Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25

PA

# A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment

**Peng Ding**[1] **and Fan Li**[2]

[1] *Department of Statistics, University of California, 425 Evans Hall, Berkeley, CA 94720, USA. Email: pengdingpku@berkeley.edu*
[2] *Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708, USA. Email: fl35@duke.edu*

## Abstract

Difference-in-differences is a widely used evaluation strategy that draws causal inference from observational panel data. Its causal identification relies on the assumption of parallel trends, which is scale-dependent and may be questionable in some applications. A common alternative is a regression model that adjusts for the lagged dependent variable, which rests on the assumption of ignorability conditional on past outcomes. In the context of linear models, Angrist and Pischke (2009) show that the difference-in-differences and lagged-dependent-variable regression estimates have a bracketing relationship. Namely, for a true positive effect, if ignorability is correct, then mistakenly assuming parallel trends will overestimate the effect; in contrast, if the parallel trends assumption is correct, then mistakenly assuming ignorability will underestimate the effect. We show that the same bracketing relationship holds in general nonparametric (model-free) settings. We also extend the result to semiparametric estimation based on inverse probability weighting. We provide three examples to illustrate the theoretical results with replication files in Ding and Li (2019).

*Keywords:* causal inference, ignorability, nonparametric, panel data, parallel trends

## 1 Introduction

Difference-in-differences is a popular evaluation strategy in the social sciences; it makes causal comparisons from observational panel data by exploiting variation across time (Ashenfelter 1978; Bertrand, Duflo, and Mullainathan 2004; Angrist and Pischke 2009; Bechtel and Hainmueller 2011; Keele and Minozzi 2013; Malesky, Nguyen, and Tran 2014; Keele *et al.* 2019; Callaway and Sant'Anna 2019). The key assumption underlying difference-in-differences is *parallel trends*, that is, the counterfactual trend behavior of treatment and control groups, in the absence of treatment, is the same, possibly conditioning on some observed covariates (Heckman, Ichimura, and Todd 1997; Abadie 2005). In practice, the parallel trends assumption can be questionable because unobserved confounders may have time-varying effects on the outcomes. A common alternative method is a regression model that adjusts for the lagged dependent variables (Ashenfelter 1978), which assumes *ignorability* conditional on past outcomes and observed covariates.

Difference-in-differences and lagged-dependent-variable adjustment—also known respectively as the gain score estimator and the analysis of covariance estimator in sociology and psychology—are two different methods relying on different identification assumptions. Extensive conceptual, empirical and numerical comparisons between the two methods have been made in the literature (e.g., Allison 1990; Maris 1998; van Breukelen 2013; Ryan, Burgess, and Dimick 2015; O'Neill *et al.* 2016). In particular, in the context of linear models, Angrist and Pischke (2009) show that difference-in-differences and lagged-dependent-variable regression estimators have

*Authors' note:* We are grateful to three reviewers, Jas Sekhon, Benjamin Lu, Olivia Angiuli, and Frank Li for helpful comments, to Avi Feller for kindly editing the paper, and to Eric Donnell for providing the transportation data. Peng Ding is supported by the National Science Foundation grant NSF-DMS1713152.

***Political Analysis (2019)***
**vol. 27:605–615**
**DOI:** 10.1017/pan.2019.25

**Published**
11 July 2019

**Corresponding author**
Peng Ding

**Edited by**
Jeff Gill

© The Author(s) 2019. Published by Cambridge University Press on behalf of the Society for Political Methodology.

Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25

a *bracketing* relationship. Namely, for a true positive effect, if the ignorability assumption is correct, then mistakenly assuming parallel trends will overestimate the effect; in contrast, if the parallel trends assumption is correct, then mistakenly assuming ignorability will underestimate the effect. The opposite holds for a true negative effect.

The bracketing relationship is important in practice. Though we usually do not know which one of the two assumptions is true in real applications, we can analyze the data under each assumption and treat the estimates as the upper and lower bounds of the true effect. However, the linear setting in Angrist and Pischke (2009) is restrictive, particularly for applications with noncontinuous outcomes. For example, binary outcomes are common in political science (e.g., Keele and Minozzi 2013; Malesky, Nguyen, and Tran 2014) and health studies (e.g., Stuart *et al.* 2014); count outcome are common in transportation safety studies where the before–after design is popular (e.g., Hauer 1997). Moreover, the parallel trends assumption is scale-dependent (Athey and Imbens 2006). Therefore, an extension to nonlinear settings is relevant for both theory and practice. In this paper, we prove that, within the canonical two-period two-group setting, the same bracketing relationship holds in general nonparametric and semiparametric settings. We give three examples to illustrate the theoretical results.

## 2 Setup

### 2.1 Difference-in-Differences

We proceed under the potential outcomes framework (Neyman 1923; Rubin 1974). We consider the basic two-period two-group panel design, where a sample of units, indexed by $i \in \{1, \ldots, n\}$, are drawn from a target population of two groups, labeled by $G_i = 0$ or 1. Each unit can potentially be assigned to a treatment $d$, with $d = 1$ for the active treatment and $d = 0$ for the control. Units in both groups are followed in two periods of time $T$, with $T = t$ and $T = t + 1$ denoting the before and after period, respectively. The treatment is only administered to the group with $G_i = 1$ in the after period. For each unit $i$, let $D_{iT}$ be the observed treatment status at time $T$. The above design implies $D_{it} = 0$ for all units and $D_{i,t+1} = 1$ for the units in group $G_i = 1$; thus $G_i = D_{i,t+1}$. Assume that each unit has two potential outcomes in each period, $\{Y_{iT}(1), Y_{iT}(0)\}$ for $T = t$ and $t+1$, and only the one corresponding to the observed treatment status, $Y_{iT} = Y_{iT}(D_{iT})$, is observed. Therefore, $Y_{it} = Y_{it}(0)$ and $Y_{i,t+1} = (1-G_i)Y_{i,t+1}(0) + G_i Y_{i,t+1}(1)$. For each unit, a vector of pretreatment covariates $X_i$ are also observed in the before period.

