

2019-04-02

The promise and pitfalls of differences-in-differences: reflections on '16 and Pregnant' and other applications

Ariella Kahn-Lang & Kevin Lang. (2019) "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications," *Journal of Business & Economic Statistics*, <https://doi.org/10.1080/07350015.2018.1546591>

<https://hdl.handle.net/2144/34895>

"Downloaded from OpenBU. Boston University's institutional repository."

The Promise and Pitfalls of Differences-in-Differences: Reflections on *16 and Pregnant* and Other Applications*

Ariella Kahn-Lang and Kevin Lang
Harvard Kennedy School, Harvard University (Kahn-Lang)
and
Department of Economics, Boston University (Lang)

October 18, 2018

Abstract

We use the exchange between Kearney/Levine and Jaeger/Joyce/Kaestner on *16 and Pregnant* to reexamine the use of DiD as a response to the failure of nature to properly design an experiment for us. We argue that 1) any DiD paper should address why the original levels of the experimental and control groups differed, and why this would not impact trends, 2) the parallel trends argument requires a justification of the chosen functional form and that the use of the interaction coefficients in probit and logit may be justified in some cases, and 3) parallel trends in the period prior to treatment is suggestive of counterfactual parallel trends, but parallel pre-trends is neither necessary nor sufficient for the parallel counterfactual trends condition to hold. Importantly, the purely statistical approach uses pretesting and thus generates the wrong standard errors. Moreover, we underline the dangers of implicitly or explicitly accepting the null hypothesis when failing to reject the absence of a differential pre-trend.

Keywords: experimental design, natural experiments, preexisting trends

*We are grateful to the five authors of the original papers on *16 and Pregnant* and to Ivan Fernandez-Val, Shulamit Kahn, Adrienne Sabety and Jonathan Roth for helpful comments and suggestions. The usual caveat applies with particular force in this case.

1 Introduction

While it has a long history going back at least to Snow’s 1855 work on death rates from cholera in London and thus predating its renaissance as a leading approach in empirical economics, differences-in-differences (DiD) is plausibly *the* showpiece of the credibility revolution in empirical economics. Angrist and Pischke 2010 describe it as “probably the most widely applicable design-based estimator.”

Describing DiD, certain instrumental variables, and regression discontinuity estimators as techniques for conducting ‘natural experiments’ can be an incredibly effective rhetorical device. However, this terminology obscures the fact that these approaches are typically not design-based estimators but rather solutions for a lack of a randomized experimental design. In a true randomized experiment, barring selective attrition or terrible luck, the baseline characteristics, including the baseline value of the outcome variable, of the experimental and control groups are similar. While controlling for baseline values may (or may not) increase the precision of the estimated treatment effect, there is no need for differences-in-differences. Simple single-differences provide an unbiased estimate of the treatment effect. We only *need* DiD because ‘nature’ did not conduct a randomized trial for us.

This is not a criticism of DiD or other ‘natural experiment’ approaches. Instrumental variables, for example, can be a very useful solution to a lack of randomized experimental design. But we note that in the case of instrumental variables, economists have put a lot of thought into the interpretation of the results when instrumental variables is used to correct this. As discussed in Lang 1993 and formalized in Angrist and Imbens 1994, under certain conditions the instrumental variables estimator captures the local effect of treatment on those individuals whose treatment status was affected by the instruments, which Imbens and Angrist called the *local average treatment effect*. This is a different experiment than the one in which treatment is randomized. This point is perhaps even more important when assessing the results of studies based on regression discontinuity. In the case of DiD, similar thought is necessary for considering the interpretation of results; this requires both logical

and statistical evaluation of the identification assumptions in DiD.

Importantly, most of the issues we discuss are familiar from standard regression analysis. Indeed, since DiD has a regression representation, it cannot *inherently* provide more compelling evidence of a causal effect than regression analysis does. We face the same issues of whether the model is properly specified and whether, conditional on the controls, the variable of interest is orthogonal to the error term. In this case, the controls include membership in the experimental or control group and time period indicators for before or after the ‘experimental’ intervention. The variable of interest is the interaction between experimental group and post-intervention. And, with limited data, we face the usual risks of overfitting and loss of power when we try to address potential correlation with the error term by including further controls.

We use the papers on *16 and Pregnant* by Kearney and Levine 2015, hereafter KL, and Jaeger et al. 2018, hereafter JJK, as a jumping off point for reexamining the use of DiD as a response to the absence of an experimental design with randomly assigned treatment. KL examine the impact of the introduction of the show on teen pregnancy rates. Using a DiD design comparing areas with differing MTV viewership prior to the introduction of *16 and Pregnant*, they find that the show led to a 4.3 percent reduction in teen births. JJK contend that the identification used by KL was flawed, primarily because pre-trends were not parallel, making the assumption of common trends under the counterfactual, as required for DiD identification, less plausible.

We choose these papers not because they provide egregious examples of errors, which they do not, but because they are thoughtful but different approaches that help to underscore some of the issues we raise. We note that KL and, therefore, JJK are somewhat unusual in the DiD literature for two reasons. First, they have a continuum of treatments rather than an experimental and a control group. Second, to address potential selection in *16 and Pregnant* viewership, they rely on prior MTV viewership in the relevant time slot as an instrument. When we relate our analysis to their exchange, we will focus on the reduced form which

is the relation between the change in teen pregnancy and potential viewership. Consistent with our approach, much of their analysis and debate concerns whether the relation between (potential) *16 and Pregnant* viewership and teen pregnancy changed after the show went on the air. Because our objective is to make general points rather than comments specifically on the exchange between KL and JJK, for consistency with the usual DiD framework, we generally discuss their findings as if there were a single treatment and therefore separate experimental and control groups.

