NBER WORKING PAPER SERIES

HOW MUCH SHOULD WE TRUST DIFFERENCES-IN-DIFFERENCES ESTIMATES?

Marianne Bertrand
Esther Duflo
Sendhil Mullainathan

Working Paper 8841
http://www.nber.org/papers/w8841

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2002

We thank Alberto Abadie, Daron Acemoglu, Joshua Angrist, Abhijit Banerjee, Victor Chernozhukov, Kei Hirano, Guido Imbens, Larry Katz, Je .rey Kling, Kevin Lang, Steve Levitt, Kevin Murphy, Emmanuel Saez, Doug Staiger, Bob Topel and seminar participants at Harvard, MIT, University of Chicago GSB, University of California at Los Angeles, University of California Santa Barbara, and University of Texas at Austin for many helpful comments. Tobias Adrian, Shawn Cole, and Francesco Franzoni provided excellent research assistance. We are especially grateful to Khaled for motivating us to write this paper. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

How Much Should We Trust Differences-in-Differences Estimates?
Marianne Bertrand, Esther Duflo and Sendhil Mullainathan
NBER Working Paper No. 8841
March 2002
JEL No. C10, C13, E24, K39

ABSTRACT

Most Difference-in-Difference (DD) papers rely on many years of data and focus on serially correlated outcomes. Yet almost all these papers ignore the bias in the estimated standard errors that serial correlation introduce4s. This is especially troubling because the independent variable of interest in DD estimation (e.g., the passage of law) is itself very serially correlated, which will exacerbate the bias in standard errors. To illustrate the severity of this issue, we randomly generate placebo laws in state-level data on female wages from the Current Population Survey. For each law, we use OLS to compute the DD estimate of its "effect" as well as the standard error for this estimate. The standard errors are severely biased: with about 20 years of data, DD estimation finds an "effect" significant at the 5% level of up to 45% of the placebo laws.

Two very simple techniques can solve this problem for large sample sizes. The first technique consists in collapsing the data and ignoring the time-series variation altogether; the second technique is to estimate standard errors while allowing for an arbitrary covariance structure between time periods. We also suggest a third technique, based on randomization inference testing methods, which works well irrespective of sample size. This technique uses the empirical distribution of estimated effects for placebo laws to form the test distribution.

Marianne Bertrand
Graduate School of Business
University of Chicago
1101 East 58th Street
Chicago, IL 60637,
NBER and CEPR
marianne.bertrand@gsb.uchicago.edu

Esther Duflo
Department of Economics
MIT, E52-252G
50 Memorial Drive
Cambridge, MA 02142,
NBER and CEPR
eduflo@mit.edu

Sendhil Mullainathan
Department of Economics
MIT, E52-380A
50 Memorial Drive
Cambridge, MA 02142
and NBER
mullain@mit.edu

1 Introduction

Difference-in-Difference (DD) estimation has become an increasingly popular way to estimate causal relationships. DD estimation consists of identifying a specific intervention or *treatment* (often the passage of law). One then compares the difference in outcomes after and before the intervention for groups affected by it to this difference for unaffected groups. For example, to identify the incentive effects of social insurance, one might first isolate states that have raised unemployment insurance benefits. One would then compare changes in unemployment duration for residents of states raising benefits to residents of states not raising benefits. 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.[1]

Obviously, DD estimation also has its drawbacks. Most of the debate around the validity of a DD estimate revolves around the possible endogeneity of the laws or interventions themselves.[2] Sensitive to this concern, researchers have developed a set of informal techniques to gauge the extent of the endogeneity problem.[3] In this paper, we address an altogether different problem with DD estimation. We assume away biases in estimating the intervention's effect and instead focus on possible biases in estimating the *standard error* around this effect.

DD estimates and standard errors for these estimates 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. Formally, let Y_{ist} be the outcome of interest for individual i in group s (such as a state) at time t and T_{st} be a dummy for whether the intervention has affected group s at time t.[4] One then typically estimates the following regression using OLS:

$$Y_{ist} = A_s + B_t + c\,X_{ist} + \beta\,T_{st} + \epsilon_{ist} \tag{1}$$

where A_s and B_t are fixed effects for the states and years and X_{ist} represents the relevant individual

[1] See Meyer (1994) for an overview.

[2] See Besley and Case (1994). Another prominent concern has been whether DD estimation ever isolates a specific behavioral parameter. See Heckman (1996) and Blundell and MaCurdy (1999). Abadie (2000) discusses how well control groups serve as a control.

[3] Such techniques include the inclusion of pre-existing trends in states passing a law, testing for an "effect" of the law before it takes effect, or using information on political parties to instrument for passage of the law (Besley and Case 1994).

[4] For simplicity of exposition, we will often refer to interventions as laws, groups as states and time periods as years in what follows. Of course this discussion generalizes to other types of DD estimates.

controls. The estimated impact of the intervention is then the OLS estimate $\hat{\beta}$. Standard errors around that estimate are OLS standard errors after accounting for the correlation of shocks within each state-year (or s-t) cell.[5]

In this paper, we argue that the estimation of equation 1 is in practice subject to a possibly severe serial correlation problem. While serial correlation is well-understood, it has been largely ignored by researchers using DD estimation. Three factors make serial correlation an especially important issue in the DD context. First, DD estimation usually relies on fairly long time series. Our survey of DD papers, which we discuss below, finds an average of 16.5 periods. Second, the most commonly used dependent variables in DD estimation are typically highly positively serially correlated. Third, and an intrinsic aspect of the DD model, the treatment variable T_{st} changes itself very little within a state over time. These three factors reinforce each other to create potentially large mis-measurement in the standard errors coming from the OLS estimation of equation 1.

To assess the extent of this bias, we examine how DD performs on placebo laws, where state and year of passage are chosen at random. Since these laws are fictitious, a significant "effect" at the 5% percent level should be found only 5% of the time. In fact, we find dramatically higher rejection rates of the null hypothesis of no effect. For example, using female wages as a dependent variable (from the Current Population Survey) and covering 21 years of data, we find a significant effect at the 5% level in as much as 45% of the simulations.[6]

We propose three different techniques to solve the serial correlation problem.[7] The first two techniques are very simple and work well for sufficiently large samples. First, one can remove the time-series dimension by aggregating the data into two periods: pre- and post-intervention. Second, one can allow for an arbitrary covariance structure over time within each state. Both of these solutions work well when the number of groups is large (e.g. 50 states) but fare poorly as

[5] This correction accounts for the presence of a common random effect at the state-year cell level. For example, economic shocks may affect all individuals in a state on an annual basis (Moulton 1990; Donald and Lang 2001). Ignoring this grouped data problem can lead to an under-statement of the standard error. In most of what follows, we will assume that the researchers estimating equation 1 have already accounted for this problem, either by allowing for appropriate random group effects or, as we do, by collapsing the data to a higher level of aggregation, such as state-year cells.

[6] Similar magnitudes arise in data manufactured to match the CPS distributions and where we can be absolutely sure that the placebo laws are not by chance picking up a real intervention.

[7] Other techniques fare poorly. Simple parametric corrections which estimate specific processes (such as an AR(1)) fare poorly because even long time series (by DD standards) are too short to allow precise estimation of the auto-correlation parameters and to identify the right assumption about the auto-correlation process. On the other hand, block bootstrap fails because the number of groups (e.g. 50 states) is not large enough.

the number of groups gets small. We propose a third (and preferred) solution which works well irrespective of sample size. This solution, based on the randomization inference tests used in the statistics literature, uses the distribution of estimated effects for placebo laws to form the test statistic.

The remainder of this paper proceeds as follows. In section 2, we assess the potential relevance of the auto-correlation problem: section 2.1 reviews why failing to take it into account will result in biased standard errors, and section 2.2 surveys existing DD papers to assess how it affects them. Section 3 examines how DD performs on placebo laws. Section 4 describes possible solutions. Section 5, discusses implications for the existing literature. We conclude in Section 6.

2 Auto-correlation and Standard Errors

2.1 Review

It will be useful to quickly review exactly why serial correlation poses a problem for OLS estimation. Consider the OLS estimation of equation 1, and denote V the vector of independent variables and α the vector of parameters. Assume that the error term ϵ has $E[\epsilon] = 0$ and $E[\epsilon\epsilon] = \Omega$. The true variance of the OLS estimate is given by:

$$\text{var}(\hat{\alpha}) = \sigma_\epsilon^2 (V'V)^{-1} V \Omega V (V'V)^{-1} \tag{2}$$

while the OLS estimate of the variance is:

$$\text{est var}(\hat{\alpha}) = \hat{\sigma}_\epsilon^2 (V'V)^{-1} \tag{3}$$

To more easily compare these expressions, let's consider a simple uni-variate time-series case in which we regress y_t on v_t with T periods of data. Suppose that the error term u_t follows an AR(1) process with auto-correlation parameter ρ and that the independent variable v_t follows an AR(1) with auto-correlation parameter $\lambda \geq 0$. In this special case, equations 2 and 3 can be simplified to:

$$var(\hat{\alpha}) = \frac{\sigma_\epsilon^2}{\sum_{t=1}^T v_t^2}(1 + 2\rho \frac{\sum_{t=1}^{T-1} v_t v_{t+1}}{\sum_{t=1}^T v_t^2} + 2\rho^2 \frac{\sum_{t=1}^{T-2} v_t v_{t+2}}{\sum_{t=1}^T v_t^2} + ... + 2\rho^{T-1} \frac{v_1 v_T}{\sum_{t=1}^T v_t^2})$$

and

$$\text{est var}(\hat{\alpha}) = \frac{\hat{\sigma}_\epsilon^2}{\sum_{t=1}^T v_t^2}$$

As $T \to \infty$, the ratio of estimated to true variance equals $\frac{1-\rho\lambda}{1+\rho\lambda}$.