In the two-period two-group panel design, the target estimand is usually the average treatment effect for the treated (ATT) (Abadie 2005; Angrist and Pischke 2009; Lechner 2011):

$$\tau_{\text{ATT}} \equiv E\{Y_{i,t+1}(1) - Y_{i,t+1}(0) \mid G_i = 1\} = \mu_1 - \mu_0, \tag{1}$$

where $\mu_1 = E\{Y_{i,t+1}(1) \mid G_i = 1\}$ and $\mu_0 = E\{Y_{i,t+1}(0) \mid G_i = 1\}$. When the outcome is discrete, ratio versions of $\tau_{\text{ATT}}$ are often of interest, such as

$$\gamma_{\text{ATT}} \equiv E\{Y_{i,t+1}(1) \mid G_i = 1\}/E\{Y_{i,t+1}(0) \mid G_i = 1\} = \mu_1/\mu_0, \tag{2}$$

which is the causal risk ratio for binary outcomes and the causal rate ratio for count outcomes.

The quantity $\mu_1$ equals $E(Y_{i,t+1} \mid G_i = 1)$, and thus is directly estimable from the observed data, e.g., by the moment estimator $\bar{Y}_{1,t+1} = \sum_{i=1}^n G_i Y_{i,t+1} / \sum_{i=1}^n G_i$. In contrast, the quantity $\mu_0$, the counterfactual outcome for the treatment group in the after period in the absence of treatment, is not observable and must rely on additional assumptions to identify. The central task in this design is to use the observed data to estimate the counterfactual $\mu_0$. Any consistent estimator of $\mu_0$ leads to consistent estimators of $\tau_{\text{ATT}}$ and $\gamma_{\text{ATT}}$.

With difference-in-differences, the key for identifying $\mu_0$ is the *parallel trends* assumption.

Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25

ASSUMPTION 1 (Parallel trends). $E\{Y_{i,t+1}(0) - Y_{it}(0) \mid X_i, G_i = 1\} = E\{Y_{i,t+1}(0) - Y_{it}(0) \mid X_i, G_i = 0\}$.

The parallel trends assumption requires that, conditional on covariates $X_i$, the average outcomes in the treated and control groups in the absence of treatment would have followed parallel paths over time. Under Assumption 1, we have the nonparametric identification formula for $\mu_0$:

$$\begin{aligned}\tilde{\mu}_{0,\text{DID}} &= E\big[E\{Y_{it}(0) \mid X_i, G_i = 1\} + E\{Y_{i,t+1}(0) - Y_{it}(0) \mid X_i, G_i = 0\} \mid G_i = 1\big] \\ &= E(Y_{it} \mid G_i = 1) + E\{E(Y_{i,t+1} - Y_{it} \mid X_i, G_i = 0) \mid G_i = 1\} \\ &= E(Y_{it} \mid G_i = 1) + \int E(Y_{i,t+1} - Y_{it} \mid X_i = x, G_i = 0) F_{X|G=1}(\mathrm{d}x), \end{aligned} \tag{3}$$

where $F_{X|G=1}(x) = \text{pr}(X \le x \mid G = 1)$ is the distribution of $X$ in the treatment group. All terms of the right hand side of (3) are identifiable from the observed data. A stronger version of Assumption 1 imposes parallel trends without conditioning on covariates, under which we can write

$$\tilde{\mu}_{0,\text{DID}} = E(Y_{it} \mid G_i = 1) + E(Y_{i,t+1} \mid G_i = 0) - E(Y_{it} \mid G_i = 0). \tag{4}$$

Based on the identification formula (4), a moment estimator of $\tau_{\text{ATT}}$ is

$$\hat{\tau}_{\text{DID}} = (\bar{Y}_{1,t+1} - \bar{Y}_{1,t}) - (\bar{Y}_{0,t+1} - \bar{Y}_{0,t}), \tag{5}$$

where $\bar{Y}_{g,T}$ is the mean observed outcome for group $g$ at time $T$ ($g = 0, 1; T = t, t+1$). The form of this estimator underlies the name "difference-in-differences".

A well-known limitation of the difference-in-differences approach is that the parallel trends assumption depends on the scale of the outcome (Athey and Imbens 2006; Lechner 2011). Specifically, the parallel trends assumption may hold for the original $Y$ but not for a nonlinear monotone transformation of $Y$, for example, $\log Y$. This scale dependence restricts the use of difference-in-differences in settings with non-Gaussian and discrete outcomes.

## 2.2 Lagged-Dependent-Variable Adjustment

In the treatment–control panel design, a class of alternative methods relies on the assumption of ignorability conditional on the lagged dependent variable, that is, in the absence of treatment, the outcomes for the treated and control groups would have the same distributions, conditional on their lagged outcome and covariates.

ASSUMPTION 2 (Ignorability). $Y_{i,t+1}(0) \perp\!\!\!\perp G_i \mid (Y_{it}, X_i)$.

Under ignorability, we have the following nonparametric identification formula of $\mu_0$:

$$\begin{aligned}\tilde{\mu}_{0,\text{LDV}} &= E\big[E\{Y_{i,t+1}(0) \mid G_i = 1, Y_{it}, X_i\} \mid G_i = 1\big] \\ &= E\{E(Y_{i,t+1} \mid G_i = 0, Y_{it}, X_i) \mid G_i = 1\} \\ &= \int E(Y_{i,t+1} \mid G_i = 0, Y_{it} = y, X_i = x) F_{Y_t,X|G=1}(\mathrm{d}y, \mathrm{d}x), \end{aligned} \tag{6}$$

where $F_{Y_t,X|G=1}(y, x)$ is the joint distribution of $(Y_t, X)$ in the treatment group. The form of $\tilde{\mu}_{0,\text{LDV}}$ is identical to the traditional identification formula for the average treatment effect for the treated in observational cross-sectional studies. We can specify a model for $E\{Y_{i,t+1}(0) \mid Y_{it}, G_i, X_i\}$, based on which we impute the counterfactual mean $\mu_0 = E\{Y_{i,t+1}(0) \mid G_i = 1\}$ by averaging over $Y_t$ and $X$ and thus obtain a consistent estimator for $\tau_{\text{ATT}}$.

Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25

In contrast to the parallel trends assumption, the ignorability assumption is scale-free. Three popular methods under the ignorability assumption are the synthetic control method (Abadie and Gardeazabal 2003; Abadie, Diamond, and Hainmueller 2015), matching (Heckman, Ichimura, and Todd 1997) or regression adjustment (Ashenfelter 1978) of the lagged dependent variable. Among these, the lagged-dependent-variable adjustment approach is the easiest to implement. Through extensive simulations, O'Neill *et al.* (2016) have found that, when the parallel trends assumption does not hold, the lagged-dependent-variable regression adjustment approach produces the most efficient and least biased estimates among these three methods.