We make three principal points:

1. Any DiD paper should address why the original levels of the experimental and control groups differed or, in other words, why the experimental design failed. The researcher should then provide justification for the assertion that the same mechanism would not impact trends. If the researcher believes that the groups did not differ before the intervention, that, too, must be established.
2. Determining that two groups would have experienced parallel trends requires a justification of the chosen functional form. It is a mistake to view this purely as a statistical process. We argue that this decision should be made such that the estimated treatment effect is consistent with the perceived counterfactual trends and note the role of ‘theory’ in this process. In particular, in contrast with Ai and Norton 2003, we argue that it is frequently appropriate to use the probit or logit coefficients on the interaction term in a DiD model. Importantly, it can matter whether we believe the ‘correct’ model is a linear probability model, probit or logit since they assume different counterfactuals.
3. Simply comparing the trends in both groups prior to intervention, or ‘pre-trend testing,’ is insufficient to establish parallel trends as the appropriate counterfactual in the treatment period. The existence of parallel trends in the period prior to treatment is suggestive of parallel trends in the treatment period, but it is neither necessary nor sufficient for the parallel trends condition to hold. We argue that the presence

of parallel trends in the pre-period does not guarantee these trends would have continued in the absence of treatment. Further, failure to reject the null hypothesis of non-parallel trends does not confirm the existence of parallel trends (type II errors). In a regression context, it is natural to consider whether adding a linear (or other) trend interacted with group membership changes the results. Related to this, we also note that the purely statistical approach uses pretesting and thus generates the wrong standard errors.

2 DiD in the Absence of a Randomized Experiment

In the potential outcomes framework, we write $E(Y_{gt}(D_1))$ to represent the expected outcome of group g in year t if it is treated and $E(Y_{gt}(D_0))$ if it is not. A standard experiment ensures that, subject to sampling variation and, assuming no attrition, potential outcomes for the control and treatment group are the same with and without treatment. Letting $S = 1$ denote membership of group g in the treatment group, this means that

$$E(Y_{gt}(D_1)|S_g = 1) = E(Y_{gt}(D_1)|S_g = 0) \tag{1}$$

and

$$E(Y_{gt}(D_0)|S_g = 1) = E(Y_{gt}(D_0)|S_g = 0). \tag{2}$$

A problem arises when persistent factors that may be correlated with the outcome of interest are correlated with membership in the experimental group, that is that (1) and/or (2) does not hold. In this case we do not have a properly designed experiment. The key to identification in a DiD is that although outcome levels differ in the pre-period, outcomes between the pre-period and the treatment period (denoted by 0 and 1) would have moved in parallel in the absence of treatment, that is

$$\begin{aligned}
& E(Y_{g1}(D_0)|S_g = 1) - E(Y_{g0}(D_0)|S_g = 1) \\
&= E(Y_{g1}(D_0)|S_g = 0) - E(Y_{g0}(D_0)|S_g = 0).
\end{aligned} \tag{3}$$

The implicit underlying assumption of this model is that there is some initial difference between the groups that shifts the dependent variable vertically without affecting the slope of the time trend. This assumption must be justified, which requires discussion of why levels varied in the pre-period and how this should influence our interpretation of likely counterfactual changes.

When we do not have an experiment, there *may* be a case that assignment is as good as random. In this case, DiD can still be useful for three reasons. First, controlling for pre-period outcomes may be beneficial in that it improves precision and therefore efficiency. Second, because DiD requires weaker assumptions than single-differences, DiD is still a valid design even if it is unnecessary. Lastly, DiD techniques used for assessing the parallel trends assumption can be used for testing the assumption of no systematic differences between groups in the absence of treatment. Our discussion of testing for pre-trends in section 4 applies in this case as well.

Although the model can be written more generally, we present the model in its more common forms.

$$E(h(Y_{gt}(D))) = \beta D_{gt} + \gamma_t + \alpha_g. \tag{4}$$

This can be represented in regression form

$$h(y_{gti}) = \beta D_{gt} + \gamma_t + \alpha_g + \varepsilon_{gti}. \tag{5}$$

In a true experiment, $E(\alpha_g D_{gt}|t) = 0$. Therefore by estimating the model only on the post-period, leaving out the group dummies, which are perfectly collinear with D , we can

obtain a consistent estimate of β . When we do not have a true experiment, the assumption that $E(\alpha_g D_{gt}|t) = 0$ is usually less compelling, and in such cases we need to control for S , membership in the (eventually) treated group. Similarly, in the case of a small number of groups, we may believe that treatment is as good as randomly assigned at the group level, but that there are still group specific time-invariant factors. With a small number of groups, orthogonality will almost always be violated in the data so that group dummies are necessary. There are also important issues regarding the correct standard errors when there is a group-time specific error (η_{gt}). See Cameron and Miller 2015 for a review. Note that t is correlated with D by construction. So unless we have strong evidence that there are no time effects, we must also control for time in the regression. In a true experiment or in DiD, including the full set of α terms, when not perfectly collinear with D , may be helpful if they absorb enough of the error variance.

A common case, for example probit or logit models, is when y is not observed but is a latent variable, y^* , where $y_{gt}^* > 0 \implies Y_{gt} = 1$ and $Y_{gt} = 0$ otherwise. For the most part, our concerns will be similar except that the incidental parameters problem precludes using the full set of α terms rather than S unless the number of observations in each group is large. If not, the α terms must be replaced with a dummy variable for being part of the experimental groups. If group membership is highly predictive of outcomes, this may substantially decrease the precision of estimates. There are some issues specific to interpreting coefficients on probit or logit which we discuss in section 3.

