Online Supplement to 'Policy Evaluation with Nonlinear Trended Outcomes: COVID-19 Vaccination Rates in the US'*

Lynn Bergeland Morgan[a], Peter C. B. Phillips[b,c,d] and Donggyu Sul[a]

[a] *University of Texas at Dallas,* [b] *Yale University,* [c] *University of Auckland,*
[d] *Singapore Management University,*

July 11, 2023

This document contains supplementary theorems on regressions with nonstationary outcomes, proofs, additional discussion concerning mechanisms for convergence clustering, further simulation results, and other supplementary regression results to those reported in Morgan et al. (2023). Proofs are provided in the Appendix to this Online Supplement. We begin by considering some pitfalls in two-way fixed effect (TWFE) regressions when trends are present.

1 Pitfalls of TWFE estimation I: nonstationary outcomes under a homogeneous trend

We illustrate some of the problems that can arise in the use of an explanatory TWFE regresion (given below in (2)) when the outcome data y_{it} employed are generated by a simple panel DGP involving individual stochastic trends with linear drifts of the following form

$$y_{it} = a_i + b_i t + \xi_{it}, \quad \xi_{it} = \xi_{it-1} + e_{it}, \quad i = 1, \cdots, N; \; t = 1, \cdots, T \tag{1}$$

Here, the primary innovations e_{it} are assumed to be cross-sectionally independent and stationary and ergodic over time with zero mean, finite variances and positive long run variances ω_{ei}^2, whose cross-sectional average $\frac{1}{n}\sum_i^n \omega_{ei}^2$ has a finite positive limit $\omega_e^2 = \lim_{n\to\infty} \frac{1}{n}\sum_i^n \omega_{ei}^2 > 0$. The policy regressors x_{it} that appear in the TWFE regression (2) below are assumed to be cross-sectionally independent, strictly exogenous (i.e., independent of the innovations e_{it} in (1)), and stationary over time with variances σ_{xi}^2, whose cross-sectional average $\frac{1}{n}\sum_i^n \sigma_{xi}^2$ has a finite positive limit $\sigma_x^2 = \lim_{n\to\infty} \frac{1}{n}\sum_i^n \sigma_{xi}^2 > 0$. Further, centered time series averages of x_{it} satisfy a

*Phillips acknowledges research support from the NSF (Grant No. SES 18-50860) at Yale University and a Kelly Fellowship at the University of Auckland. Email: lynn.morgan@utdallas.edu; peter.phillips@yale.edu; d.sul@utdallas.edu.

central limit theorem (CLT) so that $T^{-1/2}\sum_{t=1}^{T}(x_{it}-m_{xi}) \rightsquigarrow \mathcal{N}(0,\omega_{xi}^2)$, where $m_{xi}=\mathbb{E}x_{it}$ is the mean and $\omega_{xi}^2>0$ the long run variance of x_{it}. The limit of the cross section average of the products of the long run variances of x_{it} and e_{it}, viz., $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\omega_{xi}^2\omega_{ei}^2$, is also assumed to be finite and positive.

In this panel DGP framework (1) for the outcome variables y_{it}, trend slope homogeneity clearly depends on the coefficients b_i being equal for all i. If $b_i \neq b$ for all i then the y_{it} have heterogeneous deterministic linear trends. If the b_i are random, then these linear deterministic trends have stochastic slope coefficients cross section.

We suppose a TWFE regression model is conducted to explain these data, taking the form

$$y_{it} = a_i + \theta_t + \beta x_{it} + u_{it}, \tag{2}$$

where, for simplicity, no variables beyond the policy variable x_{it} are included in the regression. Under the null hypothesis that the policy variable coefficient $\beta=0$ and a maintained assumption that the time specific effect is the linear drift $\theta_t = bt$, the implied true regression error in (2) is

$$u_{it} = (b_i - b)t + \xi_{it}, \tag{3}$$

thereby making slope homogeneity on b_i relevant to the properties of the regression. To formalize the analysis we employ assumptions that distinguish the relevant two cases, starting with trend homogeneity.

Assumption 1 (Homogeneous trends): $b_i = b$ for all i.

Let β_{fe} be the TWFE estimator of β from (2) and define the demeaned data $\dot{y}_{it}$ as follows

$$\dot{y}_{it} := y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it} - \frac{1}{n}\sum_{i=1}^{n} y_{it} + \frac{1}{T}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it} =: y_{it} - \bar{y}_{i\bullet} - \bar{y}_{\bullet t} + \bar{y}_{\bullet\bullet}, \tag{4}$$

with similar definitions for $\dot{x}_{it}$ and $\dot{u}_{it}$. Then the TWFE regression equation has the form

$$\dot{y}_{it} = \beta \dot{x}_{it} + \dot{u}_{it}, \tag{5}$$

with $\dot{u}_{it} = \dot{\xi}_{it}$ and the centered TWFE estimator is

$$\hat{\beta}_{\text{fe}} - \beta = \frac{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}\dot{u}_{it}}{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^2} = \frac{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}\dot{\xi}_{it}}{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^2}. \tag{6}$$

It is useful to note that both the form and the asymptotic properties of $\hat{\beta}_{\text{fe}}$ are unaffected if the policy variable x_{it} has a homogeneous deterministic trend, such as $x_{it} = ct + x_{it}^0$, where x_{it}^0 has the same properties earlier ascribed to x_{it}, as in this case it is easy to see that composite demeaning as in (4) leads to the equivalence $\dot{x}_{it} = x_{it} - \bar{x}_{i\bullet} - \bar{x}_{\bullet t} + \bar{x}_{\bullet\bullet} =: \dot{x}_{it}^0$.

Under the stated conditions the denominator of (6) has a well defined limit in probability as $(n,T)\to\infty$. The numerator has more complex behavior due to the nonstationary characteristics of $\dot{\xi}_{it}$ and the effects of removing fixed effects in the regression by demeaning. The following result gives the limit behavior of the standardized and centered TWFE estimator $\hat{\beta}_{\text{fe}}$.

Theorem 1 Nonstationary outcome with a stationary policy *Under the conditions stated above and homogeneity in Assumption 1, as* $(n, T) \to \infty$ *the limit distribution of* $\hat{\beta}_{\text{fe}}$ *is given by*

$$\sqrt{n}\left(\hat{\beta}_{\text{fe}} - \beta\right) \rightsquigarrow \mathcal{N}\left(0, V_\beta\right), \tag{7}$$

where $V_\beta = V/\sigma_x^4$ *with*

$$V = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{\omega_{xi}^2 \omega_{ei}^2}{6}, \text{ and } \sigma_x^2 = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \sigma_{xi}^2. \tag{8}$$

The proof of Theorem 1 is given in the Appendix and here we provide an intuitive discussion of the asymptotics. For this purpose it is useful to compare the present setting with the conventional framework where the outcome variable is trend stationary with a homogeneous trend and simple *iid* innovations. In that model the TWFE estimator in (2) has the usual $\sqrt{nT}$ convergence rate, giving a typical panel data gain from both time series and cross section averaging. But, as Theorem 1 shows, when the outcome variable has a stochastic trend component some of this conventional gain is lost due to the impact of stochastic nonstationarity and this is so even under a homogeneous deterministic trend. As is apparent from (6), the effective residual in the regression is $\zeta_{it} := \dot{x}_{it}\dot{\xi}_{it}$, which carries the impact of the stochastic trend component ξ_{it} with order $O_p(\sqrt{T})$ rather than the conventional $O_p(1)$ of a panel regression residual. On the other hand, the signal from the regressor $\dot{x}_{it}^2$ retains its conventional $O_p(nT)$ strength. The net effect, therefore, is that the consistency of the TWFE estimator relies solely on the size of n and more time series observations T do not improve estimation accuracy. In effect, stochastic trend residuals have a measurable impact on panel TWFE regressions: consistency is retained by cross section averaging of the independent individual components in the panel, while the 'mildly spurious' impact of a stochastic trend in the residuals weakens the advantage of time series averaging without inducing inconsistency.

To estimate the asymptotic variance V_β in (7) consistently under the assumption that there is cross section independence in both x_{it} and ξ_{it}, the following panel robust variance estimator may be employed, viz.,

$$\hat{V}_\beta = \left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^2\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{w}_{iT}^2\right)\left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^2\right)^{-1}, \tag{9}$$

with $\hat{w}_{iT} = \frac{1}{T}\sum_{t=1}^{T}\dot{x}_{it}\hat{u}_{it}$ where $\hat{u}_{it}$ is the residual from the TWFE regression. Consistency of the variance estimator (9) is shown in the Appendix and relies on joint asymptotics with $(n, T) \to \infty$ as well as cross section independence over i. In the presence of cross-sectional dependence (including spatial dependence), however, estimators such as $\hat{V}_\beta$ are typically not consistent, although a clustering approach (Cameron and Miller, 2015) can be successfully used to achieve consistency when there is an underlying cluster form of dependence that is known to the investigator. Consistency of $\hat{V}_\beta$ and extensions to allow for cross section dependence facilitate inference about the slope coefficient β and hence the impact of the policy variable x_{it}

on the outcome variable y_{it} in the panel.

2 Pitfalls of TWFE Estimation II: nonstationary outcomes under heterogeneous trends

In formulating (1) and developing asymptotic theory for TWFE regression using (2) we now allow for heterogeneous trends.

Assumption 2 (Heterogeneous trends) $b_i \neq b$ for some i.