## 3 Theory

Our goal in this section is to establish the analytical relationship between the difference-in-differences and lagged-dependent-variable adjustment estimators under general settings. For notational simplicity, we condition on the covariates $X$ and thus ignore them in the discussion.

### 3.1 Bracketing Relationship in Linear Models

We start with the simple case of linear regressions. Specifically, the difference-in-differences approach is usually implemented via a linear fixed-effects model:

$$E(Y_{iT} \mid D_{iT}, X_i) = \alpha_i + \lambda_T + \tau D_{iT}, \quad (i = 1, \ldots, n; T = t, t+1) \tag{7}$$

where $\alpha_i$ is the individual fixed effect and $\lambda_T$ is the time-specific fixed effect. When model (7) is correct, the coefficient $\tau$ equals the estimand $\tau_{\text{ATT}}$; any consistent estimator of $\tau$ in (7) is also consistent for $\tau_{\text{ATT}}$. By taking the difference between outcomes at time points $t$ and $t+1$ in (7), we can eliminate the individual fixed-effects $\alpha_i$. Because $G_i = D_{i,t+1} - D_{it}$, we have $E(Y_{i,t+1} - Y_{it} \mid G_i) = (\lambda_{t+1} - \lambda_t) + \tau G_i$. Therefore, we can fit a linear regression of the difference $Y_{i,t+1} - Y_{it}$ on the group indicator $G_i$ to estimate $\tau$. The resulting ordinary least squares estimator is the difference between the sample means of $Y_{i,t+1} - Y_{it}$ in the treated and control groups, and thus it equals $\hat{\tau}_{\text{DID}}$ defined in (5).

The lagged-dependent-variable adjustment method can be implemented via linear models in two ways. In the first approach, motivated by (6), we can fit an ordinary least squares line $\hat{E}(Y_{t+1} \mid G = 0, Y_t = y) = \hat{\alpha} + \hat{\beta} Y_t$ using only the control units; then we obtain $\hat{\mu}_{0,\text{LDV}} = \hat{\alpha} + \hat{\beta} \bar{Y}_{1,t}$ as the sample analog of $\tilde{\mu}_{0,\text{LDV}}$ and $\hat{\tau}_{\text{LDV}} = \bar{Y}_{1,t+1} - \hat{\mu}_{0,\text{LDV}}$ as the estimate of $\tau_{\text{ATT}}$. In the second approach, as in Angrist and Pischke (2009, ch. 5.4), we can use the following linear model:

$$E(Y_{i,t+1} \mid Y_{it}, G_i) = \alpha + \beta Y_{it} + \tau G_i. \tag{8}$$

When model (8) is correct, the coefficient $\tau$ equals the causal estimand $\tau_{\text{ATT}}$, and any consistent estimator of $\tau$ is consistent for $\tau_{\text{ATT}}$. We can fit the ordinary least squares line $\hat{E}(Y_{t+1} \mid G, Y_t) = \hat{\alpha} + \hat{\tau}'_{\text{LDV}} G + \hat{\beta}' Y_t$ using all units and take the coefficient $\hat{\tau}'_{\text{LDV}}$ as an estimate of $\tau_{\text{ATT}}$. We have the following expressions for the two estimators $\hat{\tau}_{\text{LDV}}$ and $\hat{\tau}'_{\text{LDV}}$ (the proof is given in the Appendix).

PROPOSITION 1. *Without covariates, the two lagged-dependent-variable adjustment estimates are*

$$\hat{\tau}_{\text{LDV}} = (\bar{Y}_{1,t+1} - \bar{Y}_{0,t+1}) - \hat{\beta}(\bar{Y}_{1,t} - \bar{Y}_{0,t}), \quad \hat{\tau}'_{\text{LDV}} = (\bar{Y}_{1,t+1} - \bar{Y}_{0,t+1}) - \hat{\beta}'(\bar{Y}_{1,t} - \bar{Y}_{0,t}). \tag{9}$$

These two estimates in (9) differ from the moment difference-in-differences estimate $\hat{\tau}_{\text{DID}} = (\bar{Y}_{1,t+1} - \bar{Y}_{0,t+1}) - (\bar{Y}_{1,t} - \bar{Y}_{0,t})$ only in the coefficients $\hat{\beta}$ and $\hat{\beta}'$. Consider the case with $\hat{\beta}$ or $\hat{\beta}'$ larger than 0 but smaller than 1. The sign of $\hat{\tau}_{\text{DID}} - \hat{\tau}_{\text{LDV}}$ or $\hat{\tau}_{\text{DID}} - \hat{\tau}'_{\text{LDV}}$ depends on the sign of $\bar{Y}_{1,t} - \bar{Y}_{0,t}$. If the treatment group has larger lagged outcome $Y_t$ on average, then $\hat{\tau}_{\text{DID}} < \hat{\tau}_{\text{LDV}}$; if the treatment

group has smaller $Y_t$ on average, then $\hat{\tau}_{\text{DID}} > \hat{\tau}_{\text{LDV}}$. In the special case with $\hat{\beta} = 1$ or $\hat{\beta}' = 1$, they are *identical*: $\hat{\tau}_{\text{DID}} = \hat{\tau}_{\text{LDV}}$ or $\hat{\tau}'_{\text{DID}} = \hat{\tau}_{\text{LDV}}$. How much $\hat{\beta}$ or $\hat{\beta}'$ deviates from 1 indicates how different the two estimates are. We see this phenomenon in the examples in Section 4. Importantly, the discussion in this subsection holds without imposing any stochastic assumptions. That is, Proposition 1 is a purely numerical result. In contrast, the bracketing relationship in Angrist and Pischke (2009, ch. 5.4) is proven under the linear model assumptions.

Gelman (2007) pointed out that restricting $\beta = 1$ in (8) gives identical least squares estimators for $\tau$ from models (7) and (8), which is also evident from Proposition 1. However, the nonparametric identification Assumptions 1 and 2 are not nested, and the difference-in-differences estimator is not a special case of the lagged-dependent-variable adjustment estimator in general. Therefore, it is natural to investigate whether Angrist and Pischke (2009)'s result is unique to the linear models (7) and (8). In the next subsection, we generalize the bracketing relationship to model-free settings.

