Abstract Title Page
Not included in page count.

Title:

Evaluation of Alternative Difference-in-Differences Methods

Authors and Affiliations:

Bing Yu
The University of Chicago

Abstract Body

Background / Context:

Difference-in-differences (DID) strategies are particularly useful for evaluating policy effects in natural experiments in which, for example, a policy affects some schools and students but not others. However, the standard DID method may produce biased estimation of the policy effect if the confounding effect of concurrent events varies by individual characteristics and if the experimental group and the comparison group differ in such characteristics (Meyer, 1995). Three alternative DID approaches have been proposed in the literature to overcome this problem: (1) linear covariance adjusted DID models (e.g., Barnow, Cain, & Goldberger, 1980; Card & Kruger, 1993; Dynarski, 2003; Fitzpatrick, 2008); (2) propensity score-based DID analyses (Abadie, 2005; Blundell et al, 2004; Cerdá, et al, 2012; Heckman, Ichimura, Smith, & Todd, 1998; Heckman, Ichimura, & Todd, 1997); and (3) nonlinear changes-in-changes (CIC) models (Athey & Imbens, 2006). We propose a fourth alternative DID approach that utilizes prognostic scores (Authors, 2012). Each of these methods invokes a different set of identification assumptions that has implications for their relative performance.

Purpose / Objective / Research Question / Focus of Study:

This paper reviews the existing alternative DID methods and compares their identification assumptions with those of the new prognostic score-based DID strategy. Generating data that approximate typical education accountability data, we evaluate the relative strengths and limitations of the new DID strategy through a series of Monte Carlo simulations.

Significance / Novelty of study:

This study contributes to the literature by comparing alternative DID methods. We hypothesize that, in comparison with other existing DID methods, the new prognostic score-based DID strategy invokes assumptions that are relatively more plausible and is more likely to produce unbiased and efficient estimates of policy effects.

Statistical, Measurement, or Econometric Model:

Notation and causal estimand. Let Y_i denote the outcome of individual i. Let $G_i = 1,0$ denote whether the individual is in the experiment group affected by the policy or in the comparison group unaffected by the policy, respectively. Let $T_i = 1,0$ denote whether the individual was observed in the post-policy year or in the pre-policy year, respectively. Let X_i denote a vector of covariates measuring individual characteristics that are not caused by the policy. Let $Y_{iG1.T0}^{(1)}$ denote the potential outcome that individual *i* would display if in the experimental group and counterfactually having exposure to the policy in the pre-policy year; let $Y_{iG1.T0}^{(0)}$ denote the individual's potential outcome in the pre-policy year in the absence of the policy. Suppose we are interested in estimating the average policy effect for individuals in the experiment group in the pre-policy year (i.e., the treatment effect on the untreated in the experimental group), the causal estimand is $\delta_{G1.T0} = E\left[Y_{G1.T0}^{(1)} - Y_{G1.T0}^{(0)} | G = 1, T = 0\right]$.

Standard DID method. The standard DID estimator is $\{E[Y|G = 1, T = 1] - E[Y|G = 1, T = 0]\} - \{E[Y|G = 0, T = 1] - E[Y|G = 0, T = 0]\}$ and can be obtained through analyzing a linear model $Y_i = \alpha_0 + \alpha_1 T_i + \beta_0 G_i + \beta_1 G_i T_i + e_i$. The average confounding effect of the concurrent events for the experimental group is $b_{G1} = E\left[Y_{G1.T1}^{(1)} | G = 1, T = 1\right] -$

$E\left[Y_{G1.T0}^{(1)}|G=1,T=0\right]$ and that for the comparison group is $b_{G0}=E\left[Y_{G0.T1}^{(1)}|G=0,T=1\right]-E\left[Y_{G0.T0}^{(1)}|G=0,T=0\right]$. The standard DID method removes the confounding effect under the assumption $b_{G1}=b_{G0}$. This assumption will be violated and therefore the DID results will be biased, for example, if the confounding effect varies by individual characteristics and if the experimental group and the comparison group differ in such characteristics.

Linear covariance adjusted DID method. Prior research has typically employed a DID model with linear covariance adjustment for observed pretreatment characteristics, $Y_i=\alpha_0+\alpha_1T_i+\beta_0G_i+\beta_1G_iT_i+\lambda X_i+e_i$. Here β_1 is an unbiased estimate of the causal estimand $\delta_{G1.T0}$ under the key assumption that the confounding effect of concurrent events for the experimental group is the same as that for the comparison group within levels of pretreatment covariates $X=x$, that is, $b_{G1|X}=b_{G0|X}$. Moreover, this method requires correct specification of the functional form of the outcome model and relies heavily on linear extrapolation.