Under Assumption 2 and the generating mechanism (1), the regression error in (5) becomes

$$\dot{u}_{it} = \left(b_i - n^{-1}\sum_{i=1}^{n} b_i\right)\left(t - T^{-1}\sum_{t=1}^{T} t\right) + \dot{\xi}_{it} =: (b_i - \bar{b})(t - \bar{t}) + \dot{\xi}_{it},$$

so that $\dot{u}_{it}$ includes deterministic trends with zero cross section and time series sample averages in addition to $\dot{\xi}_{it}$. It follows that the asymptotic properties of the TWFE estimator now depend on whether the regressors $\dot{x}_{it}$ themselves have heterogeneous trends. When both outcome and policy regressor variables have heterogeneous trends, asymptotic properties in TWFE estimation depend on the characteristics and relationship between the slope coefficients of the deterministic trends. To fix ideas let

$$x_{it} = c_i t + x_{it}^0, \tag{10}$$

where x_{it}^0 is stationary and ergodic over time with the same properties as earlier in Section 1 with no deterministic trend dependence. Under the null of no policy impact ($\beta = 0$) and when the trend slope coefficients b_i and c_i are random *iid* sequences with respective variances σ_b^2 and σ_c^2 and zero covariance $\mathbb{C}ov(b_i, c_i) =: \sigma_{bc} = 0$, then we may expect different asymptotic results from when there is nonzero covariance between the sequences. In the latter case there is naturally induced correlation between the policy x_{it} and the outcomes y_{it}. Similar correlations occur when the trend slope coefficients b_i and c_i are fixed, nonrandom sequences for which $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \tilde{b}_i \tilde{c}_i = \sigma_{bc} \neq 0$, where $\tilde{c}_i := c_i - n^{-1}\sum_{i=1}^{n} c_i$ with matching definitions for $\tilde{b}_i$ and $\tilde{t} = t - T^{-1}\sum_{t=1}^{T} t$. In both these two cases, the induced correlation between the policy and outcome variables affects the limit theory, much in the same way as there is always (potentially spurious) correlation between deterministically trending time series (Durlauf and Phillips, 1988; Phillips, 1986). Under these conditions Theorem 1 can be extended with a similar $\sqrt{n}$ convergence rate for $\hat{\beta}_{\text{fe}}$ when $\sigma_{bc} \neq 0$ but with a different asymptotic variance that is dominated by the variances (σ_b^2 and σ_c^2) of the slope coefficient sequences. To proceed we assume for the following result that b_i and c_i are random *iid* sequences with means m_b and m_c and covariance $\mathbb{C}ov(b_i, c_i) = \sigma_{bc}$. Note that after mean transformation the policy variable has the form $\dot{x}_{it} = \tilde{c}_i \tilde{t} + \dot{x}_{it}^0$. The next result, proved in the Appendix, gives the limit behavior of the TWFE estimator when policy variable has heterogeneous trends.

Theorem 2 Nonstationary outcomes and heterogeneously trending policies *Under Assumption 2, equation* (10)*, and the conditions given above for the random sequences* (c_i, b_i)

of trend slope coefficients, as $n, T \to \infty$,

$$\sqrt{n}\left(\hat{\beta}_{\text{fe}} - \beta\right) \to^{d} \mathcal{N}\left(0, \sigma_b^2/\sigma_c^2\right). \tag{11}$$

The form of the limit theory in (11) is natural and somewhat expected in this simple two variable panel TWFE regression since the linear trend terms in the policy regressors x_{it} and the outcome variables y_{it} are dominant in the regression asymptotics even in the presence of a stochastic trend ξ_{it} in y_{it}.[1]

A further case that is relevant in some empirical settings occurs when the the outcome variable has heterogeneous trends but the policy variable x_{it} is stationary with no trend. Using the same notation as above, the TWFE estimator has the following decomposition

$$\hat{\beta}_{\text{fe}} - \beta = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}\tilde{b}_i\tilde{t}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^2} + \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}\dot{\xi}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^2}. \tag{12}$$

Compared with (6), where there are no deterministic tends in the outcome variable, the error in the TWFE estimator now contains the effect of the outcome trends in the first term on the right side of (12). This first term then dominates asymptotically because of these retained trend effects which produce high variance. The implication is that even when the policy coefficient $\beta = 0$, the TWFE estimator diverges asymptotically, producing potentially large (positive or negative) values in finite samples. The following result, proved in the Appendix, formalizes the asymptotic behavior.

Theorem 3 Impact of heterogeneous trends with no trends in policy Under *Assumption 2 and the conditions of Theorem 2, but when regressors are stationary with no trend components, as* $(n, T) \to \infty$, *the asymptotic behavior of* $\hat{\beta}_{fe}$ *has the following form*

$$\hat{\beta}_{fe} - \beta = O_p\left(\sqrt{\frac{T}{n}}\right) + O_p\left(\frac{1}{\sqrt{n}}\right), \tag{13}$$

so that $\hat{\beta}_{fe}$ *is inconsistent and divergent when* $n = o(T)$.

As shown in the Appendix, heterogeneous deterministic trends in the outcome variables y_{it} of (1) continue to dominate the asymptotics and $\hat{\beta}_{\text{fe}} - \beta = O_p\left(\sqrt{T/n}\right)$ still holds when the outcome variables y_{it} have stationary innovations ξ_{it}.

The above analysis is developed for linear deterministic trend functions, allowing also for random heterogeneous slope coefficients. The empirical applications to Covid-19 data in Morgan et al. (2023) show that a linear trend model may not always provide a good representation of the data when nonlinear trends such as U-shape or inverted U-shape patterns may apply. To assist in addressing such trend behavior in panel data Phillips and Sul (007a) (hereafter P-S)

[1] We mention that this feature of the limit theory is not necessarily the case in multiple regressor models, where stochastic trend effects can manifest in a weaker direction when the limit theory is degenerate, as shown in Park and Phillips (1988).

used the following data generating process

$$y_{it} = b_{it}\theta_t, \tag{14}$$

where θ_t is (typically) assumed to be the dominant (possibly stochastic) common trend and the b_{it} are time varying slope coefficients that induce variations in the trend patterns across the panel. The DGP in (1) can be rewritten in this form as

$$y_{it} = a_i + b_i t + \xi_{it} = (a_i/t + b_i + \xi_{it}/t)t = b_{it}t.$$

If $b_{it} \neq b$ for some i, then TWFE estimation generally suffer from pitfalls such as those considered in Theorem 3. To accommodate this type of practical problem and to allow for more general varying behavior over time and across section while permitting some (possibly unknown) commonalities Morgan et al. (2023) utlize a dynamic clustering mechanism. The next section provides a short review of the mechanism of relative convergence (Phillips and Sul, 007a) that enables this more general approach and a discussion of issues that can arise in the practical implementation of these procedures.

3 Relative convergence tests and convergence clustering

If (14) holds and $y_{it}/y_{jt} \to_p 1$ as $t \to \infty$, then y_{it} is said to relatively converge to y_{jt} over time. Let $\hat{\mu}_t$ be the sample cross-sectional average of the y_{it}. Then, if $y_{it}/\hat{\mu}_t \to_p 1$ as $t \to \infty$ for all i, then we say that the y_{it} relatively converge to their cross-sectional average. In this case, the y_{it} share the same common (stochastic) trend. Relative convergence can be tested using the following so-called $\log t$ regression developed in P-S

$$\log \frac{H_1}{H_t} - 2\log\left(L\left(t\right)\right) = a + b\log t + e_t, \tag{15}$$

where

$$H_t = \frac{1}{n}\sum_{i=1}^{n}\left(h_{it} - 1\right)^2, \text{ for } h_{it} = \frac{y_{it}}{\hat{\mu}_t}, \tag{16}$$

$L(t) = \log t$, $t = p+1, \ldots, T$, $p = \lfloor r \times T \rfloor$ with $r = 1/3$, and $\lfloor \cdot \rfloor$ denoting the floor function.

Under the null of relative convergence, H_t asymptotically converges to zero over time since $h_{it} \to 1$ as $t \to \infty$. Hence, $\log(H_1/H_t)$ is increasing over time. If the t-value for $\hat{b}$ is larger than -1.65, then the null of relative convergence is not rejected at the 5% level. In finite samples, the term $2\log\left(L\left(t\right)\right)$ serves as a penalty function in the regression (15), as explained in P-S. Under relative divergence, H_t and $\log(H_1/H_t)$ should respectively increase and decrease over time. This behavior means that $\hat{b}$ becomes significantly less than zero as t increases. When H_t is simply fluctuating over time, then the dependent variable on the left side of (15) will be decreasing over time due to the presence of the penalty function of $-2\log(L(t))$.

For practical implementation of this $\log t$ regression P-S imposed some restrictions. The first is non-negativity in $y_{it} \geq 0$ for all i and t. The second is that the first 1/3 of the time series observations be discarded in performing the test. The latter condition is naturally restrictive because discarding this much time series data can be expected to lower discriminatory power of the test.[2] Both of these conditions can be relaxed. It is particularly useful to address the second condition and, as discussed shortly, the restriction can be considerably relaxed.

3.1 Measurement units invariance

A further practical issue in performing $\log t$ regressions of the type (15) is that the relative transition parameter, h_{it}, in (16) depends on the choice of measurement units. Suppose, for instance, that y_{it} is log per capita nominal or real GDP for country i-th at time t. In 2020 US per capital nominal GDP was \$63,543.58, a value can be expressed in alternate units as \$63.5K or \$0.0635M. Let k be the measurement unit, which can be any number such as $k = \$1, \$1,000, \text{or } \$1M$. Next, suppose the data generating process is the simple mechanism $y_{it} = b_{it}\theta_t$, where b_{it} is the i-th time varying loading at time t, and θ_t is the dominant common factor trend. We assume that $\theta_t > 0$ for all t, which does not require $y_{it} \geq 0$ for all i and t. If some y_{it} are negative, then the corresponding loading $b_{it} < 0$. Now let

$$y_{it} = \log\left(Y_{it}^* \times k\right) = b_{it}^*\theta_t + \log k, \tag{17}$$

where Y_{it}^* is a latent level variable which is not observable and assume that $\min_{1\leq i\leq n, 1\leq t\leq T} \log Y_{it}^*$ is much smaller than $\log k$. The relative transition parameter h_{it} is then contaminated by the measurement unit k in the following way

$$h_{it} = \frac{y_{it}}{\frac{1}{n}\sum_{i=1}^n y_{it}} = \frac{b_{it}^* + \log k/\theta_t}{\frac{1}{n}\sum_{i=1}^n b_{it}^* + \log k/\theta_t} \neq h_{it}^*, \text{ for any } \; k \neq 1. \tag{18}$$

Further, for large $k \to \infty$, $h_{it} \to 1$, thereby distorting detective capability in testing for relative transition.