## 3.2 Nonparametric Bracketing Relationship

For notational simplicity, below we also drop the subscript $i$. Under ignorability, the nonparametric identification formula (6) of $\mu_0$ simplifies to

$$\tilde{\mu}_{0,\text{LDV}} = E\{E(Y_{t+1} \mid G = 0, Y_t) \mid G = 1\} = \int E(Y_{t+1} \mid G = 0, Y_t = y) F_{Y_t}(\mathrm{d}y \mid G = 1), \quad (10)$$

where $F_{Y_t}(y \mid G = g) = \text{pr}(Y_t \le y \mid G = g)$ is the cumulative distribution function of $Y_t$ for units in group $g$ ($g = 0, 1$). The form of $\tilde{\mu}_{0,\text{LDV}}$ is identical to the identification formula for the ATT estimand in cross-sectional studies.

To compare $\tilde{\tau}_{\text{DID}}$ and $\tilde{\tau}_{\text{LDV}}$ without imposing any functional form of the outcome model, we first obtain the following analytical difference between $\tilde{\mu}_{0,\text{DID}}$ and $\tilde{\mu}_{0,\text{LDV}}$ (the proof is given in the Appendix).

LEMMA 1. *The difference between $\tilde{\mu}_{0,\text{DID}}$ and $\tilde{\mu}_{0,\text{LDV}}$ is*

$$\tilde{\mu}_{0,\text{LDV}} - \tilde{\mu}_{0,\text{DID}} = \int \Delta(y) F_{Y_t}(\mathrm{d}y \mid G = 1) - \int \Delta(y) F_{Y_t}(\mathrm{d}y \mid G = 0),$$

*where $\Delta(y) = E(Y_{t+1} \mid G = 0, Y_t = y) - y$.*

The quantity $\Delta(y) = E(Y_{t+1} \mid G = 0, Y_t = y) - y = E(Y_{t+1} - Y_t \mid G = 0, Y_t = y)$ equals the expectation of the change in the outcome conditioning on the lagged outcome in the control group. Lemma 1 suggests that the relative magnitude of $\tilde{\mu}_{0,\text{DID}}$ and $\tilde{\mu}_{0,\text{LDV}}$ depends on (a) the expectation of the before–after difference $Y_{t+1} - Y_t$ conditional on $Y_t$ in the control group, and (b) the difference between the distribution of the before outcome $Y_t$ in the treated and control groups. Both are important characteristics of the underlying data generating process, which measures (a) the dependence of the outcome on the lagged outcome and (b) the dependence of the treatment assignment on the lagged outcome, respectively. In particular, if $Y_t \perp\!\!\!\perp G$ or equivalently $F_{Y_t}(y \mid G = 1) = F_{Y_t}(y \mid G = 0)$, then $\tilde{\mu}_{0,\text{LDV}} = \tilde{\mu}_{0,\text{DID}}$.

To reach the main conclusion, we introduce two additional conditions regarding the quantities in Lemma 1. The first is a stationarity condition on the outcome.

CONDITION 1 (Stationarity). $\partial E(Y_{t+1} \mid G = 0, Y_t = y)/\partial y < 1$ for all $y$.

In a linear model for $E(Y_{t+1} \mid G = 0, Y_t = y)$, Condition 1 requires that, in the control group, the regression coefficient of the outcome $Y_{t+1}$ on the lagged outcome $Y_t$ is smaller than 1; this is also invoked by Angrist and Pischke (2009). Its sample version is $\hat{\beta} < 1$ or $\hat{\beta}' < 1$ as in Section 3.1. In general, Condition 1 ensures that the time series of the outcomes would not grow infinitely as time, which is reasonable in most applications.



Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25

Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25

The second condition describes the treatment assignment mechanism with respect to the lagged outcome, with two opposite versions.

CONDITION 2 (Stochastic Monotonicity). (a) $F_{Y_t}(y \mid G = 1) \geq F_{Y_t}(y \mid G = 0)$ for all $y$; (b) $F_{Y_t}(y \mid G = 1) \leq F_{Y_t}(y \mid G = 0)$ for all $y$.

Condition 2(a) implies that the treated group has smaller lagged outcome compared to the control group, and Condition 2(b) implies the opposite relationship. In the case of linear models, Condition 2(a) or (b) reduces to the *average* lagged outcome in the treated group is smaller or larger than that in the control group, respectively.

Because they only involve observed variables, Conditions 1 and 2 are testable empirically. Specifically, to check Condition 1, we can estimate the derivative of the conditional mean function $E(Y_{t+1} \mid G = 0, Y_t = y)$; to check Condition 2, we can visually compare the empirical cumulative probability distributions of the outcomes in the treatment and control groups. These conditions hold in many applications, e.g., in the examples in Section 4. In contrast, Assumptions 1 and 2 are in general untestable.

Under Conditions 1 and 2, we have the following results on the bracketing relationship on $\tilde{\tau}_{\text{DID}}$ and $\tilde{\tau}_{\text{LDV}}$ in a nonparametric setting; see the Appendix for the proof.

THEOREM 1. *If Conditions 1 and 2(a) hold, then $\tilde{\mu}_{0,\text{DID}} \leq \tilde{\mu}_{0,\text{LDV}}$, and thus $\tilde{\tau}_{\text{DID}} \geq \tilde{\tau}_{\text{LDV}}$ and $\tilde{\gamma}_{\text{DID}} \geq \tilde{\gamma}_{\text{LDV}}$; if Conditions 1 and 2(b) hold, then $\tilde{\mu}_{0,\text{DID}} \geq \tilde{\mu}_{0,\text{LDV}}$, and thus $\tilde{\tau}_{\text{DID}} \leq \tilde{\tau}_{\text{LDV}}$ and $\tilde{\gamma}_{\text{DID}} \leq \tilde{\gamma}_{\text{LDV}}$.*

