Why the Two-way Fixed Effects Model is Difficult to Interpret, and What to Do About It[*]

Jonathan Kropko
University of Virginia
jkropko@virginia.edu

Robert Kubinec
University of Virginia
rmk7xy@virginia.edu

May 14, 2018

Abstract

The two-way fixed effects (FE) model, an increasingly popular method for modeling time-series cross-section (TSCS) data, is substantively difficult to interpret because the model's estimates are a complex amalgamation of variation in the over-time and cross-sectional effects. We demonstrate this complexity in the two-way FE estimate through mathematical exposition. As an illustration, we develop a novel simulation that enables us to generate TSCS data with varying over-time and cross-sectional effects and examine the behavior of the two-way FE model as these effects change. We demonstrate that the two-way FE model makes specific assumptions about TSCS datasets, and if these assumptions are not met, the model may be unidentified even if substantial variation exists along both dimensions. Because of the difficulty in interpretation, we do not recommend that applied researchers rely on the two-way FE model except for situations in which the assumptions are well-understood, such as the canonical difference-in-difference design.

[*] A previous version of this manuscript was presented at the annual meeting of the American Political Science Association, Philadelphia, PA, September 2, 2016. We thank Tom Carsey, Jonathan Katz, Marius Radean, Daniel Gingerich, Jeffery J. Harden, Justin Esarey, and Justin Kirkland for their comments and advice.

1 Introduction

Time-series cross-section (TSCS) data is collected to establish important relationships between independent and dependent variables that have wide generalizability. To that end, applied researchers select models for these datasets with the goal of producing clear and interpretable estimates, but the existing literature does not always provide clear guidance for interpreting standard variants of the widely-used linear model, in particular one and two-way fixed effects (FE) models. One reason for the proliferation of linear modeling strategies for TSCS data is that different models yield results with different interpretations, speaking to different dimensions of variance within the data.

We address this shortcoming through a novel mathematical exposition of the two-way FE model with an accompanying simulation that shows why the model's estimates are difficult to interpret in terms of the underlying dimensions of variance in the TSCS dataset. Without a clear interpretation, the original aim of the data collecting enterprise — establishing relationships between variables — is difficult to accomplish even with unbiased estimators. For many research designs, the two-way FE model does not provide an interpretation that directly addresses the research question.

One motivation for selecting a FE model, and the two-way FE model in particular, is to choose a model that addresses omitted variable bias by removing unobserved confounders that are fixed over time or across cases.[1] It follows in this approach that two-way FE models must be preferable to one-way FE models because they exclude more omitted variables by virtue of including additional intercepts. But we argue that in addition to accounting for omitted variables, FE specifications also change the interpretation of coefficients in a linear model. To be specific, one-way FE specifications, which place the FEs on cases or on time points but not both, operate unambiguously on one of the two dimensions variation in the data. Case FEs only analyze temporal variation and time point FEs only consider cross-sectional variation. As such, case FE coefficients can be interpreted as the average change in y as x increases by one-unit *over time*, and time FE coefficients can be interpreted as the average change in y as x increases by one-unit *between cases*. In contrast, we show that the two-way FE model, when identified, is a complex amalgamation of the cross-sectional and over-time effects in the data. In this paper we derive and present the correct interpretations of two-way FE coefficients in terms of the actual variation that exists in TSCS datasets.

Furthermore, we consider the relationship between the two-way FE model and the difference-in-difference estimator. We provide a separate interpretation of two-way FE coefficients from this perspective and discuss the conditions that are needed for these coefficients to be valid difference-in-difference (DiD) estimates.

As an illustration of how the two-way FE model combines cross-sectional and temporal variance, we develop a novel simulation that enables us to generate TSCS data with varying over-time and cross-sectional effects and examine the behavior of the two-way FE model as these effects change. We demonstrate through this evidence that

[1] We define a case to be the unit of analysis in the data that is observed at repeated points in time. For example, cases may be countries, individual respondents in a longitudinal survey, elected officials, and so on.

the two-way FE model makes very strong assumptions about TSCS datasets, and if these assumptions are not met, the model can be unidentified even if substantial variation exists along both dimensions. This unidentifiability has not been previously diagnosed because common statistical software packages employ techniques to fix the estimation in ways that are not transparent to the user.

We urge researchers to choose between models by taking into consideration the way in which parameters should be interpreted in addition to estimation properties such as bias, consistency, and efficiency. Because of the restrictive assumptions and difficulty in substantive interpretation of the two-way FE model, we do not recommend that applied researchers rely on this model except for situations in which the model's interpretation exactly matches the researcher's intended research question and the model's assumptions are well-understood, such as the canonical difference-in-difference design.

2 The Use of One-Way and Two-Way FEs in Political Science

We collected methodological information from all articles that appeared in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics* for the period 1976 to 2014.[2] We used the search term "fixed effect" to create a dataset of 363 articles. This set is likely a conservative selection from the population, as some researchers may employ fixed effects in their estimations without using those words. Furthermore, to ensure that we were focusing only on this particular subset of TSCS models, we excluded any hierarchical or random-effects models from the dataset.

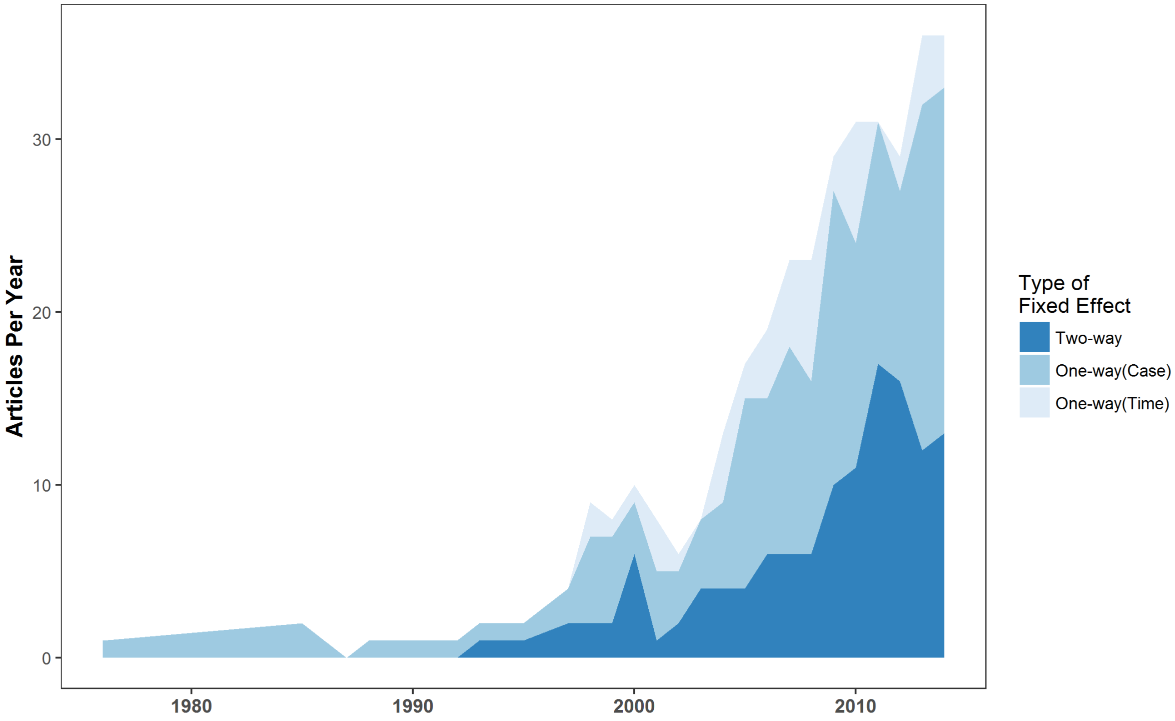
Figure 1 shows a steady rise in the usage of fixed effects models among top journals, a trend that has continued through the 2000s. The vast majority of fixed effects models use either case FEs or both time and case FEs, which we define as the two-way FE estimator. Models that only use time FEs are comparatively rare. The year-on-year growth in fixed effects models is likely due to two sources: 1) the greater availability of TSCS datasets to political scientists, which we regard as a great asset to the discipline, and 2) the growing awareness among practitioners of certain features of fixed effects models to decipher the complexity of TSCS variables. What is also of note is that fixed effects models are widely used across the empirical subfields.[3]

Given this context, it is important to note that the time-series cross-section literature as cited by these papers is heavily influenced by a limited number of articles. The first published paper on the topic in a political science journal, and a still very influential work, was written by Stimson (1985) and presented the least-squares with dummy variables (LSDV) approach. Stimson argued that researches should incorporate FEs in OLS models as a means of apportioning the variance within these datasets, although he did not provide specific recommendations as to what kind of one or two-way FEs to use. Beck and Katz (1995) and Beck, Katz, and Tucker (1998) advanced the conversation considerably by providing more specific recommendations for how political scientists should run TSCS

[2] The bulk of the articles were collected through JSTOR's Data for Research service, while the most recent years were collected through each journal separately.

[3] The full meta-analytical data will be made publicly available upon publication of this manuscript.

Figure 1: Type of Fixed Effects Employed by Political Science Articles (1976-2014)



models depending on the structure of the data, i.e., different ratios of cases versus time points. More recently, alternatives to traditional fixed effects models, such as estimation of “rarely changing variables” (Plümper and Troeger 2007) and difference-in-difference designs (Abadie 2005) have diversified fixed effects models to an extent, although the original papers by Beck and Katz and Stimson are still very influential.

