Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends

Jonathan Roth*

March 20, 2019

Abstract

When using difference-in-differences and related event-study designs, researchers often test for pre-event trends ("pre-trends"), yet estimation and inference typically does not account for this test. This paper analyzes the properties of conventional event-study estimates conditional on having survived a pre-test for parallel trends. I derive a formula for the bias in conventional treatment effects estimates after pre-testing, which generally differs from the unconditional bias when parallel trends is violated. Moreover, I prove that under homoskedasticity, the additional bias from pre-testing amplifies the bias in the treatment effects estimates from a monotone violation of parallel trends. Hence, pre-trends tests meant to mitigate bias in published work can actually exacerbate it. In addition, coverage rates of traditional confidence intervals can be above or below their nominal level conditional on surviving the pre-test. Simulations based on a review of recent papers in leading economics journals suggest that substantial distortions from pre-testing are possible in practice. To address these issues, I develop a method of constructing corrected event-study plots that removes distortions from pre-testing or from model-selection on the basis of pre-trends. I illustrate the usefulness of these corrections in simulations and in an application to Dube et al. (2010)'s study of the minimum wage.

*Department of Economics, Harvard University, jonathanroth@g.harvard.edu. I am grateful to Isaiah Andrews, Kirill Borusyak, Peter Ganong, Edward Glaeser, Nathan Hendren, Ariella Kahn-Lang, Max Kasy, Larry Katz, Sendhil Mullainathan, Jesse Shapiro, Jann Spiess, Jim Stock, and Elie Tamer for helpful comments. This material is based upon work supported by the NSF Graduate Research Fellowship under Grant DGE1144152.

1 Introduction

Difference-in-differences and related event-study designs are highly popular tools for quasi-experimental empirical research. The key identifying assumption in these research designs is that of parallel trends, meaning that had the treatment of interest not occurred the mean of the outcome would have evolved in parallel for the treatment and control groups. Although one cannot directly test what would have happened under the counterfactual, researchers often test for differences in trends prior to the date of treatment assignment – typically referred to as "pre-trends" – as a way of weeding out instances where the assumption of parallel counterfactual trends is implausible. Such tests are remarkably common: since 2014, over 70 papers in the journals of the American Economic Association have employed a so-called "event-study plot" in order to visually test for pre-trends.

A handful of recent papers have warned, however, that the common approach of testing for pre-trends will be imperfect in finite samples (Freyaldenhoven et al., 2018; Kahn-Lang and Lang, 2018; Bilinski and Hatfield, 2018). Owing to noise in the data, we may spuriously detect a violation of parallel trends when none exists, and perhaps more concerningly, we may fail to detect a violation of parallel trends when there is one. Given the imperfect performance of pre-trends tests in finite samples, it is important to understand the statistical properties of the estimates that survive a test for pre-trends. While prior work has analyzed the distribution of estimates after pre-testing in a variety of other contexts (see Related Literature below), the properties of the distribution of event-study estimates after testing for parallel trends remain poorly understood.

This paper fills that gap in the literature by analyzing the properties of event-study estimates and confidence intervals conditional on having survived a pre-test for pre-trends. I show that many of the properties we expect of conventional estimates no longer hold conditional on passing the pre-test. Perhaps most concerningly, I illustrate that pre-tests meant to mitigate bias in the treatment effects estimates can actually amplify bias in published work. I then propose a method for producing a corrected event-study plot that eliminates the distortions from pre-testing.

In Section 2, I begin with a stylized example that highlights how pre-trends testing can amplify bias in point estimates and distort the coverage rate of confidence intervals. I consider a difference-in-differences setting with three periods in which there are potentially linear violations of parallel trends. A key insight of the example is that the estimate of the pre-existing trend and the estimate of the treatment effect both depend on the estimation error of the treatment-control difference in the period before treatment. As it turns out, noise that leads us to fail to detect a violation of parallel trends also tends to produce particularly biased

estimates of the treatment effect. Thus, when the parallel trends assumption is violated in population, the bias conditional on surviving the pre-test is worse than the unconditional bias. Selection on noise in the pre-period can similarly lead confidence intervals (CIs) to have coverage rates that differ from their nominal level conditional on passing the pre-test.

In Section 3, I provide theoretical results that extend the intuition from the stylized model to a more general setting that allows for an arbitrary number of pre-periods and non-linear violations of parallel trends. I first derive general formulas for the bias and variance of event-study estimates after passing a pre-test for parallel trends. I then prove that under homoskedasticity, pre-testing necessarily amplifies bias in the treatment effects estimates whenever there is a monotone violation of parallel trends. Lastly, I show that pre-testing reduces the variance of event-study estimates under very general conditions. The distortions to both bias and variance have opposite effects on the coverage rates of conventional CIs; as a result, conventional CIs will tend to overcover when the underlying trend is close to zero (when bias is small), and will tend to undercover when the underlying trend is sufficiently large.

Section 4 evaluates the practical relevance of these distortions in simulations based on a systematic review of recent papers in three leading economics journals (the *American Economic Review*, *AEJ: Applied Economics*, and *AEJ: Economic Policy*). Although other recent papers have cautioned that pre-trends tests may have low power (Freyaldenhoven et al., 2018; Kahn-Lang and Lang, 2018; Bilinski and Hatfield, 2018), I provide the first systematic evaluation of the power of pre-trends tests in published papers. I find that, indeed, conventional pre-trends tests often have low power against meaningful violations of parallel trends. In many cases, linear trends against which conventional tests have power of only 50 percent would produce bias of a magnitude similar to the estimated treatment effect. Additionally, the bias conditional on failing to detect an underlying trend is worse than the unconditional bias in the large majority of cases, in line with the theoretical prediction for the homoskedastic case. This bias amplification can be substantial in magnitude: in some cases, the bias conditional on passing the pre-test is more than twice as large as the unconditional bias.

How can we evaluate event-study plots given the distortions from pre-testing? In Section 5, I develop a method for constructing a "corrected event-study plot" that recovers many of the properties we expect of conventional event-study estimates conditional on passing a pre-test for parallel trends. In particular, my method, which builds on work by Andrews and Kasy (2017) and Lee et al. (2016), provides median-unbiased point estimates and valid confidence intervals for the population event-study coefficients, regardless of the true underlying trend. I further show that my method can be extended to cover situations in which a researcher

selects among multiple possible research designs – e.g. using different subgroups or different sets of control variables – on the basis of the observed pre-trends. The corrected event-study plot thus gives the reader an unbiased way of evaluating the chosen event-study in light of the process by which the research design has been screened and/or selected.

In practice, I recommend that researchers who rely on pre-trends tests report corrected event-study plots along with calculations of the power of their pre-test to detect meaningful violations of parallel trends.[1] The power calculations are important because, while the corrected event-studies eliminate the bias and size distortions from pre-testing, they do not solve the issue that the pre-test may be under-powered to detect an underlying trend. In other words, while the corrected point estimates for the pre-period coefficients will be median-unbiased for the population event-study coefficients, the confidence intervals on the pre-period coefficients may be sufficiently wide that it is difficult to distinguish between no underlying trend and an underlying trend that would meaningfully bias the treatment effects estimates. The power calculations thus give the reader a sense of the probability that a violation of parallel trends would be detected via the pre-trends test, whereas the corrected event study estimates allow us to evaluate the results in light of the screening process that has occurred.

Finally, I illustrate the economic relevance of these corrections in an application to Dube et al. (2010)'s influential study of the employment effects of the minimum wage. Dube et al. (2010) present the results from two primary specifications, one of which uses all counties in the US and one which restricts only to counties on state borders. They find a sizeable and significant pre-trend in the former specification but not in the latter, and thus prefer the latter specification. I first show that the border counties specification has low power against a trend of the same magnitude as that in the specification using all counties. Next, I apply my method for creating corrected event-studies to account for the model selection on the basis of the pre-trends. The corrected point estimates indicate that following a minimum wage increase there is a moderate decrease in employment relative to 4 quarters prior to the change. Moreover, while the original Dube et al. (2010) confidence intervals can rule out the negative employment effect found by Neumark and Wascher (2000), the Neumark and Wascher estimate falls well within the confidence intervals after adjusting for the pre-test.

Related Literature. This paper contributes to a large body of work on the econometrics of difference-in-differences and related research designs (e.g. Bertrand et al. (2004); Abadie (2005); Donald and Lang (2007); Borusyak and Jaravel (2016); Abraham and Sun (2018);

[1] An interesting question for future research is the extent to which the current practice of pre-trends testing can be improved upon, either by changing the pre-testing criteria or by adopting a more continuous approach to accounting for potential confounding trends.

Athey and Imbens (2018); de Chaisemartin and D'Haultfœuille (2018a,b); Goodman-Bacon (2018); Callaway and Sant'Anna (2019)). Most closely related, recent papers by Freyaldenhoven et al. (2018), Kahn-Lang and Lang (2018), and Bilinski and Hatfield (2018) have warned that traditional pre-tests may have low power to detect meaningful violations of parallel trends. I contribute to this literature in three ways. First, I characterize the distribution of treatment effects estimates conditional on having survived a pre-test for parallel trends, and I show that pre-testing can exacerbate the bias from an underlying trend.[2] Second, I provide the first systematic evaluation of the power of pre-tests and the distortions from pre-testing in published work. Third, I develop a method for producing a corrected event-study that eliminates the additional bias and coverage distortions from pre-testing.

More broadly, this paper relates to a large literature in econometrics and statistics showing that, in a variety of contexts, problems can arise in both estimation and inference if researchers do not account for a pre-testing or model selection step (see, e.g., Giles and Giles (1993), Leeb and Pötscher (2005), Lee et al. (2016), and references therein). Recent work has examined, for instance, the implications of conditioning analysis on tests for weak identification (Andrews, 2018) or data-driven tuning parameters (Armstrong and Kolesár, 2018).

Finally, this paper relates to the literature on selective publication of scientific results (Rothstein et al. (2005) and Christensen and Miguel (2016) provide reviews). My procedure for producing corrected event-study plots builds on results developed by Andrews and Kasy (2017) to correct for publication bias, as well as earlier results from Lee et al. (2016) and Pfanzagl (1994). A particularly relevant paper on selective publication is Snyder and Zhuo (2018), who provide empirical evidence that papers with significant placebo coefficients – which they refer to as "sniff tests" – are less likely to be published. I study tests for pre-trends, a common form of sniff test, and provide estimation and inference procedures that correct for the distortions created by selecting research designs based on the result of this sniff test.

2 Intuition: The effect of pre-trends testing in a stylized three-period model

This section develops intuition for how testing for pre-trends affects the distribution of event-study estimates in a simple model with three-periods, homoskedastic errors, and (potentially)

[2]Relatedly, Daw and Hatfield (2018) and Chabé-Ferret (2015) illustrate that selecting a control group on the basis of pre-period outcomes can induce bias in difference-in-differences.

linear violations of parallel trends. A key insight is that estimation error in the treatment-control difference in the reference period ($t = 0$) enters both the pre-period and post-period event-study coefficients, so pre-testing using the pre-period coefficients affects the distribution of the post-period coefficients via its effect on the distribution of this error.

After setting up the model, we first show that pre-testing exacerbates bias when there is a linear violation of parallel trends, but does not induce any bias when parallel trends holds. We next illustrate that the coverage of traditional confidence intervals may be too high or too low, depending on the underlying trend. Finally, we show that a rule requiring researchers to pre-test for pre-trends can either increase or decrease bias in published work, with the effect depending on i) the power of the pre-test to detect a violation of parallel trends, and ii) the ex ante plausibility of the research design.

2.1 Set-up of stylized model

Suppose that we observe an outcome y_{it} for individuals i in period t for three periods $t = -1, 0, 1$. Individuals in the treatment group ($Treatment_i = 1$) receive a treatment of interest between periods 0 and 1, whereas individuals in the control group do not receive the treatment. We denote by $y_{it}(1)$ and $y_{it}(0)$ the potential outcomes for individual i in period t that would have occurred if they respectively did or did not receive treatment. For simplicity, we consider the case where there is no causal effect of treatment, i.e. $y_{it}(1) \equiv y_{it}(0)$, and the true data-generating process for $y_{it}(0)$ is given by:

$$y_{it}(0) = \alpha_i + \phi_t + Treatment_i \times g(t) + \epsilon_{it} \tag{1}$$

for $\epsilon_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. The term $Treatment_i \times g(t)$ represents a potential difference in trends between the treatment and control group. For instance, if $g(t) = t$, then the average outcome for the treatment group is increasing linearly relative to the control group, whereas if $g(t) = 0$ then the parallel trends assumption holds.

We suppose that the researcher estimates the canonical "event-study" difference-in-differences regression specification:

$$y_{it} = \alpha_i + \phi_t + \sum_{s \neq 0} \beta_s \times 1[s = t] \times Treatment_i + \epsilon_{it}. \tag{2}$$

The estimate $\hat{\beta}_1$ is the canonical difference-in-differences treatment effect estimate,

$$\hat{\beta}_1 = \Delta\bar{y}_{t=1} - \Delta\bar{y}_{t=0},$$

where $\Delta\bar{y}_t$ is the difference in sample means between the treatment and control group in period t. Likewise, the estimate $\hat{\beta}_{-1}$ is the canonical pre-period event-study coefficient,

$$\hat{\beta}_{-1} = \Delta\bar{y}_{t=-1} - \Delta\bar{y}_{t=0}.$$

An important observation is that the term $\Delta\bar{y}_{t=0}$, the estimated difference in means between treatment and control in the reference period ($t = 0$), enters the expression for both the pre-period and post-period coefficients. As a result, if we select on the observed pre-period coefficient $\hat{\beta}_{-1}$ being close to zero, this will affect the distribution of $\Delta\bar{y}_{t=0}$, which in turn will impact the distribution of $\hat{\beta}_1$. The next two sections illustrate how this selection plays out, first in a setting where parallel trends is violated and next in the case where it holds.

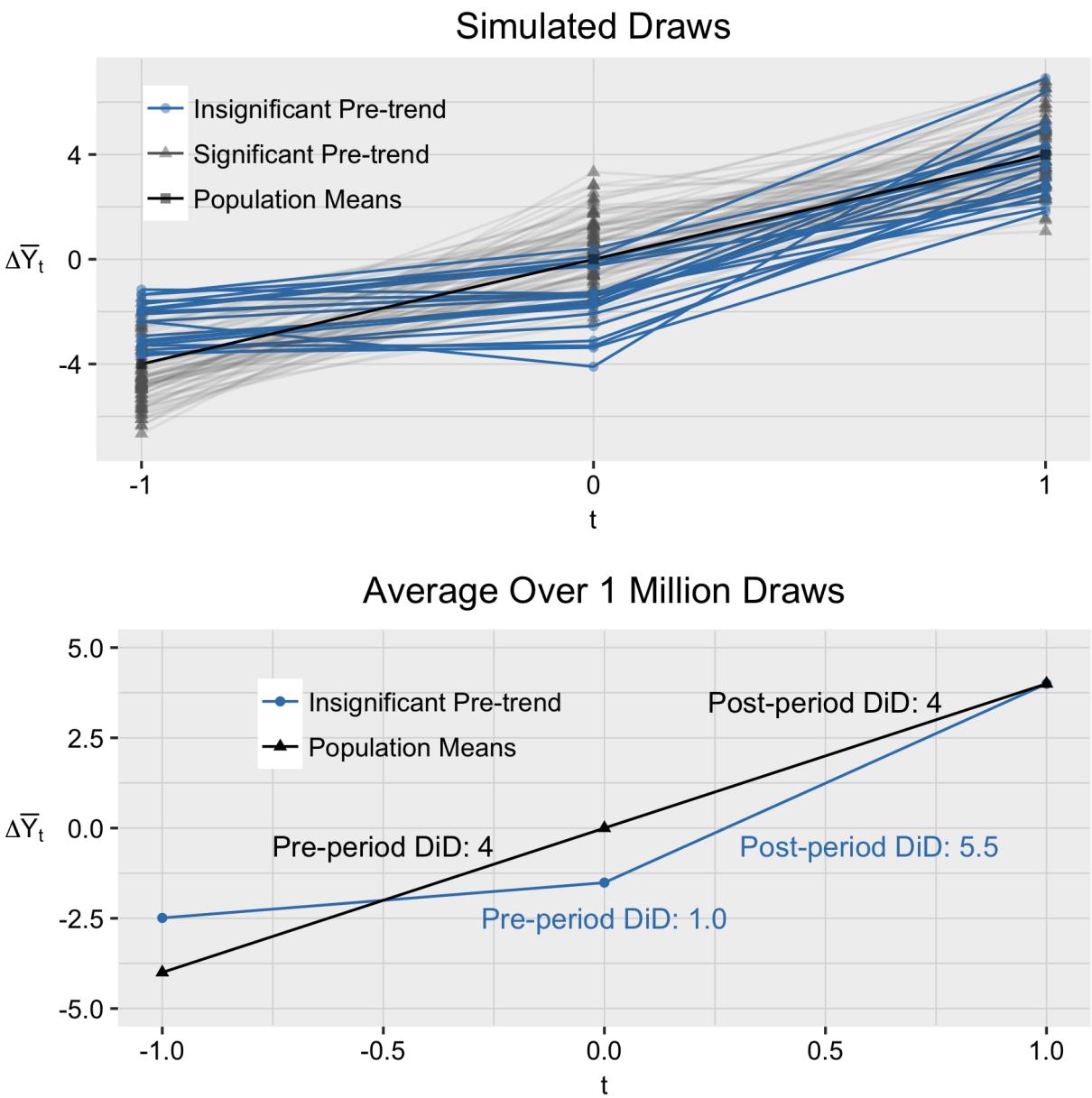
2.2 When there is an underlying trend, pre-trends testing exacerbates bias

Figure 1 provides intuition for how pre-trends testing affects the distribution of estimated event-study estimates when parallel trends is violated. The top panel of the figure shows simulations from a data-generating process where in population there is an underlying upward linear trend, i.e. $g(t) = b \cdot t$ for $b > 0$. The y-axis shows the difference in means between treatment and control in each period, which in population is a straight line.

Although in population there is an underlying pre-trend, in finite samples there will sometimes be noise that prevents us from detecting this trend. The top panel highlights such draws in blue – in particular, the highlighted draws are those for which the t-statistic on $\hat{\beta}_{-1}$ is less than 1 in absolute value, so that at conventional levels we would not detect a significant pre-trend. I use a threshold t-stat of 1 (rather than 1.96) for this stylized example because it allows for an easier visual comparison of the cases with and without significant pre-trends, although results using the standard threshold of 1.96 are qualitatively similar.[3] By definition, the slope of the line between $t = -1$ and $t = 0$ corresponds with $-\hat{\beta}_{-1}$, whereas the slope between $t = 0$ and $t = 1$ corresponds with $\hat{\beta}_1$. Thus, the highlighted draws are

[3] In Section 4, I conduct simulations based on a review of recent published papers using the more traditional threshold of 1.96.

Figure 1: Intuition for how bias is worse conditional on not detecting a significant pre-trend



Note: The top panel of the figure shows simulated draws from a DGP in which in population the outcome of interest for the treatment group is increasing linearly relative to the control group. The y-axis shows the difference in sample means between the treatment and control group in each period ($\Delta\bar{y}_t$). I highlight in blue the draws of the data in which the t-stat on the pre-period coefficient $\hat{\beta}_{-1}$ is less than one in absolute value. The bottom panel shows the average of the blue lines over 1 million draws. The figure illustrates that the cases in which we fail to detect an upward underlying trend exhibit a small observable pre-trend, but produce particularly large treatment effects estimates.

those for which the slope between $t = -1$ and $t = 0$ is close to 0.

