How much should we trust differences-in-differences estimates?

Carlos Lezama
Advanced Microeconometrics
ITAM

Spring 2022

1. **Introduction**

2. **A Survey of DD Papers**

3. **Overrejection in DD Estimation**

4. **“If hundreds of researchers analyzed the effects of various laws in the CPS, what fraction would find a significant effect even when the laws have no effect?”**

5. **Solutions**

6. **Conclusion**

Introduction

Introduction

Differences-in-Differences (DD) estimation consists of identifying a specific intervention or treatment. One then compares the difference in outcomes after and before the intervention for groups affected by the intervention to the same difference for unaffected groups.

The great appeal of DD estimation comes from its simplicity as well as its potential to circumvent many of the endogeneity problems that typically arise when making comparisons between heterogeneous individuals.

Obviously, DD estimation also has its limitations. It is appropriate when the interventions are as good as random, conditional on time and group fixed effects. Therefore, much of the debate around the validity of a DD estimate typically revolves around the possible endogeneity of the interventions themselves.

Introduction

Focus on issues relating to the *standard error* of the estimate.

DD estimates and their standard errors most often derive from using Ordinary Least Squares (OLS) in repeated cross sections (or a panel) of data on individuals in treatment and control groups for several years before and after a specific intervention.

One then typically estimates the following regression using OLS:

$$Y_{ist} = A_s + B_t + cX_{ist} + \beta I_{st} + \varepsilon_{ist}, \tag{1}$$

where the subscripts i, s, and t stand for the individual, group (such as a state), and time (such as a year), respectively. Furthermore, let Y be our outcome of interest, I be a dummy, A and B be fixed effects, X be relevant individual controls, and ε be an error term.

Introduction

The estimated impact of the intervention is then the OLS estimate $\hat{\beta}$. Standard errors used to form confidence interval for $\hat{\beta}$ are usually OLS standard errors.

Remark

Note that this is valid only under the very restrictive assumption that changes in the outcome variable over time would have been exactly the same in both treatment and control groups in the absence of the intervention.

Introduction

Three factors make serial correlation an especially important issue in the DD context:

1. DD estimation usually relies on fairly long time series.
2. The most commonly used dependent variables in DD estimation are typically highly positively serially correlated.
3. The treatment variable I_{st} changes itself very little within a state over time.

These three factors reinforce each other so that the standard error for $\hat{\beta}$ could severely understate the standard deviation of $\hat{\beta}$.

Introduction

To assess the extent of this problem, they examine how DD performs on placebo laws, where treated states and year of passage are chosen at random.

Placebo tests diagnose problems with research designs in observational studies. When a researcher estimates a treatment effect based on observational data, the estimator may be biased by confounders, model misspecification, differential measurement error, or other flaws; the researcher may also have constructed confidence intervals incorrectly, such that we would reject the null hypothesis too frequently (or infrequently) under the null. A placebo test checks for an association that should be absent if the research design is sound but not otherwise. Placebo tests can thus be seen as a strategy for checking the soundness of a research finding and, more broadly, improving causal inference.

A Survey of DD Papers

A Survey of DD Papers

Papers were classified as “DD” if they focus on specific interventions and use units unaffected by the law as a control group.

Table 1: Survey of DD papers

Number of DD papers	92
Number with more than 2 periods of data	69
Number which collapse data into before-after	4
Number with potential serial correlation problem	65
Number with some serial correlation correction	5
GLS	4
Arbitrary variance-covariance matrix	1
Number with potential clustering problem	80
Number which deal with it	36

A Survey of DD Papers

Table 2: Distribution of time span for papers with more than 2 periods

Percentile	Value
1%	3
5%	3
10%	4
25%	5.75
50%	11
75%	21.5
90%	36
95%	51
99%	83
Average	16.5

A Survey of DD papers

Table 3: Most commonly used dependent variables

Employment	18
Wages	13
Health/medical expenditure	8
Unemployment	6
Fertility/teen motherhood	4
Insurance	4
Poverty	3
Consumption/savings	3

A Survey of DD papers

Table 4: Informal techniques used to assess endogeneity

Graph dynamics of effect	15
See if effect is persistent	2
Attempt to do triple-differences (DDD)	11
Include time trend specific to treated states	7
Look for effect prior to intervention	3
Include lagged dependent variable	3

Overrejection in DD Estimation

Data

The survey above suggests that most DD papers may report standard errors that understate the standard deviation of the DD estimator. To illustrate the magnitude of the problem they turn to a sample of women's wages from the Current Population Survey (CPS).

More specifically, data on:

- Women in their fourth interview month in the Merged Outgoing Rotation Group of the CPS.
- Years: 1979 – 1999.
- Age: 25 – 50 y/o.
- Information on weekly earnings, employment status, education, age, and state of residence.

Data

Summary

The sample contains:

- Nearly 900,000 observations.
- Approximately 540,000 women report strictly positive weekly earnings.
- wage = log(weekly earnings).

This generates ($50 \times 21 = 1050$) state-year cells, with each cell containing on average a little more than 500 women with strictly positive earnings.

Methodology

The correlogram of the wage residuals was informative enough to estimate first, second, and third autocorrelation coefficients for the mean state-year residuals from a regression of wages on state and year dummies such that they equal 0.51, 0.44, and 0.31, respectively (obtained by a simple OLS regression of the residuals on the corresponding lagged residuals) — which are high and statistically significant.

Methodology

Subsequently, in the DD context:

1. Randomly, draw a year $\sim \mathcal{U}(1985, 1995)$.
2. Select exactly half states (25) at random and designate them as "affected" by the law such that

$$I_{st} = \begin{cases} 1 & \text{for all women that live in an affected state} \\ & \text{after the intervention date,} \\ 0 & \text{otherwise.} \end{cases}$$

3. Estimate equation (1) using OLS on these placebo laws.

“If hundreds of researchers analyzed the effects of various laws in the CPS, what fraction would find a significant effect even when the laws have no effect?”

Methodology

If OLS were to provide consistent standard errors, we would expect to reject the null hypothesis of no effect ($\beta = 0$) roughly 5 percent of the time when using a threshold of 1.96 for the absolute *t*-statistic.

Solutions

Parametric Methods

Text.

Block Bootstrap

Text.

Ignoring Time Series Information

Text.

Empirical Variance-Covariance Matrix

Text.

Arbitrary Variance-Covariance Matrix

Text.

Summary

Text.

Conclusion

Conclusion

Text.