A Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure

Liyang Sun and Jesse M. Shapiro*

February 2022

Abstract

Linear panel models featuring unit and time fixed effects appear in many areas of empirical economics. An active literature studies the interpretation of the ordinary least squares estimator of the model, commonly called the two-way fixed effects (TWFE) estimator, in the presence of unmodeled coefficient heterogeneity. We illustrate some implications for the case where the research design takes advantage of variation across units (say, US states) in exposure to some treatment (say, a policy change). In this case, the TWFE can fail to estimate the average (or even a weighted average) of the units' coefficients. Under some conditions, there exists *no* estimator that is guaranteed to estimate even a weighted average. Building on the literature, we note that when there is a unit totally unaffected by treatment, it is possible to estimate an average effect by replacing the TWFE with an average of difference-in-differences estimators.

Economists often seek to evaluate the effects of a certain event, such as the adoption of a policy or the arrival of an innovation, on some outcome of interest. For

*Liyang Sun is Postdoctoral Research Fellow, University of California Berkeley, Berkeley, California and Assistant Professor, CEMFI (Center for Monetary and Financial Studies), Madrid, Spain. Jesse M. Shapiro is George Gund Professor of Economics and Business Administration, Harvard University, Cambridge, Massachusetts and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are lsun20@berkeley.edu and jesse_shapiro@fas.harvard.edu.

example, how did the enactment of Medicare affect total expenditures on health care? How did the historical arrival of the potato affect population growth across the Old World? Simply comparing outcomes before and after the occurrence of the event risks conflating the effect of the event with the effect of numerous other coincident changes: think of all the other things that changed around the start of Medicare (1965) or the beginning of the Columbian exchange (1492). One way to measure the effect of these coincident changes is by looking at the outcomes of a control group totally unaffected by the event. But in some cases it is difficult to find such a pure control—Medicare was a national policy, and the arrival of the potato likely touched every part of the Old World in some way.

In such settings, it is common to take advantage of variation across geographic or other units in the extent of their *exposure* to the event. Even though all US states were affected by the introduction of Medicare, some were more affected than others, for example because they had relatively less well insured elderly populations prior to Medicare. Likewise, some regions of the Old World were relatively better suited to potato cultivation, making them better able to take advantage of the new crop's arrival.

One model of such a situation holds that the outcome is composed of a unit effect, a time effect, an interaction between a measure of the event and a measure of the unit's exposure, and an error term unrelated to the others. We can write a heuristic model like this:

$$\text{Outcome} = \text{Unit effect} + \text{Time effect} + \text{Coefficient}\,(\text{Event} \times \text{Exposure}) + \text{Error}. \qquad \text{(heuristic model)}$$

In this linear panel model, the unknown unit effect accounts for features of the unit (e.g., state or region) that are time-invariant, the unknown time effect accounts for background changes that may coincide with the event, and the unknown error term accounts for other unsystematic factors that influence different units at different times. The observed event variable varies over time and captures the event of interest. The observed exposure variable varies across units and captures units' different exposure to the event. The product of these two variables is the term of greatest interest, as it captures the fact that different units are affected differently by the event because of

their different exposure to it.

Linear panel models featuring an interaction between an event variable and an exposure variable, as in the heuristic model, appear in many areas of economics. For example, Finkelstein's (2007, equation 1) model of hospital expenses includes an interaction between time indicators (around the introduction of Medicare) and a measure of access to private insurance. Nunn and Qian's (2011, equation 3) model of Old World population growth includes an interaction between an indicator for periods following the introduction of the potato and the log of land area in a country that is suitable for growing potatoes. Dube and Vargas' (2013, equation 1) model of violence in Colombia includes an interaction between the world oil price and a measure of a region's baseline oil production intensity.[1]

Under suitable conditions on the error term, the unknown coefficient in the heuristic model can be estimated via ordinary least squares regression of the outcome on unit indicators, time indicators, and an interaction between the event variable and the exposure variable. Because the model involves two sets of fixed effects—one for units and one for time—this ordinary least squares estimator is sometimes called a two-way fixed effects (TWFE) estimator.

In this paper we consider the possibility that, in addition to the exposure variable, the effect of the policy or event itself—the coefficient in the heuristic model—differs by unit. Heterogeneous effects of this kind can arise for many reasons. For example, a given change in the fraction of elderly insured might affect expenditures more in states with a less healthy uninsured population. A given level of potato cultivation might affect population growth more in regions with better access to trade. Economists have been interested in heterogeneous effects of this kind for a long time (see, for example, surveys in Heckman and Vytlacil 2007; Imbens and Wooldridge 2009). Recently, an especially active literature has studied the effects of this form of coefficient heterogeneity on the performance and interpretation of the two-way fixed effects estimator. We draw heavily on this literature, and especially on work by de Chaisemartin and D'Haultfœuille (2018), who consider a setting similar to the one we consider here.

[1]Other examples include Zhang and Zhu's (2011, equations 2 and 3) model of social influences on contributions to Chinese Wikipedia, Dafny et al.'s (2012, equation 5) model of the effect of a merger on health insurance premiums, and Pierce and Schott's (2016, equation 2) model of the effect of trade with China on US manufacturing employment.

We will see that in general the two-way fixed effects estimator can perform very poorly when effects are heterogeneous, in the sense that it can fail to estimate the average (or even a weighted average) of the units' coefficients. This problem can be so severe that it affects any estimator, not just the TWFE estimator. And we will look at one situation—a setting with an unaffected unit—in which it is possible to estimate an average effect by replacing the TWFE with an average of exposure-adjusted difference-in-differences estimators.

A Motivating Example

To study the issues in more detail we now introduce a concrete example. We base the example loosely on Finkelstein's (2007) study of the effect of Medicare, setting aside much of the richness of Finkelstein's (2007) original analysis.