Figure 1 makes apparent that the draws where we fail to detect a significant pre-trend also have a particularly large slope between period $t = 0$ and $t = 1$, corresponding with a large value of $\hat{\beta}_1$. The reason for this is that the draws of the data in which we fail to detect the underlying tend to have below-average values of $\Delta\bar{y}_{t=0}$. Intuitively, this is because negative shocks to $\Delta\bar{y}_{t=0}$ help to mask the underlying trend and "flatten" out the observed slope in the pre-period. However, when we underestimate the difference in means between treatment and control in period 0, we tend to overestimate the growth in this difference between period $t = 0$ and $t = 1$ owing to mean reversion, and so the cases where we fail to detect the underlying trend tend to produce particularly large treatment estimates for period 1. Thus, the expected bias conditional on passing the pre-test is worse than the unconditional bias in OLS.

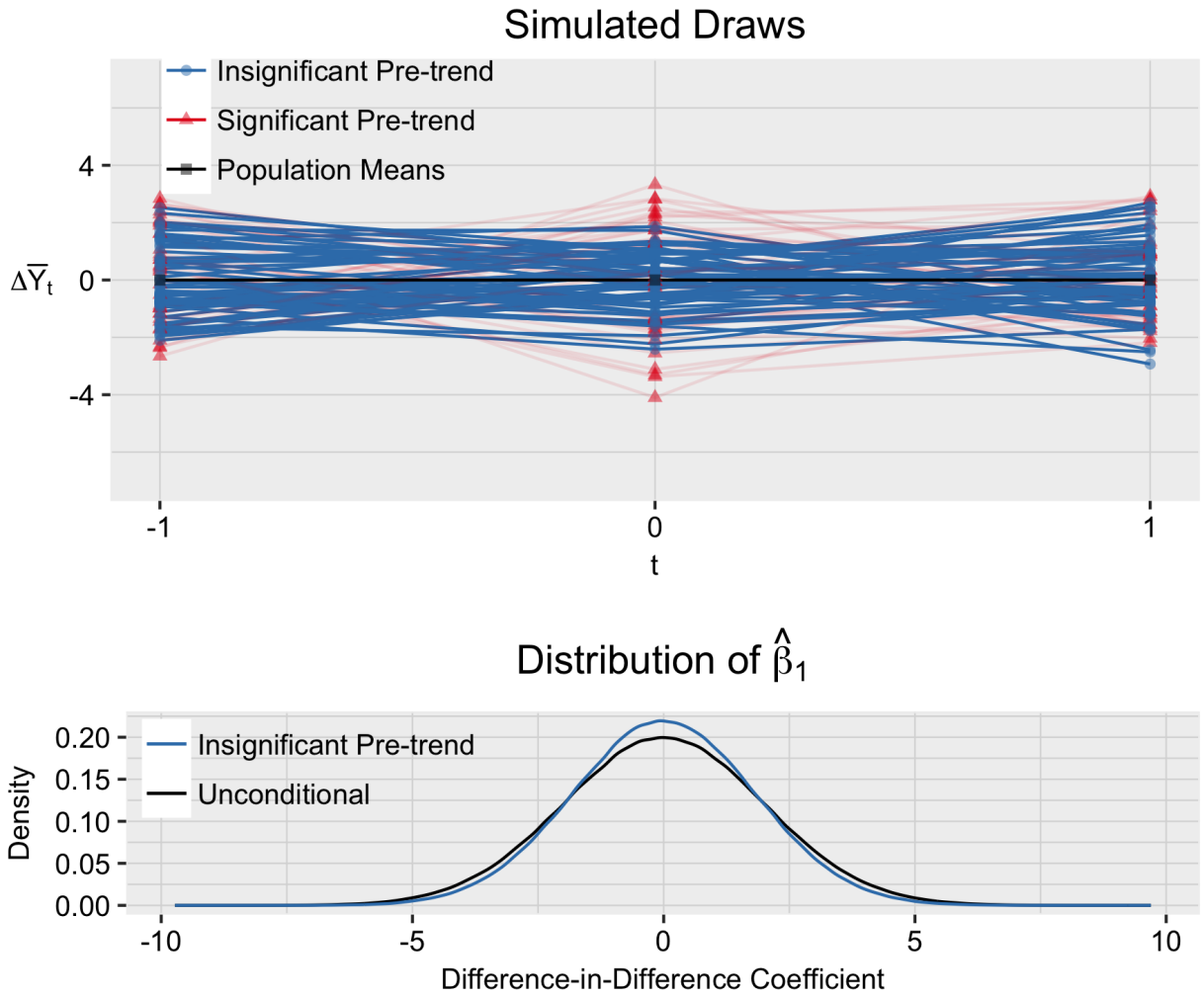
2.3 When parallel trends holds, there is still no bias after pre-trends testing

What happens if parallel trends holds in population? The top panel of Figure 2 shows draws from our DGP when parallel trends holds ($g(t) \equiv 0$). We again highlight in blue the draws in which the t-stat on $\hat{\beta}_{-1}$ is less than 1 in absolute value, and we now highlight in red the draws in which the t-stat is greater than 1 in absolute value. We see from the figure that the draws where we detect a significant pre-trend (in red) disproportionately have extreme values for $\Delta\bar{y}_{t=0}$. This is because when in population there is no underlying trend, we tend to incorrectly detect one when the noise in the reference period is large. As a result, by conditioning on having not found a significant pre-trend, we are excluding cases that on average have a high degree of noise in period 0, and consequently have noisy treatment effects estimates. This leads to a reduction in the variance of the treatment effects estimates, but since we are equally likely to throw out cases where the noise in the reference period is positive or negative, we do not induce any bias in the treatment effects estimates. This can be seen in the bottom panel of Figure 2, which shows that the distribution of estimated treatment effects conditional on finding an insignificant pre-trend is centered around zero but has lower variance than the unconditional distribution.

2.4 After pre-testing, coverage rates can be too high or too low

What do these two examples imply for the performance of traditional confidence intervals conditional on finding an insignificant pre-trend? Intuitively, when parallel trends (approximately) holds, the conditional treatment effects estimates are (approximately) unbiased but

Figure 2: Intuition for why pre-testing reduces variance but preserves unbiasedness when parallel trends holds



Note: The top panel of the figure shows simulated draws from a DGP in which parallel trends holds. The y-axis shows the difference in sample means between the treatment and control group in each period ($\Delta\bar{y}_t$). I highlight in blue the draws of the data in which the t-stat on the pre-period coefficient $\hat{\beta}_{-1}$ is less than one in absolute value, and I highlight in red the draws where the t-stat exceeds 1 in absolute value. The bottom panel shows the distribution of the treatment effects estimates unconditionally and conditional on having found an insignificant pre-trend. The figure illustrates that conditional on finding an insignificant pre-trend, the treatment effects remain unbiased but have a lower variance than unconditionally.

have lower variance, so traditional confidence intervals will tend to overcover. For larger violations of parallel trends, the bias in the OLS estimates is worse conditional on finding an insignificant pre-trend, leading to undercoverage in the conditional confidence intervals. These dynamics are captured in Figure 3, which summarizes the performance of the OLS treatment effects estimates and CIs under linear violations of parallel trends as a function of the underlying slope. We see that for underlying slopes close to zero, traditional CIs overcover both the true OLS coefficient and true treatment effect conditional on finding an insignificant pre-trend. However, as the bias gets larger, the traditional CIs undercover the true OLS coefficient conditional on finding an insignificant pre-trend. Additionally, for larger values of the underlying trend, undercoverage of the true treatment effect is substantially worse conditional on not finding a significant pre-trend.

2.5 Implications of publication rules that require pre-testing

So far, the analysis has conditioned on whether or not there is an underlying trend in population. A natural follow-up question is what happens when researchers try many different studies, and parallel trends is satisfied in some of these but not others.

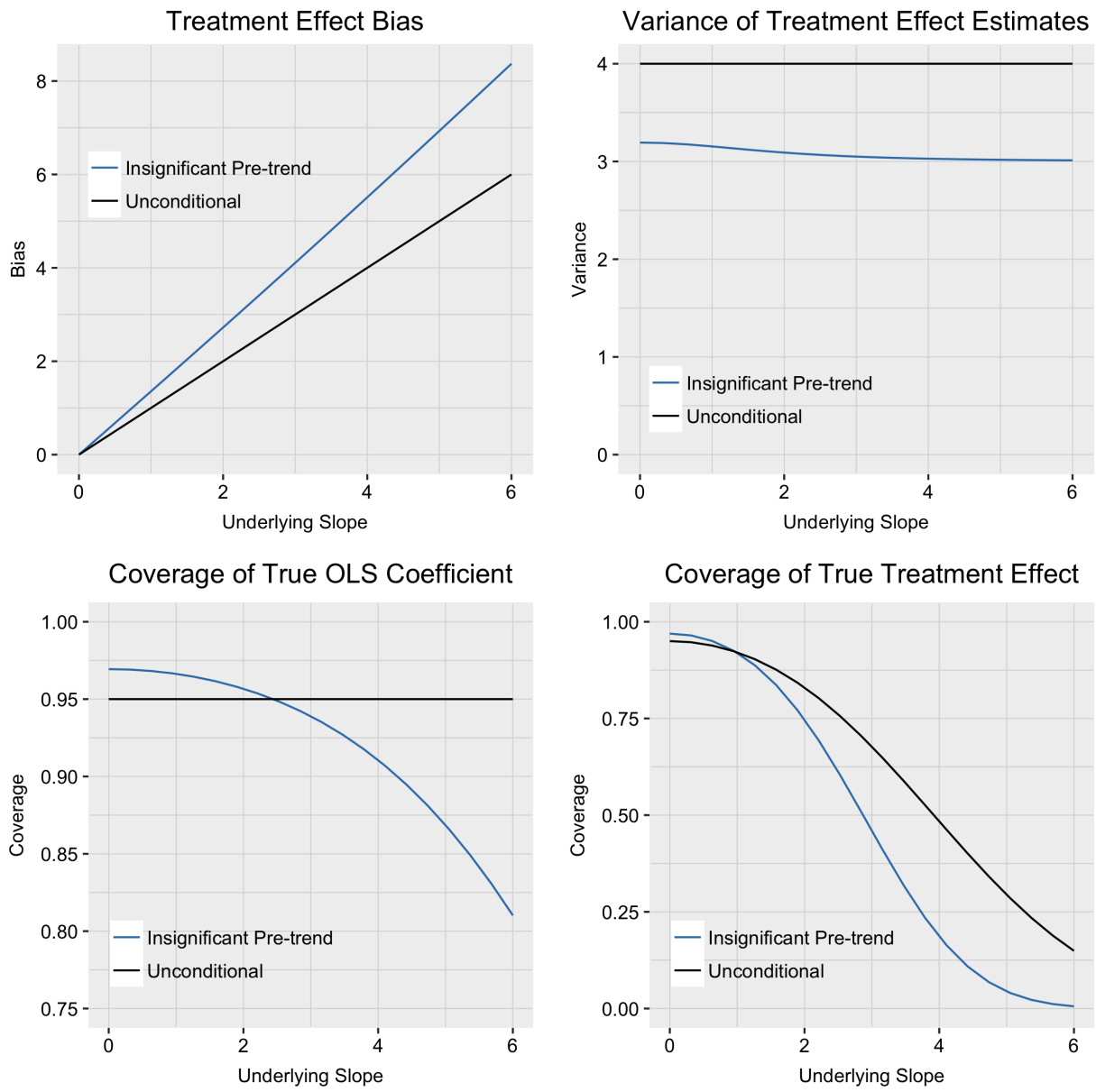
This section illustrates via a simple extension of the base model that requiring insignificant pre-trends in order to publish can either reduce or increase bias in published work in this setting. Intuitively, when we require an insignificant pre-trend in order to publish, there is a tradeoff between two effects: the parallel trends assumption holds for a higher fraction of studies, but for any given violation of parallel trends, the expected bias is worse conditional on not finding a significant pre-trend. As a result, whether pre-trends testing reduces or increases bias in published work will depend on both the power of the pre-test to detect meaningful violations of parallel trends, and the fraction of latent research designs in which parallel trends holds.

To clarify the tradeoffs of requiring insignificant pre-trends in order to publish, we consider a simple extension to the stylized model in which parallel trends holds in fraction $1-\alpha$ of latent studies, and in fraction α of latent studies there is a linear violation of parallel trends with slope $\bar{b} > 0$. We will denote the underlying slope in population by b, so that $b = \bar{b}$ when parallel trends is violated, and $b = 0$ when parallel trends holds. If we didn't test for parallel trends and published everything, the expected bias in published studies would be:

$$Bias^{\text{No test}} = P(b = \bar{b})\bar{b} = \alpha\bar{b}.$$

Likewise, if we only publish cases where we accept the pre-trend, the bias in published studies

Figure 3: Bias, Variance, and Coverage of OLS Treatment Effect Estimates Under Linear Violations of Parallel Trends



Note: This figure shows the performance of the OLS treatment effect estimate under linear violations of parallel trends, both unconditionally and conditional on not detecting a pre-trend ($|t| < 1$). The top two panels show the bias and variance of the treatment effects estimates. The bottom left panel shows the coverage of the true OLS coefficient for a nominal 95% interval; the bottom right panel shows the coverage of the true treatment effect.

is:

$$Bias^{\text{Test}} = P(b = \bar{b} \,|\, Accept)\mathbb{E}\left[bias \,|\, b = \bar{b}, Accept\right].$$

The ratio of biases across the two regimes is then:

$$\frac{Bias^{Test}}{Bias^{Notest}} = \underbrace{\frac{P(b = \bar{b} \,|\, Accept)}{P(b = \bar{b})}}_{\substack{\text{Relative fraction of} \\ \text{studies with biased} \\ \text{design } (\leq 1)}} \cdot \underbrace{\frac{\mathbb{E}\left[bias \,|\, b = \bar{b}, Accept\right]}{b}}_{\substack{\text{Ratio of bias when accept} \\ \text{biased design } (\geq 1)}}. \tag{3}$$

Equation (3) makes clear the tradeoffs involved in requiring an insignificant pre-trend in order to publish. The first term represents the relative fraction of published studies with a biased design ($b = \bar{b}$) across the two regimes. Pre-testing makes us relatively more likely to accept a study where parallel trends holds, so this term will tend to be less than 1. However, the second term represents the ratio of biases in the published studies where parallel trends does not hold in population. As demonstrated in Section 2.2, this bias is worse conditional on the pre-test, so the second term will be greater than 1.

The bias under the pre-testing regime will tend to be worse when either the fraction of latent studies with a biased design (α) is high, or if the pre-test has low power. To see why this is the case, using Bayes' rule we can re-write the first term in (3) as:

$$\frac{1}{\alpha + (1 - \alpha)BF} \tag{4}$$

where

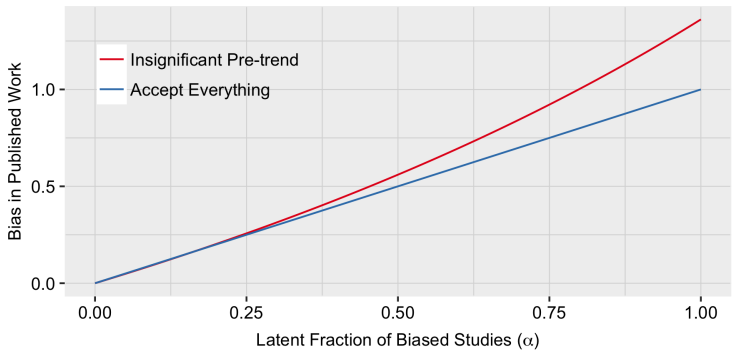
$$BF := \frac{P(Accept|b = 0)}{P(Accept|b = \bar{b})}$$

is the Bayes factor, i.e. the ratio of the likelihood of finding an insignificant pre-trend when parallel trends holds relative to when it is violated. The pre-testing regime will tend to have larger bias when the expression in (4) is close to 1. This will occur if α is close to 1, meaning that a high fraction of latent research designs are biased, or if the Bayes Factor is close to 1, meaning that the pre-test has low power.

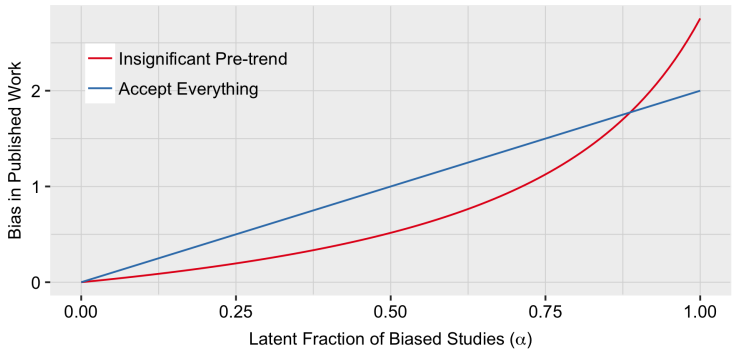
These dynamics are captured in Figure 4, which shows the (mean) bias in published studies as a function of α for three values of $\bar{b}$. These three values of $\bar{b}$ correspond with β_{pre} being 1, 2, and 3 standard errors away from zero, and lead to Bayes factors of 1.4, 4.3, and 30. The top panel shows that when the Bayes factor is small, so that the pre-test is poorly

Figure 4: Comparing bias in published studies when requiring an insignificant pre-trend to publish versus publishing everything

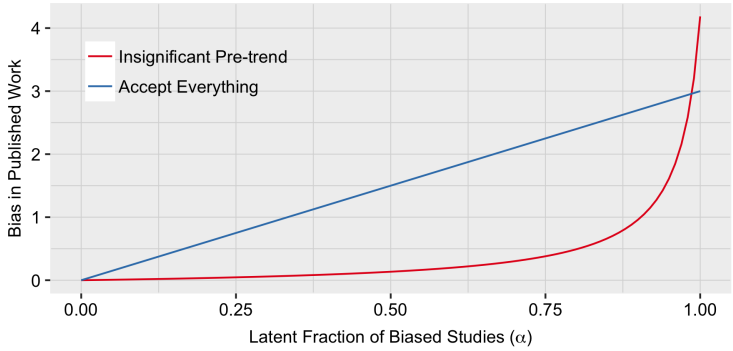
(a) $\bar{b} = \sigma_{pre}$, Bayes factor = 1.4



(b) $\bar{b} = 2\sigma_{pre}$, Bayes factor = 4.3



(c) $\bar{b} = 3\sigma_{pre}$, Bayes factor = 30



Note: Each figure shows the (mean) bias in published work in the setting described in Section 2.5 as a function of the fraction of latent studies in which parallel trends is violated (α). The Insignificant Pre-trend regime only publishes studies in which the t-stat on the pre-period is insignificant. The three panels show results for three different values of the slope of the underlying trend when parallel trends fails. See Section 2.5 for further detail.

powered, requiring an insignificant pre-test in order to publish leads to weakly larger bias in published work for all values of α, with larger differences when α is large. In the second and third panels, where the power of the pre-test is larger, we see that requiring an insignificant pre-test to publish can substantially reduce bias for lower values of α, but will nonetheless exacerbate bias if α is sufficiently large.