These formulas make transparent three well known facts about how serial correlation biases OLS estimates of standard errors. First, positive serial correlation in the error term ($\rho > 0$) will cause an under-statement of the standard error while negative serial correlation will cause an over-statement. This is intuitive: positive serial correlation means that there is less information in each new year of data than OLS assumes. Second, the magnitude of the bias also depends on how serially correlated the independent variables is. In fact, when the independent variable is not serially correlated ($\lambda = 0$), there is no bias in the estimated standard errors. This second point is important in the DD context, where the intervention variable T_{st} is in fact very serially correlated. Indeed, for affected states, the intervention variable typically equals 0 period after period until one year where it turns to 1 and then stays at 1. In other words, the variable "whether a law was passed by state s by time t" varies little over time within a state, thereby exacerbating any serial correlation in the dependent variable. Finally, the magnitude of the problem depends on the length of the time series (T). All else held constant, as T increases, the bias in the OLS estimates of the standard errors worsens.

2.2 A Survey of DD Papers

This quick review suggests the relevance of a serial correlation problem for existing DD papers depends on three factors: (1) the typical length of the time series used; (2) the serial correlation of the most commonly used dependent variables; and (3) whether any procedures are use to correct for serial correlation.

Since these factors are inherently empirical, we collected data on published DD papers. We identified all DD papers published in 6 journals between 1990 and 2000: *the American Economic Review*, *the Industrial and Labor Relations Review*, *the Journal of Political Economy*, *the Journal of Public Economics*, *the Journal of Labor Economics*, and *the Quarterly Journal of Economics*. We classified a paper as "DD" if it met two following criteria. First, the paper must focus on specific interventions. For example, we would not classify a paper that regressed wages on unemployment as a DD paper (even though it might suffer from serial correlation issues as well). Second, the paper must use units unaffected by the law as a control. We found 92 such papers. For each of these papers, we determined the number of time periods in the study, the nature of the dependent variable, and the technique(s) used to estimate standard errors.

Table 1 summarizes the results of this exercise. We start with the lengths of the time series. Sixty-nine of the 92 DD papers used more than two periods of data. Four of these papers began with more than two periods but collapsed the data into two effective periods: before and after. Table 1 reports the distribution of time periods for the remaining 65 papers.[8] The average number of periods used is 16.5 and the median is 11. More than 75% of the papers use more than 5 periods of data. As we will see in the simulations below, lengths such as these are more than enough to cause serious under-estimation of the standard errors.[9]

The most commonly used variables in DD estimation are employment and wages.[10] Eighteen papers study employment and thirteen study wages. Other labor market variables, such as retirement and unemployment also receive significant attention, as do health outcomes. Most of these variables are clearly highly auto-correlated. To cite an example, Blanchard and Katz (1992) in their survey of regional fluctuations find strong persistence in shocks to state employment, wages and unemployment. It is interesting to note that first-differenced variables, which might have a tendency to exhibit negative auto-correlation (and thereby *over-state* standard errors) are quite uncommon. In short, the bulk of DD papers focus on outcomes which are likely positively serially correlated.

How do these 65 papers correct for serial correlation? The vast majority of papers do not address the problem at all. Only five papers explicitly deal with it. Of these five, four use a parametric AR(k) GLS-based correction. As we will see later, this correction does very little in the way of adjusting standard errors. The fifth allows for arbitrary variance-covariance matrix within state, one of the solutions we suggest in Section 4.

Two additional points are worth noting. First, 80 of the original 92 DD papers have a potential problem with grouped error terms as the unit of observation is more detailed than the level of the variation.[11] Only 36 of these papers address this problem, either by clustering standard errors or

[8] When a used several data sets with different time spans, we only recorded the shortest span.

[9] A period here is whatever unit of time is used. The very long time series in the data, such as the 51 or 83 at the 95^{th} and 99^{th} percentile arise because several papers use monthly or quarterly data.

[10] When a paper studies more than one variable, we record all the variables used.

[11] For example, the effect of state level laws is studied using individual level data.

by aggregating the data. Second, several informal techniques are used for dealing with the possible endogeneity of the intervention variable. For example, three papers include a lagged dependent variable in equation 1, seven include a trend specifically for treated states, 15 plot graphs of some form to examine the dynamics of the treatment effect, 3 examine whether there is an effect before the law, two see if the effect is persistent, and eleven formally attempt to do triple-differences (DDD) by finding another control group. We return to the issue on how these informal techniques interact with the serial correlation problem in Section 5.

In summary, our review suggests that serial correlation is likely an important problem for many existing DD papers and that this problem has been poorly addressed to date.

3 Over-Rejection in DD Estimation

While the survey above shows that most DD papers are likely to report under-estimated standard errors, it does not tell us how serious the problem is in practice. To assess magnitudes, we turn to a specific data set, a sample of women's wages from the Current Population Survey (CPS).[12]

We extract data on women in their fourth interview month in the Merged Outgoing Rotation Group of the CPS for the years 1979 to 1999. We focus on all women between 25 and 50 years old. We extract information on weekly earnings, employment status, education, age, and state of residence. The sample contains nearly 900,000 observations. We define wage as log(weekly earnings). Of the 900,000 women in the original sample, approximately 300,000 report strictly positive weekly earnings. This generates (50*21=1050) state-year cells with each cell containing on average a little less than 300 women with positive weekly earnings.

The correlogram of the wage residuals is informative. We estimate first, second and third auto-correlation coefficients of the residuals from a regression of the logarithm of wages on state and year dummies (the relevant residuals since DD includes these dummies). The auto-correlation coefficients are obtained by a simple regression of the residuals on the corresponding lagged residuals. We are therefore imposing common auto-correlation parameters for all states. The estimated first order auto-correlation is 0.51, and is strongly significant. The second and third order auto-correlation

[12] The CPS is one of the most commonly used data sets in the DD literature.

are high as well (.44 and .33 respectively), and decline much less rapidly than we would expect if the residual was following an AR(1) process.[13]

3.1 Placebo Interventions

To quantify the bias induced by serial correlation in the DD context, we randomly generate laws, which affect some states and not others. We first draw at random from a uniform distribution between 1985 and 1995.[14] Second, we select exactly half the states (25) at random and designate them as "affected" by the law (even though the law does not actually have an effect). The intervention variable T_{st} is then defined as a dummy variable which equals 1 for all women that live in an affected state after the intervention date, and 0 otherwise.

We can then estimate DD (equation 1) using OLS on these placebo laws. The estimation generates an estimate of the laws' "effect" and a corresponding standard error. To understand how well DD performs we can repeat this exercise a large number of times, each time drawing new laws at random. If DD provides an appropriate estimate for the standard error, we would expect that we reject the null hypothesis of no effect ($\beta = 0$) exactly 5% of the time when we use a threshold of 1.96 for the t-statistic.[15]

This exercise tells us about Type I error. Note that a small variant also allows us to assess Type II error, or power. After constructing the placebo intervention, T_{st}, we can replace the outcome in the CPS data by the outcome plus T_{st} times whichever effect we wish to simulate. For example, we can replace log(weekly earnings) by log(weekly earnings) plus $Tst * .02$ to generate a true .02 log point (approximately 2%) effect of the intervention.[16] By repeatedly estimating DD in this data (with new laws randomly drawn each time), we can assess how often DD finds an effect when there

[13] Solon (1984) points out that in panel data, when the number of time periods is fixed, the estimates of the auto-correlation coefficients obtained using a simple OLS regression are biased. Using Solon's generalization of Nickell's (1981) formula for the bias, the first order auto-correlation coefficient of 0.51 we estimate in the wage data, with 21 time periods, would correspond to a true auto-correlation coefficient of 0.6 if the data generating process were an AR(1). However, Solon's formulas also imply that the second and third order auto-correlation coefficients would be much smaller than the coefficients we observe if the true data generating process were an AR(1) process with an auto-correlation coefficient of 0.6. To match the estimated second and third order auto-correlation parameters, the data would have to follow and AR(1) process with an auto-correlation coefficient of 0.8.

[14] We choose to limit the intervention date to the 1985-1995 period to ensure having enough observations prior and post intervention.

[15] One might argue that the rejection rate could be higher than 5% as we might accidentally be capturing some real interventions with the randomization procedure. However, we will show later on that data manufactured to track the variance structure of the CPS also produces rejection rates similar to those in the CPS.