We are interested in learning the effect of Medicare on health care expenditures. Medicare is a US government program introduced in 1965 to provide health insurance to the elderly. We observe per capita health care expenditures y_{st} on the elderly for each US state s in each of two time periods t, where we can let $t = 0$ denote the period before the introduction of Medicare and $t = 1$ denote the period after.

Because many things change over time, simply comparing expenditures at time $t = 1$ to those at time $t = 0$ may not give a reliable estimate of the effect of Medicare. It would be helpful to have a control state that did not adopt Medicare, but since Medicare was a national policy, such a state does not exist.

Instead, we can take advantage of the fact that states differ in the fraction of the elderly that were insured prior to Medicare's introduction. In a New England state, where the penetration of private insurance among the elderly was relatively high prior to the introduction of Medicare (Finkelstein 2007, Table 1), Medicare had a relatively small effect on rates of insurance coverage. In a Pacific state, where the penetration of private insurance among the elderly was relatively low prior to Medicare (Finkelstein 2007, Table 1), Medicare had a relatively large effect on rates of insurance coverage.

Let x_{st} be the fraction of elderly with health insurance in a given state s at time t. At time $t = 0$, before Medicare, we can think of x_{s0} as measuring the fraction of elderly with private or other (non-Medicare) government insurance in state s. At

time $t = 1$, after Medicare, we can think of x_{s1} as being equal to 1 for all states s due to the universal coverage afforded by Medicare.

A linear panel data model of health care expenditures – what we will refer to as the linear model – might then take the form

$$y_{st} = \alpha_s + \delta_t + \beta x_{st} + \varepsilon_{st}. \qquad \text{(linear model)}$$

Here α_s is a state fixed effect that captures time-invariant state characteristics that may affect health care expenditures, δ_t is a time fixed effect that captures state-invariant time-dependent factors that may affect health care expenditures, and ε_{st} is an error term unrelated to x_{st}.[2] The parameter β measures the causal effect of insurance coverage on health care expenditure. Specifically, it measures the effect on per capita health expenditures of going from no coverage ($x_{st} = 0$) to full coverage ($x_{st} = 1$).[3]

We can rewrite the linear model in a form that resembles the heuristic model. In particular, because $x_{s1} = 1$ for all states s, it is straightforward to show that the linear model implies that

$$y_{st} = \tilde{\alpha}_s + \delta_t + \beta\,(1 - x_{s0})\,t + \varepsilon_{st}. \qquad \text{(exposure model)}$$

In the exposure model, the term $\tilde{\alpha}_s$ plays the role of the unknown unit effect from the heuristic model.[4] The term δ_t plays the role of the unknown time effect. The term ε_{st} plays the role of the unknown error term. The term $(1 - x_{s0})$ is the observed exposure variable and the term t, which is just an indicator for whether the observation is from the post-Medicare period, is the observed event variable.

Intuitively, under the exposure model, we can learn about the coefficient β by looking at whether, following the introduction of Medicare, health care expenditures diverge between states with different levels of private insurance before Medicare (dif-

[2]Specifically we assume that each of ε_{s0} and ε_{s1} has mean zero conditional on x_{s0} and x_{s1}.

[3]The effect of Medicare on expenditures in state s is given by $\beta\,(1 - x_{s0})$, that is, the effect of insurance coverage on expenditures, β, multiplied by the effect of Medicare on insurance coverage, $(1 - x_{s0})$.

[4]To go from the linear model to the exposure model, we have redefined the state fixed effect as $\tilde{\alpha}_s = \alpha_s + \beta x_{s0}$.

ferent values of x_{s0}). If so, then because different states are affected equally by the time effect represented by δ_t, it must be that Medicare is exerting a causal effect on expenditures.

More practically, we can estimate the unknown coefficient β by regressing health expenditures on state indicators, a time indicator, and an interaction between the fraction previously uninsured $(1 - x_{s0})$ and the post-Medicare indicator t. This is a two-way fixed effects (TWFE) estimator. Call it $\hat{\beta}$. The TWFE estimator $\hat{\beta}$ has some appealing properties. For example, if the exposure model holds, and ε_{st} is unrelated to x_{st}, then $\hat{\beta}$ is centered around β, in the sense that even though in any given sample $\hat{\beta}$ may be higher or lower than β, across samples $\hat{\beta}$ will tend to be equal to β on average.

The Possibility of Heterogeneous Coefficients

According to the linear model, a given change in the fraction insured has the same effect on per capita health expenditures in every state s. But it seems plausible that health expenditures will respond differently to changes in insurance in different states. For example, a state with a less healthy uninsured population may see expenditures rise more in response to a given expansion in insurance, compared to a state with a more healthy uninsured population, because relatively less healthy insurees require more expensive care.

We can formalize this possibility by imagining that each state s has its own coefficient β_s describing the effect of insurance on expenditures in the state, much as it has its own fixed effect α_s describing its baseline level of expenditures. Keeping all other elements of the linear model yields the following heterogeneous panel model:

$$y_{st} = \alpha_s + \delta_t + \beta_s x_{st} + \varepsilon_{st}. \qquad \text{(heterogeneous model)}$$

Even though we are allowing heterogeneity in the effect of treatment, we are still maintaining that the error term ε_{st} is unrelated to the fraction of elderly with health insurance x_{st} as before, so absent changes in the insurance levels x_{st}, all states would follow identical average trends over time.

Consider a researcher who believes that the effect of insurance may differ across states as in the heterogeneous model. How reasonable would it be for the researcher to estimate the effect of added health insurance using the convenient TWFE estimator that is based on the exposure model, which assumes that all states have the same coefficient β?

A single estimator $\hat{\beta}$, by construction, cannot be centered around each of the 50 different true coefficients for each state β_s. But maybe the single estimator $\hat{\beta}$ is centered around a good summary of the true coefficients, such as an average. If so, $\hat{\beta}$ might still be a convenient way to estimate the effect of insurance on expenditures in a "typical" state.

