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

Jonathan Roth*

May 8, 2020

Abstract

Tests for pre-existing trends ("pre-trends") are a common way of assessing the plausibility of the parallel trends assumption in difference-in-differences and related research designs. This paper highlights some important limitations of pre-trends testing. From a theoretical perspective, I analyze the distribution of conventional estimates and confidence intervals conditional on surviving a pre-test for pre-trends. I show that in non-pathological cases, the bias of conventional estimates conditional on passing a pre-test can be worse than the unconditional bias. Thus, pre-tests meant to mitigate bias and coverage issues in published work can in fact exacerbate them. I empirically investigate the practical relevance of these concerns in simulations based on a systematic review of recent papers in leading economics journals. I find that conventional pre-tests are often underpowered against plausible violations of parallel trends that produce bias of a similar magnitude as the estimated treatment effect. Distortions from pre-testing can also be substantial. Finally, I discuss alternative approaches that can improve upon the standard practice of relying on pre-trends testing.

1 Introduction

When using difference-in-differences or related research designs, researchers commonly test for pre-treatment differences in trends ("pre-trends") between the treated and control units

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

as a way of assessing the plausibility of the parallel trends assumption. These tests are remarkably common: based on my review, over 70 recent papers in the journals of the American Economic Association have employed an "event-study plot" to visually test for pre-trends.

This paper provides theoretical and empirical evidence on the limitations of pre-trends testing. From a theoretical perspective, I analyze the distribution of conventional estimates and confidence intervals (CIs) after surviving a test for pre-trends. In non-pathological cases, bias and coverage rates of conventional estimates and CIs can be worse conditional on passing the pre-test. Empirically, I conduct a systematic review of recent papers in leading economics journals that test for pre-trends. Simulations based on these papers suggest that the theoretical concerns about pre-trends testing are relevant in practice: the power of conventional pre-tests is often low, and substantial distortions from pre-testing are possible. Finally, I discuss alternative approaches that can improve upon the standard practice of pre-trends testing.

I begin in Section 2 with a stylized model to illustrate the intuition for the limitations of pre-trends testing. I consider a difference-in-differences setting with three periods in which there are normal, homoskedastic errors and potentially linear violations of parallel trends. Even when parallel trends is violated, some draws of the data nonetheless "survive" the test for pre-trends. Moreover, these surviving draws have bias that is worse than would be expected based on the difference in trends alone. The intuition for this is that noise in the data that masks the pre-existing trend also exacerbates bias in the treatment effect estimate because of a mean-reversion effect. An implication of this result is that publication rules that require insignificant pre-trends may or may not reduce bias in published work. Whether they do so will depend on the latent distribution of biases in studies that researchers consider for publication, as well as the power of the pre-test against relevant alternatives. Similar results apply to the coverage rates of conventional CIs.

Section 3 provides a more general theoretical treatment of the distribution of event-study estimates after surviving a pre-test for pre-trends.[1] I derive formulas for the bias and variance of conventional estimates after pre-testing. In general, the bias after surviving a pre-test can be larger or smaller than the unconditional bias. I prove, however, that the bias after pre-testing is necessarily larger in settings with homoskedastic errors and monotone differences in trends. Thus, pre-testing can exacerbate bias in non-pathological cases. I also show under quite general conditions that the variance of conventional estimates is lower conditional on

[1]Throughout the paper, I use the phrase "event-study" to refer to a large class of specifications that estimate dynamic treatment effects along with placebo pre-treatment effects. This includes, but is not limited to, settings with staggered treatment timing; see Related Literature and Remark 2.

passing the pre-test. As a result, traditional CIs will tend to over-cover conditional on passing the pre-test when bias is small, but will generally under-cover as the bias grows larger.

Section 4 evaluates the practical relevance of these theoretical concerns in data-generating processes calibrated 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., 2019; 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 violations against which conventional tests have power of only 50 percent would produce bias of a magnitude similar to the estimated treatment effect. Under such violations of parallel trends, conventional 95% CIs fail to include the average post-treatment effect approximately half of the time in the median paper. Although homoskedasticity does not typically hold in practice, the bias conditional on failing to detect an underlying trend is nonetheless 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.

Finally, I consider different approaches for improving upon the current practice of relying on pre-trends testing in Section 5. I first consider parametric approaches, which extrapolate the pre-treatment difference in trends to the post-treatment periods via a functional form assumption. The advantage of these approaches is that they give valid causal estimates and CIs without pre-testing, provided that the functional form assumption is correct. Although off-the-shelf parametric approaches are not valid conditional on passing a pre-test, I show in Appendix B that these parametric approaches can be combined with corrections for publication bias developed in Andrews and Kasy (2019) to obtain valid estimation and inference following a pre-testing step. This modified parametric approach can be used for retrospective analysis of studies that have been selected on the basis of pre-trends testing.

Unfortunately, however, researchers are often unsure about the correct functional form for the differential trend (Wolfers, 2006; Lee and Solon, 2011). I therefore next turn my attention to two methods that relax the exact parallel trends assumption in different ways. One approach is that of Freyaldenhoven et al. (2019), who allow for parallel trends to be violated so long as there is an excluded covariate that is assumed to be affected by the same confounding factors as the outcome but unaffected by the treatment. A second approach is that of Rambachan and Roth (2019), who provide methods for valid causal inference under assumptions about how informative the true pre-existing trends are about the counterfactual

post-treatment differences in trends. They formalize these assumptions via restrictions on the smoothness of the possible violations of parallel trends, and recommend conducting sensitivity analysis with respect to these assumptions. Both of these approaches enable one to obtain valid causal inference over certain classes of violations of parallel trends without conducting pre-tests, thus avoiding the issues with pre-testing discussed in this paper.

Each of these three alternative approaches can be viewed as imposing certain assumptions about the possible ways in which the parallel trends assumption may be violated. Which method to use thus depends on what assumptions are reasonable in a particular economic context; I provide recommendations for choosing between the methods in Section 5.3. Regardless of the context, I encourage researchers to be explicit about their assumptions about how parallel trends may be violated, and to subject these assumptions to scrutiny given economic knowledge. Incorporating economic knowledge, rather than relying on the statistical significance of pre-trends tests, will improve the credibility and clarity of science in difference-in-differences and related research designs.

Related Literature. This paper relates to a large literature in econometrics and statistics showing that problems can arise in a variety of contexts 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 pre-testing for weak identification (Andrews, 2018), choosing between OLS and IV specifications on the basis of a pre-test (Guggenberger, 2010), using data-driven tuning parameters (Armstrong and Kolesár, 2018), and model selection in high-dimensional settings (Belloni et al., 2014; Farrell, 2015; Belloni et al., 2017). I show theoretically and empirically that similar issues arise with the common practice of testing for pre-trends in difference-in-differences and related research designs.

This paper also contributes to a large body of work on the econometrics of difference-in-differences and related research designs in particular. A topic of substantial recent interest has been the failure of standard two-way fixed effect models to recover a sensible causal estimand in settings with staggered treatment timing and heterogenous treatment effects, even under a suitable parallel trends assumption (Borusyak and Jaravel, 2016; Abraham and Sun, 2018; Athey and Imbens, 2018; de Chaisemartin and D'Haultfœuille, 2018; Goodman-Bacon, 2019; Callaway and Sant'Anna, 2019). For expositional clarity, the main theoretical focus of this paper is on the failures of pre-trends testing in the simpler setting with non-staggered treatment timing or homogenous treatment effects, although the concerns raised apply to the staggered case with heterogeneity as well; see Remark 2 for further discussion. Most closely related to the current paper, recent papers by Freyaldenhoven et al. (2019),

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 show that pre-testing need not necessarily reduce bias and coverage issues from violations of parallel trends.[2] Second, I provide the first systematic empirical evaluation of the power of pre-trends tests and the distortions from pre-trends testing in published work. Finally, I discuss alternatives to pre-testing and provide concrete recommendations for researchers.

Lastly, this paper relates to the literature on selective publication of scientific results (Rothstein et al. (2005) and Christensen and Miguel (2016) provide reviews). 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 theoretical and empirical results on the limitations of these tests in reducing bias and coverage issues. I also build on work by Andrews and Kasy (2019) on correcting for publication bias, as well as earlier results from Lee et al. (2016) and Pfanzagl (1994), to develop corrections to standard parametric methods that have good properties conditional on passing a pre-test for pre-trends.

2 Stylized Model

This section develops intuition for the limitations of pre-trends testing in a stylized model with three periods, homoskedastic errors, and (potentially) linear violations of parallel trends.

2.1 Stylized model set-up

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 ($D_i = 1$) receive a treatment of interest between periods 0 and 1, whereas individuals in the control group ($D_i = 0$) 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. The observed outcome can then be written as $y_{it} = D_i y_{it}(1) + (1 - D_i) y_{it}(0)$. 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

[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.

data-generating process for $y_{it}(0)$ is given by

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

where $\epsilon_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. The term $D_i \times g(t)$ 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 "event-study" difference-in-differences regression specification

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

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 section illustrates how this selection plays out in detail.

2.2 Conventional estimates and CIs after pre-testing

We now analyze the performance of the point estimates and CIs for $\hat{\beta}_1$ under the data-generating process described above. For simplicity, we focus on linear violations of parallel trends, $g(t) = \gamma \cdot t$, and vary the slope of the difference in trends γ. We choose the number of observations N and variance σ_ϵ^2 so that $\mathbb{V}\text{ar}\left[\hat{\beta}_1\right] = \mathbb{V}\text{ar}\left[\hat{\beta}_{-1}\right] = 1$.[3] We then analyze the properties of traditional point estimates and CIs both unconditionally, and conditional

[3] Note that $\mathbb{V}\text{ar}\left[\hat{\beta}_1\right] = \frac{4\sigma_\epsilon^2}{N}$, and likewise for $\hat{\beta}_{-1}$, so the variance depends only on the ratio of σ_ϵ^2 and N.

on surviving a pre-test for the null hypothesis that β_{-1} is equal to 0 at the 95% level. The results of this exercise are summarized in Figure 1.

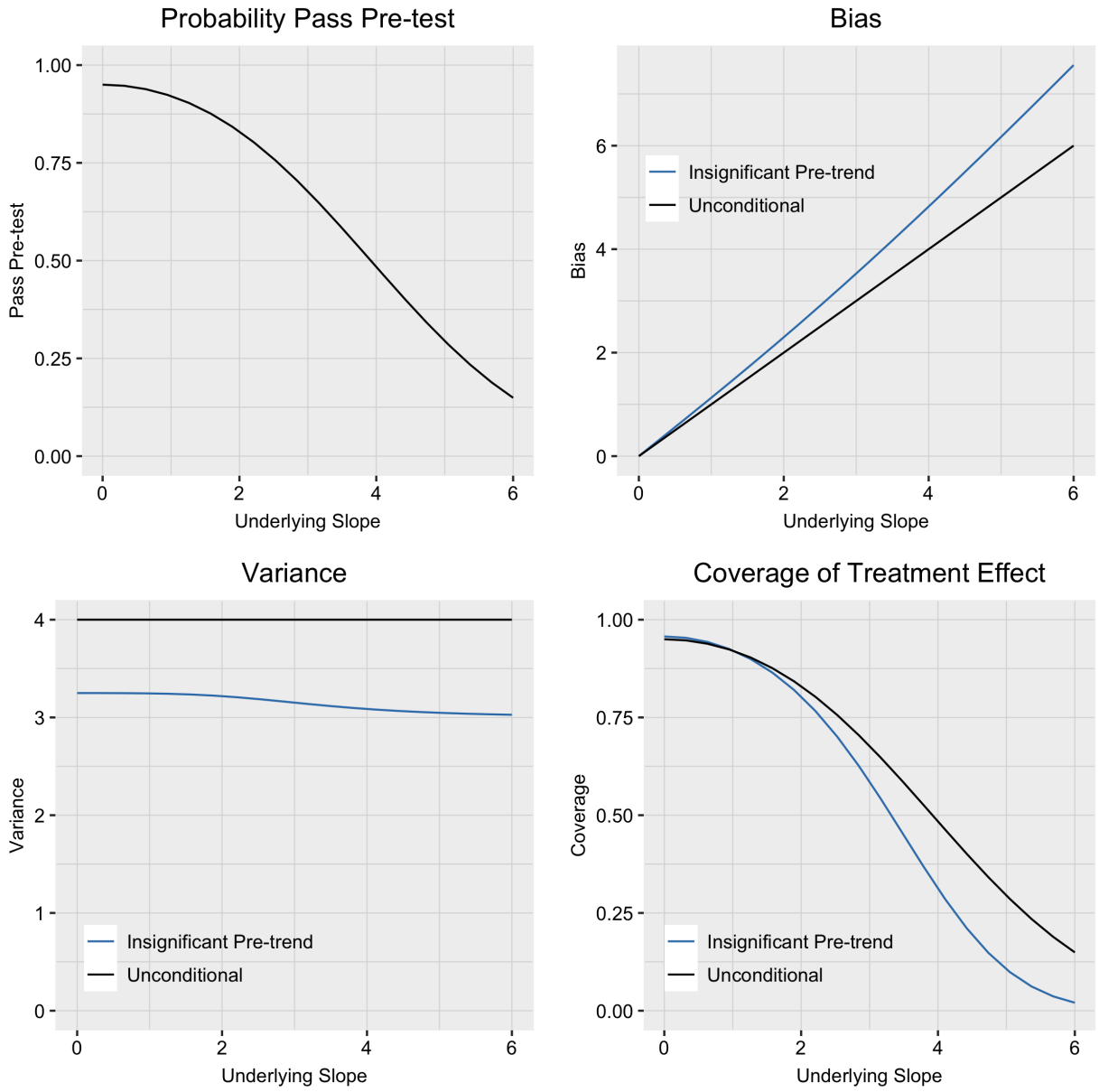
Power of pre-testing. The top left panel of Figure 1 shows the probability that we "pass" the pre-test, i.e. that we do not find a significant pre-period coefficient $\hat{\beta}_{-1}$ at the 95% level. Naturally, we fail to find a significant pre-period coefficient 95% of the time when parallel trends holds ($\gamma = 0$). Note, however, that we also do not find a significant pre-trend in a substantial fraction of cases for certain violations of parallel trends. Consider, for instance, the case when $\gamma = 2$, so that the mean of $\hat{\beta}_{-1}$ is two standard errors below 0. We will fail to reject the null if $\hat{\beta}_{-1}$ is within 2 standard errors of 0, which occurs if $\hat{\beta}_{-1}$ is between 0 and 4 standard errors above its mean. This occurs with probability $\Phi(4) - \Phi(0) \approx 0.5$, so we pass the pre-test about half the time when the magnitude of γ is twice the standard error of $\hat{\beta}_{-1}$.

Bias. The top right panel of Figure 1 shows the bias of $\hat{\beta}_1$ for the treatment effect in period 1. When we do not condition on the result of the pre-test, by construction the bias is just the slope of the differential trend, γ, which is shown in black. In blue, we plot the bias of $\hat{\beta}_1$ in realizations of the data in which we do not detect a significant pre-trend. We see that when parallel trends is violated, the bias conditional on passing the pre-test is *larger* than the unconditional bias; that is, the realizations of the data in which we fail to detect a violation of parallel trends tend to produce estimates $\hat{\beta}_1$ that are more biased than would be expected based on the differential trend of slope γ alone.

To understand the intuition for this bias exacerbation, consider Figure 2. The left panel shows realizations of $\Delta\bar{y}_t$, the difference in sample means between the treated and control group in period t, simulated from our DGP when $\gamma = 3$. Note that the slope of the line between period 0 and period 1 corresponds with the magnitude of the pre-period coefficient $\hat{\beta}_{-1}$, whereas the slope between period 0 and period 1 corresponds with $\hat{\beta}_1$. Highlighted in blue are the draws of the data in which we do not detect a significant pre-trend. By construction, these blue lines have small slopes between period -1 and period 0, corresponding with small values of $\hat{\beta}_{-1}$. Note, however, that these blue lines also tend to have larger slopes between period 0 and period 1, corresponding with more biased realizations of $\hat{\beta}_1$. (The right panel of the figure shows the average over 1 million draws.) This is because the blue draws of the data tend to have below-average values of $\Delta\bar{y}_{t=0}$, since negative shocks in period 0 "flatten" out the observed slope in the pre-period. Owing to a mean-reversion effect, there is then a larger change between period 0 and period 1, leading these insignificant draws of the data to be particularly biased for the treatment effect in period 1.[4]

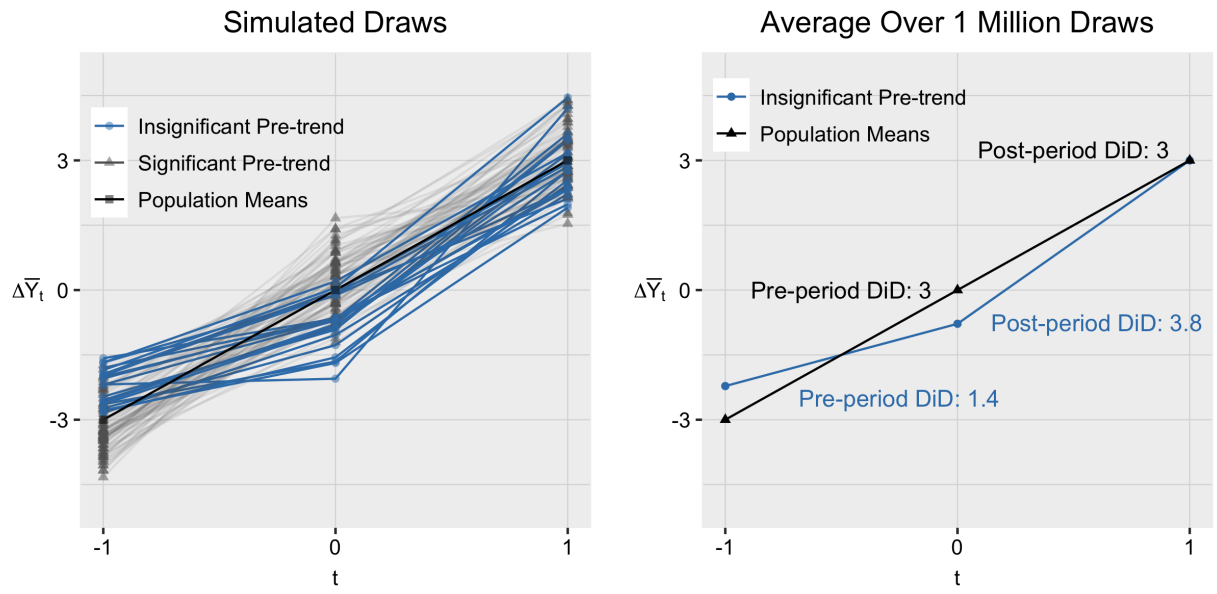
[4]Daw and Hatfield (2018) show that a similar mean-reversion effect can produce bias when creating a

Figure 1: Probability Pass Pre-test, and 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 significant pre-trend at the 5% level. The top left panel shows the probability of passing the pre-test; the top right and bottom left panels respectively show the bias and variance of $\hat{\beta}_1$. The bottom right panel shows the coverage of the treatment effect for a nominal 95% interval.

Figure 2: Intuition for how bias is worse conditional on not detecting a significant pre-trend



Note: The left 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 pre-period coefficient $\hat{\beta}_{-1}$ is insignificant at the 95% level. The right panel shows the average of the blue lines over 1 million draws.

Variance. The bottom left panel of Figure 1 shows the variance of $\hat{\beta}_1$, both unconditionally and conditional on passing the pre-test. We see that the variance of $\hat{\beta}_1$ is lower conditional on passing the pre-test for all values of γ. The intuition for this is that $\hat{\beta}_1 = \Delta\bar{y}_{t=1} - \Delta\bar{y}_{t=0}$, and conditioning on passing the pre-test tends to reduce the variance of $\Delta\bar{y}_{t=0}$. This can be seen, for instance, in the left panel of Figure 2, in which the blue dots at $t = 0$ are substantially less dispersed than the gray triangles.

Coverage of CIs. The bottom right panel of Figure 1 plots the rate at which traditional CIs for $\hat{\beta}_1$ cover the treatment effect. The unconditional coverage rate starts at the nominal 95% level when parallel trends holds ($\gamma = 0$), but declines as γ increases because of the bias from the differential trend. Owing to the symmetry of our stylized example between period -1 and period 1, the unconditional coverage rate is identical to the probability of passing the pre-test. Thus, under the slope γ for which we detect a pre-trend only half the time, a conventional 95% CI would (unconditionally) reject the true treatment effect half the time. The coverage rates conditional on passing the pre-test start out slightly higher than the nominal coverage rate (96% under parallel trends), but can be substantially lower than

control group based on matching on pre-period outcomes, rather than pre-testing for significance.

the unconditional coverage rates for larger violations of parallel trends. This is a result of a trade-off between two effects: on the one hand, the variance of $\hat{\beta}_1$ is smaller conditional on passing the pre-test; on the other, pre-testing introduces additional bias when parallel trends is violated. Conditional on surviving the pre-test, conventional CIs thus over-cover for values of γ close to zero, where the variance effect dominates, but substantially undercover for larger values of γ, for which bias dominates.

2.3 Implications of publication rules that require pre-testing

So far, we have evaluated the implications of pre-testing on the performance of conventional estimates and CIs for fixed values of γ. We now extend the baseline model to understand the implications of these results for publication regimes in an environment where researchers try many different studies, and parallel trends is satisfied in some of these but not others. Intuitively, when we require an insignificant pre-trend to publish, there is a tradeoff between two effects. First, requiring an insignificant pre-trend increases the fraction of published studies in which parallel trends holds ($\gamma = 0$). Second, pre-testing affects the distribution of estimates that survive for any given violation of parallel trends. As shown in the previous section, bias and coverage rates may be worse conditional on passing the pre-test for a fixed value of γ.