To avoid this dependence Sul (2019) suggests normalizing the data by the minimum value of y_{it}, viz.,$y_{\min} = \min_{1\leq i\leq n, 1\leq t\leq T} y_{it}$. Note that

$$y_{\min} = \min_t \min_i (b_{it}^*\theta_t + \log k) = \min_t \min_i (b_{it}^*\theta_t) + \log k = y_{\min}^* + \log k,$$

so that the normalized observations y_{it}^+ satisfy

$$y_{it}^+ = y_{it} - y_{\min} = b_{it}^*\theta_t - y_{\min}^*.$$

With this normalization, y_{it}^+ no longer depends on the nuisance parameter k and

$$h_{it}^+ = \frac{b_{it}^* - y_{\min}^*/\theta_t}{\frac{1}{n}\sum_{i=1}^n b_{it}^* - y_{\min}^*/\theta_t} \approx h_{it}^*. \tag{19}$$

Further, as is apparent in (19), h_{it}^* is close to h_{it}^+ whenever $y_{\min}^*$ is small and/or the trend factor

[2]See Kwak (2021) for a detailed discussion.

θ_t is large. This modification therefore provides a simple relaxation of the first restriction that $y_{it} \geq 0$ for all i and t since by construction we always have $y_{it}^{+} \geq 0$.

3.2 Fixed rule for initial discarding

The reason it is useful to discard some initial observations in the $\log t$ regression is the nature of the $\log t$ function itself. Both the dependent variable, $\log \frac{H_1}{H_t}$, and the $\log t$ regressor originate at zero when $t = 1$, whereas the penalty function $2 \log (L(t))$ with $L(t) = \log t$ is undefined at $t = 1$. Further, by virtue of its concave shape $\log t$ increases rapidly for small t: for the sequence $t = 1, 2, 3, 4, 5,$ we have $\log t = 0, 0.693, 1.099, 1.386, 1.609, 1.792,$ so that the growth rates in the first three values are relatively large compared with those for the larger t values. Hence, if some initial observations are not discarded, estimation of the coefficient b in (15) can become inaccurate, as pointed out in P-S, Remark 3.

We now propose a fixed rule for discarding the data using m, defined as the first sample observation used in the regression, in place of r, the fraction of the sample observations that are discarded. As P-S showed in their Figure 3, and as demonstrated in the next section, the t-ratio test statistic of $\hat{b}$ typically increases with m, rising as more observations m are discarded. Then, under the alternative of divergence, as more data are discarded, the null becomes harder to reject and test power is weaker. The recommendation for the setting $r = 1/3$ in P-S was based on simulation experiments conducted with a simple data generating process under the null. The next section, investigates appropriate settings for the parameter m comparing finite sample performance in more realistic conditions.

There is another benefit to using a fixed m value in practical work. Panel data typically evolve over time. For instance, panel observations that are initially divergent may begin to converge over time or vice versa. To capture such evolution requires some re-examination of relative convergence behavior over the sample. Two common ways of assessing dynamic evolution are through rolling and recursive regression. Convergence analysis based on rolling samples relies on choice of the rolling window and convergence results can fluctuate considerably over the sub-samples. More stable regression outcomes can be obtained with recursive sampling by fixing a starting sample and monitoring the evolution of convergence patterns over time in the cross section one observation at a time. If instead of a fixed starting sample a fixed fraction r of the time series is eliminated, as in P-S, the size of the discarded sample rises with the number of observations included in the regression, leading to less robust use of the data for detecting time variation in the panel.[3] For this reason we recommend using a fixed m value in determining the size of the discarded sample observations.

3.3 Robust estimation of a core convergence club

The original clustering algorithm suggested by P-S requires finding a core convergence club within the panel. Using a cross section ordering based on the final time series observation, the

[3] Movements in the observed t-ratio test for $\hat{b}$ may then be due to discarding more initial observations rather than the inclusion of more sample information.

first two individual series are selected and a $\log t$ regression is run to see if these series relatively converge. If so, an additional individual series is added, based on the last observation ordering, and a second $\log t$ regression is run, testing for relative convergence. This procedure is repeated until the addition of an individual series leads to the rejection of the null of convergence. The core club is identified as the group with the maximum t-ratio from the individual series obtained from this sieve procedure. We call this method the '$\max -t$ rule'. Once this core club is fixed, $\log t$ regressions are run by adding each non-core individual to the core club, one at a time. If the t-ratio in this regression exceeds a certain threshold value τ, then the added individual series is identified as a member of the first convergence club. Successive repetition of this procedure identifies members of the first convergent club. For the remaining series in the panel, the same procedure is repeated to identify a second convergence club, and so on. The overall procedure is called an 'convergence clustering mechanism' (CCM) and readers are referred to P-S for full details concerning its implementation.

Let C_{it} be the club membership for the i-th individual at time t. In earlier discussion we explained the use of recursive sampling in the implementation of the $\log t$ regression test. In the determination of core convergence club membership, the final sample observation ordering may change over time in recursive sampling, which may in turn lead to a change in core convergence club membership. To prevent this ambiguity, we suggest using the full time series sample to determine core club membership.

The steps in applying the dynamic automatic clustering technique are as follows:[4]

Step 1: Using the full sample, find the core group using the $\max t$ rule described in P-S
Step 2: With the fixed core from Step 1, run the CCM with a sub-sample of the data, $t = 1 : T_0$ and record the club membership.
Step 3: Increase the subsample to $t = 1 : T_0 + 1$ and re-run CCM using the fixed core from Step 1
Step 4: Repeat, adding one additional observation at a time until the whole sample is included.

The CCM is built upon an initial core convergent club and its asymptotic justification, as P-S showed, is based on the assumption that the number of core club members goes to infinity as $n \to \infty$. In a finite sample, the size of a core club is fixed, which can be a problem if the size is extremely small, such as only two members.[5] If this outcome arises, the estimated club membership C_{it} can be unstable due to the extremely small size of core members. In this case, users need to check whether the DCCM leads to robust club membership over time. If club memberships fluctuate significantly over time, then the size of the core members should be increased.[6]

[4]A Stata program is available for this 'dynamic convergence clustering mechanism' (DCCM) upon request.

[5]As discussed above, the $\max -t$ rule is initially based on two series of which the last observations are largest. In extremely rare cases there is a chance that adding any other individual does not lead to an increase the $t-$ratio. In such a case, the $\max -t$ rule selects only two series as a core club.

[6]It is easy to alter the $\max -t$ rule by considering only $t-$ratios with more than two series. Consider the following hypothetical example. The $t-$ratio which initially selected two individuals is -6.5. Adding an additional individual – now the size of core becomes three –, of which the last observation is third highest, leads to a $t-$ratio of -4.5. Adding another individual leads to a $t-$ratio to -3.2. Then rather than selecting the initial two individuals, users can select the first three individuals as a core club. If the DCCM leads to robust

4 Monte Carlo Simulations

We consider the following data generating process (DGP) to compare the fixed discarding rule with the fractional rule

$$y_{it} = a_i + b_{it}t + \epsilon_{it}, \text{ with } b_{it} = b_i - (c_0 a_i)t^{-\alpha}, \tag{20}$$

where $\epsilon_{it} \sim \mathcal{N}(0, 0.5)$. Note that the time varying loading, b_{it}, is an inverse function of the fixed effect, a_i, which leads to so-called β-convergence unless c_0=0. If $b_i = b$ for all i, then the y_{it} are relatively convergent as long as $\alpha > 0$, where α is the speed of convergence parameter. The DGP in (20) is slightly different from the DGP considered by P-S in the sense that there is no slowly time varying function $L(t)^{-1}$ in (20). If this function is added in b_{it}, then the two DGPs are asymptotically equivalent. Since P-S did not consider the case under the absence of the $L(t)^{-1}$ function, we exclude the $L(t)^{-1}$ term in the DGP on purpose, but include it in the $\log t$ regression. Technically speaking, the null of convergence without including the $L(t)^{-1}$ term is not well defined asymptotically when $\alpha = 0$. However, the DGP in (20) is more conveniently interpreted such that convergence holds only when $\alpha > 0$. This exclusion may lead to slight over-rejection of the null of relative convergence. As we will show shortly, the null is more often rejected with smaller m and T values. Naturally, this setting leads to a more conservative choice of m. For robustness, we also introduced serial dependence in ϵ_{it}, but we found that there was little difference in convergence results.

We let $c_0 = 0.01$ and $a_i \sim_{iid} \mathcal{U}[0, 5]$. Under the null of convergence, we let $b_i = 0.2$. Under the alternative, we let $b_i = 0.2$ if $i \leq n/2$, and $b_i = 0.4$ otherwise. Two values of the speed of convergence parameter α are assigned: $\alpha = 0.05$ (slow convergence) and $\alpha = 1.1$ (rapid convergence). Next, to evaluate the impact of the data discard parameter m on the size and power of the test, we consider the following $\log t$ regression.

$$\log H_1/H_t - 2\log L(t) = a + bt + u_t \text{ for } t = m+1, \cdots, T. \tag{21}$$

In our experiments here we focus on small sample sizes. If T is large (e.g., $T > 30$), 1/3 of the sample can be discarded since the discard parameters r or m have relatively little influence on the size and the power of the test with large T. We set $n = 100$. The simulation results have little variation with different values of n. The simulation is repeated 5,000 times.

Table 1 reports the frequencies of rejection for the null of convergence under both the null and the alternative. Since the DGP in (20) does not include the slowly varying function, the rejection rate under the null should be higher than the nominal size with a small T. With $m =$ 2 or 3, there is huge size distortion when T is less than 20. The size distortion is worse with a

club membership, there is no need to increase the core membership further. Otherwise, increase the numbers in the core club by including additional individuals until the DCCM produces a robust club classification. As long as the core convergent club is relatively converging, the clustering method is asymptotically well justified.

higher speed of convergence. More interestingly, the size of the test has a U-shape over m when $T < 22$. As T increases, the minimum size of the test is often found with large m values. This result is consistent with the P-S finding that the t-ratio of $\hat{b}$ typically increases as more m are discarded, as mentioned in Section 2.2. With higher t-ratios, we reject the null less frequently, thus minimizing the size of the test. (Recall that this is a one-sided test and the null is rejected when $t_{\hat{b}} < -1.65$.) Under the alternative, the power of the test is nearly 100 with small m. As m increases, the power gets worse, again coinciding with the results that the t-ratio increases with m.

To investigate choice of an optimal value of m we combine power with size for each m. As is shown in Table 1, the optimal value of m is 5 or 6. Note that we cannot calculate the exact size-adjusted power since we intentionally do not include the $L(t)$ function in our DGP.