In certain situations, the estimator $\hat{\beta}$ will indeed be centered on an average of the true state-level coefficients β_s. One such situation is where β_s is unrelated to (i.e., statistically independent of) all the other terms in the heterogeneous model. In this case, results in the online appendix imply that $\hat{\beta}$ is centered around an average of the coefficients β_s, and therefore might still be considered an appealing estimator.

However, the situation where the coefficient β_s is unrelated to the other terms in the model is somewhat special. Suppose, for example, that β_s is greater in states with a less healthy uninsured population. Suppose, further, that the uninsured population is less healthy in states with greater insurance penetration prior to Medicare, say because in such states only the least healthy elderly remain uninsured. In this case, β_s will tend to be positively related to x_{s0}. Such a relationship between β_s and x_{s0} can cause the two-way fixed effects estimator $\hat{\beta}$ to behave rather badly.

To illustrate, consider a hypothetical numerical example of the heterogeneous model. In this example, we let the index s of the states run from 1 to 50. We let the fraction of elderly with insurance before Medicare be given by $x_{s0} = 0.245 + s/100$, so that the fraction runs from 0.255 ($s = 1$) through 0.745 ($s = 50$) in increments of 0.01, with an average value of 0.5.

In this numerical example, we also let the coefficient β_s vary across states according to the equation

$$\beta_s = 1 + 0.5\lambda - \lambda x_{s0}. \qquad \text{(numerical example)}$$

Here, λ is a parameter that governs how much the coefficient β_s varies across states, and how the state-level coefficient β_s is related to the fraction of elderly with insurance

before Medicare. When λ is 0, the coefficient β_s is equal to 1 in all states regardless of prior insurance penetration. When λ is less than 0, states with greater insurance penetration prior to Medicare have a larger coefficient β_s. When λ is greater than 0, states with greater insurance penetration prior to Medicare have a smaller coefficient β_s.

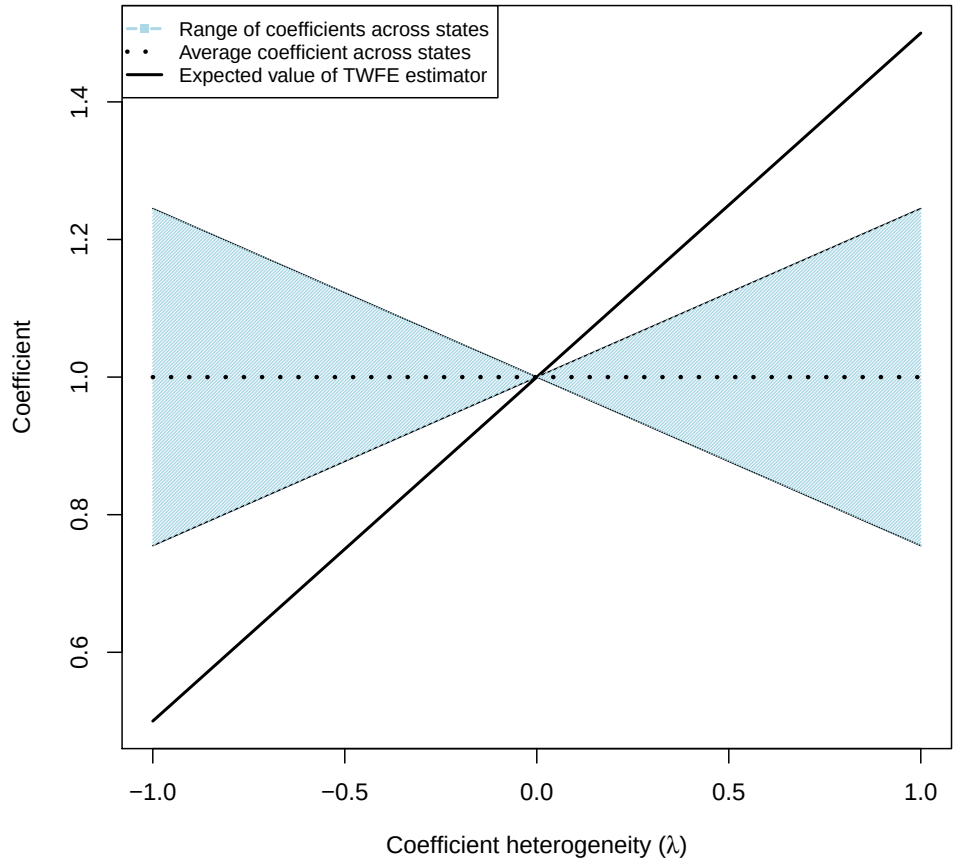
We have constructed the numerical example so that, no matter the value of λ, the average value of β_s across all states is always 1. By varying λ, we can therefore vary the relationship between β_s and x_{s0} while holding constant the average value of β_s.

Figure 1 illustrates the behavior of the two-way fixed effects estimator $\hat{\beta}$ in this numerical example. The horizontal axis shows the parameter λ, which controls the strength of the relationship between β_s and x_{s0}, and hence the degree of heterogeneity in the coefficient β_s. We consider values of λ ranging from -1 to $+1$. The shaded region shows the range of coefficients across the states s for each given value of λ. As λ departs from zero, this range widens, but remains centered around the average value of 1, which is illustrated with a dotted line. The solid line shows the value around which the TWFE estimator $\hat{\beta}$ is centered. Specifically, the line shows the average or expected value of $\hat{\beta}$ across repeated samples of the data. Except when $\lambda = 0$, this value, which is derived in the online appendix, does not coincide with the average value of β_s.

Perhaps more surprising, and more concerning, is that, when λ is not equal to zero, $\hat{\beta}$ is centered outside the shaded region that depicts the range of true coefficients β_s. When λ is less than zero, $\hat{\beta}$ is centered on a value smaller than any of the true coefficients β_s. When λ is greater than zero, $\hat{\beta}$ is centered on a value larger than any of the true coefficients β_s. A researcher using $\hat{\beta}$ to estimate an average or typical effect of insurance on health expenditure would, in these situations, end up with a very misleading estimate, one that is centered on a value outside the range of the true coefficients β_s.