Propensity score-based DID method. Heckman and colleagues (1997, 1998) used propensity score matching to identify the common support in the observed covariates *X* and to equate the distribution of *X* between the experimental group and the comparison group. Abadie (2005) proposed using propensity score-based inverse-probability-of-treatment weighting (IPTW) to equate the distribution of *X* between the two groups. Let $\phi(X)=pr(G=1|X)$ be the propensity score representing the conditional probability that an individual would be assigned to the experimental group given *X*. Propensity score matching and IPTW assume that $Y_{G1.t}^{(1)}$ and $Y_{G0.t}^{(1)}$ are independent of G given the propensity score $\phi(X)$ in year t for $t=0,1$. This assumption requires that $E\left(Y_{G1.T1}^{(1)}|G=1,T=1,X=x\right)=E\left(Y_{G0.T1}^{(1)}|G=0,T=1,X=x\right)$ and $E\left(Y_{G1.T0}^{(1)}|G=1,T=0,X=x\right)=E\left(Y_{G0.T0}^{(1)}|G=0,T=0,X=x\right)$. In contrast, the assumption $b_{G1|X}=b_{G0|X}$ holds even when the average potential outcomes of the experimental group and the comparison group are unequal in a given year.

Nonlinear CIC Adjustment. When the experimental group and the comparison group are different in unobserved pretreatment characteristics *U*, to estimate the treatment effect on the treated, Athey and Imbens (2006) proposed a nonlinear CIC model estimating the entire distribution of the counterfactual outcome for the experimental group based on the observed change in the outcome distribution of the comparison group. Key assumptions of the CIC method include (a) A single index model in year t represented by $h(u,t)$ and common change in production function from $h(u,0)$ to $h(u,1)$ regardless group membership; (b) The production function $h(u,t)$ is strictly increasing in u given *t*; (c) Time invariance in the distribution of *U* within groups; (d) The support for the distribution of *U* in the experimental group is a subset of that in the comparison group.

Prognostic score-based DID method. This new method aims to equate the predicted amount of confounding of concurrent events across the pre-policy experimental group, the post-policy experimental group, the pre-policy comparison group, and the post-policy comparison group within subclasses of units. A pair of prognostic scores (Hansen, 2008) per unit, denoted by ψ_0 and ψ_1, represent the predicted pre-policy outcome and the predicted post-policy outcome under the comparison condition in the absence of policy change. DID analysis within each homogeneous subpopulation defined by ψ_0 and ψ_1 is expected to generate an unbiased estimate of the policy effect of interest under the identification assumption that $b_{G1|\psi_0,\psi_1}=b_{G0|\psi_0,\psi_1}$ where $b_{G1|\psi_0,\psi_1}=E\left[Y_{G1.T1}^{(1)}|G=1,T=1,\psi_0,\psi_1\right]-E\left[Y_{G1.T0}^{(1)}|G=1,T=0,\psi_0,\psi_1\right]$ and

$b_{G0|\psi_0,\psi_1} = E\left[Y_{G0.T1}^{(1)}|G=0,T=1,\psi_0,\psi_1\right] - E\left[Y_{G0.T0}^{(1)}|G=0,T=0,\psi_0,\psi_1\right]$. This assumption implies that (a) $\psi_t = f(x,t)$, for $t = 0,1$ defines the function for the counterfactual outcome under the comparison condition at time *t* regardless of one's actual treatment group membership, and that (b) the support for the observed covariates *X* in the comparison school, denoted by $\mathbb{X}_0$, encompasses the support in the experimental school, denoted by $\mathbb{X}_1$.

Usefulness / Applicability of Method:

Comparing the identification assumptions across the above DID methods, we highlight some important strengths of the prognostic score-based DID method. These include:

(1) Unlike the linear covariance adjusted DID models, the prognostic score-based DID strategy does not assume that, conditional on the observed pretreatment covariates *X*, the average treatment effect for the untreated in the experimental group is the same as that for the entire population should all units have been untreated. This advantage is shared by the propensity score-based DID methods and the nonlinear CIC models.

(2) Unlike linear covariance adjusted DID, prognostic score-based DID does not assume invariant *x*-*y* relationships across time.

(3) Prognostic score-based DID assumes that, if an experimental unit had counterfactually been assigned to the comparison condition in a given time period, the *x*-*y* relationship would have been the same as that of the comparison units with the same observed pretreatment characteristics. In contrast, the nonlinear CIC method requires applying a single production function to the outcomes of both the experimental group and the comparison group in a given time period. The latter seems implausible because a change in the distribution of *U* will likely change the u-y relationship.

(4) Unlike the nonlinear CIC models, the prognostic score-based DID models do not require strict monotonicity. Nor do they require that the outcome contain no measurement error. These advantages are shared by linear covariance adjusted DID and propensity score-based DID.

(5) While the nonlinear CIC models assume time invariance within the experimental group and the comparison group, this is not a requirement for all the other DID methods including prognostic score-based DID.