To address the issue of a measurement unit, and how well the minimum rule works, we treat y_{it} as a latent value, and the actual observations are generated as follows

$$y_{1,it} = y_{it} + \log 10^2, \;\; y_{2,it} = y_{it} + \log 10^3, \;\; y_{3,it} = y_{it} + \log 10^6. \tag{22}$$

Let, $y_{it}^+ = y_{j,it} - \min y_{j,it}$ which is free from measurement units. We treat y_{it} as an unobserved latent variable, y_{it}^*. We select m=5 and 6, following the results in Table 1.

Table 2 reports the simulation results. The performance with $\min y_{it}$ or y_{it}^+ is just slightly worse than that of the infeasible case, y_{it}^*, in terms of the size and power of the test. As k gets larger, the size of the test gets worse. When $k = \log 10^6$, the size becomes almost unity, which implies that the null is falsely rejected 100% of the time. This indicates that measurement unit could in fact be very problematic, particularly with large k, and a strategy was needed to correct for the potential distortion. Overall, the simple modification with y_{it}^+ appears to resolve the measurement unit problem and performs well.

Next, we investigate how the max $-t$ rule affects the finite sample performance of the automatic clustering mechanism. We create two distinct clubs using the DGP in (20), where $a_i \sim_{iid} \mathcal{U}[0,5]$ and $b_i = 0.6$ if $i \in \mathcal{G}_1$, otherwise $a_i \sim_{iid} \mathcal{U}[0,3]$, and $b_i = 0.4$. We then randomly select ten members from $\mathcal{G}_1$ and set them as the core. We call this selection of core members 'infeasible' since in reality true club membership is unknown to the researcher. To highlight the effectiveness of the max $-t$ rule and the automatic clustering mechanism, we then apply both to the created dataset. We allow $n = [50, 100, 200]$ and $T = [40, 50, 70, 100]$, and set the speed of convergence parameter to $\alpha = 0.05$. (The performance with high α values is very similar,so these results are not reported.)

Table 3 shows the results. We calculate the acceptance frequencies of the null of convergence when an additional individual is added to the core club. If the max $-t$ rule and CCM work well, this acceptance rate should be close or equal to one when the added individual is in $\mathcal{G}_1$, thereby accurately detecting club membership. The rate should be zero, or close to it if the added individual is in $\mathcal{G}_2$, since members of $\mathcal{G}_2$ do not converge with the core of $\mathcal{G}_1$.[7] As Table 3 shows, both the max $-t$ rule for selecting the core group and infeasible random selection of core

[7]We found that after filtering the first convergent club, the remaining series always relatively converged. Hence, we did not need to run the automatic clustering method with the rest of the samples.

members lead to perfect performance regardless of the choice of threshold values. A threshold value in this case is the minimum t-ratio that would qualify an individual as having membership in the same club as the core group. Meanwhile, the acceptance frequencies with an individual in $\mathcal{G}_2$ are different. With the max $-t$ rule, the acceptance frequencies approach zero very rapidly as either n or T increases. The random but infeasible rule, however, shows poor performance even though the acceptance rates are decreasing over T and n. This evidence shows how accurate the original CCM is at detecting actual club membership. Also, it is worth noting that the threshold value affects the acceptance rates, as we would expect. With a high value ($\tau = 0$), the acceptance rate is much smaller than with a low value ($\tau = -1.65$).

Last, we examine how recursive estimation effectively detects club membership over time. To investigate this, we change the DGP to the following system with

$$b_{it} = \begin{cases} b_1 + (c_0 a_i)t^{-\alpha} & \text{if } i \in \mathcal{G}_1, \\ b_2 + (c_0 a_i)t^{-\alpha} & \text{if } i \in \mathcal{G}_2, \ \& \ t < T/2 \\ b_2 + (b_1 - b_2)(1 - \tau_t^{-2}) + (c_0 a_i)t^{-\alpha}, & \text{if } i \in \mathcal{G}_2, \ \& \ t \geq T/2 \end{cases} \tag{23}$$

where $x_t = 0$ if $t < T/2$, and $i \in \mathcal{G}_2$, and $x_t = 1 - \tau_t^{-2}$ with $\tau_t = t - T/2$ if $t \geq T/2$, and $i \in \mathcal{G}_2$ initially. Hence, as t exceeeds $T/2$, x_t approaches 1 rapidly so that i joins $\mathcal{G}_1$ after $t > T/2$. We set m=5 and examine the effectiveness of the club clustering technique with differing convergence speeds, α. As we discussed before, once the core club is selected by using the whole sample, the convergence club members are estimated for $T \geq 20$. When $T = 20$, the initial club clustering is estimated and based on this filtering, the previous membership is determined. For example, if y_{1t} is assigned to Club 1 by using $T = 1, ..., 20$, then the first individual is assigned to Club 1 from T=1 to T=20, and the same for Club 2.

Denote $\hat{C}_{it}$ and C_{it} as the estimated and true club memberships for the ith individual at time t, respectively. Note that C_{it} is ambiguous during the transition periods. We assign $b_1 = 0.6$, $b_2 = 0.4$, and $C_{it} = 1$ if $b_{it} \geq 0.55$, otherwise $C_{it} = 2$. The accuracy of club clustering is measured by the following mean and variance losses. Let $B_{it} = \hat{C}_{it} - C_{it}$. Then we have

$$M_1 = \frac{2}{nT} \sum_{i=1}^{n} \sum_{t=1}^{t=T/2} B_{it}, \tag{24}$$

$$V_1 = \frac{2}{nT} \sum_{i=1}^{n} \sum_{t=1}^{t=T/2} [B_{it} - M_1]^2. \tag{25}$$

M_1 measures the biasedness and V_1 shows the efficiency. Note that in (24) and (25) for each t, half of C_{it} is equal to 1 and the other half is equal to 2. To examine the accuracy after treatment, we use the second half of the samples and calculate M_2 and V_2 from $t = T/2 + 1, \cdots, T$.

Table 4 reports the estimated M_j and V_j by Monte Carlo simulation. As is shown in Table 4, a small false inclusion rate – including non-convergent members in Club 1 – occurs with small T. This small mistake blows up by setting club membership from $t = 1$ to $t = 20$. Also as n gets larger, the chances of assigning a wrong membership increase. Due to this problem, a downward bias presents with M_1 when the convergence speed, α, is small. As α increases, this downward bias disappears very quickly. Meanwhile, when $t \geq T/2$, the club memberships are accurately estimated. The estimated M_2 shows little bias even with small T and n. The variances of M_1 and M_2 are small, which implies that the club membership estimation is done accurately. All programs are performed in MATLAB and the codes are available upon request. (including automatic clustering mechanism).

Table 1: Rejection Frequencies with Various m values

	$\alpha = 0.05$					$\alpha = 1.1$				
m/T	18	20	22	24	26	18	20	22	24	26
Size										
2	43.78	15.12	2.64	0.16	0.00	82.14	58.44	33.08	11.70	2.52
3	7.82	0.96	0.14	0.02	0.00	34.40	13.50	3.22	0.70	0.02
4	1.36	0.28	0.00	0.00	0.00	9.60	2.28	0.40	0.10	0.02
5	0.36	0.02	0.00	0.00	0.00	3.20	0.72	0.14	0.00	0.00
6	0.34	0.10	0.00	0.00	0.00	1.78	0.28	0.04	0.00	0.00
7	0.26	0.04	0.00	0.00	0.00	1.30	0.30	0.04	0.00	0.00
8	0.28	0.08	0.00	0.00	0.00	1.50	0.30	0.02	0.00	0.00
9	0.40	0.00	0.00	0.00	0.00	1.36	0.38	0.02	0.00	0.00
10	0.72	0.04	0.02	0.00	0.00	2.08	0.44	0.10	0.00	0.00
Power										
2	100	100	100	100	100	100	100	100	100	100
3	99.96	99.98	100	100	100	99.94	100	100	100	100
4	99.32	99.88	100	99.98	100	99.48	99.84	99.98	100	100
5	98.46	99.74	99.94	100	100	98.54	99.56	99.94	100	99.96
6	96.58	99.16	99.78	99.96	100	96.56	98.94	99.7	99.98	100
7	93.22	97.66	99.64	99.92	100	93.12	97.7	99.34	99.92	100
8	89.08	96.7	99.44	99.92	99.98	89.04	96.14	99.18	99.86	99.92
9	83.3	93.82	98.4	99.78	99.98	83.34	93.16	98.3	99.5	99.86
10	76.8	91.48	97.56	99.56	99.94	75.22	88.76	96.46	99.12	99.9
Power-Size										
2	56.22	84.88	97.36	99.84	**100**	17.86	41.56	66.92	88.30	97.48
3	92.14	99.02	99.86	99.98	100	65.54	86.50	96.78	99.30	99.98
4	97.96	99.60	**100**	99.98	100	89.88	97.56	99.58	99.90	99.98
5	**98.10**	**99.72**	99.94	**100**	100	**95.34**	**98.84**	**99.80**	**100**	99.96
6	96.24	99.06	99.78	99.96	100	94.78	98.66	99.66	99.98	**100**
7	92.96	97.62	99.64	99.92	100	91.82	97.40	99.30	99.92	100
8	88.80	96.62	99.44	99.92	99.98	87.54	95.84	99.16	99.86	99.92
9	82.90	93.82	98.40	99.78	99.98	81.98	92.78	98.28	99.50	99.86
10	76.08	91.44	97.54	99.56	99.94	73.14	88.32	96.36	99.12	99.90

Table 1 Continued –

	$\alpha = 0.05$					$\alpha = 1.1$				
m/T	10	12	14	16	18	10	12	14	16	18
Size										
2	99.02	97.14	91.14	74.20	43.66	99.64	99.26	97.92	94.28	82.32
3	82.68	70.72	50.72	24.90	6.96	89.16	84.08	74.58	55.28	32.20
4	52.88	39.22	20.36	7.46	1.56	62.26	52.98	39.78	23.08	9.44
5	35.50	22.66	10.56	2.96	0.68	42.28	33.40	21.22	10.34	3.66
6	29.52	17.66	6.88	1.58	0.18	32.88	23.88	14.46	5.70	1.64
Power										
2	99.86	99.90	100	99.96	99.98	99.96	99.96	99.98	100	100
3	96.48	98.34	99.34	99.60	99.82	97.82	99.00	99.46	99.80	99.92
4	82.64	91.34	95.88	98.38	99.30	85.68	93.30	97.12	98.88	99.56
5	64.92	78.32	88.74	95.24	98.40	68.96	81.70	91.36	95.42	98.78
6	51.34	65.34	78.94	90.26	95.92	54.18	68.52	81.26	91.06	96.48
Power - Size										
2	0.84	2.76	8.86	25.76	56.32	0.32	0.70	2.06	5.72	17.68
3	13.80	27.62	48.62	74.70	92.86	8.66	14.92	24.88	44.52	67.72
4	**29.76**	52.12	75.52	90.92	**97.74**	23.42	40.32	57.34	75.80	90.12
5	29.42	**55.66**	**78.18**	**92.28**	97.72	**26.68**	**48.30**	**70.14**	85.08	**95.12**
6	21.82	47.68	72.06	88.68	95.74	21.30	44.64	66.80	**85.36**	94.84