To understand why the two-way fixed effects estimator behaves this way, consider the case of $\lambda > 0$ and recall that $\hat{\beta}$ is the ordinary least squares estimate of the coefficient β on the interaction term $(1 - x_{s0})\, t$ in the exposure model. This estimate will tend to be larger when, following Medicare's introduction, expenditure grows more in states that experience a larger increase in insurance coverage, $(1 - x_{s0})$. When

Figure 1: Expected Value of the Two-way Fixed Effects Estimator Under Coefficient Heterogeneity



Source: Illustrative calculations by the authors.
Notes: This figure illustrates the behavior of the two-way fixed effects (TWFE) estimator $\hat{\beta}$ of the parameter β in the exposure model for a hypothetical numerical example described in Section "The Possibility of Heterogeneous Coefficients." The horizontal axis corresponds to the parameter λ which governs how much and in what way the coefficient β_s in the heterogeneous model varies across states. For each given value of λ, the shaded region shows the range of coefficients $(\beta_1, \ldots, \beta_{50})$ across the 50 states, the dotted line shows the average value of β_s, and the solid line shows expected value of the two-way fixed effects estimator, $\hat{\beta}$.

$\lambda > 0$, states with a larger increase in insurance coverage, $(1 - x_{s0})$, also have larger coefficients β_s. Following Medicare's introduction, expenditure therefore grows more in states with larger $(1 - x_{s0})$ both because these states experience a larger increase in insurance coverage and because these states experience a larger change in expenditure for a given change in insurance coverage. The exposure model accounts for only the first of these effects, so the corresponding ordinary least squares estimator $\hat{\beta}$ conflates them, thus overstating the effect of insurance on expenditure. In the numerical example, this conflation is so severe that the expected value of the TWFE estimator falls outside the range of the true coefficients β_s.

The numerical example proves that the two-way fixed effects estimator cannot, in general, be guaranteed to be centered around a value inside the range of the true coefficients β_s in the heterogeneous model. In fact, we prove in the online appendix that there is *no* estimator that can be guaranteed, regardless of the coefficients β_s and the pre-Medicare insurance levels x_{s0}, to be centered around a value inside the range of the true coefficients β_s in the heterogeneous model. It follows that there is no estimator guaranteed to be centered around the average β_s across the states.

A Difference-in-Differences Perspective

Another way to build intuition about the impact of coefficient heterogeneity is to consider the behavior of some difference-in-differences type estimators. To relate to the classical difference-in-differences estimator, imagine that Medicare had been adopted in one treatment state, say state s, and not adopted in another control state, say state s'. Imagine further that no one had health insurance to begin with in either state, so that Medicare increased the fraction of the elderly with health insurance from 0 to 1 in the treatment state s, and left the fraction at 0 in the control state s'. In this case, by computing the difference in the change in the outcome y between the treatment and control states, $(y_{s1} - y_{s0}) - (y_{s'1} - y_{s'0})$, we would, on average, isolate the effect of Medicare, and arrive at a difference-in-differences estimator centered around the true effect β, much as in Card and Krueger's (1994) classic study of the effect of the minimum wage.

In this hypothetical situation, we have one treatment state that is strongly affected

by the introduction of Medicare, and another control state that is totally unaffected. In the more realistic situation where all states were affected by the introduction of Medicare, simply comparing the change in the outcome y between a more affected state s and a less affected state s' seems incomplete, because such a comparison does not account for the different changes in insurance rates x induced by Medicare in the two states. The following exposure-adjusted difference-in-differences estimator provides one possible way to account for changes in insurance rates:

$$\hat{\beta}^{DID}_{s,s'} = \frac{(y_{s1} - y_{s0}) - (y_{s'1} - y_{s'0})}{(1 - x_{s0}) - (1 - x_{s'0})}.$$

de Chaisemartin and D'Haultfœuille (2018) call $\hat{\beta}^{DID}_{s,s'}$ a Wald-difference-in-differences estimator because it consists of the ratio of the difference-in-differences estimator for the outcome (in our case, expenditures) to the one for exposure (insurance).

The estimator $\hat{\beta}^{DID}_{s,s'}$ is intuitive, but suffers from limitations similar to those of the TWFE estimator. In particular, $\hat{\beta}^{DID}_{s,s'}$ can be centered around a value that is larger or smaller than both β_s, the true coefficient for state s, and $\beta_{s'}$, the true coefficient for state s'. For a concrete example, if we take $s = 1$ and $s' = 50$ from the earlier numerical example, and say that $\lambda = 1$, then based on the formula we derived, the estimator $\hat{\beta}^{DID}_{s,s'}$ is centered around the value 1.5, which is greater than both $\beta_1 = 1.245$ and $\beta_{50} = 0.755$. One way to build an intuition for this behavior is to note that $\hat{\beta}^{DID}_{s,s'}$ is equivalent to the TWFE estimator $\hat{\beta}$ in the case where we have only two states in the sample, s and s'. Just like the TWFE estimator, $\hat{\beta}^{DID}_{s,s'}$ cannot be guaranteed to be centered around a value inside the range of β_s and $\beta_{s'}$.[5]

Suppose, though, that in state s' Medicare had no effect on insurance rates, for example because all elderly in the state were insured prior to Medicare, $x_{s'0} = 1$. That would take us closer to the classical difference-in-differences setting of Card and Krueger (1994) and others, and in that case, $\hat{\beta}^{DID}_{s,s'}$ is centered around β_s, the true coefficient for the affected state s. In fact, by taking an average of $\hat{\beta}^{DID}_{s,s'}$ across all of the affected states s, always treating state s' as the comparison, we arrive at an estimator that is centered around the average value of β_s across all affected states s.[6]

[5] In the online appendix, we establish the equivalence of $\hat{\beta}^{DID}_{s,s'}$ and $\hat{\beta}$ in the case of two states, and derive the expected value of $\hat{\beta}^{DID}_{s,s'}$.