Theorem 1 is a result on the relative magnitude between the two quantities $\tilde{\tau}_{\text{DID}}$ and $\tilde{\tau}_{\text{LDV}}$ (and between $\tilde{\gamma}_{\text{DID}}$ and $\tilde{\gamma}_{\text{LDV}}$). On the one hand, Theorem 1 holds without requiring either Assumption 1 or 2. Specifically, under Stationarity and Stochastic Monotonicity (a), $\tilde{\tau}_{\text{DID}}$ is larger than or equal to $\tilde{\tau}_{\text{LDV}}$. Both of them can be biased for the true causal effect $\tau_{\text{ATT}}$: if $\tilde{\tau}_{\text{DID}} \geq \tilde{\tau}_{\text{LDV}} \geq \tau_{\text{ATT}}$, then $\tilde{\tau}_{\text{DID}}$ overestimates $\tau_{\text{ATT}}$ more than $\tilde{\tau}_{\text{LDV}}$; if $\tau_{\text{ATT}} \geq \tilde{\tau}_{\text{DID}} \geq \tilde{\tau}_{\text{LDV}}$, then $\tilde{\tau}_{\text{LDV}}$ underestimate $\tau_{\text{ATT}}$ more than $\tilde{\tau}_{\text{DID}}$; if $\tilde{\tau}_{\text{DID}} \geq \tau_{\text{ATT}} \geq \tilde{\tau}_{\text{LDV}}$, then $\tilde{\tau}_{\text{DID}}$ and $\tilde{\tau}_{\text{LDV}}$ are the upper and lower bounds on $\tau_{\text{ATT}}$. Analogous arguments apply under Stationarity and Stochastic Monotonicity (b). On the other hand, only under Assumption 1 or 2, the quantities $\tilde{\tau}_{\text{DID}}$ and $\tilde{\tau}_{\text{LDV}}$ have the interpretation as the nonparametric identification formulas of the causal estimand $\tau_{\text{ATT}}$. We stress that the bracket $(\tilde{\tau}_{\text{DID}}, \tilde{\tau}_{\text{LDV}})$ provides bounds for the true effect $\tau_{\text{ATT}}$ if either Assumption 1 or 2 holds; however, it does not answer the question about whether the true effect falls inside or, if outside, which side of the bracket when neither Assumption 1 nor 2 holds. The relationship under such a scenario is dependent on the specific true data generating model.

For discrete outcomes, Equation (10) reduces to $\tilde{\mu}_{0,\text{LDV}} = \sum_y E(Y_{t+1} \mid G = 0, Y_t = y)\text{pr}(Y_t = y \mid G = 1)$, and the stationary condition becomes $E(Y_{t+1} \mid G = 0, Y_t = y + 1) - E(Y_{t+1} \mid G = 0, Y_t = y) < 1$ for all values of $y$. For the case of binary outcome, the stationary condition always holds because $0 \leq E(Y_{t+1} \mid G = 0, Y_t = y) \leq 1$ for $y = 0, 1$. Therefore, we only need to check the sign of the empirical counterpart of $\text{pr}(Y_t = 0 \mid G = 1) - \text{pr}(Y_t = 0 \mid G = 0)$. Specifically, if $\text{pr}(Y_t = 0 \mid G = 1) \geq \text{pr}(Y_t = 0 \mid G = 0)$, then $\tilde{\tau}_{\text{DID}} \geq \tilde{\tau}_{\text{LDV}}$ and $\tilde{\gamma}_{\text{DID}} \geq \tilde{\gamma}_{\text{LDV}}$; if $\text{pr}(Y_t = 0 \mid G = 1) \leq \text{pr}(Y_t = 0 \mid G = 0)$, then $\tilde{\tau}_{\text{DID}} \leq \tilde{\tau}_{\text{LDV}}$ and $\tilde{\gamma}_{\text{DID}} \leq \tilde{\gamma}_{\text{LDV}}$.

## 3.3 Semiparametric Bracketing Relationship

Under the parallel trends Assumption 1, Abadie (2005) proposed a semiparametric inverse probability weighting estimator for $\tau_{\text{ATT}}$ based on the following identification formula of $\mu_0$:

$$\tilde{\mu}'_{0,\text{DID}} = E\left\{GY_t + \frac{e(1-G)(Y_{t+1} - Y_t)}{1-e}\right\} \Big/ \text{pr}(G = 1), \tag{11}$$

where the propensity score is defined as $e = \text{pr}(G = 1)$. Abadie (2005)'s estimator based on $\tilde{\mu}_{0,\text{DID}}$ shares the same form as the inverse probability weighting estimator for the ATT in the

Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25

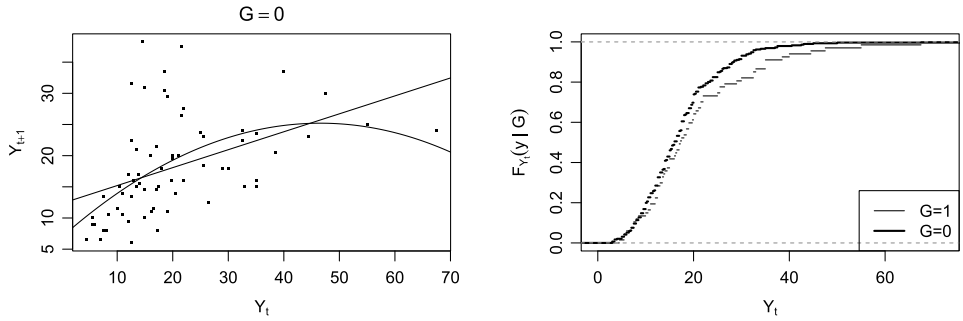


**Figure 1.** Card and Krueger (1994) study. Left: linear and quadratic fitted lines of $E(Y_{t+1} \mid G = 0, Y_t)$. Right: empirical distribution functions $F_{Y_t}(y \mid G = g)$ $(g = 0, 1)$ satisfy Stochastic Monotonicity.

cross-sectional setting, but replaces the outcome in the treatment group by the before–after difference $Y_{t+1} - Y_t$. Similarly, under Assumption 2, we can construct a semiparametric estimator based on

$$\tilde{\mu}'_{0,\text{LDV}} = E\left\{\frac{e(Y_t)}{1-e(Y_t)}(1-G)Y_{t+1}\right\} \Big/ \text{pr}(G=1), \tag{12}$$

where the propensity score is defined as $e(Y_t) = \text{pr}(G = 1 \mid Y_t)$.

Because (11) and (12) are alternative identification formulas for $\mu_0$, we can show that $(\tilde{\mu}_{0,\text{DID}}, \tilde{\mu}_{0,\text{LDV}}) = (\tilde{\mu}'_{0,\text{DID}}, \tilde{\mu}'_{0,\text{LDV}})$ and thus have the following corollary of Theorem 1.