An implication of this section is that researchers should consider both the power of their pre-test to reject meaningful violations of parallel trends as well as the ex ante plausibility of the research design (as proxied by $1-\alpha$). If either of these is low, then pre-testing for an insignificant pre-trend in the usual way will likely be ineffective, and can even increase bias in published work.

3 Theory: The effect of pre-trends testing in the more general set-up

Section 2 considered the performance of conventional treatment effects estimates after pre-testing in a stylized setting with 3 periods, i.i.d. shocks to the outcome across periods, and linear violations of parallel trends. This section formalizes the intuition from Section 2 and extends the analysis to allow for additional periods, more complicated covariance structures, and non-linear violations of parallel trends.

3.1 The generalized set-up

I consider a setting where the researcher observes a vector of pre-period and post-period coefficients that is jointly normally distributed with known variance:

$$\begin{pmatrix} \hat{\beta}_{post} \\ \hat{\beta}_{pre} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta_{post} \\ \beta_{pre} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

I denote by K the dimension of the pre-period coefficient vector $\hat{\beta}_{pre}$, and by M the dimension of the post-period coefficients $\hat{\beta}_{post}$. For ease of notation, I will consider the case where $M = 1$ unless noted otherwise; all of the results for $M = 1$ will then apply to each individual post-period coefficient in the case when $M > 1$.

I will analyze the properties of the distribution of $\hat{\beta}_{post}$ conditional on a pre-test of the pre-trends coefficients – i.e. conditional on the event that $\hat{\beta}_{pre} \in B$ for some set B. For instance, researchers often test to see whether any of the pre-period coefficients is individually statistically significant at the 5% level, which is captured by the event $\hat{\beta}_{pre} \in B_{NIS} := \{\hat{\beta}_{pre} : |\hat{\beta}_{pre,j}|/\sqrt{\Sigma_{jj}} \leq 1.96 \text{ for all } j\}$.

It will sometimes be useful to decompose the population mean as

$$\begin{pmatrix} \beta_{post} \\ \beta_{pre} \end{pmatrix} = \begin{pmatrix} \tau_{post} \\ 0 \end{pmatrix} + \begin{pmatrix} \delta_{post} \\ \delta_{pre} \end{pmatrix},$$

where τ is the true causal parameter of interest, and δ represents the (unconditional) bias in conventional estimates from an underlying trend. For instance, in the example in Section 2, the true treatment effect was $\tau_{post} = 0$, but the researcher estimating regression (2) would have bias from the underlying trend given by $\begin{pmatrix} \delta_{post} \\ \delta_{pre} \end{pmatrix} = \begin{pmatrix} g(1) - g(0) \\ g(-1) - g(0) \end{pmatrix}$.

The finite-sample normal model specified above will hold exactly if we assume normal errors, as in the example in Section 2, and can also be thought of as an asymptotic approximation, since a wide variety of estimation procedures will yield asymptotically normal coefficients via the central limit theorem.[4] For instance, the traditional two-way fixed effects model (2) will lead to asymptotically normal coefficients as N grows large under mild regularity conditions.

The finite-sample normal model is sufficiently broad to cover other procedures as well. For example, Freyaldenhoven et al. (2018) propose an alternative method for estimating dynamic treatment effects in the linear panel setting, which identifies the causal parameter of interest if there is a pre-trend in a covariate that is not affected by the treatment but is affected by the same confounds as the outcome of interest. Their estimator can be written as a GMM estimator, and hence will be asymptotically normal. The results here can be used to analyze the properties of their estimator conditional on having not found significant pre-period placebo effects of the treatment on the outcome of interest.

3.2 A formula for how pre-testing impacts treatment effect bias

My first result provides a formula for the bias in the treatment effect estimate conditional on passing a pre-test for parallel trends.

Proposition 1. *For any conditioning set B,*

$$\begin{aligned} \mathbb{E}\left[\hat{\beta}_{post} \,|\, \hat{\beta}_{pre} \in B\right] &= \beta_{post} + \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] - \beta_{pre}\right) \\ &= \tau_{post} + \delta_{post} + \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] - \beta_{pre}\right). \end{aligned}$$

The formula in Proposition 1 illustrates that the bias in the treatment effect estimate is equal to the unconditional bias (δ_{post}) *plus* an additional term, which I will refer to as the

[4] More specifically, the normal model above can be thought of as the limiting experiment of a sequence of distributions in which the bias δ is on the order of $\sqrt{n}$. The local-to-0 approximation captures the fact that in finite samples pre-trends tests may have power strictly between 0 and 1.

"pre-test bias." The pre-test bias depends on the distortion to the mean of the pre-period coefficients from pre-testing, as well as on the normalized covariance between the pre-period and post-period coefficients.

3.3 Unbiasedness after pre-testing when parallel trends holds

In Section 2, we saw that when the parallel trends assumption was true, $\hat{\beta}_{post}$ was unbiased conditional on not finding a significant pre-trend. A corollary of Proposition 1 is that this result extends to the general set-up, so long as the pre-test is symmetric around 0 in the sense that we reject the hypothesis of parallel pre-trends for $\hat{\beta}_{pre}$ iff we reject the hypothesis for $-\hat{\beta}_{pre}$. This property holds for any two-sided test of significance.

Proposition 2 (No pre-test bias under parallel trends)**.** *Suppose that parallel trends holds, so that $\delta_{pre} = \delta_{post} = 0$. If the pre-test B is such that $\hat{\beta}_{pre} \in B$ iff $-\hat{\beta}_{pre} \in B$, then*

$$\mathbb{E}\left[\hat{\beta}_{post} \,|\, \hat{\beta}_{pre} \in B\right] = \tau_{post}.$$

3.4 Conditions under which pre-testing amplifies bias

In the stylized example in Section 2, we saw that when in population there was a linear underlying trend, the bias in the treatment effect estimate was worse conditional on having not detected a significant pre-period – i.e. the pre-test bias and the bias from trend went in the same direction. In this section, I show that under certain conditions on the covariance structure, this feature holds regardless of the number of pre-periods or the functional form of the underlying trend. I first state the formal result, and then discuss the assumptions on the covariance structure, which amount to homoskedasticity in the general multiple-period case but are less restrictive in other situations.

Assumption 1. *Let K denote the dimension of $\hat{\beta}_{pre}$.*

1. *If $K = 1$, then we assume that $\Sigma_{12} = \text{Cov}\left(\hat{\beta}_{pre}, \hat{\beta}_{post}\right) > 0$.*
2. *If $K > 1$, we assume that Σ has a common term σ^2 on the diagonal and a common term $\rho > 0$ off of the diagonal, with $\sigma^2 > \rho$.*

Proposition 3 (Sign of bias under upward pre-trend)**.** *Suppose that there is an upward pre-trend in the sense that $\delta_{pre} < 0$ (elementwise) and $\delta_{post} > 0$. If Assumption 1 holds, then*

$$\mathbb{E}\left[\hat{\beta}_{post} \,|\, \hat{\beta}_{pre} \in B_{NIS}\right] > \beta_{post} > \tau_{post}.$$

The analogous result holds replacing "$>$" with "$<$" and vice versa.

Assumption 1 is implied by a suitable homoskedasticity assumption in the canonical two-way fixed effects model. To see this, suppose that the data is generated from the model

$$y_{it} = \alpha_i + \phi_t + \sum_{s \neq 0} \underbrace{\beta_s}_{\tau_s + \delta_s} \times Treatment_i + \epsilon_{it}.$$

If the researcher estimates regression (2), then the estimated coefficients will be given by

$$\hat{\beta}_s = \beta_s + \Delta\bar{\epsilon}_s - \Delta\bar{\epsilon}_0,$$

where $\Delta\bar{\epsilon}_t$ is the difference in the average residuals for the treatment and control groups in period t, i.e. $\frac{1}{\#\{i|Treatment_i=1\}} \sum_{i \in \{i|Treatment_i=1\}} \epsilon_{it} - \frac{1}{\#\{i|Treatment_i=0\}} \sum_{i \in \{i|Treatment_i=0\}} \epsilon_{it}$. It follows that $\text{Cov}\left(\hat{\beta}_j, \hat{\beta}_k\right) = \text{Cov}\left(\Delta\bar{\epsilon}_j - \Delta\bar{\epsilon}_0, \Delta\bar{\epsilon}_k - \Delta\bar{\epsilon}_0\right)$. Hence, Assumption 1 will hold if $\Delta\bar{\epsilon}_t$ is *iid* across time, since we will have $\mathbb{V}\text{ar}\left[\hat{\beta}_k\right] = 2\sigma^2$ and $\text{Cov}(\hat{\beta}_k, \hat{\beta}_j) = \sigma^2$ for $\sigma^2 := \mathbb{V}\text{ar}\left[\Delta\bar{\epsilon}_t\right]$. A sufficient condition for $\Delta\bar{\epsilon}_t$ to be *iid* across time is for the individual-level errors ϵ_{it} to be *iid*.

With a one-dimensional pre-period coefficient, the requirements of Assumption 1 are less restrictive, as we only require the pre-period and post-period coefficients to be positively correlated. One can show, for instance, that if ϵ_{it} follows an AR(1) process, then Assumption 1 will hold so long as the AR coefficient is strictly less than 1. Although in practice having only one pre-period may be rare, I note that it is not unusual for researchers to test for a pre-trend using a parametric linear trend, such as

$$y_{it} = \alpha_i + \phi_t + \beta_{trend} \times t \times Treatment_i + \sum_{s>0} \beta_s \times 1[s = t] \times Treatment_i + \epsilon_{it}. \tag{5}$$

In this case, testing the significance of β_{trend} amounts to testing a one-dimensional pre-period coefficient.

What happens if Assumption 1 is not satisfied? One can construct examples using a covariance matrix that violates Assumption 1 in which the conditional bias is less than the unconditional bias, so there is no universal guarantee that the bias is exacerbated with arbitrary covariance structures. However, it is straightforward to calculate whether pre-testing will exacerbate bias for any particular underlying trend using Proposition 1. More specifically, Proposition 1 implies that the conditional bias in the treatment effect estimate will be worse than the unconditional bias iff the bias from trend δ_{post} and the pre-test bias $\Sigma_{12}\Sigma_{22}^{-1}\left(\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] - \beta_{pre}\right)$ have the same sign. For the common pre-test based on

the individual significance of the pre-period coefficients ($\hat{\beta}_{pre} \in B_{NIS}$), $\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right]$ can be calculated using the results of Manjunath and Wilhelm (2012). In Section 4, I apply this approach to calculate the pre-test bias under linear violations of parallel trends in a sample of published papers. I show that although in practice homoskedasticity typically does not hold, in most published papers the pre-test bias nonetheless goes in the same direction as the underlying trend.

A second limitation of the result in Proposition 3 is that the result applies only to the pre-test that no individual coefficient is statistically significant, as opposed to an arbitrary pre-test. It seems likely that similar results may be available for tests of joint significance using the results on elliptically-truncated normal from Tallis (1963) and Arismendi Zambrano and Broda (2016), but I leave this to future work. However, as in the previous paragraph, I note that the researcher interested in the pre-testing bias from a particular violation of parallel trends can calculate it using Proposition 1.[5]

3.5 How pre-testing affects treatment effect variance

Having analyzed the properties of the mean of the treatment effect estimate conditional on passing the pre-test for parallel trends, we now turn to analyzing its variance. We begin with a general formula, which expresses the conditional variance of the treatment effect in terms of its unconditional variance and the distortion to the variance of the pre-period coefficients.

Proposition 4.

$$\mathbb{V}ar\left[\hat{\beta}_{post}|\hat{\beta}_{pre} \in B\right] = \mathbb{V}ar\left[\hat{\beta}_{post}\right] + (\Sigma_{12}\Sigma_{22}^{-1})\left(\mathbb{V}ar\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] - \mathbb{V}ar\left[\hat{\beta}_{pre}\right]\right)(\Sigma_{12}\Sigma_{22}^{-1})'.$$

In the model in Section 2, we found that the variance of the treatment effect estimate conditional on passing the pre-test for parallel trends was smaller than the unconditional variance. We now show that that this feature holds more broadly under very mild conditions. In particular, we only require that the pre-test is convex, meaning that if we don't reject parallel trends for $\hat{\beta}_{pre,1}$ and $\hat{\beta}_{pre,2}$, then for $\alpha \in (0,1)$, we also will not reject parallel trends for $\alpha\hat{\beta}_{pre,1} + (1-\alpha)\hat{\beta}_{pre,2}$. This property holds for most common pre-tests – including tests of individual statistical significance, joint tests for significance, and tests for significant linear slopes.

[5]For non-rectangular pre-tests, the formula of Manjunath and Wilhelm (2012) can no longer be used to compute the conditional expectation. Nonetheless, there may be analytical formulas for $\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right]$ – e.g. Arismendi Zambrano and Broda (2016) for elliptical conditioning sets – and if not the truncated expectation can be calculated via simulation.

Proposition 5 (Pre-testing reduces variance)**.** *Suppose that B is a convex set. Then* $\mathbb{V}ar\left[\hat{\beta}_{post} \,|\, \hat{\beta}_{pre} \in B\right] \leq \mathbb{V}ar\left[\hat{\beta}_{post}\right]$.

4 The practical relevance of pre-testing distortions: evidence from a review of recent papers

Sections 2 and 3 illustrate that pre-trends testing can lead to undesirable distortions in the distribution of treatment effects estimates, particularly in the case when parallel trends is violated. The extent to which these distortions are consequential in practice, however, will depend on the power of pre-tests against meaningful violations of parallel trends. The practical relevance of theses concerns will also depend on whether the covariance matrices, which may not be homoskedastic in practice, are typically such that bias is exacerbated under monotone violations of parallel trends.

This section provides evidence that the theoretical concerns raised in the previous sections are relevant in practice. First, in a systematic review of recent papers in three leading economics journals, I illustrate that conventional pre-tests for parallel trends often have low power even against linear violations of parallel trends. This suggests that the possibility of failing to detect a meaningful violation of parallel trends is not merely a theoretical concern. Second, I show that while homoskedasticity typically does not hold in practice, the bias from pre-testing nonetheless typically amplifies the bias from an underlying trend, and can be of a substantial magnitude.

4.1 Selecting the sample of papers

I searched on Google Scholar for occurrences of the phrase "event study" in papers published in the *American Economic Review*, *AEJ: Applied Economics*, and *AEJ: Economic Policy* between 2014 and June 2018. I chose the phrase "event study" since papers that evaluate pre-trends often do so in a so-called "event study plot." The search returned 70 total papers that include a figure displaying the results from what the authors describe as an event-study.

For my analysis, I further restricted to papers meeting the following criteria:

1. The data to replicate the event-study plot was publicly available.

2. The event-study plot shows point estimates and confidence intervals for dynamic treatment effects relative to some reference period, which is normalized to zero.

3. The authors do not explicitly reject a causal interpretation of the event-study.

Meets criteria:	Number of Papers
Contains event study plot	70
& Replication data available	27
& Provides standard errors	18
& Normalizes a period to 0	15
& Doesn't reject causal interpretation	12

Table 1: Number of papers meeting criteria for inclusion in review of papers

Table 1 shows the number of papers that were eliminated by each of the criteria. Unfortunately, the constraint that the data be publicly available eliminated roughly two-thirds of the original sample of papers.[6] The second constraint eliminated two groups of papers. First, some papers portray the time-series of the outcome of interest for the treatment group and control group, typically without standard errors. I omit these papers, since I would like to rely on the author's determination of what the appropriate clustering scheme is for standard errors. Second, the restriction that a pre-period be normalized to zero primarily rules out a handful of papers employing a more traditional finance event-study, which examines the time-series of cumulative abnormal return around some event of interest. The final constraint eliminated a handful of papers in which the authors recognize that the pre-trends do not appear to be flat, and either subsequently add time-varying controls or suggest a non-causal interpretation.

Twelve papers contained event-study plots that matched all of the above criteria. Some of these papers present multiple event-study plots, many of which show robustness checks or heterogeneity analyses. I focus here on the first event-study plot presented in each paper, which I view as a reasonable proxy for the main specification in the paper.

4.2 What pre-tests are researchers using?

It is not entirely clear in practice what criteria researchers are using to evaluate pre-trends. By far the most commonly mentioned criterion is that none of the pre-period coefficients is individually statistically significant – e.g. "the estimated coefficients of the leads of treatments, i.e., δ_k for all $k \leq -2$ are statistically indifferent from zero" (He and Wang, 2017). However, many papers do not specify the exact criteria that they are using to evaluate pre-trends. Moreover, it is clear that a statistically significant pre-period coefficient does not necessarily preclude publication. As shown in Table 2, there is at least one statistically significant pre-trends coefficient in three of the 12 papers in my final sample, and in two

[6] I also omit one paper in which the replication code produced different results from the actual paper.

papers the pre-trends coefficients are also jointly significant.[7]

Paper	# Pre-periods	# Significant	Max |t|	Joint p-value	|t| for slope
Bailey and Goodman-Bacon (2015)	5	0	1.674	0.540	0.381
Bosch and Campos-Vazquez (2014)	11	2	2.357	0.137	0.446
Deryugina (2017)	4	0	1.090	0.451	1.559
Deschenes et al. (2017)	5	1	2.238	0.014	0.239
Fitzpatrick and Lovenheim (2014)	3	0	0.774	0.705	0.971
Gallagher (2014)	10	0	1.542	0.166	0.855
He and Wang (2017)	3	0	0.884	0.808	0.720
Kuziemko et al. (2018)	2	0	0.474	0.825	0.474
Lafortune et al. (2017)	5	0	1.382	0.522	1.390
Markevich and Zhuravskaya (2018)	3	0	0.850	0.591	0.676
Tewari (2014)	10	0	1.061	0.948	0.198
Ujhelyi (2014)	4	1	2.371	0.003	1.954

Table 2: Summary of Pre-period Event Study Coefficients

Note: This table provides information about the pre-period event-study coefficients in the papers reviewed. The table shows the number of pre-periods in the event-study, the fraction of the pre-period coefficients that are significant at the 95% level, the maximum t-stat among those coefficients, the p-value for a chi-squared test of joint significance, and the t-stat for the slope of the linear trend through the pre-period coefficients. See Section 4 for more detail on the sample of papers reviewed.

4.3 Evaluating power and pre-test bias in practice

In this section, I evaluate two questions: i) to what extent might pre-trends tests fail to detect meaningful violations of parallel trends?, and ii) to what extent will pre-testing exacerbate bias in practice?