Nothing in this section is original to us, but it serves to fix ideas and notation. Presenting DiD in a standard regression framework helps us make our points more simply and allows for an easy extension to a continuous treatment and even instrumental variables as in the studies of *16 and Pregnant*. Importantly, identification of a causal treatment effect is dependent on the assumption that $E(\varepsilon D|t, S) = 0$ (or $E(\varepsilon Z|t, S) = 0$ for the case of IV), the regression counterpart to the parallel trends assumption.

The debate between KL and JJK focuses on whether the pre-period experience suggests

that a common counterfactual trend is plausible. Therefore, both papers implicitly accept the view that they do not have a well-defined experiment and that this, in turn, creates challenges for causal inference that must be addressed. This means that the credibility with which we can interpret β as a causal effect depends on how confident we are that $E(\varepsilon D|t, g) = 0$. Both sets of authors are aware of this requirement for causal interpretation but come to different conclusions about its plausibility.

3 Choosing a Counterfactual Functional Form

In this section, we will proceed under the assumption that a researcher has identified a setting in which, despite having different initial levels, in the absence of the intervention, the variable of interest would have changed similarly in the control and experimental groups. But what does ‘similarly’ mean when the groups are different initially? When using a DiD model, we need not only to establish that the two groups would have moved similarly, but also that these patterns would have been consistent with the functional form of the chosen model specification. As Meyer 1995 points out, unless the distribution of outcomes is initially the same for the experimental and control groups, the effect of any changes associated with time cannot be the same both if the model is specified in, for example, levels and if it is specified in logarithms. Note that this is true even if the means are initially equal provided that the distributions differ. Choosing the functional form for (5) is a key decision on the part of the researcher and requires justification.

Some common functional form assumptions in DiD models are that group outcomes would have moved by the same absolute amount, by the same percentage, or according to a logit/probit model. While all of these assumptions are plausible in certain contexts, it is often far from obvious which functional form properly represents the counterfactual. For example, if the probability of some event in the control group increases from .80 before the treatment period to .82 after treatment, it is not obvious what our counterfactual should be. Say the

treatment group had a probability of .5 of this event in the pre-period. Since, on net, two percentage points of the control group shifted, perhaps we should expect a counterfactual probability of .52 in the treatment group. This would be consistent with the standard linear (probability) model, in which groups move by the same absolute amount under the counterfactual. Alternatively, since (at least on net) 10 percent of those in the control group who had a value of zero in the pre-period shifted to a value of one, we might expect 10 percent of the zeros in the experimental group to have shifted so that our counterfactual is a rise from .50 to .55. Logit and probit fall in between these two counterfactuals. Based on the shape of the logistic and normal distributions, both predict a counterfactual of approximately .53. We note that probit assumes that in the counterfactual, the control and treatment groups would have moved by the same number of standard deviations (of the standard normal error) while logit makes a similar assumption in terms of logits.

Which counterfactual is appropriate depends on the setting. For example, one might believe based on historical trends that unemployment rates for two groups would have moved proportionally. In contrast, if we examined the effect of a policy that affected third but not fourth graders, we might expect the same progress, measured in grade equivalents (which measures performance on a fixed exam in terms of the grade at which that performance is normally achieved), in both grades in the absence of the policy. Absent a theory, it is hard to make a strong case for any of these three counterfactuals. If they give different answers, we should have less confidence about drawing strong conclusions unless we have a strong justification for the chosen counterfactual.

In a highly-cited and influential paper, Ai and Norton 2003 point out that the coefficients from the probit or logit estimation of a DiD model give the ‘wrong answer.’ They make their point more generally, but for simplicity we will use logit to illustrate it. Let

$$y_{igt}^* = c + \beta D_{gt} + \alpha d_g + \gamma_t + \varepsilon_{igt} \tag{6}$$

where y^* is an unobserved latent variable such as the tendency to give birth as a teen, d is a

dummy for membership in the experimental group, $post$ is a dummy for the post experiment period and D is an interaction between being in the experimental group and the post-period. We observe a birth, $y_{igt} = 1$, if and only if $y_{igt}^* > 0$. Note that this equation is analogous to (5) but has a single variable for the experimental group (to avoid the incidental parameters problem). Given that we are using logit, we have assumed that

$$P(y_{igt} = 1) = \frac{e^{c+\beta D_{gt}+\gamma t+\alpha d_{ig}}}{1 + e^{c+\beta D_{gt}+\gamma t+\alpha d_{ig}}}. \quad (7)$$

The DiD effect on P is given by

$$DiD = \left(\frac{e^{c+\beta+\gamma+\alpha}}{1 + e^{c+\beta+\gamma+\alpha}} - \frac{e^{c+\alpha}}{1 + e^{c+\alpha}} \right) - \left(\frac{e^{c+\gamma}}{1 + e^{c+\gamma}} - \frac{e^c}{1 + e^c} \right). \quad (8)$$

After rearranging terms, the numerator of this expression is given by

$$N = e^{c+\alpha} (e^{\beta+\gamma} - 1) (1 + e^{c+\gamma}) (1 + e^c) - e^c (e^\gamma - 1) (1 + e^{c+\beta+\gamma+\alpha}) (1 + e^{c+\alpha}). \quad (9)$$

Now suppose that $\beta = 0$. Then the numerator reduces to

$$N = e^c (1 - e^\gamma) (1 - e^\alpha) (1 - e^{2c+\gamma+\alpha}). \quad (10)$$

Thus when β is 0, the ‘true’ DiD estimate will only be 0, when γ equals 0, α equals 0 or $2c+\alpha+\gamma$ equals 0. The first case corresponds to one in which we could do a simple difference over time using only the experimental group and the second to one in which we could use the simple difference between the experimental and control groups, while the third is obviously special. Ai and Norton argue that because the DiD estimate can be non-zero when $\beta = 0$ (or conversely, the DiD estimate can be 0 when $\beta \neq 0$), testing whether β equals 0 in (6) is wrong.

