Abstract Title Page

Not included in page count.

Title:

Prognostic Score-Based Difference-in-Differences Strategy for Multilevel Multi-Cohort Data

Authors and Affiliations:

Guanglei Hong
The University of Chicago

Abstract Body

Background / Context:

When using time series accountability data to evaluate system-wide education policies, concurrent changes often pose threats to internal validity. The standard difference-in-differences (DID) method resorts to a non-equivalent comparison group whose average outcome change is due to such confounding. This strategy relies on the strong assumption that the average confounding impact of concurrent events is the same for the comparison group unaffected by the policy and the experimental group affected by the policy. 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 (Meyer, 1995). Prior research has typically employed a DID model with linear covariance adjustment for observed pretreatment characteristics (e.g., Barnow, Cain, & Goldberger, 1980; Card & Kruger, 1993; Dynarski, 2003; Fitzpatrick, 2008). More recently, researchers have attempted to equate the covariate distribution of the comparison group with that of the experimental group through propensity score matching or weighting before conducting DID analyses (Abadie, 2005; Blundell et al, 2004; Cerdá, et al, 2012; Heckman, Ichimura, Smith, & Todd, 1998; Heckman, Ichimura, & Todd, 1997). Another approach is to nonlinearly estimate the distribution of the counterfactual outcome of the experimental group resembling the outcome change in the comparison group (Athey & Imbens, 2006). Each of these strategies invokes a set of strong assumptions that may not hold in a particular application.

Purpose / Objective / Research Question / Focus of Study:

We propose an alternative strategy that extends the Peters-Belson method (Belson, 1956; Peters, 1941) to the DID context. The use of prognostic scores (Hansen, 2008) in the causal inference literature can be viewed as the latest development of the Peters-Belson method. Our strategy involves a pair of prognostic scores per unit representing the predicted pre-policy outcome and the predicted post-policy outcome under the comparison condition in the absence of policy change. The difference between the two prognostic scores is the predicted amount of confounding attributable to concurrent events for each unit. Subsequent difference-in-differences analyses within subclasses defined by this pair of prognostic scores allow for a calibrated adjustment. Our rationale is 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. We show that the average within-subclass DID estimate of the policy effect can be obtained through analyzing a weighted outcome model.

This study provides the theoretical rationale for the prognostic score-based DID strategy, clarifies its identification assumptions, and develops an analytic procedure. The new strategy is then extended to multilevel multi-cohort education accountability data. We illustrate with an evaluation of a policy adopted by the Chicago Public Schools requiring all ninth graders to take algebra. We define the causal estimand and develop statistical models for investigating whether the policy effect was enhanced as its implementation became mature or whether the effect faded out over time as the reform lost its momentum after the initial period.

Setting / Intervention / Program / Practice:

The algebra-for-all policy adopted by the Chicago Public Schools (CPS) in 1997 was intended to eliminate remedial math courses for low-achieving students and thereby improving high school math achievement across the board. However, CPS students experienced a number of important policy changes during those same years. The impact of replacing remedial math with algebra was likely confounded by the impacts of those other concurrent interventions.

Among the 59 neighborhood high schools in Chicago that existed both before and after 1997, 45 schools offered remedial math to low-achieving students prior to 1997 and replaced remedial math with algebra after 1997; 14 schools offered algebra to all 9th graders prior to 1997 and thus were unaffected by the policy. This provides a possibility of using the DID strategy to remove the confounding of concurrent events. Unlike most DID studies in which individuals in the experimental group started to experience the policy at a certain time while those in the comparison group never experienced the policy, in this application, the policy had already been implemented in the comparison schools before it was adopted by the experimental schools.

Significance / Novelty of study:

A major challenge to DID analyses is that the confounding effect of concurrent events may vary by pretreatment covariates that are distributed differently across the experimental group and the comparison group. This study contributes to the literature by developing a new alternative DID strategy that makes use of prognostic scores to adjust for the confounding effect of concurrent events. This new strategy invokes assumptions that are distinct from and often weaker than most of the existing DID methods. In particular, by using a prognostic score-based weighting adjustment, the outcome model is non-parametric in nature and hence is exempt from strong model-based assumptions that make conventional DID analyses prone to bias.