Table 2: Rejection Frequencies with Various k values

Size	$m=5$					$m=6$				
	y_{it}^{*}	$\log 10^2$	$\log 10^3$	$\log 10^6$	y_{it}^{+}	y_{it}^{*}	$\log 10^2$	$\log 10^3$	$\log 10^6$	y_{it}^{+}
T=10	34.66	77.08	82.82	88.06	46.36	28.56	57.18	62.58	68.26	35.2
T=12	23.06	83.72	90.40	94.98	40.34	17.52	65.80	74.70	82.82	28.32
T=14	10.66	87.10	94.62	98.50	26.84	6.62	66.12	78.26	89.04	16.32
T=16	3.12	85.84	95.24	99.06	13.96	1.62	65.20	81.38	92.76	7.70
T=18	0.38	81.80	94.94	99.32	5.32	0.32	59.58	81.28	94.86	2.46
T=20	0.04	75.86	94.06	99.44	1.48	0.02	48.88	77.50	94.58	0.38
T=22	0	59.74	90.34	99.48	0.20	0	34.72	69.88	93.98	0.10
T=24	0	43.08	83.40	99.12	0.10	0	20.12	59.16	92.82	0
T=26	0	23.46	73.10	98.26	0	0	8.66	43.78	89.84	0
Power										
T=10	65.16	96.6	98.22	99.18	76.06	52.78	85.12	89.20	93.08	61.16
T=12	78.68	99.8	99.96	100	89.08	65.22	98.14	99.24	99.64	76.90
T=14	88.94	100	100	100	96.02	79.50	99.90	100	100	90.34
T=16	95.22	100	100	100	98.86	89.74	99.98	100	100	96.24
T=18	98.20	100	100	100	99.68	96.28	100	100	100	99.26
T=20	99.38	100	100	100	99.96	99.06	100	100	100	99.88
T=22	99.80	100	100	100	100	99.76	100	100	100	99.98
T=24	99.98	100	100	100	100	99.98	100	100	100	100
T=26	100	100	100	100	100	100	100	100	100	100
Power - Size										
T=10	30.50	19.52	15.40	11.12	29.70	24.22	27.94	26.62	24.82	25.96
T=12	55.62	16.08	9.56	5.02	48.74	47.70	32.34	24.54	16.82	48.58
T=14	78.28	12.90	5.38	1.50	69.18	72.88	33.78	21.74	10.96	74.02
T=16	92.10	14.16	4.76	0.94	84.90	88.12	34.78	18.62	7.24	88.54
T=18	97.82	18.20	5.06	0.68	94.36	95.96	40.42	18.72	5.14	96.80
T=20	99.34	24.14	5.94	0.56	98.48	99.04	51.12	22.50	5.42	99.50
T=22	99.80	40.26	9.66	0.52	99.80	99.76	65.28	30.12	6.02	99.88
T=24	99.98	56.92	16.60	0.88	99.90	99.98	79.88	40.84	7.18	100
T=26	100	76.54	26.90	1.74	100	100	91.34	56.22	10.16	100

Table 3: Effectiveness of $max - t$ Rule and CCM in Detecting Club Membership

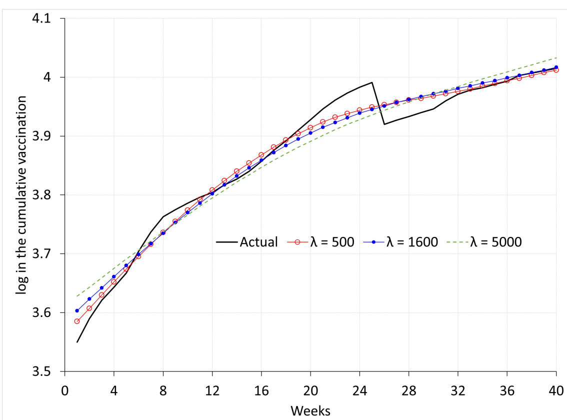
		No. of	$\mathbb{P}\left[t_{\hat{\phi}_i} < \tau \mid i \in \mathcal{G}_1\right]$				$\mathbb{P}\left[t_{\hat{\phi}_i} < \tau \mid i \in \mathcal{G}_2\right]$			
		Core	$\max t$		infeasible		$\max t$		infeasible	
T	n	Members	$\tau = 0$	$\tau = -1.65$	$\tau = 0$	$\tau = -1.65$	$\tau = 0$	$\tau = -1.65$	$\tau = 0$	$\tau = -1.65$
40	50	8.97	1	1	1	1	0.033	0.143	0.678	0.848
50	50	9.32	1	1	1	1	0.021	0.088	0.676	0.837
70	50	9.46	1	1	1	1	0.008	0.032	0.534	0.711
100	50	9.64	1	1	1	1	0.002	0.008	0.285	0.473
40	100	8.59	1	1	1	1	0.002	0.022	0.588	0.785
50	100	9.05	1	1	1	1	0.001	0.009	0.561	0.759
70	100	9.33	1	1	1	1	0.000	0.001	0.416	0.613
100	100	9.55	1	1	1	1	0.000	0.000	0.228	0.400
40	200	8.49	1	1	1	1	0.000	0.008	0.522	0.743
50	200	8.97	1	1	1	1	0.000	0.002	0.496	0.712
70	200	9.22	1	1	1	1	0.000	0.000	0.336	0.530
100	200	9.50	1	1	1	1	0.000	0.000	0.171	0.336

Table 4: Difference between DGP and $\log t$ Club Membership

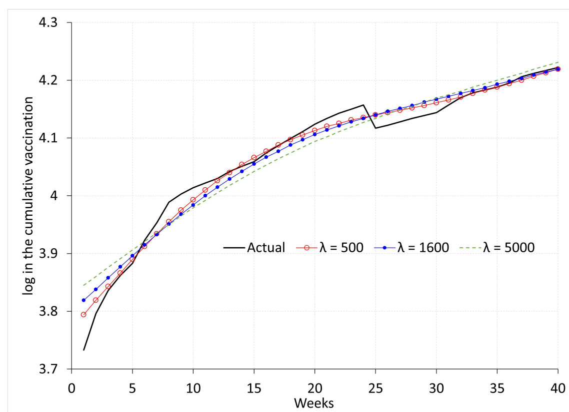
	$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 1.1$		
T/n	50	100	200	50	100	200	50	100	200
M_1									
40	-0.03	-0.01	0.01	-0.04	-0.01	0.03	-0.04	0.01	0.04
50	-0.02	-0.01	0.00	-0.03	0.00	0.02	-0.02	0.01	0.03
70	-0.02	-0.04	-0.06	-0.02	-0.04	-0.02	0.01	0.02	0.03
100	-0.06	-0.12	-0.15	-0.07	-0.11	-0.14	0.00	0.01	0.01
M_2									
40	0.02	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.03
50	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
70	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
100	0.00	-0.01	-0.01	0.00	0.00	-0.01	0.00	0.00	0.00
V_1									
40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
70	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
100	0.03	0.03	0.03	0.02	0.02	0.02	0.00	0.00	0.00
V_2									
40	0.02	0.02	0.01	0.02	0.02	0.01	0.01	0.01	0.01
50	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
70	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00
100	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00

5 Smoothing parameter choice in the HP filter

As mentioned in the main paper, we applied the HP filter to the cumulative vaccination rate data due to a handful of states having small presumed data errors. Here we report how the choice of the HP smoothing parameter affects the DCCM results and the unconditional logit regressions. Figure 1 shows log cumulative vaccination rates for Kentucky and New Mexico. The cumulative vaccination rate in Kentucky suddenly dropped at week 25, and that of New Mexico fell at week 26, which are presumably data errors since cumulative rates cannot decrease. To smooth out these series, we applied the HP filter with the smoothing parameter (λ) of 1600. As Hamilton (2018) and Phillips and Jin (2021) point out, the HP filter does not estimate either the trend or cyclical components consistently. But a boosting iteration (Phillips and Shi, 2021) does successfully estimate trend and cycle consistently and this modification of HP is often well captured by a single extra interation (HP2) or by suitable changes in the smoothing parameter. We therefore used the HP filter in our empirical work and to raise robustness we investigated a range of smoothing parameter choices. As is shown in Figure 1, with a large λ value (greater than 2500), the filtered series look like a simple linear trend, whereas with $\lambda = 500$, the estimated growth components are too close to the original series given the presence of measurement error. In view of these findings we decided to retain the use of the setting $\lambda = 1600$.



(a) Kentucky



(b) New Mexico

Figure 1: Filtered Vaccination Rates with Various λ's

Table 5 shows how the choice of λ changes the results of the DCCM groupings. The overall patterns are very similar: over time, the size of convergent club 1 gets larger. By the 17-th week, all states merge into a single convergent club. There are some slight differences during the transition periods, particularly weeks 14, 15 and 16. Between the settings $\lambda = 1300$ and $\lambda = 1700$, these sensitivities are minor.