[16] The 2% effect was chosen so that there is sufficient power in the data sets we study.

actually is one.[17]

3.2 Basic Rejection Rates

The first row of Table 2 presents the result of this exercise when performed in the CPS micro data, without any correction for grouped error terms. We re-estimate equation 1 as described above for at least 200 independent draws of placebo laws. The control variables X_{ist} include 4 education dummies (less than high school, high school, some college and college and more) and a quartic in age as controls. We report the fraction of simulations in which the absolute value of the t-statistic was greater than 1.96, i.e. the fraction of simulations where the null hypothesis of no intervention effect was rejected at the 5% level.

The first column of row 1 shows the results of this exercise in the unaltered micro CPS data. Here, even though there is no true effect of the placebo laws, we find that the null of no effect is rejected a stunning 67.5% of the time. Thus, in this setup, DD is over-rejecting by a factor of thirteen.[18] The second column of this row performs a similar exercise but on the CPS data altered to contain a 2% effect (we added $.02 * T_{st}$ to the data). We find that, in this case, we reject the null hypothesis of no effect in 85.5% of the cases.

One important reason for this gross over-rejection has been described by Donald and Lang and (2001), who apply Moulton's (1990) general arguments to DD inference. The estimation above does not account for correlation within state-year cell; it does not allow for aggregate year-to-year shocks that affect all the observations within a state. In other words, 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 state and year cell. As noted earlier, while 80 papers suffer from this problem, only 36 correct for it.

To properly account for such shocks, one can assume that there is a random effect for each state-year cell in equation 1:

$$Y_{ist} = A_s + B_t + c\,X_{it} + \beta\,T_{st} + \nu_{st} + \epsilon_{it} \tag{4}$$

[17] In this case, we will count only correct rejections, i.e. the number of DD estimates that are significant and positive.

[18] The average of the estimated coefficients $\hat{\beta}$ was 0. Thus while OLS overestimates standard errors, the estimated coefficients are unbiased.

where ν_{st} are group random effects. The standard error can be corrected for the correlation at the group level introduced by the random effect. Row 2 reports rejection rates when we allow for state-year random effects using the White correction (1984) which allows for an arbitrary intra-group correlation matrix. We continue to find a 44% rejection rate.[19]

In row 3, we take a more drastic approach to solve this problem. We aggregate the data into state-year cells to construct a panel of states over time. To aggregate, we first regress log weekly earnings on the controls (education and age) and form residuals. We then compute means of these residuals by state and year. This leaves us with 50 times 21 state-year cells in the aggregate data. Using this aggregate data, we estimate:

$$\bar{Y}_{st} = \alpha_s + \gamma_t + \beta T_{st} + \epsilon_{st} \tag{5}$$

where the bar above the variables refers to the aggregation.

If the correlation within state-year cells were the only reason for over-rejection, aggregation ought to fully solve the problem, since it would make the variance-covariance matrix for the error term diagonal. Row 3 displays the results of multiple estimations of equation 5 for placebo laws. The rejection rate of the null of no effect is almost as high here as when use the micro data and correct the standard errors for clustering. In about 44% of the simulations, we reject the null hypothesis of no effect.[20]

These magnitudes suggest that failing to account for serial correlation when computing standard errors generates a dramatic bias. As we saw in equation 2.1, one important factor in the DD context is the serial correlation of the intervention variable T_{st} itself. In fact, we would expect the un-corrected estimates of the standard errors for the intervention variable to be consistent in any variation of the DD model where the intervention variable is not serially correlated. To illustrate this point, we construct a different type of intervention variable. As before, we randomly select half of the states to form the treatment group. However, instead of randomly choosing one date

[19] Practically, this is usually implemented by the "cluster" command in STATA. We also applied the correction procedure suggested in Moulton (1990). That procedure allows for a random effect for each group, which puts structure on the intra-cluster correlation matrices and therefore may perform better in finite samples. This is especially true when the number of clusters is small (if in fact the assumption of a constant correlation is a good approximation). The rate of rejection of the null hypothesis of no effect was not statistically different under the Moulton technique, possibly reflecting the fact that the number of clusters is large in this application.

[20] One might worry that the aggregation process, while it deals with the clustering problem, introduces heteroskedasticity in the data. However, the results in Table 2 do not change if we use standard heteroskedasticity-correction techniques.

after which all the states in the treatment group are affected by the law, we randomly select 10 dates between 1979 and 1999. The law is now defined as 1 if the observation relates to a state that belongs to the treatment group at one of the 10 intervention dates, 0 otherwise. In other words, the intervention variable is now repeatedly turned on and off, with its value yesterday telling us nothing about its value today, and thereby eliminating the serial correlation in T_{st}. In row 4, we see now that the null of no effect is rejected only 6% of the time in this case. Removing the serial correlation in the law removes the over-rejection. This strongly suggests that the bias in the standard errors is due to serial correlation, rather than other properties of the error terms.

In rows 5 and 6, we examine how rejection rates vary as we modify the number of affected states. When 12 states or 36 states are affected, the rejection rates are 39% and 43% respectively.

The placebo laws so far have been constructed in such a way that the intervention variable affects all treated states in the same year. In row 7, we create placebo laws such that the date of passage can differ across treated states. We randomly choose half of the states to form the treatment group but now randomly choose a passage date separately for each state (uniformly drawn between 1985 and 1995) in the treatment group. The T_{st} variable is still defined to equal 1 if state s has passed the law by time t, and 0 otherwise. The rejection rates continue to be high in this case, with the null of no effect being rejected about 35% of the time.

One might still worry at this point that factors other than serial correlation give rise to these large rejection rates. Perhaps we are by chance detecting actual laws (or other relatively discrete changes). Or perhaps other features of the wage data, such as state specific trends or other characteristics of the distribution of the error term, give rise to the over-rejection. To directly address all of these problems, we replicate our analysis in manufactured data. Specifically, we generate data whose variance structure in terms of relative contribution of state and year fixed effects matches the empirical variance decomposition in the CPS. The data is normally distributed and follows an AR(1) process with an auto-correlation parameter ρ. By construction, we can therefore be sure that there are no ambient trends and that the laws truly have no effect. Yet, in row 8, where we assume that ρ equals .8, we find a rejection rate roughly equal to what we found in the CPS, 37%.

3.3 Magnitude of Effects

These excess rejection rates are stunning, but they might not be particularly problematic if the estimated "effects" of the placebo laws are economically insignificant. We examine the magnitudes of the statistically significant intervention effects in Table 3. Using the aggregate data, we perform 200 independent simulations of equation 5 for placebo laws defined such as in row 3 of Table 2. Table 3 reports the empirical distribution of the estimated effects $\hat{\beta}$, when these effects are significant. Not surprisingly, this distribution appears quite symmetric: half of the false rejections are negative and half are positive. On average, the absolute value of the effects is roughly .02, which corresponds roughly to a 2 percent effect. Nearly 60% fall in the 1 to 2 % range. About 30% fall in the 2 to 3% range, and the remaining 10% are larger than 3%.

These magnitudes are especially large considering that DD estimates are often represented as elasticities.[21] For example, suppose that the law under study corresponds to be a 5% increase in the child-care subsidy. An increase in log earnings of .02 would correspond to an elasticity of .4. Similarly, in many DD estimates, the affected group is often only a fraction of the sample, meaning a measured 2% effect on the full sample would indicate a much larger effect on the treated subsample.[22]

To summarize, our findings suggest that the standard errors from the OLS estimation of equations 1 grossly understate the true standard errors, even if one accounts for group data effects in the micro data. In other words, reporting simple DD estimates and their standard errors without accounting for serial correlation will generate many spurious results.[23]

3.4 Varying Serial Correlation

How does over-rejection vary with the serial correlation in the dependent variable? We address this question in two different ways. First, we use other outcome variables in the CPS as left-hand side

[21] The DD estimates are normalized using the magnitude of the change in the policy variable of interest.

[22] For example, when studying the effects of changes unemployment insurance benefits on job search, one would examine the full sample of the unemployed, not only those who actually took up the program. This use of all people *eligible* for a program rather than all recipients is often motivated by an attempt to avoid the endogeneity caused by selective take-up.

[23] Mapping these findings to the existing literature on DD requires more care. Researchers often use informal techniques along with the mechanical DD procedure we have described. For example, to check for endogeneity, they will test whether a law appears to have an "effect" before it was passed. We discuss whether these informal techniques interact with the over-rejection rates in Section 5.

variables. Second, we experiment with various auto-correlation parameters in the manufactured data.

The first part of Table 4 estimates equation 5 on employment rate, average weekly working hours and change in log wages, as well as the original log weekly earnings. As before, the table displays the rejection rate of the null hypothesis of no effect in 200 independent simulations with random draws of the intervention variable, both in raw data and in data altered to create a 2% effect of the intervention. We also report in Table 4 estimates of the first, second and third order auto-correlation coefficients for each of these variables. As we see, the false rejection problem diminishes with the serial correlation in the dependent variable. As expected, when the estimate of the first-order auto-correlation is negative (as it is the case for change in log wages), we find that the conventional standard errors tend to *underestimate* the precision of the estimated treatment effect.