But the answer is wrong only in the sense that the change in absolute probability between

the pre-period and treatment period can differ between the control and experimental groups even when the interaction term in probit or logit is zero. Similarly, a significant interaction term from probit or logit need not mean that the change in absolute probability differed between the two groups.

Implicit in Ai and Norton’s argument is that the counterfactual should be a constant absolute change in the probability of an event between the pre-period and treatment period, consistent with the standard linear model. We have no disagreement with the formal part of Ai and Norton’s argument, but we believe that this is more a question of choosing the appropriate counterfactual functional form than a simple question of ‘right’ or ‘wrong.’ If researchers choose to use a nonlinear model to estimate the DiD, presumably they believe that this model accurately captures the effect of the explanatory variables and thus that an interaction term of 0 is consistent with the counterfactual. In this case, adjusting the coefficient on the interaction term to test the counterfactual implicit in the linear model would be a mistake. Blundell and Costa Dias 2009 point out that the interaction term can be translated into a more intuitively accessible metric of the effect in the treated group by calculating $F(\beta_0 + \beta_1 \text{treated_group} + \beta_2 \text{post} + \beta_3 \text{treated} * \text{post}) - F(\beta_0 + \beta_1 \text{treated_group} + \beta_2 \text{post})$ where F is the relevant CDF.

Athey and Imbens 2006 propose an approach they call changes-in-changes (CIC). One first calculates the change in the outcome at each quantile of the outcome variable in the control group. This gives the counterfactual at each initial value of the outcome variable. If we are interested in the mean of the counterfactual, we can reweight these changes to match the pre-period distribution of outcomes in the experimental group. Somewhat more formally, we write

$$Y_{igt} = g(u_i, t|D). \tag{11}$$

This means that the outcome, Y , is a function of time, treatment, and a stochastic term, u_i . Importantly, Y_{igt} does not directly depend on group membership. Therefore, in the

pre-period, if a member of the control group has the same Y as a member of the treatment group, she also has the same u . In addition, Athey and Imbens assume that g is strictly increasing in u and that the distribution of u within a group does not change over time. This ensures that if we observe two individuals from the same group at different times but at the same quantile of the group's distribution, they have the same u regardless of whether they are the same person. This, in turn, means that we can identify the treatment effect using DiD on a set of people all of whom happen to have the same draw of u , which we describe as a pseudo-person. Thus, if for example, the 10th percentile outcome in the control group is 50 in the pre-period and 65 in the post-period, we then ascribe a counterfactual outcome of 65 for every pseudo-person in the experimental group with a pre-period outcome of 50.

In the simple case when individuals keep the same u across periods, we mechanically have a constant distribution of u within group. Since the no-treatment outcome is monotonic in u , we have rank invariance. Athey and Imbens do not require that each individual maintain the same u , as long as the distribution of u stays constant within group over time. Thus two individuals with the same u in different periods are not necessarily the same person, but a pseudo-person. As a result, if each person draws a new u from the distribution in each period, we have rank invariance among pseudo-people but not in the population, and thus we cannot estimate the distribution of individual effects.

A crucial feature of CIC is that the counterfactual can be very different from the standard DiD counterfactual. To see this, suppose that in the control group the distribution of outcomes widens but the mean remains constant. The standard counterfactual is therefore no change. But if, in the pre-treatment period, the experimental group had Y s drawn disproportionately from the upper end of the control group distribution, the CIC counterfactual would be a positive change. It is not self-evident which counterfactual is correct. Dropping the assumption of rank invariance and thus allowing individuals to get new draws leaves the CIC approach particularly open to this concern.

CIC strikes us as a helpful approach, but, as discussed in the original paper, we cannot

use any observations in the control group without a counterpart in the experimental group in the pre-period since using them requires parametric assumptions, and we do not have a counterfactual for any experimental observations without a counterpart in the control group in the pre-period. Note that when the outcome is binary, deriving a point estimate of the counterfactual requires strong assumptions. Under these assumptions, the counterfactual in our earlier example with binary outcomes is given by the rise from .50 to .55.

In the example of the papers on *16 and Pregnant*, because the dependent variable is the logarithm of the teen birth rate, the assumption is that in the absence of the intervention the teen pregnancy rate in all DMAs (designated market areas) would have fallen by a constant proportion. For instance, a DMA with an initial rate of 45 per thousand would have experienced a decline one and a half times that of a DMA with an initial rate of 30. This is neither obviously right nor obviously wrong. Assessing its validity requires understanding why the teen birthrate was dropping, a topic about which Kearney and Levine are certainly more knowledgeable than we are.

But we can imagine many alternative hypotheses. For example, if teen pregnancies consist of two components, intended pregnancies which are not dropping and unintended pregnancies which are, then we might expect teen pregnancies to fall by a higher proportion in high birth-rate areas. Whether this is true obviously depends on the correlation between intended and unintended pregnancies and their relative frequencies. Alternatively if the rate of intended pregnancies were much more similar across DMAs and intended pregnancies were declining, then a constant percentage point change would be a more accurate counterfactual.