To clarify these tradeoffs more formally, we consider a simple extension to the stylized model in which parallel trends holds in fraction $1 - \theta$ of latent studies, and in fraction θ of latent studies there is a linear violation of parallel trends with slope $\bar{\gamma} > 0$. If we did not test for parallel trends and published everything, the expected bias in published studies would be:

$$Bias^{\text{No test}} = P(\gamma = \bar{\gamma})\bar{\gamma} = \theta\bar{\gamma}.$$

Likewise, if we only accept studies that pass the pre-test, the bias in published studies is:

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

The ratio of biases across the two regimes is then:

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

Equation (3) makes clear the tradeoffs involved in requiring an insignificant pre-trend to publish. The first term represents the relative fraction of published studies with a biased design ($\gamma = \bar{\gamma}$) 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, in our simple model this bias is worse conditional on surviving 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 (θ) 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:

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

where

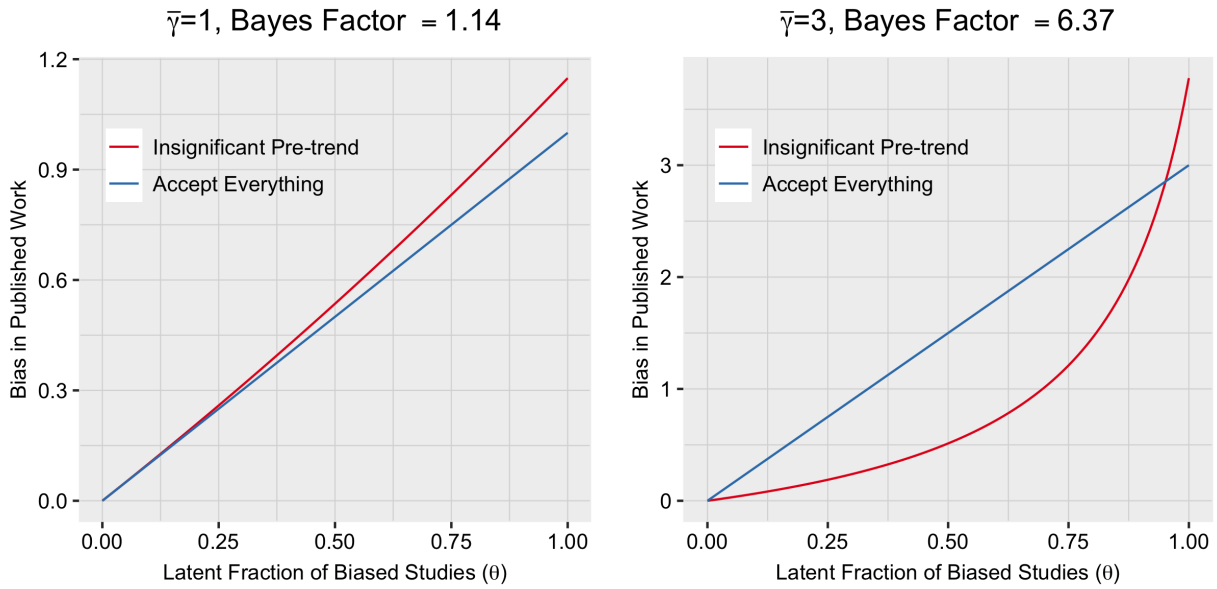
$$BF := \frac{P(Accept|\gamma = 0)}{P(Accept|\gamma = \bar{\gamma})}$$

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 θ 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 3, which shows the (mean) bias in published studies as a function of θ for $\bar{\gamma} = 1$ and $\bar{\gamma} = 3$, which correspond with Bayes factors of 1.1 and 6.4. When $\bar{\gamma} = 1$ and the Bayes factor is small, we see that requiring an insignificant pre-test to publish leads to weakly larger bias in published work for all values of θ, with the pre-testing regime doing substantially worse when θ is large. For $\bar{\gamma} = 3$, where the pre-test is better powered, we see that requiring an insignificant pre-test to publish can substantially reduce bias for lower values of θ, but will nonetheless exacerbate bias if θ is sufficiently large.

Appendix Figure D1 shows the analogous results for size control in published work, rather than bias. Again, the pre-testing publication regime may be ineffective in controlling size, with size contol in published work substantially exceeding the nominal level of 5% for many parameter values. The pre-testing regime can even produce worse results than accepting all papers if the pre-test is sufficiently underpowered or the fraction of latent biased studies is high.

An implication of this section is that requiring insignificant pre-trends to publish need not necessarily be effective in reducing bias or controlling size in published work. Indeed, if the

Figure 3: Comparing bias in published studies when requiring an insignificant pre-trend versus publishing everything



Note: Each figure shows the (mean) bias in published work in the setting described in Section 2.3 as a function of the fraction of latent studies in which parallel trends is violated (θ). The Insignificant Pre-trend regime only publishes studies in which $\hat{\beta}_{-1}$ is statistically insignificant. The two panels show results for different values of the slope of the differential trend (γ) when parallel trends fails. See Section 2.3 for further detail.

pre-test is underpowered or if a high fraction of latent research designs are biased, requiring pre-testing may even exacerbate these issues. These results should make researchers cautious of relying solely on the significance of pre-tests, unless they have context-specific knowledge about the power of the pre-test against the relevant alternatives or the latent credibility of the research design.

3 Theory: Pre-testing in a more general model

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 Model

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:

$$\begin{pmatrix} \hat{\beta}_{post} \\ \hat{\beta}_{pre} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta_{post} \\ \beta_{pre} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right). \tag{5}$$

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 (or linear combinations thereof) in the case when $M > 1$.

We decompose the population mean as

$$\begin{pmatrix} \beta_{post} \\ \beta_{pre} \end{pmatrix} = \begin{pmatrix} \tau_{post} \\ 0 \end{pmatrix} + \begin{pmatrix} \delta_{post} \\ \delta_{pre} \end{pmatrix}, \tag{6}$$

where τ_{post} 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 2.2, the true treatment effect was $\tau_{post} = 0$, but the researcher estimating regression (2) would have bias from the underlying trend given by $(\delta_{post}, \delta_{pre}) = (\gamma, -\gamma)$. If parallel trends holds, then $\delta = 0$. We assume that the treatment of interest has no causal effect prior to its implementation, so that $\beta_{pre} = \delta_{pre}$.

I will analyze the properties of the distribution of $\hat{\beta}_{post}$ conditional on a pre-test based on the pre-period 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 $\hat{\beta}_{pre} \in B_{NIS} := \{\hat{\beta}_{pre} : |\hat{\beta}_{pre,j}|/\sqrt{\Sigma_{jj}} \leq 1.96 \text{ for all } j\}$.

Remark 1. The finite-sample normal model specified above will hold exactly if we assume normal errors, as in the example in Section 2, but can more reasonably be thought of as an asymptotic approximation, since a wide variety of estimation procedures will yield asymptotically normal coefficients via the central limit theorem. For instance, the traditional two-way fixed effects model (2) will lead to asymptotically normal coefficients as N grows large under mild regularity conditions. Other procedures, such as the GMM estimator proposed by Freyaldenhoven et al. (2019), will also have an asymptotically normal distribution under suitable regularity conditions, and thus the results here can also be used to analyze the asymptotic distribution of treatment effects estimates conditional on not finding significant pre-period placebo coefficients in these models as well. Appendix C shows that the

results derived in the finite sample normal model hold uniformly over a wide range of data-generating processes under which the probability of passing the pre-test does not disappear asymptotically.[5] ■

Remark 2. An active recent literature has examined the interpretation of the coefficients β_{pre} and β_{post} from the two-way fixed effects regression (2) when there is staggered adoption of treatment timing (Abraham and Sun, 2018; Borusyak and Jaravel, 2016; Callaway and Sant'Anna, 2019).[6] The results in Abraham and Sun (2018) imply that the coefficients β_{pre} and β_{post} may not have a sensible interpretation if there is treatment effect heterogeneity, *even* under a strong generalization of the parallel trends assumption to the staggered timing case. Specifically, they show that there may be pollution across lags, such that β_{pre} may be non-zero even if parallel trends holds, and the coefficient β_t for $t > 0$ may be affected by treatment effects in periods other than t. On the other hand, their results suggest that if cohorts that adopt treatment at different times have the same dynamic path of treatment effects, then the coefficients β can sensibly be decomposed as in (6), where τ_{post} is the dynamic profile of treatment effects and $\delta_{pre} \neq 0$ only if parallel trends is violated in the pre-period.

Moreover, Abraham and Sun (2018) and Callaway and Sant'Anna (2019) propose alternative estimators that have sensible interpretations as weighted averages of causal effects at a particular lag since treatment even under heterogeneous treatment effects. These estimators can be used to construct estimates of dynamic causal effects, as well as placebo pre-treatment estimates. Since these estimators yield asymptotically normal coefficients under suitable regular conditions, the results here can also be used to analyze the asymptotic distribution of those estimates following a pre-test of these placebo coefficients. ■

3.2 Bias After Pre-testing

I begin by analyzing the bias of $\hat{\beta}_{post}$ for τ_{post} conditional on passing the pre-test. The following result, which follows from standard arguments using the conditional distributions of multivariate normals, provides a formula for the conditional bias.

Proposition 3.1. *For any conditioning set B,*

[5] The condition that the probability of passing the pre-test does not disappear asymptotically requires that the pre-period trend δ_{pre} be shrinking with the sample size. This local-to-0 approximation captures the fact that in finite samples the pre-trend may be of a similar order of magnitude as the sampling uncertainty in the data. In a model with fixed δ_{pre}, the probability of rejecting the pre-test would be either 0 or 1 asymptotically, which does not capture the fact that in practice we are often uncertain whether the pre-trend is zero or not.

[6] Relatedly, several other recent papers focus on the interpretation of "static" specifications with only a post-treatment indicator, rather than leads and lags of treatment timing. See the Related Literature section for references.

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

The formula in Proposition 3.1 illustrates that the expectation of $\hat{\beta}_{post}$ conditional on passing the pre-test is the sum of i) the treatment effect of interest τ_{post}, ii) the unconditional bias δ_{post}, and iii) an additional "pre-test bias" term, which 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.

3.2.1 Sufficient conditions for bias exacerbation

In the stylized example in Section 2, we saw that the bias of $\hat{\beta}_{post}$ for τ_{post} was worse conditional on passing the pre-test when there were linear violations of parallel trends. We now show that this result extends to arbitrary monotone violations of parallel trends under certain restrictions on the covariance matrix. These restrictions follow from a homoskedasticity assumption in the general case with multiple periods.

We introduce the following restrictions on the covariance matrix, which depend on the number of pre-treatment coefficients, K.

Assumption 1. *Σ satisfies the following restrictions:*

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 $\sigma^2 > \rho$.*

Remark 3. With one pre-treatment period, Assumption 1 imposes only that the pre-period coefficient $\hat{\beta}_{pre}$ has positive covariance with the treatment effect estimate $\hat{\beta}_{post}$. Although in practice having only one pre-period may be rare, when there are multiple pre-periods researchers may test for a pre-trend using a parametric linear trend, such as

$$y_{it} = \alpha_i + \phi_t + \beta_{trend} \times t \times D_i + \sum_{s>0} \beta_s \times 1[s = t] \times D_i + \epsilon_{it}. \tag{7}$$

In this case, testing the significance of $\hat{\beta}_{trend}$ is equivalent to testing a one-dimensional pre-period coefficient. Assumption 1 thus will apply whenever the coefficient $\hat{\beta}_{trend}$ is positively correlated with the estimate for a post-treatment coefficient of interest (e.g., $\hat{\beta}_1$). ■

Remark 4. With multiple pre-treatment periods, Assumption 1 is implied by a suitable homoskedasticity assumption in the canonical two-way fixed effects difference-in-differences

model. To see this, suppose that the data is generated from the model

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

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. 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*. ■

We now show that under Assumption 1, the bias after testing for significant pre-treatment coefficients is worse than the unconditional bias under arbitrary monotone violations of parallel trends.

Proposition 3.2 (Sign of bias under monotone trend)**.** *Suppose that there is an upward pre-trend in the sense that $\delta_{pre} < 0$ (elementwise) and $\delta_{post} > 0$.*[7] *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.

Remark 5. In many difference-in-differences settings, researchers are worried that treatment may be correlated with ongoing secular economic trends, in which case the monotonicity of δ may be a reasonable assumption. Several recent papers suggest that any violations of parallel trends would be monotone. For instance, Lovenheim and Willen (2019) argue that violations of parallel trends cannot explain their results because "pre-[treatment] trends are either zero or in the wrong direction (i.e., opposite to the direction of the treatment effect)." Likewise, Greenstone and Hanna (2014) estimate upward-sloping pre-existing trends and argue that their estimates would be upward biased "if the pre-trends had continued." Nonetheless, there are economic settings in which we do not expect monotonicity to hold, with the so-called

[7] Recall that specifications such as (2) implicitly normalize $\delta_0 = 0$. Thus, if the difference in trends is monotonically increasing, i.e. $\delta_{-K} \leq ... \leq \delta_0 = 0 \leq ... \leq \delta_M$, then the restrictions on δ imposed by Proposition 3.2 hold. Note, however, that the restriction that $\delta_{pre} < 0$ and $\delta_{post} > 0$ is actually somewhat weaker than monotonicity. It allows, for instance, for $\delta_{-3} > \delta_{-2}$, so long as both are less than 0.

Ashenfelter's dip expected in job-training programs as a notable example (Ashenfelter, 1978). ■

Remark 6. The homoskedasticity assumption required to obtain the bias exacerbation in Proposition 3.2 is of course strong and unlikely to hold in most practical applications. 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. Nonetheless, the fact that bias is exacerbated under homoskedasticty and arbitrary monotone violations of parallel trends indicates that pre-testing can exacerbate bias in non-pathological cases.

Moreover, it is straightforward to calculate whether pre-testing will exacerbate bias for any particular underlying trend and covariance matrix using the formula in Proposition 3.1. In Section 4, I apply this approach to calculate the pre-test bias under linear violations of parallel trends in a sample of recently 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. ■

Remark 7. Another limitation of the result in Proposition 3.2 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 distributions from Tallis (1963) and Arismendi Zambrano and Broda (2016), although I leave this to future work. ■

3.2.2 Unbiasedness after pre-testing when parallel trends holds

In the simple example in Section 2, we saw that when parallel trends was satisfied, $\hat{\beta}_{post}$ remained unbiased for the treatment effect conditional on not finding a significant pre-trend. From Proposition 3.1, we see that $\hat{\beta}_{post}$ is conditionally unbiased for τ_{post} if $\delta_{post} = 0$ and $\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] = \beta_{pre}$. One can show that if $\delta_{pre} = 0$, then $\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] = \beta_{pre}$ provided that the pre-test B is symmetric in the sense that we reject the hypothesis of parallel pre-trends for $\hat{\beta}_{pre}$ if and only if we reject the hypothesis for $-\hat{\beta}_{pre}$, a property which holds for any two-sided test of significance. It follows that, as in the simple example, $\hat{\beta}_{post}$ is unbiased for the treatment effect of interest after pre-testing when parallel trends holds, so long as the pre-test is symmetric.

Corollary 3.1 (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 $\hat{\beta}_{pre} \in B$ if and only if $-\hat{\beta}_{pre} \in B$, then*

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

3.3 Pre-testing reduces the variance of estimates

Having analyzed the properties of the mean of the treatment effect estimate conditional on passing a 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 3.3.

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

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 for a large class of pre-tests. In particular, we only require that the pre-test is convex, meaning that if we do not reject parallel trends for $\hat{\beta}_{pre,1}$ and $\hat{\beta}_{pre,2}$, then for $\theta \in (0,1)$, we also will not reject parallel trends for $\theta\hat{\beta}_{pre,1} + (1-\theta)\hat{\beta}_{pre,2}$. This property holds for most common pre-tests – including tests of individual statistical significance, joint tests for significance, and tests for significant linear slopes.

Proposition 3.4 (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

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 substantial linear violations of parallel trends. Second, I show that bias and coverage issues can be substantially different, and in many cases worse, conditional on surviving a pre-test. Although homoskedasticity typically does not hold in practice, the bias from pre-testing nonetheless amplifies the bias from a monotone trend in most cases, 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 CIs 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.[8] 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

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

heterogeneity analyses. I focus here on the first plot presented in the paper that meet the criteria above, which I view as a reasonable proxy for paper's main specification.

Two caveats are in order with regards to the sample of papers considered here. First, my sample by construction only includes papers that made it through the publication process at leading economics journals. To the extent that papers with insignificant pre-trends are more likely to be published, the sample may be biased towards papers where the power of pre-tests is low. Second, several papers in my sample use dynamic two-way fixed effects specifications in settings with staggered treatment timing. As discussed in Remark 2, pre-testing for values of $\delta_{pre} \neq 0$ has a sensible interpretation only under certain homogeneity assumptions about the dynamic path of treatment effects. For the remainder of the section, I suppose that the authors are willing to impose some homogeneity assumptions such that the pre-testing step using their baseline specification is sensible. I then evaluate the power of common pre-tests and distributions of conventional estimates in data-generating processes calibrated to these specifications.

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-period coefficient in three of the 12 papers in my final sample, and in two papers the pre-period coefficients are also jointly significant.[9]

4.3 Evaluating power and pre-test bias in practice

I now evaluate the power of conventional pre-tests and the distortions from pre-testing in data-generating processes calibrated to my survey of recent papers.

Data-generating processes. I calibrate the finite-sample normal model (5) by setting Σ to be the estimated variance-covariance matrix from the specification reported by the

[9] 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-statistic of 1.95.

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.

authors, using whatever clustering method was specified by the authors. Using the finite-sample normal DGP suppresses any complications arising from non-normality of the OLS estimates or difficulties with estimating Σ in finite sample, and thus focuses attention solely on the issues related to pre-testing for violations of parallel trends.

Power calculations. For each study in my sample, I evaluate the power of common pre-trends tests to detect linear violations of parallel trends. In light of the emphasis in published work on the individual statistical significance of the pre-period coefficients, I base my calculations on pre-tests using this criterion (using 95% CIs). Specifically, I consider linear violations of parallel trends with a slope of γ – that is, the element of δ corresponding with period t is $\delta_t = \gamma \cdot t$. I then compute the value of γ for which the probability of passing the pre-test, $\mathbb{P}\left(\hat{\beta}_{pre} \in B_{NIS}\right)$, is equal to 50 or 80 percent. I choose 80 percent since this is often used as a benchmark for an acceptable degree of power in power analyses (Cohen, 1988). I refer to the resulting values, $\gamma_{0.5}$ and $\gamma_{0.8}$, as the slopes against which we have 50 or 80 percent power.[10]

[10]The power of the pre-test under a slope γ could easily be calculated via simulation. However, under the normality assumption, these probabilities can actually be calculated analytically using the formulas of Manjunath and Wilhelm (2012), which I implement using the R package tmvtnorm.

Focusing on linear violations of parallel trends is a reasonable benchmark for several reasons. First, when researchers are worried about possible violations of parallel trends, they frequently control parametrically for linear differential trends (e.g., Wolfers (2006); Dobkin et al. (2018); Goodman-Bacon (2018)), which indicates that authors perceive linear violations of parallel trends to be reasonable in many cases. In other cases, researchers may not be confident that the linear functional form is correct – which is perhaps why they do not include parametric linear controls – but they may nonetheless be worried about secular differences in trends that evolve smoothly over time. Focusing on the performance of conventional pre-tests under linear violations of parallel trends thus gives a lower bound on the worst-case performance of pre-tests over smooth classes of violations that include linear trends. For instance, a common way to parameterize the smoothness of a (discrete) curve is via bounds on (the disrete analog of) its second derivative, otherwise known as a second-order Holder class (e.g., Armstrong and Kolesar (2018); Kolesar and Rothe (2018); Rambachan and Roth (2019)). Since linear functions are included in any second-order Holder class, the results in this section can alternatively be viewed as a lower-bound on the worst-case performance of the pre-testing method over these classes of smoothly-evolving differences in trends. These calculations likewise provide a lower bound on the worst-case performance over the class of polynomial violations.

Nonetheless, linear violations of parallel trends will not be an appropriate benchmark in all cases. In Appendix D, I conduct a similar exercise under data-generating processes in which there are differential stochastic shocks to the treated and control groups. I again find poor performance of standard pre-testing methods in controlling size distortions from the differential trends.

Bias and size calculations. I evaluate the performance of conventional estimators and CIs under data-generating processes with linear violations of parallel trends with slopes $\gamma_{0.5}$ or $\gamma_{0.8}$. I focus on estimation and inference for two scalar causal estimands, the first period treatment effect τ_1, and the average treatment effect across the post-treatment periods, $\bar{\tau} = \frac{1}{M}(\tau_1 + ... + \tau_M)$. Note that both of these estimands can be written as linear combinations of the vector of treatment effects, $\tau_* = l'\tau_{post}$. I therefore evaluate the performance of the estimator $\hat{\tau}_* = l'\hat{\beta}_{post}$, as well as the 95% confidence interval, $CI_{\tau_*} = \hat{\tau}_* \pm 1.96\sigma_{\tau_*}$, where $\sigma^2_{\tau_*} = l'\Sigma l$.

Specifically, I calculate the unconditional bias $\mathbb{E}\left[\hat{\tau}_* - \tau\right]$, and the bias conditional on passing the pre-test $\mathbb{E}\left[\hat{\tau}_* - \tau \mid \hat{\beta}_{pre} \in B_{NIS}\right]$. Note that the unconditional bias for τ_1 is merely γ, whereas the unconditional bias for $\bar{\tau}$ is $\frac{1}{M}(\gamma + ... + M\gamma)$. I compute the conditional bias using the formula in Proposition 3.1, making use of the results in Manjunath and Wilhelm (2012)

to calculate the expectation of the multivariate truncated normal analytically. Likewise, I compute the size (i.e. null rejection probability) of $CI_{\tau*}$ both unconditionally and conditionally, $\mathbb{P}(\tau_* \notin CI_{\tau_*})$ and $\mathbb{P}(\tau_* \notin CI_{\tau_*} \,|\, \beta_{pre} \in B_{NIS})$.[11] These probabilities are calculated analytically using the tmtvnorm package in R, which implements results on the marginal distribution of truncated normals from Cartinhour (1990).[12]

Results. I find that the magnitude of the violations of parallel trends against which we have 50 and 80 percent power can be sizable relative to the magnitude of the estimated treatment effect. Figure 4 plots in green the magnitude of the unconditional bias for $\bar{\tau}$ under the linear violations of parallel trends against which we have 80% power ($\gamma_{0.8}$). The bias from such trends is often of a magnitude comparable to, and in some cases larger than, the estimated treatment effect. Appendix Figure D2 shows the equivalent results for the trends against which we have 50 percent power ($\gamma_{0.5}$); even at the lower power threshold, the biases are of a comparable magnitude to the estimated treatment effects in several cases. Appendix Figures D3 and D4 present the analogous results when the estimand is the first period treatment effect (τ_1); the patterns are similar although a bit less extreme. This is intuitive, given that the bias from a linear trend grows over time, and we would thus expect it to be larger for later periods.

I likewise find that the bias conditional on passing the pre-test can substantially differ from the unconditional bias, and is worse in most cases. Figure 4 plots in red the conditional bias for $\bar{\tau}$. I also summarize the additional bias from pre-testing as a percentage of the unconditional bias 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.[13] The analogous values are even larger when looking at the trend against which we have 80 percent power. Moreover, the pre-test bias and the bias from trend go in the same direction in all but two of the studies in the sample when the estimand is $\bar{\tau}$, and all but three of the studies when it is τ_1. Thus, although not always true, the prediction of the direction of the bias from the homoskedastic case holds in most cases I consider.