[6] Because the effect of insurance β_s does not vary with time, the heterogeneous model satisfies

The presence of a totally unaffected state therefore makes it possible to construct an estimator centered around the true coefficient for any affected state, such as $\hat{\beta}^{DID}_{s,s'}$, and one centered around the average of true coefficients for all affected states, such as the average of $\hat{\beta}^{DID}_{s,s'}$. It is important to note, however, that the presence of a totally unaffected state does not repair the problems we highlighted earlier with the TWFE estimator $\hat{\beta}$. Calculations in the online appendix show that even if we add a totally unaffected state to the sample, the TWFE estimator remains centered outside of the range of treatment effects β_s in the numerical example. Thus, while the presence of a totally unaffected state means that it is possible to find estimators that are centered around the average coefficient, it does not guarantee that all estimators are centered around an average coefficient.

Some economic situations do not feature a totally unaffected unit that can serve as a comparison for affected units. In such situations, researchers may still be able to make progress by using economic assumptions to impose further structure on the coefficients β_s. For example, suppose that a researcher is willing to posit a linear relationship between β_s and x_{s0} of the form in the numerical example, but does not know the value of the parameter λ that governs this relationship. In this case, it is possible to substitute the expression for β_s into the heterogeneous model to arrive at a linear panel model whose unknown parameter, λ, can be estimated by a two-way fixed effects estimator, thus allowing the researcher to estimate averages of the coefficients β_s.

Suggestions for Further Reading

Recently there has been a surge in interest in the role of treatment effect heterogeneity in the sorts of settings we discuss here, where policies are introduced with different intensities, or at different times, to different units. This is a very active area and it is not our intention to survey it fully. However, we can point to some published or

the stable treatment effect assumption of de Chaisemartin and D'Haultfœuille (2018). Because the state s' is unaffected by Medicare, state s' satisfies the stable group assumption of de Chaisemartin and D'Haultfœuille (2018). Theorem 1 of de Chaisemartin and D'Haultfœuille (2018) implies that, under other standard conditions, the average of $\hat{\beta}^{DID}_{s,s'}$ is centered on the average coefficient among states affected by the policy change.

forthcoming articles that readers may find helpful.

de Chaisemartin and D'Haultfœuille (2018) consider a setting closely related to the one we discuss here. They consider the possibility that treatment effects vary by unit and over time, and formalize issues that can arise with exposure-adjusted difference-in-differences estimators. They propose two alternative estimators, one of which corrects the exposure-adjusted difference-in-differences estimator directly for diverging trends due to differential exposure. de Chaisemartin and D'Haultfœuille (2020) extend the analysis to a more general setting with multiple time periods, and again propose a time-corrected difference-in-differences estimator that can help avoid issues of the sort we illustrate above.[7] The Stata packages `fuzzydid` and `did_multipledgt` implement both alternative estimators. Related to de Chaisemartin and D'Haultfœuille (2020), Imai and Kim (2020) characterize the relationship between a two-way fixed effects estimator and the difference-in-differences estimator, and use this to illustrate some pitfalls of the two-way fixed effects estimator.

A related but distinct setting is one of staggered adoption, where different units (e.g., US states) adopt a policy (e.g., unilateral divorce) at different times. In this setting, when policy effects may differ over time or across units based on when they adopt the policy, the two-way fixed effects estimator experiences issues similar to those we illustrate above. Goodman-Bacon (2021) proposes diagnostics for the performance of a two-way fixed effects estimator in such situations. The Stata package `bacondecomp` implements these diagnostics. Sun and Abraham (2021) propose an estimator that avoids some of the drawbacks of the two-way fixed effects estimator by taking advantage of the presence of never-treated units in the sample. Callaway and Sant'Anna (2021) propose a similar estimator that uses not-yet-treated units as control, and can efficiently adjust for covariates using approaches developed in Sant'Anna and Zhao (2020). The Stata package `eventstudyinteract` and the R package `did` implement these two estimators respectively. Athey and Imbens (2022) consider the interpretation and variability of the difference-in-differences estimator in situations in which a unit's date of adoption is randomly assigned.

[7]Both articles by de Chaisemartin and D'Haultfœuille (2018, 2020) include applications to an earlier paper of Shapiro's (Gentzkow et al. 2011). So, Shapiro is here to take advice as well as give it.

We thank our dedicated research assistants for their contributions to this project. We are grateful to Ethan Lewis, Dan Levy, two anonymous reviewers, and the editors, Timothy Taylor and Heidi Williams, for helpful comments and suggestions. Liyang Sun gratefully acknowledges support from the Jerry A. Hausman Graduate Dissertation Fellowship and Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010. Jesse Shapiro gratefully acknowledges support from the Eastman Professorship, the Population Studies and Training Center, and the JP Morgan Chase Research Assistantship, all at Brown University, and from the National Science Foundation under Grant No. 1949047.

References

Athey, Susan and Guido Imbens, "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 2022, *226* (1), 62–79.

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230. Themed Issue: Treatment Effect 1.

Card, David and Alan B. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, September 1994, *84* (4), 772–793.

Dafny, Leemore, Mark Duggan, and Subramaniam Ramanarayanan, "Paying a Premium on Your Premium? Consolidation in the US Health Insurance Industry," *American Economic Review*, April 2012, *102* (2), 1161–1185.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Fuzzy Differences-in-Differences," *Review of Economic Studies*, April 2018, *85* (2), 999–1028.

_ **and** _ , "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, September 2020, *110* (9), 2964–2996.

Dube, Oeindrila and Juan F. Vargas, "Commodity Price Shocks and Civil Conflict: Evidence from Colombia," *Review of Economic Studies*, October 2013, *80* (4), 1384–1421.