The second part of Table 4 uses manufactured data. The error structure is designed to follow an AR(1) process. As before, the manufactured data has the same number of states and years as the CPS data and is constructed so that the relative importance of state effects, year effects and the error term in the total variance matches the variance composition of the wage data in the CPS. Not surprisingly, the false rejection rates increase with the auto-correlation parameter in the AR(1) process. There is exactly the right rejection rate when there is no auto-correlation. But even at moderate levels, such as a ρ of .2 or .4, rejection rates are already two to four times as large as they should be. As noted earlier, with an AR(1) parameter of 0.8, the rejection rate using the standard OLS formula is close to what we observe in the CPS data. And again, when the auto-correlation is negative, there is under-rejection.

3.5 Varying the Number of States and Time Periods

The stylized exercise above focused data with 51 states and 21 time periods. Many DD papers use fewer states (or treated and control units), either because of data limitations or a desire to focus only on comparable controls. For similar reasons, several DD papers use fewer time periods. In Table 5, we examine how the over-rejection rate varies with these two important parameters. As before, we use the CPS as well as manufactured data to analyze these effects. We also examine rejection rates when we have added a treatment effect to the data.

Rows 1-4 and 10-13 show that varying the number of states does not change the extent of over-rejection.[24] Rows 5-9 and 14-18 vary the number of years. As expected, the extent of over-rejection falls as the time span gets shorter, but it does so at a surprisingly slow rate. For example, even with only 7 years of data, the over-rejection rate is 16% in the CPS, three times too large. Around 70% of the DD papers in our survey use at least that many periods. With 5 years of data, the rejection rate varies between 8% (CPS data) and 17% (manufactured data). When T=50, the rejection rate rises to nearly 50% in the manufactured data.

4 Solutions

4.1 Parametric Methods

A first natural solution to the serial correlation problem would be to specify the auto-correlation structure for the error term, estimate its parameters, and use equation (2) to estimate true standard errors. We implement several variations of this basic correction method in Table 6.

Row 2 performs the simplest of these parametric corrections, wherein an AR(1) process is estimated in the data.[25] This technique does little to solve the serial correlation problem: the rejection rate is still 34.5%. The failure here is in part due to the under-estimation of the auto-correlation coefficient. As is well understood, with short time-series the OLS estimation of the auto-correlation parameter is biased downwards. In the CPS data, OLS estimates a first-order auto-correlation coefficient of only 0.4. Similarly, in the manufactured data where we know that the auto-correlation parameter is .8, a $\hat{\rho}$ of .62 is estimated (row 7). If we impose a first-order autocorrelation of .8 in the CPS data (row 3), the rejection rate goes down to 12.5%, a clear but only partial improvement.

In row 8, we establish an upper-bound on the power of any potential correction, by examining

[24] In the CPS data, we vary the number of states by randomly selecting some states and discarding others. In both types of data, we continue to treat exactly half the states.

[25] Computationally, we first estimate the first order auto-correlation coefficient of the residual by regressing the residual on its lag, and then uses this estimated coefficient to form an estimate of the variance-covariance matrix of the residual. The matrix is block diagonal, with a matrix of the form Ω (in equation 2) in each block. The results are the same whether or not we assume each state has its own auto-correlation parameter.

how well the *true* variance-covariance matrix does. In other words, we use the variance-covariance matrix implied by an AR(1) process with ρ of .8 to estimate standard errors. As expected, the rejection rate is now indistinguishable from 5% when there is no effect. More interestingly, when there is a 2% effect, the rejection rate is now 32.3%. This will be a useful benchmark for other corrections.

Another problem with this parametric correction may be that we have not correctly specified the auto-correlation process. As noted earlier, an AR(1) does not fit the correlogram of wages in CPS. In rows 9 and 10, we use the manufactured data to see the effect of such mis-specification of the autocorrelation process. In row 9, we generate data according to an AR(2) process with $\rho_1 = .55$ and $\rho_2 = .35$. These parameters were chosen because they match well the estimated first, second and third auto-correlation parameters in the wage data when we apply the formulas to correct for small sample bias given in Solon (1984). We then correct the standard error assuming that the error term follows an AR(1) process. The rejection rate rises significantly with this mis-specification of the auto-correlation structure (30.5%).

In row 10, we use a process that provides an even better match of the time-series properties of the CPS data: the sum of an AR(1) (with auto-correlation parameter 0.95) and a white noise (the variance of the white noise is 13 % of the total variance of the residual).[26] When trying to correct the auto-correlation in this data by fitting an AR(1), we reject the null in about 39% of the case, close to what we found for the CPS data in row 2.

However, attempting to correct for the auto-correlation by specifying different processes does not look like a plausible option: it is clearly difficult to find the right process. In rows 4 and 5, we correct the standard errors in the CPS data by imposing the specific AR(2) and AR(1) plus white noise processes that we have seen match fairly well the CPS data. The rejection rates remain high.[27]

[26] Note that an AR(1) plus white noise seems a priori a very reasonable process for the CPS data. Indeed, even if the wage follows an AR(1) in the population, the repeated cross-section in the CPS implies that a non-persistent sampling error is added to the error term each year.

[27] Similar results hold if we estimate the parameters of the processes, as in row 2, rather than impose them.

4.2 Block Bootstrap

An alternative correction method with which we experiment is block bootstrap (Efron and Tibshirani, 1994). Block bootstrap is a variant of bootstrap which maintains the auto-correlation structure by keeping all the residuals of a state together. We implement block bootstrap as follows. We first estimate equation 1 and compute the residuals. This gives a vector of residuals ϵ_i for each state. We then draw for each state a new residual vector from this distribution (with replacement). Adding this residual back to the original predicted value gives us a new outcome variable Y_{it}^1. We then estimate equation 1 for this outcome. By repeatedly sampling from the residual distribution of ϵ_i, we can form different Y_{it}^k and repeat this exercise to estimate a sequence of $\hat{\beta}_k$. The distribution of these parameters then gives us a test statistic.

The results of the block bootstrap estimation are reported in Table 7. They are not encouraging. The rejection rates are still high: 35% in the CPS data (row 2) 29% in the manufactured data as well (row 3). The problem appears to be the small number of blocks or states. In row 4, when we allow for 400 states (and 200 states passing the law), block bootstrap delivers a close to correct rejection rate. Since very few applications in practice can rely on that many groups, block bootstrap does not appear to be a realistic solution to the serial correlation problem.

4.3 Empirical Variance-Covariance Matrix

Both the most parametric and the most non-parametric methods seem to fail because of lack of data: there are not enough time periods to estimate the time series process and perform a parametric correction, and not enough states for block bootstrap. However, the techniques we tried above did not make use of the fact that we have a large number of states that can be used to estimate the auto-correlation process in a flexible way. Specifically, suppose that the auto-correlation process is the same in all states. In this case, if the data is sorted by states and (by decreasing order of) years, the variance-covariance matrix of the error term is block diagonal, with 50 identical blocks of size T by T (where T is the number of time periods). Each of these block is symmetric, and the element $(i, i+j)$ is the correlation between ϵ_i and ϵ_{i-j}. We can therefore use the variation across the 50 states to estimate each element of this matrix, and use this estimated matrix to compute standard errors from equation 2. This is equivalent to treating the problem as a system of T seemingly unrelated equations estimated jointly (1 for each year), with 50 data points (one for each state) in each year. We implement this technique in Table 8. The rejection rate we obtain (in 200 simulations) is 7.75% in the CPS (row 2) and 10% in the manufactured data (row 9), a significant improvement over the two previous methods.

This correction method however has important limitations in practice. First, as the number of states drops, the rejection rates increase (rows 4, 6 and 8). A second obvious issue is that this method does not deliver consistent estimates of the standard error if the data generating process is not the same across all states. Finally, the technique has low power. In column 2 of row 2, we see that the rejection rate is only 8.5% when a true 2% effect is associated with the laws.

4.4 Arbitrary Variance-Covariance Matrix

This procedure can be generalized to an estimator of the variance covariance matrix which is consistent in the presence of *any* correlation pattern within states over time. Of course, we cannot consistently estimate each element of the matrix Ω in this case, but we can use a generalized White-like formula to compute the standard errors.[28] This estimator for the variance-covariance matrix is given by:

$$V = (X'X)^{-1} \left(\sum_{j=1}^{n_c} u_j' u_j \right) (X'X)^{-1}$$

where n_c is the total number of states, X is matrix of independent variables and u_j is defined for each state to be:

$$u_j = \sum_{t=1}^{T} e_{jt} x_{jt}$$

where the summation is over all elements in the state, e_{jt} is the residual at time t (in that particular state) and x_{jt} is a row vector of dependent variables (including the constant).[29] This estimator of the variance-covariance matrix is consistent as the number of states tends to infinity.

The results for this estimation procedure are shown in Table 9. Despite its generality, the arbitrary variance-covariance matrix does quite well. The rejection in the unaltered CPS data is

[28] This is analogous to applying the Newey-West correction (Newey and West 1987) in the panel context where we allow for all lags to potentially be important.

[29] This is implemented in a straightforward way by using the cluster command in STATA and choosing entire states (and not only state-year cells) as clusters.