I also find that traditional CIs perform poorly under these violations of parallel trends

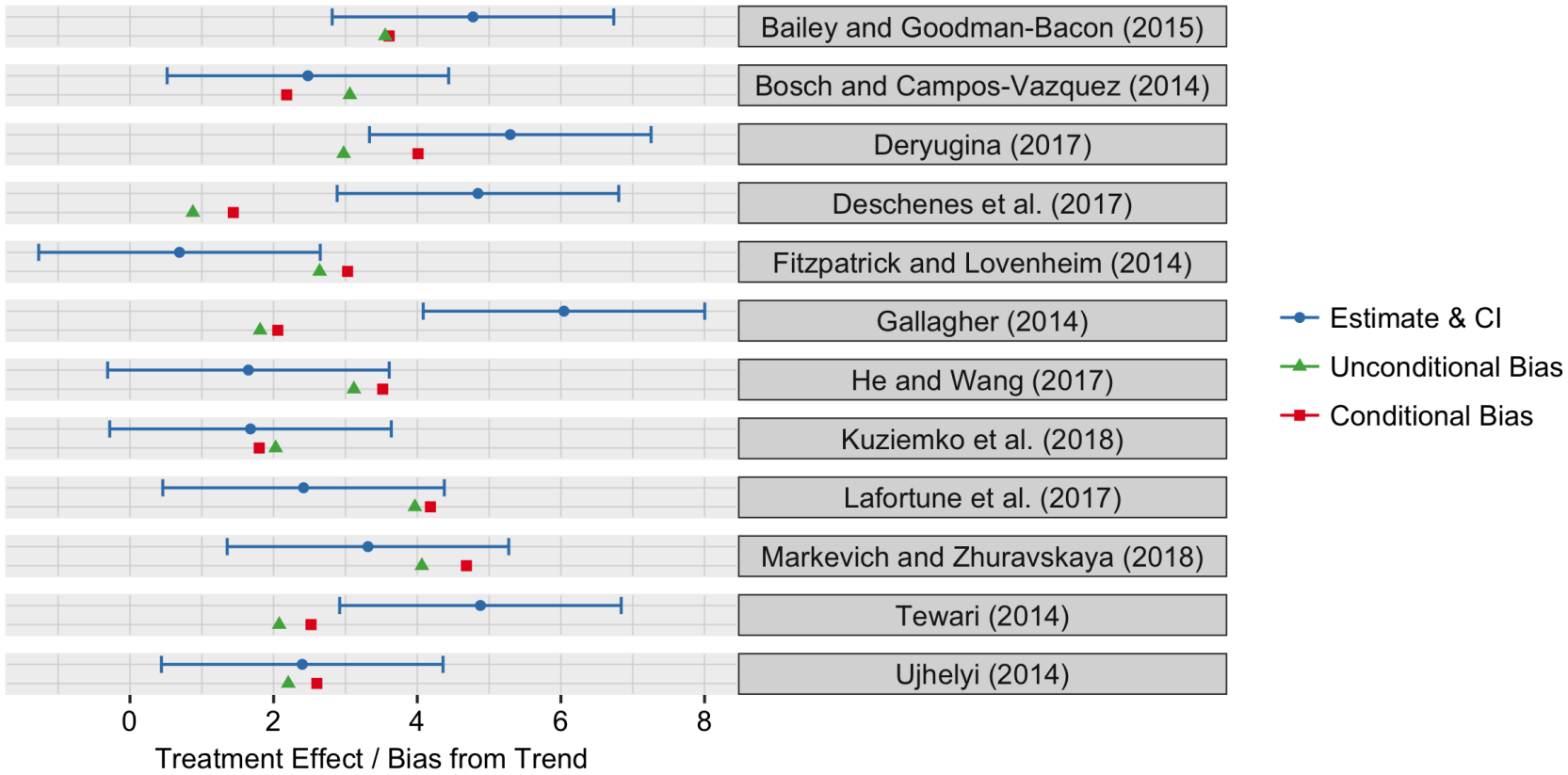
[11] Note that both the estimates and confidence intervals are equivariant in τ_{post}, so that the bias and coverage probabilities do not depend on the true value of the treatment effect. τ_{post} thus does not need to be specified in our calibrations.

[12] I have verified that calculating both the bias and size results via simulation yields similar results to the analytic formulas discussed above.

[13] 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).

that conventional pre-tests are marginally powered to detect. Table 4 shows the probability that a 95% confidence interval for $\bar{\tau}$ fails to include the true value of the parameter. Although the true parameter should nominally fall outside the confidence interval no more than 5% of the time, for many of the specifications the null rejection rate is over 50%. Table D1 shows the analogous results when the estimand is the first period after treatment. Although less extreme, null rejection probabilities again often substantially exceed their nominal levels.

Figure 4: OLS Estimates and Bias from Linear Trends for Which We Have 80 Percent Power – Average Treatment Effect



Note: I calculate the linear trend against which we would have a rejection probability of 80 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 estimand is the average of the treatment effects in all periods after treatment began, $\bar{\tau}$.

	Treatment Effect:			
	1st Period		All Periods	
	Slope of differential trend:			
Paper	$\gamma_{0.5}$	$\gamma_{0.8}$	$\gamma_{0.5}$	$\gamma_{0.8}$
Bailey and Goodman-Bacon (2015)	51	56	1	2
Bosch and Campos-Vazquez (2014)	-29	-34	-25	-29
Deryugina (2017)	103	120	30	35
Deschenes et al. (2017)	88	119	48	64
Fitzpatrick and Lovenheim (2014)	25	30	12	15
Gallagher (2014)	57	62	11	14
He and Wang (2017)	29	34	11	13
Kuziemko et al. (2018)	-16	-20	-9	-11
Lafortune et al. (2017)	-9	-10	5	5
Markevich and Zhuravskaya (2018)	52	62	13	15
Tewari (2014)	90	102	19	21
Ujhelyi (2014)	51	59	15	18

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 percentage of the unconditional bias, i.e. $100 \cdot \frac{ConditionalBias - UnconditionalBias}{UnconditionalBias}$. Biases are calculated under linear violations of parallel trends with slopes $\gamma_{0.5}$ and $\gamma_{0.8}$, against which conventional pre-tests have 50 or 80% power.

	Conditional on passing pre-test?			
	No		Yes	
	Slope of differential trend:			
	$\gamma_{0.5}$	$\gamma_{0.8}$	$\gamma_{0.5}$	$\gamma_{0.8}$
Bailey and Goodman-Bacon (2015)	0.61	0.94	0.62	0.95
Bosch and Campos-Vazquez (2014)	0.49	0.86	0.28	0.61
Deryugina (2017)	0.49	0.84	0.75	1.00
Deschenes et al. (2017)	0.09	0.14	0.10	0.25
Fitzpatrick and Lovenheim (2014)	0.41	0.75	0.50	0.87
Gallagher (2014)	0.19	0.44	0.22	0.54
He and Wang (2017)	0.54	0.88	0.63	0.95
Kuziemko et al. (2018)	0.28	0.53	0.20	0.42
Lafortune et al. (2017)	0.71	0.98	0.76	0.99
Markevich and Zhuravskaya (2018)	0.76	0.98	0.87	1.00
Tewari (2014)	0.20	0.55	0.25	0.72
Ujhelyi (2014)	0.29	0.60	0.36	0.76

Table 4: Null Rejection Probabilities for Nominal 5% Test of Average Treatment Effect Under Linear Trends Against Which We Have 50 or 80% Power

Note: This table shows null rejection probabilities for nominal 5% significant level tests using data-generating processes under which there are linear violations of parallel trends that conventional pre-tests would detect 50 or 80% of the time ($\gamma_{0.5}$ and $\gamma_{0.8}$). The first two columns show unconditional null rejection probabilities, whereas the latter two columns condition on passing the pre-test. The estimand is the average of the post-treatment causal effects, $\bar{\tau}$.

5 Alternative Approaches

In light of the issues highlighted with the common approach of pre-testing for pre-trends, this section discusses alternative approaches to estimation and inference in settings where we are concerned that parallel trends may be violated. Each of these methods can be viewed as imposing different restrictions on the way that parallel trends may be violated. The appropriate method will therefore depend on which assumptions are reasonable in a given context. I first discuss each of the methods, and provide recommendations for choosing between them in Section 5.3.

5.1 Parametric Approaches

I first consider parametric approaches, which impose particular functional form restrictions on the way that parallel trends can be violated. Extrapolations of pre-treatment data can then be used to estimate the counterfactual post-treatment difference in trends, thus removing the bias in conventional estimates. One common approach in the literature is to assume linearity of the differential trend (e.g., Wolfers (2006); Dobkin et al. (2018); Goodman-Bacon (2018)). That is, we assume that $\delta_t = t \times M_T \delta_{pre}$, where $M_T = (\boldsymbol{t}'\boldsymbol{t})^{-1}\boldsymbol{t}'$ is the matrix that projects the pre-period coefficients onto a linear time trend $\boldsymbol{t} = (-K, ..., -1)$. Under this assumption, $\tau_t = \beta_t - t \times M_T \beta_{pre}$, and it is thus straightforward to do estimation and inference using the sample analog, $\hat{\tau}_t = \hat{\beta}_t - t \times M_T \hat{\beta}_{pre}$, with standard errors calculated via the delta method.[14] An advantage of this approach is that it provides valid causal estimates without pre-testing, provided that the functional form restriction is correct.

Under the appropriate functional form restrictions, the simple method discussed so far yields valid causal estimates from an ex ante sampling perspective. However, it does not yield valid causal estimates conditional on passing a pre-test. For instance, it is clear from Figure 2 that extrapolating a linear trend in the DGP considered in Section 2.2 would yield valid causal estimates unconditionally, but fail to properly control for the pre-trends conditional on passing the pre-test. Standard parametric methods are thus not well-suited for retrospective analysis, i.e. evaluating previously published studies, if we think that these were selected on the basis of pre-trends tests. In Appendix B, I show that parametric methods can be adapted to obtain estimates and CIs with good properties conditional on passing a pre-test (or model selection on the basis of pre-trends), provided that we know the

[14] We consider here the approach where one first estimates a usual "event-study" specification and then projects out a linear trend fitted to the estimated coefficients. A closely related approach directly estimates a version of equation (2) that includes an interaction of treatment status with time. The two approaches are identical in the balanced panel, non-staggered treatment timing difference-in-differences case. See Goodman-Bacon (2019, Supplement, p.20) for additional discussion regarding the staggered case.

pre-testing rule. By adapting publication-bias corrections from Andrews and Kasy (2019), I provide median-unbiased estimates and CIs that are valid after pre-testing under parametric functional form restrictions on the difference in trends. These adapted methods are thus suitable for parametric, retrospective analyses of published studies that relied on pre-trends testing.

An issue with the parametric approach, however, is that we are often unsure of the correct functional form for the difference in trends (Wolfers, 2006; Lee and Solon, 2011). To address this, researchers sometimes report specifications using different functional forms, e.g. linear and quadratic specifications. This, in my view, is an improvement upon the status quo of relying only on the significance of pre-trends tests, as results that do not survive a simple linear or quadratic adjustment, despite an insignificant pre-trend, should rightly be viewed with caution. Nonetheless, this approach is a bit unsatisfying from a theoretical perspective, as we usually do not believe that either a linear or quadratic extrapolation is exactly correct.

5.2 Alternative relaxations of the parallel trends assumption

In light of this, we now turn our attention to alternative approaches that relax the exact parallel trends assumption without taking as strong a stance on the functional form of possible violations of parallel trends.

One such approach is that of Freyaldenhoven et al. (2019). Their approach allows for their to be differential trends in an outcome y_{it}, so long as i) differences in trends in y_{it} are driven by some unobserved factors η_{it} that affect both y_{it} and treatment status D_{it}, and ii) there exists an alternative outcome x_{it}, which is assumed to also be affected by η_{it} (possibly with different factor loadings) but is unaffected by the treatment. In a minimum wage context, y could be youth employment, D an indicator for whether the minimum wage was raised, x adult employment (which is plausibly unaffected by the minimum wage), and η factors related to local labor demand. Freyaldenhoven et al. (2019) show that under these assumptions (and certain other technical conditions), the causal effect of the treatment is point identified, and it can be consistently estimated via GMM. An advantage of this approach is that it allows for complicated forms of pre-trends that cannot be represented by a simple functional form. A challenge is that it is often difficult to determine an appropriate excluded outcome.

Rambachan and Roth (2019) provide an alternative approach to robust inference that attempts to formalize many of the implicit intuitions behind pre-trends testing. They provide methods for inference on components of τ under the assumption that the difference in trends δ lies in some pre-specified class Δ. They propose as a default classes Δ that impose smoothness restrictions on the possible difference in trends, e.g. bounds on (the discrete ana-

log of) the second derivative.[15] This formalizes the intuition behind pre-trends testing that the pre-treatment differences are informative about the counterfactual post-treatment differences in trends, since without such a smoothness restriction, the difference in trends could be arbitrarily "jumpy," and thus values of δ_{pre} close to 0 would provide little information about the post-treatment bias δ_{post}. The parametric linear specifications discussed above can be viewed as a special case of a smoothness restriction in which we impose that the differences in trends are exactly linear. Rambachan and Roth (2019) recommend conducting sensitivity analysis with respect to the allowed degree of non-linearity in the differential trends, an exercise which enables the researcher to make precise what needs to be assumed about the possible differences in trends in order to draw particular conclusions. Their framework also allows researcher to incorporate sign and shape restrictions, such as monotonicity, that may be motivated by context-specific knowledge.

An advantage of the approaches of Freyaldenhoven et al. (2019) and Rambachan and Roth (2019) is that they provide uniformly valid inference under certain restrictions about the way in which parallel trends may be violated. Under these restrictions, we can therefore obtain valid causal inference without conducting pre-tests, thus avoiding the issues with pre-testing discussed in this paper.

A disadvantage of the approaches of Freyaldenhoven et al. (2019) and Rambachan and Roth (2019) is that both provide valid inference only from an *ex ante* sampling perspective, and not conditional on passing a pre-test. Thus, while both are valuable for prospective analyses, they cannot currently be used for retrospective analysis of work that has already been screened for pre-existing trends. An interesting question for future research is the extent to which these procedures can be modified to obtain valid inference following a pre-test, analogous to the modifications discussed above and in Appendix B for the parametric approach.

5.3 Recommendations

Given the issues with pre-testing discussed in this paper, I strongly recommend against relying solely on tests for pre-trends in contexts in which there is concern that the parallel trends assumptions may be violated. Depending on the context and assumptions the researcher is willing to impose, each of the three alternative approaches discussed above may be reasonable. The parametric approach is a sensible option in settings where researchers have a strong

[15] A closely related predecessor to Rambachan and Roth (2019) is Manski and Pepper (2018), who consider how the identified set of parameters changes under different assumptions about how parallel trends is violated. In contrast to Rambachan and Roth (2019), however, they do not formally consider incorporating information from pre-trends or conducting inference.

view about the relevant functional form of a violation of parallel trends. When combined with the corrections discussed in Appendix B, it is also currently the only viable method to obtain valid estimation and inference for retrospective analyses of papers that have been screened on the basis of pre-trends. Likewise, in settings where the researcher does not have a strong prior about the functional form restriction, but does know of an excluded outcome affected by the same confounds, the approach of Freyaldenhoven et al. (2019) is sensible. Finally, I recommend the sensitivity analysis described in Rambachan and Roth (2019) for settings where the researcher has neither a strong prior on the functional form of differential trends or an excluded outcome.

Regardless of the exact method, I urge researchers to bring economic knowledge to bear in evaluating the assumptions necessary to obtain valid causal inference. To be valid, each of these alternative methods requires a particular set of assumptions on how parallel trends can potentially be violated. Bringing economic knowledge to bear on how parallel trends may be violated, and thus the plausibility of these assumptions, will yield stronger, more credible inferences than relying on the statistical significance of pre-trends tests alone. This argument echoes the sentiment of Kahn-Lang and Lang (2018), who encourage researchers to consider the economic content of the parallel trends assumption.

6 Conclusion

This paper illustrates the limitations of tests for pre-trends, which are common in applied work. I show both theoretically and in simulations based on a survey of recent papers that pre-trends testing may be ineffective in guarding against potential violations of parallel trends, both because power may be low and because of distortions in the distribution of conventional estimates from pre-testing. I discuss alternative econometric approaches for settings where there is concern that parallel trends may be violated. I encourage researchers to be transparent about their assumptions about how parallel trends may be violated, and to use economic knowledge to evaluate the content of these assumptions.

References

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. (2019). Identification of and Correction for Publication Bias. *American Economic Review*, 109(8):2766–2794.

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. and Kolesar, M. (2018). Optimal inference in a class of regression models. 86:655–683.

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

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics*, 60(1):47–57.

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

Bailey, M. J. and Goodman-Bacon, A. (2015). The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans. *American Economic Review*, 105(3):1067–1104.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program Evaluation and Causal Inference With High-Dimensional Data. *Econometrica*, 85(1):233–298. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA12723.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

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.

Bosch, M. and Campos-Vazquez, R. M. (2014). The Trade-Offs of Welfare Policies in Labor Markets with Informal Jobs: The Case of the "Seguro Popular" Program in Mexico. *American Economic Journal: Economic Policy*, 6(4):71–99.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic Press. Google-Books-ID: YleCAAAAIAAJ.

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. (2018). Two-way fixed effects estimators with heterogeneous treatment effects. *arXiv:1803.08807 [econ]*.

Deryugina, T. (2017). The Fiscal Cost of Hurricanes: Disaster Aid versus Social Insurance. *American Economic Journal: Economic Policy*, 9(3):168–198.

Deschênes, O., Greenstone, M., and Shapiro, J. S. (2017). Defensive Investments and the Demand for Air Quality: Evidence from the NOx Budget Program. *American Economic Review*, 107(10):2958–2989.

Dobkin, C., Finkelstein, A., Kluender, R., and Notowidigdo, M. J. (2018). The economic consequences of hospital admissions. 108(2):308–352.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Fitzpatrick, M. D. and Lovenheim, M. F. (2014). Early retirement incentives and student achievement. *American Economic Journal: Economic Policy*, 6(3):120–154.

Freyaldenhoven, S., Hansen, C., and Shapiro, J. M. (2019). Pre-event Trends in the Panel Event-Study Design. *American Economic Review*, 109(9):3307–3338.

Gallagher, J. (2014). Learning about an Infrequent Event: Evidence from Flood Insurance Take-Up in the United States. *American Economic Journal: Applied Economics*, 6(3):206–233.

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). Public insurance and mortality: Evidence from medicaid implementation. *Journal of Public Economics*, 126(1):216–262.

Goodman-Bacon, A. (2019). Difference-in-Differences with Variation in Treatment Timing. Working paper.

Greenstone, M. and Hanna, R. (2014). Environmental Regulations, Air and Water Pollution, and Infant Mortality in India. *American Economic Review*, 104(10):3038–3072.

Guggenberger, P. (2010). THE IMPACT OF A HAUSMAN PRETEST ON THE ASYMPTOTIC SIZE OF A HYPOTHESIS TEST. *Econometric Theory*, 26(2):369–382.

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.

Kolesar, M. and Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. 108(8):2277–2304.

Kuziemko, I., Meckel, K., and Rossin-Slater, M. (2018). Does Managed Care Widen Infant Health Disparities? Evidence from Texas Medicaid. *American Economic Journal: Economic Policy*, 10(3):255–283.

Lafortune, J., Rothstein, J., and Schanzenbach, D. W. (2018). School Finance Reform and the Distribution of Student Achievement. *American Economic Journal: Applied Economics*, 10(2):1–26.

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.

Lee, J. Y. and Solon, G. (2011). The fragility of estimated effects of unilateral divorce laws on divorce rates. *The B.E. Journal of Economic Analysis & Policy*, 11(1).

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

Lovenheim, M. F. and Willen, A. (2019). The long-run effects of teacher collective bargaining. *American Economic Journal: Economic Policy*, 11(3):292–324.

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

Manski, C. F. and Pepper, J. V. (2018). How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, 100(2):232–244.

Markevich, A. and Zhuravskaya, E. (2018). The Economic Effects of the Abolition of Serfdom: Evidence from the Russian Empire. *American Economic Review*, 108(4-5):1074–1117.

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

Rambachan, A. and Roth, J. (2019). An honest approach to parallel trends.

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.

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.

Tewari, I. (2014). The Distributive Impacts of Financial Development: Evidence from Mortgage Markets during US Bank Branch Deregulation. *American Economic Journal: Applied Economics*, 6(4):175–196.

Ujhelyi, G. (2014). Civil Service Rules and Policy Choices: Evidence from US State Governments. *American Economic Journal: Economic Policy*, 6(2):338–380.

Wolfers, J. (2006). Did unilateral divorce laws raise divorce rates? a reconciliation and new results. *American Economic Review*, 96:1802–1820.

Supplement to the paper

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

For online publication

Jonathan Roth

May 8, 2020

This supplement contains proofs and additional results for the paper "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends." Section A provides proofs for the results in the main text. Section B introduces corrections to parametric approaches that have good properties conditional on surviving a test for pre-trends. Section C states and proves asymptotic results. Section D provides additional simulation results in which the treatment and control group receive stochastic common shocks. Finally, Section E contains additional figures.

A Proofs for Results in the Main Text

This section collects proofs for the results in the main text, as well as some auxiliary lemmas. We begin with a lemma, which will be useful in the following proofs.

Lemma A.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

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

which completes the proof.

□

Proof of Proposition 3.1 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 A.1, and the third and fourth use the definition of $\tilde{\beta}_{post}$, β_{post}, and β_{pre}. Since $\beta_{post} = \tau_{post} + \delta_{post}$ by definition, the result follows. □

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 A.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. Note that if $Y \sim \mathcal{N}(0, \Sigma)$, then $-Y$ is also distributed $\mathcal{N}(0, \Sigma)$. Using this, combined with the fact that $(-Y) \in B$ iff $Y \in B$ by assumption, we have

$$\begin{aligned}
\mathbb{E}[Y \,|\, Y \in B] &= \mathbb{E}[-Y \,|\, (-Y) \in B] \\
&= \mathbb{E}[-Y \,|\, Y \in B] \\
&= -\mathbb{E}[Y \,|\, Y \in B],
\end{aligned}$$

which implies that $\mathbb{E}[Y \,|\, Y \in B] = 0$.

□

Proof of Corollary 3.1 From Proposition 3.1, it suffices to show that $\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] - \beta_{pre} = 0$. However, $\beta_{pre} = 0$ by the assumption of parallel trends, and $\mathbb{E}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B\right] = 0$ by Lemma A.2. □

We now prove a series of Lemmas leading up to the proof of Proposition 3.2.

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

Proof.[16]
Define the function $H : \mathbb{R}^k \to \mathbb{R}$ by

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

for $\phi_\Sigma(x) = (2\pi)^{-\frac{k}{2}} det(\Sigma)^{-\frac{1}{2}} exp(-\frac{1}{2} x' \Sigma^{-1} x)$ 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 $g_2(y, \mu) = 1[y \in B]$. 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.

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

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

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

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

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

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

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

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,

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

Thus, applying Leibniz's rule again along with the product rule,

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

Recall that

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

The first term on the right-hand side of (9) thus becomes

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

Applying the chain-rule, the second term on the right-hand side of (9) becomes

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

Substituting the expressions in the previous two displays back into (9), we have

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

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

Lemma A.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 $c_2 = \rho c_1 > 0$.

□

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

Proof. For any $x \in \mathbb{R}^K$ such that $x_j \leq b_j$ for all j, define $B^X(x) = \{y \in \mathbb{R}^K \,|\, x_j \leq y \leq b_j \text{ for all } j\}$. Let $b = (b_1, \ldots, b_K)$. Note that $B^X(-b)$ is a symmetric rectangular truncation around 0, so from Lemma A.2, we have that $\mathbb{E}\left[Y \,|\, Y \in B^X(-b)\right] = 0$. 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 $g(a) > 0$. By the mean-value theorem, for some $t \in (0, 1)$,

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

By assumption, $(a + b)$ is elementwise greater than 0. It thus suffices to show that all elements of $\nabla g\,(x^t)$ are positive. Without loss of generality, 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

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

where $\phi_\Sigma(y)$ denotes the PDF of a multivariate normal with mean 0 and variance Σ, and the second line uses the quotient rule. 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,

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

It is clear that $\mathbb{E}\left[Y_K\,|\,x_j^t\le Y_j\le b_j,\forall j\right] > x_K^t$, 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 A.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, $\Sigma_{-K,K}\Sigma_{K,K}^{-1} = \rho/\sigma^2$ 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}, and then apply Lemma A.4 to obtain the desired result. To do this, note that 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 $\tilde{\Sigma} = \Sigma_{-K,-K} - \Sigma_{-K,K}\Sigma_{K,K}^{-1}\Sigma_{K,-K}$ 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 A.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 the previous display implies $\mathbb{E}\left[\iota' Y_{-K} \,|\, x_j^t \leq Y_j \leq b_j \text{ for } j < K,\, Y_K = y_K\right] = \iota' h(c y_K \iota)$. 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 A.3; the third line uses Lemma A.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 3.2 From Proposition 3.1, 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 A.4, $\Sigma_{12}\Sigma_{22}^{-1} = c_1\iota'$ 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, observe that $\hat{\beta}_{pre} \in B_{NIS} = \{\hat{\beta}_{pre} : |\hat{\beta}_{pre,j}|/\sqrt{\Sigma_{jj}} \leq c_\alpha \text{ for all } j\}$ 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 A.5.