In light of the emphasis in published work on the individual statistical significance of the pre-period coefficients, I base my calculations on this criterion. For each study in my sample, I evaluate the power of the pre-trends test to detect linear violations of parallel trends. Of course, we might also be concerned about non-linear violations of parallel trends – otherwise, we would just control parametrically for a linear time trend. However, it seems reasonable in most settings to consider the possibility of a linear violation of parallel trends, potentially alongside other functional forms for the violation. In this case, my calculations can be viewed as an upper bound on the ability of pre-trends tests to detect meaningful

[7] In none of the papers is the slope of the best-fit line through the pre-period coefficients significant at the 5% level. However, no paper mentions this as a criterion of interest, and one case falls just short of significance with a t-stat of 1.95.

violations of parallel trends.[8]

To evaluate the extent to which pre-testing will detect meaningful violations of parallel trends, I compute the linear slope for which the pre-test will achieve 50 or 90 percent power. To do this, I obtain the estimated variance-covariance matrix Σ for the pre- and post-period coefficients from each of the event-study regressions studied. Let Σ_{pre} denote the component of the covariance matrix corresponding with the pre-trends coefficients, and let $\beta_{pre}(\alpha) = \alpha \cdot (-K, ..., -1)$ denote a vector of means with an upward slope of α. I then ask the following question: if the pre-trends coefficients $\hat{\beta}_{pre} \sim \mathcal{N}(\beta_{pre}(\alpha), \Sigma)$, for what value of α would we reject the research design 50 (90) percent of the time? I refer to this slope as the slope against which we have 50 (90) percent power.

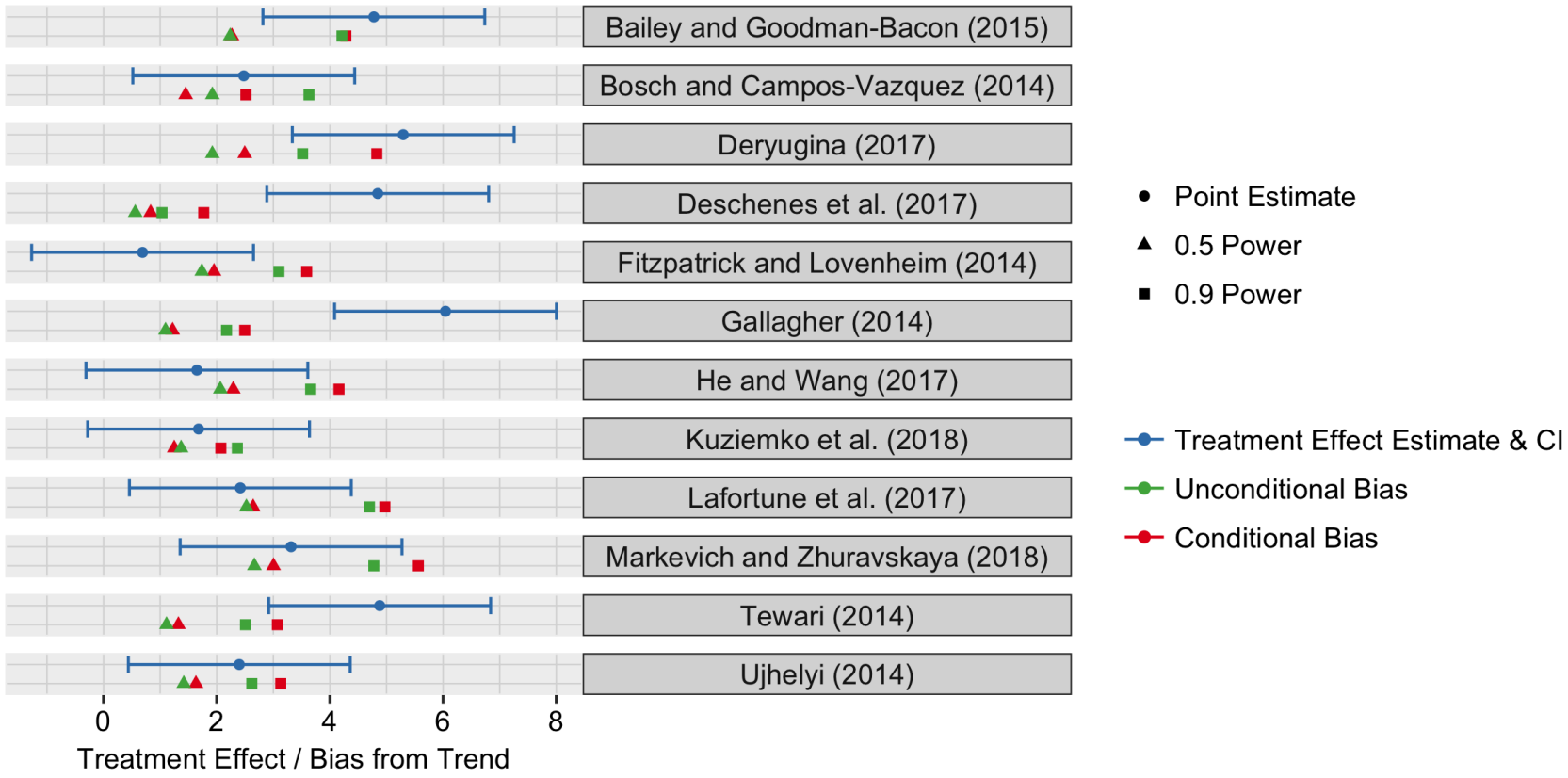
The magnitude of the violations of parallel trends against which we have 50 and 90 percent power can be sizeable, as illustrated in Figures 5 and 6. In each figure, I plot in green the (unconditional) bias in the treatment effect estimate that would be induced by each violation of parallel trends – Figure 5 shows the bias for the average of the post-treatment periods, whereas Figure 6 shows the bias for the first period after treatment only. I compare the magnitudes of the bias to the estimated treatment effect and confidence interval. In many cases, the magnitude of the bias from the trends for which we have only 50 or 90 percent power is a sizeable fraction of the treatment effect estimate, and even exceeds the magnitude of the estimated treatment effect in some cases.

I also find that the bias conditional on passing the pre-test can be substantially worse than the unconditional bias. In Figures 5 and 6, I plot in red the average bias that would be induced from the trends against which we have 50 and 90 percent power, *conditional* on not finding any significant pre-period coefficient. I calculate these values using the formula in Proposition 1. I also summarize the bias from pre-testing as a percentage of the bias from trend in Table 3. For the trend against which we have 50 percent power, the pre-test bias can be as much as 103 percent of the bias from the trend for the first period after treatment, and as much as 48 percent for the average of the post-periods.[9] Moreover, the pre-test bias and the bias from trend go in the same direction for the average treatment effect in all but two of the studies in the sample, and all but three of the studies for the first period. Thus, although not always true, the prediction of the direction of the bias from the homoskedastic case holds in most cases I consider.

[8]Formally, the worst-case bias over any class of potential violations that includes the class of linear violations is weakly worse than the worst-case bias over the class of linear violations.

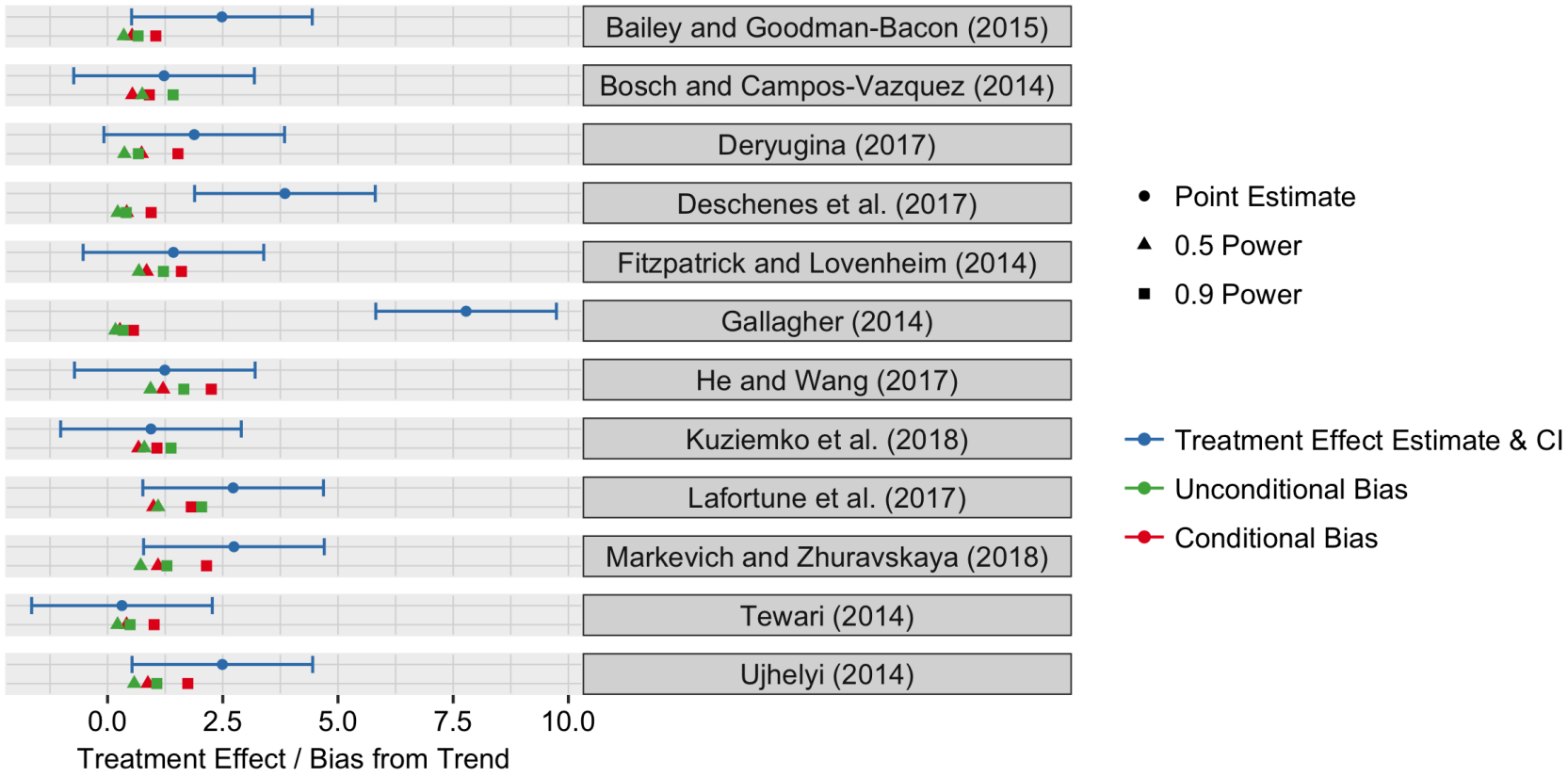
[9]We expect the bias from pre-testing to be a larger fraction of the bias from the trend in periods closer to treatment, since the bias from the trend grows linearly in the number of periods after treatment, whereas the pre-test bias need not grow over time (whether it does depends on the covariance between the pre-period and post-period coefficients).

Figure 5: OLS Estimates and Bias from Linear Trends for Which We Have 50 and 90 Percent Power – Average Treatment Effect



Note: I calculate the linear trend against which we would have a rejection probability of 50 and 90 percent if we rejected the research design whenever any of the pre-period event-study coefficients was statistically significant at the 5% level. I plot in red the bias that would result from such a trend conditional on not rejecting the research design; I plot in green the unconditional bias from such a trend. In blue, I plot the original OLS estimates and 95% CIs. All values are normalized by the standard error of the estimated treatment effect and so the OLS treatment effect estimate is positive. The parameter of interest is the average of the treatment effects in all periods after treatment began.

Figure 6: OLS Estimates and Bias from Linear Trends for Which We Have 50 and 90 Percent Power – First Period Treatment Effect



Note: I calculate the linear trend against which we would have a rejection probability of 50 and 90 percent if we rejected the research design whenever any of the pre-period event-study coefficients was statistically significant at the 5% level. I plot in red the bias that would result from such a trend conditional on not rejecting rejecting the research design; I plot in green the unconditional bias from such a trend. In blue, I plot the original OLS estimates and 95% CIs. All values are normalized by the standard error of the estimated treatment effect and so the OLS treatment effect estimate is positive. The parameter of interest is the treatment effect in the first period after treatment.

	Treatment Effect:			
	1st Period		All Periods	
	Power Against Slope:			
Paper	0.5	0.9	0.5	0.9
Bailey and Goodman-Bacon (2015)	51	58	1	2
Bosch and Campos-Vazquez (2014)	-29	-36	-25	-31
Deryugina (2017)	103	129	30	37
Deschenes et al. (2017)	88	133	48	71
Fitzpatrick and Lovenheim (2014)	25	32	12	16
Gallagher (2014)	57	66	11	15
He and Wang (2017)	29	36	11	14
Kuziemko et al. (2018)	-16	-22	-9	-12
Lafortune et al. (2017)	-9	-11	5	6
Markevich and Zhuravskaya (2018)	52	67	13	16
Tewari (2014)	90	107	19	22
Ujhelyi (2014)	51	63	15	19

Table 3: Percent Additional Bias Conditional on Passing Pre-test

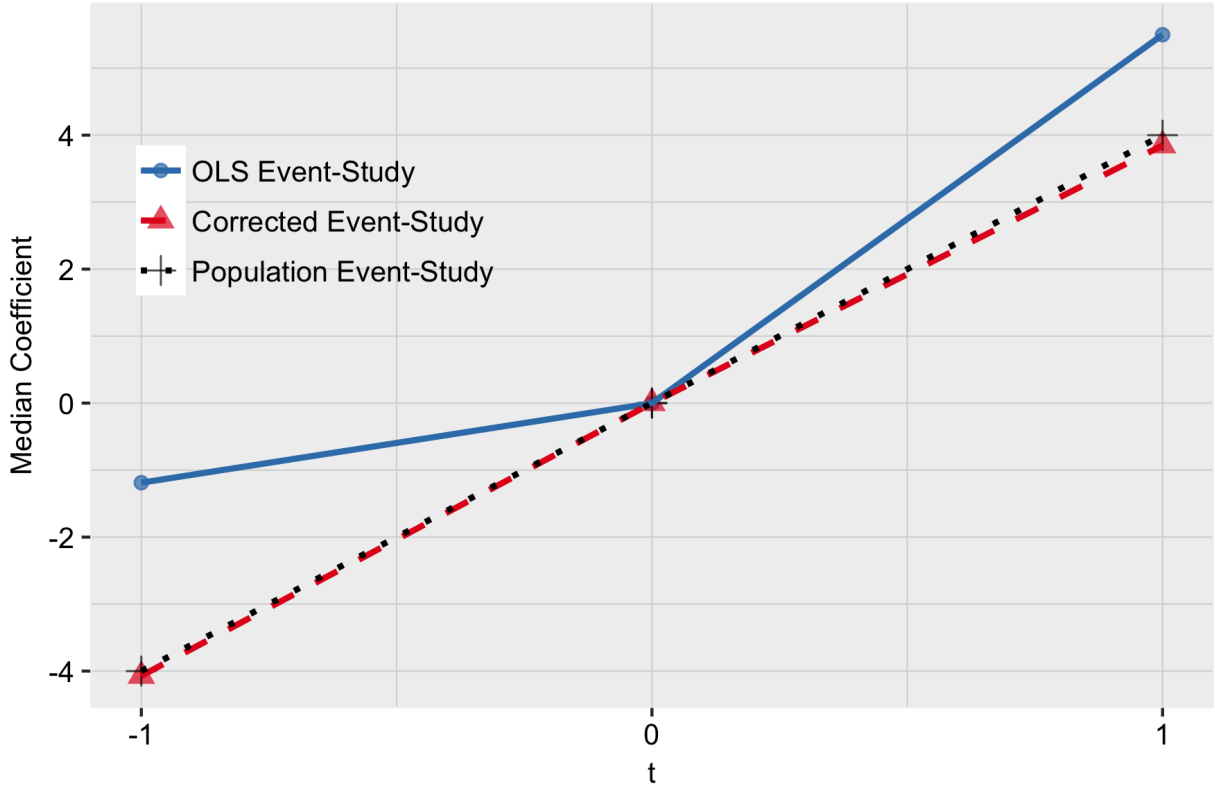
Note: This table shows the additional bias from conditioning on none of the pre-period coefficients being statistically significant as a fraction of the unconditional bias, i.e. $100 \cdot \frac{ConditionalBias - UnconditionalBias}{UnconditionalBias}$. See notes to Figures 5 and 6 for additional detail on the simulation exercise.

5 Pre-test Corrected Event Studies

The results developed so far suggest that many of the properties we expect of conventional event-study estimates and confidence intervals do not hold in the cases when they're actually analyzed – i.e. conditional on not detecting a significant pre-trend. How can we evaluate event-studies in light of the pre-test that we know has occurred? In this section, I derive alternative estimators, along with associated confidence intervals, that correct for the bias and coverage distortions induced by conditioning on having not observed a significant pre-trend.

The estimator I develop provides median-unbiased estimates for the true event-study coefficients, conditional on passing the pre-test. To make this concrete, consider the example from Section 2 in which in population there was a linear violation of parallel trends. Figure 7 shows in blue the median OLS event-study coefficients from estimating specification (2), conditional on not finding a significant pre-period coefficient. We see from the figure that conditional on passing the pre-test, we tend to observe a relatively flat pre-trend, and a large kink after treatment – despite the fact that in population the trend is a straight line. In other words, the pre-test tends to "mask" the underlying trend. By contrast, the estimator developed in this section, plotted in red in Figure 7, on-median shows a straight line in this

Figure 7: Median OLS and corrected event-study estimates, conditional on not finding a significant pre-trend, in the example from Section 2



Note: This figure shows the median OLS and corrected-event study estimates, conditional on not finding a significant pre-trend, in the example considered in Figure 1. The results are based on 1,000 simulations from the DGP considered in Figure 1. See Section 2 for additional details on the DGP considered, and Section 5 for details on the construction of the corrected event-study.

context.

5.1 Construction of the corrected estimator

In this section, I discuss the construction of an optimal median-unbiased estimator and confidence intervals for a parameter of the form $\eta'\beta$ conditional on passing the pre-test for parallel trends, where $\beta = (\beta_{pre}, \beta_{post})$, and η is a vector of the appropriate length. By setting η to be the appropriate basis vector, we can therefore do estimation and inference for each individual event-study coefficient, and we can then collect the results to form a corrected event-study. If we have M periods after treatment, we could also for instance do inference on the average post-period treatment effect by setting η to put weight 0 on the pre-period coefficients and weight $1/M$ on each of the post-period coefficients.

I begin with a derivation for the case where the researcher does a pre-test for a fixed specification. I then show that the results can be extended to cases where the researcher searches over multiple specifications – e.g. with different sets of control variables or focusing on different subpopulations – and selects the final specification on the basis of the observed pre-trends.

5.1.1 Correcting for a pre-test with a fixed specification

We begin by deriving the distribution of $\eta'\hat{\beta}$ conditional on the event $\hat{\beta} \in B$.[10] In general, the conditional distribution of $\eta'\hat{\beta}$ will depend on the full parameter vector β. We will therefore condition also on a minimal sufficient statistic for the other components of β.

Proposition 6 (Conditional distribution of $\eta'\hat{\beta}$)**.** *Let $\hat{\beta} = (\hat{\beta}_{post}, \hat{\beta}_{pre})$ and $\eta \neq 0$ be in $\mathbb{R}^{K+M}$. Define $c = \Sigma\eta/(\eta'\Sigma\eta)$ and $Z = (I - c\eta')\hat{\beta}$. Then*

$$\eta'\hat{\beta} \,|\, \hat{\beta} \in B, Z = z \sim \xi \,|\, \xi \in \Xi(z),$$

for $\xi \sim \mathcal{N}(\eta'\beta,\, \eta'\Sigma\eta)$, and $\Xi(z) := \{x : \exists \hat{\beta} \in B \text{ s.t. } x = \eta'\hat{\beta} \text{ and } z = (I - c\eta')\hat{\beta}\}$.