To some extent, when sufficient data exist, these concerns can be addressed by examining the patterns in the data prior to the ‘experiment,’ although we raise concerns about pre-trend testing in the next section. By comparing trends for the treatment and control groups prior to treatment, it may be possible to identify the functional form of the relation. However, in practice, this requires many pre-periods with precise estimates in each period. Most researchers will not have the necessary power to distinguish between alternative functional

forms in the pre-period. Regardless, in the absence of a compelling case for a particular functional form for the counterfactual, economists should consider how robust their results are to alternative choices.

4 Testing for Preexisting Trends

As DiD has grown in popularity, researchers have become increasingly aware that seemingly similar groups may not always exhibit truly parallel trends. This may arise either because the control group really is not a very good control group or, as discussed in the previous section, because the functional form for the counterfactual is incorrect. Consequently, whenever data permit, authors are now effectively required to test the assumption of parallel trends using a test of ‘pre-trends.’ Angrist and Pischke 2010 maintain that “The most compelling differences-in-differences-type studies report outcomes for treatment and control observations for a period long enough to show the underlying trends, with attention focused on how deviations from trend relate to changes in policy.”

The logic is that if the two groups truly would have exhibited parallel trends in the absence of treatment, we should find that this model fits the data in the periods prior to treatment. There are two ways that researchers generally test for parallel trends in periods prior to treatment, or ‘pre-trends.’ The first, frequently used when only a modest number of time periods are available, is to replicate the model on two periods prior to the treatment. In this case, the second period becomes the placebo treatment period. The second approach, used by KL, considers the year prior to treatment to be the “base year” and estimates the difference between the control and treatment groups (or, more precisely in the case of KL, an IV coefficient) in each previous year relative to the base year. This allows the researcher to test the null hypothesis that outcomes prior to the treatment year exhibited parallel trends. KL and JJK differ in their choice of base year and in how they group years, but these differences appear to us to be minor.

Testing for parallel pre-trends arises naturally in the potential treatments approach to DiD, which assumes that we can write potential outcomes for a group, g , in time t as $Y_{gt}(D) = f(D, u_{gt})$ where u_{gt} is a group and time error term which incorporates group specific shocks. This framework therefore requires that $f(0, u_{g1}) - f(0, u_{g0})$ is mean independent of D_g . Given that we do not observe $f(0, u_{g1})$, it is natural to test for mean independence by asking whether this holds in $f(0, u_{g0}) - f(0, u_{g-1})$, or more generally, whether $f(0, u_{g1}) - f(0, u_{g\tau})$, $\tau \leq 0$ is correlated with D_g .

KL and JJK both follow this approach and test whether the coefficients on the treatment group in the pre-period (relative to time 0) are individually or jointly statistically significantly different from zero. The logic of their approach is that if, for example, the coefficients on periods prior to the base year are positive relative to the base year, then the difference was already declining in the pre-period. But what are they and others who use this approach really testing? They are testing the null hypothesis that the relative differences are all zero against the alternative that the coefficients are not equal to zero. Given the unknown nature of the pre-trends in this setting, it would certainly be troubling if the pre-trend coefficients failed this test. However, passing this test does not mean that the researcher should feel content that significant pre-trends have been ruled out. In the case of KL, all of the coefficients are positive and decreasing, suggesting a potential linear trend in the pre-period. In this case, a one-sided chi-squared test evaluating whether the coefficients are all positive or directly testing for a linear pre-trend is appropriate. In other words, if the pre-trends fail some tests of significance but not all tests, this still suggests there is serious reason for concern on the part of the researcher.

Increasingly, researchers point to a statistically insignificant pre-trend test to argue that they therefore accept the null hypothesis of parallel trends. There is no doubt that testing for a common pre-trend plays an important role in validating the parallel trends assumption underlying DiD. However, failing to reject that outcomes in years *prior* to treatment exhibit parallel trends, should not be confused with establishing the validity of the parallel trends

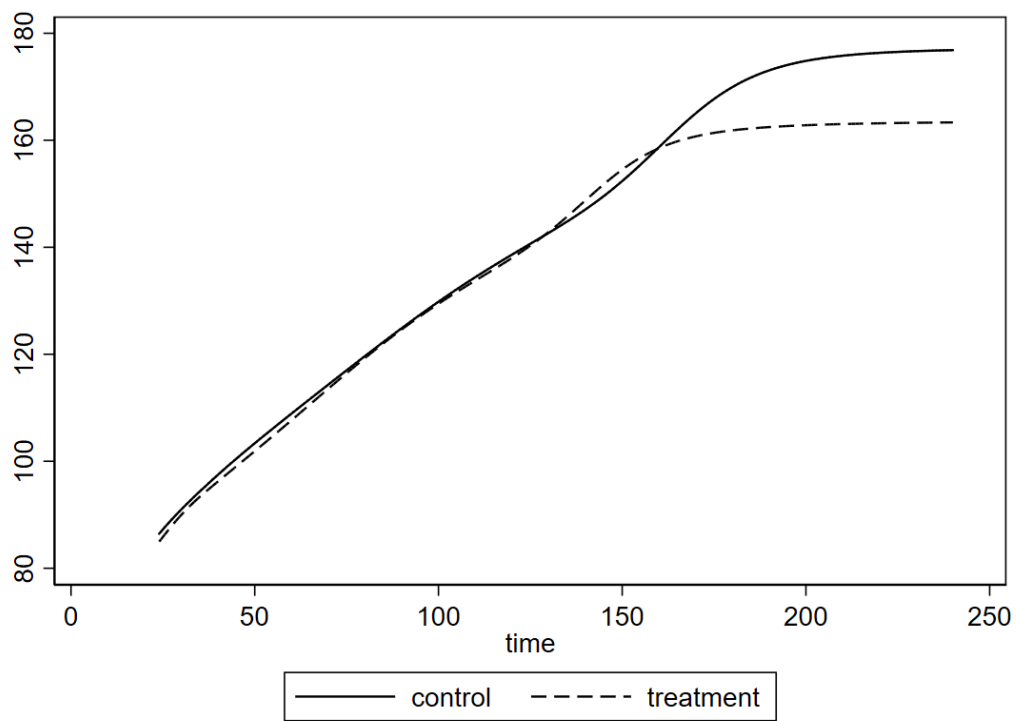
counterfactual. Moreover, clearly, not rejecting the null hypothesis is not equivalent to confirming it. Yet, in such settings, economists may be tempted to, in Brad DeLong's terms, "seize the high ground of the null hypothesis," which can lead to incorrect estimation of treatment effects. Second, what we would like to establish is not that there were parallel trends in the pre-treatment period, but rather that there would have been in the post-period in the absence of the treatment, something which is unfortunately unknowable. A parallel trend in the periods prior to treatment does not guarantee that a parallel trend would have continued in the absence of treatment. The researcher should consider whether there is reason to believe this pattern would have continued. This requires understanding why the groups diverged in levels but otherwise followed similar trends, in other words, a good understanding of the subject matter.