Table 6 reports the regression results with various λ values. Overall results are consistent: none of the state level vaccination policy coefficients are significantly different from zero. Only the combined federal mandates and employer mandates variable impact club membership. Among cross section variables, political variables do not influence club membership regardless of

Table 5: Role of HP Filter Smoothing Parameter on DCCM

Week #	λ=500∼1200	λ=1300∼1500	λ=1600∼1700	λ=1800∼2500
13	$\mathcal{G}_0$	$\mathcal{G}_0$	$\mathcal{G}_0$	$\mathcal{G}_0$
14	$\mathcal{G}_1$	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_1$
15	$\mathcal{G}_1$ + MS	$\mathcal{G}_1$	$\mathcal{G}_3$	$\mathcal{G}_3$ + NM
16	All	$\mathcal{G}_3$+ MS	$\mathcal{G}_3$ + MS	$\mathcal{G}_3$ + MS+ NM
17	All	All	All	All

Notes: $\mathcal{G}_0$ = CT, DC, FL, MA, MD, ME, NH, NY, OR, RI, UT, VT, WA.
$\mathcal{G}_1 = \mathcal{G}_0$ + CA,
$\mathcal{G}_2 = \mathcal{G}_1$ + NJ,
$\mathcal{G}_3 = \mathcal{G}_2$ + VA

the value of λ. Significance levels for other control variables vary slightly but these are all cross section variables, so they are not responsible for the dynamic patterns of club memberships.

Table 6: Impact of HP Filter Smoothing Parameter on Logit Regressions
Dependent Variable: Club Membership ($\hat{C}_{it}$)

	λ			
	500∼1200	1300∼1500	1600∼1700	1800∼2500
Federal Mandate and Employer Mandates	205.5*	142.4*	101.0*	102.0*
State Incentives				
Lottery	0.218	0.229	-1.468	-1.137
Cash	-2.298	-2.398	-0.084	0.902
Community Service	-1.687	-1.903	-2.440	-2.383
State Policies				
Vaccine Mandate State Employees	-7.057	-7.149	-1.919	-1.710
Indoor Vaccine Mandate	46.54	18.63	14.33	11.45
Mask Mandate	-2.721	-3.951	2.096	3.220
Ban on Proof of Vaccination	-3.902	-5.469	-3.164	-3.102
Mask Mandate Ban	-1.558	-2.021	-1.810	-1.412
Political				
Percent of State House that is Republican	4.542	9.702	7.037	12.00
Percent of Vote for Trump 2020	-17.13	-16.95	-13.66	-17.07
State Characteristics				
Population Density	11.61*	13.98†	11.36	12.53†
Median Household Income	3.459†	3.723	2.379	3.786†
Percent Foreign Born	14.35*	17.82*	13.26†	12.42*
Percent of People Employeed by Industry				
Health care and social assistance	3.069†	3.516†	3.626†	3.293†
Government and government enterprises	-3.045*	-3.106†	-2.055	-2.378†
Retail trade	16.13*	16.01*	11.90*	13.53*
Wholesale trade	-15.50*	-16.87†	-13.07†	-15.68*
Transportation and warehousing	-10.04*	-10.36†	-9.513	-9.013†

Notes: Median household income is measured in tens of thousands of dollars and population density is per 1,000 square miles. The binary club membership obtained from the dynamic P-S club clustering technique, $\hat{C}_{it}$, is the dependent variable in the logit regressions. The coefficient on the federal-level mandates in the unconditional logit model is large and all state level policies had no impact on the likelihood of being in the high vaccination club. $^{\dagger}p < .05, ^{*}p < .01$

6 Appendix

Proof of Theorem 1: The numerator in (6) has elements involving the demeaned quantity $\dot{\xi}_{it}$ that depend on the stochastic trend $\xi_{it} = \sum_{s=1}^{t} e_{is} = O_p(\sqrt{t})$ for each i. Under the stated conditions, $\frac{1}{\sqrt{T}}\xi_{it=\lfloor Tr \rfloor} \rightsquigarrow B_i(r)$ as $T \to \infty$ for $r \in (0,1]$, where the limit process $B_i(r)$ is Brownian motion with variance ω_{ei}^2, the long run variance of e_{it}, again for each i. It is assumed that the cross-sectional average long run variance $\frac{1}{n}\sum_i^n \omega_{ei}^2$ has a finite positive limit $\omega_e^2 = \lim_{n\to\infty} \frac{1}{n}\sum_i^n \omega_{ei}^2 > 0$. Standard methods give $T^{-3/2}\sum_{t=1}^{T} \xi_{it} \rightsquigarrow \int_0^1 B_i(r)dr$, so that $\frac{1}{\sqrt{T}}\xi_{it=\lfloor Tr \rfloor} - \frac{1}{T^{3/2}}\sum_{t=1}^{T}\xi_{it} \rightsquigarrow B_i(r) - \int_0^1 B_i(r)dr =: \tilde{B}_i(r)$, the demeaned functional of Brownian motion $B_i(r)$. Further, due to cross section independence over i and the fact that $B_i(r) \sim_{i.ni.d} \mathcal{N}(0, \omega_{ei}^2)$, together with the finite limit $\omega_e^2 = \lim_{n\to\infty} \frac{1}{n}\sum_i^n \omega_{ei}^2 > 0$, it follows that as $(n,T) \to \infty$

$$\frac{1}{n\sqrt{T}}\sum_{i=1}^{n}\xi_{it=\lfloor Tr \rfloor} \sim_a \frac{1}{n}\sum_{i=1}^{n} B_i(r) = \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{N}(0,\omega_{ei}^2)\right) = O_p\left(\frac{1}{\sqrt{n}}\right), \tag{26}$$

so that $\frac{1}{n}\sum_{i=1}^{n}\xi_{it} \sim_a O_p\left(\frac{\sqrt{T}}{\sqrt{n}}\right)$ at most, for all $t \leq T$. In a similar fashion and using the fact that $\mathbb{E}\left(\int_0^1 B_i(r)dr\right)^2 = \omega_{ei}^2/3$, we deduce that

$$\frac{1}{T}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\xi_{it} \sim_a \frac{\sqrt{T}}{n}\sum_{i=1}^{n}\int_0^1 B_i(r)dr = \frac{\sqrt{T}}{\sqrt{n}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{N}(0,\omega_{ei}^2/3)\right) = O_p\left(\frac{\sqrt{T}}{\sqrt{n}}\right). \tag{27}$$

With these results in hand and expanding $\dot{\xi}_{it}$ we have the decomposition

$$\begin{aligned}
\dot{\xi}_{it} &:= \xi_{it} - \frac{1}{T}\sum_{t=1}^{T}\xi_{it} - \frac{1}{n}\sum_{i=1}^{n}\xi_{it} + \frac{1}{T}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\xi_{it}, \\
&= \xi_{it} - O_p(\sqrt{T}) - O_p(\frac{\sqrt{t}}{\sqrt{n}}) + O_p(\frac{\sqrt{T}}{\sqrt{n}}) \\
&= O_p(\sqrt{T}) \quad \text{ for all } t \leq T \text{ as } (n,T) \to \infty.
\end{aligned} \tag{28}$$

It follows that when $(n,T) \to \infty$ and $r \in (0,1]$

$$\frac{\dot{\xi}_{i,t=\lfloor Tr \rfloor}}{\sqrt{T}} \sim_a \frac{\xi_{i,t=\lfloor Tr \rfloor}}{\sqrt{T}} - \frac{1}{T}\sum_{t=1}^{T}\frac{\xi_{i,t}}{\sqrt{T}} \rightsquigarrow B_i(r) - \int_0^1 B_i(r)dr =: \tilde{B}_i(r) \tag{29}$$

where $\tilde{B}_i(r)$ is demeaned Brownian motion $B_i(r)$. Result (28) and the limit theory $\frac{\dot{\xi}_{i,t=\lfloor Tr \rfloor}}{\sqrt{T}} \rightsquigarrow \tilde{B}_i(r)$ in (29) rely on joint asymptotics as $(n,T) \to \infty$.

Proceeding with the main argument and noting that, under the maintained conditions on x_{it}, $\frac{1}{T}\sum_{t=1}^{T} x_{it} \sim_a m_{xi} + O_p(T^{-1/2})$, we now find that, as $(n,T) \to \infty$,

$$\dot{x}_{it} := x_{it} - \frac{1}{T}\sum_{t=1}^{T} x_{it} - \frac{1}{n}\sum_{i=1}^{n} x_{it} + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T} x_{it},$$

$$
\begin{aligned}
&= x_{it} - \Big(m_{xi} + O_p(T^{-1/2}\Big) - \frac{1}{n}\sum_{i=1}^{n} m_{xi} - \frac{1}{n}\sum_{i=1}^{n}(x_{it} - m_{xi}) + \frac{1}{n}\sum_{i=1}^{n} m_{xi} + O_p(T^{-1/2}) \\
&= x_{it} - m_{xi} - \frac{1}{n}\sum_{i=1}^{n}(x_{it} - m_{xi}) + O_p(T^{-1/2}) = x_{it} - m_{xi} + O_p(n^{-1/2}) + O_p(T^{-1/2}), \qquad (30)
\end{aligned}
$$

since under the maintained conditions the centered variables $e_{xit} := x_{it} - m_{xi}$ satisfy the CLT $T^{-1/2}\sum_{t=1}^{T}(x_{it} - m_{xi}) \rightsquigarrow \mathcal{N}(0, \omega_{xi}^2)$, where $m_{xi} = \mathbb{E}x_{it}$ is the mean and $\omega_{xi}^2 > 0$ the long run variance of x_{it} for each i, so that $\frac{1}{n}\sum_{i=1}^{n} e_{xit} = O_p(\frac{1}{\sqrt{n}})$. It then follows by virtue of (29) and by standard weak convergence arguments to a stochastic integral that, as $T \to \infty$,

$$
\frac{1}{T}\sum_{t=1}^{T} e_{xit}\dot{\xi}_{it} \sim_a \sum_{t=1}^{T} \frac{\dot{\xi}_{it}}{\sqrt{T}}\frac{e_{xit}}{\sqrt{T}} \rightsquigarrow \int_0^1 \tilde{B}_i(r)dB_{xi}(r) \equiv \mathcal{MN}\left(0, \omega_{xi}^2 \int_0^1 \tilde{B}_i(r)^2 dr\right), \qquad (31)
$$

where, as $T \to \infty$, $\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr \rfloor} e_{xit} \rightsquigarrow B_{xi}(r)$, Brownian motion with variance ω_{xi}^2 for all i, and where $\mathcal{MN}$ denotes mixed normality. Since the limit variates are independent over i with zero mean, variances $\int_0^1 \mathbb{E}\tilde{B}_i(r)^2 dr$, and finite fourth moments by virtue of the properties of Brownian motion, it follows that as $n \to \infty$

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \tilde{B}_i(r)dB_{xi}(r) \rightsquigarrow \mathcal{N}(0, V), \qquad (32)
$$

where

$$
V = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\omega_{xi}^2\int_0^1 \mathbb{E}\tilde{B}_i(r)^2 dr = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\omega_{ei}^2\omega_{xi}^2}{6}, \qquad (33)
$$

which limit exists if $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\omega_{xi}^2\omega_{ei}^2$ exists. Standard calculations, given below in (35), lead to $\mathbb{E}\tilde{B}_i(r)^2 = \omega_{ei}^2\left(\frac{1}{3} - r(1-r)\right)$ and then $\int_0^1 \mathbb{E}\tilde{B}_i(r)^2 dr = \frac{\omega_{ei}^2}{6}$, leading to the stated variance in (33) above. It now follows from results (29) - (32) that as $(n, T) \to \infty$

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{it}\dot{\xi}_{it} \sim_a \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\dot{\xi}_{it}}{\sqrt{T}}\frac{e_{xit}}{\sqrt{T}} \sim_a \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \tilde{B}_i(r)dB_{xi}(r) \rightsquigarrow \mathcal{N}(0, V), \qquad (34)
$$

with V given by (33) above. This result gives us the limit behavior of the standardized numerator in (6).