Having derived the conditional distribution $\eta'\hat{\beta}$, we can then make use of results on optimal quantile-unbiased estimators and inference for exponential family distributions, which were originally developed by Pfanzagl (1994). Similar techniques have been used recently in papers by Andrews and Kasy (2017) on publication bias, Lee et al. (2016) on inference for the LASSO, and Andrews et al. (2018) on inference for "winners".

I first state the result for pre-tests of a general form, and then show how the results simplify for common pre-tests.

Proposition 7 (Optimal quantile-unbiased estimation)**.** *Let $\eta \neq 0$ be in $\mathbb{R}^{K+M}$. Assume that $\hat{\beta} \in B$ with positive probability, and that Σ is full rank. Let F^{Ξ}_{μ,σ^2} denote the CDF of the normal distribution with mean μ and variance σ^2 truncated to the set Ξ. Define $\hat{b}_\alpha(\eta'\hat{\beta}, z)$ to be the value of x that solves $F^{\Xi(z)}_{x,\eta'\Sigma\eta}(\eta'\hat{\beta}) = \alpha$, for $\Xi(z)$ as defined in Proposition 6. Then for any $\alpha \in (0,1)$,*

$$P\left(\hat{b}_\alpha(\eta'\hat{\beta}, Z) \leq \eta'\beta \,|\, \hat{\beta} \in B\right) = \alpha.$$

Further, suppose that the parameter space for β is an open set, and that the distribution of $\eta'\hat{\beta} \,|\, Z, \hat{\beta} \in B$ is continuous for almost every Z. Then $\hat{b}_\alpha$ is uniformly most concentrated

[10] In a slight change of notation, I will now refer to B as the conditioning set for the full parameter vector $\hat{\beta} = (\hat{\beta}_{pre}, \hat{\beta}_{post})$ rather than for $\hat{\beta}_{pre}$ only. Note that we can write the event $\hat{\beta}_{pre} \in B_{pre} \subset \mathbb{R}^K$ as $\hat{\beta} \in B = \{(b_{pre}, b_{post}) \in \mathbb{R}^{K+M} \,|\, b_{pre} \in B_{pre}\}$.

in the class of level-α quantile-unbiased estimators, in the sense that for any other level-α quantile unbiased estimator $\tilde{b}_\alpha$, and any loss function $L(x, \eta'\beta)$ that attains its minimum at $x = \eta'\beta$ and is increasing as x moves away from $\eta'\beta$,

$$\mathbb{E}\left[L\left(\hat{b}_\alpha(\eta'\beta, Z), \eta'\beta\right) \mid \hat{\beta} \in B\right] \leq \mathbb{E}\left[L\left(\tilde{b}_\alpha(\eta'\beta, Z), \eta'\beta\right) \mid \hat{\beta} \in B\right].$$

Thus, conditional on $\hat{\beta} \in B$, $\hat{b}_{0.5}(\eta'\hat{\beta}, Z)$ is an (optimal) median-unbiased estimate of $\eta'\beta$, and the interval $[\hat{b}_{\alpha/2}(\eta'\hat{\beta}, Z), \hat{b}_{1-\alpha/2}(\eta'\hat{\beta}, Z)]$ is a $1-\alpha$ confidence interval for $\eta'\beta$.

I now show that the form of $\Xi(z)$ is particularly simple for the common test that none of the pre-period coefficients is statistically significant. To do this, we first note that we can define a matrix A^{NS} and vector b^{NS} such that $\hat{\beta}_{pre} \in B_{NIS}$ iff $A^{NS}\hat{\beta} \leq b^{NS}$. In particular, it is easy to verify that this holds for $A^{NS} = \begin{pmatrix} I_{K\times K} & 0_{1\times K} \\ -I_{K\times K} & 0_{1\times K} \end{pmatrix}$ and $b^{NS} = \begin{pmatrix} c_\alpha \times \sqrt{diag(\Sigma)} \\ c_\alpha \times \sqrt{diag(\Sigma)} \end{pmatrix}$. We can therefore make use of a result by Lee et al. (2016) regarding polyhedral conditioning sets.

Proposition 8 (Application to polyhedral conditioning sets)**.** *Suppose that the conditioning set $B = \{\beta \,|\, A\beta \leq b\}$ for A an $R \times K + M$ matrix and b an $R \times 1$ vector. Then $\Xi(z)$, as defined in Proposition 6, is an interval in $\mathbb{R}$, with endpoints $V^-(z)$ and $V^+(z)$ given by:*

$$V^-(z) = \max_{\{j:(Ac)_j<0\}} \frac{b_j - (Az)_j}{(Ac)_j} \tag{6}$$

$$V^+(z) = \min_{\{j:(Ac)_j>0\}} \frac{b_j - (Az)_j}{(Ac)_j}. \tag{7}$$

Additionally, if $\mathbb{P}\left(\hat{\beta} \in B\right) > 0$, then the conditions for the optimality of the α-quantile estimator in Proposition 7 are met.

I also derive the form of $\Xi(z)$ for tests based on a quadratic form of the parameters, such as tests based on the joint significance or the euclidean norm of the pre-period coefficients. In particular, $\hat{\beta}_{pre}$ will be insignificant iff the F-statistic $\hat{\beta}'_{pre}\Sigma_{pre}\hat{\beta}_{pre}$ is less than a threshold determined by the F-distribution. Likewise, the euclidean norm of $\hat{\beta}_{pre}$ is less than a threshold c iff $\hat{\beta}'_{pre}\hat{\beta}_{pre} < c^2$.

Proposition 9. *Let $B = \{\beta \,|\, \beta' A\beta \leq b\}$ for A an $(K+M) \times (K+M)$ matrix and b a scalar. Let $\mathcal{A} = c'Ac$, $\mathcal{B} = 2c'Az$, $\mathcal{C} = z'Az - b$, and $\mathcal{D} = \mathcal{B}^2 - 4\mathcal{A}\cdot\mathcal{C}$, for c and z as defined in Proposition 6. Then:*

1. *If $\mathcal{A} > 0, \mathcal{D} \geq 0$, $\Xi(z) = \left[\frac{-\mathcal{B}-\sqrt{\mathcal{D}}}{2\mathcal{A}}, \frac{-\mathcal{B}+\sqrt{\mathcal{D}}}{2\mathcal{A}}\right]$.*

2. *If* $\mathcal{A} < 0, \mathcal{D} \geq 0$, $\Xi(z) = \left(-\infty, \frac{-\mathcal{B}+\sqrt{\mathcal{D}}}{2\mathcal{A}}\right] \cup \left[\frac{-\mathcal{B}-\sqrt{\mathcal{D}}}{2\mathcal{A}}, \infty\right)$.

3. *If* $\mathcal{A} < 0, \mathcal{D} < 0$, $\Xi(z) = \mathbb{R}$.

4. *If* $\mathcal{A} > 0, \mathcal{D} < 0$, *then* $\Xi(z) = \emptyset$.

5. *If* $\mathcal{A} = 0$, $\mathcal{B} > 0$ *then* $\Xi(z) = (-\infty, -\frac{c}{b}]$.

6. *If* $\mathcal{A} = 0$, $\mathcal{B} < 0$, $\Xi(z) = [-\frac{c}{b}, \infty)$.

7. *If* $\mathcal{A} = 0$, $\mathcal{B} = 0$, *then* $\Xi(z) = \mathbb{R}$ *if* $\mathcal{C} \leq 0$ *and* $\Xi(z) = \emptyset$ *if* $\mathcal{C} > 0$.

   Additionally, if $\mathbb{P}\left(\hat{\beta} \in B\right) > 0$, *then the conditions for the optimality of the* α*-quantile estimator in Proposition 7 are met.*

5.1.2 Correcting for specification search using pre-trends

So far, I have considered the case where the researcher accepts or rejects a fixed research design on the basis of pre-trends. In practice, however, researchers may choose among multiple specifications on the basis of pre-trends. For instance, a researcher may first evaluate tests for pre-trends in a large sample, and then upon finding a significant pre-trend, restrict to a subsample in which the pre-trends appear to be better. Likewise, a researcher may evaluate the pre-trends both with and without certain controls in their regression, and choose the specification with the flattest observable pre-trends.

The machinery developed so far can easily be adapted to handle selection among a finite number of specifications on the basis of the pre-trends. Suppose that we have M models, each with estimated event-study coefficients $\hat{\beta}^m = (\hat{\beta}^m_{pre}, \hat{\beta}^m_{post})$. Let $\hat{\beta}^{stacked} = (\hat{\beta}^1, ..., \hat{\beta}^M)$ denote the stacked vector of coefficients across the M models. For OLS, the stacked vector of coefficients can be estimated using generalized least squares (also known as Seemingly Unrelated Regressions (SUR) in this context), and so will typically be asymptotically normal. Under a normal approximation for $\hat{\beta}^{stacked}$, we can immediately apply the results from the previous section to obtain median-unbiased estimates and valid confidence intervals for each of the elements of β^m, conditional on model m being chosen. That is, letting B^*_m denote the set of values for $\hat{\beta}$ such that model m is chosen, we can obtain adjusted estimates with the property that $\mathbb{P}\left(\hat{b}^m_j < \beta^m_j \mid \hat{\beta} \in B^*_m\right) = 0.5$.

Implementing these corrected estimates in practice requires calculation of the set $\Xi(z)$ accounting for the model selection rule. In Appendix B, I show that $\Xi(z)$ can be easily calculated for many applicable model selection rules, including decision rules where the researcher tries a series of models and stops when she finds one without a significant pre-trend, or where she chooses the model with the smallest pre-trend.

5.2 Practical Recommendations

In practice, I recommend that researchers who rely on pre-tests for parallel trends report corrected event-studies along with power calculations for what they perceive to be economically relevant violations of parallel trends. For a given pre-trend of interest δ_{pre}, the power of the pre-test B is the probability that a $\mathcal{N}(\delta_{pre}, \Sigma_{pre})$ variable falls outside the region B, which can be easily calculated by simulation or, for instance, using the R package mvtnorm (Genz et al., 2018). The power calculations are important because although the corrections proposed in this section correct the bias from pre-testing, they do not solve the issue that the power of the pre-test against violations of parallel trends may be low. In other words, while the corrected point estimates for the pre-period event-study coefficients will be median-unbiased, the confidence intervals for the pre-period coefficients may be sufficiently wide that it is difficult to distinguish between no underlying trend and an underlying trend that would meaningfully bias the treatment effects estimates.

The relevant functional form for the potential violation of parallel trends depends on the economic context.[11] In many contexts, we may be worried that the outcome of interest was growing steadily at a different rate relative to the control group, in which case considering linear violations of parallel trends would be a sensible approach. In other contexts, we may worry that the policy of interest was implemented in response to shocks that occurred in close proximity to the timing of treatment, in which case the researcher should consider violations of parallel trends where the treatment and control group between to diverge only in the pre-periods close to treatment.

Finally, an important caveat to these corrections is that journal editors may be inclined to only publish papers in which the corrected event-study plot still shows a clear break from trend at the time of treatment. However, the guarantees of unbiasedness and proper coverage are valid conditional on the original pre-test for parallel trends, *not* conditional on passing the original pre-test *and* having the event-study "survive" the correction. Thus, if all authors apply these corrections to their working papers, these desireable properties will hold over the sample of working papers, but likely will not hold conditional on the paper being published. In a sense, this is analogous to the application of Bonferroni adjustments to studies with multiple hypothesis tests. An author who Bonferroni-adjusts all of her studies will falsely

[11] Kahn-Lang and Lang (2018) rightly observe that the parallel trends assumption necessary for identification differs depending on the functional form used for the outcome (e.g. levels versus logs), and argue that the choice of functional form should be motivated by context-specific economic knowledge. I argue here that when authors test for parallel trends, they are trying to differentiate between different models of the world, in some of which the parallel trends assumption holds and in some of which it is violated. As in Kahn-Lang and Lang (2018), the relevant models to consider should likewise be determined by context-specific economic knowledge.

reject the null in no more than 5% of the studies she runs (assuming a 5% significance threshold). However, if journals are more likely to accept the papers where the results are significant after Bonferroni-adjustment, then she may falsely reject the null in more than 5% of her published papers. Likewise, if a researcher applies my corrections for pre-testing to each project where she does not reject the research design on the basis of pre-trends, median-unbiasedness and proper coverage will hold over the event-study plots in her studies where she accepts the research design. (Likewise, if she corrects for specification search on the basis of pre-trends, these properties will hold in the event study plots for the chosen model.) However, if the studies that show a clear break in trend in the corrected event-study are more likely to be published, then these properties will not hold over the sample of the author's published papers. I refer the reader interested in correcting published estimates to the supplement of Andrews and Kasy (2017), who show how such corrections can be done provided that one knows the probability of publishing as a function of the both the estimated pre-trend coefficients and the estimated treatment effects.

5.3 Simulations based on the survey of papers

I evaluate the performance of the corrected event-studies in simulations based on the survey of papers discussed in Section 4. As in Section 4, for each paper I consider a setting where the unconditional distribution of the event-study coefficients $\hat{\beta}$ is $\mathcal{N}(\beta, \Sigma)$, where Σ is the estimated variance covariance matrix in the regression in the paper. I consider the case where in population parallel trends holds, so that $\beta = 0$, and where β captures a linear upward trend such that the power of the pre-test is 0.5 or 0.9 (see Section 4.3 for details on this calculation). For each value of the underlying trend, I conduct 10,000 Monte Carlo simulations, and I evaluate the performance of conventional estimates and of the corrected event-studies in the cases that pass the pre-test – i.e. where no significant pre-trend coefficient is detected.

I calculate the median bias for each estimator relative to the population event-study coefficients β, with the results shown in Table 4. For comparability between studies with different units, I measure the median bias in the estimate for the coefficient β_i as the difference between the exceedance probability $\mathbb{P}\left(\hat{\beta}_i > \beta_i\right)$ and 0.5, as in Andrews et al. (2018). A median bias of 0 therefore represents unbiasedness, whereas the maximum value of 0.5 represents the case where $\hat{\beta}_i$ always exceeds β_i. I compute these probabilities for each individual event-study coefficient, and then report the average across all pre-periods and all post-periods for each study.

The first column of Table 4 shows that when parallel trends holds, conventional estimates

are median-unbiased for the true event-study coefficients conditional on passing the pre-test. However, when there is a positive pre-trend – so the pre-period coefficients have a population mean below zero – the conventional pre-period estimates are biased upwards conditional on passing the pre-test. These biases can be quite sizeable, with many of the values approaching the maximum of 0.5. Likewise, in the large majority of papers, the post-period coefficients are also upward biased when there is an upward underlying trend. These results confirm the intuition from before that – conditional on passing the pre-test – conventional event-study estimates mask an underlying trend: although in population the event-study should show a straight line, the cases that pass the pre-test tend to show a kink at $t = 0$, with pre-trends coefficients above their population mean (towards 0) and post-period coefficients also above their mean (away from zero). By contrast, the corrected event-study estimates are median unbiased for the true event-study coefficients conditional on passing the pre-test, regardless of the underlying trend.

Table 5 shows that the confidence intervals for the corrected event-study estimates also have (approximately) correct coverage for the population event-study coefficients, whereas the coverage of traditional CIs can be above or below the nominal rate. In particular, when parallel trends holds, coverage rates conditional on passing the pre-test are too low for conventional CIs, since point estimates are unbiased but – as discussed in Section 3.5 – the variance of the conventional estimator conditional on passing the pre-test is lower than the unconditional variance, on which the standard errors are based. When the underlying trend is non-zero, the conditional variance again remains lower than the unconditional variance, but the CIs are non-centered, so coverage can be either too high or too low.

Table 6 illustrates that the improved bias and coverage properties of the corrected event-study estimates do come at a cost in terms of the width of the confidence intervals. The table shows the median ratio of the width of the 95% CI for the corrected event-study estimates relative to the conventional CIs. A notable feature of the corrected event-study estimates, however, is that the inflation in the CI widths tend to be smaller in the case when parallel trends holds in population. When parallel trends holds, the inflation in the post-period CIs is between 0 and 26 percent, and the inflation in the pre-period CIs is between 33 and 61 percent. By contrast, the inflation in the CIs is often above 100 percent in both the pre- and post-periods when we consider the data-generating process with a linear slope against which we have 90 percent power. This disparity occurs because the adjusted CIs become large as the vector of pre-period coefficients approaches the rejection boundary, and this occurs more frequently the larger is the underlying trend in population.

Paper	Estimator: $\hat{\beta}^{Conventional}$			$\hat{b}_{0.5}$		
	Slope of Pre-trend:					
	0 Slope	0.5 Power	0.9 Power	0 Slope	0.5 Power	0.9 Power
Bailey and Goodman-Bacon (2015)	-0.00	0.22	0.44	-0.00	0.00	0.00
Bosch and Campos-Vazquez (2014)	-0.00	0.33	0.49	-0.00	-0.01	0.00
Deryugina (2017)	-0.00	0.20	0.44	-0.00	-0.01	0.01
Deschenes et al. (2017)	-0.01	0.21	0.44	-0.01	-0.00	-0.00
Fitzpatrick and Lovenheim (2014)	-0.00	0.23	0.43	-0.00	-0.00	0.01
Gallagher (2014)	-0.00	0.10	0.33	-0.00	0.00	0.02
He and Wang (2017)	0.00	0.29	0.48	0.00	-0.01	0.00
Kuziemko et al. (2018)	0.00	0.20	0.50	0.00	-0.01	-0.02
Lafortune et al. (2017)	0.01	0.27	0.48	0.01	-0.00	0.00
Markevich and Zhuravskaya (2018)	0.00	0.22	0.43	0.00	-0.01	0.01
Tewari (2014)	-0.00	0.15	0.37	-0.00	-0.00	0.01
Ujhelyi (2014)	0.01	0.27	0.50	0.01	-0.01	-0.01

(a) Median Bias for Pre-period Event-Study Coefficients, i.e. $\mathbb{P}\left(\hat{\beta}_{pre} > \beta_{pre} \,|\, \hat{\beta}_{pre} \in B\right) - 0.5$

Paper	Estimator: $\hat{\beta}^{Conventional}$			$\hat{b}_{0.5}$		
	Slope of Pre-trend:					
	0 Slope	0.5 Power	0.9 Power	0 Slope	0.5 Power	0.9 Power
Bailey and Goodman-Bacon (2015)	-0.00	0.01	0.03	-0.00	-0.00	0.01
Bosch and Campos-Vazquez (2014)	0.00	-0.18	-0.40	0.00	0.01	-0.01
Deryugina (2017)	-0.00	0.18	0.38	-0.00	-0.01	0.01
Deschenes et al. (2017)	-0.00	0.11	0.31	-0.00	-0.00	0.02
Fitzpatrick and Lovenheim (2014)	-0.00	0.07	0.15	0.00	0.00	-0.01
Gallagher (2014)	-0.00	0.04	0.09	-0.00	0.00	0.02
He and Wang (2017)	0.00	0.07	0.17	-0.00	-0.01	0.01
Kuziemko et al. (2018)	-0.00	-0.06	-0.14	-0.00	-0.01	0.01
Lafortune et al. (2017)	0.00	0.02	0.03	0.00	0.00	-0.01
Markevich and Zhuravskaya (2018)	0.00	0.12	0.26	0.00	0.00	0.03
Tewari (2014)	-0.00	0.06	0.16	-0.00	-0.01	0.01
Ujhelyi (2014)	-0.00	0.09	0.20	0.00	-0.00	-0.00

(b) Median Bias for Post-period Event-Study Coefficients, i.e. $\mathbb{P}\left(\hat{\beta}_{post} > \beta_{post} \,|\, \hat{\beta}_{pre} \in B\right) - 0.5$

Table 4: Median Bias for Event-Study Coefficients Conditional on Not Finding a Significant Pre-trend Coefficient

Note: This table shows the median bias of conventional and corrected event-study estimates in simulations based on recent papers in the AEA journals. All results are shown conditional on not finding a significant pre-period event study coefficient. For comparability across studies, I measure median bias as the probability the parameter estimate is greater than the population event-study coefficient, minus one half, as in Andrews et al. (2018). For each study, I calculate median bias for each individual pre-period and post-period coefficient, and then average over all of the pre periods (top panel) or post periods (bottom panel). I present results for the case where parallel trends holds in population (0 Slope), and when in population there is a linear violation of parallel trends against which we 0.5 or 0.9 power. See Section 5.3 for details.