Finkelstein, Amy, "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare," *Quarterly Journal of Economics*, February 2007, *122* (1), 1–37.

Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson, "The Effect of Newspaper Entry and Exit on Electoral Politics," *American Economic Review*, December 2011, *101* (7), 2980–3018.

Goodman-Bacon, Andrew, "Difference-in-differences with Variation in Treatment Timing," *Journal of Econometrics*, 2021, *225* (2), 254–277. Themed Issue: Treatment Effect 1.

Heckman, James J. and Edward J. Vytlacil, "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in James J. Heckman and Edward E. Leamer, eds., *Handbook of Econometrics*, Vol. 6, Elsevier, January 2007, pp. 4875–5143.

Imai, Kosuke and In Song Kim, "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data," *Political Analysis*, 2020, p. 1–11.

Imbens, Guido W. and Jeffrey M. Wooldridge, "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, March 2009, *47* (1), 5–86.

Nunn, Nathan and Nancy Qian, "The Potato's Contribution to Population and Urbanization: Evidence From A Historical Experiment," *Quarterly Journal of Economics*, May 2011, *126* (2), 593–650.

Pierce, Justin R. and Peter K. Schott, "The Surprisingly Swift Decline of US Manufacturing Employment," *American Economic Review*, July 2016, *106* (7), 1632–1662.

Sant'Anna, Pedro H. C. and Jun Zhao, "Doubly Robust Difference-in-differences Estimators," *Journal of Econometrics*, November 2020, *219* (1), 101–122.

Sun, Liyang and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021, *225* (2), 175–199. Themed Issue: Treatment Effect 1.

Zhang, Xiaoquan (Michael) and Feng Zhu, "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia," *American Economic Review*, June 2011, *101* (4), 1601–1615.

Online Appendix for

A Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure

Liyang Sun, *UC Berkeley and CEMFI*
Jesse M. Shapiro, *Harvard University and NBER*[8]

February 2022

This appendix formalizes claims made in the paper.

Claim 1. In the setting of Section "The Possibility of Heterogeneous Coefficients," the expected value of the two-way fixed effects (TWFE) estimator of the exposure model, given the data $x = \{x_{10}, ..., x_{S0}\}$ for states $s \in \{1, ..., S\}$, is given by

$$\mathrm{E}\left(\hat{\beta}|x\right) = \frac{\mathrm{Cov}\left(\beta_s\left(1 - x_{s0}\right), \left(1 - x_{s0}\right)\right)}{\mathrm{Var}\left(1 - x_{s0}\right)}$$

where $\mathrm{Cov}\left(\cdot, \cdot\right)$ and $\mathrm{Var}\left(\cdot\right)$ denote the sample covariance and variance, respectively, and the expectation $\mathrm{E}\left(\hat{\beta}|x\right)$ is taken with respect to the distribution of the errors ε_{st} conditional on the data $x = \{x_{10}, ..., x_{S0}\}$.

Proof. With only two time periods the TWFE estimator of the exposure model is equivalent to an OLS estimator of the first-differenced model

$$y_{s1} - y_{s0} = \delta_1 - \delta_0 + \beta\left(1 - x_{s0}\right) + \varepsilon_{s1} - \varepsilon_{s0}.$$

Therefore the TWFE estimator based on the given sample is

$$\hat{\beta} = \frac{\mathrm{Cov}\left(y_{s1} - y_{s0}, 1 - x_{s0}\right)}{\mathrm{Var}\left(1 - x_{s0}\right)}.$$

[8]E-mail: lsun20@berkeley.edu, jesse_shapiro@fas.harvard.edu.

From the heterogeneous model we have that

$$y_{s1} - y_{s0} = \delta_1 - \delta_0 + \beta_s (1 - x_{s0}) + \varepsilon_{s1} - \varepsilon_{s0}$$

and therefore

$$\hat{\beta} = \frac{\text{Cov}\left(\beta_s (1 - x_{s0}), 1 - x_{s0}\right)}{\text{Var}\left(1 - x_{s0}\right)} + \frac{\text{Cov}\left(\varepsilon_{s1} - \varepsilon_{s0}, 1 - x_{s0}\right)}{\text{Var}\left(1 - x_{s0}\right)}.$$

If $(\varepsilon_{s1} - \varepsilon_{s0})$ is mean zero conditional on $(1 - x_{s0})$ then the expected value of $\hat{\beta}$ conditional on the data $x = \{x_{10}, ..., x_{S0}\}$ is

$$\text{E}\left(\hat{\beta}|x\right) = \frac{\text{Cov}\left(\beta_s (1 - x_{s0}), 1 - x_{s0}\right)}{\text{Var}\left(1 - x_{s0}\right)}.$$

□

Corollary 1. *In the setting of Section "The Possibility of Heterogeneous Coefficients," if β_s is independent of x_{s0} across states s, then the expected value of the two-way fixed effects (TWFE) estimator of the exposure model, given the data $x = \{x_{10}, ..., x_{S0}\}$ for states $s \in \{1, ..., S\}$, is given by*

$$\text{E}\left(\hat{\beta}|x\right) = \text{E}\left(\beta_s\right)$$

for $\text{E}(\beta_s)$ the expected value of β_s. Here the expectation $\text{E}\left(\hat{\beta}|x\right)$ is taken with respect to the distribution of the errors ε_{st} and coefficients β_s conditional on the data x.