Expanding the variance calculations given above we now show that $\mathbb{E}\tilde{B}_i(r)^2 = \omega_{ei}^2\left(\frac{1}{3} - r(1-r)\right)$. Observe that

$$
\begin{aligned}
\mathbb{E}\tilde{B}_i(r)^2 &= \mathbb{E}B_i(r)^2 - 2\int_0^1 \mathbb{E}B_i(r)B_i(s)ds + \mathbb{E}\left(\int_0^1 B_i(s)ds\right)^2 \\
&= \omega_{ei}^2 r - 2\omega_{ei}^2\int_0^1 r\wedge s ds + 2\omega_{ei}^2\int_0^1\int_0^r s ds dr \\
&= \omega_{ei}^2 r - 2\omega_{ei}^2\int_0^r s ds - 2\omega_{ei}^2 r\int_r^1 ds + \omega_{ei}^2\int_0^1 r^2 dr
\end{aligned}
$$

$$
\begin{aligned}
&= \omega_{ei}^2 r - \omega_{ei}^2 r^2 - 2\omega_{ei}^2 r(1-r) + \frac{1}{3}\omega_{ei}^2 \\
&= \omega_{ei}^2 r + \omega_{ei}^2 r^2 - 2\omega_{ei}^2 r + \frac{1}{3}\omega_{ei}^2 \\
&= \omega_{ei}^2 \left(\frac{1}{3} - r(1-r)\right), \qquad (35)
\end{aligned}
$$

as required. Then, $\int_0^1 \mathbb{E}\tilde{B}_i(r)^2 dr = \omega_{ei}^2 \int_0^1 \left(\frac{1}{3} + r^2 - r\right) dr = \omega_{ei}^2 \left(\frac{2}{3} - \frac{1}{2}\right) = \frac{\omega_{ei}^2}{6}$, and

$$
\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \omega_{xi}^2 \int_0^1 \mathbb{E}\tilde{B}_i(r)^2 dr = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \frac{\omega_{xi}^2 \omega_{ei}^2}{6},
$$

confirming (33).

From (6) we have the suitably standardized and centered TWFE estimator

$$
\sqrt{n}(\hat{\beta}_{fe} - \beta) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \dot{x}_{it} \dot{\xi}_{it}}{\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \dot{x}_{it}^2}.
$$

The limit distribution of the numerator is given in (34). To obtain the limit distribution of $\sqrt{n}(\hat{\beta}_{fe} - \beta)$, we combine this limit behavior with that of the denominator, which is given by

$$
\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \dot{x}_{it}^2 \sim_a \frac{1}{n} \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T e_{xit}^2 \sim_a \frac{1}{n} \sum_{i=1}^n \sigma_{xi}^2 \to_p \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \sigma_{xi}^2 =: \sigma_x^2, \qquad (36)
$$

and assumed to exist. It follows from results (34) and (36) and by continuous mapping that $\sqrt{n}(\hat{\beta}_{fe} - \beta) \rightsquigarrow \mathcal{N}\left(0, V/\sigma_x^4\right)$, giving the stated result of the theorem. ■

Validation of the consistency of $\hat{V}_\beta$

First, in view of the decomposition (30) and since x_{it} is stationary and ergodic with finite second moments we have

$$
\frac{1}{n} \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \dot{x}_{it}^2 \to_p \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}(x_{it} - m_{xi})^2 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \sigma_{xi}^2 = \sigma_x^2. \qquad (37)
$$

Second, $\hat{\dot{u}}_{it} = \dot{y}_{it} - \hat{b}\dot{x}_{it} = \dot{u}_{it} - (\hat{b} - b)\dot{x}_{it} = \dot{\xi}_{it} + O_p(\frac{1}{\sqrt{n}})$ so that

$$
\hat{\dot{w}}_{iT} = \frac{1}{T} \sum_{t=1}^T \dot{x}_{it} \hat{\dot{u}}_{it} \sim_a \frac{1}{T} \sum_{t=1}^T \dot{x}_{it} \dot{\xi}_{it} \rightsquigarrow \int_0^1 \tilde{B}_i(r) dB_{xi}(r)
$$

and then

$$
\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \hat{\dot{w}}_{iT}^2 \sim_a \frac{1}{n} \sum_{i=1}^n \left(\int_0^1 \tilde{B}_i(r) dB_{xi}(r)\right)^2 \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left(\int_0^1 \tilde{B}_i(r) dB_{xi}(r)\right)^2 + \frac{1}{n} \sum_{i=1}^n \left[\left\{\left(\int_0^1 \tilde{B}_i(r) dB_{xi}(r)\right)^2 - \mathbb{E}\left(\int_0^1 \tilde{B}_i(r) dB_{xi}(r)\right)^2\right\}\right]
\end{aligned}
$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\int_0^1 \tilde{B}_i(r)dB_{xi}(r)\right)^2 + O_p\left(\frac{1}{\sqrt{n}}\right) \to \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\omega_{ei}^2\omega_{xi}^2}{6} \tag{38}$$

Combining (37) and (38) we have

$$\hat{V}_\beta = \left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^2\right)^{-2}\frac{1}{n}\sum_{i=1}^{n}\hat{w}_{iT}^2 = \left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^2\right)^{-2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{it}\hat{u}_{it}\right)^2$$
$$\to \frac{1}{\sigma_x^4}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\omega_{ei}^2\omega_{xi}^2}{6} = V_\beta, \tag{39}$$

as required.

Proof of Theorem 2 We first consider the standardized signal in the denominator of (6). Expanding the standardized signal in the mean adjusted policy variable and using the fact that $\dot{x}_{it} = \dot{x}_{it}^0 + \tilde{c}_i\tilde{t}$, we have the following components in the denominator

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^2 = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{c}_i^2\tilde{t}^2 + \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\dot{x}_{it}^0\right)^2 + 2\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{c}_i\tilde{t}\dot{x}_{it}^0. \tag{40}$$

For the first component of (40) observe that $\tilde{c}_i = c_i - \bar{c} = c_i - m_c + O_p(1/\sqrt{n})$ where $m_c = \mathbb{E}c_i$, so that by virtue of cross section independence and the existence of the moments of c_i we have $\frac{1}{n}\sum_{i=1}^{n}\tilde{c}_i^2 \sim_a \frac{1}{n}\sum_{i=1}^{n}(c_i - m_c)^2 \to_p \sigma_c^2$, as $n \to \infty$. Also $\frac{1}{T}\sum_{s=1}^{T}\frac{s}{T} \to \int_0^1 r dr = \frac{1}{2}$, giving

$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\tilde{t}}{T}\right)^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{t-\bar{t}}{T}\right)^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{t}{T} - \frac{1}{T}\sum_{s=1}^{T}\frac{s}{T}\right)^2 \to \int_0^1 (r - 1/2)^2\, dr = \frac{1}{12},$$

as $T \to \infty$. It follows that

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{c}_i^2\tilde{t}^2 = \frac{T^2}{n}\sum_{i=1}^{n}\tilde{c}_i^2\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\tilde{t}}{T}\right)^2 \sim_a T^2\frac{1}{n}\sum_{i=1}^{n}\tilde{c}_i^2\int_0^1 (r-1/2)^2 dr \sim_a T^2\frac{\sigma_c^2}{12}.$$

In a similar fashion for the third component of (40), assuming that c_i and x_{it} are correlated with $\sigma_{cx} = \mathbb{E}(c_i - m_c)(x_{it}^0 - m_{xi})$, we find that

$$\frac{2}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{c}_i\tilde{t}\dot{x}_{it}^0 \sim_a 2T\frac{1}{T}\sum_{t=1}^{T}\left(\frac{t-\bar{t}}{T}\right)\frac{1}{n}\sum_{i=1}^{n}(c_i - m_c)(x_{it}^0 - m_{xi})$$
$$= 2T\frac{1}{T}\sum_{t=1}^{T}\left(\frac{t-\bar{t}}{T}\right)\sigma_{cx} + \frac{2T}{\sqrt{n}}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{t-\bar{t}}{T}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(c_i - m_c)(x_{it}^0 - m_{xi}) - \sigma_{cx}]$$
$$= 2T\left(\int_0^1 (r - 1/2)dr + O\left(\frac{1}{T}\right)\right)\sigma_{cx} + \frac{2T}{\sqrt{n}}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{t-\bar{t}}{T}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(c_i - m_c)(x_{it}^0 - m_{xi}) - \sigma_{cx}]$$
$$= O(1) + \frac{2T}{\sqrt{n}}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{t-\bar{t}}{T}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(c_i - m_c)(x_{it}^0 - m_{xi}) - \sigma_{cx}] \tag{41}$$

which is at most $O_p\left(\frac{T}{\sqrt{n}}\right)$ since $\int_0^1 (r-1/2)dr = 0$ and by cross section independence and central limit theory $\frac{1}{\sqrt{n}}\sum_{i=1}^n [(c_i - m_c)(x_{it}^0 - m_{xi}) - \sigma_{cx}] \rightsquigarrow \Xi_{cx} \equiv_d \mathcal{N}(0, V_{cx})$ where $V_{cx} = \mathbb{V}ar\left((c_i - m_c)(x_{it}^0 - m_{xi})\right)$ as $n \to \infty$. Finally, for the second component of (40) we have, as in (37) above, $\frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \left(\dot{x}_{it}^0\right)^2 \to_p \sigma_x^2$. Combining these three components of the denominator we deduce that

$$\frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \dot{x}_{it}^2 \sim_a T^2 \frac{\sigma_c^2}{12}. \tag{42}$$