COROLLARY 1. *Theorem 1 holds if* $(\tilde{\mu}_{0,\text{DID}}, \tilde{\mu}_{0,\text{LDV}})$ *are replaced by* $(\tilde{\mu}'_{0,\text{DID}}, \tilde{\mu}'_{0,\text{LDV}})$.

Corollary 1 shows that the bracketing relationship between $\tilde{\tau}_{\text{DID}}$ and $\tilde{\tau}_{\text{LDV}}$ does not depend on the forms of identification formulas and estimators.

## 4 Examples

### 4.1 Minimum Wages and Employment

We reanalyze part of the classic Card and Krueger (1994) study on the effect of a minimum wage increase on employment. Data were collected on the employment information at fast food restaurants in New Jersey and Pennsylvania before and after a minimum wage increase in New Jersey in 1992. The outcome is the number of full-time-equivalent employees at each restaurant.

The difference-in-differences estimate is $\hat{\tau}_{\text{DID}} = 2.446$, and the lagged-dependent-variable adjustment estimates are $\hat{\tau}_{\text{LDV}} = 0.302$ and $\hat{\tau}'_{\text{LDV}} = 0.865$ with coefficients of the lag outcome $\hat{\beta} = 0.288 < 1$ and $\hat{\beta}' = 0.475 < 1$. Meanwhile, because the sample means satisfy $\bar{Y}_{1,t} - \bar{Y}_{0,t} = 17.289 - 20.299 < 0$, our theoretical result predicts that $\hat{\tau}_{\text{DID}} > \hat{\tau}_{\text{LDV}}$(or $\hat{\tau}'_{\text{LDV}}$), which exactly matches the relative magnitude of the empirical estimates above. In addition, if we adopt a quadratic specification of $E(Y_{t+1} \mid G = 0, Y_t)$, the lagged-dependent-variable regression estimate becomes $\hat{\tau}_{\text{LDV}} = 0.275$, which is also smaller than $\hat{\tau}_{\text{DID}}$. This is again coherent with our theory because Stationarity and Stochastic Monotonicity hold, depicted in Figure 1. In this example, the differences between $\hat{\tau}_{\text{DID}}$ and $\hat{\tau}_{\text{LDV}}$(or $\hat{\tau}'_{\text{LDV}}$) are significant at level 0.05.

### 4.2 Electoral Returns to Beneficial Policy

We reanalyze the Bechtel and Hainmueller (2011) study on electoral returns to beneficial policy. We focus on the short-term electoral returns by analyzing the causal effect of disaster relief aid due to the 2002 Elbe flooding in Germany. The before period is 1998 and the after period is 2002. The units of analysis are electoral districts, the treatment is the indicator whether a district is affected by the flood, and the outcome is the vote share that the Social Democratic Party attains in that district.

Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25

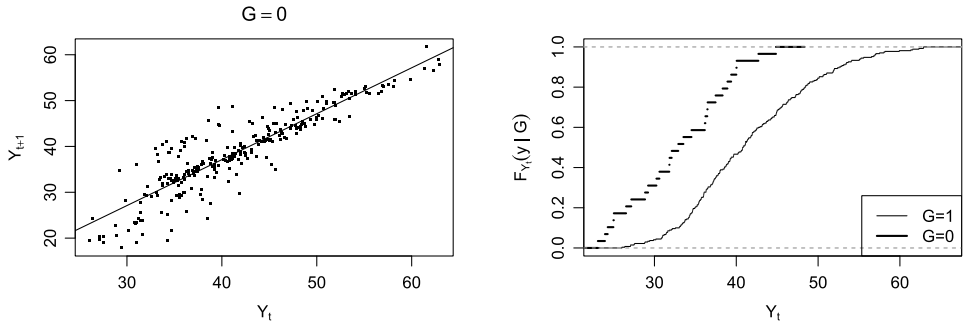


**Figure 2.** Bechtel and Hainmueller (2011) study. Left: linear fitted lines of $E(Y_{t+1} \mid G = 0, Y_t)$. Right: empirical distribution functions $F_{Y_t}(y \mid G = g)$ $(g = 0, 1)$ satisfy Stochastic Monotonicity.

**Table 1.** Crash counts in the 1986 road sites in Pennsylvania (3+ means 3 or more crashes).

| **(a) Control group $G = 0$** | | | | | | | **(b) Treated group $G = 1$** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $Y_{t+1}$ | | | | | | $Y_{t+1}$ | | | | | |
| | | **0** | **1** | **2** | **3+** | **Total** | | **0** | **1** | **2** | **3+** | **Total** |
| $Y_t$ | 0 | 789 | 238 | 57 | 18 | 1102 | $Y_t$ | 183 | 39 | 7 | 3 | 232 |
| | 1 | 235 | 95 | 40 | 15 | 385 | | 40 | 22 | 5 | 2 | 69 |
| | 2 | 61 | 37 | 11 | 6 | 115 | | 16 | 4 | 0 | 1 | 21 |
| | 3+ | 26 | 21 | 4 | 2 | 53 | | 2 | 6 | 0 | 1 | 9 |
| | Total | 1111 | 391 | 112 | 41 | 1655 | | 241 | 71 | 12 | 7 | 331 |

The difference-in-differences estimate is $\hat{\tau}_{\text{DID}} = 7.144$, and the lagged-outcome adjustment estimates are $\hat{\tau}_{\text{LDV}} = 7.160$ and $\hat{\tau}'_{\text{LDV}} = 7.121$ with coefficients of the lag outcome $\hat{\beta} = 1.002 > 1$ and $\hat{\beta}' = 0.997 < 1$. The relative magnitudes match our theory in Section 3.1. However, these estimates are almost identical because the coefficients of $Y_t$ are extremely close to 1. In this example, even though the empirical distributions of $F_{Y_t}(y \mid G = 1)$ and $F_{Y_t}(y \mid G = 0)$ differ significantly as Figure 2 shows, the analysis is not sensitive to the choice between the difference-in-differences and lagged-dependent-variable adjustment estimates.