Proof. Based on a similar proof for Claim 1, we have that

$$\text{E}\left(\hat{\beta}|x\right) = \frac{\text{E}\left(\text{Cov}\left(\beta_s (1 - x_{s0}), 1 - x_{s0}\right)\right)}{\text{Var}\left(1 - x_{s0}\right)}$$

where the expectation is now taken with respect to the distribution of the errors ε_{st} as well as β_s conditional on the data $x = \{x_{10}, ..., x_{S0}\}$. By the independence of β_s and x_{s0}, we have that

$$\text{E}\left(\text{Cov}\left(\beta_s (1 - x_{s0}), 1 - x_{s0}\right)\right) = \text{Cov}\left(\text{E}\left(\beta_s\right)(1 - x_{s0}), 1 - x_{s0}\right) = \text{E}\left(\beta_s\right)\text{Var}\left(1 - x_{s0}\right),$$

and therefore that

$$\mathrm{E}\left(\hat{\beta}|x\right) = \mathrm{E}\left(\beta_s\right).$$

□

Corollary 2. *In the numerical example of Section "The Possibility of Heterogeneous Coefficients," the expected value of the two-way fixed effects (TWFE) estimator of the exposure model, given the data* $x = \{x_{10}, ..., x_{S0}\}$ *for states* $s \in \{1, ..., S\}$*, lies outside the range of coefficients* $[\min_s \beta_s, \max_s \beta_s]$ *if and only if* $\lambda \neq 0$*. The same continues to hold when the sample is extended to include a totally unaffected state.*

Proof. From Claim 1 we have that

$$\mathrm{E}\left(\hat{\beta}|x\right) = \frac{\mathrm{Cov}\left(\beta_s\left(1 - x_{s0}\right), 1 - x_{s0}\right)}{\mathrm{Var}\left(1 - x_{s0}\right)}.$$

Because in the numerical example $\beta_s = 1 + 0.5\lambda - \lambda x_{s0}$, we have that

$$\mathrm{E}\left(\hat{\beta}|x\right) = 1 + 0.5\lambda - \lambda C$$

for

$$C = \frac{\mathrm{Cov}\left(x_{s0}\left(1 - x_{s0}\right), \left(1 - x_{s0}\right)\right)}{\mathrm{Var}\left(1 - x_{s0}\right)}.$$

In the setting of Section "The Possibility of Heterogeneous Coefficients," given the data $x = \{x_{10}, ..., x_{S0}\}$ where $x_{s0} = 0.245 + s/100$ for $s = 1, \ldots, 50$, by direct calculation we have that $C = 0$, which means that

$$\mathrm{E}\left(\hat{\beta}|x\right) = 1 + 0.5\lambda.$$

If we add to the sample a totally unaffected state $s = 0$ with $x_{00} = 1$, and the remaining states $s = 1, \ldots, 50$ continue to follow $x_{s0} = 0.245 + s/100$, by direct calculation we have that $C \approx 0.087$, which means that

$$\mathrm{E}\left(\hat{\beta}|x\right) \approx 1 + 0.413\lambda.$$

Therefore, with or without a totally unaffected state, when $\lambda > 0$ we have $\mathrm{E}\left(\hat{\beta}|x\right) > \beta_s$ for all s because $\max_s \beta_s = 1 + 0.245\lambda$. Similarly, with or without a totally unaffected state, when $\lambda < 0$ we have $\mathrm{E}\left(\hat{\beta}|x\right) < \beta_s$ for all s because

$\min_s \beta_s = 1 + 0.245\lambda$. Finally, with or without a totally unaffected state, when $\lambda = 0$ we have $\mathrm{E}\left(\hat{\beta}|x\right) = 1 = \mathrm{E}\left(\beta_s\right) = \max_s \beta_s = \min_s \beta_s$. □

Claim 2. In the setting of Section "The Possibility of Heterogeneous Coefficients," there exists no estimator $\hat{\beta}'$ that can be expressed as a function of the data $\{(x_{s0}, y_{s0}, y_{s1})\}_{s=1}^{S}$ and whose expected value is guaranteed to be contained in $[\min_s \beta_s, \max_s \beta_s]$ for any heterogeneous model and any $\{x_{s0}\}_{s=1}^{S}$.

Proof. It is sufficient to establish this claim for a special case with $S = 2$, some x_{s0}'s with $0 < x_{20} \leq x_{10} < 1$, $\beta_1 < \beta_2$, and δ_0 known to be zero. The model for the data is then

$$y_{s0} = \alpha_s + \beta_s \cdot x_{s0} + \varepsilon_{s0}$$
$$y_{s1} = \alpha_s + \delta_1 + \beta_s + \varepsilon_{s1}$$

with parameters $\theta = \left(\{(\alpha_s, \beta_s)\}_{s=1}^{2}, \delta_1, F_{\varepsilon|X}\right)$, for $F_{\varepsilon|X}$ the distribution of $(\varepsilon_{s0}, \varepsilon_{s1})$ conditional on x_{s0}. Pick some estimator $\hat{\beta}'$. Given any parameter θ, define the distinct parameter $\theta' = \left(\left\{\left(\alpha_s', \beta_s'\right)\right\}_{s=1}^{2}, \delta_1', F_{\varepsilon|X}\right)$ given by

$$\theta' = \left(\left\{\left(\alpha_s + \frac{\Delta \cdot x_{s0}}{1 - x_{s0}}, \beta_s - \frac{\Delta}{1 - x_{s0}}\right)\right\}_{s=1}^{2}, \delta_1 + \Delta, F_{\varepsilon|X}\right)$$

for some $\Delta > (\beta_2 - \beta_1) \cdot (1 - x_{20}) > 0$.