Next consider the numerator of (6). Using $\dot{x}_{it} = \dot{x}_{it}^0 + \tilde{c}_i \tilde{t}$ and $\dot{u}_{it} = \dot{\xi}_{it} + \tilde{b}_i \tilde{t}$, we have the following four components

$$\begin{aligned}
\frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \dot{x}_{it}\dot{u}_{it} &= \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \tilde{c}_i \tilde{b}_i \tilde{t}^2 + \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \tilde{b}_i \tilde{t} \dot{x}_{it}^0 + \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \tilde{c}_i \tilde{t} \dot{\xi}_{it} + \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \dot{x}_{it}^0 \dot{\xi}_{it} \\
&=: I + II + III + IV.
\end{aligned}$$

Considering each term in turn, we have

$$I = \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \tilde{c}_i \tilde{b}_i \tilde{t}^2 = \frac{1}{n}\sum_{i=1}^n \tilde{c}_i \tilde{b}_i \frac{1}{T}\sum_{t=1}^T \tilde{t}^2 = O_p\left(n^{-1/2}\right) \times O\left(T^2\right), \tag{43}$$

when $\mathbb{C}ov(c_i, b_i) = 0$. Otherwise, $I = O_p(T^2)$ and the TWFE estimator $\hat{\beta}_{\text{fe}}$ is inconsistent, as shown below. Next

$$II = \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \tilde{b}_i \tilde{t} \dot{x}_{it}^0 = \frac{T}{\sqrt{n}}\frac{1}{T}\sum_{t=1}^T \frac{t-\bar{t}}{T}\frac{1}{\sqrt{n}}\sum_{i=1}^n (b_i - \bar{b})\dot{x}_{it}^0 = O_p\left(\frac{T}{\sqrt{n}}\right), \tag{44}$$

$$\begin{aligned}
III &= \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \tilde{c}_i \tilde{t} \dot{\xi}_{it} = \frac{T^{3/2}}{n}\sum_{i=1}^n (c_i - \bar{c})\frac{1}{T}\sum_{t=1}^T \frac{t-\bar{t}}{T}\frac{\dot{\xi}_{it}}{\sqrt{T}} \sim_a \frac{T^{3/2}}{n}\sum_{i=1}^n (c_i - \bar{c})\int_0^1 \tilde{r}\tilde{B}_i(r)\,dr \\
&= \frac{T^{3/2}}{\sqrt{n}}\int_0^1 \tilde{r}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (c_i - \bar{c})\tilde{B}_i(r)\,dr\right) = O_p\left(\frac{T^{3/2}}{\sqrt{n}}\right),
\end{aligned} \tag{45}$$

since c_i and ξ_{it} are independent and, as $T \to \infty$, $T^{-5/2}\sum_{t=1}^T (t-\bar{t})\,\dot{\xi}_{it} \to \int \tilde{r}\tilde{B}_i(r)\,dr$, where $\tilde{B}_i(r)$ is given in (29). Finally, as in (34), we have

$$IV = \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \dot{x}_{it}^0 \dot{\xi}_{it} = \frac{1}{n}\sum_{i=1}^n \left(\sum_{t=1}^T \frac{\dot{x}_{it}^0}{\sqrt{T}}\frac{\dot{\xi}_{it}}{\sqrt{T}}\right) = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{46}$$

Combining components (43) – (46), evidently term I dominates the numerator and we have

$$\begin{aligned}
\frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \dot{x}_{it}\dot{u}_{it} &= \frac{T^2}{n}\sum_{i=1}^n \tilde{c}_i \tilde{b}_i \frac{1}{T}\sum_{t=1}^T \left(\frac{\tilde{t}}{T}\right)^2 + O_p\left(\frac{T}{\sqrt{n}}\right) + O_p\left(\frac{T^{3/2}}{\sqrt{n}}\right) + O_p\left(\frac{1}{\sqrt{n}}\right) \\
&\sim_a \frac{T^2}{\sqrt{n}}\int_0^1 \tilde{r}^2 dr \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{c}_i \tilde{b}_i = \frac{1}{12}\frac{T^2}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{c}_i \tilde{b}_i.
\end{aligned} \tag{47}$$

Using the dominant terms (42) and (47) from the denominator and numerator, we obtain the following reduced expression and limit theory for the standardized and centered TWFE estimator

$$\sqrt{n}\left(\hat{\beta}_{\text{fe}}-\beta\right)=\frac{\frac{1}{\sqrt{n}T}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}\dot{u}_{it}}{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^{2}}\sim_{a}\frac{\frac{T^{2}}{12}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{c}_{i}\tilde{b}_{i}}{T^{2}\frac{\sigma_{c}^{2}}{12}}$$
$$=\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{c}_{i}\tilde{b}_{i}}{\sigma_{c}^{2}}\rightsquigarrow\mathcal{N}\left(0,\frac{\sigma_{b}^{2}}{\sigma_{c}^{2}}\right), \tag{48}$$

The limit theory (48) holds when c_i and b_i are independent and each sequence is independent over i. But when the covariance $\mathbb{C}(c_i,b_i)=\mathbb{E}(c_i-m_c)(b_i-m_b)=\sigma_{cb}\neq 0$, we have instead as $n,T\to\infty$,

$$\hat{\beta}_{\text{fe}}-\beta=\frac{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}\dot{u}_{it}}{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^{2}}\to_{p}\frac{\sigma_{cb}}{\sigma_{c}^{2}}, \tag{49}$$

and $\hat{\beta}_{\text{fe}}$ is inconsistent. The same result applies when the trend slope coefficients c_i and b_i are non-random sequences, in which case inconsistency continues to apply with the same limit form given in (49), but with $\sigma_{cb}=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(c_i-\bar{c})(b_i-\bar{b})$ and $\sigma_c^2=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(c_i-\bar{c})^2$ rather than limits in probability. In this case, the inconsistency of $\hat{\beta}_{\text{fe}}$ may be interpreted as the result of spurious trend regression in a panel context. ■

Proof of Theorem 3 From (12) we have the estimation error

$$\hat{\beta}_{\text{fe}}-\beta=\frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}\tilde{b}_{i}\tilde{t}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^{2}}+\frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}\dot{\xi}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^{2}}, \tag{50}$$

The denominator asymptotics in both terms of follow from (36), giving

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{it}^{2}\to_{p}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sigma_{xi}^{2}=\sigma_{x}^{2}, \tag{51}$$

In view of Theorem 1 and the fact that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{it}\dot{\xi}_{it}\rightsquigarrow\mathcal{N}\left(0,V\right),$$

from (34), the second term of (50) has the following asymptotic behavior

$$\frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}\dot{\xi}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^{2}}=\frac{1}{\sqrt{n}}\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{it}\dot{\xi}_{it}}{\frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}^{2}}=O_{p}\left(\frac{1}{\sqrt{n}}\right). \tag{52}$$

Upon standardization and using (30), the numerator of the first term of (50) is

$$\frac{1}{\sqrt{n}T^{3/2}}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{x}_{it}\tilde{b}_{i}\tilde{t}\sim_{a}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(b_i-\bar{b})\sum_{t=1}^{T}\frac{(x_{it}-m_{xi})}{\sqrt{T}}\frac{t-\bar{t}}{T}$$

$$\sim_a \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (b_i - \bar{b}) \int_0^1 (r - \frac{1}{2}) dB_{xi}(r) \rightsquigarrow \mathcal{N}\left(0, \sigma_b^2 \omega_x^2 \int_0^1 (r - 1/2)^2 dr\right) \tag{53}$$

$$= \mathcal{N}\left(0, \frac{\sigma_b^2 \omega_x^2}{12}\right), \tag{54}$$

by standard central limit theory for independent variates using the fact that the b_i and x_{it} (and hence the limit Brownian motions B_{xi}) in (53) are independent. It follows that

$$\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \dot{x}_{it} \tilde{b}_i \tilde{t} = O_p\left(\sqrt{\frac{T}{n}}\right).$$

Combining (50), (51), (52), and (54) we have

$$\hat{\beta}_{\text{fe}} - \beta = O_p\left(\sqrt{\frac{T}{n}}\right) + O_p\left(\frac{1}{\sqrt{n}}\right), \tag{55}$$

so that $\hat{\beta}_{\text{fe}}$ is inconsistent and divergent when $n = o(T)$. It is clear from (50) and the above analysis that heterogeneous deterministic trends in y_{it} continue to dominate and $\hat{\beta}_{\text{fe}} - \beta = O_p\left(\sqrt{T/n}\right)$ still holds when the outcome variables y_{it} have stationary innovations ξ_{it}. ■

References

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372.

Durlauf, S. N. and Phillips, P. C. B. (1988). Trends versus random walks in time series analysis. *Econometrica*, 56:1333–1354.

Hamilton, J. D. (2018). Why you should never use the Hodrick-Prescott filter. *Review of Economics and Statistics*, 100(5):831–843.

Kwak, J. (2021). A new approach to the relative convergence test. *Applied Economics Letters*, pages 1–7.

Morgan, L. B., Phillips, P. C. B., and Sul, D. (2023). Policy evaluation with nonlinear trended outcomes: Covid-19 vaccination rates in the us. Mimeo, University of Texas at Dallas.

Park, J. Y. and Phillips, P. C. B. (1988). Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory*, 4(3):468–497.

Phillips, P. C. B. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics*, 33(3):311–340.

Phillips, P. C. B. and Jin, S. (2021). Business cycles, trend elimination, and the hp filter. *International Economic Review*, 62(2):469–520.

Phillips, P. C. B. and Shi, Z. (2021). Boosting: Why you can use the hp filter. *Inter-*

national Economic Review, 62(2):521–570.

Phillips, P. C. B. and Sul, D. (2007a). Transition modeling and econometric convergence tests. *Econometrica*, 75(6):1771–1855.

Sul, D. (2019). *Panel data econometrics: Common factor analysis for empirical researchers.* Routledge.