Paper	Estimator: $\hat{\beta}^{Conventional}$			$\hat{b}_{0.5}$		
	Slope of Pre-trend:					
	0 Slope	0.5 Power	0.9 Power	0 Slope	0.5 Power	0.9 Power
Bailey and Goodman-Bacon (2015)	0.00	0.04	0.14	0.05	0.05	0.05
Bosch and Campos-Vazquez (2014)	0.00	0.04	0.22	0.05	0.05	0.05
Deryugina (2017)	0.00	0.04	0.13	0.05	0.05	0.07
Deschenes et al. (2017)	0.00	0.03	0.11	0.05	0.05	0.05
Fitzpatrick and Lovenheim (2014)	0.00	0.04	0.14	0.05	0.05	0.06
Gallagher (2014)	0.00	0.02	0.04	0.05	0.05	0.04
He and Wang (2017)	0.00	0.04	0.17	0.05	0.05	0.06
Kuziemko et al. (2018)	0.00	0.03	0.07	0.05	0.05	0.05
Lafortune et al. (2017)	0.00	0.04	0.17	0.05	0.05	0.05
Markevich and Zhuravskaya (2018)	0.00	0.04	0.12	0.05	0.05	0.05
Tewari (2014)	0.00	0.03	0.10	0.05	0.05	0.04
Ujhelyi (2014)	0.00	0.04	0.17	0.05	0.05	0.05

(a) Rejection Probability for True Pre-period Event-Study Coefficients

Paper	Estimator: $\hat{\beta}^{Conventional}$			$\hat{b}_{0.5}$		
	Slope of Pre-trend:					
	0 Slope	0.5 Power	0.9 Power	0 Slope	0.5 Power	0.9 Power
Bailey and Goodman-Bacon (2015)	0.05	0.05	0.05	0.05	0.05	0.05
Bosch and Campos-Vazquez (2014)	0.03	0.05	0.14	0.05	0.06	0.05
Deryugina (2017)	0.03	0.04	0.12	0.05	0.05	0.07
Deschenes et al. (2017)	0.03	0.03	0.06	0.05	0.05	0.04
Fitzpatrick and Lovenheim (2014)	0.05	0.05	0.06	0.05	0.05	0.05
Gallagher (2014)	0.04	0.04	0.05	0.05	0.05	0.05
He and Wang (2017)	0.04	0.05	0.06	0.05	0.05	0.06
Kuziemko et al. (2018)	0.04	0.03	0.02	0.05	0.05	0.05
Lafortune et al. (2017)	0.05	0.05	0.06	0.05	0.05	0.05
Markevich and Zhuravskaya (2018)	0.04	0.05	0.08	0.05	0.05	0.05
Tewari (2014)	0.04	0.04	0.07	0.05	0.05	0.05
Ujhelyi (2014)	0.04	0.04	0.05	0.05	0.05	0.04

(b) Rejection Probability for True Post-period Event-Study Coefficients

Table 5: Rejection Probabilities for Population Event-Study Coefficients, Conditional on Not Finding a Significant Pre-trend Coefficient

Note: This table shows the probability that a 95% confidence interval rejects the population event-study coefficient, using simulations based on a sample of recent papers in the AEJ journals. The first three columns show results for conventional CIs, whereas the latter three columns show for the corrected event-study estimates. The rejection rates are calculated for each individual event-study point estimate, and then the average is calculated pooling all pre-period coefficients (top panel) and post-period coefficients (bottom panel). I present results for the case where parallel trends holds in population (0 Slope), and when in population there is a linear violation of parallel trends against which we 0.5 or 0.9 power. See Section 5.3 for details.

Paper	Pre-periods			Post-periods		
	Slope of Pre-trend:					
	0 Slope	0.5 Power	0.9 Power	0 Slope	0.5 Power	0.9 Power
Bailey and Goodman-Bacon (2015)	1.50	1.76	2.47	1.00	1.00	1.01
Bosch and Campos-Vazquez (2014)	1.61	1.96	2.82	1.21	1.43	2.04
Deryugina (2017)	1.46	1.76	2.40	1.26	1.54	2.11
Deschenes et al. (2017)	1.50	1.81	2.52	1.14	1.39	2.06
Fitzpatrick and Lovenheim (2014)	1.43	1.71	2.29	1.01	1.04	1.18
Gallagher (2014)	1.55	1.77	2.37	1.04	1.07	1.20
He and Wang (2017)	1.45	1.81	2.58	1.01	1.05	1.22
Kuziemko et al. (2018)	1.33	1.62	2.38	1.10	1.46	2.44
Lafortune et al. (2017)	1.52	1.82	2.58	1.01	1.03	1.12
Markevich and Zhuravskaya (2018)	1.40	1.69	2.29	1.03	1.16	1.50
Tewari (2014)	1.57	1.73	2.16	1.08	1.13	1.31
Ujhelyi (2014)	1.49	1.79	2.64	1.03	1.13	1.44

Table 6: Median Ratio of CI Widths (Corrected Event Study CIs relative to Conventional CIs)

Note: This table shows the median ratio of the width of the 95% confidence interval for the corrected event-study estimates relative to conventional CIs, using simulations based on a sample of recent papers in the AEJ journals. The ratios are calculated for each individual event-study point estimate, and then the median is calculated pooling all pre-period coefficients (first 3 columns) and post-period coefficients (last 3 columns). I present results for the case where parallel trends holds in population (0 Slope), and when in population there is a linear violation of parallel trends against which we 0.5 or 0.9 power. See Section 5.3 for details.

6 Application to Dube et al. (2010)

In a highly influential paper, Dube et al. (2010) study the employment effects of the minimum wage, pooling all state-level minimum wages changes between 1990 and 2006. Dube et al. (2010) present results from two primary specifications. The first, the All Counties specification, is a regression at the county level with two-way county and time fixed effects, and thus effectively runs difference-in-differences between counties in states that changed their minimum wages and those that didn't. The second specification, which the authors prefer, restricts to border-county pairs and includes border-county-pair by time fixed effects. This Border Counties specification effectively pools difference-in-differences analyses between border counties in states that changed the minimum wage and their neighbors.

Dube et al. (2010) test for pre-trends by regressing current county-level employment on the change in the minimum wage 1 year in the future and 2-3 years in the future. Formally, their regression specification is

$$\begin{aligned}\ln Employment_{it} = & \underbrace{\eta_t \ln(MW_{it})}_{\text{Contemporaneous minimum wage}} + \\ & \underbrace{\eta_{t-4}(\ln(MW_{i,t+4} - MW_{i,t}))}_{\text{Minimum wage change 1-4 quarters in future}} + \\ & \underbrace{\eta_{t-12}(\ln(MW_{i,t+12} - MW_{i,t+4}))}_{\text{Minimum wage change 4-12 quarters in future}} + \\ & Controls + \epsilon_{it}\end{aligned}$$

the results of which are shown in the first two columns of Table 7. They write:

> "[W]e find evidence of a preexisting negative trend in restaurant employment for the [All Counties] specification. Restaurant employment was clearly low and falling during the $(t-12)$ to $(t-4)$ period. The η_{t-4} coefficient and the trend estimate $(\eta_{t-4} - \eta_{t-12})$ are both negative ... and significant at the 10% level. In contrast, none of the employment lead terms are ever significant or sizable in our contiguous county specification"

Although it is true that the pre-trends coefficients in the Border Counties specification are smaller in magnitude and not statistically significant, a closer examination reveals that the pre-trends test in the Border Counties specification may have low power against meaningful violations of parallel trends. Examining columns 1 and 2 of Table 7, we see that the

	All Counties	Border Counties	Border Counties (Adjusted)
η_{t-12}	-0.071	0.009	0.023
	$[-0.183, 0.041]$	$[-0.122, 0.14]$	$[-0.126, 0.237]$
η_{t-4}	-0.194*	0.05	0.151
	$[-0.419, 0.03]$	$[-0.288, 0.388]$	$[-0.478, 1.042]$
Trend	-0.124*	0.041	0.127
	$[-0.262, 0.015]$	$[-0.222, 0.304]$	$[-0.391, 0.862]$
η_t	-0.076	0.015	0.001
	$[-0.21, 0.058]$	$[-0.175, 0.205]$	$[-0.285, 0.218]$

Table 7: Original and Adjusted Estimates from Dube et al. (2010)

Note: This table shows the original point estimates and 95% confidence intervals for the two main specifications in Dube et al. (2010), as well as adjusted estimates for the Border Counties specification that account for the pre-test for pre-trends. The original results for the All Counties and Border Counties specification are replications of the 1st and 5th columns of Table 3 in Dube et al. (2010). See Section 6 for additional details.

confidence intervals for the pre-trends coefficients in the Border County specification include the point estimates from the All Counties specification, which the authors deemed to be "sizeable." Further, we can conduct power calculations along the lines of those recommended in Section 5.2. If the true coefficients in the Border Counties specification were equal to the point estimates from the All Counties specification, we would fail to find any significant pre-trend coefficient at the 10% level 57% of the time.[12] Likewise, the magnitude of the point estimates for η_{t-4} and the trend $(\eta_{t-4} - \eta_{t-12})$ would be smaller than that deemed sizeable in the All Counties specification 41% of the time. By contrast, if parallel trends held in the Border Counties specification, we would fail to find any significant pre-trend coefficient 81% of the time, and we would fail to find a sizeable coefficient 61% of the time, so the tests for significance and sizeable coefficients respectively have a Bayes Factor of 1.4 and 1.5. The power of the pre-test here is thus similar to that in the example in Figure 4 (a) in which requiring pre-testing exacerbated bias in published work.

Although it is not entirely clear under what conditions the authors would have chosen the Border Counties specification as their main specification, it seems reasonable to suppose that they would not have done so if the point estimates for η_{t-4} and $\eta_{t-4} - \eta_{t-12}$ were of the same magnitude as their estimates in the All Counties specification, which they allude to as "sizeable."

Column 3 of Table 7 provides adjusted (median-unbiased) estimates and CIs that account for this selection, using the adjusted estimators described in Section 5.[13] The adjusted

[12] I use a 10% significance level, since the authors note that the All Counties pre-trends coefficients are significant at the 10% level. We would fail to find any coefficient significant at the 5% level 70% of the time.

[13] In particular, I condition on the event $|\hat{\eta}_{t-4}| < .194$ and $|\hat{\eta}_{t-4} - \hat{\eta}_{t-12}| < .124$, where .194 and .124

point estimates for the pre-trend are of a similar magnitude to those in the All Counties specification, but in the opposite direction, suggesting that employment was growing faster in the counties that raised their minimum wages prior to the policy change. Further, the adjusted point estimates suggest that following a minimum wage change, employment in border counties that raised the minimum wage decreases relative to the four quarters prior to the change (i.e. the estimate for η_{t-4} is larger than η_t).

Importantly, the confidence intervals on the adjusted employment elasticity η_t are also substantially wider than before. The authors directly compare their estimates to those of Neumark and Wascher (2000), whose employment elasticity of -0.22 falls outside of the 95% confidence interval in the authors' preferred specification. However, the Neumark and Wascher estimates fall well within the 95% confidence interval that adjusts for the pre-test for parallel trends. Thus, after adjusting for the test for pre-trends, we are not longer able to rule out previous estimates finding a substantial negative employment effect of the minimum wage.

7 Conclusion

This paper illustrates the importance of accounting for tests for pre-trends, which are common in applied work. I provide both theoretical and empirical evidence that pre-trends testing can lead to undesirable distortions in the distribution of conventional treatment effect estimates, and potentially even amplify bias in published work. I introduce a method for correcting event-study plots for pre-testing or model selection using pre-trends, and I encourage researchers relying on pre-testing to report these corrected estimates along with simple calculations of the power of their pre-test against economically relevant violations of parallel trends.

I conclude by highlighting two potential directions for future research. First, while I have analyzed the implications of researchers pre-testing or choosing among a finite number of models on the basis of pre-trends, the synthetic control method (Abadie et al., 2010) essentially automates the selection of a control group on the basis of pre-trends. Future research might investigate the extent to which the synthetic control method can cause bias amplification similar to that under simple pre-testing when parallel trends is violated. Second, while I have analyzed the current practice of testing for pre-trends, an interesting question for future research is the extent to which the current pre-testing regime could be improved upon, either by modifying the rules used for the pre-test or adopting a different approach to account for possible violations of the parallel trends assumption.

respectively correspond with the estimates of $\hat{\eta}_{t-4}$ and $\hat{\eta}_{t-4} - \hat{\eta}_{t-12}$ from the All Counties regression.

References

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, 72(1):1–19.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505.

Abraham, S. and Sun, L. (2018). Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects. SSRN Scholarly Paper ID 3158747, Social Science Research Network, Rochester, NY.

Andrews, I. (2018). Valid Two-Step Identification-Robust Confidence Sets for GMM. *The Review of Economics and Statistics*, 100(2):337–348.

Andrews, I. and Kasy, M. (2017). Identification of and Correction for Publication Bias. Working Paper 23298, National Bureau of Economic Research.

Andrews, I., Kitagawa, T., and McCloskey, A. (2018). Inference on winners. Technical Report CWP31/18, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Arismendi Zambrano, J. and Broda, S. A. (2016). Multivariate Elliptical Truncated Moments. SSRN Scholarly Paper ID 2841401, Social Science Research Network, Rochester, NY.

Armstrong, T. B. and Kolesár, M. (2018). A Simple Adjustment for Bandwidth Snooping. *The Review of Economic Studies*, 85(2):732–765.

Athey, S. and Imbens, G. (2018). Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption. *arXiv:1808.05293 [cs, econ, math, stat]*.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Bilinski, A. and Hatfield, L. A. (2018). Seeking evidence of absence: Reconsidering tests of model assumptions. *arXiv:1805.03273 [stat]*.

Borusyak, K. and Jaravel, X. (2016). Revisiting Event Study Designs. SSRN Scholarly Paper ID 2826228, Social Science Research Network, Rochester, NY.

Callaway, B. and Sant'Anna, P. H. C. (2019). Difference-in-Differences with Multiple Time Periods. SSRN Scholarly Paper ID 3148250, Social Science Research Network, Rochester, NY.

Cartinhour, J. (1990). One-dimensional marginal density functions of a truncated multivariate normal density function. *Communications in Statistics-theory and Methods - COMMUN STATIST-THEOR METHOD*, 19:197–203.

Chabé-Ferret, S. (2015). Analysis of the bias of Matching and Difference-in-Difference under alternative earnings and selection processes. *Journal of Econometrics*, 185(1):110–123.

Christensen, G. S. and Miguel, E. (2016). Transparency, Reproducibility, and the Credibility of Economics Research. Working Paper 22989, National Bureau of Economic Research.

Daw, J. R. and Hatfield, L. A. (2018). Matching and Regression to the Mean in Difference-in-Differences Analysis. *Health Services Research*.

de Chaisemartin, C. and D'Haultfœuille, X. (2018a). Fuzzy Differences-in-Differences. *The Review of Economic Studies*, 85(2):999–1028.

de Chaisemartin, C. and D'Haultfœuille, X. (2018b). Two-way fixed effects estimators with heterogeneous treatment effects. *arXiv:1803.08807 [econ]*.

Donald, S. G. and Lang, K. (2007). Inference with Difference-in-Differences and Other Panel Data. *The Review of Economics and Statistics*, 89(2):221–233.

Dube, A., Lester, T. W., and Reich, M. (2010). Minimum wage effects across state borders: Estimates using contiguous counties. *The Review of Economics and Statistics*, 92(4):945–964.

Freyaldenhoven, S., Hansen, C., and Shapiro, J. M. (2018). Pre-event Trends in the Panel Event-study Design. Working Paper 24565, National Bureau of Economic Research.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2018). *Mvtnorm: Multivariate Normal and t Distributions*.

Giles, J. A. and Giles, D. E. A. (1993). Pre-Test Estimation and Testing in Econometrics: Recent Developments. *Journal of Economic Surveys*, 7(2):145–197.

Goodman-Bacon, A. (2018). Difference-in-Differences with Variation in Treatment Timing. Working Paper 25018, National Bureau of Economic Research.

He, G. and Wang, S. (2017). Do College Graduates Serving as Village Officials Help Rural China? *American Economic Journal: Applied Economics*, 9(4):186–215.

Kahn-Lang, A. and Lang, K. (2018). The Promise and Pitfalls of Differences-in-Differences: Reflections on '16 and Pregnant' and Other Applications. Working Paper 24857, National Bureau of Economic Research.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Leeb, H. and Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory*, 21(1):21–59.

Manjunath, B. and Wilhelm, S. (2012). Moments Calculation For the Doubly Truncated Multivariate Normal Density. *arXiv:1206.5387 [stat]*.

Neumark, D. and Wascher, W. (2000). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment. *American Economic Review*, 90(5):1362–1396.

Pfanzagl, J. (1994). *Parametric Statistical Theory*. W. de Gruyter. Google-Books-ID: 1S20QgAACAAJ.

Rothstein, H. R., Sutton, A. J., and Borenstein, M. (2005). Publication Bias in Meta-Analysis. In Co-Chair, H. R. R., Co-Author, A. J. S., and PI, M. B. D. A. L., editors, *Publication Bias in Meta-Analysis*, pages 1–7. John Wiley & Sons, Ltd.

Saumard, A. and Wellner, J. A. (2014). Log-concavity and strong log-concavity: A review. *arXiv:1404.5886 [math, stat]*.

Snyder, C. and Zhuo, R. (2018). Sniff Tests in Economics: Aggregate Distribution of Their Probability Values and Implications for Publication Bias. Working Paper 25058, National Bureau of Economic Research.

Tallis, G. M. (1963). Elliptical and Radial Truncation in Normal Populations. *The Annals of Mathematical Statistics*, 34(3):940–944.

