhaps slightly down for females without.

Table V interestingly shows no significant response in annual hours and annual weeks from reform. This result got a lot of attention.

One last thing: the analysis says nothing about the overall costs (to the taxpayer) for introducing the program.

A note on weighting data

Sampling weights are often used to correct for imperfections in the sample that might lead to bias and other departures between the sample and reference population.

Be aware of how your data was collected.

e.g. Census: no missing observations. Why? (hot deck: allocation flags to catch this) Imputed observations
e.g. PSID: over-samples low income families

Why weight?

1) to compensate for unequal probabilities of selection (non random sample) (known)
2) 2) to compensate for non response (missing observations): use known distributions of observable observations (e.g. gender) to reweight sample so weighted sample in line with known distribution.
3) c) to adjust weighted sample distribution to make it conform to a known population distribution (make the data 'add up' to known population"

To compute any counts or means, must use weights

Example:

There is a population of 100,000 people, and only enough money to interview 1,000 people. The population Is divided into 2 regions, A and B. The percentage of low income people in the total population is 20%. We want to do some separate analysis for the low income group, and 200 people may not generate a large enough sample. Suppose we know Region A has 25,000 people, 50% low income people. Region B has only 10% low income people. If we sample 500 people from each region, we can expect to sample 500*.5 + 500*.1 = 300, instead of 200 from sampling a random sample across both regions.

The chance of a person in region A being selected is 500/25,000=.02. The chance of a person in region B being selected is 500/75,000 = .00666667. To create weights, we assign the inverse probability of being selected. People in region A get a weight of 1/.02 = 50. Each person in region A represents 50 people. People in B get a weight of 1/..00666667 = 150.

$$\bar{x}_i = \frac{\sum_n x_i w_i}{\sum_n w_i} = \frac{\sum_g \bar{x}_g N_g w_i}{\sum_n w_i} = \sum_g \bar{x}_g \Pr(g)$$

For regression it's less clear whether we should use weights. If our data is of cell means and we know the sample size in each cell, we would definitely want to weight the regression. If the variance of each individual observation is normally distributed: $e_i \sim N(0,\sigma^2)$, then the variance for cell mean observations is $e_g \sim N(0,\frac{\sigma^2}{n_g})$, where n_g is the number of observations in each cell. The appropriate correction is:

Let $v_g = \frac{N_g n_g}{\sum_g n_g}$. $\mathbf{D}$ is the diagonal matrix whose diagonal elements are the elements of v. Ng is total number of cells. Then, regression weights the equations by the observations:

$$\mathbf{X'X} = \mathbf{X'DX}$$
$$\mathbf{X'y} = \mathbf{X'Dy}$$

$$\hat{\beta} = \frac{\sum_g (y_i - \bar{y})(x_i - \bar{x})v_g}{\sum_n (x_i - \bar{x})^2 v_g}.$$

This is the computation made when using aweights in STATA.

Note, this is the equivalent to multiplying every variable in the regression by $\sqrt{n_g}$ and carrying out the unweighted regression of: $Y_i\sqrt{n_g} = \beta_0\sqrt{n_g} + \beta_1 X_i\sqrt{n_g} + e_i\sqrt{n_g}$

The justification for using probability weights when the survey over samples some groups is less clear. If variance for each observation is $e_i \sim N(0,\sigma^2)$, the variance for each over or under sampled observation is still $e_i \sim N(0,\sigma^2)$. There is no heteroskedasticity problem like the case with cell mean observations. One could argue for using this approach with probability weights instead (w_i as defined above) for efficiency reasons. In addition, if you believe B is different for different groups, you should weight if you are after the population average effect.

In practice, doesn’t seem to matter much if proportion in population similar to proportion in sample. See, for example, Angrist and Krueger, table 12. With regression, conditioning on X variables used to group and compute weights, don’t require weights. Some statisticians have even questioned whether weights should be used at all with regression. I have yet to see a paper that rests on the weighting assumptions, but the standard practice is to weight.