## 4.3 A Traffic Safety Intervention on Crashes

Outside the political science literature, the before–after treatment–control design is the state-of-the-art method in traffic safety evaluations (Hauer 1997), where count outcomes are common. Here we provide an example of evaluating the effects of rumble strips on vehicle crashes. Crash counts were collected on $n = 1986$ road segments in Pennsylvania before (year 2008) and after (year 2012) the rumble strips were installed in 331 segments between year 2008 to 2012. The control group consists of 1655 sites matched to the treated sites on covariates including past accident counts, road characteristics, traffic volume. Table 1 presents the crash counts classified by $Y_t$ and $Y_{t+1}$ for control and treatment groups, respectively.

We first examine the dichotomized outcome of whether there has been at least one crash in that site. As noted after Theorem 1, Condition 1 automatically holds for a binary outcome. We can verify that Condition 2(a) holds because the empirical means suggest $\widehat{\text{pr}}(Y_t = 0 \mid G = 1) - \widehat{\text{pr}}(Y_t = 0 \mid G = 0) = 232/331 - 1102/1655 = 0.701 - 0.666 > 0$. Therefore, applying Theorem 1, we predict that $\tilde{\tau}_{\text{DID}} > \tilde{\tau}_{\text{LDV}}$ and $\tilde{\gamma}_{\text{DID}} > \tilde{\gamma}_{\text{LDV}}$. Now we calculate the nonparametric estimate of $\mu_0$ under ignorability to be $\hat{\mu}_{0,\text{LDV}} = \sum_{y=0,1} \hat{E}(Y_{t+1} \mid G = 0, Y_t = y)\widehat{\text{pr}}(Y_t = y \mid G = 1) = 0.324$, and under parallel trends to be $\hat{\mu}_{0,\text{DID}} = 0.294$. Therefore, the empirical estimates suggest $\hat{\tau}_{\text{DID}} > \hat{\tau}_{\text{LDV}}$ and $\hat{\gamma}_{\text{DID}} > \hat{\gamma}_{\text{LDV}}$, which matches the theoretical prediction.

Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25

We then examine the original count outcome in Table 1. The sample means $\hat{E}(Y_{t+1} \mid G = 0, Y_t = y)$ are $0.374, 0.572, 0.670, 0.660$ for $y = 0, 1, 2, 3+$, respectively. Therefore, Condition 1 holds for all $y$. We can also verify that Condition 2(a) holds because the sample probabilities are $\widehat{\text{pr}}(Y_t \le y \mid G = 1) = 0.700, 0.909, 0.973$ and $\widehat{\text{pr}}(Y_t \le y \mid G = 0) = 0.666, 0.898, 0.968$ for $y = 0, 1, 2$, respectively. Therefore, applying Theorem 1, we predict that $\tilde{\tau}_{\text{DID}} > \tilde{\tau}_{\text{LDV}}$ and $\tilde{\gamma}_{\text{DID}} > \tilde{\gamma}_{\text{LDV}}$. Now we calculate the nonparametric estimate of $\mu_0$ under ignorability to be $\hat{\mu}_{0,\text{LDV}} = 0.438$, and under parallel trends to be $\hat{\mu}_{0,\text{DID}} = 0.395$. Therefore, the empirical estimates suggest $\hat{\tau}_{\text{DID}} > \hat{\tau}_{\text{LDV}}$ and $\hat{\gamma}_{\text{DID}} > \hat{\gamma}_{\text{LDV}}$, which matches the theoretical prediction.

In this example, the differences between the $\hat{\gamma}_{\text{DID}}$'s and the $\hat{\gamma}_{\text{LDV}}$'s are not significant at level 0.05.

## 5 Discussion

We established a model-free bracketing relationship between the difference-in-differences and lagged-dependent-variable adjustment estimators in the canonical two-period two-group setting. In practice, we cannot use the data to validate the assumptions that justify these approaches. Therefore, a practical suggestion is to report results from both approaches and ideally to conduct sensitivity analyses allowing for violations of these assumptions.

Several directions are worth investigating. First, in the setting with $K + 1$ time periods, we may consider a model that incorporates both Model (7) and (8): $E(Y_{i,T} \mid X_i, Y_{i,T-1}, G_i) = \alpha_i + \lambda_T + \beta Y_{i,T-1} + \tau G_i + \theta^{\mathrm{T}} X_i$ for $T = t + 1, \ldots, t + K$. However, Nickell (1981) and Angrist and Pischke (2009, Section 5.3) pointed out that identification and estimation under this model require much stronger assumptions. It is of interest to extend the bracketing relationship to this setting. Second, we focused on the average treatment effect on the treated; we can extend the result to other types (e.g. categorical and ordinal) of outcomes for which the averages are less interpretable (Lu, Ding, and Dasgupta 2018).

## Appendix Proofs

Proof of Proposition 1. First, the ordinary least squares fit $\hat{E}(Y_{t+1} \mid G = 0, Y_t = y) = \hat{\alpha} + \hat{\beta} Y_t$ using the control units must satisfy $\hat{\alpha} = \bar{Y}_{0,t+1} - \hat{\beta}\bar{Y}_{0,t}$. Therefore,

$$\hat{\tau}_{\text{LDV}} = \bar{Y}_{1,t+1} - \hat{\mu}_{0,\text{LDV}} = \bar{Y}_{1,t+1} - \hat{\alpha} - \hat{\beta}\bar{Y}_{1,t} = \bar{Y}_{1,t+1} - (\bar{Y}_{0,t+1} - \hat{\beta}\bar{Y}_{0,t}) - \hat{\beta}\bar{Y}_{1,t} = (\bar{Y}_{1,t+1} - \bar{Y}_{0,t+1}) - \hat{\beta}(\bar{Y}_{1,t} - \bar{Y}_{0,t}).$$

Second, the coefficient $\hat{\tau}'_{\text{LDV}}$ in the ordinary least squares fit $\hat{E}(Y_{t+1} \mid G, Y_t) = \hat{\alpha} + \hat{\tau}'_{\text{LDV}} G + \hat{\beta}' Y_t$ using all units equals the difference-in-means of $Y_{i,t+1} - \hat{\alpha} - \hat{\beta}' Y_{it}$ in the treated and control groups. Therefore,

$$\hat{\tau}'_{\text{LDV}} = (\bar{Y}_{1,t+1} - \hat{\alpha} - \hat{\beta}'\bar{Y}_{1,t}) - (\bar{Y}_{0,t+1} - \hat{\alpha} - \hat{\beta}'\bar{Y}_{0,t}) = (\bar{Y}_{1,t+1} - \bar{Y}_{0,t+1}) - \hat{\beta}'(\bar{Y}_{1,t} - \bar{Y}_{0,t}).$$