A Proofs

Lemma 1. *Let $\tilde{\beta}_{post} = \hat{\beta}_{post} - \Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}$. Then $\tilde{\beta}_{post}$ and $\hat{\beta}_{pre}$ are independent.*

Proof. Note that by assumption, $\hat{\beta}_{post}$ and $\hat{\beta}_{pre}$ are jointly normal. Since $\tilde{\beta}_{post}$ is a linear combination of $\hat{\beta}_{post}$ and $\hat{\beta}_{pre}$, it follows that $\hat{\beta}_{pre}$ and $\tilde{\beta}_{post}$ are jointly normal. It thus suffices to show that $\hat{\beta}_{pre}$ and $\tilde{\beta}_{post}$ are uncorrelated. We have

$$\begin{aligned}
\operatorname{Cov}\left(\hat{\beta}_{pre}\,,\,\tilde{\beta}_{post}\right) &= \mathbb{E}\left[\left(\hat{\beta}_{pre} - \beta_{pre}\right)\left((\hat{\beta}_{post} - \beta_{post}) - \Sigma_{12}\Sigma_{22}^{-1}(\hat{\beta}_{pre} - \beta_{pre})\right)\right] \\
&= \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} \\
&= 0
\end{aligned}$$

which completes the proof.

□

Proof of Proposition 1 Note that by construction, $\hat{\beta}_{post} = \tilde{\beta}_{post} + \Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}$. It follows that

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}_{post} \,|\, \hat{\beta}_{pre} \in B\right] &= \mathbb{E}\left[\tilde{\beta}_{post} \,|\, \hat{\beta}_{pre} \in B\right] + \Sigma_{12}\Sigma_{22}^{-1}\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] \\
&= \mathbb{E}\left[\tilde{\beta}_{post}\right] + \Sigma_{12}\Sigma_{22}^{-1}\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] \\
&= \mathbb{E}\left[\hat{\beta}_{post} - \Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}\right] + \Sigma_{12}\Sigma_{22}^{-1}\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] \\
&= \beta_{post} - \Sigma_{12}\Sigma_{22}^{-1}\beta_{pre} + \Sigma_{12}\Sigma_{22}^{-1}\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] \\
&= \beta_{post} + \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] - \beta_{pre}\right)
\end{aligned}$$

where the second line uses the independence of $\tilde{\beta}_{post}$ and $\hat{\beta}_{pre}$ from Lemma 1, and the third and fourth use the definition of $\tilde{\beta}_{post}$, β_{post}, and β_{pre}. □.

Definition 1 (Symmetric Truncation About 0)**.** *We say that $B \subset \mathbb{R}^K$ is a symmetric truncation around 0 if $\beta \in B$ iff $-\beta \in B$.*

Lemma 2. *Suppose $Y \sim \mathcal{N}(0, \Sigma)$ is a K-dimensional multivariate normal, and B is a symmetric truncation around 0. Then $\mathbb{E}[Y \,|\, Y \in B] = 0$.*

Proof. Since $(-Y) \in B$ iff $Y \in B$, we have

$$
\begin{aligned}
\mathbb{E}\left[Y \mid Y \in B\right] &= \mathbb{E}\left[-Y \mid (-Y) \in B\right] \\
&= \mathbb{E}\left[-Y \mid Y \in B\right] \\
&= -\mathbb{E}\left[Y \mid Y \in B\right]
\end{aligned}
$$

which implies that $\mathbb{E}\left[Y \mid Y \in B\right] = 0$.

□

Proof of Proposition 2 From Proposition 1, it suffices to show that $\mathbb{E}\left[\hat{\beta}_{pre} \mid \hat{\beta}_{pre} \in B\right] - \beta_{pre} = 0$. However, $\beta_{pre} = 0$ by the assumption of parallel trends, and it then follows that $\mathbb{E}\left[\hat{\beta}_{pre} \mid \hat{\beta}_{pre} \in B\right] = 0$ by Lemma 2. □

We now prove a series of Lemmas leading up to the proof of Proposition 3.

Lemma 3. *Suppose Y is a k-dimensional multivariate normal, $Y \sim \mathcal{N}(\mu, \Sigma)$, and let $B \subset \mathbb{R}^k$ be a convex set such that $\mathbb{P}(Y \in B) > 0$. Letting D_μ denote the Jacobian operator with respect to μ, we have*

1. $D_\mu \mathbb{E}\left[Y \mid Y \in B, \mu\right] = \mathbb{V}ar\left[Y \mid Y \in B, \mu\right] \Sigma^{-1}$.
2. $\mathbb{V}ar\left[Y \mid Y \in B\right] - \Sigma$ *is negative semi-definite.*

Proof.[14]

Define the function $H : \mathbb{R}^k \to \mathbb{R}$ by

$$
H(\mu) = \int_B \phi_\Sigma(y - \mu) dy
$$

for $\phi_\Sigma(x) = det(2\pi\Sigma)^{-\frac{1}{2}} exp(-\frac{1}{2} x' \Sigma^{-1} x)$ the PDF of the $\mathcal{N}(0, \Sigma)$ distribution. We now argue that H is log-concave in μ. Note that we can write $H(\mu) = \int_{\mathbb{R}^k} g_1(y, \mu) g_2(y, \mu) dy$ for $g_1(y, \mu) = \phi_\Sigma(y - \mu)$ and $g_2(y, \mu) = 1\left[y \in B\right]$. The normal PDF is log-concave, and g_1 is the composition of the normal PDF with a linear function, and hence log-concave as well. Likewise, g_2 is log-concave since B is a convex set. The product of log-concave functions is log-concave, and the marginalization of a log-concave function with respect to one of its arguments is log-concave by Prekopa's theorem (see, e.g. Theorem 3.3 in Saumard and Wellner (2014)), from which it follows that H is log-concave in μ.

[14] I am grateful to Alecos Papadopolous, whose answer on StackOverflow to a related question inspired this proof.

Now, applying Leibniz's rule, we have that the $1 \times k$ gradient of $\log H$ with respect to μ is equal to

$$\begin{aligned} D_\mu \log H &= \frac{\int_B D_\mu \phi_\Sigma(y-\mu) dy}{\int_B \phi_\Sigma(y-\mu) dy} \\ &= \frac{\int_B \phi_\Sigma(y-\mu)(y-\mu)'\Sigma^{-1} dy}{\int_B \phi_\Sigma(y-\mu) dy} \\ &= (\mathbb{E}\left[Y \mid Y \in B\right] - \mu)'\Sigma^{-1}. \end{aligned}$$

where the second line takes the derivative of the normal PDF, $D_\mu \phi_\Sigma(y-\mu) = \phi_\Sigma(y-\mu) \cdot (y-\mu)'\Sigma^{-1}$, and the third uses the definition of the conditional expectation. It follows that

$$\mathbb{E}\left[Y \mid Y \in B, \mu\right] = \mu + \Sigma(D_\mu \log H)'.$$

Differentiating again with respect to μ, we have that the $k \times k$ Jacobian of $\mathbb{E}\left[Y \mid Y \in B, \mu\right]$ with respect to μ is given by

$$D_\mu \mathbb{E}\left[Y \mid Y \in B, \mu\right] = I + \Sigma D_\mu (D_\mu \log H)'. \tag{8}$$

Since H is log-concave, $D_\mu(D_\mu \log H)'$ is the Hessian of a concave function, and thus is negative semi-definite. Next, note that by definition,

$$\mathbb{E}\left[Y \mid Y \in B, \mu\right] = \frac{\int_B y\, \phi_\Sigma(y-\mu)\, dy}{\int_B \phi_\Sigma(y-\mu)\, dy}.$$

Thus, by the product rule,

$$\begin{aligned} D_\mu \mathbb{E}\left[Y \mid Y \in B, \mu\right] = & \frac{\int_B y\, D_\mu \phi_\Sigma(y-\mu)\, dy}{\int_B \phi_\Sigma(y-\mu)\, dy} + \\ & \left[\int_B y\, \phi_\Sigma(y-\mu)\, dy\right] \cdot D_\mu \left[\int_B \phi_\Sigma(y-\mu)\, dy\right]^{-1}. \end{aligned} \tag{9}$$

Recall that

$$D_\mu \phi_\Sigma(y-\mu) = \phi_\Sigma(y-\mu) \cdot (y-\mu)'\Sigma^{-1}.$$

The first term in (9) thus becomes

$$\frac{\int_B y(y-\mu)'\phi_\Sigma(y-\mu)\,dy}{\int_B \phi_\Sigma(y-\mu)\,dy}\Sigma^{-1} =$$
$$\left(\mathbb{E}\left[YY' \,|\, Y \in B, \mu\right] - \mathbb{E}\left[Y \,|\, Y \in B, \mu\right]\mu'\right)\Sigma^{-1}.$$

Applying the chain-rule, the second term in (9) becomes

$$-\frac{\int_B y\,\phi_\Sigma(y-\mu)\,dy \cdot \int_B (y-\mu)'\,\phi_\Sigma(y-\mu)\,dy}{\left[\int_B \phi_\Sigma(y-\mu)\,dy\right]^2}\Sigma^{-1} =$$
$$\left(-\mathbb{E}\left[Y \,|\, Y \in B, \mu\right]\mathbb{E}\left[Y \,|\, Y \in B, \mu\right]' + \mathbb{E}\left[Y \,|\, Y \in B, \mu\right]\mu'\right)\Sigma^{-1}.$$

Substituting back into (9), we have

$$\begin{aligned} D_\mu \mathbb{E}\left[Y \,|\, Y \in B, \mu\right] &= \left(\mathbb{E}\left[YY' \,|\, Y \in B, \mu\right] - \mathbb{E}\left[Y \,|\, Y \in B, \mu\right]\mathbb{E}\left[Y \,|\, Y \in B, \mu\right]'\right)\Sigma^{-1} \\ &= \mathbb{V}\mathrm{ar}\left[Y \,|\, Y \in B, \mu\right]\Sigma^{-1} \end{aligned} \tag{10}$$

which establishes the first result. Additionally, combining (8) and (10), we have that

$$\mathbb{V}\mathrm{ar}\left[Y \,|\, Y \in B, \mu\right]\Sigma^{-1} = I + \Sigma D_\mu (D_\mu \log H)' \tag{11}$$

which implies that

$$\mathbb{V}\mathrm{ar}\left[Y \,|\, Y \in B, \mu\right] - \Sigma = \Sigma\, D_\mu (D_\mu \log H)'\,\Sigma. \tag{12}$$

Thus, for any vector $x \in \mathbb{R}^k$,

$$\begin{aligned} x'\left(\mathbb{V}\mathrm{ar}\left[Y \,|\, Y \in B, \mu\right] - \Sigma\right)x &= x'\left(\Sigma\, D_\mu (D_\mu \log H)'\,\Sigma\right)x \\ &= (\Sigma x)'\,\left(D_\mu (D_\mu \log H)'\right)\,(\Sigma x) \\ &\leq 0 \end{aligned}$$

where the inequality follows from the fact that $D_\mu (D_\mu \log H)'$ is negative semi-definite. Since

$\mathbb{V}\text{ar}\left[Y \mid Y \in B, \mu\right] - \Sigma$ is symmetric, it follows that it is negative semi-definite, as we desired to show. $\square$

Lemma 4. *Suppose that Σ satisfies Assumption 1. Then for ι the vector of ones and some $c_1 > 0$, $\iota'\Sigma_{22}^{-1} = c_1\iota'$. Additionally, $\Sigma_{12}\Sigma_{22}^{-1} = c_2\iota'$, for a constant $c_2 > 0$.*

Proof. First, note that if $K = 1$, then Σ_{12} and Σ_{22} are each positive scalars, and the result follows trivially. For the remainder of the proof, we therefore consider $K > 1$. Note that we can write $\Sigma_{22} = \Lambda + \rho\iota\iota'$, where $\Lambda = (\sigma^2 - \rho)I$. It follows from the Sherman-Morrison formula that:

$$\begin{aligned}\Sigma_{22}^{-1} &= \Lambda^{-1} - \frac{\rho^2\Lambda^{-1}\iota\iota'\Lambda^{-1}}{1 + \rho^2\iota'\Lambda^{-1}\iota} \\ &= (\sigma^2 - \rho)^{-1}I - \frac{\rho^2(\sigma^2 - \rho)^{-2}\iota\iota'}{1 + \rho^2(\sigma^2 - \rho)^{-1}\iota'\iota}.\end{aligned}$$

Thus:

$$\begin{aligned}&\iota'\Sigma_{22}^{-1} = \\ &\iota'\left((\sigma^2 - \rho)^{-1}I - \frac{\rho^2(\sigma^2 - \rho)^{-2}\iota\iota'}{1 + \rho^2(\sigma^2 - \rho)^{-1}\iota'\iota}\right) = \\ &(\sigma^2 - \rho)^{-1}\left(1 - \frac{\rho^2(\sigma^2 - \rho)^{-1}\iota'\iota}{1 + \rho^2(\sigma^2 - \rho)^{-1}\iota'\iota}\right)\iota' = \\ &\underbrace{(\sigma^2 - \rho)^{-1}\left(\frac{1}{1 + \rho^2(\sigma^2 - \rho)^{-1}\iota'\iota}\right)}_{:=c_1}\iota'.\end{aligned}$$

Since $\sigma^2 - \rho > 0$, all of the terms in c_1 are positive, and thus $c_1 > 0$, as needed. Finally, note that Assumption 1 implies that $\Sigma_{12} = \rho\iota'$. It follows that $\Sigma_{12}\Sigma_{22}^{-1} = \rho c_1\iota' = c_2\iota'$ for $c_2 = \rho c_1 > 0$.

$\square$

Lemma 5. *Suppose $Y \sim N(0, \Sigma)$ is K-dimensional normal, with Σ satisfying the requirements on Σ_{22} imposed by Assumption 1. Let $B = \{y \in \mathbb{R}^K \mid a_j \leq y \leq b_j$ for all $j\}$, where $-b_j < a_j < b_j$ for all j. Then for ι the vector of ones, $\mathbb{E}\left[\iota'Y \mid Y \in B\right] = \mathbb{E}\left[Y_1 + \ldots + Y_k \mid Y \in B\right]$ is elementwise greater than 0.*

Proof. For any $x \in \mathbb{R}^K$ such that $x_j \leq b_j$ for all j, define $B^X(x) = \{y \in \mathbb{R}^K \mid x_j \leq y \leq b_j$ for all $j\}$. Let $b = (b_1, \ldots, b_K)$. Note that $B^X(-b)$ is a symmetric rectangular truncation around 0, so from Lemma 2, we have that $\mathbb{E}\left[Y \mid Y \in B^X(-b)\right] = 0$. Now, define

$$g(x) = \mathbb{E}\left[\iota' Y \,|\, Y \in B^X(x)\right].$$

From the argument above, we have that $g(-b) = 0$, and we wish to show that $g(a) > 0$. Note that by the mean-value theorem, for some $t \in (0,1)$,

$$\begin{aligned} g(a) &= g(-b) + (a - (-b))\, \nabla g\left(ta + (1-t)(-b)\right) \\ &= (a+b)\nabla g\left(ta + (1-t)(-b)\right) \\ &=: (a+b)\nabla g(x^t). \end{aligned}$$

By assumption, $(a+b)$ is elementwise greater than 0. It thus suffices to show that all elements of $\nabla g\left(x^t\right)$ are positive. WLOG, we show that $\dfrac{\partial g(x^t)}{\partial x_K} > 0$.

Using the definition of the conditional expectation and Leibniz's rule, we have

$$\begin{aligned} &\frac{\partial g(x^t)}{\partial x_K} = \\ &\frac{\partial}{\partial x_K^t}\left[\left(\int_{x_1^t}^{b_1} \cdots \int_{x_K^t}^{b_K} (y_1 + \ldots + y_K)\, \phi_\Sigma(y)\, dy_1 \ldots dy_K\right)\left(\int_{x_1^t}^{b_1} \cdots \int_{x_K^t}^{b_K} \phi_\Sigma(y)\, dy_1 \ldots dy_K\right)^{-1}\right] = \\ &\left(\int_{x_1^t}^{b_1} \cdots \int_{x_K^t}^{b_K} (y_1 + \ldots + y_K)\, \phi_\Sigma(y)\, dy_1 \ldots dy_K \times \int_{x_1^t}^{b_1} \cdots \int_{x_{K-1}^t}^{b_{K-1}} \phi_\Sigma\left(\begin{pmatrix} y_{-K} \\ x_K^t \end{pmatrix}\right) dy_1 \ldots dy_{K-1}\right. \\ &\left. - \int_{x_1^t}^{b_1} \cdots \int_{x_{K-1}^t}^{b_{K-1}} (y_1 + \ldots + y_{K-1} + x_K^t)\, \phi_\Sigma\left(\begin{pmatrix} y_{-K} \\ x_K^t \end{pmatrix}\right) dy_1 \ldots dy_{K-1} \times \int_{x_1^t}^{b_1} \cdots \int_{x_K^t}^{b_K} \phi_\Sigma(y)\, dy_1 \ldots dy_K\right) \\ &\times \left(\int_{x_1^t}^{b_1} \cdots \int_{x_K^t}^{b_K} \phi_\Sigma(y)\, dy_1 \ldots dy_K\right)^{-2} \end{aligned} \tag{13}$$

where $\phi_\Sigma(y)$ denotes the PDF of a multivariate normal with mean 0 and variance Σ. It follows from (13) that $\dfrac{\partial g(x^t)}{\partial x_K} > 0$ if and only if

$$\frac{\int_{x_1^t}^{b_1} \cdots \int_{x_k^t}^{b_K} (y_1 + \ldots + y_K)\ \phi_\Sigma(y)\ dy_1 \ldots dy_K}{\int_{x_1^t}^{b_1} \cdots \int_{x_k^t}^{b_K}\ \phi_\Sigma(y)\ dy_1 \ldots dy_K} >$$

$$\frac{\int_{x_1^t}^{b_1} \cdots \int_{x_{K-1}^t}^{b_{K-1}} (y_1 + \cdots + y_{K-1} + x_K^t)\ \phi_\Sigma\left(\begin{pmatrix} y_{-K} \\ x_K^t \end{pmatrix}\right)\ dy_1 \ldots dy_{K-1}}{\int_{x_1^t}^{b_1} \cdots \int_{x_{K-1}^t}^{b_{K-1}}\ \phi_\Sigma\left(\begin{pmatrix} y_{-K} \\ x_K^t \end{pmatrix}\right)\ dy_1 \ldots dy_{K-1}}$$

or equivalently,

$$\mathbb{E}\left[Y_1 + \ldots + Y_K \,|\, x_j^t \leq Y_j \leq b_j, \forall j\right] > \mathbb{E}\left[Y_1 + \ldots + Y_K \,|\, x_j^t \leq Y_j \leq b_j,\ \text{for } j < K,\, Y_K = x_K^t\right].$$

It is clear that $\mathbb{E}\left[Y_K \,|\, x_j^t \leq Y_j \leq b_j, \forall j\right] > x_K^t$, since $x_K^t < b_K$ and the Kth marginal density of the rectangularly-truncated normal distribution is positive for all values in $[x_K^t, b_K]$ (see Cartinhour (1990)). This completes the proof for the case where $K = 1$. For $K > 1$, it suffices to show that

$$\mathbb{E}\left[Y_1 + \ldots + Y_{K-1} \,|\, x_j^t \leq Y_j \leq b_j, \forall j\right] \geq \mathbb{E}\left[Y_1 + \ldots + Y_{K-1} \,|\, x_j^t \leq Y_j \leq b_j,\ \text{for } j < K,\, Y_K = x_K^t\right]. \tag{14}$$