Proof of Proposition 3.3 Note that since $\hat{\beta}_{post} = \tilde{\beta}_{post} + \Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}$, for any set S,

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

where we use the independence of $\tilde{\beta}_{post}$ and $\hat{\beta}_{pre}$ from Lemma A.1 to obtain that $\mathbb{V}\text{ar}\left[\tilde{\beta}_{post} \,|\, \hat{\beta}_{pre} \in B\right] = \mathbb{V}\text{ar}\left[\tilde{\beta}_{post}\right]$ 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

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

which gives the desired result.

Proof of Proposition 3.4 By Proposition 3.3, it suffices to show that

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

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

B Pre-test Corrected Parametric Approaches

Section 5.1 discusses how, under functional form restrictions about the possible violations of parallel trends, one can obtain estimates and CIs that have good properties from an ex ante sampling perspective. However, even if the imposed functional form restrictions are correct, these parametric specifications will not yield unbiased point estimates or CIs with correct coverage conditional on surviving a pre-test for parallel trends. Given how common these pre-tests are under the status quo, for retrospective analyses of published papers we may wish to obtain estimates with good properties conditional on surviving a pre-test.

In this section I develop methods which, under the same functional form restrictions as standard parametric approaches, deliver median-unbiased estimates and valid CIs *conditional* on passing a pre-test for pre-trends. In particular, I discuss the construction of a median-unbiased estimator and CIs for a scalar parameter of the form $\tau^* = \eta'\beta$, where $\beta = (\beta_{pre}, \beta_{post})$ and η is a vector of the appropriate length. This allows for estimating causal effects under a variety of functional form assumptions about the difference in trends. For instance, in Section 5.1, we showed that under a linearity assumption about the possible differences in trends, we could write $\tau_t = \beta_t - t \times M_T \beta_{pre}$, where M_T is the matrix that projects β_{pre} onto a linear time trend. We could likewise accommodate other types of functional form restrictions by projecting β_{pre} onto different functions of time, e.g. higher-order polynomials or sinusoidal functions.

I begin with a derivation of median-unbiased estimates and CIs that are valid conditional on surviving 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.

B.1 Construction of the corrected estimator and CIs

B.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 $\hat{\beta} \in B$.[17] In general, the conditional distribution of $\eta'\hat{\beta}$ will depend on the full parameter vector β, and we will therefore condition also on a minimal sufficient statistic for the other components of β. The following result extends Theorem 5.2 in Lee et al. (2016), who show the result for the particular case where B is a polyhedron.

Lemma B.1 (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 $\Xi(z) := \{x : \exists \hat{\beta} \in B \text{ s.t. } x = \eta'\hat{\beta} \text{ and } z = (I - c\eta')\hat{\beta}\}$.

Proof of Lemma B.1

[17] 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 $\hat{\beta} \in B = \{(\beta_{pre}, \beta_{post}) \in \mathbb{R}^{K+M} \,|\, \beta_{pre} \in B_{pre}\}$.

Proof. 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 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 $\Xi(z) := \{x : \exists \hat{\beta} \in B \text{ s.t. } x = \eta'\hat{\beta} \text{ and } z = (I - c\eta')\hat{\beta}\}$. □

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). The following result is a restatement of Proposition D.2 of the supplement to Andrews and Kasy (2019); similar results have been obtained recently by Lee et al. (2016) on inference for the LASSO, and Andrews et al. (2018) on inference for "winners".

Proposition B.1 (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 $F^{\Xi(z)}_{x,\eta'\Sigma\eta}(\eta'\hat{\beta}) = 1-\alpha$, for $\Xi(z)$ as defined in Lemma B.1. Then for any $\alpha \in (0,1)$, $\hat{b}_\alpha$ is α-quantile unbiased conditional on $\hat{\beta} \in B$,*

$$P\left(\hat{b}_\alpha(\eta'\hat{\beta}, Z) \leq \eta'\beta \,|\, \hat{\beta} \in B\right) = 1-\alpha.$$

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 in the class of α-quantile-unbiased estimators, in the sense that for any other 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 $\eta'\beta$,

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

Corollary B.1. *Under the conditions of Proposition B.1, conditional on $\hat{\beta} \in B$, $\hat{b}_{0.5}(\eta'\hat{\beta}, Z)$ is a uniformly most-concentrated median-unbiased estimate of $\eta'\beta$, and the interval $\mathcal{C}_{1-\alpha} := [\hat{b}_{\alpha/2}(\eta'\hat{\beta}, Z), \hat{b}_{1-\alpha/2}(\eta'\hat{\beta}, Z)]$ is a $1-\alpha$ level confidence interval for $\eta'\beta$.*

Proof of Corollary B.1

Proof. Median-unbiasedness of $\hat{b}_{0.5}(\eta'\hat{\beta}, Z)$ follows immediately from Proposition B.1. To show that $\mathcal{C}_{1-\alpha}$ controls size, note that $\eta'\beta \notin \mathcal{C}_{1-\alpha}$ only if either $\hat{b}_{\alpha/2}(\eta'\hat{\beta}, Z) > \eta'\beta$ or $\hat{b}_{1-\alpha/2}(\eta'\hat{\beta}, Z) < \eta'\beta$. However, Proposition B.1 implies that each of these events occurs with probability bounded above by $\alpha/2$, and thus $\eta'\beta \notin \mathcal{C}_{1-\alpha}$ with probability bounded above by α. □

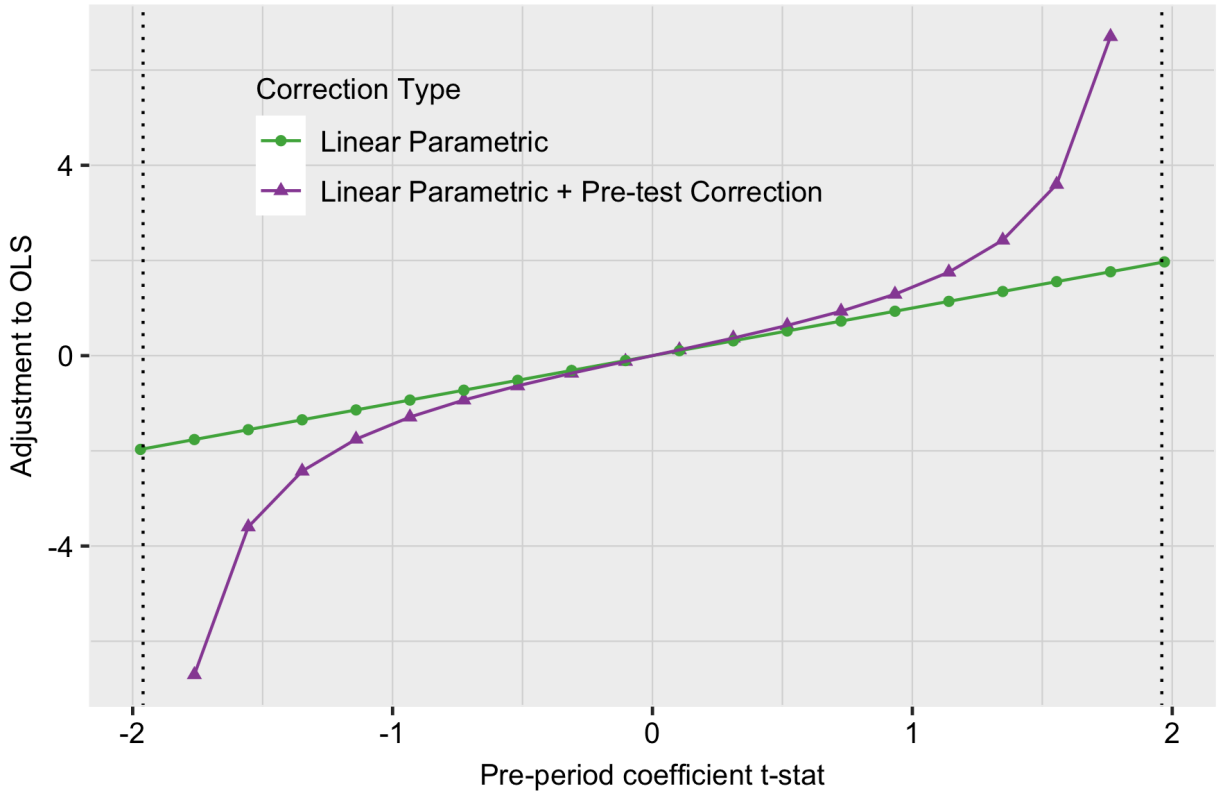
Applying these results in practice requires calculation of the set $\Xi(z)$. In Section B.2, I derive easy-to-calculate formulas for $\Xi(z)$ for the cases where B is based on linear or quadratic restrictions on $\hat{\beta}$, which covers the common cases of tests based on individual or joint significance.

Intuitively, the median-unbiased estimator proposed above chooses the value $\hat{b}_{0.5}$ so that if the parameter of interest $\eta'\beta$ were equal to $\hat{b}_{0.5}$, then the observed value $\eta'\hat{\beta}$ would be at the median of the distribution conditional on passing the pre-test. Figure 5 shows how this pre-test corrected median-unbiased estimator works in the stylized example in Section 2. We treat as the estimand the value of the post-treatment event-study coefficient minus a linear projection of the pre-period trend, $\tau^* = \beta_1 + \beta_{-1}$, which corresponds with τ_1 if we impose that the differential trend is linear ($\delta_1 = -\delta_{-1}$). Figure 5 shows the difference $\hat{\beta}_1 - \hat{b}_{0.5}$ as a function of the pre-period coefficient $\hat{\beta}_{-1}$. It also shows the equivalent difference for the naive linear projection that does not adjust for pre-testing ($\hat{\beta}_1 - (\hat{\beta}_1 + \hat{\beta}_{-1}) = -\hat{\beta}_{-1}$). We see that $\hat{b}_{0.5}$ looks similar to the parametric estimator that does not adjust for pre-testing when $\hat{\beta}_{-1}$ is close to 0. However, for values of $\hat{\beta}_{-1}$ closer to the rejection boundary of ± 1.96, the pre-test corrected estimator makes a larger adjustment to $\hat{\beta}_{-1}$ than the traditional estimator. Intuitively, this is because the naive estimate of the slope of a linear pre-trend will generally be biased towards 0 conditional on passing the pre-test. In order to correct for this, the pre-tested adjusted estimator inflates the estimates of the slope of the pre-trend. It uses a larger inflation factor the closer $\hat{\beta}_{-1}$ is to the inflation boundary, since conditional on passing the pre-test, $\hat{\beta}_{-1}$ is more likely to be near the decision boundary when β_{-1} is large in magnitude.

B.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

Figure 5: Adjustment to conventional estimates using pre-test corrected linear parametric estimates



Note: For an estimator $\hat{\tau}_1$, this figure shows the difference $\hat{\beta}_1 - \hat{\tau}_1$ as a function of $\hat{\beta}_{-1}$, in the stylized model considered in Section 2. The two estimators considered are the naive estimator that adjusts parametrically for a linear trend, $\hat{\tau}_1 = \hat{\beta}_1 + \hat{\beta}_{-1}$, and the adjusted linear parametric estimator $\hat{b}_{0.5}$ that corrects for the pre-test that $\hat{\beta}_{-1}$ is statistically insignificant.

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 Seemingly Unrelated Regressions (SUR), 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 CIs for parameters of the form $\tau^{*,m} = \eta'\beta^m$, conditional on model m being chosen. That is, letting B^*_m denote the set of values for $\hat{\beta}^{stacked}$ such that model m is chosen, we can obtain adjusted estimates with the property that $\mathbb{P}\left(\hat{b}^m_j \leq \tau^{*,m} \,|\, \hat{\beta}^{stacked} \in B^*_m\right) = 0.5$. Conditional coverage of the corrected CIs can be defined analogously.

Implementing these corrected estimates and CIs in practice requires calculation of the set $\Xi(z)$ accounting for the model selection rule. In Section B.2, I show that $\Xi(z)$ can be easily calculated for a variety of model selection rules, including 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.

B.2 Computing Ξ For Common Pre-tests

Applying the corrections discussed above in practice requires computation of the set $\Xi(z)$. I now derive the form of $\Xi(z)$ for polyhedral and quadratic pre-tests, which respectively cover the cases of pre-tests based on individual and joint significance. I then derive the form of $\Xi(z)$ for a variety of model selection criteria.

B.2.1 Calculating $\Xi(z)$ for polyhedral pre-tests

We first consider the case where $B = \{\beta \,|\, A\beta \leq b\}$. Note that the test that no pre-period coefficient is significant can be written in this form, $B_{NIS} = \{\beta \,|\, A^{NIS}\beta \leq b^{NIS}\}$, for $A^{NIS} = \begin{pmatrix} I_{K\times K} & 0_{1\times M} \\ -I_{K\times K} & 0_{1\times M} \end{pmatrix}$ and $b^{NIS} = \begin{pmatrix} c_\alpha \times \sqrt{diag(\Sigma)} \\ c_\alpha \times \sqrt{diag(\Sigma)} \end{pmatrix}$. For the polyhedral case, the form of $\Xi(z)$ follows immediately from the results of Lee et al. (2016).

Lemma B.2 (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 Lemma B.1, is an interval in $\mathbb{R}$, with endpoints $V^-(z)$ and $V^+(z)$ given by:*

$$V^-(z) = \max_{\{j:(Ac)_j<0\}} \frac{b_j - (Az)_j}{(Ac)_j} \tag{16}$$

$$V^+(z) = \min_{\{j:(Ac)_j>0\}} \frac{b_j - (Az)_j}{(Ac)_j}. \tag{17}$$

Additionally, if $\mathbb{P}\left(\hat{\beta} \in B\right) > 0$*, then the conditions for the optimality of the* α*-quantile-unbiased estimator in Proposition B.1 are met.*

B.2.2 Calculating $\Xi(z)$ for quadratic pre-tests

I next 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.

Lemma B.3. *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 Lemma B.1. 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* $\Xi(z) = (-\infty, -\frac{\mathcal{C}}{\mathcal{B}}]$.
6. *If* $\mathcal{A} = 0$*,* $\mathcal{B} < 0$*,* $\Xi(z) = [-\frac{\mathcal{C}}{\mathcal{B}}, \infty)$.
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-unbiased estimator in Proposition B.1 are met.*

B.2.3 Calculating $\Xi(z)$ after model selection

I now discuss the computation of $\Xi(z)$ after selection among a finite number of models, as discussed in Section B.1.2.

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 unions of intervals for $\Xi(z)$.[18] Further, Propositions B.2 and B.3 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 $\hat{\beta}'(A_m - A_{m'})\hat{\beta} \leq 0$ 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.

B.2.4 Proofs for the results on $\Xi(z)$

Proof of Lemma B.2

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 Lemma B.1, $\eta'\hat{\beta} \,|\, Z = z, A\hat{\beta} \leq b$ is truncated normal with truncation points $V^-(z)$ and $V^+(z)$ and untruncated variance $\eta'\Sigma\eta$. The untruncated variance is strictly positive since Σ is positive definite and $\eta \neq 0$, and so the conditional distribution of $\eta'\hat{\beta}$ is 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 $V^-(z) = \eta'\hat{\beta}$.

[18] 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.

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

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

Next, observe that 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 Lemma B.3 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 by solving for the region where the parabola $\mathcal{A}x^2 + \mathcal{B}x + \mathcal{C} \leq 0$ using 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 $\eta'\hat{\beta} \,|\, \hat{\beta} \in B, Z = z$ 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 Lemma B.2, this occurs with probability zero, which completes the proof. □

C Uniform Asymptotic Results

In the main text of the paper, I consider a finite sample normal model for the event-study coefficients. I evaluate the distribution of the event-study estimates conditional on passing a pre-test for the pre-period coefficients. In Appendix B, I derive corrections to parametric methods that have good properties after pre-testing in the context of this model. In this section, I show that these finite-sample results translate to uniform asymptotic results over a large class of data-generating processes in which the probability of passing the pre-test does not go to zero asymptotically, i.e. when the pre-trend is $O(n^{-\frac{1}{2}})$. I focus here on results for polyhedral pre-tests, which include the common pre-test that no pre-period coefficient be individually statistically significant.

C.1 Assumptions

We consider a class of data-generating processes $\mathcal{P}$. Let $\hat{\beta}_n = \sqrt{n}\hat{\beta}$ be the event-study estimates $\hat{\beta} = \begin{pmatrix} \hat{\beta}_{post} \\ \hat{\beta}_{pre} \end{pmatrix}$ scaled by $\sqrt{n}$. Likewise, let $\tau_{P,n} = \sqrt{n}\begin{pmatrix} \tau_{post}(P) \\ 0 \end{pmatrix}$ be the scaled vector of treatment effects under data-generating process $P \in \mathcal{P}$, where we assume there is no true effect of treatment in the pre-periods.

Assumption 2 (Unconditional uniform convergence)**.** *Let BL_1 denote the set of Lipschitz functions which are bounded by 1 in absolute value and have Lipschitz constant bounded by 1. We assume*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{f\in BL_1} \left|\left|\mathbb{E}_P\left[f(\hat{\beta}_n - \tau_{P,n})\right] - \mathbb{E}\left[f(\xi_{P,n})\right]\right|\right| = 0,$$

where $\xi_{P,n} \sim \mathcal{N}(\delta_{P,n}, \Sigma_P)$.

Convergence in distribution is equivalent to convergence in bounded Lipschitz metric (see Theorem 1.12.4 in van der Vaart and Wellner (1996)), so Assumption 2 formalizes the notion of uniform convergence in distribution of $\hat{\beta}_n - \tau_{P,n}$ to a $\mathcal{N}(\delta_{P,n}, \Sigma_P)$ variable under P. Note that we allow δ to depend both on P and the sample size n.

We next assume that we have a uniformly consistent estimator of the variance Σ_P, and that the eigenvalues of Σ_P are bounded above and away from singularity.

Assumption 3 (Consistent estimation of Σ_P)**.** *Our estimator $\hat{\Sigma}$ is uniformly consistent for Σ_P,*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(||\hat{\Sigma}_n - \Sigma_P|| > \epsilon\right) = 0,$$

for all $\epsilon > 0$.

Assumption 4 (Assumptions on Σ_P)**.** *We assume that there exists $\bar{\lambda} > 0$ such that for all $P \in \mathcal{P}$, $\Sigma_P \in \mathcal{S} := \{\Sigma \,|\, 1/\bar{\lambda} \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq \bar{\lambda}\}$, where $\lambda_{min}(A)$ and $\lambda_{max}(A)$ denote the minimal and maximal eigenvalues of a matrix A.*

Next, we assume that the pre-test takes the form of a polyhedral restriction on the vector of pre-period coefficients. Recall that the test that no pre-period coefficient be individually significant can be written in this form.

Assumption 5 (Assumptions on B)**.** *We assume that the conditioning set $B(\Sigma)$ is of the form $B(\Sigma) = \{(\beta_{post}, \beta_{pre}) \,|\, A_{pre}(\Sigma)\beta_{pre} \leq b(\Sigma)\}$ for continuous functions A_{pre} and b. We further assume that for all Σ on an open set containing $\mathcal{S}$, $B(\Sigma)$ is bounded and has non-empty interior, and $A_{pre}(\Sigma)$ has no all-zero rows.*

For ease of notation, it will be useful to define $A(\Sigma) = [0,\, A_{pre}(\Sigma)]$, so that $\beta \in B(\Sigma)$ iff $A(\Sigma)\beta \leq b(\Sigma)$.

C.2 Main uniformity results

Our first result concerns the asymptotic distribution of the event-study coefficients *conditional* on passing the pre-test.

Proposition C.1 (Uniform conditional convergence in distribution)**.** *Under Assumptions 2-5,*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{f\in BL_1} \left|\left| \mathbb{E}_P \left[f(\hat{\beta}_n - \tau_{P,n}) \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n) \right] - \mathbb{E}\left[f(\xi_{P,n}) | \xi_{P,n} \in B(\Sigma_P) \right] \right|\right| \mathbb{P}_P \left(\hat{\beta}_n \in B(\hat{\Sigma}_n) \right) = 0,$$

where $\xi_{P,n} \sim \mathcal{N}\left(\delta_{P,n},\, \Sigma_P\right)$.

Note that if we removed the $\mathbb{P}_P \left(\hat{\beta}_n \in B(\hat{\Sigma}_n) \right)$ term from the statement of Proposition C.1, then the proposition would imply uniform convergence in distribution of $(\hat{\beta}_n - \tau_{P,n}) | \hat{\beta}_n \in B(\hat{\Sigma}_n)$ to $\xi_{P,n} | \xi_{P,n} \in B(\Sigma_P)$. The Proposition thus guarantees such convergence in distribution along any sequence of distributions for which the probability of passing the pre-test is not going to zero.

Although Proposition C.1 gives uniform convergence of the treatment effect estimates conditional on passing the pre-test, it is well known that convergence in distribution need not imply convergence in expectations. Our next result shows that under the additional assumption of asymptotic uniform integrability, we also obtain uniform convergence in expectations, provided that the probability of passing the pre-test is not going to zero.

Proposition C.2 (Uniform convergence of expectations)**.** *Suppose Assumptions 2-5 hold. Let* $\beta_{P,n} = \tau_{P,n} + \delta_{P,n}$*. Assume that* $\hat{\beta}_n - \beta_{P,n}$ *is asymptotically uniformly integrable over the class* $\mathcal{P}$*,*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[||\hat{\beta}_n - \beta_{P,n}|| \cdot 1[||\hat{\beta}_n - \beta_{P,n}|| > M] \right] = 0.$$

Then, for any $\epsilon > 0$*,*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} 1\left[\left|\left| \mathbb{E}_P \left[\hat{\beta}_n - \tau_{P,n} \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n) \right] - \mathbb{E}\left[\xi_{P,n} \,|\, \xi_{P,n} \in B(\Sigma_P) \right] \right|\right| > \epsilon \right] \mathbb{P}_P \left(\hat{\beta}_n \in B(\hat{\Sigma}_n) \right) = 0,$$

where $\xi_{P,n} \sim \mathcal{N}\left(\delta_{P,n},\, \Sigma_P\right)$*.*

Finally, our last main result concerns the asymptotic validity of the pre-test corrected parametric estimator and CIs.

Proposition C.3 (Uniform asymptotic α-quantile unbiasedness)**.** *Let* $\eta \neq 0$*, and consider* $\hat{b}_\alpha(\hat{\beta}_n, \hat{\Sigma}_n)$ *the* α*-quantile-unbiased estimator of* $\eta'\beta$ *conditional on* $\hat{\beta}_n \in B(\hat{\Sigma}_n)$*. Define* $\beta_{P,n} = \tau_{P,n} + \delta_{P,n}$*. Then under Assumptions 2-5,*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \left|\left| \mathbb{P}_P \left(\hat{b}_\alpha(\hat{\beta}_n, \hat{\Sigma}_n) \leq \eta'\beta_{P,n} \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n) \right) - (1-\alpha) \right|\right| \mathbb{P}_P \left(\hat{\beta}_n \in B(\hat{\Sigma}_n) \right) = 0.$$

Proposition C.3 states that the corrected α-quantile-unbiased estimator $\hat{b}_\alpha$ is uniformly α-quantile unbiased along any sequence of distributions such that the limiting probability of passing the pre-test is not going to zero. It follows immediately that under any such sequence $\hat{b}_{0.5}$ is asymptotically median-unbiased and the interval $[\hat{b}_{\alpha/2}, \hat{b}_{1-\alpha/2}]$ is a valid $1-\alpha$ level confidence interval.