6% (row 2, column 1). Moreover, the procedure has relatively high power compared to what we found in Table 8 (row 2, column 2). In the manufactured data, we can also see how well it does relative to the upper-bound. We saw in Tables 4 6, that with the correct covariance matrix, the rejection rate in the case of a 2% effect was 78% in manufactured data with no auto-correlation and 32% in AR(1) data with $\rho = .8$. The arbitrary variance-covariance matrix comes nearly these upper-bounds, achieving rejection rates of 74% and 27.5% respectively.

Again, however, rejection rates increase significantly above 5% when the number of groups declines (15% with 6 states, 8.5% with 10 states). This method, therefore, seems to work well, when the number of treated units is large enough.

4.5 Ignoring Time Series Information

Another possible solution to the serial correlation problem is to simply ignore the time series component in the estimation and when computing the standard errors.[30] To do this, one could simply average the data before and after the law and run equation 1 on this averaged outcome variable as a panel of length 2. The results of this exercise are reported in Table 10. The rejection rates are approximately correct now at 6%. Moreover, the power of .295 in row 2 is quite high relative to other techniques.

Taken literally, however, this solution will work only for laws that are passed for all the treated states at the same time. If states pass the law at different times, "before" and "after" are no longer the same and are not even defined for states that never pass the law. One can however slightly modify the technique in the following way. First, one regress Y_{st} on state fixed effects, year dummies, and any relevant covariates. One then divides the residuals of the *treatment states only* into two groups: residuals from years before the law, and residuals from years after the law. The estimate and the standard error come from an OLS regression of this two-period panel on an after dummy. This procedure does as well as the simple aggregation (row 3 vs. row 2) for laws that are all passed at the same time. It also does well when the laws are staggered over time (row 4).

The downside of these procedures (both raw and residual aggregation) is that they do poorly when the number of states is small. With 20 states, residual aggregation has a rejection rate of

[30] One could still use time-series data for specification checks.

9.5%, nearly twice too large. With only 6 states, the rejection rate is as high as 31.5%.[31] The extent of over-rejection in fact appears to increase faster in this case than in the clustering method presented above.

4.6 Randomization Inference

We have so far isolated two procedures that appear to do fairly well when the number of states is sufficiently large. In this section, we present an estimation technique that does well irrespective of the sample size. The principle behind the test is simple, and motivated by the simulation exercises carried out throughout this paper. To compute the standard error for a specific experiment, we propose to compute DD estimates for a large number of randomly generated placebo laws and to use the empirical distribution of the estimated effects for these placebo laws to form significance test for the true law.[32]

This test is closely related to the randomization inference test (or Fisher's exact test), discussed in the statistical literature (see Rosenbaum (1996) for an overview of this test and its applications). Before laying out the test in more details, it will be useful to give a simple, motivating illustration of the way randomization inference works. Suppose we have outcome data (such as sick days) on individuals, half of whom have received a flu shot. Define Y_i to be the outcome and T_i be the indicator for vaccination, both for individual i. Suppose that the we make the hypothesis that the treatment effect is constant:

$$Y_i = a_i + \theta T_i$$

Now, let us assume that the flu shot was administered to a random group, so that a_i is orthogonal to T_i. To estimate the effect of the shot, we could take the difference in mean outcomes between those receiving the shot and those not receiving it. Call this estimator $\hat{\theta}(T_i, Y_i)$. How can we test hypotheses about this parameter? We would use the OLS estimate of the standard error but this estimates rests on strong assumptions, such as homoskedasticity, normal distribution of the error and independence between observations. Instead, to test whether the coefficient is different from 0,

[31] At these small samples (as small as 12), the normal approximation to the t-distribution will clearly not work. But this alone is not causing the over-rejection since for small degrees of freedom, a threshold of 1.96 should not produce this high of a rejection rate. For example, for 6 degrees of freedom, a 1.96 threshold should only produce roughly a 10% rejection rate. Donald and Lang (2001) discuss inference in small-sample aggregated data sets.

[32] The code needed to perform this test is available from the authors upon request.

one can use the fact that under the null of no effect, people in the treatment and control groups are statistically the same. One can, therefore, generate a set of placebo interventions $\tilde{T}_i$ and estimate the "effect" of these placebos. Repeated estimation will give us a distribution of effects for such placebos. We can then observe where the original estimate $\hat{\theta}$ lies in this distribution to test the hypothesis that the coefficient is 0.

For example, to form a two-tailed test for whether the initial estimate is significantly different from 0 at the 5% level, we would use the 2.5% and 97.5% values in the distribution $\hat{\theta}(T_i, Y_i)$ as our cut-off value. The intuition behind this test is quite simple. Since the flu shot is assumed to be random, we can generate other *truly random*, zero-effect shots and see what estimates these produce *in the data we have*. Notice the advantages of this procedure. We are making no assumptions about the error term (except for the randomness of the flu shot itself). It need not be homoskedastic, normal or even independent across individuals. Further, this test does not rest on any large sample approximation: it is therefore valid for any sample size. Under the assumption that the treatment effect is constant across unit (i.e θ is not indexed by i), we could test other hypotheses in the same way: if we want to test whether the estimate is different from θ_0, we first form $Y_{i0} = a_i - \theta_0 T_{i0}$, to make the treatment and the control comparable under the null, and we apply the same technique to the transformed data.[33]

The statistical inference test we propose for the DD model follows naturally from this example. To form the estimate of the law's effect, we estimate by OLS the usual aggregate equation:

$$\bar{Y}_{st} = \alpha_s + \gamma_t + \beta T_{st} + \epsilon_{st}$$

We then generate a placebo law that affect 25 *randomly chosen* states in a *random* year, P_{st} and estimate using OLS:

$$\bar{Y}_{st} = \alpha_s + \gamma_t + \gamma P_{st} + \epsilon_{st}$$

Repeating this procedure many times for random placebo laws produces a distribution of $\hat{\gamma}$. Define $G(.)$ to be this distribution and $\hat{\gamma}$ to be the random variable drawn from this distribution. To test the hypothesis that our estimate is statistically different from 0, we simply need to ask where $\hat{\beta}$ lies in the $G(.)$ distribution. For example, to form a two-tailed test of level p, we would simply

[33] One can use the same technique to form entire confidence intervals.

identify the $\hat{\gamma}$ at the $\frac{p}{2}$ lower and upper tail of the distribution and use these values as cutoffs: If the estimated coefficient lies outside these two cutoff value, we reject the hypothesis that it is equal to 0, otherwise we accept it. As before, note that we have not used any information about the error term. Instead, we have relied solely on the random assignment of the laws.[34]

In Table 11, we assess how well this procedure performs in the aggregated CPS data as well as in manufactured data with a known auto-correlation structure. Again, we randomly generate intervention variables. For *each* randomly generated intervention T_{st} , we construct the test statistic for a 5% cutoff. We perform 200 independent draws of T_{st}. For each of these draws, we perform 400 draws of P_{st} to construct the distribution $G(\hat{\gamma})$.

We see that the basic randomization inference procedure leads us to reject the null hypothesis of no effect in 6% of the simulations (row 2, column 1). Notice that the procedure also seems fairly powerful, leading to rejecting in 23% of the cases when there is an effect (row 2, column 2). As before, in the manufactured data, we can compare its performance to the upper-bound. We saw in Tables 4 6, that with the correct covariance matrix, the rejection rate in the case of a 2% effect was 78% in manufactured data with no auto-correlation and 32% in AR(1) data with $\rho = .8$. The performance of randomization inference, 30% and 84 % is comparable to these upper-bounds.

There is an additional assumption behind this test, namely the requirement that we know the exact statistical process determining which units get treated (e.g. in the flu shot example, the statistician know that half of the sample was randomly selected to get the treatment). In practice, we will not know the true process by which laws are generated. If a particular law took place in 1988, and we are estimating its effect, should we assume when generating the placebo laws that the law could only have been passed in 1988, that it could have been passed at any random time between 1971 and 2000, or something else? Rows 3 to 5 consider the effect of using different (and simpler) assumptions about the law generating process. Each row allows the laws for the test distribution to lie in a different window around the actual law date. In row 3 for example, we assume that it lies in a 5 year window centered on the actual date. In row 4, we assume a 3 year window. In row 5, we force the placebo laws to occur in the same year as the actual laws; in other

[34] The theoretical justification for this test rests on the proof of the validity of using the randomization distribution. These tests use the strong assumption of randomization of the treatment to map out the distribution of the test statistic. The only difference in our case is that state laws are not random unconditionally, but instead that they are random *conditional* on the state fixed effects and year dummies.

words, we permute only which states were affected by the laws. As can be seen, both in terms of type I and type II errors, all the procedures produce very similar results.[35] The next few rows repeat the exercise using a progressively smaller number of states. Even with as few as 6 states, the procedure produces exactly the right rejection rates.

To summarize, Table 11 makes it clear that, of all the techniques we have considered, randomization inference performs the best. It removes the over-rejection problem and does so independently of sample size. Moreover, it appears to have power comparable to that of the other tests. Although randomization inference testing is well known in statistics, it is largely ignored in the econometrics literature. Difference in differences estimation, which deals with small effective sample size, and complicated error distribution, seems a particularly fertile ground for the application of this testing technique.