Despite the wide uniformity of citations in fixed effects papers, there exists substantial diversity in the interpretation and justification of FE models. Generally speaking, the econometric notion that fixed effects estimators exist to correct for “unobserved heterogeneity” (Wooldridge 2010, 251) has meant that authors have come to prefer fixed effects for its ability to assist in causal identification, that is to say, to remove potential confounders. However, some scholars have struggled to incorporate FE models even though they value this benefit of FE analysis. During our literature review, we encountered several papers that were removed from our dataset because the author wanted to use a FE model, but instead employed a different TSCS model because the FE model failed to produce estimates, usually because of a lack of variation in the variable of interest. As a consequence, we believe that even more authors would use fixed effects in their analysis if their data permitted it.

To take one example, Donno (2013) does not use a one-way case FE model because “fixed effects leads to too many observations dropping from the analysis” (p. 709). In Donno’s case, the independent variable of interest is electoral system, a variable which will only rarely change over time within countries. As a result, the inclusion of case FEs results in all countries dropping from the model that did not experience at least one change in electoral system, such as the United States. As we explain elsewhere in this paper, the problem is not that variation in electoral systems does not exist, but rather that it mostly does not exist within cases, which is the dimension of variation on which the case FE estimator operates. This preference for case FEs, even in situations when the variation appears more appropriate for time FEs, is what we argue explains the small number of time FE models in Figure 1.

Regardless, time FE models have been and continue to be used in political science research, especially within certain research domains. For example, in the American political science literature on courts, it is common to include fixed effects for the tenure of a court, administration or Congress (Jenkins and Monroe 2012; Hall 2014; Boyd, Epstein, and Martin 2010; Lazarus 2009). We demonstrate below that the time FE estimator unambiguously compares cases to one another at the same point in time, a comparison that speaks directly to many important substantive research questions.

In recent years, the preference for case FEs has broadened to include the two-way FE model, which is seen as an even more stringent standard for causal inference. For example, Gabel et al. (2012) argue that they need both case and time FEs because they want to control for "unobserved national factors" (case FEs) and "temporal shocks" (time FEs) (p. 1132) simultaneously. Similarly, Scheve and Stasavage (2012) use two-way FEs within their difference-in-difference framework because case FEs "control for time-constant unobserved country-level heterogeneity" while time FEs will account for "common shocks" (p. 82-83). In other words, the reason for the growing use of two-way FEs is not because it is seen as a method suited for a particular form of variation, but rather as a form of control that can exclude unmeasured covariates. The logical conclusion of this thinking is that the two-way FE model presents the gold standard for TSCS causal identification, a sentiment we found increasing over time in the literature. For example, Blaydes and Kayser (2011) use estimates from both the one-way case FE and two-way FE models while positing that the two-way FE model "sets a high bar for most models to clear" precisely because it absorbs so much variation (p. 899).

While related to concerns over causal identification, a second driver behind the rising use of two-way FE models relates to a distinct causal paradigm, that of the difference-in-difference (DiD) design (Bechtel and Hainmueller 2011; Anzia and Berry 2011; Condra and Shapiro 2012; McGhee et al. 2014; Truex 2014). Although two-way FEs and DiD models are often referred to interchangeably, they do not have any necessary connection to each other, as one is a statistical model and the other a research design. The canonical DiD research design posits two groups (treatment and control) observed at two time points, the pre-treatment and post-treatment observations (Khandker, Koolwal, and Samad 2010). Two-way FE models, as we explain later, represent a particular combination of case and time variation that returns a DiD estimate under this canonical design, but does not estimate the DiD in more general contexts unless very strong model-based assumptions are used.[4]

[4] Furthermore, two-way FEs are not the only way that researchers can estimate a DiD. The standard DiD design can also be estimated by either a difference of means t-test or a one-way time FE model depending on how the treatment variable is coded in the dataset (Khandker, Koolwal, and Samad 2010).

3 Interpretation of One-way FE Coefficients

There are several ways to implement fixed effects that are equivalent to including dummy variables for cases or for time points. The one-way case FE estimator can be derived by subtracting the mean across observations within each case (Cameron and Trivedi 2005, 726; Greene 2012, 361; Wooldridge 2010, 269). In other words, we transform the outcome as follows:

$$y_{it}^* = y_{it} - \bar{y}_i = y_{it} - \frac{1}{|T_i|} \sum_{t \in T_i} y_{it}. \tag{1}$$

Here T_i is the set of time points observed for case i and $|T_i|$ is the number of time points in this set. This model removes all possible covariates that vary across cases but are fixed across time, regardless of whether or not those covariates are observed. Consider for example an outcome given by a linear model

$$y_{it} = \alpha + \beta x_{it} + \delta u_i + \varepsilon_{it} \tag{2}$$

that includes an independent variable x_{it} that varies both across cases and time points, and an independent variable u_i that varies across cases but is fixed across time. If we apply the transformation in equation 1 to this linear equation, the result is

$$\begin{aligned} y_{it} - \frac{1}{|T_i|} \sum_{t \in T_i} y_{it} &= \Big(\alpha + \beta x_{it} + \delta u_i + \varepsilon_{it}\Big) - \frac{1}{|T_i|} \sum_{t \in T_i} \Big(\alpha + \beta x_{it} + \delta u_i + \varepsilon_{it}\Big) \\ &= \beta \Big(x_{it} - \frac{1}{|T_i|} \sum_{t \in T_i} x_{it}\Big) + \Big(\varepsilon_{it} - \frac{1}{|T_i|} \sum_{t \in T_i} \varepsilon_{it}\Big). \end{aligned}$$

Importantly, the time-fixed covariate u_i drops out of the model. The coefficient β in equation 2 is then estimated by OLS to be

$$\hat{\beta}_{\text{caseFE}} = \frac{\displaystyle\sum_{i=1}^{N} \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\displaystyle\sum_{i=1}^{N} \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i)^2}. \tag{3}$$

Likewise, the one-way time FE estimator subtracts the mean across observations within each time point,

$$y_{it}^* = y_{it} - \bar{y}_t = y_{it} - \frac{1}{|N_t|} \sum_{i \in N_t} y_{it} \tag{4}$$

where N_t is the set of cases observed at time point t and $|N_t|$ is the number of cases in this set. Unlike the case FE model, the time FE model cannot eliminate a time-fixed covariate u_i; however it does eliminate variables v_t that vary over time but are fixed across cases. The OLS estimate of a coefficient β on a covariate x_{it} in a time FE model is

$$\hat{\beta}_{\text{timeFE}} = \frac{\displaystyle\sum_{t=1}^{T}\sum_{i=1}^{N_t}(x_{it} - \bar{x}_t)(y_{it} - \bar{y}_t)}{\displaystyle\sum_{t=1}^{T}\sum_{i=1}^{N_t}(x_{it} - \bar{x}_t)^2}. \tag{5}$$

Going forward, we refer to this operationalization of fixed effects as the *mean-centering* approach.

Another way to implement fixed effects is what we call the *data subsetting* approach. Case FEs only work with the time series — not the cross-sections — in the data because a coefficient from this model is a weighted average of the coefficients we obtain by subsetting the data by case. To demonstrate this point, note that we can multiply and divide a factor of $\sum_{t=1}^{T_i}(x_{it} - \bar{x}_i)^2$ within the summation across cases in the numerator in equation 3,

$$\hat{\beta}_{\text{caseFE}} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i}(x_{it} - \bar{x}_i)^2 \frac{\sum_{t=1}^{T_i}(x_{it}-\bar{x}_i)(y_{it}-\bar{y}_i)}{\sum_{t=1}^{T_i}(x_{it}-\bar{x}_i)^2}}{\sum_{i=1}^{N}\sum_{t=1}^{T_i}(x_{it} - \bar{x}_i)^2}, \tag{6}$$

which can be rewritten as

$$\hat{\beta}_{\text{caseFE}} = \frac{\sum_{i=1}^{N}\omega_i\hat{\beta}_i}{\sum_{i=1}^{N}\omega_i} \tag{7}$$

where $\hat{\beta}_i$ is the OLS coefficient estimated using only the data within one case, and

$$\omega_i = \sum_{t=1}^{T_i}(x_{it} - \bar{x}_i)^2 = T_i \times V(x_{it}). \tag{8}$$

This result implies that running a case FE model is equivalent to taking these steps:

1. Consider only the observations for case 1. Since all of these observations come from the same case, all of the variation exists over time.
2. Regress y_{1t} on x_{1t} using this subset of the data, and record the coefficient β_1.
3. Calculate the variance of the values of x for case 1 and the sample size within this subset T_1, and record the product $\omega_1 = T_1 \times V(x_{1t})$.
4. Repeat steps 1, 2, and 3 for every case in the data.
5. Calculate the case FE coefficient using equation 7 by taking the average of every case-specific coefficient β_i, weighted by ω_i.

Likewise, the time FE estimator produces a coefficient estimate that is a weighted (by variance and sample size) average of the coefficients we calculate for each cross-section.[5]

The data subsetting approach to operationalizing fixed effects leads to a clear interpretation of one-way FE coefficients. Within the data for one case, for instance, all variation must occur over time, so a regression coefficient

[5] We present this argument again as a formal mathematical proof in Appendix A.1 in the supplemental material.

within this subset must be interpreted as the average effect of a unit-increase in x on y as each variable changes over time for this specific case. Because case FE coefficients average these corresponding coefficients across all cases, a case FE coefficient represents the average effect of a unit-increase in x on y as each variable changes over time, generalized to all cases. Similarly, within one time point in the data all variation is cross-sectional, so a regression coefficient within this subset must be interpreted as the average effect of a unit-increase in x on y as each variable changes from case to case at this specific point in time. Therefore time FE coefficients represent the average effect of a unit-increase in x on y as each variable changes from case to case, generalized across all time points.