Corollary C.1 (Median unbiasedness and coverage of equal-tailed CIs)**.** *Suppose the conditions of Proposition C.3 hold. Then*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \left|\left| \mathbb{P}_P \left(\hat{b}_{0.5}(\hat{\beta}_n, \hat{\Sigma}_n) \leq \eta'\beta_{P,n} \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n) \right) - 0.5 \right|\right| \mathbb{P}_P \left(\hat{\beta}_n \in B(\hat{\Sigma}_n) \right) = 0$$

and, for $\mathcal{C}_{1-\alpha}(\hat{\beta}_n, \hat{\Sigma}_n) = [\hat{b}_{\alpha/2}(\hat{\beta}_n, \hat{\Sigma}_n), \hat{b}_{1-\alpha/2}(\hat{\beta}_n, \hat{\Sigma}_n)]$*,*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \left|\left| \mathbb{P}_P \left(\eta'\beta_{P,n} \in \mathcal{C}_{1-\alpha}(\hat{\beta}_n, \hat{\Sigma}_n) \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n) \right) - (1-\alpha) \right|\right| \mathbb{P}_P \left(\hat{\beta}_n \in B(\hat{\Sigma}_n) \right) = 0.$$

C.3 Proofs for main uniformity results

Proof of Proposition C.1 Towards contradiction, suppose that the proposition is false. Then there exists an increasing sequence of sample sizes n_m and data-generating processes P_{n_m} such that

$$\liminf_{m\to\infty} \sup_{f\in BL_1} \left|\left|\mathbb{E}_{P_n}\left[f(\hat{\beta}_n - \tau_{P_{n_m},n_m}) \,|\, \hat{\beta}_{n_m} \in B(\hat{\Sigma}_{n_m})\right] - \mathbb{E}\left[f(\xi_{P_{n_m},n_m})|\xi \in B(\Sigma_{P_{n_m}})\right]\right|\right| \times \mathbb{P}_{P_{n_m}}\left(\hat{\beta}_{n_m} \in B(\hat{\Sigma}_{n_m})\right) > 0. \tag{18}$$

Since the interval $[0,1]$ is compact, there exists a subsequence of increasing sample sizes, n_q, such that

$$\lim_{q\to\infty} \mathbb{P}_{P_{n_q}}\left(\hat{\beta}_{n_q} \in B(\hat{\Sigma}_{n_q})\right) = p^*,$$

for $p^* \in [0,1]$.

Suppose first that $p^* = 0$. Note that by definition, a function $f \in BL_1$ is bounded in absolute value by 1. It then follows from the triangle inequality that for all $f \in BL_1$,

$$\left|\left|\mathbb{E}_{P_{n_q}}\left[f(\hat{\beta}_{n_q} - \tau_{P_{n_q},n_q}) \,|\, \hat{\beta}_{n_q} \in B(\hat{\Sigma}_{n_q})\right] - \mathbb{E}\left[f(\xi_{P_{n_q},n_q})|\xi_{P_{n_q},n_q} \in B(\Sigma_{P_{n_q}})\right]\right|\right| \leq 2$$

for all q. But this implies that

$$\liminf_{q\to\infty} \sup_{f\in BL_1} \left|\left|\mathbb{E}_{P_{n_q}}\left[f(\hat{\beta}_{n_q} - \tau_{P_{n_q},n_q}) \,|\, \hat{\beta}_{n_q} \in B(\hat{\Sigma}_{n_q})\right] - \mathbb{E}\left[f(\xi_{P_{n_q}})|\xi_{P_{n_q}} \in B(\Sigma_{P_{n_q}})\right]\right|\right| \mathbb{P}_{P_{n_q}}\left(\hat{\beta}_{n_q} \in B(\hat{\Sigma}_{n_q})\right) \leq 2p^* = 0,$$

which contradicts (18).

Now, suppose $p^* > 0$. Note that by Assumption 4, Σ_P falls in the set $\mathcal{S} = \{\Sigma | 1/\bar{\lambda} \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq \bar{\lambda}\}$, which is compact (e.g., in the Frobenius norm). Thus, we can extract a further subsequence of increasing sample sizes, n_r, such that

$$\lim_{r\to\infty} \Sigma_{P_{n_r}} = \Sigma^*,$$

for some $\Sigma^* \in \mathcal{S}$.

Additionally, since $p^* > 0$, Lemma C.4 implies that $\delta^{pre}_{P_{n_r},n_r}$ is bounded, and thus we can extract a further subsequence n_s along which

$$\lim_{s\to\infty} \delta^{pre}_{P_{n_s},n_s} = \delta^{pre,*}.$$

By Lemma C.3, for $\delta^+_{n_s} = \begin{pmatrix} \delta^{post}_{P_{n_s},n_s} \\ 0 \end{pmatrix}$, $\delta^* = \begin{pmatrix} 0 \\ \delta^{pre,*} \end{pmatrix}$, and $\xi^* \sim \mathcal{N}(\delta^*, \Sigma^*)$, we have

$$(\hat{\beta}_{n_s} - \tau_{P,n_s} - \delta^+_{n_s})|\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \xrightarrow{d} \xi^*|\xi^* \in B(\Sigma^*),$$

and

$$(\xi_{P_{n_s}} - \delta^+_{n_s})|\xi_{P_{n_s}} \in B(\Sigma_{P_{n_s}}) \xrightarrow{d} \xi^*|\xi^* \in B(\Sigma^*).$$

Recalling the convergence in distribution is equivalent to convergence in bounded Lipschitz metric, we see that

$$\lim_{s\to\infty} \sup_{f\in BL_1} \left|\left|\mathbb{E}_{P_{n_s}}\left[f(\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta^+_{n_s}) \,|\, \hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s})\right] - \mathbb{E}\left[f(\xi^*)|\xi^* \in B(\Sigma^*)\right]\right|\right| = 0 \quad (19)$$

and

$$\lim_{s\to\infty} \sup_{f\in BL_1} \left|\left|\mathbb{E}\left[f(\xi_{P_{n_s}} - \delta^+_{n_s})|\xi_{P_{n_s}} \in B(\Sigma_{P_{n_s}})\right] - \mathbb{E}\left[f(\xi^*)|\xi^* \in B(\Sigma^*)\right]\right|\right| = 0. \quad (20)$$

Equations (19) and (20) together with the triangle inequality then imply that

$$\lim_{s\to\infty} \sup_{f\in BL_1} \left|\left|\mathbb{E}_{P_{n_s}}\left[f(\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta^+_{n_s}) \,|\, \hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s})\right] - \mathbb{E}\left[f(\xi_{P_{n_s}} - \delta^+_{n_s})|\xi_{P_{n_s}} \in B(\Sigma_{P_{n_s}})\right]\right|\right| = 0.$$

However, BL_1 is closed under horizontal transformation (i.e. $f(x) \in BL_1$ implies $f(x-c) \in BL_1$), and so this implies that

$$\lim_{s\to\infty} \sup_{f\in BL_1} \left|\left|\mathbb{E}_{P_{n_s}}\left[f(\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s}) \,|\, \hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s})\right] - \mathbb{E}\left[f(\xi_{P_{n_s}})|\xi_{P_{n_s}} \in B(\Sigma_{P_{n_s}})\right]\right|\right| = 0,$$

which contradicts (18). □

Proof of Proposition C.2 Towards contradiction, suppose the proposition is false. Then there exists an increasing sequence of sample sizes n_m and data-generating processes P_{n_m} such that for some $\epsilon > 0$,

$$\liminf_{m\to\infty} 1\left[\left|\left|\mathbb{E}\left[\hat{\beta}_{n_m} - \tau_{P_{n_m},n_m} \,|\, \hat{\beta}_{n_m} \in B(\hat{\Sigma}_{n_m})\right] - \mathbb{E}\left[\xi_{P_{n_m}} \,|\, \xi_{P_{n_m}} \in B(\Sigma_{P_{n_m}})\right]\right|\right| > \epsilon\right] \times \mathbb{P}_{P_{n_m}}\left(\hat{\beta}_{n_m} \in B(\hat{\Sigma}_{n_m})\right) > 0. \tag{21}$$

Since the interval $[0,1]$ is compact, we can extract a subsequence of increasing sample sizes, n_q, along which

$$\lim_{q\to\infty} \mathbb{P}_{P_{n_q}}\left(\hat{\beta}_{n_q} \in B(\hat{\Sigma}_{n_q})\right) = p^*$$

for $p^* \in [0,1]$.

First, suppose $p^* = 0$. Since the indicator function is bounded by 1,

$$\liminf_{s\to\infty} 1\left[\left|\left|\mathbb{E}\left[\hat{\beta}_{n_q} - \tau_{P_{n_q},n_q} \,|\, \hat{\beta}_{n_q} \in B(\hat{\Sigma}_{n_q})\right] - \mathbb{E}\left[\xi_{P_{n_q}} \,|\, \xi_{P_{n_q}} \in B(\Sigma_{P_{n_q}})\right]\right|\right| > \epsilon\right] \mathbb{P}_{P_{n_q}}\left(\hat{\beta}_{n_q} \in B(\hat{\Sigma}_{n_q})\right) \leq \liminf_{s\to\infty} \mathbb{P}_{P_{n_q}}\left(\hat{\beta}_{n_q} \in B(\hat{\Sigma}_{n_q})\right) = p^* = 0,$$

which contradicts (21).

Now, suppose $p^* > 0$. As argued in the proof to Proposition C.1, we can iteratively extract subsequences to obtain a subsequence, n_s, along which

$$\begin{aligned}
&\lim_{s\to\infty} \Sigma_{P_{n_s}} = \Sigma^*, \\
&\lim_{s\to\infty} \delta^{pre}_{P_{n_s},n_s} = \delta^{pre,*}, \\
&\lim_{s\to\infty} \mathbb{P}_{P_{n_s}}\left(\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s})\right) = p^* > 0,
\end{aligned}$$

where $\Sigma^* \in \mathcal{S}$.

Let $\delta^-_{n_s} = \begin{pmatrix} 0 \\ \delta^{pre}_{P_{n_s},n_s} \end{pmatrix}$ and $\delta^* = \begin{pmatrix} 0 \\ \delta^{pre,*} \end{pmatrix}$ be the vectors with zeros for the post-period coefficients and $\delta^{pre}_{P_{n_s},n_s}$ and $\delta^{pre,*}$, respectively, for the pre-period coefficients. Similarly, let $\delta^+_{n_s} = \begin{pmatrix} \delta^{post}_{P_{n_s},n_s} \\ 0 \end{pmatrix}$ be the vector with zeros for the pre-period coefficients and $\delta^{post}_{P_{n_s},n_s}$ for the

post-period coefficients. From Lemma C.3, $(\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta^+_{n_s})|\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \xrightarrow{d} \xi^*|\xi^* \in B(\Sigma^*)$, for $\xi^* \sim \mathcal{N}(\delta^*, \Sigma^*)$.

Additionally, from uniform integrability, we have

$$\lim_{M\to\infty} \limsup_{s\to\infty} \mathbb{E}_{P_{n_s}} \left[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| \cdot 1[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| > M] \right] = 0.$$

Observe that

$$\begin{aligned}
&\mathbb{E}_{P_{n_s}} \left[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| \cdot 1[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| > M] \right] = \\
&\mathbb{E}_{P_{n_s}} \left[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| \cdot 1[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| > M] \,|\, \hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right] \cdot \mathbb{P}_{P_{n_s}} \left(\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right) + \\
&\mathbb{E}_{P_{n_s}} \left[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| \cdot 1[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| > M] \,|\, \hat{\beta}_{n_s} \notin B(\hat{\Sigma}_{n_s}) \right] \cdot \mathbb{P}_{P_{n_s}} \left(\hat{\beta}_{n_s} \notin B(\hat{\Sigma}_{n_s}) \right) \geq \\
&\mathbb{E}_{P_{n_s}} \left[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| \cdot 1[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| > M] \,|\, \hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right] \cdot \mathbb{P}_{P_{n_s}} \left(\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right),
\end{aligned}$$

and hence

$$\lim_{M\to\infty} \limsup_{s\to\infty} \mathbb{E}_{P_{n_s}} \left[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| \cdot 1[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| > M] \,|\, \hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right] \cdot \mathbb{P}_{P_{n_s}} \left(\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right) = 0.$$

Further, since $\mathbb{P}_{P_{n_s}} \left(\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right) \to p^* > 0$, it follows that

$$\lim_{M\to\infty} \limsup_{s\to\infty} \mathbb{E}_{P_{n_s}} \left[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| \cdot 1[||\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}|| > M] \,|\, \hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right] = 0,$$

so $\hat{\beta}_{n_s} - \beta_{P_{n_s},n_s}$ is uniformly asymptotically integrable conditional on $\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s})$. Note that $\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta^+_{n_s} = \hat{\beta}_{n_s} - \beta_{P_{n_s},n_s} + \delta^-_{n_s}$, and $\delta^-_{n_s} \to \delta^*$ as $s \to \infty$. It then follows from Lemma C.6 that $\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta^+_{n_s}$ is uniformly asymptotically integrable conditional on $\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s})$.

Convergence in distribution along with uniform asymptotic integrability implies convergence in expectation (see Theorem 2.20 in van der Vaart (2000)), and thus

$$\lim_{s\to\infty} \left|\left| \mathbb{E}_{P_{n_s}} \left[\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta^+_{n_s} \,|\, \hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right] - \mathbb{E}\left[\xi^* \,|\, \xi^* \in B(\Sigma^*) \right] \right|\right| = 0.$$

Likewise, Lemma C.5 gives that

$$\lim_{s\to\infty} \left|\left| \mathbb{E}\left[\xi_{P_{n_s}} - \delta^+_{n_s} \,|\, \xi_{P_{n_s}} \in B(\Sigma_{P_{n_s}}) \right] - \mathbb{E}\left[\xi^* \,|\, \xi^* \in B(\Sigma^*) \right] \right|\right| = 0.$$

It then follows from the triangle inequality that

$$\lim_{s\to\infty}\left|\left|\mathbb{E}_{P_{n_s}}\left[\hat{\beta}_{n_s}-\tau_{P_{n_s},n_s}-\delta_{n_s}^{+}\,|\,\hat{\beta}_{n_s}\in B(\hat{\Sigma}_{n_s})\right]-\mathbb{E}\left[\xi_{P_{n_s}}-\delta_{n_s}^{+}\,|\,\xi_{P_{n_s}}\in B(\Sigma_{P_{n_s}})\right]\right|\right|=0.$$

Cancelling the $\delta_{n_s}^{+}$ terms gives

$$\lim_{s\to\infty}\left|\left|\mathbb{E}_{P_{n_s}}\left[\hat{\beta}_{n_s}-\tau_{n_s,P_{n_s}}\,|\,\hat{\beta}_{n_s}\in B(\hat{\Sigma}_{n_s})\right]-\mathbb{E}\left[\xi_{P_{n_s}}\,|\,\xi_{P_{n_s}}\in B(\Sigma_{P_{n_s}})\right]\right|\right|=0,$$

which contradicts (21). □

Proof of Proposition C.3 Towards contradiction, suppose that the proposition is false. Then there exists an increasing sequence of sample sizes n_m and data-generating processes P_{n_m} such that

$$\liminf_{m\to\infty}\left|\left|\mathbb{P}_{P_{n_m}}\left(\hat{b}_\alpha(\hat{\beta}_{n_m},\hat{\Sigma}_{n_m})\le\eta'\beta_{P_{n_m},n_m}\,|\,\hat{\beta}_{n_m}\in B(\hat{\Sigma}_{n_m})\right)-(1-\alpha)\right|\right|\mathbb{P}_{P_{n_m}}\left(\hat{\beta}_{n_m}\in B(\hat{\Sigma}_{n_m})\right)>0. \tag{22}$$

Since the interval $[0,1]$ is compact, there exists a subsequence of increasing sample sizes, n_q, such that

$$\lim_{q\to\infty}\mathbb{P}_{P_{n_q}}\left(\hat{\beta}_{n_q}\in B(\hat{\Sigma}_{n_q})\right)=p^*,$$

for $p^*\in[0,1]$.

First, suppose $p^*=0$. Note that

$$\left|\left|\mathbb{P}_{P_{n_q}}\left(\hat{b}_\alpha(\hat{\beta}_{n_q},\hat{\Sigma}_{n_q})\le\eta'\beta_{n_q,P_{n_q}}\,|\,\hat{\beta}_{n_q}\in B(\hat{\Sigma}_{n_q})\right)-(1-\alpha)\right|\right|\le 1,$$

and hence

$$\liminf_{q\to\infty}\left|\left|\mathbb{P}_{P_{n_q}}\left(\hat{b}_\alpha(\hat{\beta}_{n_q},\hat{\Sigma}_{n_q})\le\eta'\beta_{P_{n_q},n_q}\,|\,\hat{\beta}_{n_q}\in B(\hat{\Sigma}_{n_q})\right)-(1-\alpha)\right|\right|\mathbb{P}_{P_{n_q}}\left(\hat{\beta}_{n_q}\in B(\hat{\Sigma}_{n_q})\right)\le p^*=0,$$

which contradicts (22).

Next, suppose $p^*>0$. As argued in the proof to Proposition C.1, we can extract a subsequence of increasing sample sizes, n_s, such that

$$\lim_{s\to\infty} \Sigma_{P_{n_s}} = \Sigma^*, \tag{23}$$

$$\lim_{s\to\infty} \delta^{pre}_{P_{n_s},n_s} = \delta^{pre,*}, \tag{24}$$

$$\lim_{s\to\infty} \mathbb{P}_{P_{n_s}} \left(\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right) = p^*, \tag{25}$$

where $\Sigma^* \in \mathcal{S}$ and $p^* \in [0,1]$.

We wish to obtain a contradiction of (22) by showing that

$$\lim_{s\to\infty} \left|\left| \mathbb{P}_{P_{n_s}} \left(\hat{b}_\alpha(\hat{\beta}_{n_s}, \hat{\Sigma}_{n_s}) \leq \eta' \beta_{P_{n_s},n_s} \,|\, \hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \right) - (1-\alpha) \right|\right| = 0.$$

Let $\delta^+_{n_s} = \begin{pmatrix} \delta^{post}_{P_{n_s},n_s} \\ 0 \end{pmatrix}$, $\delta^-_{n_s} = \begin{pmatrix} 0 \\ \delta^{pre}_{P_{n_s},n_s} \end{pmatrix}$, and $\delta^* = \begin{pmatrix} 0 \\ \delta^{pre,*} \end{pmatrix}$. From Lemma C.7, it suffices to show that, for $\hat{\beta}^*_{n_s} = \hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta^+_{n_s}$,

$$\lim_{s\to\infty} \left|\left| \mathbb{P}_{P_{n_s}} \left(\hat{b}_\alpha(\hat{\beta}^*_{n_s}, \hat{\Sigma}_{n_s}) \leq \eta' \delta^-_{n_s} \,|\, \hat{\beta}^*_{n_s} \in B(\hat{\Sigma}_{n_s}) \right) - (1-\alpha) \right|\right| = 0.$$

Further, Lemma C.8 implies that this is equivalent to:

$$\lim_{s\to\infty} \left|\left| \mathbb{P}_{P_{n_s}} \left(g(\hat{\beta}^*_{n_s}, \hat{\Sigma}_{n_s}, \delta^-_{n_s}) \leq 1-\alpha \,|\, \hat{\beta}^*_{n_s} \in B(\hat{\Sigma}_{n_s}) \right) - (1-\alpha) \right|\right| = 0,$$

for g as defined in Lemma C.8.

Note that by construction, for $\xi^* \sim \mathcal{N}(\delta^*, \Sigma^*)$,

$$g(\xi^*, \Sigma^*, \delta^*) \,|\, \xi^* \in B(\Sigma^*) \sim U[0,1].$$

Thus,

$$\mathbb{P}\left(g(\xi^*, \Sigma^*, \delta^*) \leq 1-\alpha \,|\, \xi^* \in B(\Sigma^*) \right) = 1-\alpha,$$

and the distribution of $g(\xi^*, \Sigma^*, \delta^*) \,|\, \xi^* \in B(\Sigma^*)$ is continuous at $1-\alpha$. Additionally, Lemma C.3, along with (23) and (24), imply that

$$(\hat{\beta}^*_{n_s}, \hat{\Sigma}_{n_s}, \delta^-_{n_s}) \,|\, \hat{\beta}^*_{n_s} \in B(\hat{\Sigma}_{n_s}) \xrightarrow{d} (\xi^*, \Sigma^*, \delta^*) \,|\, \xi^* \in B(\Sigma^*).$$

Since Lemma C.12 gives that the function g is continuous for almost every $(\xi^*, \Sigma^*, \delta^*)$, conditional on $\xi^* \in B(\Sigma^*)$, the result then follows from the Continuous Mapping Theorem. □

Proof of Corollary C.1

Proof. The result for $\hat{b}_{0.5}$ is immediate from Proposition C.3. To show the second result, note that

$$\mathbb{P}_P\left(\eta'\beta_{P,n} \notin \mathcal{C}_{1-\alpha}(\hat{\beta}_n, \hat{\Sigma}_n) \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n)\right) =$$
$$\mathbb{P}_P\left(\hat{b}_{\alpha/2}(\hat{\beta}_n, \hat{\Sigma}_n) > \eta'\beta_{P,n} \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n)\right) + \mathbb{P}_P\left(\hat{b}_{1-\alpha/2}(\hat{\beta}_n, \hat{\Sigma}_n) < \eta'\beta_{P,n} \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n)\right),$$

since $\eta'\beta_{P,n}$ falls outside of $\mathcal{C}_{1-\alpha}$ only if it is greater than the upper bound or less than the lower bound, and both of these events cannot occur simultaneously. Applying the result in the previous display along with the triangle inequality and the fact that for any event E, $\mathbb{P}(E) = 1 - \mathbb{P}(E^c)$, we obtain that

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \left|\left|\mathbb{P}_P\left(\eta'\beta_{P,n} \in \mathcal{C}_{1-\alpha}(\hat{\beta}_n, \hat{\Sigma}_n) \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n)\right) - (1-\alpha)\right|\right| \mathbb{P}_P\left(\hat{\beta}_n \in B(\hat{\Sigma}_n)\right) \leq$$
$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \left|\left|\mathbb{P}_P\left(\hat{b}_{1-\alpha/2}(\hat{\beta}_n, \hat{\Sigma}_n) \leq \eta'\beta_{P,n} \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n)\right) - \alpha/2\right|\right| \mathbb{P}_P\left(\hat{\beta}_n \in B(\hat{\Sigma}_n)\right) +$$
$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \left|\left|\mathbb{P}_P\left(\hat{b}_{\alpha/2}(\hat{\beta}_n, \hat{\Sigma}_n) \leq \eta'\beta_{P,n} \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n)\right) - (1-\alpha/2)\right|\right| \mathbb{P}_P\left(\hat{\beta}_n \in B(\hat{\Sigma}_n)\right) +$$
$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \left|\left|\mathbb{P}_P\left(\hat{b}_{1-\alpha/2}(\hat{\beta}_n, \hat{\Sigma}_n) = \eta'\beta_{P,n} \,|\, \hat{\beta}_n \in B(\hat{\Sigma}_n)\right)\right|\right| \mathbb{P}_P\left(\hat{\beta}_n \in B(\hat{\Sigma}_n)\right)$$

The first two terms on the right-hand side of the previous display converge to 0 by Proposition C.3. That the final term is 0 can be shown using an argument analogous to that in the proof of Proposition C.3. Specifically, using the notation from the proof of Proposition C.3, note that $\hat{b}_\alpha(\hat{\beta}_{n_s}, \hat{\Sigma}_{n_s}) = \eta'\beta_{P_{n_s},n_s}$ iff $g(\hat{\beta}^*_{n_s}, \hat{\Sigma}_{n_s}, \delta^-_{n_s}) = 1-\alpha$. However, we show in the proof to Proposition C.3 that for $\xi^* \sim \mathcal{N}(\delta^*, \Sigma^*)$, $g(\xi^*, \Sigma^*\delta^*)|\xi^* \in B(\Sigma^*)$ is uniformly distributed, and thus equal to $1-\alpha$ with probability 0. The desired result then follows from an application of the continuous mapping theorem as in the proof to Proposition C.3.