(6) A major difference between propensity score-based DID and prognostic score-based DID is that the latter does not require equating the pretreatment composition of the experimental group and the comparison group. The same advantage is shared by linear covariance adjusted DID and nonlinear CIC.

(7) Similar to propensity score-based DID, prognostic score-based DID emphasizes and verifies the common support between the experimental group and the comparison group with regard to observed covariates *X*, which effectively avoids unwarranted extrapolation. However, propensity score-based DID may suffer if some pretreatment covariates unrelated to the outcome lead to a shrinkage in the common support. The nonlinear CIC models make a similar assumption with regard to the unobserved covariates *U* that cannot be empirically verified.

(8) Similar to the propensity score-based DID, the prognostic score-based DID greatly reduces the dimensionality of covariates for adjustment, which is a major advantage over the linear covariance adjusted DID.

(9) Both prognostic score-based DID and DID with propensity score-based matching enable researchers to detect heterogeneity in the confounding effects of concurrent events as well as in the policy effect.

We also emphasize some potential limitations of the prognostic score-based DID method:

(1) A unique feature of the nonlinear CIC models is that it does not require explicit specification of the outcome model. All other DID strategies require explicit modeling that involves functional forms. The prognostic score-based DID models are no exception. To alleviate the impact of misspecifying the functional form of the model, researchers may employ various semi-parametric or nonparametric approaches, a topic beyond the scope of this paper.

(2) Both linear covariance-adjusted DID and propensity score-based DID would suffer if the experimental group and the conditional group differ in the distribution of unobserables U and if the amount of confounding of concurrent events is a function of U. The same type of unobserables U, if independent of the observed covariates X, would also bias the prognostic score-based DID estimate of the policy effect. However, if the confounding does not depend on U or if the distribution of U is the same between the experimental group and the comparison group conditioning on X, then omitting U would not introduce bias. The special implication for the prognostic score-based DID method is that the estimated policy effect could possibly be unbiased even when the prognostic score models have low predictive power.

Research Design:

We examine the following research questions in the simulation study:

(1) In the best possible world in which all the assumptions required by the standard DID method hold, how does the prognostic score-based DID result compare with those of other DID methods in terms of bias reduction, precision, and mean square error?

(2) If the pretreatment composition differs either between the experimental group and the comparison group or between the pre-policy and the post-policy years, or both, how does the prognostic score-based DID result compare with those of other DID methods?

(3) If the covariate-outcome relationship conditional on other observed covariates changes over time within each group yet remains the same across the experimental group and the comparison group at a given time under the same policy, how does the prognostic score-based DID result compare with those of other DID methods?

(4) How severe are the consequences when the linear covariance adjusted-DID model, the propensity score model, and the prognostic score model are misspecified in their respective functional forms?

(5) When the confounding effect of concurrent events is a function of U and when the distribution of U differs between the experimental group and the comparison group and, additionally, when the distribution of U either remains the same or changes over time, how sensitive is the prognostic score-based DID result to the omission of U when compared with other DID methods?

Findings / Results:

Results are forthcoming.

Conclusions:

Empirical findings obtained from simulation studies will inform our understanding of the relative performance of the new prognostic score-based DID method in comparison with the existing DID methods under a wide array of scenarios often plausible in educational policy evaluations with accountability data. The results will contribute to the statistics and econometrics literature and will provide practical guidance for applied researchers.

Appendices

Appendix A. References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*. *72*, 1-19.

Athey, S. & Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431-497.

Authors (2012).

Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. *Evaluation Studies*, *5*, 43-59.

Blundell, R., Costa Dias, M., Meghir, C. & Van Reenen, J. (2004), Evaluating the employment impact of a mandatory job search assistance program, *Journal of the European Economics Association*, *2*(4), 596-606.

Card, D. & Kruger, A. (1993). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, *84*, 772-784.

Cerdá, M., Morenoff, J. D., Hansen, B. B., Tessari Hicks, K. J., Duque, L. F., Restrepo, A., & Diez-Roux, A. V. (2012). Reducing violence by transforming neighborhoods: A natural experiment in Medellín, Colombia. *American Journal of Epidemiology*. Advance access DOI: 10.1093/aje/kwr428.

Dynarski, S. (2003). Does aid matter? Measuring the effect of student aid on college attendance and completion. *The American Economic Review, 93*, 279-288.

Fitzpatrick, M. D. (2008). Starting school at four: The effect of universal prekindergarten on children's academic achievement. *The B.E. Journal of Economic Analysis & Policy*, *8*(1) (Advances), Article 46.

Heckman, J. Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*(5), 1017-1098.

Heckman, J. Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies* Special Issue: Evaluation of Training and Other Social Programmes, *64*(4), 605-654.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*(2), 481-488.

Meyer, B. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, *13*, 151-161.