For example, consider the ongoing debate about whether economic development in a country affects the quality of that country's democracy (Acemoğlu and Robinson 2006; Acemoğlu et al. 2008; Andersen and Ross 2014; Boix 2011, 2003; Haber and Menaldo 2011; Houle 2009; Inglehart and Welzel 2005; Kennedy 2010; Limongi and Przeworski 1997). With TSCS data we can investigate individual countries over time or particular cross-sections of countries in a specific year. If we look only at the time series for India, we ask "as GDP increases for India over time, how does the quality of its democracy change?" If we use case FEs, we generalize this question across countries: "as GDP increases for *a country* over time, how does the quality of its democracy change?" If instead we look at the cross-section that exists in 1990, we ask "how much more democratic are wealthier countries than poorer countries in 1990?" Time FEs generalize this question to the entire time frame under analysis, and simply ask "how much more democratic are wealthier countries than poorer countries?" If a researcher intends to compare one case to itself over time, it is appropriate to examine individual time series and to use case FEs; if a researcher intends to compare one case to another at the same point in time, it is appropriate to examine cross-sections and to use time FEs.

Employing a one-way FE model in a way that can answer the research question, on the other hand, does not guarantee that it will do so. That is, selecting a model with a correct interpretation is a necessary but not a sufficient condition for successful statistical analysis. Indeed, time series in TSCS data may have all of the well known problems of time series in non-panel contexts: seasonality, non-stationarity, stochastic volatility, and so on. Likewise, cross-sections in TSCS data may exhibit reverse causality, heteroskedasticity, multicollinearity, etc. Both time series and cross-sections can also suffer from omitted variable bias if there are unmeasured confounders along the dimension of variance in the model. While these issues must be addressed, they must be addressed in a way that still allows researchers to interpret the model.

4 Interpretation of Two-way FE Coefficients

The two-way FE model includes dummy variables for every case and for every time point in the same equation,

$$y_{it} = \alpha + \beta x_{it} + \sum_{j=2}^{N} \gamma_j I_{i=j} + \sum_{w=2}^{T} \lambda_w I_{t=w} + \varepsilon_{it}, \tag{9}$$

where $I_{i=j} = 1$ if observation (i, t) belongs to case j and 0 otherwise, $I_{t=w} = 1$ if observation (i, t) exists at time point w and 0 otherwise, and one of each set of dummy variables is omitted as a reference group.

As with one-way FE models, we can alternatively express the same model by using a mean-centering or a data subsetting approach to operationalizing the fixed effects along each dimension. First, we mean-center the variables within both cases and time points by subtracting the mean along one dimension, then subtracting the mean of these differences along the other dimension. If the panels in the data are balanced,[6] then Greene (2012, 364) and others write the two-way FE coefficient estimate that we denote $\hat{\beta}_{TW}$ as

$$\hat{\beta}_{TW} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \bar{y}_i - \bar{y}_t + \bar{y})(x_{it} - \bar{x}_i - \bar{x}_t + \bar{x})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i - \bar{x}_t + \bar{x})^2}. \tag{10}$$

Covariates that are fixed across time are removed from the model through the subtraction of the case-means of each variable. Simultaneously, covariates that are fixed across cases are removed through the subtraction of the time-means.

The most useful interpretation of coefficients from the two-way FE model is revealed when we apply mean-centering to one dimension and data subsetting on the other.[7] Without loss of generality, we mean-center with regard to cases and we subset with regard to the time points. Given N cases, T time points, and data y_{it} and x_{it}, mean-centering requires transforming the variables as follows:

$$y^*_{it} = y_{it} - \bar{y}_i, \qquad x^*_{it} = x_{it} - \bar{x}_i.$$

Then, to apply data subsetting on time points, we consider each time point $t \in \{1, \ldots, T\}$ individually and for each we calculate a coefficient β_t from the equation

$$y^*_{it} = \alpha_t + \beta_t x^*_{it} + \varepsilon^*_{it}. \tag{11}$$

A two-way FE coefficient has an interpretation that generalizes the interpretation of β_t in equation 11 across time. So, to understand what a two-way FE coefficient means, we must understand what β_t means.

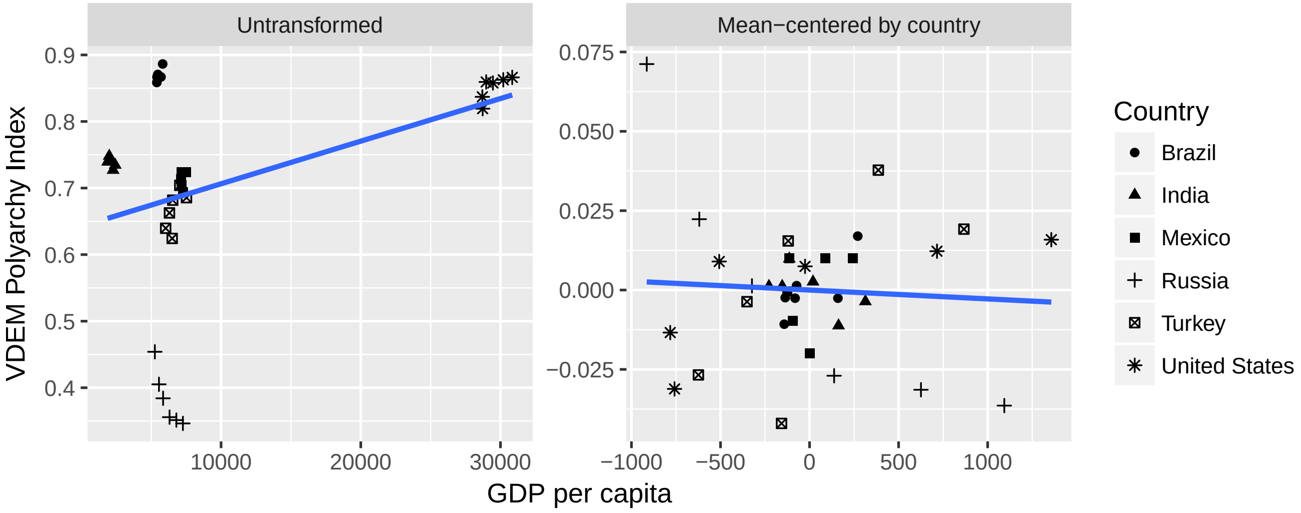
For clarity, we use real data to illustrate the proper interpretation of the two-way FE coefficient. We return to the example of GDP and democracy: we employ the Varieties of Democracy polyarchy index for the quality of a country's democracy, and the measure of GDP per capita included in the VDEM data (Coppedge et al. 2017). We keep six countries — Brazil, India, Mexico, Russia, Turkey, and the United States — and the six years from 2000

[6] Balanced means that exactly the same time points are observed for every case. We dispel this assumption in appendix B.3 in the supplemental material.

[7] It is not possible to use data subsetting on both cases and time points, because by definition in TSCS data there is only one observation per case and time, which is insufficient to identify a regression coefficient.

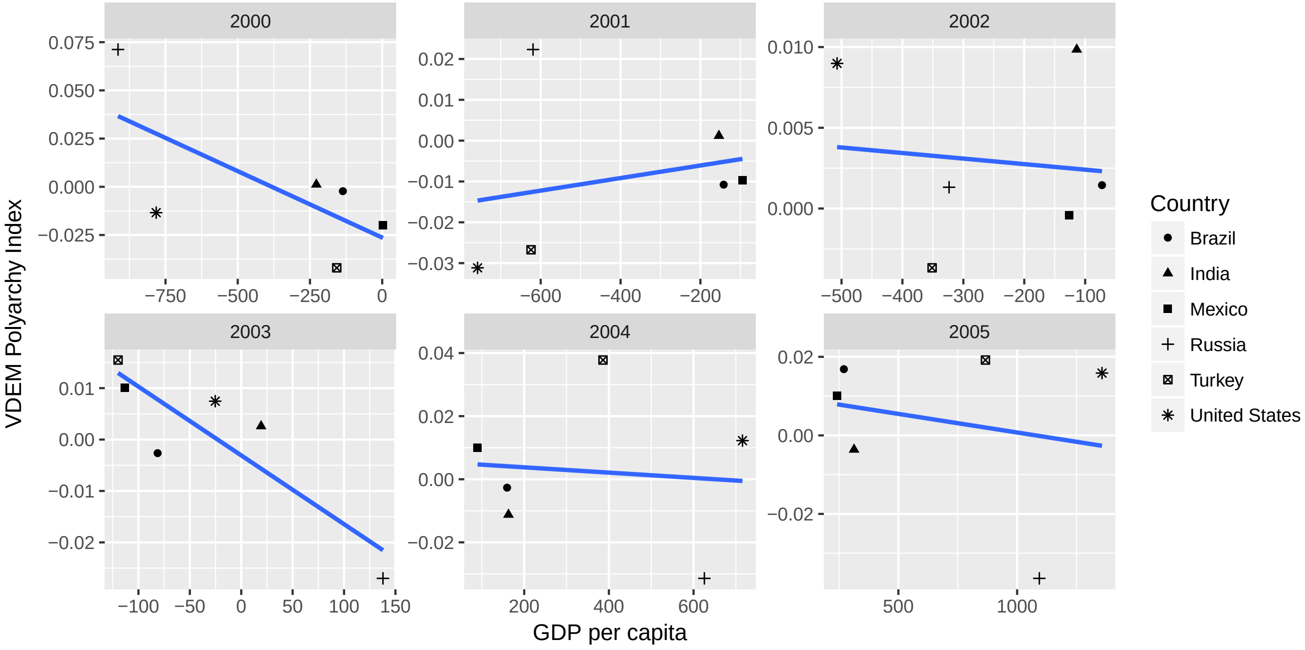
through 2005. Figure 2 includes two panels. The left-hand panel is a scatterplot of each variable, untransformed. The slope of the best-fit line in the left-hand panel is the pooled OLS coefficient. In the right-hand panel, the country-specific means have been subtracted from both democracy and per capita GDP. The slope of the best-fit line in the right-hand panel is the coefficient from the case FE model.