Take, for example, the following study relying on some or all of the data in Figure 1. We will tell you shortly what is on the y-axis, but in the interest of honest reflection and introspection, we discourage you from reading ahead and to spend a brief period looking at the figure.

It is tempting to begin hypothesizing about what caused the sudden break from the common trend at roughly $T=160$. We trust that the reader would not come up with an explanation and then test it using these same data. KL certainly deserve credit for first noticing a pattern like that in figure 1 and then asking how to test it without relying directly on the data that led to their hypothesis.

Assume a researcher reads about an intervention that affected some of the treatment group but none of the control group at $T=160$. We assume our researcher is honest and did not know the pattern in the data before reading this. If she looks at the long-term pre-trend, she will almost definitely conclude that the control and experimental groups have followed a similar pre-trend (with some small unexplained departure around $T=140$). If she looks only at the shorter period, she will probably conclude that the difference in pre-trend is a problem with which she must grapple. It is not obvious which is correct. Here we agree with

Figure 1: Results of a 'Natural Experiment'



the Kearney and Levine 2016 response to JJK. A long-run trend is not necessarily better than a short-run trend, but the converse is also not true. Consequently, the presence of a long-run pre-trend does not prove that the analysis based on the absence of a short-run pre-trend is incorrect. But, in the absence of a compelling analysis of why we should prefer one piece of evidence rather than the other, it does detract from our confidence about either conclusion. And, as will be clear, the similar long-run trend for much of the sample period in our example is misleading.

The two groups in figure 1 are males (the control group) and females (the experimental group). We have plotted average height in centimeters against age measured in months. For reasons that we well understand (or at least we would if we had studied enough biology), height first diverges between the sexes when females hit puberty and then diverges in the opposite direction when males hit puberty at a later age. Despite exhibiting parallel trends in height prior to $T=135$, male and female height should not be expected to continue at the same rate of growth.

We are confident that no economics journal would publish our fictional study, but we are less confident that they do not publish well-intended research on complicated psychological or social phenomena based on a belief in the superiority of economists' statistical methods for uncovering causal effects. As the disagreement between KL and JJK suggests, the evidence that in the absence of the intervention there would have been a common trend in the post-intervention period should not and cannot be purely statistical.

For example, JJK point out that *16 and Pregnant* premiered during the onset of the Great Recession. They argue that the recession impacted populations and areas of the country differently, suggesting that even if trends had been parallel prior to treatment, they might have diverged at the time of treatment even in the absence of treatment. This argument raises a valid concern regardless of what one concludes about the pre-trend. Here again, understanding why teen birth rates differed prior to treatment helps us assess whether the coincidence of the timing of the show and of the Great Recession is a significant cause for

concern.

While the absence of a pre-trend is neither a necessary nor a sufficient condition for the absence of divergence in the counterfactual, it would be foolish to suggest that establishing that there is no pre-trend lends no support to the null hypothesis. But when should we conclude that there is no pre-trend? Even if we accept that we should rely on the more recent period for testing for a pre-trend, the relation between teen pregnancy rates and predicted viewership appears to be decreasing prior to the introduction of *16 and Pregnant* despite its statistical insignificance, which as we have previously noted might have proved to be more statistically significant if the authors' had used a more powerful test. Although we do not offer a complete solution to this problem, we propose the following two considerations.

The first is to test statistically the null hypothesis of non-parallel trends sufficient to eliminate an experimental effect. This requires considering a functional form of pre-period trends, for example linear, and testing directly whether the confidence interval includes a sufficiently large difference in trends to eliminate the purported experimental effect. If we cannot reject a trend that, if continued into the post-period, would eliminate the experimental effect, we do not have strong support for that effect.

We can also think of this in the context of the standard regression equation. Here, it seems more natural to ask whether it changes our estimate of the 'causal' effect if we include group-specific trends or allow trends to depend on group characteristics. And, indeed some papers have taken this approach, including KL in their 2016 response to JJK. If doing so changes the interpretation of the coefficient of interest, we must be appropriately circumspect about our conclusions. Of course, all of the usual caveats about adding additional controls also apply in this setting. In particular, if we add sufficiently nonlinear group-specific trends, we will inevitably render the effect of the intervention statistically insignificant. Even if we restrict ourselves to linear trends, the loss of degrees of freedom may dramatically reduce our power to detect a true effect of the intervention.