Further references:

http://www.amstat.org/sections/srms/Proceedings/papers/1981_135.pdf

http://www2.chass.ncsu.edu/garson/pa765/sampling.htm

http://unstats.un.org/unsd/demographic/meetings/egm/Sampling_1203/docs/no_5.pdf

STATA 8 User reference 23.16

A note on the need to 'cluster' standard errors

OLS assumes no serial correlation or autocorrelation in the error terms when estimating the variance (and standard errors) of the coefficients. This can lead to downward bias in the standard errors if, instead, the errors, or at least some of them, are positively correlated. The bias can sometimes be severe.

Consider the variance of the ordinary least squares regression. Let $\mathbf{X}$ be the $n \times p$ design matrix and $\mathbf{y}$ be the $n \times 1$ vector of dependent values. The regression model is: $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$, so any fixed effects are defined as dummy variables contained in the $\mathbf{X}$ matrix, and $\mathbf{y}$ and $\mathbf{X}$ are deviations from their means. The ordinary linear regression estimates are $(\mathbf{X'X})^{-1}\mathbf{X}'\mathbf{y}$, and the variance is:

$$\text{var}(\mathbf{b}) = (\mathbf{X'X})^{-1}\mathbf{X}'E[(\mathbf{y} - E\mathbf{y})(\mathbf{y} - E\mathbf{y})']\mathbf{X}(\mathbf{X'X})^{-1}$$

$$\text{var}(\mathbf{b}) = (\mathbf{X'X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X'X})^{-1}$$

where $E(\varepsilon_i) = 0$ and $E(\varepsilon_i \varepsilon_j) = \Omega$, (the variance-covariance matrix for all i and j observations)

The standard OLS assumption to estimate the variance is $\Omega = \sigma^2 \mathbf{I}$, and $\hat{\sigma}^2 = \frac{1}{N}\sum_1^N e_i^2$:

$$\text{var}(\mathbf{b}) = \hat{\sigma}_\varepsilon^2 (\mathbf{X'X})^{-1}$$

OLS assumes that the variance matrix for the error term is diagonal while in practice it might be block diagonal, with a constant correlation coefficient within each group and time cell. When we want to identify an aggregate group/time effect, within group/time correlation can be substantial. In practice, the correlation is often positive, which leads the OLS results to underestimate the standard error, making it more likely to reject the null hypothesis. It is reasonable to expect that units sharing observable characteristics such as being from the same industry, state, marital status, time period and location, also share unobservable characteristics that would lead the regression disturbances to be positively correlated. With Monte Carlo experiments, several recent papers have suggested using OLS standard error estimates can bias standard errors downwards and lead to rejection that the coefficient is zero, when in fact, it is.

Fortunately, White (and earlier Eicher and Huber) found a way to estimate robust standard errors, regardless of the form Ω takes (provided that Ω is well defined). White pointed out that we do not need to estimate every component in the n x n Ω matrix, an obviously impossible task when only n observations are available. But this way of looking at the problem is misleading. What is actually required is to estimate

$$\text{var}(\mathbf{b}) = (\mathbf{X'X})^{-1} E[\mathbf{X}^{'}\mathbf{ee'X}](\mathbf{X'X})^{-1}$$

(White, 84, Aymptotic Theory for econometricians)

The robust variance-covariance matrix estimator is:

$$\text{var}(\mathbf{b}) = (\mathbf{X'X})^{-1}\left(\sum_{1}^{N}[(y_i - \hat{y}_i)\mathbf{x_i}]'[(y_i - \hat{y}_i)\mathbf{x_i}]\right)(\mathbf{X'X})^{-1}$$

where $\hat{y}_i$ is the estimated error term, and the sum is over all observations. This variance is computed when the 'robust' option is specified in STATA. When prior knowledge leads the researcher to believe the error terms may be serially correlated within groups, but independent across groups, the variance can be calculated as:

$$\text{var}(\mathbf{b}) = (\mathbf{X'X})^{-1}\left(\sum_{1}^{G}u_k^{'}u_k\right)(\mathbf{X'X})^{-1}, \text{ where } u_k = \sum_{j\in k}[(y_j - \hat{y}_j)\mathbf{x_j}]'[(y_j - \hat{y}_j)\mathbf{x_j}]$$

This variance estimate is computed with STATA's 'cluster' command, specifying groups G.

This estimator is consistent for any arbitrary heteroskedasticity or serial correlation, but it is not efficient when prior information about the form of the matrix is known.