Figure 2: GDP and Democracy Data from the Varieties of Democracy Dataset, 2000-2005



To proceed from the case FE coefficient to the two-way FE coefficient, we subset the data in the right-hand panel of Figure 2 by year. These six scatterplots are displayed in Figure 3.

Figure 3: Subsetting the Country-Mean Centered Data by Year



Each best-fit line in each panel represents another entry for β_t in equation 11. The two-way FE coefficient is an average of these slopes, weighted by the amount of data in each scatterplot times the variance of the x-values in each plot. But to understand the substantive meaning of the two-way FE coefficient, it is necessary to describe the substantive meaning of the slope in each of the plots in Figure 3. Consider the plot for the year 2000. The x-axis represents how, in the year 2000, a single country's GDP per capita compares to that country's mean GDP per capita from 2000-2005. Likewise, the y-axis represents how, in the year 2000, a single country's democracy index compares to that country's mean democracy index from 2000-2005. The negative slope in 2000 means that, on average, relative to a country with a GDP per capita that is farther below its own over-time mean, another country with a GDP per capita that is closer to *its* over-time mean will have a democracy index that is farther below that other country's democracy index's over-time mean. The two-way FE coefficient generalizes this interpretation to all years between 2000 and 2005 by calculating the weighted average of the six slopes that appear in Figure 3. The previous two sentences represent our best effort to provide an intuitive expression of $\hat{\beta}_{TW}$ in equation 10.

This interpretation will often be difficult to communicate and to understand. The difficulty arises because the interpretation requires two dimensions of comparison, not just one. GDP per capita is negative *relative to the country's over-time average*, so we compare a country to itself as it changes over time. But then, by regressing relative democracy on relative GDP per capita for the six countries, the two-way FE coefficient ultimately expresses how one country's GPD per capita and democracy, relative to itself, compares to another country's GDP per capita and democracy, relative to *it*self. If this interpretation does not match the question the model is intended to answer, then we suggest that applied researchers employ methods with interpretations that directly answer the research question.

This challenge to interpretation is not meant to imply that the two-way FE model cannot be profitably applied to TSCS data. There are interpretations of the two-way FE model that can speak directly to research designs employed by political scientists, most notably the difference-in-differences design, which also generates results that express two levels of comparison. In fact, the two-way FE model has often itself been described as a DiD estimator although there are important differences between a two-way FE model and a DiD design. We consider the similarities and differences between two-way FE models and DiD designs in section 4.1.

Our discussion of two-way FE coefficients means that these coefficients combine the effects we calculate by comparing cases to one another and by comparing how each case changes relative to itself over time. In other words, the two-way FE model is a complex amalgamation of cross-sectional and temporal effects in TSCS data. This idea accords with Imai and Kim (2016), who show that two-way FE coefficients can be expressed as a weighted average of the coefficients from case FEs, time FEs, and pooled OLS. Specifically, in balanced panels this average is

$$\beta_{TW} = \frac{\omega_{\text{caseFE}}\,\beta_{\text{caseFE}} + \omega_{\text{timeFE}}\,\beta_{\text{timeFE}} - \omega_{\text{pooled}}\,\beta_{\text{pooled}}}{\omega_{\text{caseFE}} + \omega_{\text{timeFE}} - \omega_{\text{pooled}}}$$

where $\omega_{\text{caseFE}} = (T-1)\sum_{i=1}^{N} \hat{V}_i(x)$, $\omega_{\text{timeFE}} = (N-1)\sum_{t=1}^{T} \hat{V}_t(x)$, and $\omega_{\text{pooled}} = (NT-1)\hat{V}(x)$. This estimator also algebraically reduces to[8]

$$\hat{\beta}_{TW} = \frac{\displaystyle\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(y_{it}-\bar{y}_t)}{\displaystyle\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_t)}, \quad \text{or equivalently to} \quad \hat{\beta}_{TW} = \frac{\displaystyle\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_t)(y_{it}-\bar{y}_i)}{\displaystyle\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_t)}. \tag{12}$$

The numerator of this estimator mean-centers the outcome on one dimension to the outcome as it mean-centers the covariate on the other dimension.

Furthermore, although the two-way FE estimator removes case-fixed and time-fixed omitted variables, it does not isolate either the variation across cases or the variation across time in TSCS data. While the cross-sectional variance is removed from case FEs this variance is present for time FEs, so it must be present in the two-way FE model as well. Likewise, while the temporal variation is omitted from time FEs it exists in case FEs, so it is present in two-way FEs. If researchers have the goal of removing "problematic" variation from the dependent variable, as we discussed in our literature review earlier, then the two-way FE estimate paradoxically accomplishes the opposite of what these researchers intend.

4.1 How the Two-Way FE Estimator Compares to a Difference-in-Difference Design

As we discuss in section 2, the two-way FE model is often defended as a difference-in-difference (DiD) design. However, in order to be an estimator for a DiD effect, the two-way FE model must make assumptions that are more unrealistic than researchers who are interested in causal inference would typically be willing to accept when the data contain more than two time points, the treatment is not binary, or when the treatment is not zero for all cases in the first time point.

A DiD design approximates random assignment in a time period by subtracting from the outcome the value of the outcome at a prior time point (Morgan and Winship 2007, 253-254). The canonical application of a DiD estimator is to data that contain two time points, $t \in \{1, 2\}$, and a binary treatment variable X_{it} that is 0 for all cases in time point 1, 0 for the control group in time point 2, and 1 for the treatment group in time point 2. If we denote the outcomes for cases in the control group to be y_{1t} and the outcomes for the cases in the treatment group to be y_{2t}, then the DiD estimate is given by

$$\delta = E(y_{22}) - E(y_{21}) - E(y_{12}) + E(y_{11}). \tag{13}$$

[8] The proof of the theorem that demonstrates this estimator is in appendix A.2 in the supplemental material.

Consider the two-way FE regression

$$y_{it} = \alpha_{TW} + \beta_{TW} x_{it} + u_i + v_t + \varepsilon_{it}, \tag{14}$$

where u_i represents $\sum_{j=2}^{N} \gamma_j I_{i=j}$ and v_t represents $\sum_{w=2}^{T} \lambda_w I_{t=w}$ as defined in equation 9. Under the same conditions as the canonical DiD application, this regression yields a coefficient estimate that is equal to δ. Again, assume there are two time periods, a control group represented by $i = 1$ and a treatment group represented by $i = 2$, and a treatment given by

$$x_{it} = \begin{cases} 1 & \text{if } i = 2 \text{ and } t = 2, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Then the DiD statistic is

$$\begin{aligned} \delta &= E(y_{22}) - E(y_{21}) - E(y_{12}) + E(y_{11}) \\ &= (\alpha_{TW} + \beta_{TW} + u_2 + v_2) - (\alpha_{TW} + u_2 + v_1) - (\alpha_{TW} + u_1 + v_2) + (\alpha_{TW} + u_1 + v_1) \\ &= \beta_{TW} + (\alpha_{TW} - \alpha_{TW} - \alpha_{TW} + \alpha_{TW}) + (u_2 - u_2) + (u_1 - u_1) + (v_2 - v_2) + (v_1 - v_1) \\ &= \beta_{TW}. \end{aligned}$$

Therefore, under these ideal conditions, the two-way FE estimator is a DiD estimator.

The two-way FE estimator is often proposed as a DiD design even when there are multiple time points and when the treatment is continuous rather than binary. Our goal is to characterize how exactly a DiD estimate must be defined in this general context in order for β_{TW} to remain equal to δ. Without loss of generality, let there be any two distinct time points s and t that are not necessarily just one unit of time apart, and let there be two distinct cases i and j. If $(x_{it} - x_{js}) - (x_{it} - x_{js}) = d$, then the generalized DiD estimate is

$$\begin{aligned} \delta &= E(y_{it}) - E(y_{is}) - E(y_{jt}) + E(y_{js}) \\ &= \beta_{TW}(x_{it} - x_{is} - x_{jt} + x_{js}) + (\alpha_{TW} - \alpha_{TW} - \alpha_{TW} + \alpha_{TW}) + (u_i - u_i) + (u_j - u_j) + (v_t - v_t) + (v_s - v_s) \\ &= d\beta_{TW}. \end{aligned}$$

In particular, if $d = 1$ then $\beta_{TW} = \delta$.[9] Therefore the two-way FE coefficient has the following substantive interpretation:

> Consider two distinct cases i and j and two distinct time points s and t, and let $d = x_{it} - x_{is} - x_{jt} + x_{js}$. A one-unit increase in d is associated with a β_{TW} change in the DiD statistic $\delta = E(y_{it}) - E(y_{is}) -$

[9] Note that this property of δ also holds when the true model is one-way case FEs, one-way time FEs, or pooled OLS, as these models are special cases of the two-way FE model where one or both sets of FEs are zero.

$E(y_{jt)}) + E(y_{js})$, on average.

Let's consider the tacit assumptions one makes when using this coefficient to characterize an effect. In addition to the usual difference-in-difference assumption of parallel paths (see, for example, Baltagi 2011), this interpretation requires an assumption of homogeneity across both cases and time points. Homogeneity across cases is often non-controversial, especially if there is no reason to expect an interaction to exist in the cross-section; it is the assumption we make any time we allow an effect to generalize across cases. But this DiD effect also assumes homogeneity across time in a way that does not model temporal processes. The two time points s and t may exist at any point in the time series and they do not have to be adjacent. In other words, this effect generalizes across all time differences, regardless of whether they occur early in the time series, later in the time series, whether they are one year apart, or 100 years apart. Because the effect we estimate generalizes across cases and time points, we must rely heavily on the linearity of the model itself to impute the comparisons that we do not observe directly. Therefore an analyst must be convinced that the true causal estimate is both homogenous across time and linear in order for the two-way FE coefficient to be an accurate representation of the DiD statistic.