5 Implications for Existing Papers

This paper has highlighted an important problem with DD estimation and proposed several solutions to deal with it. What are the implications for the existing stock of papers which use the DD estimation technique but do not explicitly correct for serial correlation? We have already seen that most of these papers use long time series and variables which are likely quite auto-correlated. One possibility, however, is that some of the informal techniques used in these papers might indirectly help alleviate the auto-correlation problem.

As we noted earlier, researchers have developed a set of diagnostic tests that are often performed in conjunction with the estimation of equations 1 or 5. These tests are meant to assess the endogeneity of the interventions, something that is not a problem in our setup, as we construct random interventions that are by definition exogenous. It is possible that these tests may incidentally lessen the auto-correlation problem. In Table 12, we report rejection rates when we perform the OLS estimation of equations 1 and 5 in combination with several commonly used diagnostic

[35] Similar problems may arise in determining how many states were affected. If we see 25 of 50 states affected, was the process one in which each state had a 50% iid chance of being affected or was it one in which exactly 25 states will be affected? Simulations which vary this additional element show that the results are not affected by which selection process is assumed (even if it differs from the actual process). It is worth noting that while these results tell us that in simulations, using incorrect approximations do not make a difference, this may not be a general theoretical result.

techniques.[36] We concentrate on 3 data sets: micro CPS wage data (with clustering of the error term by state-year cell), aggregate CPS wage data, and manufactured data where the error term follows an AR(1) process with an autocorrelation parameter of 0.8.

The first diagnostic test looks for pre-existing "effects" of the law. We re-estimate the original DD with the T_{st} variable but also include a dummy for "this state will pass a law next year". We then reject the null hypothesis of no effect only if the coefficient on the law is significant *and* the coefficient on the pre-law dummy is either insignificant or opposite signed. This diagnostic test implies only a small reduction in the rejection rates.

The second diagnostic test examines persistence of the effect. We reject the null hypothesis of no effect if the estimated coefficient is statistically significant under OLS *and* the effect persists three years after the intervention date. Again, this does not lead to a significant reduction in the rejection rate in any of the data sets.

The third test looks for pre-existing trends in the treatment sample. We perform a regression in the pre-period and estimate whether there is a significant time trend in these years for the difference between control and treatment states). Under this test, we reject the null of no effect if the OLS coefficient is statistically significant *and* if there is no statistically significant pre-existing treatment trend or a treatment trend of opposite sign of the intervention effect. The rejection rates are lower under this test but remain above 20%.

Finally, in the last test we allow for a treatment specific trend in the data. Under this test, we reject the null if the OLS coefficient is significant and stays significant and of the same size after controlling in the regression for a yearly trend interacted with the treatment dummy. This diagnostic test leads to a bigger reduction in the rejection rate, especially in the aggregate data (.105 and .150), though this is still at least twice as large as we would want.[37] In short, the results in Table 12 therefore confirm that the existing stock of DD papers is very likely affected by the

[36] Of course, we can only study the formal procedures which people use. One can always argue that informal procedures (such as looking at the data) will lead one to avoid the serial correlation problem. Such a claim is by construction hard to test using simulations. The only way to defend it would be go back to the original papers and submit them to the procedures discussed above.

[37] Moreover, it is not clear that these tests are rejecting the "right" ones. Any stringent criterion will reduce the rejection rate, but how are we to assess whether the reduction is sensible? We investigated this by computing the marginal rejection rate of the randomization inference test, *conditional on passing the treatment trend diagnostic test.* The odds that, having passed the trend test, an intervention would pass the randomization inference test are only 26% suggesting that the treatment trend reduces the rejection rate but not in a way that alleviates the serial correlation problem.

estimation problem discussed in this paper.[38]

6 Conclusion

Our results suggest that, because of serial correlation, DD estimation as it is commonly performed grossly under-states the standard errors around the estimated intervention effect. While the bias induced by serial correlation is well understood in theory, the sheer magnitude of this problem in the DD context should come as a surprise to most readers. Since a large fraction of the published DD papers we surveyed report t-statistics around 2, our results suggest that the findings in many of these papers may not be as precise as originally thought to be and that far too many false rejections of the null hypothesis of no effect have taken place.

We propose three solutions to deal with the serial correlation problem in the DD context. Collapsing the data into pre- and post- periods or allowing for an arbitrary covariance matrix within state over time have been shown to be simple viable solutions when sample sizes are sufficiently large. Alternatively, a simple adaptation of the randomization inference testing techniques developed in the statistics literature appears to fully correct standard errors irrespective of sample size.

[38] We have attempted variants on these diagnostic tests, such as changing what constitutes an "effect" in the pre-period or what constitutes a trend. We have also attempted all the tests together. The results were qualitatively similar.

References

Abadie, Alberto, "Semiparametric Difference-in-Differences Estimators," Working Paper, Kennedy School of Government, Harvard University, 2000.

Blundell, Richard and Thomas MaCurdy, "Labor Supply," in *Handbook of Labor Economics*, Orley Ashenfelter and David Card (eds.), December 1999.

Efron, Bradley and Robert Tibshirani *An Introduction to the Bootstrap* Monograph in Applied Statistics and Probability, no 57, Chapman and Hall, 1994

Heckman, James, "Discussion," in *Empirical Foundations of Household Taxation*, Martin Feldstein and James Poterba (eds.), 1996.

Donald, Stephen and Kevin Lang, "Inferences with Difference in Differences and Other Panel Data," Boston University Working Paper, 2001.

MaCurdy, Thomas E., "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis," *Journal of Econometrics*, v. 18 (1982): 83-114.

Meyer, Bruce, "Natural and Quasi-Natural Experiments in Economics," *Journal of Business and Economic Statistics*, v. 13 (1995): 151-162.

Moulton, Brent R., "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables in Micro Units," *Review of Economics and Statistics*, v. 72 (1990): 334-338.

Newey, Whitney and K. D. West, "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent-Covariance Matrix," *Econometrica*, v. 55, pp. 703-708, 1987.

Nickell, Stephen, "Biases in Dynamic Models with Fixed Effects," *Econometrica*, v. 49 (1981): 1417-1426.

Rosenbaum, Paul, "Hodges-Lehmann Point Estimates of Treatment Effect in Observational Studies," *Journal of the American Statistical Association*, v. 88 (1993): 1250-1253.

Rosenbaum, Paul, "Observational Studies and Nonrandomized Experiments," S. Ghosh and C.R. Rao, eds, *Handbook of Statistics*, v. 13 (1996).

Solon, Gary, "Estimating Auto-correlations in Fixed-Effects Models," NBER Technical Working Paper no. 32, 1984.

White, Halbert, *Asymptotic Theory for Econometricians*, Academic Press, 1984.

Table 1: Survey of DD Papers[a]

Number of DD papers	92	
Number with more than 2 periods of data	69	
Number which collapse data into before-after	4	
Number with potention serial correlation problem	65	
Number with some serial correlation correction	5	
GLS	4	
Arbitrary variance-covariance matrix	1	
Distribution of time-span for papers with more than 2 periods	**Average**	16.5
	Percentile	**Value**
	1%	3
	5%	3
	10%	4
	25%	5.75
	50%	11
	75%	21.5
	90%	36
	95%	51
	99%	83
Informal manipulations of data	**Number**	
Graph time series of effect	15	
See if effect persists	2	
Examine lags of law to see timing of effect	2	
DDD	11	
Include trend specific to passing states	7	
Explicitly include lead to look for effect prior to law	3	
Include laggged dependent variable	3	
Number which have clustering problem	80	
Number which deal with it	36	
Most commonly used variables		
Employment	18	
Wages	13	
Health/Medical Expenditure	8	
Unemployment	6	
Fertility/Teen Motherhood	4	
Insurance	4	
Poverty	3	
Consumption/Savings	3	

[a] Notes: Data comes from a survey of all articles in six journals between 1990 and 2000:*American Economic Review*; *Industrial Labor Relations Review*; *Journal of Labor Economics*; *Journal of Political Economy*; *Journal of Public Economics*; and *Quarterly Journal of Economics*. We define an article as "Difference-in-Difference" if it: (1) examines the effect of a specific interventions and (2) uses units unaffected by the intervention as a control group.

Table 2: DD Rejection Rates for Placebo Laws[a]

Data	Law Type	Technique	Rejection Rate	
			No Effect	2% Effect
		A. REAL DATA		
CPS micro	25 states, one date	OLS	.675	.855
			(.027)	(.020)
CPS micro	25 states, one date	Cluster	.44	.74
			(.029)	(.025)
CPS aggregate	25 states, one date	OLS	.435	.72
			(.029)	(.026)
CPS aggregate	Serially uncorrelated laws	OLS	.06	.895
			(.014)	(.018)
CPS aggregate	12 states, one date	OLS	.433	.673
			(.029)	(.027)
CPS aggregate	36 states, one date	OLS	.398	.668
			(.028)	(.027)
CPS aggregate	25 states, multiple dates	OLS	.48	.71
			(.029)	(.026)
		B. MANUFACTURED DATA		
AR(1), $\rho = .8$	25 states, one date	OLS	.373	.725
			(.028)	(.026)

[a] Notes:

1. Each cell represents the rejection rate of the null hypothesis of no effect (at the 5% significance level) on the intervention (law) variable for randomly generated placebo interventions. The number of simulations for each cell is at least two hundred.
2. CPS data are data for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable unless otherwise specified is log weekly earnings. In row 3 to 7, data are aggregated to state-year level cells after controlling for demographic variables (education and age). Manufactured data are data generated so that the variances match the CPS variances. The ρ refers to the auto-correlation parameter in manufactured data.
3. All regressions also include, in addition to the intervention variable, state and year fixed effects. In the individual level regression they include the demographic controls as well.
4. Standard errors are in parenthesis and computed using the number of simulations.
5. "Effect" specifies whether an effect of the placebo law has been added to the data.