Statistical, Measurement, or Econometric Model:

For simplicity, we start by focusing on the mean difference in the math outcome between a pre-policy cohort and a post-policy cohort, contrasting a hypothetical experimental school with a hypothetical comparison school. We will then extend the results to multiple schools and multi-cohort data.

Notation. Let Y_i denote the math outcome of student i at the end of the 9th grade measured on a continuous scale. Let $G_i = 1$ if the student attended an experimental school affected by the policy; let $G_i = 0$ if the student attended a comparison school unaffected by the policy. Let $T_i = 1$ if the student was enrolled in the 9th grade during the post-policy year and 0 if the student was enrolled in the pre-policy year. Let X_i denote a vector of covariates measuring student characteristics that are not caused by the policy.

Causal estimand. We are interested in estimating the average policy effect for students attending the experimental school in the pre-policy year (i.e., the treatment effect on the untreated in the experimental group), considering the possibility that the policy could have been introduced in an earlier year. Let $Y_{iG1.T0}^{(1)}$ denote the potential outcome that student i would display if attending the experimental school and counterfactually having exposure to the policy in the pre-policy year; let $Y_{iG1.T0}^{(0)}$ denote the student's potential outcome in the pre-policy year in the absence of the policy. Here the superscript indicates policy exposure while the subscript indicates school membership and cohort membership. 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 and bias. 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 standard method will generate a biased estimate of the policy effect if the average confounding effect of the concurrent events differs between the experimental school and the comparison school.

Theoretical rationale for prognostic score-based DID. Across all four *G*-by-*T* groups of students, we attempt to identify a subpopulation of students defined by $X=x$ who would experience the same amount of confounding if attending the comparison school. We have observed $Y_{G0.T1}^{(1)}$ of post-policy students in the comparison school and $Y_{G0.T0}^{(1)}$ of pre-policy students in the comparison school. Prognostic score model specifications will allow us to predict this pair of potential outcomes for all students. We define $\psi_0^{G0}(X)$ and $\psi_1^{G0}(X)$ as the predicted pre-policy and post-policy outcomes respectively of a student if assigned to the comparison school, which are henceforth denoted by ψ_0 and ψ_1, respectively.Within each homogeneous subpopulation defined by ψ_0 and ψ_1, DID analysis is expected to generate an unbiased estimate of the policy effect of interest.

Identification assumption. Suppose that ψ_0 and ψ_1 are based on true models for $Y_{G0.T0}^{(1)}$ and $Y_{G0.T1}^{(1)}$ respectively. Then we have that $Y_{G0.T0}^{(1)} \perp X|\psi_0,\psi_1$ and $Y_{G0.T1}^{(1)} \perp X|\psi_0,\psi_1$. Our key identification assumption is that

$$\begin{aligned} &E\left[Y_{G1.T1}^{(1)}|G=1,T=1,\psi_0,\psi_1\right]-E\left[Y_{G1.T0}^{(1)}|G=1,T=0,\psi_0,\psi_1\right] \\ &=E\left[Y_{G0.T1}^{(1)}|G=0,T=1,\psi_0,\psi_1\right]-E\left[Y_{G0.T0}^{(1)}|G=0,T=0,\psi_0,\psi_1\right]. \end{aligned}$$

The above 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; (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 $\mathbb{X}_1$. A proof in Appendix B1 shows that, conditioning on ψ_0 and ψ_1, the policy effect $\delta_{G1.T0}$ can be estimated without bias under these assumptions.

Analytic procedure. We specify a prognostic score model for the comparison school students in the pre-policy year and a second prognostic score model for the comparison school students in the post-policy year. We then apply these two models to all students in all four *G*-by-*T* combinations and predict a pair of prognostic scores ψ_0 and ψ_1 for every student. To conduct DID within cells jointly defined by ψ_0 and ψ_1, we may divide the sample into three strata on the basis of ψ_0 and then subdivide each stratum into three on the basis of ψ_1. We then conduct a standard DID analysis within each of the nine cells and pool the results to obtain an estimate of the policy effect. This procedure allows the DID estimate to differ across different levels of ψ_0 and ψ_1. Let D_s for $s=1,\dots,9$ denote the nine cells. Through analyzing the model $Y_i=\alpha_1 T_i+\beta_0 G_i+\beta_1 G_i T_i+\sum_{s=1}^{9}\lambda_{si}D_{si}+e_i$, we obtain β_1 as an estimate of the average policy effect. This is equivalent to analyzing a weighted model $Y_i=\alpha_0+\alpha_1 T_i+\beta_0 G_i+\beta_1 G_i T_i+e_i$. Appendix B2 shows the weights to be computed for estimating $\delta_{G1.T0}$. The weighted model is relatively convenient to use in multilevel multi-cohort analysis.