We show that the two parameter values θ and θ' are observationally equivalent, which means the expected value of $\hat{\beta}'$ must be the same under θ and θ'. To see this, note that the distribution of (y_{s0}, y_{s1}) conditional on x_{s0} is the same under θ and θ':

$$
\begin{aligned}
&F_{Y_0,Y_1|X}\left(y_0, y_1 \mid x_{s0} = x; \theta\right) \\
=&\Pr\left\{\varepsilon_{s0} \leq y_0 - \alpha_s - \beta_s \cdot x,\ \varepsilon_{s1} \leq y_1 - \alpha_s - \delta_1 - \beta_s \mid x_{s0} = x; \theta\right\} \\
=&\Pr\left\{\varepsilon_{s0} \leq y_0 - \alpha_s - \beta_s \cdot x,\ \varepsilon_{s1} - \varepsilon_{s0} \leq y_1 - y_0 - \delta_1 - \beta_s\left(1 - x\right) | x_{s0} = x; \theta\right\} \\
=&\Pr\left\{\begin{array}{c} \varepsilon_{s0} \leq y_0 - \left(\alpha_s + \frac{\Delta \cdot x}{1-x}\right) - \left(\beta_s - \frac{\Delta}{1-x}\right) \cdot x, \\ \varepsilon_{s1} - \varepsilon_{s0} \leq y_1 - y_0 - \left(\delta_1 + \Delta\right) - \left(\beta_s - \frac{\Delta}{1-x}\right)\left(1 - x\right) \end{array} \middle| x_{s0} = x; \theta \right\} \\
=&\Pr\left\{\begin{array}{c} \varepsilon_{s0} \leq y_0 - \alpha_s' - \beta_s' \cdot x, \\ \varepsilon_{s1} - \varepsilon_{s0} \leq y_1 - y_0 - \delta_1' - \beta_s'\left(1 - x\right) \end{array} \middle| x_{s0} = x; \theta' \right\} \\
=&F_{Y_0,Y_1|X}\left(y_0, y_1 \mid x_{s0} = x; \theta'\right).
\end{aligned}
$$

However, the Δ is chosen such that $\beta_1' = \beta_1 - \frac{\Delta}{1-x_{10}} < \beta_2 - \frac{\Delta}{1-x_{20}} = \beta_2' < \beta_1 < \beta_2$. Therefore the expected value of $\hat{\beta}'$ cannot be contained in both $[\beta_1, \beta_2]$ and $[\beta_1', \beta_2']$, because these intervals do not intersect. □

Claim 3. In the setting of Section "A Difference-in-Differences Perspective," the exposure-adjusted difference-in-differences estimator $\hat{\beta}_{s,s'}^{DID}$ is equivalent to the TWFE estimator $\hat{\beta}$ based on the two states s and s'. Moreover, the expected value of $\hat{\beta}_{s,s'}^{DID}$, given the data $x = \{x_{s0}, x_{s'0}\}$ for states s and s', is given by

$$
\mathrm{E}\left(\hat{\beta}_{s,s'}^{DID} | x\right) = \frac{\left(1 - x_{s0}\right)\beta_s - \left(1 - x_{s'0}\right)\beta_{s'}}{x_{s'0} - x_{s0}}
$$

where the expectation $\mathrm{E}\left(\hat{\beta}_{s,s'}^{DID} | x\right)$ is taken with respect to the distribution of the errors ε_{st} conditional on the data $x = \{x_{s0}, x_{s'0}\}$.

Proof. For the first part of the claim, note that from the proof of Claim 1 we have

$$
\hat{\beta} = \frac{\mathrm{Cov}\left(y_{s1} - y_{s0}, 1 - x_{s0}\right)}{\mathrm{Var}\left(1 - x_{s0}\right)}
$$

where $\mathrm{Cov}\left(\cdot, \cdot\right)$ and $\mathrm{Var}\left(\cdot\right)$ denote the sample covariance and variance, respectively.

Since the sample includes only two states s and s', for the numerator we have

$$\begin{aligned}
&\mathrm{Cov}\left(y_{s1}-y_{s0}, 1-x_{s0}\right) \\
&= \frac{1}{4}\left(\left(y_{s1}-y_{s0}\right)-\left(y_{s'1}-y_{s'0}\right)\right)\left(1-x_{s0}\right)+\frac{1}{4}\left(\left(y_{s'1}-y_{s'0}\right)-\left(y_{s1}-y_{s0}\right)\right)\left(1-x_{s'0}\right) \\
&= \frac{1}{4}\left(\left(1-x_{s0}\right)-\left(1-x_{s'0}\right)\right)\left(\left(y_{s1}-y_{s0}\right)-\left(y_{s'1}-y_{s'0}\right)\right)
\end{aligned}$$

where the first equality applies the definition of sample covariance and $a-\frac{a+b}{2}=\frac{a-b}{2}$. Similarly, for the denominator we have

$$\mathrm{Var}\left(1-x_{s0}\right)=\frac{1}{4}\left(\left(1-x_{s0}\right)-\left(1-x_{s'0}\right)\right)^2.$$

Plugging the above expressions into $\hat{\beta}$ gives the equivalence to $\hat{\beta}^{DID}_{s,s'}$.

Given the equivalence between $\hat{\beta}$ and $\hat{\beta}^{DID}_{s,s'}$ when the sample includes only two states s and s', we apply Claim 1 to derive the expected value of $\hat{\beta}^{DID}_{s,s'}$. Specifically, Claim 1 implies that given the data $x=\{x_{s0}, x_{s'0}\}$ for states s and s', we have

$$\mathrm{E}\left(\hat{\beta}^{DID}_{s,s'}|x\right)=\frac{\mathrm{Cov}\left(\beta_s\left(1-x_{s0}\right), 1-x_{s0}\right)}{\mathrm{Var}\left(1-x_{s0}\right)}.$$

Based on a similar simplification to the expression of $\hat{\beta}^{DID}_{s,s'}$, we have

$$\mathrm{Cov}\left(\beta_s\left(1-x_{s0}\right), 1-x_{s0}\right)=\frac{1}{4}\left(\left(1-x_{s0}\right)-\left(1-x_{s'0}\right)\right)\left(\left(1-x_{s0}\right)\beta_s-\left(1-x_{s'0}\right)\beta_{s'}\right)$$

and therefore

$$\frac{\mathrm{Cov}\left(\beta_s\left(1-x_{s0}\right), 1-x_{s0}\right)}{\mathrm{Var}\left(1-x_{s0}\right)}=\frac{\left(1-x_{s0}\right)\beta_s-\left(1-x_{s'0}\right)\beta_{s'}}{x_{s'0}-x_{s0}}.$$

□