Another concern with including group-specific trends is that the treatment effect may not

be fixed over time. In this case, the original regression equation is misspecified, and adding further controls can increase rather than decrease the bias. Thus, if the causal effect of the intervention increases (or decreases) with time since implementation, the DiD approach will give the average effect of the intervention over the post-period. Borusyak and Jaravel 2017 discuss in depth the challenges of using DiD in cases with group-specific time trends and time-specific treatment effects. We illustrate through a simple example where adding a variable for treatment group*time would result in the coefficient on treatment having the wrong sign.

Consider the following example of a treatment with a delayed effect. There are five pre-periods and five post-periods. Outcomes for the treatment and control group are shown in Figure 2. Observe that the dependent variable for both the experimental and treatment group increases linearly (from one to five) over the pre-period. Thereafter, it remains at five for the control group. For the experimental group it takes the values five in periods six and seven, reflecting a delayed effect. In period eight, when the effect starts to kick in, the treatment group experiences an outcome value of six, and then seven in periods nine and ten when the full effect is felt.

Compare the results we get here with and without including group-specific time trends. Let d be a dummy for membership in the treatment group. If we regress the dependent variable without the group-specific time trend and dropping subscripts for simplicity, we get

$$y = 1.05 - 0.0d - 1.25post + 1.0d * post + 0.65time. \tag{12}$$

Without the group-specific time trend, the coefficient on the interaction term, 1, is the average effect over the post-policy period. Now we add a group-specific trend, which in this case consists simply of adding an interaction between experimental group and time. The resulting regression results are

$$y = 1.5 - 0.9d - 0.5post - 0.5d * post + 0.5t + 0.3d * t. \tag{13}$$

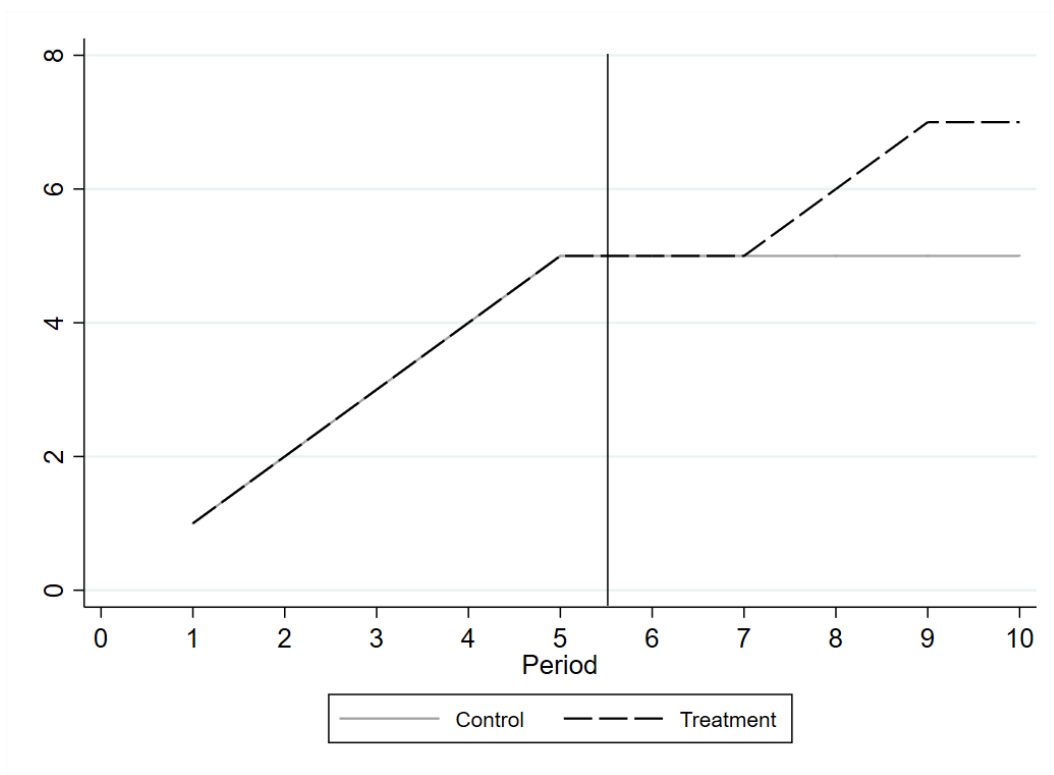
We now find a negative coefficient of $-.5$ for the treatment impact. The investigator is likely to conclude incorrectly that the experiment had no effect when in reality it quite clearly had a positive effect.

Why should we have anticipated the reversal of sign? Consider what happens if we detrend the dependent variable separately for the two groups. Because the experimental group eventually shows an increase while the control group stops increasing following the experiment, the ‘trend’ is steeper for the experimental group. Consequently, right after the onset of the experiment, the detrended variable shows a sharper drop for the experimental group than for the control group. By the end of the experiment, the detrended variable is less negative for the experimental group than for the control group, but the difference for the earlier period outweighs the later period in the example.

Our second recommendation is to take seriously our earlier discussion of the experimental design. For example, this will include an analysis of what factors might explain the differences in levels in the period prior to treatment. If we understand why the experimental and control groups differ in levels, we may better understand whether to anticipate common or divergent trends. For example, JJK show that the share of the population that identifies as Hispanic is substantially lower in areas with high MTV viewership. This suggests that there are meaningful differences between high and low viewership areas which may generate divergent trends. One option here is to include race or ethnicity specific time trends or, similarly, interactions between race/ethnicity and period. If this substantively changes the interpretation of treatment effects, this should raise serious concerns about our estimates of the treatment effect. On the other hand, if, subject to all our other caveats, accounting for the racial and ethnic composition of the DMAs eliminates both the initial difference and differential trends, we will feel considerably more confident that we have solved the problem with nature’s design of the experiment.