We argue that these assumptions are more restrictive than researchers who attempt to achieve causal identification would be willing to accept. Causal inference methodology is founded on the principle that before any estimation can occur, a researcher must clearly describe the causal effect of interest. In the case of a continuous treatment, the researcher must be clear about how different regions on the continuum impact the outcome differently, or else defend the linearity assumption. In the case of time dependent data, this specification must be clear about when an effect occurs and how long it takes to occur. Morgan and Winship (2007, 274) make this point explicit with regard to TSCS data:

> Defendable assumptions about the treatment assignment process must be specified. And, to use longitudinal data to its maximum potential, researchers must carefully consider the dynamic process that generates the outcome, clearly define the causal effect of interest, and then use constrained models only when there is reason to believe that they fit the underlying data.

Difference-in-difference estimation is a causal inference strategy, but the practice of employing a two-way FE model as a difference-in-difference design does not responsibly address the concerns regarding the model's assumptions. Other causal inference strategies, such as matching, dispel the assumption of linearity that the two-way FE estimator strongly makes. Approaches such as the synthetic control (Abadie and Gardeazabal 2003; Xu 2017) are explicit with regard to the comparisons they make, both cross-sectionally and over time. We therefore refer researchers to methods like these that are more careful about the assumptions they employ.

5 Illustrating How Two-Way FE Coefficients Combine Cross-Sectional and Temporal Effects

In section 4 we show that two-way FE coefficients combine cross-sectional and temporal effects in the data. In this section we consider each within-case slope — the slope of a best-fit line using only the data for one case — and each within-time slope — the slope of a best-fit line using only the data for one time point. Within cases all variance must occur over time, so within-case slopes speak to variance over time in the data. Likewise, within time points all variance occurs between cases, so within-time slopes speak to cross-sectional variance. Our goal is to illustrate what happens to two-way FE coefficients when the within-case and within-time slopes in the data change.

We emphasize that our purpose is not to demonstrate bias in the two-way FE coefficients. Indeed, because a two-way FE model can be estimated with OLS, coefficient estimates from this model are BLUE, and we can easily demonstrate this fact by generating data from equation 9 and running the two-way FE model. We also do not claim that the within-case or within-time slopes are necessarily the quantities that applied researchers need to report as primary results. However, the advantage of the within-case and within-time slopes is that they have clear interpretations as temporal and cross-sectional effects. Given the conceptual difficulty of interpreting two-way FE coefficients, understanding the relationship between two-way FE coefficients and the within-case and within-time slopes can help applied researchers understand the substantive meaning of two-way FE coefficients.

To examine the relationship between two-way FE coefficients and within-case and within-time slopes, we need a method for generating TSCS data in which we can set both the within-case and within-time slopes to prior, known values. The technique we employ to generate data with this property is novel and is discussed first in section 5.1.[10]

5.1 Simulating TSCS Data With Known Within-Case and Within-Time Slopes

To express the within-time slopes, we use a simple bivariate regression model that implies the following expected value for y_{it}:

$$E(y_{it}) = \alpha_t + \beta_t x_{it}. \tag{16}$$

In this model each cross-section has its own intercept α_t and its own coefficient, or within-time slope, on x_{it}, which is β_t. If the coefficients on x_{it} within each cross-section are all the same, then $\beta_t = \beta, \forall t$, which corresponds to a standard one-way FE regression with fixed effects on time points.

To express the within-case slopes, we use another simple bivariate regression model that implies the following expected value for y_{it}:

$$E(y_{it}) = \alpha_i + \gamma_i x_{it}. \tag{17}$$

As with equation 16, the regressions in each time series may have unique intercepts, α_i, and within-case slopes, γ_i,

[10] All of the code to run each of the simulations in R is listed in appendix C in the supplemental material.

and if $\gamma_i = \gamma, \forall i$, then we have a standard one-way FE regression with fixed effects on cases. Since equations 16 and 17 describe the same data while focusing on different dimensions within those data, both equations can be true for a given TSCS dataset. We need then a simulation of this data that corresponds to the situation that many applied researchers find themselves in, in which they can fit either model to a given TSCS dataset and must choose one of them.

To accomplish this task, we generate data from both equations 16 and 17 simultaneously. Accordingly, we set up a system of simultaneous equations

$$\begin{cases} E(y_{it}) = \alpha_t + \beta_t x_{it}, \\ E(y_{it}) = \alpha_i + \gamma_i x_{it}, \end{cases} \tag{18}$$

and solve it for $E(y_{it})$ and x_{it}:[11]

$$x_{it} = \frac{\alpha_i - \alpha_t}{\beta_t - \gamma_i} \tag{19}$$

and

$$E(y_{it}) = \frac{\beta_t \alpha_i - \gamma_i \alpha_t}{\beta_t - \gamma_i}. \tag{20}$$

To include a stochastic component in the dependent variable, we add an exogenous error term to $E(y_{it})$ to generate the outcome,

$$y_{it} = \frac{\beta_t \alpha_i - \gamma_i \alpha_t}{\beta_t - \gamma_i} + \varepsilon_{it}. \tag{21}$$

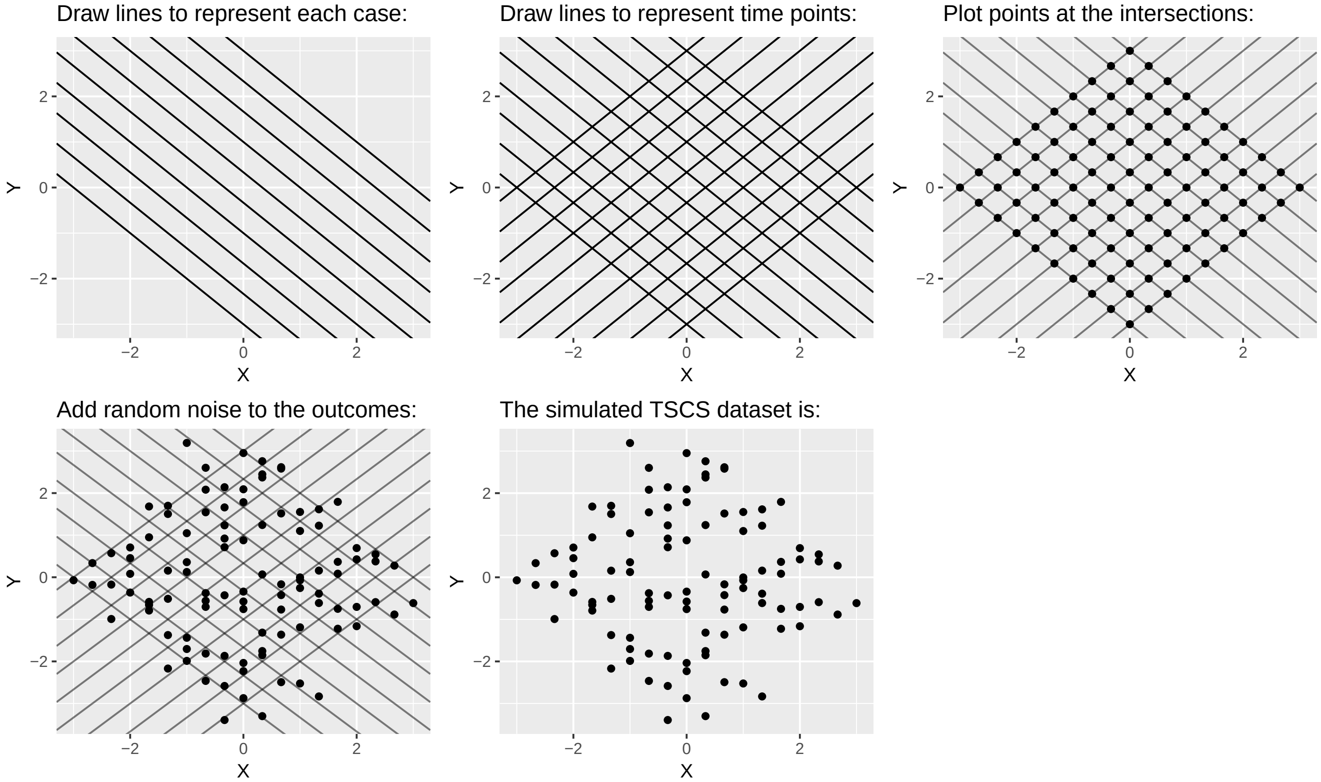
In the following simulations, we generate data that corresponds both to equations 16 and 17 by generating x_{it} from equation 19 and y_{it} from equation 21.

While our simulation is certainly unusual, we maintain that it is the simplest method for simulating TSCS data in which we can set the within-case and within-time slopes and fit different 1-way and 2-way FE models to the same data. To more clearly demonstrate the logic of our approach, we next provide an example of our simulation in which we simulate 10 cases and 10 time points for 100 total observations. We set the case-specific intercepts and the time-specific intercepts to values from -3 to 3 in increments of 0.6. We set the within-case slopes to each be -1, and the within-time slopes to each be 1. This example is illustrated in Figure 4.

First, in the upper-left-hand panel of Figure 4, we draw lines to represent the best-fit lines within each case. The slope of each line is equivalent to the coefficient from running a bivariate regression within each time point or case. In this specific example, the lines are parallel because we set each slope to -1. Next, in the upper-middle-panel, we draw lines to represent the best-fit lines within each time point. Again, in this specific example, the lines are parallel because we set each slope to 1. We plot points exactly at the intersections in this lattice in the upper-right-hand panel because TSCS data has the specific restriction that every observation exists in exactly one case and in exactly one time point. The only way to generate data with this property is to draw points at the intersections.

[11] See appendix A.3 in the supplemental material for the derivation of this solution.

Figure 4: Example of the Generation of One TSCS Dataset



Finding these intersections is equivalent to solving the system of equations in 18. Finally, in the bottom-left-hand panel, we add exogenous noise to the outcome of each datapoint — because the errors are uncorrelated with X, no linear model run on these data involves an endogeneity bias. Then, as shown in the bottom-middle-panel, we have constructed simulated a TSCS dataset.