To see why (14) holds, let $\tilde{Y}_{-K} = Y_{-K} - \Sigma_{-K,K}\Sigma_{K,K}^{-1} Y_K$, where a "$-K$" subscript denotes all of the indices except for K. By an argument analogous to that in the Proof of Lemma 1 for $\tilde{\beta}_{post}$, one can easily verify that $\tilde{Y}_{-K}$ is independent of Y_K and $\tilde{Y}_{-K} \sim \mathcal{N}\left(0,\, \tilde{\Sigma}\right)$ for $\tilde{\Sigma} = \Sigma_{-K,-K} - \Sigma_{-K,K}\Sigma_{K,K}^{-1}\Sigma_{K,-K}$. By construction, $Y_{-K} = \tilde{Y}_{-K} + \Sigma_{-K,K}\Sigma_{K,K}^{-1} Y_K$, from which it follows that

$$Y_{-K} \,|\, Y_K = y_k \sim \mathcal{N}\left(\Sigma_{-K,K}\Sigma_{K,K}^{-1} y_K,\, \tilde{\Sigma}\right).$$

We now argue that $\Sigma_{-K,K}\Sigma_{K,K}^{-1} y_K = c\, y_k\, \iota$ for a positive constant c. If $K = 2$, then by Assumption 1, $\Sigma_{-K,K}\Sigma_{K,K}^{-1}$ is the product of two positive scalars, and can thus be trivially written as $c\iota$. For $K > 2$, we verify that $\tilde{\Sigma}$ meets the requirements that Assumption 1 places on Σ_{22}. To do this, note by Assumption 1, Σ has common terms σ^2 on the diagonal and ρ on the off-diagonal, and thus the same holds for $\Sigma_{-K,-K}$. Additionally, under Assumption 1,

$\Sigma_{-K,K} = \rho\iota$ and $\Sigma_{K,K}^{-1} = \frac{1}{\sigma^2}$, so $\Sigma_{-K,K}\Sigma_{K,K}^{-1}\Sigma_{K,-K}$ equals ρ^2/σ^2 times $\iota\iota'$, the matrix of ones. The diagonal terms of $\tilde{\Sigma}$ are thus equal to $\tilde{\sigma}^2 = \sigma^2 - \rho^2/\sigma^2$, and the off-diagonal terms are equal to $\tilde{\rho} = \rho - \rho^2/\sigma^2$, or equivalently $\tilde{\rho} = \rho(1 - \rho/\sigma^2)$. Since by Assumption 1, $0 < \rho < \sigma^2$, it is clear that $\tilde{\sigma}^2 > \tilde{\rho}$. Additionally, $0 < \rho < \sigma^2$ implies that $1 - \rho/\sigma^2 > 0$, and hence $\tilde{\rho} > 0$, which completes the proof that $\tilde{\Sigma}$ satisfies the requirements of Assumption 1 for Σ_{22}. Hence, $\Sigma_{-K,K}\Sigma_{K,K}^{-1} y_K = c\, y_k\, \iota$ by Lemma 4. We can therefore write

$$Y_{-K} \,|\, Y_K = y_k \sim \mathcal{N}\left(c\, y_k\, \iota,\, \tilde{\Sigma}\right).$$

Let $h(\mu) = \mathbb{E}\left[X | X \in B_{-K},\, X \sim \mathcal{N}\left(\mu,\, \tilde{\Sigma}\right)\right]$ for $B_{-K} = \{\tilde{x} \in \mathbb{R}^{K-1} | x_j^t \leq \tilde{x}_j \leq b_j,\ \text{for } j = 1, \ldots, K-1\}$. Then $\mathbb{E}\left[\iota' Y_{-K} \,|\, x_j^t \leq Y_j \leq b_j,\ \text{for } j < K,\, Y_K = y_k\right] = \iota' h(c y_k \iota)$. Hence,

$$\begin{aligned}
\frac{\partial}{\partial y_k}\mathbb{E}\left[\iota' Y_{-K} \,|\, x_j^t \leq Y_j \leq b_j,\ \text{for } j < K,\, Y_K = y_k\right] &= \iota' \left(D_\mu h|_{\mu = c y_k \iota}\right) \iota \cdot c \\
&= \iota' \mathbb{V}\text{ar}\left[Y_{-K} \,|\, Y_{-K} \in B_{-K}, Y_K = y_k\right] \tilde{\Sigma}^{-1} \iota c \\
&= \iota' \mathbb{V}\text{ar}\left[Y_{-K} \,|\, Y_{-K} \in B_{-K}, Y_K = y_k\right] \iota c_1 c \\
&\geq 0
\end{aligned}$$

where the second line follows from Lemma 3; the third line uses Lemma 4 to obtain that $\tilde{\Sigma}^{-1}\iota = \iota c_1$ for $c_1 > 0$ (if $K = 2$, this holds trivially); and the inequality follows from the fact that $\mathbb{V}\text{ar}\left[Y_{-K} \,|\, Y_{-K} \in B_{-K}, Y_K = y_k\right]$ is positive semi-definite and c_1 and c are positive by construction. Thus, for all $y_k \in [x_k^t, b_k]$,

$$\begin{aligned}
&\mathbb{E}\left[Y_1 + \ldots + Y_{K-1} \,|\, x_j^t \leq Y_j \leq b_j \text{ for } j < K,\, Y_K = y_k\right] \geq \\
&\mathbb{E}\left[Y_1 + \ldots + Y_{K-1} \,|\, x_j^t \leq Y_j \leq b_j \text{ for } j < K,\, Y_K = x_k^t\right].
\end{aligned}$$

By the law of iterated expectations, we have

$$\begin{aligned}
&\mathbb{E}\left[Y_1 + \ldots + Y_{K-1} \,|\, x_j^t \leq Y_j \leq b_j, \forall j\right] = \\
&\mathbb{E}\left[\mathbb{E}\left[Y_1 + \ldots + Y_{K-1} \,|\, x_j^t \leq Y_j \leq b_j \text{ for } j < K, Y_K\right] \,|\, x_j^t \leq Y_j \leq b_j, \forall j\right] \geq \\
&\mathbb{E}\left[\mathbb{E}\left[Y_1 + \ldots + Y_{K-1} \,|\, x_j^t \leq Y_j \leq b_j \text{ for } j < K, Y_K = x_K^t\right] \,|\, x_j^t \leq Y_j \leq b_j, \forall j\right] = \\
&\mathbb{E}\left[Y_1 + \ldots + Y_{K-1} \,|\, x_j^t \leq Y_j \leq b_j \text{ for } j < K, Y_K = x_K^t\right]
\end{aligned}$$

as we wished to show.

□

Proof of Proposition 3 From Proposition 1, the desired result is equivalent to showing that

$$\Sigma_{12}\Sigma_{22}^{-1}\,\mathbb{E}\left[\hat{\beta}_{pre}-\beta_{pre}\,|\,\hat{\beta}_{pre}\in B\right]>0.$$

By Lemma 4, Assumption 1 implies that $\Sigma_{12}\Sigma_{22}^{-1}=c_1\iota'$ for $c_1>0$, so it suffices to show that $\iota'\mathbb{E}\left[\hat{\beta}_{pre}-\beta_{pre}\,|\,\hat{\beta}_{pre}\in B\right]$ is elementwise greater than zero. Note that by assumption $(\hat{\beta}_{pre}-\beta_{pre})\sim\mathcal{N}(0,\,\Sigma_{22})$. Additionally, $\hat{\beta}_{pre}\in B_{NIS}=\{\hat{\beta}_{pre}:|\hat{\beta}_{pre,j}|/\sqrt{\Sigma_{jj}}\leq c_\alpha$ for all $j\}$ iff $(\hat{\beta}_{pre}-\beta_{pre})\in\tilde{B}_{NIS}=\{\beta:a_j\leq\beta_j\leq b_j\}$ for $a_j=-c_\alpha\sqrt{\Sigma_{jj}}-\beta_{pre,j}$ and $b_j=c_\alpha\sqrt{\Sigma_{jj}}-\beta_{pre,j}$. Since $\beta_{pre,j}<0$ for all j, we have that $-b_j<a_j<b_j$ for all j. The result then follows immediately from Lemma 5.

Proof of Proposition 4 Note that since $\hat{\beta}_{post}=\tilde{\beta}_{post}+\Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}$, for any set S,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\hat{\beta}_{post}\,|\,\hat{\beta}_{pre}\in S\right]&=\mathbb{V}\mathrm{ar}\left[\tilde{\beta}_{post}\,|\,\hat{\beta}_{pre}\in S\right]+\mathbb{V}\mathrm{ar}\left[\Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}\,|\,\hat{\beta}_{pre}\in S\right]\\&\quad+2\,\mathrm{Cov}\left(\tilde{\beta}_{post}\,,\,\Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}\,|\,\hat{\beta}_{pre}\in S\right)\\&=\mathbb{V}\mathrm{ar}\left[\tilde{\beta}_{post}\right]+\mathbb{V}\mathrm{ar}\left[\Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}\,|\,\hat{\beta}_{pre}\in S\right],\end{aligned}\tag{15}$$

where we use the independence of $\tilde{\beta}_{post}$ and $\hat{\beta}_{pre}$ from Lemma 1 to obtain that $\mathbb{V}\mathrm{ar}\left[\tilde{\beta}_{post}\,|\,\hat{\beta}_{pre}\in B\right]=\mathbb{V}\mathrm{ar}\left[\tilde{\beta}_{post}\right]$ and that the covariance term equals 0. Applying equation (15) for $S=B$ and for $S=\mathbb{R}^K$, and then taking the difference between the two, we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\hat{\beta}_{post}|\hat{\beta}_{pre}\in B\right]-\mathbb{V}\mathrm{ar}\left[\hat{\beta}_{post}\right]&=\mathbb{V}\mathrm{ar}\left[\Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}\,|\,\hat{\beta}_{pre}\in B\right]-\mathbb{V}\mathrm{ar}\left[\Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}\right]\\&=(\Sigma_{12}\Sigma_{22}^{-1})\left(\mathbb{V}\mathrm{ar}\left[\hat{\beta}_{pre}\,|\,\hat{\beta}_{pre}\in B\right]-\mathbb{V}\mathrm{ar}\left[\hat{\beta}_{pre}\right]\right)(\Sigma_{12}\Sigma_{22}^{-1})',\end{aligned}$$

which gives the desired result.

Proof of Proposition 5 By Proposition 4, it suffices to show that

$$(\Sigma_{12}\Sigma_{22}^{-1})\left(\mathbb{V}\mathrm{ar}\left[\hat{\beta}_{pre}\,|\,\hat{\beta}_{pre}\in B\right]-\mathbb{V}\mathrm{ar}\left[\hat{\beta}_{pre}\right]\right)(\Sigma_{12}\Sigma_{22}^{-1})'\leq 0.$$

The result then follows immediately from the fact that $\mathbb{V}\mathrm{ar}\left[\hat{\beta}_{pre}\,|\,\hat{\beta}_{pre}\in B\right]-\mathbb{V}\mathrm{ar}\left[\hat{\beta}_{pre}\right]$ is negative semi-definite by Lemma 3. □

Proof of Proposition 6

Proof. The proof follows the logic of Theorem 5.2 in Lee et al. (2016) in Lee, who consider the case where B is a polydedron.

Note that by construction, $\eta'\hat{\beta}$ and Z are jointly normal and uncorrelated, hence independent. Thus, without conditioning on $\hat{\beta} \in B$, we have

$$\eta'\hat{\beta} \,|\, Z = z \sim \mathcal{N}\left(\eta'\beta,\, \eta'\Sigma\eta\right).$$

Conditioning on further on $\hat{\beta} \in B$ implies that $\eta'\hat{\beta} \in \Xi(z)$, but owing to the (unconditional) independence of Z and $\eta'\hat{\beta}$, provides no additional information about $\eta'\hat{\beta}$. It follows that

$$\eta'\hat{\beta} \,|\, \hat{\beta} \in B, Z = z \sim \xi \,|\, \xi \in \Xi(z),$$

for $\xi \sim \mathcal{N}\left(\eta'\beta,\, \eta'\Sigma\eta\right)$, and $\Xi(z) := \{x : \exists \hat{\beta} \in B \text{ s.t. } x = \eta'\hat{\beta} \text{ and } z = (I - c\eta')\hat{\beta}\}$. □

Proof of Proposition 7

Proof. The result follows immediately from Theorem 5 of the supplement to Andrews and Kasy (2017). □

Proof of Proposition 8

Proof. The form for $\Xi(z)$ follows immediately from Lemma 5.1 in Lee et al. (2016).

We now verify that the distribution of $\eta'\hat{\beta} \,|\, Z, A\hat{\beta} \leq b$ is continuous for almost every Z. Note that by Proposition 6, $\eta'\hat{\beta} \,|\, Z = z, A\hat{\beta} \leq b$ is truncated normal with truncation points $V^-(z)$ and $V^+(z)$, and hence, continuous if $V^-(z) < V^+(z)$. Since conditional on $A\hat{\beta} \leq b$ and $Z = z$, $V^-(z) \leq \eta'\hat{\beta} \leq V^+(z)$, we have $V^-(z) = V^+(z)$ only if $V^-(z) = \eta'\hat{\beta}$.

It thus suffices to show that $\mathbb{P}\left(\eta'\hat{\beta} = V^-(Z) \,|\, A\hat{\beta} \leq b\right) = 0$. Note though that

$$\begin{aligned}\mathbb{P}\left(\eta'\hat{\beta} = V^-(Z)\right) &= \mathbb{E}\left[\mathbb{P}\left(\eta'\hat{\beta} = V^-(z) \,|\, Z = z\right)\right] \\ &= 0\end{aligned}$$

where for any fixed value z, $\mathbb{P}\left(\eta'\hat{\beta} = V^-(z) \,|\, Z = z\right) = 0$ since $\eta'\hat{\beta}$ and Z are independent by construction and the distribution of $\eta'\hat{\beta}$ is continuous since $\hat{\beta}$ is normally distributed, Σ is full rank, and $\eta \neq 0$. It follows that

$$
\begin{aligned}
0 &= \mathbb{P}\left(\eta'\hat{\beta} = V^-(Z)\right) \\
&= \mathbb{P}\left(\eta'\hat{\beta} = V^-(Z) \,|\, A\hat{\beta} \leq b\right) \mathbb{P}\left(A\hat{\beta} \leq b\right) + \mathbb{P}\left(\eta'\hat{\beta} = V^-(Z) \,|\, A\hat{\beta} \not\leq b\right) \mathbb{P}\left(A\hat{\beta} \not\leq b\right) \\
&\geq \mathbb{P}\left(\eta'\hat{\beta} = V^-(Z) \,|\, A\hat{\beta} \leq b\right) \mathbb{P}\left(A\hat{\beta} \leq b\right).
\end{aligned}
$$

Since $\mathbb{P}\left(A\hat{\beta} \leq b\right) > 0$ by assumption, it follows that $\mathbb{P}\left(\eta'\hat{\beta} = V^-(Z) \,|\, A\hat{\beta} \leq b\right) = 0$, as needed. □

Proof of Proposition 9 Note that by $\hat{\beta} \in B$ iff $\hat{\beta}'A\hat{\beta} - b \leq 0$. Further, by construction $\hat{\beta} = z + c\eta'\hat{\beta}$, so

$$
\begin{aligned}
\hat{\beta}'A\hat{\beta} - b &= \left(z + c\eta'\hat{\beta}\right)' A \left(z + c\eta'\hat{\beta}\right) - b \\
&= \underbrace{(c'Ac)}_{:=\mathcal{A}}(\eta'\hat{\beta})^2 + \underbrace{2c'Az}_{:=\mathcal{B}}(\eta'\hat{\beta}) + \underbrace{(z'Az - b)}_{:=\mathcal{C}},
\end{aligned}
$$

which is a quadratic in $(\eta'\hat{\beta})$. The first part of the result then follows from the quadratic formula.

To verify the conditions for optimality, note that the first part of the result implies that $\Xi(Z)$ is the finite union of intervals on the real line. (We can safely ignore the situations in which $\Xi(Z) = \emptyset$, since conditional on $\hat{\beta} \in B$, $\Xi(Z)$ is non-empty with probability 1). Since $\eta'\hat{\beta} \,|\, \hat{\beta} \in B, Z$ is truncated normal, it will be continuous unless $\Xi(z)$ collapses to a set of measure 0. However, examining the possible cases, we see that this could only occur if $\mathcal{A} > 0, \mathcal{D} \geq 0$ and the interval collapses to a point. By the same argument as in the proof to Proposition 8, this occurs with probability zero, which completes the proof. □

B Calculating $\Xi(z)$ after model selection

Section 5.1.2 showed that Proposition 7 could be applied to corrected event-study plot estimates after selection among a finite number of models. In this Appendix, I discuss how to compute $\Xi(z)$ in this scenario.

The form of Ξ will of course depend on the criteria for the specification search, but I note that a wide variety of specification searches will generate a Ξ that is the union of intervals in $\mathbb{R}$. To see why this is the case, note first that from the definition of $\Xi(z)$, it follows easily

that if $B = B_1 \cup B_2$, then $\Xi_B(z) = \Xi_{B_1}(z) \cup \Xi_{B_2}(z)$, and likewise, if $B = B_1 \cap B_2$, then $\Xi_B(z) = \Xi_{B_1}(z) \cap \Xi_{B_2}(z)$. $\Xi_B(z)$ will therefore take the form of a union of intervals if the conditioning set B can be written as the union and intersection of a sequence of conditioning sets that themselves produce intervals for $\Xi(z)$.[15] Further, Propositions 8 and 9 show that when conditioning on linear or quadratic restrictions on $\hat{\beta}$, $\Xi(z)$ is the union of intervals. Note also that the norm of $\hat{\beta}^m_{pre}$ is less than that of $\hat{\beta}^{m'}_{pre}$ if and only if $\hat{\beta}(A_m - A_{m'}) \leq 0$ for A_m the matrix with 1s on the diagonal in the positions corresponding with the elements of $\hat{\beta}^m_{pre}$ and zero otherwise.

Thus, any selection rule that depends on logical combinations of the (individual or joint) significance of the pre-trends coefficients from each model and/or the relative magnitudes of the models will generate a $\Xi(z)$ that is the union of intervals. A few examples are of note. First, suppose the researcher considers models sequentially and stops at the first model that has an insignificant pre-trend (either jointly, or based on the significance of each of the individual coefficients). Then if the mth model is chosen, $\Xi(z)$ is the intersection of the sets on which models $1, ..., m-1$ have a significant pre-trend, intersected with the set on which model m does not have a significant pre-trend. Second, suppose the researcher chooses the model that minimizes the norm of the pre-period coefficients. Then $\Xi(z)$ is the intersection of the sets on which the chosen model m^* has a lower norm than model m' for each candidate m'. Finally, suppose that the researcher first tests model 1 on the full population, and then if it has a significant pre-trend, chooses whichever has the smaller pre-trend among models 2 and 3, which each restrict to different subsets of the population. Then $\Xi(z)$ for the event model 2 is selected will correspond with the union of intervals on which model 1 is significant intersected with the interval(s) corresponding with the event that the norm of model 2 is less than that of model 3.

[15] Note that the complement of a collection of intervals is also a collection of intervals, and the intersection of collections of intervals can therefore be re-cast as a union of intervals using DeMorgan's laws.