□

C.4 Auxiliary lemmas and proofs

Lemma C.1. *Suppose* $(\xi_n, \Sigma_n) \xrightarrow{d} (\xi^*, \Sigma^*)$, *for* $\xi^* \sim \mathcal{N}(\delta^*, \Sigma^*)$ *and* $\Sigma^* \in \mathcal{S}$. *Then, if* B *satisfies Assumption 5,*

$$\mathbb{P}_{P_n}(\xi_n \in B(\Sigma_n)) \longrightarrow \mathbb{P}(\xi^* \in B(\Sigma^*)).$$

Proof. By definition, $\xi_n \in B(\Sigma_n)$ iff $A(\Sigma_n)\xi_n \leq b(\Sigma_n)$. Now, consider the function

$$h(\xi, \Sigma) = 1[A(\Sigma)\xi \leq b(\Sigma)].$$

Note that since $A(\cdot)$ and $b(\cdot)$ are continuous by Assumption 5, h is continuous at all (ξ, Σ) such that for all j, $(A(\Sigma)\xi)_j \neq b(\Sigma)_j$. However, the jth element of $A(\Sigma^*)\xi^*$ is normally distributed with variance $A(\Sigma^*)_{(j,\cdot)}\Sigma^* A(\Sigma^*)'_{(j,\cdot)}$, where $X_{(j,\cdot)}$ denotes the jth row of a matrix X. Since $A(\Sigma^*)$ has no non-zero rows by Assumption 5, and $\Sigma^* \in \mathcal{S}$ implies that Σ^* is positive definite, $A(\Sigma^*)_{(j,\cdot)}\Sigma^* A(\Sigma^*)'_{(j,\cdot)} > 0$. This implies that for each j, $(A(\Sigma^*)\xi^*)_j = b(\Sigma^*)_j$ with probability zero, and hence $(A(\Sigma^*)\xi^*)_j \neq b(\Sigma^*)_j$ for all j with probability 1. Thus, h is continuous at (ξ^*, Σ^*) for almost every ξ.

Since $(\xi_n, \Sigma_n) \xrightarrow{d} (\xi^*, \Sigma^*)$, the Continuous Mapping Theorem gives that $1[A(\Sigma_n)\xi_n \leq b(\Sigma_n)] \xrightarrow{d} 1[A(\Sigma^*)\xi^* \leq b(\Sigma^*)]$. Since the indicator functions are bounded, it follows that

$$\mathbb{P}\left(\xi_n \in B(\Sigma_n)\right) = \mathbb{E}\left[1[A(\Sigma_n)\xi_n \leq b(\Sigma_n)]\right] \longrightarrow \mathbb{E}\left[1[A(\Sigma^*)\xi^* \leq b(\Sigma^*)]\right] = \mathbb{P}\left(\xi^* \in B(\Sigma^*)\right),$$

which completes the proof. □

Lemma C.2. *Suppose that* $(\xi_n, \Sigma_n) \xrightarrow{d} (\xi^*, \Sigma^*)$*, for* $\xi^* \sim \mathcal{N}(\delta^*, \Sigma^*)$ *and* $\Sigma^* \in \mathcal{S}$*. Suppose further that* $\mathbb{P}(\xi^* \in B(\Sigma^*)) = p^* > 0$ *for* $B(\Sigma)$ *satisfying Assumption 5. Then*

$$\xi_n \,|\, \xi_n \in B(\Sigma_n) \xrightarrow{d} \xi^* \,|\, \xi^* \in B(\Sigma^*).$$

Proof. By the Portmanteau Lemma (see Lemma 2.2. in van der Vaart (2000)),

$$\xi_n \,|\, \xi_n \in B(\Sigma_n) \xrightarrow{d} \xi^* \,|\, \xi^* \in B(\Sigma^*)$$

iff $\mathbb{E}\left[f(\xi_n) \,|\, \xi_n \in B(\Sigma_n)\right] \longrightarrow \mathbb{E}\left[f(\xi^*) \,|\, \xi^* \in B(\Sigma^*)\right]$ for all bounded, continuous functions f.

Let f be a bounded, continuous function. Since $(\xi_n, \Sigma_n) \xrightarrow{d} (\xi^*, \Sigma^*)$, the Continuous Mapping Theorem together with the Dominated Convergence Theorem imply that $\mathbb{E}\left[g(\xi_n, \Sigma_n)\right] \xrightarrow{p} \mathbb{E}\left[g(\xi^*, \Sigma^*)\right]$ for any bounded function g that is continuous for almost every (ξ^*, Σ^*). It follows that

$$\mathbb{E}\left[f(\xi_n) \cdot 1\left[\xi_n \in B(\Sigma_n)\right]\right] \longrightarrow \mathbb{E}\left[f\left(\xi^*\right) \cdot 1\left[\xi^* \in B(\Sigma^*)\right]\right],$$

where we use the fact that the function $1[\xi \in B(\Sigma)]$ is continuous at (ξ^*, Σ^*) for almost every ξ^*, as shown in the proof to Lemma C.1, and that the product of bounded and continuous functions is bounded and continuous. Additionally, by Lemma C.1, we have that

$$\mathbb{P}\left(\xi_n \in B(\xi_n)\right) \longrightarrow \mathbb{P}\left(\xi^* \in B(\Sigma^*)\right) = p^* > 0.$$

We can thus apply the Continuous Mapping Theorem to obtain

$$\frac{\mathbb{E}\left[f(\xi_n) \cdot 1\left[\xi_n \in B(\Sigma_n)\right]\right]}{\mathbb{P}\left(\xi_n \in B(\Sigma_n)\right)} \longrightarrow \frac{\mathbb{E}\left[f\left(\xi^*\right) \cdot 1\left[\xi^* \in B(\Sigma^*)\right]\right]}{\mathbb{P}\left(\xi^* \in B(\Sigma^*)\right)},$$

which by the definition of the conditional expectation, implies

$$\mathbb{E}\left[f(\xi_n) \,|\, \xi_n \in B(\Sigma_n)\right] \longrightarrow \mathbb{E}\left[f(\xi^*) \,|\, \xi^* \in B(\Sigma^*)\right],$$

as needed. □

Lemma C.3. *Suppose Assumptions 2-5 hold, and n_s is an increasing sequence of sample sizes such that*

$$\begin{aligned}
&\lim_{s\to\infty} \Sigma_{P_{n_s}} = \Sigma^*, \\
&\lim_{s\to\infty} \delta^{pre}_{P_{n_s},n_s} = \delta^{pre,*}, \\
&\lim_{s\to\infty} \mathbb{P}_{P_{n_s}}\left(\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s})\right) = p^* > 0
\end{aligned}$$

for $\Sigma^* \in \mathcal{S}$. Let $\delta^+_{n_s} = \begin{pmatrix} \delta^{post}_{P_{n_s},n_s} \\ 0 \end{pmatrix}$ be the vector with elements corresponding with $\delta_{P_{n_s},n_s}$ for the post-period coefficients, and zeros for the pre-period coefficients. Likewise, let $\delta^* = \begin{pmatrix} 0 \\ \delta^{pre,*} \end{pmatrix}$ be the vector with zeros for the post-period coefficients and $\delta^{pre,*}$ for the pre-period coefficients. Then

$$(\hat{\beta}_{n_s} - \tau_{P,n_s} - \delta^+_{n_s}) | \hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s}) \xrightarrow{d} \xi^* | \xi^* \in B(\Sigma^*)$$

and

$$\left(\xi_{P_{n_s},n_s} - \delta^+_{n_s}\right) | \, \xi_{P_{n_s},n_s} \in B(\Sigma_{P_{n_s}}) \xrightarrow{d} \xi^* | \xi^* \in B(\Sigma^*),$$

for $\xi^* \sim \mathcal{N}\left(\delta^*, \Sigma^*\right)$.

Proof. By assumption, $\xi_{P_{n_s}} \sim \mathcal{N}\left(\delta_{P_{n_s}}, \Sigma_{P_{n_s}}\right)$, and thus $\xi_{P_{n_s}} - \delta^+_{n_s} \sim \mathcal{N}\left(\delta^-_{n_s}, \Sigma_{P_{n_s}}\right)$. Since by

construction $\delta_{n_s}^- \longrightarrow \delta^*$ and $\Sigma_{P_{n_s}} \longrightarrow \Sigma^*$, it follows that $\xi_{P_{n_s}} - \delta_{n_s}^+ \xrightarrow{d} \xi^*$, for $\xi^* \sim \mathcal{N}(\delta^*, \Sigma^*)$. Convergence in distribution is equivalent to convergence in bounded Lipschitz metric, so

$$\lim_{s\to\infty} \sup_{f\in BL_1} \left|\left|\mathbb{E}\left[f(\xi_{P_{n_s}} - \delta_{n_s}^+)\right] - \mathbb{E}\left[f(\xi^*)\right]\right|\right| = 0. \tag{26}$$

Additionally, Assumption 2 gives that

$$\lim_{s\to\infty} \sup_{f\in BL_1} \left|\left|\mathbb{E}_{P_{n_s}}\left[f(\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s})\right] - \mathbb{E}\left[f(\xi_{P_{n_s}})\right]\right|\right| = 0.$$

Since the class of BL_1 functions is closed under horizontal transformations, it follows that

$$\lim_{s\to\infty} \sup_{f\in BL_1} \left|\left|\mathbb{E}_{P_{n_s}}\left[f(\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta_{n_s}^+)\right] - \mathbb{E}\left[f(\xi_{P_{n_s}} - \delta_{n_s}^+)\right]\right|\right| = 0. \tag{27}$$

Equations (26) and (27), together with the triangle inequality, imply that

$$\lim_{s\to\infty} \sup_{f\in BL_1} \left|\left|\mathbb{E}_{P_{n_s}}\left[f(\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta_{n_s}^+)\right] - \mathbb{E}\left[f(\xi^*)\right]\right|\right| = 0, \tag{28}$$

or equivalently, $(\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta_{n_s}^+) \xrightarrow{d} \xi^*$. By Assumption 5, the pre-test is invariant to shifts that only affect the post-period coefficients, and so $\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s})$ iff $(\hat{\beta}_{n_s} - \tau_{n_s,P_{n_s}} - \delta_{n_s}^+) \in B(\hat{\Sigma}_{n_s})$. Lemma C.1 thus implies that $\lim_{s\to\infty} \mathbb{P}_{P_{n_s}}\left(\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s})\right) = \mathbb{P}\left(\xi^* \in B(\Sigma^*)\right)$, and hence $\mathbb{P}\left(\xi^* \in B(\Sigma^*)\right) = p^* > 0$. We have thus shown that $(\hat{\beta}_{n_s} - \tau_{P_{n_s},n_s} - \delta_{n_s}^+, \hat{\Sigma}_{n_s}) \xrightarrow{d} (\xi^*, \Sigma^*)$, $(\xi_{P_{n_s}} - \delta_{n_s}^+, \Sigma_{P_{n_s}}) \xrightarrow{d} (\xi^*, \Sigma^*)$, and $\mathbb{P}\left(\xi^* \in B(\Sigma^*)\right) > 0$. The result then follows immediately from Lemma C.2.

□

Lemma C.4. *Suppose that Assumptions 2-5 hold. Then for any increasing sequence of sample sizes n_q and corresponding data-generation processes P_{n_q} such that*

$$\lim_{q\to\infty} ||\delta_{P_{n_q},n_q}^{pre}|| = \infty,$$

we have

$$\lim_{q\to\infty} \mathbb{P}_{P_{n_q}}\left(\hat{\beta}_{n_q} \in B(\hat{\Sigma}_{n_q})\right) = 0.$$

Proof. Towards contradiction, suppose that there exists a sequence n_q such that

$$\lim_{q\to\infty} ||\delta^{pre}_{P_{n_q},n_q}|| = \infty,$$

and

$$\liminf_{q\to\infty} \mathbb{P}_{P_{n_q}}\left(\hat{\beta}_{n_q} \in B(\hat{\Sigma}_{n_q})\right) > 0. \tag{29}$$

Since $\mathcal{S}$ is compact, we can extract a subsequence n_r along which $\Sigma_{P_{n_r}} \to \Sigma^*$ for some $\Sigma^* \in \mathcal{S}$. Assumption 3 then implies that $\hat{\Sigma}_{n_r} \xrightarrow{p} \Sigma^*$.

By Assumption 5, $B_{pre}(\Sigma)$ is bounded for every Σ. Let $\tilde{M}(\Sigma) = \sup_{\beta_{pre}\in B_{pre}(\Sigma)} ||\beta_{pre}||$. Assumption 5 implies that $B_{pre}(\Sigma)$ is a compact-valued continuous correspondence, and so $\tilde{M}(\Sigma)$ is a continuous function by the theorem of the maximum. It follows that for any Σ in a sufficiently small neighborhood of Σ^*, $\tilde{M}(\Sigma) \leq \tilde{M}(\Sigma^*) + 1 =: \bar{M}$. Since $\hat{\Sigma}_{n_r} \xrightarrow{p} \Sigma^*$, it follows that $\tilde{M}(\hat{\Sigma}_{n_r}) \to_p \tilde{M}(\Sigma^*)$, and thus for r sufficiently large, $\tilde{M}(\hat{\Sigma}_{n_r}) \leq \bar{M}$ with probability 1. Thus, for r sufficiently large, $\mathbb{P}_{P_{n_r}}\left(\hat{\beta}_{n_r} \in B(\hat{\Sigma}_{n_r})\right) \leq \mathbb{P}_{P_{n_r}}\left(\hat{\beta}_{n_r} \in B_{\bar{M}}\right)$, where $B_{\bar{M}} = \{(\beta_{post}, \beta_{pre}) \,|\, ||\beta_{pre}|| \leq \bar{M}\}$. It follows that

$$\begin{aligned}\liminf_{r\to\infty} \mathbb{P}_{P_{n_r}}\left(\hat{\beta}_{n_r} \in B(\hat{\Sigma}_{n_r})\right) &\leq \liminf_{r\to\infty} \mathbb{P}_{P_{n_r}}\left(\hat{\beta}_{n_r} \in B_{\bar{M}}\right) \\ &= 1 - \limsup_{r\to\infty} \mathbb{P}_{P_{n_r}}\left(\hat{\beta}_{n_r} \in B^c_{\bar{M}}\right).\end{aligned}$$

We now show that $\limsup_{r\to\infty} \mathbb{P}_{P_{n_r}}\left(\hat{\beta}_{n_r} \in B^c_{\bar{M}}\right) = 1$, which along with the display above implies that $\liminf_{r\to\infty} \mathbb{P}_{P_{n_r}}\left(\hat{\beta}_{n_r} \in B(\hat{\Sigma}_{n_r})\right) = 0$, contradicting (29).

Consider the function $h(\beta) = \min(d(\beta, B_{\bar{M}}), 1)$, where for a set S we define $d(\beta, S) = \inf_{\tilde{\beta}\in S} ||\beta - \tilde{\beta}||$. It is easily verified that $h \in BL_1$, and that $h(\beta) \leq 1[\beta \in B^c_{\bar{M}}]$ for all β. Thus,

$$\limsup_{r\to\infty} \mathbb{P}_{P_{n_r}}\left(\hat{\beta}_{n_r} \in B^c_{\bar{M}}\right) \geq \limsup_{r\to\infty} \mathbb{E}_{P_{n_r}}\left[h(\hat{\beta}_{n_r})\right]. \tag{30}$$

Note that $d(\hat{\beta}, B_{\bar{M}})$ depends only on the components of $\hat{\beta}$ corresponding with the pre-period, and thus $h(\hat{\beta}) = h(\hat{\beta} - \tau)$ for any value $\tau = \begin{pmatrix} \tau_{post} \\ 0 \end{pmatrix}$ that has zeros in the positions corresponding with β_{pre}. This, along with Assumption 2, implies that

$$\lim_{r\to\infty} \left|\left|\mathbb{E}_{P_{n_r}}\left[h(\hat{\beta}_{n_r})\right] - \mathbb{E}\left[h(\xi_{P_{n_r},n_r})\right]\right|\right| = 0.$$

Using the triangle inequality and the fact that h is a non-negative function, we have

$$\mathbb{E}_{P_{n_r}}\left[h(\hat{\beta}_{n_r})\right] \geq \mathbb{E}\left[h(\xi_{P_{n_r},n_r})\right] - \left|\left|\mathbb{E}_{P_{n_r}}\left[h(\hat{\beta}_{n_r})\right] - \mathbb{E}\left[h(\xi_{P_{n_r},n_r})\right]\right|\right|.$$

It then follows that

$$\limsup_{r\to\infty} \mathbb{E}_{P_{n_r}}\left[h(\hat{\beta}_{n_r})\right] \geq \limsup_{r\to\infty} \mathbb{E}\left[h(\xi_{P_{n_r},n_r})\right]. \tag{31}$$

Now, since $\lim_{r\to\infty} ||\delta^{pre}_{P_{n_r},n_r}|| = \infty$, there exists at least one component j of $\delta^{pre}_{P_{n_r},n_r}$ that diverges. Let $\delta^{pre}_{j,r}$ denote the jth element of $\delta^{pre}_{P_{n_r},n_r}$, and suppose WLOG that $\delta^{pre}_{j,r} \to \infty$. Likewise, let $\xi^{pre}_{j,r}$ denote the jth element of $\xi^{pre}_{P_{n_r},n_r}$. Note that $h(\xi_{P_{n_r},n_r}) = 1$ whenever $\xi^{pre}_{j,r} > \bar{M}+1$, and thus $\mathbb{E}\left[h(\xi_{P_{n_r},n_r})\right] \geq \mathbb{E}\left[1[\xi^{pre}_{j,r} > \bar{M}+1]\right]$. Hence,

$$\limsup_{r\to\infty} \mathbb{E}\left[h(\xi_{P_{n_r},n_r})\right] \geq \limsup_{r\to\infty} \mathbb{E}\left[1[\xi^{pre}_{j,r} > \bar{M}+1]\right]. \tag{32}$$

Since $\xi^{pre}_{j,r} \sim \mathcal{N}\left(\delta^{pre}_{j,r}, \sigma^2_{j,r}\right)$, for $\sigma^2_{j,r}$ the jth diagonal element of $\Sigma_{P_{n_r}}$, we have

$$\mathbb{E}\left[1[\xi^{pre}_{j,r} > \bar{M}+1]\right] = 1 - \Phi\left(\frac{\bar{M}+1-\delta^{pre}_{j,r}}{\sigma_{j,r}}\right).$$

However, by construction $\sigma_{j,r} \to \sigma^*_j$ as $r \to \infty$, where σ^{*2}_j is the jth diagonal element of Σ^*. Additionally, $\sigma^*_j > 0$ by Assumption 4. Thus, since $\delta^{pre}_{j,r} \to \infty$, we have that $\Phi\left(\frac{\bar{M}+1-\delta^{pre}_{j,r}}{\sigma_{j,r}}\right) \to 0$, and hence $\mathbb{E}\left[1[\xi^{pre}_{j,r} > \bar{M}+1]\right] \to 1$. This, combined with the inequalities (30), (31), (32), gives the desired result.

□

Lemma C.5. *Suppose Assumptions 2-5 hold. Consider a subsequence of increasing sample sizes, n_s, such that*

$$\lim_{s\to\infty} \Sigma_{P_{n_s}} = \Sigma^*, \tag{33}$$

$$\lim_{s\to\infty} \delta^{pre}_{P_{n_s},n_s} = \delta^{pre,*}, \tag{34}$$

$$\lim_{s\to\infty} \mathbb{P}_{P_{n_s}} \left(\hat{\beta}_{n_s} \in B(\hat{\Sigma}_{n_s})\right) = p^* > 0 \tag{35}$$

for $\Sigma^* \in \mathcal{S}$. *Then*

$$\lim_{s\to\infty} \left|\left|\mathbb{E}\left[\xi_{P_{n_s},n_s} - \delta^+_{n_s} \,|\, \xi_{P_{n_s},n_s} \in B(\Sigma_{P_{n_s}})\right] - \mathbb{E}\left[\xi^* \,|\, \xi^* \in B(\Sigma^*)\right]\right|\right| = 0,$$

for $\xi^* \sim \mathcal{N}(\delta^*, \Sigma^*)$, *where* $\delta^* = \begin{pmatrix} 0 \\ \delta^{pre,*} \end{pmatrix}$ *and* $\delta^+_{n_s} = \begin{pmatrix} \delta^{post}_{P_{n_s},n_s} \\ 0 \end{pmatrix}$

Proof. Let $\xi_{j,s}$ denote the jth element of $\xi_{P_{n_s},n_s} - \delta^+_{n_s}$. We show that $\mathbb{E}\left[\xi_{j,s} \,|\, \xi_{P_{n_s},n_s} \in B(\hat{\Sigma}_{P_{n_s}})\right] \longrightarrow \mathbb{E}\left[\xi^*_j \,|\, \xi^* \in B(\Sigma^*)\right]$ for each element j, which implies the desired result.