Proof of Lemma 1. The conclusion follows from the law of total probability. We can write $\tilde{\mu}_{0,\text{LDV}} - \tilde{\mu}_{0,\text{DID}}$ as

$$\begin{aligned}
&-E(Y_{t+1} \mid G = 0) + E(Y_t \mid G = 0) - E(Y_t \mid G = 1) \\
&\quad + \int E(Y_{t+1} \mid G = 0, Y_t = y) F_{Y_t}(\mathrm{d}y \mid G = 1) \\
&= - \int E(Y_{t+1} \mid G = 0, Y_t = y) F_{Y_t}(\mathrm{d}y \mid G = 0) \\
&\quad + \int y F_{Y_t}(\mathrm{d}y \mid G = 0) - \int y F_{Y_t}(\mathrm{d}y \mid G = 1) \\
&\quad + \int E(Y_{t+1} \mid G = 0, Y_t = y) F_{Y_t}(\mathrm{d}y \mid G = 1)
\end{aligned}$$

$$= \int \{E(Y_{t+1} \mid G = 0, Y_t = y) - y\} F_{Y_t}(\mathrm{d}y \mid G = 1)$$
$$- \int \{E(Y_{t+1} \mid G = 0, Y_t = y) - y\} F_{Y_t}(\mathrm{d}y \mid G = 0). \quad \square$$

Proof of Theorem 1. The proof relies on a lemma on stochastic ordering in Shaked and Shanthikumar (2007). Specifically, for two random variables $A$ and $B$, $\mathrm{pr}(A \le x) \ge \mathrm{pr}(B \le x)$ for all $x$ if and only if $E\{u(A)\} \ge E\{u(B)\}$ for all nonincreasing functions $u(\cdot)$.

Under Condition 1, we have $\partial\Delta(y)/\partial y = \partial E(Y_{t+1} \mid G = 0, Y_t = y)/\partial y - 1 < 0$, i.e., $\Delta(y)$ is a nonincreasing function of $y$. Therefore, combining Lemma 1, Condition 2(1) implies $\tilde{\mu}_{0,\mathrm{DID}} \le \tilde{\mu}_{0,\mathrm{LDV}}$, and Condition 2(2) implies $\tilde{\mu}_{0,\mathrm{DID}} \ge \tilde{\mu}_{0,\mathrm{LDV}}$. $\square$

## References


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

Abadie, A., A. Diamond, and J. Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59:495–510.

Abadie, A., and J. Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *The American Economic Review* 93:113–132.

Allison, P. D. 1990. "Change Scores as Dependent Variables in Regression Analysis." *Sociological Methodology* 20:93–114.

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

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

Athey, S., and G. W. Imbens. 2006. "Identification and Inference in Nonlinear Difference in Differences Models." *Econometrica* 74:431–497.

Bechtel, M. M., and J. 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* 55:852–868.

Bertrand, M., E. Duflo, and S. Mullainathan. 2004. "How much should we Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics* 119:249–275.

Callaway, B., and P. H. Sant'Anna. 2019. "Difference-in-Differences with Multiple Time Periods." Preprint, arXiv:1803.09015.

Card, D., and A. B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 82:772–793.

Ding, P., and F. Li. 2019. "Replication Data for: A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment." https://doi.org/10.7910/DVN/2F8CWX, Harvard Dataverse.

Gelman, A. 2007. "Difference-in-Difference Estimators are a Special Case of Lagged Regression." https://andrewgelman.com/2007/02/15/differenceindif/.

Hauer, E. 1997. *Observational Before-After Studies in Road Safety. Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Bingley, UK: Emerald Group Publishing.

Heckman, J. J., H. Ichimura, and P. E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of Economic Studies* 64:605–654.

Keele, L., and W. Minozzi. 2013. "How much is Minnesota like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data." *Political Analysis* 21:193–216.

Keele, L., D. S. Small, J. Y. Hsu, and C. B. Fogarty. 2019. "Patterns of Effects and Sensitivity Analysis for Differences-in-Differences." ArXiv, page 1901-0186.

Lechner, M. 2011. "The Estimation of Causal Effects by Difference-in-Difference Methods." *Foundations and Trends in Econometrics* 4:165–224.

Lu, J., P. Ding, and T. Dasgupta. 2018. "Treatment Effects on Ordinal Outcomes: Causal Estimands and Sharp Bounds." *Journal of Educational and Behavioral Statistics* 43:540–567.

Malesky, E. J., C. V. Nguyen, and A. Tran. 2014. "The Impact of Recentralization on Public Services: A Difference-in-Differences Analysis of the Abolition of Elected Councils in Vietnam." *American Political Science Review* 108:144–168.

Maris, E. 1998. "Covariance Adjustment Versus Gain Scores—Revisited." *Psychological Methods* 3:309–327.

Neyman, J. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles (with discussion). Section 9 (translated). Reprinted ed." *Statistical Science* 5:465–472.

Nickell, S. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49:1417–1426.

O'Neill, S., N. Kreif, R. Grieve, M. Sutton, and J. S. Sekhon. 2016. "Estimating Causal Effects: Considering Three Alternatives to Difference-in-Differences Estimation." *Health Services and Outcomes Research Methodology* 16:1–21.




Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25

Rubin, D. B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688.

Ryan, A. M., J. F. Burgess Jr, and J. B. Dimick. 2015. "Why we should not be Indifferent to Specification Choices for Difference-in-Differences." *Health Services Research* 50:1211–1235.

Shaked, M., and J. G. Shanthikumar. 2007. *Stochastic Orders*. New York: Springer.

Stuart, E. A., H. A. Huskamp, K. Duckworth, J. Simmons, Z. Song, M. E. Chernew, and C. L. Barry. 2014. "Using Propensity Scores in Difference-in-Differences Models to Estimate the Effects of a Policy Change." *Health Services and Outcomes Research Methodology* 14:166–182.

van Breukelen, G. J. 2013. "ANCOVA Versus CHANGE from Baseline in Nonrandomized Studies: The Difference." *Multivariate Behavioral Research* 48:895–922.

Downloaded from https://www.cambridge.org/core. IP address: 73.158.62.196, on 09 Oct 2019 at 16:48:46, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/pan.2019.25