Extension to multilevel multi-cohort data. In multilevel data, a student's potential outcome is a function of student-level pretreatment covariates *X* and school-level pretreatment covariates *W*. One may specify a pair of two-level prognostic score models with students at level 1 and schools at level 2. In theory, a student might have multiple prognostic scores depending on which comparison school the student might have counterfactually attended. We define the prognostic scores as a student's predicted outcome of attending a typical comparison school, which can be viewed as the average of the school-specific prognostic scores for the student. In accountability systems in which repeated assessments of student academic achievement have been equated vertically, one may model the growth trajectories of students as well. Suppose that the data include three pre-policy cohorts denoted by $t = -2, -1, 0$ and three post-policy cohorts denoted by $t' = 1, 2, 3$. The average first-year policy effect on the math learning of pre-policy students attending experimental schools is defined as $\sum_{t=-2}^{0} E\left(Y_{11}^{(1)} - Y_{1t}^{(0)} \right) pr(t|G = 1)$. Here $pr(t|G = 1)$ is the proportion of pre-policy students in experimental schools entering the ninth grade in year *t*. To investigate whether the policy effect may depend on the maturity of implementation, we may combine the results of multiple pair-wise DID analyses. Each DID analysis contrasts one pre-policy cohort in year t with one post-policy cohort in year t' and is based on the corresponding pair of prognostic scores ψ_t and $\psi_{t'}$. Let $I(t' - t)$ for $t' = 1, 2, 3$ be the indicator for the subset of data used in the DID analysis for estimating the policy effect after t' years of implementation. Let $Z = 1$ denote the post-policy years and 0 for the pre-policy years. A weighted outcome model for student *i* attending school *j* will be

$$Y_{ij} = \sum_{t'=1}^{3} I_{ij}(t' - t)\left(\alpha_{0t'} + \alpha_{1t'}Z_{ij} + \beta_{0t'}G_{ij} + \beta_{1t'}G_{ij}Z_{ij}\right) + u_j + e_{ij}.$$

Here β_{11}, β_{12}, and β_{13} estimate the policy effects after one year, two years, and three years of implementation, respectively.

Usefulness / Applicability of Method:

The theoretical results and the analytic procedure presented above apply to a continuous outcome such as student achievement data as well as a binary outcome such as whether a student eventually graduates from high school. Various semi-parametric and non-parametric strategies can be employed in specifying the prognostic score models. Issues related to model misspecifications are beyond the scope of the current paper. However, by allowing the models for ψ_0 and ψ_1 to be different functions of *X* under the comparison condition, *T* and *G* each take a fixed value in a prognostic score model. Hence there is no need to consider *T*-by-*X* interaction, *G*-by-*X* interaction, *T*-by-*G* interaction, and *T*-by-*G*-by-*X* three-way interactions in any given model. Finally, the prognostic score-basis DID method does not preclude covariance adjustment in the outcome model for further bias removal and precision improvement. We demonstrate the usefulness of this new method through simulations and an application study.

Conclusions:

Past DID applications have relied heavily on model-based assumptions with regard to the temporal trend in the data in the absence of policy change. Applying the prognostic score-based DID strategy to multi-cohort data, we define the causal estimands non-parametrically and therefore do not impose a linear time trend. This new strategy greatly reduces the dimensionality of covariates for adjustment, which is a major advantage over the linear covariance adjusted DID. The stratification procedure enables researchers to detect heterogeneity in the confounding effects of concurrent events as well as in the policy effect. Yet like most other DID methods, this new strategy shows limitations when the experimental group and the conditional group differ in the distribution of an unobservable and when the amount of confounding of concurrent events is a function of the unobservable. Sensitivity analysis may be developed to assess the amount of bias associated with a possible unobservable covariate.

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.