Table 3: Magnitude of DD Estimates[a]

CPS Aggregate Data

	No Effect		2% Effect	
	Positive	Negative	Positive	Negative
Rejection Rate	.18	.17	.715	.0075
	(.022)	(.022)	(.026)	(.005)
Average Coefficient	.02	-.02	.026	-.017
Fraction of effects < .01	0	0	0	0
	(.000)	(.000)	(.000)	(.000)
in (.01,.02]	.59	.58	.33	1
	(.028)	(.028)	(.027)	(.000)
in (.02,.03]	.31	.3	.39	0
	(.027)	(.026)	(.028)	(.000)
in (.03,.04]	.084	.12	.16	0
	(.016)	(.019)	(.021)	(.000)
> .04	.014	0	.12	0
	(.007)	(.000)	(.019)	(.000)

[a] Notes:

1. The positive (negative) columns report results for estimated effects of interventions which are positive (negative).
2. CPS data are data for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable is log weekly earnings. Data are aggregated to state-year level cells after controlling for the demographic variables (education and age).
3. All regressions also include, in addition to the intervention variable, state and year fixed effects.
4. Standard errors are in parenthesis and computed using the number of simulations.

Table 4 Varying Auto-correlation[a]

Data and Dependent Variable		Technique	Rejection Rate	
			No Effect	2% Effect
	A. REAL DATA			
	$\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3$			
CPS agg, Log wage	.509, .440, .332	OLS	.435	.72
			(.029)	(.026)
CPS agg, Employment	.470, .418, .367	OLS	.415	.698
			(.028)	(.010)
CPS agg, Hours worked	.151, .114, .063	OLS	.263	.265
			(.025)	(.025)
CPS agg, Changes in log wage	-.046, .032, 002	OLS	0	.978
			(.000)	(.009)
	B. MANUFACTURED DATA			
	ρ_1			
AR(1)	0	OLS	.053	.783
			(.013)	(.024)
AR(1)	.2	OLS	.123	.738
			(.019)	(.025)
AR(1)	.4	OLS	.19	.713
			(.023)	(.026)
AR(1)	.6	OLS	.333	.700
			(.027)	(.026)
AR(1)	.8	OLS	.373	.725
			(.028)	(.026)
AR(1)	-.4	OLS	.008	.7
			(.005)	.026)

[a] Notes:

1. Each cell represents the rejection rate of the null hypothesis of no effect (at the 5% significance level) on the intervention variable for randomly generated interventions. The number of simulations for each cell is at least two hundred.
2. CPS data are data for women between 25 and 50 in fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable unless otherwise specified is log weekly earnings. Data are aggregated to state-year level cells, after controlling for the demographic variables (education and age). Manufactured data are data generated so that the variances match the CPS variances.
3. All CPS regressions also include, in addition to the intervention variable, state and year fixed effects.
4. Standard errors are in parenthesis and computed using the number of simulations.
5. The variables $\hat{\rho}_i$ refer to the estimated auto-correlation parameters of lag i.

Table 5 Varying N and T[a]

	N	T	Rejection rate	Rejection rate
			No Effect	2% Effect
A. REAL DATA				
CPS aggregate	50	21	.435	.72
			(.029)	(.026)
CPS agg	20	21	.36	.53
			(.028)	(.029)
CPS agg	10	21	.425	.525
			(.029)	(.029)
CPS agg	6	21	.45	.433
			(.029)	(.029)
CPS agg	50	11	.29	.675
			(.026)	(.027)
CPS agg	50	7	.16	.63
			(.021)	(.028)
CPS agg	50	5	.08	.503
			(.016)	(.029)
CPS agg	50	3	.0775	.39
			(.015)	(.028)
CPS agg	50	2	.073	.315
			(.015)	(.027)
B. MANUFACTURED DATA				
AR(1), ρ=.8	50	21	.35	.638
			(.028)	(.028)
AR(1), ρ=.8	20	21	.35	.538
			(.028)	(.029)
AR(1), ρ=.8	10	21	.3975	.505
			(.028)	(.029)
AR(1), ρ=.8	6	21	.393	.5
			(.028)	(.029)
AR(1), $\rho = .8$	50	11	.335	.588
			(.027)	(.028)
AR(1), ρ=.8	50	5	.175	.5525
			(.022)	(.029)
AR(1), ρ=.8	50	3	.09	.435
			(.017)	(.029)
AR(1), ρ=.8	50	50	.4975	.855
			(.029)	(.020)

[a] Notes:

1. Each cell represents the rejection rate of the null hypothesis of no effect (at the 5% significance level) on the intervention variable for randomly generated interventions. The number of simulations for each cell is at least two hundred.
2. CPS data are data for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable unless otherwise specified is log weekly earnings. Data are aggregated to state-year level cells after controlling for the demographic variables (education and age). Manufactured data are data generated so that the variances match the CPS variances. The parameter ρ measures the auto-correlation.
3. All CPS regressions also include, in addition to the intervention variable state and year fixed effects.
4. Standard errors are in parenthesis and computed using the number of simulations.
5. N refers to the number of states used in the simulation and T refers to the number of years. When the CPS data is used, we randomly drop states or years to fulfill the criterion.

Table 6: Parametric Solutions[a]

Data	Technique	Estimated ρ	Rejection rate	
			No Effect	2% effect
		A. REAL DATA		
CPS agg	OLS		.435 (.029)	.72 (.026)
CPS agg	Standard AR(1) correction	.368	.345 (.027)	.705 (.026)
CPS agg	AR(1) correction imposing ρ=.8		.12 (.019)	.4435 (.029)
CPS agg	AR(2) correction imposing $\rho_1 = .55$ and $\rho_2 = .35$		.228 (.024)	.5725 (.029)
CPS agg	AR(1) + White Noise $\rho = .95$ and n/s=.13		.335 (.027)	.638 (.028)
		B. MANUFACTURED DATA		
AR(1), ρ=.8	OLS		.373 (.028)	.765 (.024)
AR(1), $\rho = .8$	Standard AR(1) correction	.622	.205 (.023)	.715 (.026)
AR(1), $\rho = .8$ imposing ρ=.8	AR(1) correction		.06 (.023)	.323
AR(2), $\rho_1 = .55$ $\rho_2 = .35$	Standard AR(1) correction	.444	.305 (.027)	.625 (.028)
AR(1)+ white noise $\rho = .95$, noise/signal=.13	Standard AR(1) correction	.301	.385 (.028)	.4 (.028)

[a] Notes:

1. Each cell represents the rejection rate of the null hypothesis of no effect (at the 5% significance level) on the intervention variable for randomly generated interventions. The number of simulations for each cell is at least two hundred.
2. CPS data are data for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable unless otherwise specified is log weekly earnings. Data are aggregated to state-year level cells, after controlling for the demographic variables (education and age). Manufactured data are data generated so that the variances match the CPS variances. An AR(1) + white noise process is the sum of an AR(1) plus an iid process, where the auto-correlation for the AR(1) component is given by ρ_i and the relative variance of the components is given by the noise to signal ratio (n/s).
3. All CPS regressions also include, in addition to the intervention variable, state and year fixed effects.
4. Standard errors are in parenthesis and computed using the number of simulations.
5. the AR(1) correction is implemented in stata using the xtgls command.

Table 7 Block Bootstrap[a]

Data	Technique	N	Rejection rate	
			No Effect	2% Effect
			A. CPS DATA	
CPS aggregate	OLS	50	.435 (.029)	.72 (.026)
CPS aggregate	Block Bootstrap	50	.35 (.028)	.60 (.027)
			B. MANUFACTURED DATA	
AR(1), ρ=.8	OLS	50	.38 (.028)	.735 (.025)
	Block Bootstrap	50	.285 (.026)	.645 (.028)
AR(1), $\rho = .8$	OLS	400	.415 (.028)	1 (.000)
	Block Bootstrap	400	.075 (.015)	.98 (.008)

[a] Notes:

1. Each cell represents the rejection rate of the null hypothesis of no effect (at the 5% significance level) on the intervention variable for randomly generated interventions. The number of simulations for each cell is at least two hundred.
2. CPS data are data for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable unless otherwise specified is log weekly earnings. Data are aggregated to state-year level cells after controlling for the demographic variables (education and age). Manufactured data are data generated so that the variances match the CPS variances.
3. All CPS regressions also include, in addition to the intervention variable, state and year fixed effects.
4. Standard errors are in parenthesis and computed using the number of simulations.