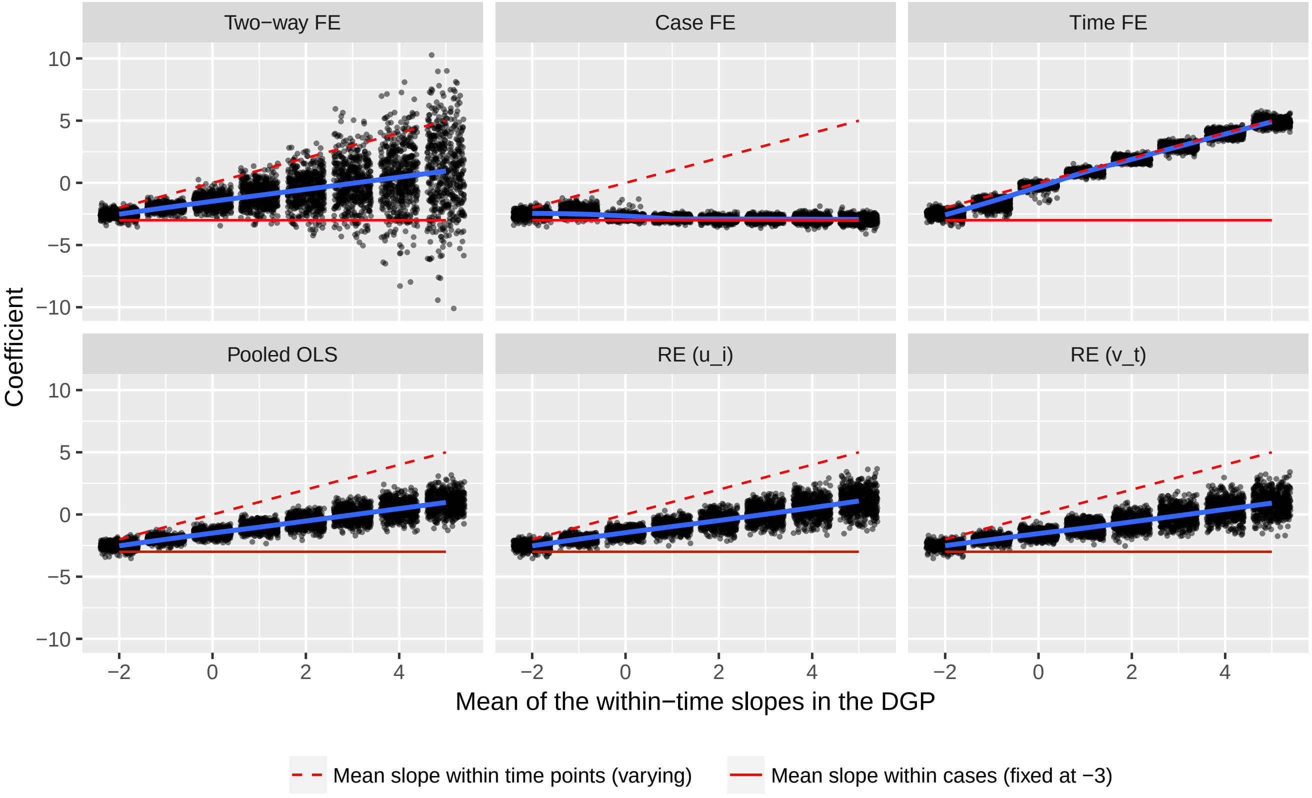
Using this framework we can set the intercepts and slopes to any value so long as the within-case and within-time slopes are not equal (otherwise two lines would have infinitely many intersection points). In addition, the lines for each case and for each time point need not be parallel. In section 5.2, we allow the within-case and within-time slopes to vary; we generate each set of slopes from a normal distribution with a variance of 0.25. In section 5.3, we demonstrate that the two-way FE estimator is unidentified when both sets of slopes are parallel: that is, when the within-case slopes do not vary across cases and the within-time slopes do not vary over time. In short, this simulation method's significant advantage is that it allows us sufficient command over the both the within-case and within-time slopes in TSCS data.

5.2 Two-way FE Coefficients Average the Within-case and Within-time Slopes in the Data

We repeatedly generate new TSCS data from equations 19 and 21. In each iteration, we set the number of cases N and the number of time points T each to be 30. We generate the case-specific intercepts α_i, the time-specific intercepts α_t, and the exogenous errors ε_{it} from standard normal distributions. We generate varying within-case slopes γ_i in equation 17 from a normal distribution with a mean of -3 and a variance of 0.25. We also draw the within-time slopes, β_t in equation 16, from normal distributions with variances of 0.25.

As the experimental treatment, we iteratively set the mean of the within-time slopes to be -2, -1, 0, 1, 2, 3, 4, and 5, and we repeat each simulation 500 times for each of these conditions. We plot these mean values of β_t on the x-axis in Figure 5, and we plot the coefficient estimates from each model under consideration on the y-axis. We compare the coefficients returned by two-way, case, and time FEs, pooled OLS, and random effects (RE). We run two versions of RE: one that integrates over a case-fixed intercept u_i, and one that integrates over a time-fixed intercept v_t.[12] The results for each model are aligned in a 3 × 2 grid in Figure 5.

Figure 5: Simulation results, varying the mean of the within-time slopes.



Note: for clarity, a small amount of random noise is added to the x-coordinate of each point.

As expected, the case FE coefficient is always about -3, equal to the average within-case slope, despite the fact

[12] We implement random effects using the `plm` package in R.

that the mean of the within-time slopes changes. The time FE coefficients fall along the 45-degree line, indicating that this estimator returns the average within-time slope on average. In contrast, the two-way FE, pooled OLS, and random effects coefficients tend to be estimated in the intermediate space between the mean within-case and mean within-time slopes.[13]

Figure 5 shows that case FEs are successful in removing the cross-sectional variation so that results clearly describe relationships between variables over time because changes in the within-time slopes do not affect the case FE coefficients. In addition, because time FEs eliminate the temporal variation and model the cross-sectional variation, the time FE coefficients accurately estimate this changing within-time slope regardless of its mean value. In contrast, two-way FE appears to be a pooling estimator like pooled OLS or RE, but with considerably less efficiency.

This result should lead us to reconsider the idea popular in applied work that two-way FEs account for both cross-sectional and temporal variation in the same way that one-way FE models do. As we have shown, two-way FE models are fundamentally different than their one-way cousins despite similarities in the estimating equations. For example, although two-way FEs include a dummy variable for every case, the two-way FE coefficients change along with the within-time slope, as do the coefficients from pooled OLS and both random effects models. Thus, by including time dummies in addition to case dummies, two-way FEs differ substantially from one-way case FEs because this model is once again dependent on both the cross-sectional and temporal variation.[14]

5.3 (Un-)Identifiability of Two-Way FE Estimates

When we use a one-way FE model, we estimate a set of lines — one line for each case or for each time point — with the same slope but with different intercepts, so that these lines are parallel. This means that, unless an interaction is used or the coefficient is explicitly modeled as random, the case FE model assumes that the within-case slopes are fixed across cases and the time FE model assumes that the within-time slopes are fixed across time points.

Whether or not these two assumptions hold, the case FE model estimates a weighted average of the within-case slopes and the time FE model returns a weighted average of the within-time slopes. In contrast, the two-way FE model is unidentified when the within-case slopes are fixed across cases at the same time as the within-time slopes are fixed across time points. The proof that the two-way FE coefficient is unidentified in this case is as follows:

Proof. If the within-time slopes are fixed across time points, then $\beta_t = \beta, \forall t$ in equation 16, and if the within-case

[13] The pooled OLS estimator is more efficient in this simulation, but we would note that this does not necessarily demonstrate that the pooled OLS and and random effects estimators are superior to the two-way FE estimator in general as this simulation does not cover the full range of possible kinds of TSCS data, including unbalanced panels that are known to affect pooled OLS and RE.. Our intention rather is to show how the two-way FE model is substantively similar to pooled OLS and random effects in that it pools variation across both dimensions.

[14] Appendix B in the supplemental material contains four additional simulations. Section B.1 demonstrates that the two-way FE estimator is neither dependent on sample size nor the amount of temporal autocorrelation in the errors. Section B.2 demonstrates that panel unbalance has a bigger impact on two-way FE estimates than on one-way FE estimates. Section B.3 shows that two-way FE coefficients resemble case FE coefficients when the variance of the within-case slopes is much less than the variance of the within-time slopes, and likewise, two-way FE coefficients resemble time FE coefficients when the variance of the within-time slopes is much less than the variance of the within-case slopes.

slopes are fixed across cases, then $\gamma_i = \gamma, \forall i$ in equation 17. Then the system of equations in 18 becomes

$$\begin{cases} E(y_{it}) = \alpha_t + \beta x_{it}, \\ E(y_{it}) = \alpha_i + \gamma x_{it}, \end{cases} \tag{22}$$

and the solution to this system is

$$x_{it} = \frac{\alpha_i - \alpha_t}{\beta - \gamma}, \qquad E(y_{it}) = \frac{\beta\alpha_i - \gamma\alpha_t}{\beta - \gamma}. \tag{23}$$

Consider the version of the two-way FE estimator in balanced panels that is listed in equation 8:

$$\hat{\beta}_{TW} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \bar{y}_i - \bar{y}_t + \bar{y})(x_{it} - \bar{x}_i - \bar{x}_t + \bar{x})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i - \bar{x}_t + \bar{x})^2}. \tag{24}$$

Next, consider just the denominator:

$$\begin{aligned} &\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i - \bar{x}_t + \bar{x})^2 \\ =&\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \frac{\sum_{t=1}^{T} x_{it}}{T} - \frac{\sum_{i=1}^{N} x_{it}}{N} + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}}{NT}\right)^2. \end{aligned} \tag{25}$$