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

Belson, W. A. (1956). A technique for studying the effects of a television broadcast. *Journal of the Royal Statistical Society*, *Series C* (Applied Statistics), *5*(3), 195-202.

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.

Peters, C. C. (1941). A method of matching groups for experiment with no loss of population. *Journal of Educational Research*, *34*(8), 606-612.

Appendix B1.

Here we prove that the DID estimator integrated over the joint distribution of the two prognostic scores ψ_0 and ψ_1 is an unbiased estimate of $\delta_{G1.T0}$ under the identification assumptions (8), (9a), and (9b).

$$\begin{aligned}
&\iint_{\psi_1\psi_0} \left(D_{G1|\psi_1,\psi_0} - D_{G0|\psi_1,\psi_0}\right) f(\psi_0|\psi_1) f(\psi_1) d\psi_0 d\psi_1 \\
&= \iint_{\psi_1\psi_0} (\{E[Y|G = 1, T = 1, \psi_1, \psi_0] - E[Y|G = 1, T = 0, \psi_1, \psi_0]\} \\
&- \{E[Y|G = 0, T = 1, \psi_1, \psi_0] - E[Y|G = 0, T = 0, \psi_1, \psi_0]\}) f(\psi_0|\psi_1) f(\psi_1) d\psi_0 d\psi_1 \\
&= \iint_{\psi_1\psi_0} \left(\left\{E\left[Y^{(1)}_{G1.T1}|G = 1, T = 1, \psi_1, \psi_0\right] - E\left[Y^{(0)}_{G1.T0}|G = 1, T = 0, \psi_1, \psi_0\right]\right\}\right. \\
&\left. - \left\{E\left[Y^{(1)}_{G0.T1}|G = 0, T = 1, \psi_1, \psi_0\right] - E\left[Y^{(1)}_{G0.T0}|G = 0, T = 0, \psi_1, \psi_0\right]\right\}\right) f(\psi_0|\psi_1) f(\psi_1) d\psi_0 d\psi_1 \\
&= \iint_{\psi_1\psi_0} \left\{E\left[Y^{(1)}_{G1.T0}|G = 1, T = 1, \psi_1, \psi_0\right]\right. \\
&\left. - E\left[Y^{(0)}_{G1.T0}|G = 1, T = 0, \psi_1, \psi_0\right]\right\} f(\psi_0|\psi_1) f(\psi_1) d\psi_0 d\psi_1 \\
&+ \iint_{\psi_1\psi_0} \left(\left\{E\left[Y^{(1)}_{G1.T1}|G = 1, T = 1, \psi_1, \psi_0\right] - E\left[Y^{(1)}_{G1.T0}|G = 1, T = 0, \psi_1, \psi_0\right]\right\}\right. \\
&\left. - \left\{E\left[Y^{(1)}_{G0.T1}|G = 0, T = 1, \psi_1, \psi_0\right] - E\left[Y^{(1)}_{G0.T0}|G = 0, T = 0, \psi_1, \psi_0\right]\right\}\right) f(\psi_0|\psi_1) f(\psi_1) d\psi_0 d\psi_1 \\
&= \delta_{G1.T0}.
\end{aligned}$$

Appendix B2.

To obtain an estimate of the policy effect on the untreated pre-policy students in the experimental school $\delta_{G1.T0}$, the DID estimate in each of the nine cells is to be weighted by the cell-specific proportion of pre-policy students in the experimental school.

when $G = 1, T = 1, S = s$,

$$\omega = \frac{pr(G = 1|T = 1)}{pr(G = 1|T = 1, S = s)} \times \frac{pr(G = 1|T = 0, S = s)}{pr(G = 1|T = 0)};$$

when $G = 0, T = 1, S = s$,

$$\omega = \frac{pr(G = 0|T = 1)}{pr(G = 0|T = 1, S = s)} \times \frac{pr(G = 1|T = 0, S = s)}{pr(G = 1|T = 0)};$$

when $G = 1, T = 0, S = s$,

$\omega = 1;$

when $G = 0, T = 0, S = s$,

$$\omega = \frac{pr(G = 0|T = 0)}{pr(G = 0|T = 0, S = s)} \times \frac{pr(G = 1|T = 0, S = s)}{pr(G = 1|T = 0)}.$$