Note that $\xi_{P_{n_s},n_s} \sim \mathcal{N}\left(\delta_{P_{n_s},n_s}, \Sigma_{P_{n_s}}\right)$, so $\xi_{P_{n_s},n_s} - \delta^+_{n_s} \sim \mathcal{N}\left(\delta^-_{n_s}, \Sigma_{P_{n_s}}\right)$, where $\delta^-_{n_s} = \begin{pmatrix} 0 \\ \delta^{pre}_{P_{n_s},n_s} \end{pmatrix}$. Since by construction $\delta^-_{n_s} \longrightarrow \delta^*$ and $\Sigma_{P_{n_s}} \longrightarrow \Sigma^*$, it follows that $\xi_{P_{n_s},n_s} - \delta^+_{n_s} \xrightarrow{d} \xi^*$. The continuous mapping theorem then gives that $(\xi_{P_{n_s},n_s} - \delta^+_{n_s}) \cdot 1[\xi_{P_{n_s},n_s} \in B(\hat{\Sigma}_{P_{n_s}})] \xrightarrow{d} \xi^* 1[\xi^* \in B(\Sigma^*)]$, where the function is continuous for almost every ξ^* as shown in the proof to Lemma C.1, and we use the fact that $\xi_{P_{n_s},n_s} \in B(\hat{\Sigma}_{P_{n_s}})$ iff $\xi_{P_{n_s},n_s} - \delta^+_{n_s} \in B(\hat{\Sigma}_{P_{n_s}})$ by Assumption 5. Next, observe that

$$|\xi_{j,s} \cdot 1[\xi_{P_{n_s},n_s} \in B(\hat{\Sigma}_{P_{n_s}})]| \leq |\xi_{j,s}|.$$

Since the absolute value function is continuous and $\xi_{j,s} \xrightarrow{d} \xi^*_j$, $|\xi_{j,s}| \xrightarrow{d} |\xi^*_j|$ by the continuous mapping theorem. Further, each $|\xi_{j,s}|$ has a folded-normal distribution, as does $|\xi^*_j|$, and since the mean of a folded-normal distribution is finite and continuous in the mean and variance parameters, we have $\mathbb{E}\left[|\xi_{j,s}|\right] \to \mathbb{E}\left[|\xi^*_j|\right] < \infty$. Thus, by the generalized dominated convergence theorem,

$$\mathbb{E}\left[\xi_{j,s} \cdot 1[\xi_{P_{n_s},n_s} \in B(\hat{\Sigma}_{P_{n_s}})]\right] \xrightarrow{d} \mathbb{E}\left[\xi^*_j \cdot 1[\xi^* \in B(\Sigma^*)]\right].$$

However, by Lemma C.1 we have that

$$\mathbb{P}\left(\xi_{P_{n_s}} \in B(\hat{\Sigma}_{P_{n_s},n_s})\right) \longrightarrow \mathbb{P}\left(\xi^* \in B(\Sigma^*)\right) = p^* > 0.$$

Thus, by the continuous mapping theorem,

$$\frac{\mathbb{E}\left[\xi_{j,s} \cdot 1[\xi_{P_{n_s}} \in B(\hat{\Sigma}_{P_{n_s}})]\right]}{\mathbb{P}\left(\xi_{P_{n_s},n_s} \in B(\hat{\Sigma}_{P_{n_s}})\right)} \longrightarrow \frac{\mathbb{E}\left[\xi_j^* \cdot 1[\xi^* \in B(\Sigma^*)]\right]}{\mathbb{P}\left(\xi^* \in B(\Sigma^*)\right)},$$

as we wished to show. □

Lemma C.6. *Suppose that a sequence of random variables Y_n is asymptotically uniformly integrable,*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[||Y_n|| \cdot 1[||Y_n|| > M]\right] = 0.$$

If c_n is a sequence of constants with $c_n \to c$ and $Y_n - c_n$ converges in distribution, then $Y_n - c_n$ is also asymptotically uniformly integrable.

Proof. Note that $||Y_n - c_n|| \leq ||Y_n|| + ||c_n||$. Thus,

$$\begin{aligned}
&\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[||Y_n - c_n|| \cdot 1[||Y_n - c_n|| > M]\right] \leq \\
&\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[||Y_n|| \cdot 1[||Y_n - c_n|| > M]\right] + \lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[||c_n|| \cdot 1[||Y_n - c_n|| > M]\right].
\end{aligned} \tag{36}$$

We now show that each of the two terms on the right hand side of (36) is zero. To see why the first term is zero, note that since $c_n \to c$, for n sufficiently large, $||c_n|| \leq ||c+1||$. By the triangle inequality, $||Y_n - c_n|| \leq ||Y_n|| + ||c_n||$ and so for n sufficiently large, $1[||Y_n - c_n|| > M] \leq 1[||Y_n|| > M - ||c+1||]$. Thus,

$$\begin{aligned}
\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[||Y_n|| \cdot 1[||Y_n - c_n|| > M]\right] &\leq \lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[||Y_n|| \cdot 1[||Y_n|| > M - ||c+1||]\right] \\
&= \lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[||Y_n|| \cdot 1[||Y_n|| > M]\right],
\end{aligned}$$

and $\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[||Y_n|| \cdot 1[||Y_n|| > M]\right] = 0$ by assumption.

To show that the second term in (36) is zero, note again that since $c_n \longrightarrow c$, for n sufficiently large, $||c_n|| \leq ||c+1||$, and thus

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[||c_n|| \cdot 1[||Y_n - c_n|| > M]\right] \leq ||c+1|| \lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[1[||Y_n - c_n|| > M]\right].$$

However, since $Y_n - c_n$ converges in distribution, Prohorov's theorem gives that $Y_n - c_n$ is uniformly tight, so

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[1[||Y_n - c_n|| > M]\right] = 0.$$

□

Lemma C.7. *Suppose Assumption 5 holds. Suppose that* $\tau = \begin{pmatrix} \tau_{post} \\ 0 \end{pmatrix}$ *has zeros in the positions corresponding with* $\hat{\beta}_{pre}$*. Then for any* $\hat{\beta}$ *and* $\hat{\Sigma}$*,*

$$\hat{b}_\alpha(\hat{\beta} - \tau, \hat{\Sigma}) = \hat{b}_\alpha(\hat{\beta}, \hat{\Sigma}) - \eta'\tau.$$

Proof. Recall that for any values $(\tilde{\beta}, \tilde{\Sigma})$, $\hat{b}_\alpha(\tilde{\beta}, \tilde{\Sigma})$ solves

$$\frac{\Phi\left(\frac{\eta'\tilde{\beta} - \hat{b}_\alpha}{\tilde{\sigma}}\right) - \Phi\left(\frac{V^-(\tilde{Z}, \tilde{\Sigma}) - \hat{b}_\alpha}{\tilde{\sigma}}\right)}{\Phi\left(\frac{V^+(\tilde{Z}, \tilde{\Sigma}) - \hat{b}_\alpha}{\tilde{\sigma}}\right) - \Phi\left(\frac{V^-(\tilde{Z}, \tilde{\Sigma}) - \hat{b}_\alpha}{\tilde{\sigma}}\right)} = 1 - \alpha, \tag{37}$$

where $\tilde{Z}$ is shorthand for $Z(\tilde{\beta}, \tilde{\Sigma}) = (I - \tilde{c}\eta')\tilde{\beta}$, for $\tilde{c} = \tilde{\Sigma}\eta'/(\eta'\tilde{\Sigma}\eta)$, $\tilde{\sigma} = \sqrt{\eta'\tilde{\Sigma}\eta}$, and the functions V^+ and V^- are as defined in Lemma B.2 (replacing Σ with $\tilde{\Sigma}$). Let $\hat{\beta}^* = \hat{\beta} - \tau$, $\hat{Z} = Z(\hat{\beta}, \hat{\Sigma})$, and $\hat{Z}^* = Z(\hat{\beta}^*, \hat{\Sigma})$. We now show that

(i) $V^-(\hat{Z}^*, \hat{\Sigma}) = V^-(\hat{Z}, \hat{\Sigma}) - \eta'\tau$

(ii) $V^+(\hat{Z}^*, \hat{\Sigma}) = V^-(\hat{Z}, \hat{\Sigma}) - \eta'\tau$

(iii) $\eta'\hat{\beta}^* = \eta'\hat{\beta} - \eta'\tau$,

which together imply that $\hat{b}_\alpha$ solves (37) for $(\hat{\beta}, \hat{Z})$ iff $\hat{b}^*_\alpha = \hat{b}_\alpha - \eta'\tau$ solves (37) for $(\hat{\beta}^*, \hat{Z}^*)$, from which the claim follows.

To establish (i), note that $\hat{Z}^* - \hat{Z} = -(I - \hat{c}\eta')\begin{pmatrix} \tau_{post} \\ 0 \end{pmatrix}$, where $\hat{c}$ depends only on $\hat{\Sigma}$ and η, and not on $\hat{\beta}$. From this we see that

$$\begin{aligned} A\hat{Z}^* &= A\hat{Z} + A(\hat{Z}^* - \hat{Z}) \\ &= A\hat{Z} - A(I - \hat{c}\eta')\begin{pmatrix} \tau_{post} \\ 0 \end{pmatrix} \\ &= A\hat{Z} + (A\hat{c})\eta'\begin{pmatrix} \tau_{post} \\ 0 \end{pmatrix} \end{aligned}$$

where $A \begin{pmatrix} \tau_{post} \\ 0 \end{pmatrix} = 0$ since $A = [0,\, A_{post}]$. Additionally, from the definition of V^-,

$$V^-(\hat{Z}, \hat{\Sigma}) = \max_{\{j:(A\hat{c})_j<0\}} \frac{b_j - (A\hat{Z})_j}{(A\hat{c})_j}.$$

It follows from the previous two displays that $V^-(\hat{Z}^*) = V^-(\hat{Z}) - \eta' \begin{pmatrix} \tau_{post} \\ 0 \end{pmatrix}$. An analogous argument establishes (ii), and (iii) follows immediately from the definition of $\hat{\beta}^*$. □

Lemma C.8. *Fix $\eta \neq 0$. For any $(\hat{\beta}, \hat{\Sigma})$ and $x \in \mathbb{R}$, let $F^{\Xi(\hat{\beta},\hat{\Sigma})}_{x,\eta'\hat{\Sigma}\eta}(\cdot)$ denote the CDF of a $\mathcal{N}\left(x,\, \eta'\hat{\Sigma}\eta\right)$ variable truncated to the set $\Xi(\hat{\beta}, \hat{\Sigma}) = [V^-(Z(\hat{\beta}, \hat{\Sigma}), \hat{\Sigma}), V^+(Z(\hat{\beta}, \hat{\Sigma}), \hat{\Sigma})]$, where the functions V^-, V^+, and Z are as defined in Lemma C.10 below. Define $g(\hat{\beta}, \hat{\Sigma}, \delta) = F^{\Xi(\hat{\beta},\hat{\Sigma})}_{\eta'\delta,\eta'\hat{\Sigma}\eta}(\eta'\hat{\beta})$. Then $\hat{b}_\alpha(\hat{\beta}, \hat{\Sigma}) \leq \eta'\delta$ iff $g(\hat{\beta}, \hat{\Sigma}, \delta) \leq 1 - \alpha$.*

Proof. By definition, $\hat{b}_\alpha$ solves:

$$F^{\Xi(\hat{\beta},\hat{\Sigma})}_{\hat{b}_\alpha,\eta'\hat{\Sigma}\eta}(\eta'\hat{\beta}) = 1 - \alpha.$$

However, $F^{\Xi(\hat{\beta},\hat{\Sigma})}_{x,\eta'\hat{\Sigma}\eta}(\eta'\hat{\beta})$ is weakly decreasing in x (see, e.g. Lemma A.1 in Lee et al. (2016)), from which the result follows immediately.

□

Lemma C.9. *Suppose Σ is a positive definite matrix such that for some j, $(Ac)_j = 0$ for $c = \Sigma\eta/(\eta'\Sigma\eta)$. Let $\xi \sim \mathcal{N}(\delta,\, \Sigma)$ and $Z = (I - c\eta')\xi$. Let $B = \{\beta \,|\, A\beta \leq b\}$ such that $\mathbb{P}(\xi \in B) > 0$. Assume further that none of the rows of A are zero. Then $\mathbb{P}(b_j - (AZ)_j > 0 \,|\, \xi \in B) = 1$.*

Proof. By Lemma 5.1 in Lee et al. (2016), $\xi \in B$ only if $b_j - (AZ)_j \geq 0$. It thus suffices to show that $\mathbb{P}((Az)_j = b_j \,|\, \xi \in B) = 0$. Note that

$$\begin{aligned}(AZ)_j &= (A_{(j,\cdot)} - (Ac)_j\eta')\xi \\ &= A_{(j,\cdot)}\xi\end{aligned}$$

where $A_{(j,\cdot)}$ denotes the jth row of A, and we use the fact that $(Ac)_j = 0$. Since by assumption Σ is positive definite and $A_{(j,\cdot)} \neq 0$, it follows that $(Az)_j = A_{(j,\cdot)}\xi$ is normal with variance $A_{(j,\cdot)}\Sigma A'_{(j,\cdot)} > 0$. Hence, $\mathbb{P}((Az)_j = b_j) = 0$. Since $\mathbb{P}(\xi \in B) > 0$, it follows that $\mathbb{P}((Az)_j = b_j \,|\, \xi \in B) = 0$, as needed. □

Lemma C.10. *Fix $\eta \neq 0$. Let $\xi^* \sim \mathcal{N}(\delta^*,\, \Sigma^*)$ for $\Sigma^* \in \mathcal{S}$ such that $\mathbb{P}(\xi^* \in B(\Sigma^*)) = p^* > 0$. Let $Z(\xi, \Sigma) = (I - c(\Sigma)\eta')\xi$ for $c(\Sigma) = \Sigma\eta/(\eta'\Sigma\eta)$, and let $V^-(Z, \Sigma)$ and $V^+(Z, \Sigma)$*

be as defined in Lemma B.2. Suppose $B(\Sigma)$ satisfies Assumption 5. Then for almost every $\xi^* \,|\, \xi^* \in B(\Sigma^*)$,

(i) $V^-(Z(\xi,\Sigma),\Sigma)$ *is continuous at* (ξ^*,Σ^*) *as a function into* $\mathbb{R}\cup\{-\infty\}$*, where we define the max over the empty set to be* $-\infty$*.*

(ii) $V^+(Z(\xi,\Sigma),\Sigma)$ *is continuous at* (ξ^*,Σ^*) *as a function into* $\mathbb{R}\cup\{\infty\}$ *for almost every* $\xi \,|\, \xi \in B(\Sigma)$*, where we define the min over the empty set to be* ∞*.*

(iii) $V^-(Z(\xi^*,\Sigma^*),\Sigma^*) < V^+(Z(\xi^*,\Sigma^*),\Sigma^*)$*.*

Proof. To prove (i), begin by fixing a value ξ^*. Consider a sequence (ξ_h,Σ_h) that converges to (ξ^*,Σ^*) as $h\to\infty$. Let $z_h = Z(\xi_h,\Sigma_h)$, and note that $z_h \longrightarrow z^* := Z(\xi^*,\Sigma^*)$, since the function Z is clearly continuous for values of Σ where the denominator in $c(\Sigma)$ is non-zero, i.e. when $\eta'\Sigma\eta > 0$, and this holds for Σ^* since $\Sigma^* \in \mathcal{S}$ and thus positive definite.

Suppose first that there is no j such that $(A(\Sigma^*)c(\Sigma^*))_j = 0$. The function $A(\Sigma)$ is continuous at Σ^* by Assumption 5, and we just argued that $c(\Sigma)$ is continuous at Σ^* as well. Thus, for h sufficiently large, $\{((A(\Sigma_h)c(\Sigma_h))_j > 0\} = \{((A(\Sigma^*)c(\Sigma^*))_j > 0\}$. Likewise, the function $b(\Sigma)$ is continuous at Σ^* by Assumption 5, and so $(b_j(\Sigma) - (AZ(\xi,\Sigma))_j)/(A(\Sigma)c(\Sigma))_j$ is continuous at (ξ^*,Σ^*), from which it follows that

$$\max_{\{j:(A(\Sigma_h)c(\Sigma_h))_j<0\}} \frac{b(\Sigma_h)_j - (A(\Sigma_h)z_h)_j}{(A(\Sigma_h)c(\Sigma_h))_j} \to \max_{\{j:(A(\Sigma^*)c(\Sigma^*))_j<0\}} \frac{b(\Sigma^*)_j - (A(\Sigma^*)z^*)_j}{(A(\Sigma^*)c(\Sigma^*))_j} \tag{38}$$

when $\{((A(\Sigma^*)c(\Sigma^*))_j > 0\}$ is non-empty. By an analogous argument, if $\{((A(\Sigma^*)c(\Sigma^*))_j > 0\}$ is empty, then for h sufficiently large, $\{((A(\Sigma_h)c(\Sigma_h))_j > 0\}$ is empty as well, and so (38) holds regardless of whether $\{((A(\Sigma^*)c(\Sigma^*))_j > 0\}$ is empty.

Now, let $\mathcal{J} = \{j \,|\, ((A(\Sigma^*)c(\Sigma^*))_j = 0\}$. Note that by the same argument as in the previous paragraph,

$$\max_{\{j\notin\mathcal{J}:(A(\Sigma_h)c(\Sigma_h))_j<0\}} \frac{b(\Sigma_h)_j - (A(\Sigma_h)z_h)_j}{(A(\Sigma_h)c(\Sigma_h))_j} \to \max_{\{j:(A(\Sigma^*)c(\Sigma^*))_j<0\}} \frac{b(\Sigma^*)_j - (A(\Sigma^*)z^*)_j}{(A(\Sigma^*)c(\Sigma^*))_j}. \tag{39}$$

Additionally, by Lemma C.9, for $j\in\mathcal{J}$, $(b(\Sigma^*)_j - A(\Sigma^*)z^*)_j > 0$ for almost every value of ξ^*. For such a ξ^*, it follows from the continuous mapping theorem that for h sufficiently large, $(b(\Sigma_h) - A(\Sigma_h)z_h)_j > 0$. Thus for any $j\in\mathcal{J}$ and any subsequence $\{h_1\}\subset\{h\}$ for which $(A(\Sigma_{h_1})c(\Sigma_{h_1}))_j < 0$, we have

$$\frac{b(\Sigma_{h_1})_j - (A(\Sigma_{h_1})z_{h_1})_j}{(A(\Sigma_{h_1})c(\Sigma_{h_1}))_j} \to -\infty.$$

This implies that

$$\max_{\{j \in \mathcal{J}:(A(\Sigma_h)c(\Sigma_h))_j<0\}} \frac{b(\Sigma_h)_j - (A(\Sigma_h)z_h)_j}{(A(\Sigma_h)c(\Sigma_h))_j} \to -\infty,$$

and thus

$$\lim_{h\to\infty} \max_{\{j \notin \mathcal{J}:(A(\Sigma_h)c(\Sigma_h))_j<0\}} \frac{b(\Sigma_h)_j - (A(\Sigma_h)z_h)_j}{(A(\Sigma_h)c(\Sigma_h))_j} = \lim_{h\to\infty} \max_{\{j:(A(\Sigma_h)c(\Sigma_h))_j<0\}} \frac{b(\Sigma_h)_j - (A(\Sigma_h)z_h)_j}{(A(\Sigma_h)c(\Sigma_h))_j}.$$

Result (i) then follows from (39). The proof of result (ii) is analogous to that for proof (i), replacing *max* with *min* and $-\infty$ with ∞. Result (iii), that $V^- < V^+$, follows from the same argument as in the proof of Lemma B.2.

□

Lemma C.11. *Let* $\tilde{g}(\xi, \Sigma, V^-, V^+, \delta) = F^{[V^-,V^+]}_{\eta'\delta,\eta'\Sigma\eta}(\eta'\xi)$. *Then* $\tilde{g}$ *is continuous in* $(\xi, \Sigma, V^-, V^+, \delta)$ *on the set* $\{(\xi, \Sigma, V^-, V^+, \delta) \,|\, \xi \in \mathbb{R}^{K+M}, \Sigma \in B(\mathcal{S}), V^- \in \mathbb{R} \cup \{-\infty\}, V^+ \in \mathbb{R} \cup \{\infty\}, \delta \in \mathbb{R}^{K+M}\}$, *for* $B(\mathcal{S})$ *an open set of positive definite matrices containing* $\mathcal{S}$.

Proof. By definition,

$$\tilde{g}(\xi, \Sigma, V^-, V^+, \delta) = \frac{\Phi\left(\frac{\eta'\xi - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{V^- - \eta'\delta}{\sigma}\right)}{\Phi\left(\frac{V^+ - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{V^- - \eta'\delta}{\sigma}\right)},$$

for $\sigma = \sqrt{\eta'\Sigma\eta}$. Since $\Sigma \in B(\mathcal{S})$ implies that Σ is full rank, and hence $\sigma > 0$, it is immediate from the functional form that $\tilde{g}$ is continuous when all of the values are finite.

Moreover, for V^- finite and $(\xi, \Sigma, V^-, \delta)$ fixed,

$$\lim_{V^+\to\infty} \frac{\Phi\left(\frac{\eta'\xi - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{V^- - \eta'\delta}{\sigma}\right)}{\Phi\left(\frac{V^+ - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{V^- - \eta'\delta}{\sigma}\right)} = \frac{\Phi\left(\frac{\eta'\xi - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{V^- - \eta'\delta}{\sigma}\right)}{\Phi\left(\frac{\infty}{\sigma}\right) - \Phi\left(\frac{V^- - \eta'\delta}{\sigma}\right)}.$$

Moreover, for V^+ finite and $(\xi, \Sigma, V^+, \delta)$ fixed,

$$\lim_{V^- \to -\infty} \frac{\Phi\left(\frac{\eta'\xi - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{V^- - \eta'\delta}{\sigma}\right)}{\Phi\left(\frac{V^+ - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{V^- - \eta'\delta}{\sigma}\right)} = \frac{\Phi\left(\frac{\eta'\xi - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{-\infty}{\sigma}\right)}{\Phi\left(\frac{V^+ - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{-\infty}{\sigma}\right)}.$$

Finally, for (ξ, Σ, δ) fixed,

$$\lim_{(V^-, V^+) \to (-\infty, \infty)} \frac{\Phi\left(\frac{\eta'\xi - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{V^- - \eta'\delta}{\sigma}\right)}{\Phi\left(\frac{V^+ - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{V^- - \eta'\delta}{\sigma}\right)} = \frac{\Phi\left(\frac{\eta'\xi - \eta'\delta}{\sigma}\right) - \Phi\left(\frac{-\infty}{\sigma}\right)}{\Phi\left(\frac{\infty}{\sigma}\right) - \Phi\left(\frac{-\infty}{\sigma}\right)}.$$

□

Lemma C.12. *Suppose the function $B(\Sigma)$ satisfies Assumption 5. Let $\xi^* \sim \mathcal{N}(\delta^*, \Sigma^*)$ for $\Sigma^* \in \mathcal{S}$ such that $\mathbb{P}(\xi^* \in B(\Sigma^*)) = p > 0$. Then $g(\xi^*, \Sigma^*, \delta^*)$ is continuous for almost every $\xi^* \,|\, \xi^* \in B(\Sigma^*)$ for the function g as defined in Lemma C.8.*

Proof. Observe that

$$g(\xi, \Sigma, \delta) = \tilde{g}(\xi, \Sigma, V^-(\xi, \Sigma), V^+(\xi, \Sigma), \delta),$$

for the function $\tilde{g}$ as defined in Lemma C.11. Lemma C.10 gives that for almost every value of $\xi^* \,|\, \xi^* \in B(\Sigma^*)$, the functions V^- and V^+ are continuous in (ξ, Σ) at (ξ^*, Σ^*) with $V^-(\xi^*, \Sigma^*) < V^+(\xi^*, \Sigma^*)$. Lemma C.11 gives that $\tilde{g}$ is continuous on $\{(\xi, \Sigma, V^-, V^+, \delta) \,|\, \xi \in \mathbb{R}^{K+M}, \Sigma \in B(\mathcal{S}), V^- \in \mathbb{R} \cup \{-\infty\}, V^+ \in \mathbb{R} \cup \{\infty\}, \delta \in \Delta\}$, for $B(\mathcal{S})$ an open set containing $\mathcal{S}$. The result then follows immediately from the fact that the composition of continuous functions is continuous. □

D Power Calculations Under Stochastic Differential Trends

This section considers data-generating processes in which there are stochastic differential trends between the treated and control groups. In particular, we consider the following hierarchical model:

$$\delta \sim \mathcal{N}(0, V) \tag{40}$$

$$\hat{\beta} \,|\, \delta \sim \mathcal{N}(\delta + \tau, \Sigma). \tag{41}$$

The distribution for $\hat{\beta}|\delta$ in (41) is identical to the model considered in Section 3. However, we now treat δ as stochastic, rather than as a fixed parameter (e.g. linear in event-time). Treating δ as stochastic is sensible in situations in which we think that there may be common shocks to the treated and control groups (e.g. if each of these is a state, and there are macro-level shocks).