We substitute x_{it} with the solution for x_{it} in equation 23,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\frac{\alpha_i-\alpha_t}{\beta-\gamma}-\frac{\sum_{t=1}^{T}\frac{\alpha_i-\alpha_t}{\beta-\gamma}}{T}-\frac{\sum_{i=1}^{N}\frac{\alpha_i-\alpha_t}{\beta-\gamma}}{N}+\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\alpha_i-\alpha_t}{\beta-\gamma}}{NT}\right)^2\\
=&\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\frac{\alpha_i-\alpha_t}{\beta-\gamma}-\frac{\sum_{t=1}^{T}\alpha_i-\alpha_t}{T(\beta-\gamma)}-\frac{\sum_{i=1}^{N}\alpha_i-\alpha_t}{N(\beta-\gamma)}+\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_i-\alpha_t}{NT(\beta-\gamma)}\right)^2\\
=&\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(NT(\alpha_i-\alpha_t)-N\sum_{t=1}^{T}(\alpha_i-\alpha_t)-T\sum_{i=1}^{N}(\alpha_i-\alpha_t)+\sum_{i=1}^{N}\sum_{t=1}^{T}(\alpha_i-\alpha_t)\right)^2}{N^2T^2(\beta-\gamma)^2}\\
=&\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(NT\alpha_i-NT\alpha_t-N\sum_{t=1}^{T}\alpha_i+N\sum_{t=1}^{T}\alpha_t-T\sum_{i=1}^{N}\alpha_i+T\sum_{i=1}^{N}\alpha_t+\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_i-\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_t\right)^2}{N^2T^2(\beta-\gamma)^2}\\
=&\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(NT\alpha_i-NT\alpha_t-NT\alpha_i+NT\bar{\alpha}_t-NT\bar{\alpha}_i+NT\alpha_t+NT\bar{\alpha}_i-NT\bar{\alpha}_t\right)^2}{N^2T^2(\beta-\gamma)^2}\\
=&\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left([NT\alpha_i-NT\alpha_i]+[NT\alpha_t-NT\alpha_t]+[NT\bar{\alpha}_t-NT\bar{\alpha}_t]+[NT\bar{\alpha}_i-NT\bar{\alpha}_i]\right)^2}{N^2T^2(\beta-\gamma)^2}=0. \qquad (26)
\end{aligned}$$

Since the denominator of the two-way FE estimator must be 0 under these conditions, it follows that the two-way FE model is unidentified. ■

This un-identifiability will manifest itself as a non-full-rank matrix that will result in an error in statistical software packages when matrix inversion is attempted. This problem will occur even if substantial variation exists in both the over-time and cross-section dimensions of the dataset; in other words, this problem is not a degrees of freedom issue.

The reason that this un-identifiability has not been recognized before is because Stata and R can deal with the fact that this model is unidentified by automatically selecting one of the fixed effects to drop from the model, as can also occur with the more common issues of multicollinearity and missing data. As such, this particular problem has likely been ignored when it occurs as it is difficult to impossible to diagnose without running regressions within each time point or case. Generally speaking, if the dropped FE happens to be a dummy variable for a case, then the resulting coefficient on x resembles the time FE coefficient on x. If the dropped FE happens to be a dummy variable for a time point, the coefficient resembles the case FE coefficient on x. In both Stata and R, we have noticed that the FE to drop in the case of non-identification is determined in part by the order in which the FEs are entered into the formula to run the linear model, although this of course depends on which estimation command is used. It is important for researchers to pay close attention to whether or not any FEs are dropped in the final model results, as this issue may be due to model non-identification.

Furthermore, this un-identifiability can manifest itself in very unstable estimates of $\hat{\beta}_{TW}$ when the within-case slopes are nearly equal across cases and when the within-time slopes are nearly equal across time points. That is, when the variance of the within-case slopes across cases and the variance of the within-time slopes across time points are both close to 0, the variance of $\hat{\beta}_{TW}$ approaches infinity. To demonstrate this behavior, we use another simulation.

We use the procedure described in section 5.1 to generate TSCS datasets with 30 cases and 30 time points each, where case-specific and time-specific intercepts and exogenous error are drawn from standard normal distributions. For each dataset, we set the mean of the within-case slopes to be -3 and the mean of the within-time slopes to be 3. In this simulation, we change the standard deviation of the within-case and within-time slopes as an experimental treatment. We draw the standard deviations of the slopes from exponential distributions with rate parameters set at 25 so that the randomly generated standard deviations are clustered at or near zero. For each pair of drawn standard deviations, we generate 100 TSCS datasets, we run the two-way FE model on each dataset, and we record the 100 coefficient estimates. We report the standard deviation of these coefficients, conditional on the standard deviations of the within-case and within-time slopes, in Figure 6. We repeat the process 1000 times.

Figure 6 is a scatterplot in which the x-axis represents the standard deviation of the within-time slopes and the y-axis represents the standard deviation of the within-case slopes. The size of each dot on this graph represents the standard deviation of the two-way FE coefficients estimated across the 100 TSCS datasets with the values of the standard deviations for the within-case and within-time slopes along the axes. As can be seen, as the standard deviation of the within-case and within-time slopes approaches zero, the standard deviation of the coefficients $\hat{\beta}_{TW}$ from the two-way FE model converges to infinity. With truly fixed slopes, the model matrix is singular and it is not possible to return a coefficient.

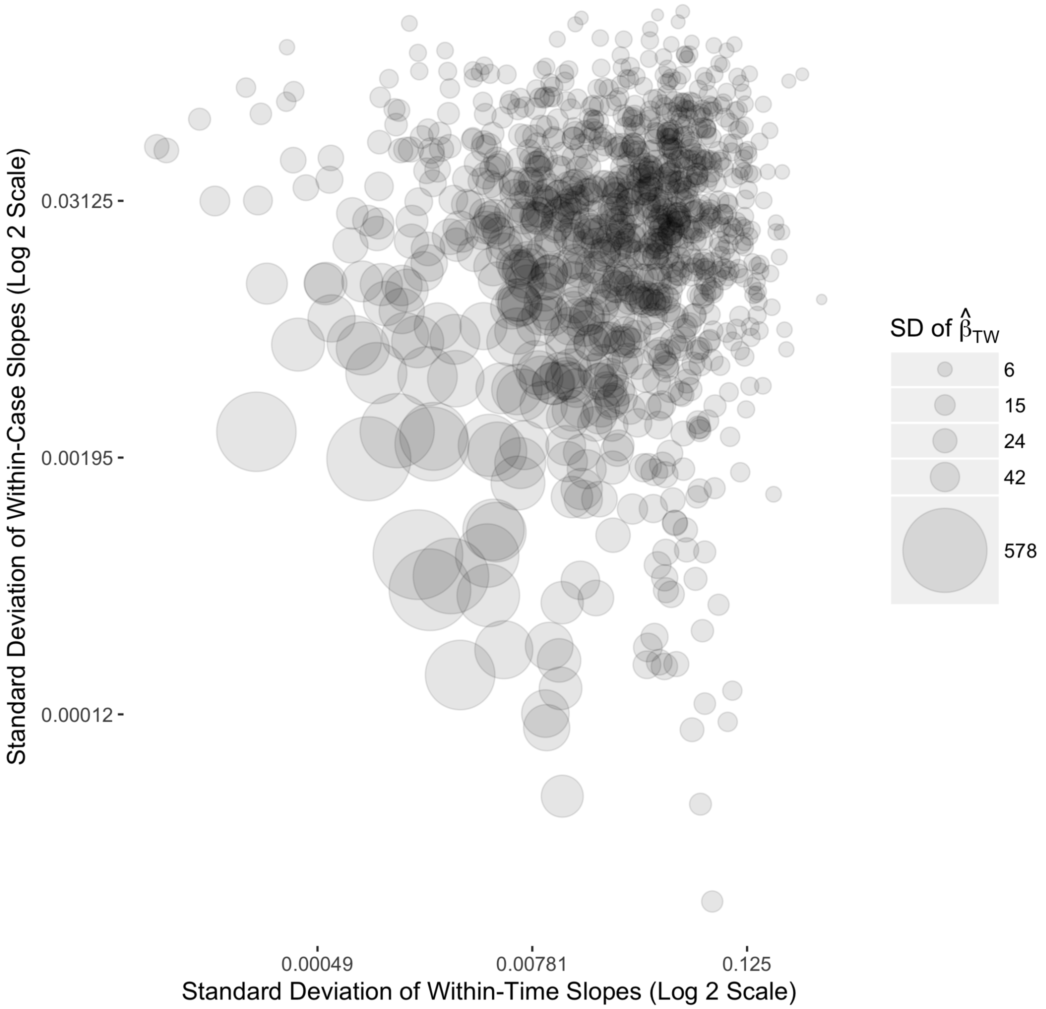
What this simulation reveals is that the variance of two-way FE coefficients will increase dramatically as slopes are nearly fixed across cases and time points, and this could well result in very unstable estimates for a particular dataset. For these reasons, in addition to the issues we have brought up in this paper, we would urge applied researchers to be very careful when employing this model for any TSCS dataset without strong prior knowledge about effect heterogeneity in the cross-section or over-time dimensions of variance.

6 Discussion: Taking Interpretation Seriously

Fixed effects models address omitted variable bias by accounting for time-fixed covariates, case-fixed covariates, or both. But they also change the research question being evaluated by the model. It is important for researchers to be aware of both of these implications of a FE specification, and to pay attention to how coefficients from the model should be interpreted.