Experimental design becomes even more important in the case when pre-trend tests fail. Although failing a pre-trend test is a cause for serious concern, it is not always an

Figure 2: Time Trends Example



immediate disqualification for a DiD study. If a researcher is testing the possibility that the treatment of interest may have caused two outcomes that would otherwise have moved in parallel to diverge, it is hard to imagine that there cannot be other possible sources of divergence. However, only after understanding the source of all deviations should the researcher feel confident continuing. Understanding the impact and nature of the shock allows the researcher to critically evaluate whether the assumption of counterfactual parallel trends is plausible, despite the observed deviation.

Finally, we note that testing for a pre-trend is a special case of pre-testing in econometrics, and, as is well known, the default standard errors are incorrect when we rely on pre-testing. Roth 2018 discusses using pre-trend testing to test the parallel trends assumption as a special case of pre-testing. He explains that using pre-trend tests as a test of parallel trends not only risks accepting misidentified studies, but can also exacerbate the bias from violations of parallel trends and lead to severe over-rejection of the null hypothesis. He proposes a corrected estimator to adjust for the fact that the pre-trend test has occurred. This is certainly an improvement, and we believe that the Roth estimator should be used in DiD designs that rely on pre-trend testing. However, using this estimator does not eliminate the need for logical reasoning related to parallel trends. In the case when parallel trends appear plausible but not certain, the researcher should also perform a thorough comparison of the differences between the treatment and control groups including demographic composition, other factors that could have differentially affected each group, and comparison of trends as far back as possible. JJK provide a good example of how a thorough comparison can help the researcher logically determine whether the parallel trends assumption is credible.

5 Conclusion

In a perfect world, we would have well designed experiments which would provide clear evidence on the causal impact of all potential policies and similar interventions. In reality,

we are far from that world. In cases without a true randomized experiment, tools like differences-in-differences broaden the range of ‘natural experiments’ we can use to identify causal effects, but we should not allow the use of this term to fool us. The use of the term ‘experiment’ seems to imply a level of credibility that will rarely if ever be commensurate with what can be expected of empirical research based on DiD. Using DiD properly requires taking all necessary precautions - both logically and in terms of methods - to ensure that the assumptions of DiD are met. We should not conclude from this that well-executed papers relying on DiD are somehow invalid. Instead, we hope that our discussion highlights some of the broader issues regarding the assumption of parallel trends required for identification in a DiD design and what should be considered necessary for justifying this assumption. Although we do not provide a full solution to this question, we hope to have provided a framework for thinking about the assumption of parallel trends.

In this paper, we highlight the discussion between KL and JJK, but our message should not be viewed primarily as applying to estimating the impact of *16 and Pregnant*, but rather as a discussion about the implementation of differences-in-differences designs more broadly. KL employ statistical methods that are used frequently in empirical work. Although JJK chose to focus their analysis on the KL paper, it is likely that there are a range of other well-known papers for which a similar analysis could be performed.

We encourage researchers to use this as a framework for thinking about identification in all empirical work in which there is a potential failure in experimental design, not just DiD. This means asking, “Why did nature’s ‘experimental design’ fail, and how could that impact my identification strategy?” Any available tests of identification should be seen as a complement to, not a substitute for, logical reasoning. This also highlights the need for additional research on how to evaluate identification strategies in cases where there has been a failure of experimental design.

References

- Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, *80*(1), 123–129. doi:10.1016/S0165-1765(03)00032-6
- Angrist, J., & Imbens, G. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, *62*(2), 467–475. doi:10.2307/2951620
- Angrist, J., & Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30. doi:10.1257/jep.24.2.3
- Athey, S., & Imbens, G. (2006). Identification and Inference in Nonlinear Difference-in-Differences Models. *Econometrica*, *74*(2), 43–497.
- Blundell, R., & Costa Dias, M. (2009). Alternative Approaches to Evaluation in Empirical Microeconomics. *The Journal of Human Resources*, *44*(3), 565–640.
- Borusyak, K., & Jaravel, X. (2017). Revisiting Event Study Designs. *Harvard University Working Paper*.
- Cameron, A. C., & Miller, D. (2015). A Practitioner’s Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372.
- Jaeger, D., Joyce, T., & Kaestner, R. (2018). Did Reality TV Really Cause a Decline in Teenage Childbearing? A Cautionary Tale of Evaluating Identifying Assumptions. *Journal of Business and Economic Statistics*, *forthcoming*.
- Kearney, M., & Levine, P. (2015). Media Influences on Social Outcomes : The Impact of MTV’s *16 and Pregnant* on Teen Childbearing. *105*(12), 3597–3632.
- Kearney, M., & Levine, P. (2016). Does Reality TV Induce Real Effects? A Response to Jaeger, Joyce, and Kaestner. *IZA DP No. 10318*.
- Lang, K. (1993). Ability bias, discount rate bias and the return to education. *Munich Personal RePEc Archive*, *paper 24651*.
- Meyer, B. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics*, *13*(2), 151–161.

Roth, J. (2018). Should We Adjust for the Test for Pre-trends in Difference-in-Difference Designs? *arXiv, 1804.01208*.

Snow, J. (1855). On the mode of communication of cholera. *J. Churchill, London, England*
2nd Ed.