Table 8: Empirical Variance-Covariance Matrix[a]

Data	Technique	N	No Effect	2% Effect
			A. REAL DATA	
CPS agg	OLS	50	.435 (.029)	.72 (.026)
CPS agg	Empirical variance	50	.0775 (.015)	.085 (.016)
CPS agg	OLS	20	.36 (.028)	.53 (.029)
CPS agg	Empirical variance	20	.0825 (.016)	.08 (.016)
CPS agg	OLS	10	.425 (.029)	.525 (.029)
CPS agg	Empirical variance	10	.0825 (.016)	.0975 (.017)
CPS agg	OLS	6	.45 (.029)	.433 (.029)
CPS agg	Empirical variance	6	.165 (.021)	.1825 (.022)
			B. MANUFACTURED DATA	
AR(1), rho=.8	Empirical variance	50	.105 (.018)	.16 (.021)

[a] Notes:

1. Each cell represents the rejection rate of the null hypothesis of no effect (at the 5% significance level) on the intervention variable for randomly generated interventions. The number of simulations for each cell is at least two hundred.
2. CPS data are data for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable unless otherwise specified is log weekly earnings. Data are aggregated to state-year level cells after controlling for demographic variables (education and age). Manufactured data are data generated so that the variances match the CPS variances.
3. All CPS regressions also include, in addition to the intervention variable, state and year fixed effects.
4. Standard errors are in parenthesis and computed using the number of simulations.

Table 9: Arbitrary Variance-Covariance Matrix[a]

Data	Technique	# States	Rejection Rate No Effect	 2% Effect
A. REAL DATA				
CPS agg	OLS	51	.435 (.029)	.72 (.026)
CPS agg	Cluster	51	.06 (.014)	.27 (.026)
CPS agg	OLS	20	.36 (.028)	.53 (.029)
CPS agg	Cluster	20	.0625 (.014)	.1575 (.021)
CPS agg	OLS	10	.425 (.029)	.525 (.029)
CPS agg	Cluster	10	.085 (.016)	.1025 (.018)
CPS agg	OLS	6	.450 (.029)	.433 (.029)
CPS agg	Cluster	6	.15 (.021)	.1875 (.023)
B. MANUFACTURED DATA				
AR(1), ρ=.8	Cluster	50	.045 (.012)	.275 (.026)
AR(1), ρ=0	Cluster	50	.035 (.011)	.74 (.025)

[a] Notes:

1. Each cell represents the rejection rate of the null hypothesis of no effect (at the 5% significance level) on the intervention variable for randomly generated interventions. The number of simulations for each cell is at least two hundred.
2. CPS data are data for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable unless otherwise specified is log weekly earnings. Data are aggregated to state-year level cells after controlling for the demographic variables (education and age). Manufactured data are data generated so that the variances match the CPS variances.
3. All CPS regressions also include, in addition to the intervention variable, state and year fixed effects.
4. Standard errors are in parenthesis and computed using the number of simulations.

Table 10 Ignoring Time Series Data[a]

Data	Technique	N	Rejection rate	
			No Effect	2% Effect
		A. REAL DATA		
CPS agg Federal	OLS	50	.435	.72
			(.029)	(.026)
Federal	Simple Aggregation	50	.060	.295
			(.014)	(.026)
Federal	Residual Aggregation	50	.053	.210
			(.013)	(.024)
Staggered	Residual Aggregation	50	.048	.335
			(.012)	(.027)
CPS agg Federal	OLS	20	.36	.53
			(.028)	(.029)
Federal	Simple Aggregation	20	.060	.188
			(.014)	(.023)
Federal	Residual Aggregation	20	.095	.193
			(.017)	(.023)
Staggered	Residual Aggregation	20	.073	.210
			(.015)	(.024)
CPS agg Federal	OLS	10	.425	.525
			(.029)	(.029)
Federal	Simple Aggregation	10	.078	.095
			(.015)	(.017)
Federal	Residual Aggregation	10	.095	.198
			(.017)	(.023)
Staggered	Residual Aggregation	10	.103	.223
			(.018)	(.024)
CPS agg Federal	OLS	6	.450	.433
			(.029)	(.029)
Federal	Simple Aggregation	6	.130	.138
			(.019)	(.020)
Federal	Residual Aggregation	6	.315	.388
			(.027)	(.028)
Staggered	Residual Aggregation	6	.275	.335
			(.026)	(.027)
	B. MANUFACTURED DATA			
AR(1), ρ=.8 Federal	Simple Aggregation	50	.050	.243
			(.013)	(.025)
Federal	Residual Aggregation	50	.045	.235
			(.012)	(.024)
Staggered	Residual Aggregation	50	.075	.355
			(.015)	(.028)
AR(1), ρ=0 Federal	Simple Aggregation	50	.053	.713
			(.013)	(.026)
Federal	Residual Aggregation	50	.045	.773
			(.012)	(.024)
Staggered	Residual Aggregation	50	.105	.860
			(.018)	(.020)

[a] Notes:

1. Each cell represents the rejection rate of the null hypothesis of no effect (at the 5% significance level) on the intervention variable for randomly generated interventions. The number of simulations for each cell is at least two hundred.
2. CPS data are data for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable unless otherwise specified is log weekly earnings. Data are aggregated to state-year level cells after controlling for demographic variables (education and age). Manufactured data are data generated so that the variances match the CPS variances. All CPS regressions also include, in addition to the intervention variable, state and year fixed effects. Standard errors are in parenthesis and computed using the number of simulations.

Table 11: Randomization Inference[a]

Data	**Technique**	**N**	**Rejection rate** **No Effect**	**Rejection rate** **2% Effect**
A. CPS DATA				
CPS agg	OLS	51	.435	.72
			(.029)	(.026)
CPS agg	Randomization Inference	51	.065	.235
			(.014)	(.024)
CPS agg	RI 5 year window	51	.06	.23
			(.014)	(.024)
CPS agg	RI 3 year window	51	.06	.23
			(.014)	(.024)
CPS agg	RI same year	51	.04	.24
			(.011)	(.025)
CPS agg	OLS	20	.36	.53
			(.028)	(.029)
CPS agg	Randomization Inference	20	.045	.1
			(.012)	(.017)
CPS agg	RI 5 year window	20	.04	.125
			(.011)	(.019)
CPS agg	RI 3 year window	20	.065	.095
			(.014)	(.017)
CPS agg	RI same year	20	.045	.115
			(.012)	(.018)
CPS agg	OLS	10	.425	.525
			(.029)	(.029)
CPS agg	Randomization Inference	10	.055	.115
			(.013)	(.018)
CPS agg	RI 5 year window	10	.055	.095
			(.013)	(.017)
CPS agg	RI 3 year window	10	.05	.125
			(.013)	(.019)
CPS agg	RI same year	10	.07	.115
			(.015)	(.018)
CPS agg	OLS	6	.450	.433
			(.029)	(.029)
CPS agg	Randomization Inference	6	.04	.07
			(.011)	(.015)
CPS agg	RI 5 year window	6	.035	.065
			(.011)	(.014)
CPS agg	RI 3 year window	6	.03	.06
			(.010)	(.014)
CPS agg	RI same year	6	.055	.07
			(.013)	(.015)
B. MANUFACTURED DATA				
AR(1), rho=0.8	Randomization Inference	50	.05	.3
			(.011)	(.025)
iid data	Randomization Inference	50	.08	.84
			(.019)	(.026)

[a] Notes:

1. Each cell represents the rejection rate of the null hypothesis of no effect (at the 5% significance level) on the intervention variable for randomly generated interventions.
2. CPS data are data for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable unless otherwise specified is log weekly earnings. Data are aggregated to state-year level cells after controlling for education and age. Manufactured data are data generated so that the variances match the CPS variances. All CPS regressions also include, in addition to the intervention variable, state fixed effects and year fixed effects. Standard errors are in parenthesis and computed using the number of simulations.

Table 12: Effect of Informal Tests[a]

Data:	CPS Aggregate		AR(1), $\rho = .8$	
	No Effect	2% effect	No Effect	2% effect
$Ho : \beta = 0$ **Rejected if:**				
OLS coef significant (OLS)	.360 (.034)	.725 (.031)	.345 (.034)	.625 (.031)
OLS+No effect before the law	.400 (.033)	.634 (.032)	.325 (.033)	.565 (.035)
OLS+Persistence of effect	.378 (.033)	.602 (.031)	.305 (.033)	.565 (.035)
OLS+No treatment specific trend in pre-period	.248 (.029)	.518 (.035)	.235 (.030)	.530 (.035)
OLS+Coef significant with treatment specific trend (TREND)	.106 (.018)	.418 (.018)	.150 (.025)	.375 (.034)

[a] Notes:

1. Each cell represents the rejection rate of the null hypothesis of no effect (at the 5% significance level) on the intervention variable for randomly generated interventions. The number of simulations for each cell is at least two hundred.
2. CPS data are data for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1979 to 1999. The dependent variable unless otherwise specified is log weekly earnings. Data are aggregated to state-year level cells after controlling for demographic variables (age and education). Manufactured data are data generated so that the variances match the CPS variances.
3. All CPS regressions also include, in addition to the intervention variable, state and year fixed effects.
4. Standard errors are in parenthesis and computed using the number of simulations.