One-way FE specifications remove omitted variables because of the fact that these models isolate one dimension

Figure 6: Two-way FE Estimates with Almost-fixed Within-case and Within-time Slopes



Note: the value of $\hat{\beta}_{TW}$ is equal to the coefficient on x obtained from a two-way FE linear regression of the simulated data. The standard deviation of the $\hat{\beta}_{TW}$ estimates is calculated from 100 random replicates for the given values of the standard deviation of the within-case and within-time slopes along the x and y axis.

of variance in the data: the case FE model only analyzes temporal variation and the time FE model only analyzes cross-sectional variation. A time-fixed omitted variable drops out of the case FE model, for example, because this variable cannot possibly explain how an outcome changes over time. The two-way FE model appears, at first, to be an elegant option for achieving greater degrees of causal identification with observational data because it removes omitted variables that are fixed across cases and omitted variables that are fixed over time. However, the two-way FE model does not isolate either the cross-sectional or temporal variance in the data, but rather averages across the two dimensions. As a result, while it is possible to interpret coefficient point estimates from two-way FEs, these interpretations are usually difficult to conceptualize and to communicate and may not correspond to the question the researcher intended to address.

We therefore issue a recommendation to applied researchers. We suggest that research questions be phrased in a way that makes the principal comparison being made explicitly clear. We then urge researchers to choose a model whose interpretation matches the intended research question. In particular, we encourage researchers to phrase research questions in a way that is more specific than "what is the effect of x on y?" In TSCS data, any answer to this general question pools across a comparison of cases and a comparison of time points, and we have to assume that the way that cases compare to one another — whether they are countries, U.S. states, political parties, institutions, elites, or individual survey respondents — is equal to the way that they compare to themselves over time. Such an assumption does not respect the nuanced theories that social scientists develop about these cases or about how they change over time.

The interpretation of a time FE model corresponds to a research question that involves cross-sectional comparisons, and the interpretation of a case FE model corresponds to a research question that involves comparisons over time. We emphasize, however, that this heuristic does not solve the estimation problems that can be manifest in TSCS data. The time FE estimator places the analysis in a cross-sectional context, and all of the problems of cross-sectional work may be present. The case FE estimator places the analysis in a time series context, and statistical issues with time series data must be addressed. Given that some of these methods tend to add significant complexity to models, we believe it is best to start with a solid basis of what the effect of X is in a well-defined estimation so that the changes produced by any model extensions can be understood with reference to a simpler base model. Once we start with the goal of estimating an effect in the cross-sectional or time dimension rather than forcing the two to be averaged together to produce one estimate, then it is easier to see what problems still exist that might obscure the relationship under study.

This approach also provides a useful framework for elaborating upon the one-way FE models to describe more complex and interactive comparisons. For example, interactive fixed effects models have become a useful way of exploring conditional relationships in panel data (Xu 2017), and our analysis helps elucidate these more sophisticated approaches as well. Furthermore, it encourages the development of new TSCS methods since a method may have a useful application without having to be the best, catch-all approach for estimation on both dimensions and any combination of the dimensions.

7 Conclusion

The two-way fixed effects model, an increasingly popular method for modeling TSCS data, is substantively difficult to interpret because the model's estimates are a complex amalgamation of variation in the over-time and cross-sectional effects. We examine the mathematical transformations that lead to different FE specifications, and show that one-way FE models can be understood as generalizations of the effects that exist within one case or within one time point. In contrast, the two-way FE model can be understood as a generalization of the effect of deviations from the case-means at a particular point in time, or equivalently, as a generalization of the effect of deviations from the time-means for each particular case. We suggest that this interpretation of two-way FE coefficients is usually difficult to conceptualize and to communicate, and seldom matches the questions researchers intend to answer.

In addition, we demonstrate through evidence from a simulation that the two-way FE model makes very specific assumptions about TSCS datasets, and if these assumptions are not met, the model can be unidentified even if substantial variation exists along both dimensions.

Because of the restrictive assumptions and difficulty in substantive interpretation, we do not recommend that applied researchers rely on the two-way FE model except for situations in which the assumptions are well-understood, such as the canonical difference-in-difference design.

We hope that an increased emphasis on interpretation leads methodologists and substantive researchers to think about TSCS data in a new way. Instead of beginning with the standard linear model and applying a myriad of corrections to account for the features of TSCS data, we suggest an approach that builds up to a complete and meaningful model from simple constituent parts. A researcher must first define the essential comparison in the data: the difference between two cases at a particular point in time, the difference between two points in time for a particular case, or a more complex and interactive comparison if the question calls for one. The researcher must then choose how to pool across all of these comparisons to generalize a finding and employ the power in the data. Future work to develop TSCS methods will be most useful to applied researchers if the method is clear about what it compares and how it generalizes across comparisons.

References

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

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Region." *The American Economic Review* 93 (1): 113–132.

Acemoğlu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared. 2008. "Income and Democracy." *American Economic Review* 98 (3): 808–842.

Acemoğlu, Daron, and James Robinson. 2006. *Economic Origins of Dictatorship and Democracy.* New York: Cambridge University Press.

Andersen, Jorgen J., and Michael L. Ross. 2014. "The Big Oil Change: A Closer Look at the Haber-Menaldo Analys." *Comparative Political Studies* 47 (7): 993–1021.

Anzia, Sarah F., and Christopher R. Berry. 2011. "The Jackie (And Jill) Robinson Effect: Why Do Congresswomen Outperform Congressmen?" *American Journal of Political Science* 55 (3): 478–493.

Baltagi, Badi H. 2011. *Econometrics.* Fifth. New York: Springer.

Bechtel, Michael M., and Jens Hainmueller. 2011. “How Lasting is Voter Gratitude? An Analysis of the Short- and Long-Term Electoral Returns to Beneficial Policy.” *American Journal of Political Science* 55 (4): 852–868.

Beck, Nathaniel, and Jonathan N. Katz. 1995. “What To Do (and Not to Do) with Time-Series Cross-Section Data.” *American Political Science Review* 89 (3): 634–647.

Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker. 1998. “Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable.” *American Journal of Political Science* 42 (4): 1260–1288.

Blaydes, Lisa, and Mark Andreas Kayser. 2011. “Counting Calories: Democracy and Distribution in the Developing World.” *International Studies Quarterly* 55 (4): 887–908.

Boix, Carles. 2003. *Democracy and Redistribution.* New York: Cambridge University Press.

———. 2011. “Democracy, Development, and the International System.” *American Political Science Review* 105 (4): 809–828.

Boyd, Christina L., Lee Epstein, and Andrew D. Martin. 2010. “Untangling the Causal Effects of Sex on Judging.” *American Journal of Political Science* 54 (2): 389–411.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* New York: Cambridge University Press.

Condra, Luke N., and Jacob N. Shapiro. 2012. “Who Takes the Blame? The Strategic Effects of Collateral Damage.” *American Journal of Political Science* 56 (1): 167–187.

Coppedge, Michael, John Gerring, Staffan I. Linberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, et al. 2017. “V-Dem [Country-Year/Country-Date] Dataset v7.1.” Varieties of Democracy (V-DEM) Project.

Donno, Daniela. 2013. “Elections and Democratization in Authoritarian Regimes.” *American Journal of Political Science* 57 (3): 703–716.

Gabel, Matthew J., Clifford J. Carruba, Caitlin Ansley, and Donald M. Beaudette. 2012. “Of Courts and Commerce.” *The Journal of Politics* 74 (4): 1125–1137.

Greene, William H. 2012. *Econometric Analysis.* Seventh. Upper Saddle River, NJ: Prentice Hall.

Haber, Stephen, and Victor Menaldo. 2011. “Do Natural Resources Fuel Authoritarianism? A Reappraisal of the Resource Curse.” *American Political Science Review* 105 (1): 1–26.

Hall, Matthew E. K. 2014. “The Semiconstrained Court: Public Opinion, the Separation of Powers, and the U.S. Supreme Court’s Fear of Nonimplementation.” *American Journal of Political Science* 58 (2): 352–366.

Houle, Christian. 2009. “Inequality and Democracy.” *World Politics* 61 (4): 589–622.

Imai, Kosuke, and In Song Kim. 2016. "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" Online. `http://imai.princeton.edu/research/files/FEmatch.pdf`.

Inglehart, Ronald, and Christian Welzel. 2005. *Modernization, Cultural Change and Democracy: The Human Development Sequence.* New York: Cambridge University Press.

Jenkins, Jeffrey A., and Nathan W. Monroe. 2012. "Buying Negative Agenda Control in the U.S. House." *American Journal of Political Science* 56 (4): 897–912.

Kennedy, Ryan. 2010. "The Contradiction of Modernization: A Conditional Model of Endogenous Democratization." *Journal of Politics* 72 (3): 785–798.

Khandker, Shahidur R., Gayatri B. Koolwal, and Hussain A. Samad. 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices.* Washington D.C.: World Bank.

Lazarus, Jeffrey. 2009. "Party, Electoral Vulnerability, and Earmarks in the U.S. House of Representatives." *The Journal of Politics* 71 (3): 1050–1061.

Limongi, Fernando, and Adam Przeworski. 1997. "Modernization: Theories and Facts." *World Politics* 49 (2): 155–183.

McGhee, Eric, Seth Masket, Boris Shor, Steven Rogers, and Nolan McCarty. 2014. "A Primary Cause of Partisanship? Nomination Systems and Legislator Ideology." *American Journal of Political Science* 58 (2).

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* New York: Cambridge University Press.

Plümper, Thomas, and Vera E. Troeger. 2007. "Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects." *Political Analysis* 15 (2): 124–139.

Scheve, Kenneth, and David Stasavage. 2012. "Democracy, War and Wealth: Lessons from Two Centuries of Inheritance Taxation." *American Political Science Review* 106 (1): 81–102.

Stimson, James A. 1985. "Regression in Space and Time: A Statistical Essay." *American Journal of Political Science* 29 (4): 914–947.

Truex, Rory. 2014. "The Returns to Office in a 'Rubber Stamp' Parliament." *American Political Science Review* 108 (2): 235–251.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25 (1): 57–76.