I now evaluate the power of pre-tests against such stochastic shocks in data-generating processes calibrated to the sample of papers reviewed in Section 4. For a given value of (V, Σ), we define the power of the pre-test to be the probability, $\mathbb{P}_{\delta,\hat{\beta}}\left(\hat{\beta}_{pre} \in B(\Sigma)\right)$, where $\mathbb{P}_{\delta,\hat{\beta}}(\cdot)$ denotes the probability taken over the realization of the joint distribution of $(\delta, \hat{\beta})$. We explicitly write the pre-test acceptance region as $B(\Sigma)$ to denote that the pre-test region depends on Σ (but not V). We again set Σ to be the estimated variance-covariance matrix from each of the papers in the sample. Calibrating the covariance matrix V for the common stochastic shocks is more difficult, as it cannot be consistently estimated from the data. For simplicity, I set $V = c \cdot \Sigma$ for a constant $c > 0$. Under this specification, the marginal distribution of $\hat{\beta}$ under the hierarchical model defined above is $\mathcal{N}(0, (1+c)\Sigma)$. The parameter c can thus be interpreted as the factor by which we have underestimated the variance matrix by treating δ as fixed and ignoring common stochastic shocks.

I then calculate the values of c for which the pre-test rejects 50 or 80% percent of the time, which I denote $c_{0.5}$ and $c_{0.8}$. As in Section 4, I use the pre-test criterion that no pre-period coefficient is significant at the 95% level. I compute the null rejection probabilities of conventional confidence intervals for the average post-treatment effect $\bar{\tau}$ and the first-period treatment effect τ_1 under the DGPs with $c_{0.5}$ and $c_{0.9}$. The null rejection probabilities are computed over the joint distribution of $(\hat{\beta}, \delta)$.[19] As in Section 4, I report these probabilities both unconditionally, and conditional on surviving the pre-test. Tables 5 and 6 show the results for τ_1 and $\bar{\tau}$, respectively. Across all specifications, the null rejection probabilities substantially exceed the nominal level of 5% for most of the papers. Conditioning on passing the pre-test generally reduces the null rejection probability, but only moderately so in most cases. Conditional on passing the pre-test, null rejection probabilities are often many multiples of the nominal size. The results thus suggest that conventional pre-tests may be underpowered against detecting common stochastic shocks, in addition to the linear secular trends considered in the main text.

I do not report results for bias as in the main text, since δ is mean-zero and so $\hat{\beta}$ is unbiased when the expectation is taken over the joint distribution of $(\hat{\beta}, \delta)$.

[19] Recall that $\hat{\beta} \sim \mathcal{N}(0, (1+c)\Sigma)$. Thus, this is the probability that τ falls inside a confidence interval based on the assumption that $\hat{\beta} \sim \mathcal{N}(\tau, \Sigma)$ distribution when in fact $\hat{\beta} \sim \mathcal{N}(\tau, (1+c)\Sigma)$.

	Conditional on passing pre-test?			
	No		Yes	
	Scaling factor for stochastic variance			
	$c_{0.5}$	$c_{0.8}$	$c_{0.5}$	$c_{0.8}$
Bailey and Goodman-Bacon (2015)	0.17	0.34	0.16	0.33
Bosch and Campos-Vazquez (2014)	0.19	0.38	0.12	0.27
Deryugina (2017)	0.19	0.38	0.04	0.09
Deschenes et al. (2017)	0.17	0.35	0.10	0.19
Fitzpatrick and Lovenheim (2014)	0.23	0.45	0.21	0.43
Gallagher (2014)	0.14	0.30	0.12	0.26
He and Wang (2017)	0.26	0.48	0.23	0.46
Kuziemko et al. (2018)	0.29	0.55	0.20	0.42
Lafortune et al. (2017)	0.19	0.38	0.18	0.37
Markevich and Zhuravskaya (2018)	0.22	0.44	0.18	0.38
Tewari (2014)	0.10	0.22	0.08	0.18
Ujhelyi (2014)	0.22	0.43	0.18	0.36

Table 5: Null Rejection Probabilities for Nominal 5% Test of Average Treatment Effect Under Stochastic Trends Against Which We Have 50 or 80% Power

Note: This table shows null rejection probabilities for nominal 5% significant level tests using data-generating processes under which there are stochastic violations of parallel trends that conventional pre-tests would detect 50 or 80% of the time ($c_{0.5}$ and $c_{0.8}$). The first two columns show unconditional null rejection probabilities, whereas the latter two columns condition on passing the pre-test. The estimand is the average of the post-treatment causal effects, $\bar{\tau}$. See Section D for details on the data-generating process.

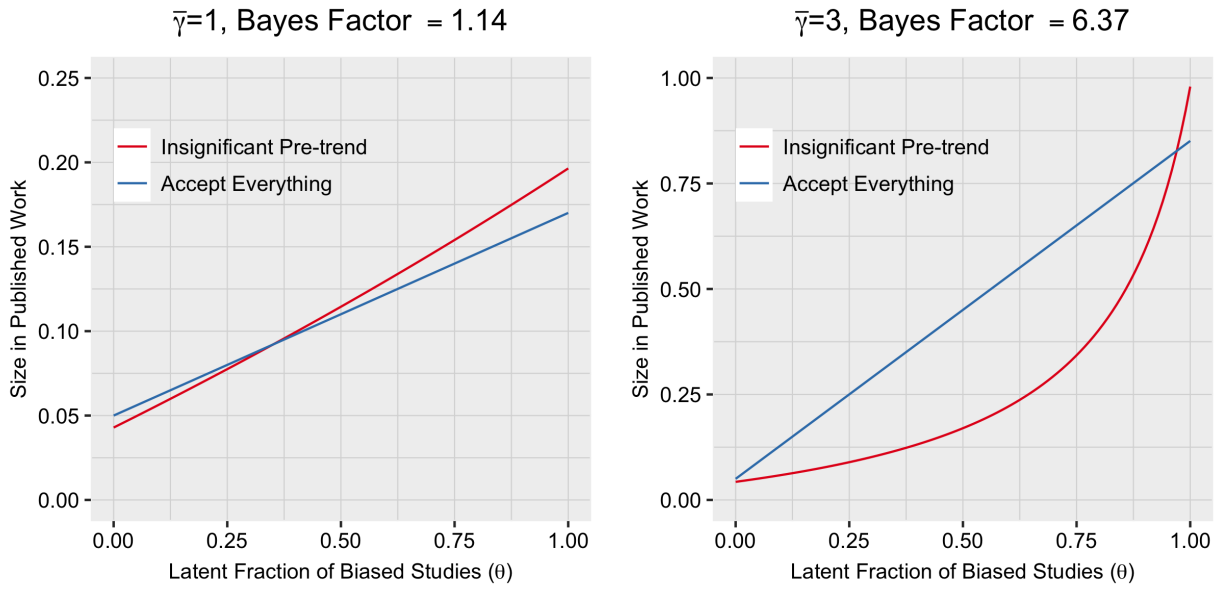
	Conditional on passing pre-test?			
	No		Yes	
	Scaling factor for stochastic variance			
	$c_{0.5}$	$c_{0.8}$	$c_{0.5}$	$c_{0.8}$
Bailey and Goodman-Bacon (2015)	0.17	0.34	0.14	0.30
Bosch and Campos-Vazquez (2014)	0.19	0.38	0.17	0.35
Deryugina (2017)	0.19	0.38	0.13	0.29
Deschenes et al. (2017)	0.17	0.35	0.11	0.22
Fitzpatrick and Lovenheim (2014)	0.23	0.45	0.22	0.44
Gallagher (2014)	0.14	0.30	0.08	0.19
He and Wang (2017)	0.26	0.48	0.23	0.45
Kuziemko et al. (2018)	0.29	0.55	0.21	0.45
Lafortune et al. (2017)	0.19	0.38	0.18	0.37
Markevich and Zhuravskaya (2018)	0.22	0.44	0.17	0.36
Tewari (2014)	0.10	0.22	0.08	0.19
Ujhelyi (2014)	0.22	0.43	0.17	0.35

Table 6: Null Rejection Probabilities for Nominal 5% Test of First Period Treatment Effect Under Stochastic Trends Against Which We Have 50 or 80% Power

Note: This table shows null rejection probabilities for nominal 5% significant level tests using data-generating processes under which there are stochastic violations of parallel trends that conventional pre-tests would detect 50 or 80% of the time ($c_{0.5}$ and $c_{0.8}$). The first two columns show unconditional null rejection probabilities, whereas the latter two columns condition on passing the pre-test. The estimand is the causal effect for the first period after treatment, τ_1. See Section D for details on the data-generating process.

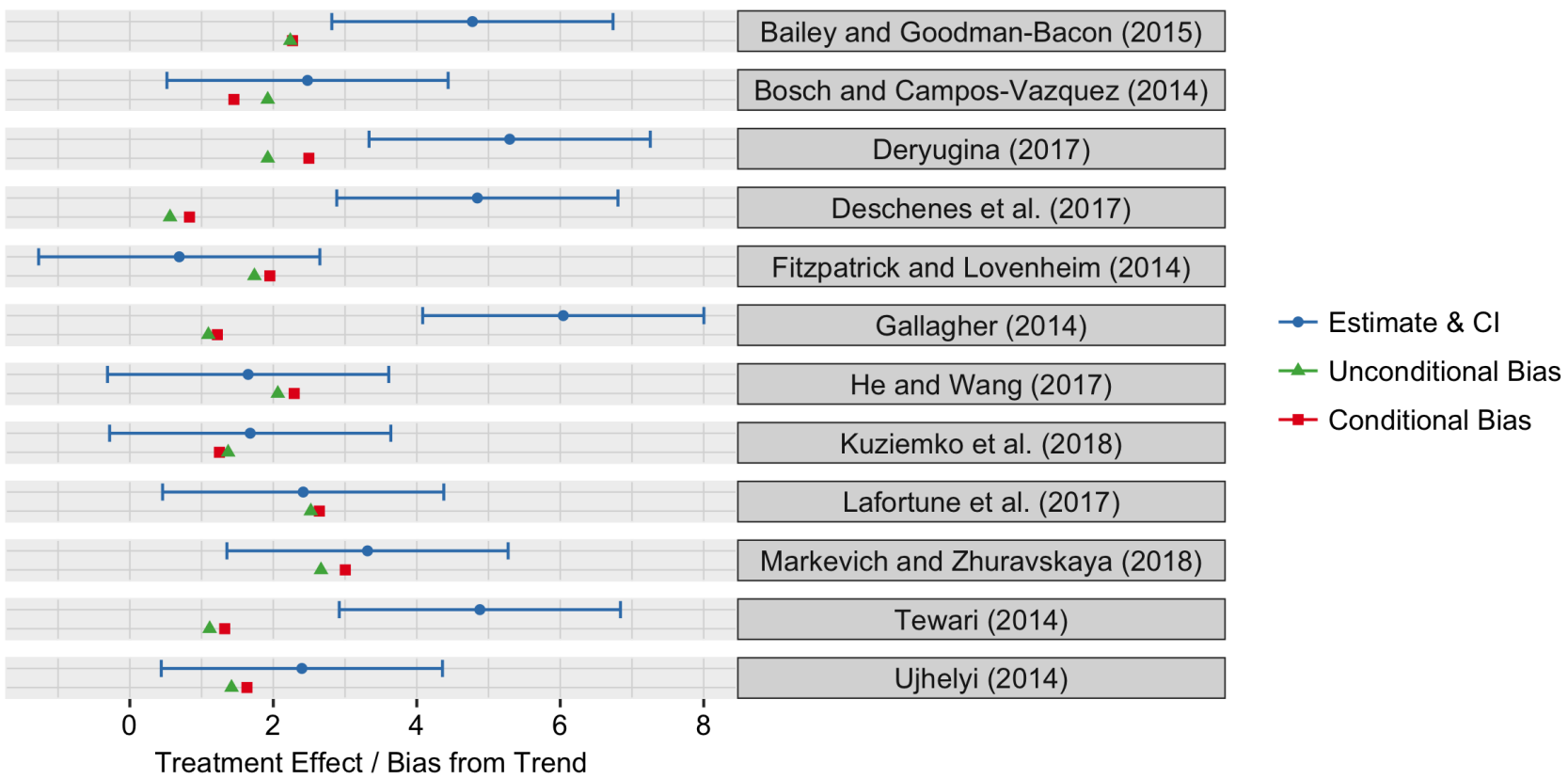
E Additional tables and figures

Figure D1: Comparing size in published studies when requiring an insignificant pre-trend versus publishing everything



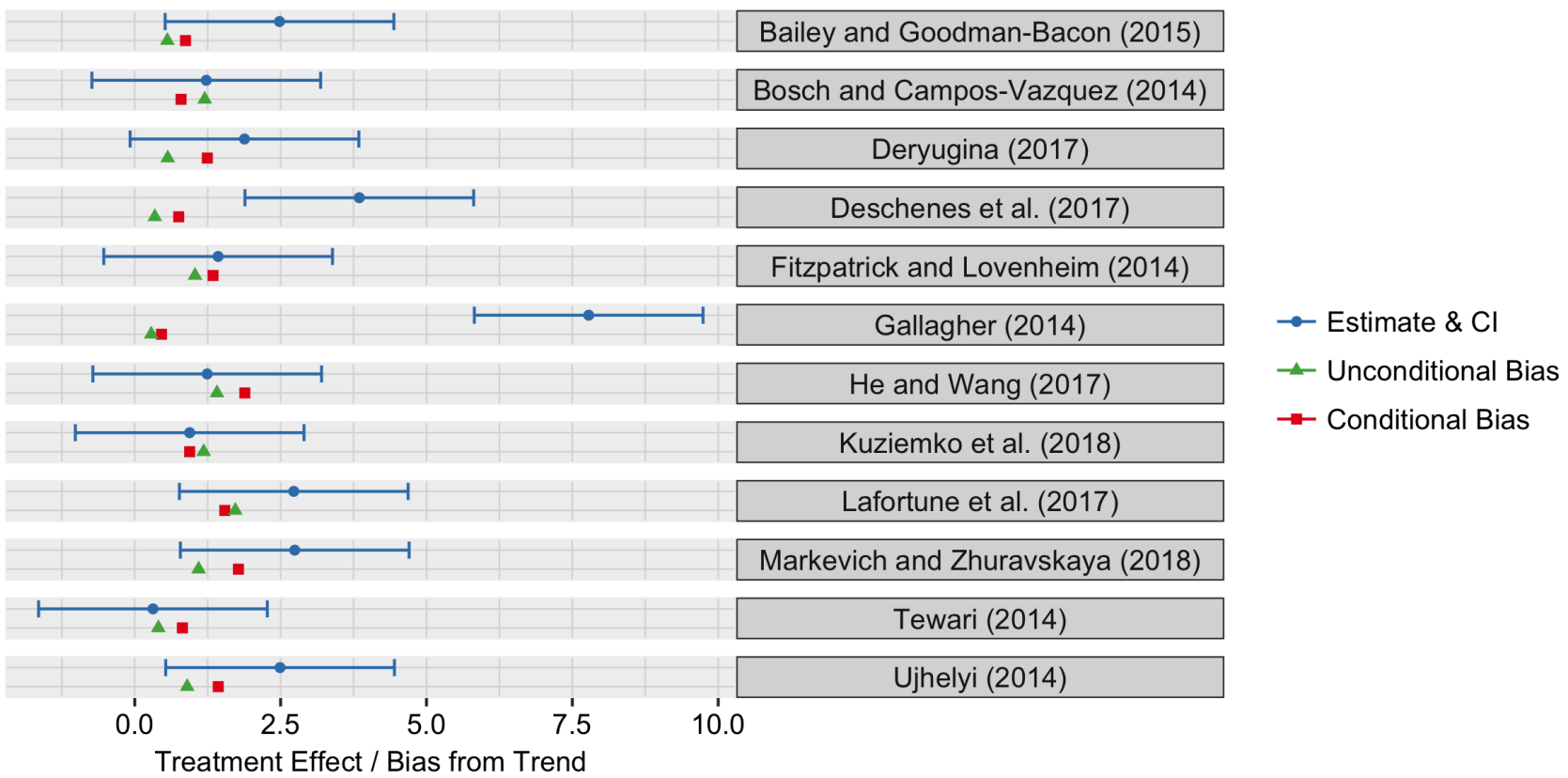
Note: Each figure shows the size (null rejection probability) in published work in the setting described in Section 2.3 as a function of the fraction of latent studies in which parallel trends is violated (θ). The Insignificant Pre-trend regime only publishes studies in which $\hat{\beta}_{-1}$ is statistically insignificant. The two panels show results for different values of the slope of the differential trend (γ) when parallel trends fails. See Section 2.3 for further detail.

Figure D2: OLS Estimates and Bias from Linear Trends for Which We Have 50 Percent Power – Average Treatment Effect



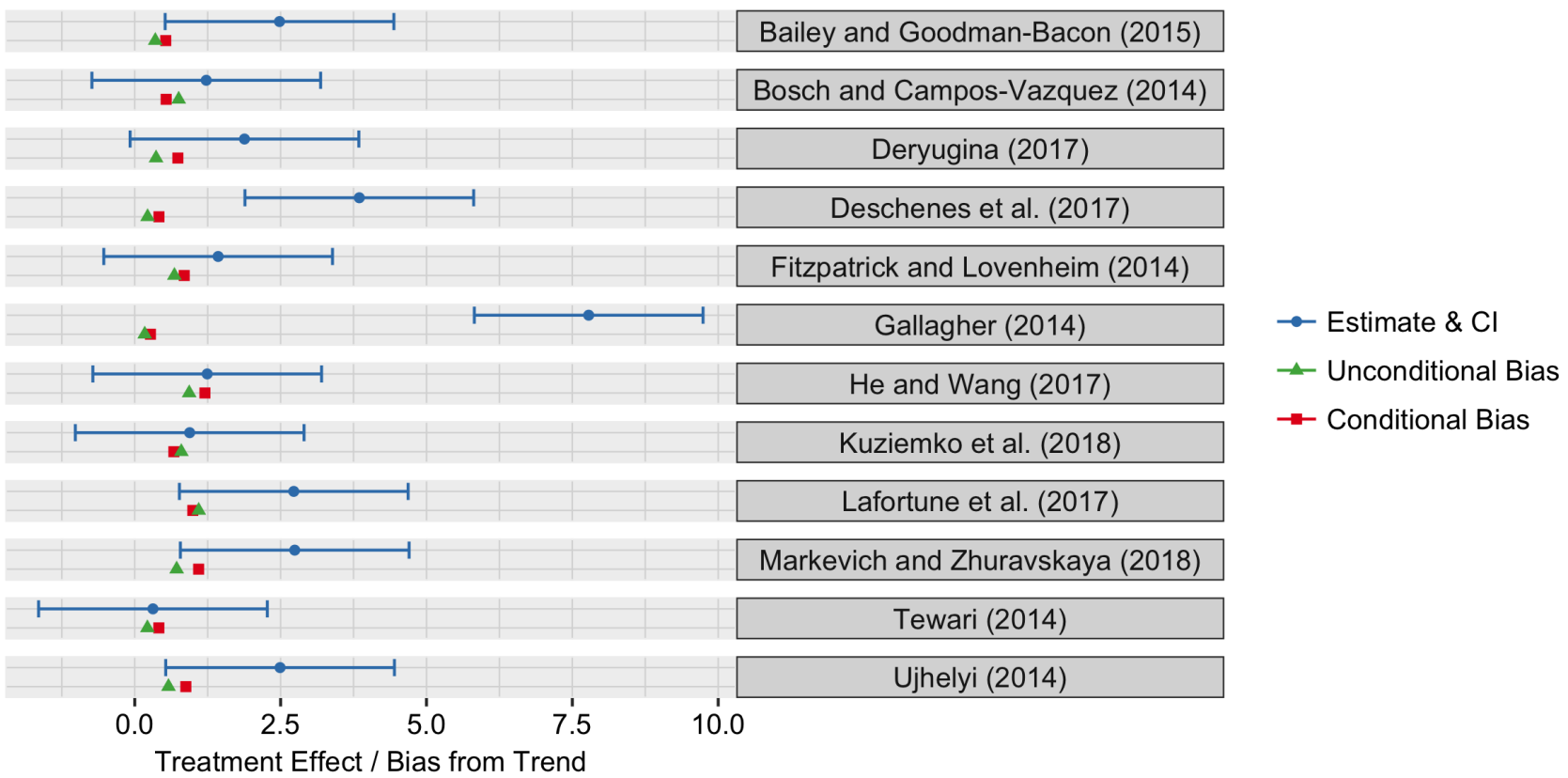
Note: I calculate the linear trend against which we would have a rejection probability of 50 percent if we rejected the research design whenever any of the pre-period event-study coefficients was statistically significant at the 5% level. I plot in red the bias that would result from such a trend conditional 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 estimand is the average of the treatment effects in all periods after treatment began, $\bar{\tau}$.

Figure D3: OLS Estimates and Bias from Linear Trends for Which We Have 80 Percent Power – First Period Treatment Effect



Note: I calculate the linear trend against which we would have a rejection probability of 80 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 estimand is the treatment effect in the first period after treatment began, τ_1.

Figure D4: OLS Estimates and Bias from Linear Trends for Which We Have 50 Percent Power – First Period Treatment Effect



Note: I calculate the linear trend against which we would have a rejection probability of 50 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 estimand is the treatment effect in the first period after treatment began, τ_1.

	Conditional on passing pre-test?			
	No		Yes	
	Slope of differential trend:			
	$\gamma_{0.5}$	$\gamma_{0.8}$	$\gamma_{0.5}$	$\gamma_{0.8}$
Bailey and Goodman-Bacon (2015)	0.06	0.09	0.07	0.13
Bosch and Campos-Vazquez (2014)	0.12	0.22	0.08	0.11
Deryugina (2017)	0.07	0.09	0.09	0.21
Deschenes et al. (2017)	0.06	0.06	0.05	0.08
Fitzpatrick and Lovenheim (2014)	0.10	0.18	0.13	0.26
Gallagher (2014)	0.05	0.06	0.04	0.05
He and Wang (2017)	0.15	0.29	0.21	0.47
Kuziemko et al. (2018)	0.13	0.22	0.07	0.11
Lafortune et al. (2017)	0.19	0.41	0.17	0.34
Markevich and Zhuravskaya (2018)	0.11	0.19	0.17	0.42
Tewari (2014)	0.06	0.07	0.06	0.11
Ujhelyi (2014)	0.09	0.15	0.12	0.28

Table D1: Null Rejection Probabilities for Nominal 5% Test of First Period Treatment Effect Under Linear Trends Against Which We Have 50 or 80% Power

Note: This table shows null rejection probabilities for nominal 5% significant level tests using data-generating processes under which there are linear violations of parallel trends that conventional pre-tests would detect 50 or 80% of the time ($\gamma_{0.5}$ and $\gamma_{0.8}$). The first two columns show unconditional null rejection probabilities, whereas the latter two columns condition on passing the pre-test. The estimand is the treatment effect in the first period after treatment, τ_1.

References

Andrews, I. and Kasy, M. (2019). Identification of and Correction for Publication Bias. *American Economic Review*, 109(8):2766–2794.

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.

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.

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.

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

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

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media. Google-Books-ID: seH8dMrEgggC.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press. Google-Books-ID: UEuQEM5RjWgC.