To give you a little intuition for the need to cluster, consider the following example. Suppose we are evaluating the relationship between education attainment and state compulsory school laws. Let S_{is} be years of schooling for individual i in state S, and Z_S is the dropout age that an individual faced when in high school, from state S. So the independent variable is the same for everyone from that state. The OLS regression equation is:

$$S_{is} = \beta Z_S + e_{iS} \ ,$$

It's certainly plausible that individuals from the same state are related in other ways. There could still be no omitted variables bias: $E(Z_S, e_{iS}) = 0$, but the error terms are serially correlated among individuals from the same state: $E(e_{iS}, e_{jS} \mid S = \overline{S}) \neq 0$.

One extreme example is we have 100 individuals, 2 from each state. Z_S is the same for each two individuals from the same state. Suppose also that S_{is} is the same for both. So what we have is 2 sets of the same 50 values for S and Z. Normalize the standard deviation to 1: $E(e_{iS}^2) = 1$. If the variance-covariance matrix is $\mathbf{\Omega} = \mathbf{I}$, as in OLS, the variance of $\hat{\beta}$ is:

$$\text{var}(\hat{\beta}) = \frac{1}{2\sum_1^{50} Z_i^2} 2\sum_1^{50} Z_i^2 \frac{1}{2\sum_1^{50} Z_i^2} = \frac{1}{2\sum_1^{50} Z_i^2}$$

If, instead, e_{iS} is perfectly correlated within state, $E(e_{iS}, e_{jS} \mid S = \overline{S}) = 1$ and zero otherwise. Recognizing the, the true variance of $\hat{\beta}$ is

$$\mathrm{var}(\hat{\beta}) = \frac{1}{2\sum_{1}^{50} Z_i^2} 4\sum_{1}^{50} Z_i^2 \frac{1}{2\sum_{1}^{50} Z_i^2} = \frac{1}{\sum_{1}^{50} Z_i^2}$$

If the second covariance matrix is correct, we falsely underestimate the variance of $\hat{\beta}$ using OLS. The second individual in each state adds no new information. If e_{iS} was only partially correlated within state, the variance would be smaller, but still larger than OLS. Using White's clustering approach leads to a consistent estimate of the variance of $\hat{\beta}$, no matter what shape underlies $\mathbf{\Omega}$.

One should note that this estimator applies asymptotically (as the sample size and the number of groups approaches infinity). Monte carlo experiments reveal that the estimator works reasonably well when the sample size within groups is not especially large relative to the number of groups. Unfortunately, the number of groups is very small, relying on asymptotics can be very misleading. What is small? The references below suggest even groups as high as 40 or 50 can lead to poor estimates. A conservative solution is to aggregate the data up to the group level and run the regressions using the grouped means, weighted by the sample size. In our example, this would be:

$$\bar{S}_s = \beta \bar{Z}_S + e_S,$$

which will generate the same estimate for B and the variance of B in our simple example. If there is no cluster effect (no serial correlation within groups), then aggregating to the group level removes information and increases the variance unnecessarily. In practice, results are far more convincing if you can produce robust and significant results with this aggregated approach (if it's applicable).

Note, in the diff-in diff example above, if we aggregated, we only would have 4 observations. And indeed, one criticism that has been put out by some researchers is that the diff in diff approach is just in essence comparing 2 groups over time and we can't be sure that any observed significant difference in means is due entirely to the policy change.

Useful references for this topic:

Wooldridge, AER, May 2003, p 133, “Cluster Sample methods in Applied Econometrics”

Donald, Stephen, and Kevin Lang, “Inference with Difference in difference and other panel data,’ mimeo, 2001

White, Halbert, “Asymptotic Theory for Econometricians,” 1984

Bertrand, Duflo, and Mullainathan, ‘how much should we trust differences-in-differences estimates,’ QJE, Feb 2004-09-13

Arellano, M. “Computing Robust Standard errors for within groups estimators,’ oxford bulletin of economics and statistics, 49, 4 (1987)

White, Halbert, “a heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity,’ econometrica, 1980

Kezdi, Gabor, ‘robust standard error estimation in fixed effects panel models,’ university of Michigan mimeo, 2002.