Received: 4 July 2022 | Accepted: 12 September 2022

DOI: 10.1111/biom.13785

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

DISCUSION

Discussion on: Instrumented difference-in-differences, by Ting Ye, Ashkan Ertefaie, James Flory, Sean Hennessy and Dylan S. Small

Karla DiazOrdaz

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, United Kingdom

Correspondence
Karla DiazOrdaz, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, United Kingdom.
Email: karla.diaz-ordaz@ucl.ac.uk

Funding information
Wellcome Trust, Grant/Award Number: Sir Henry Dale Fellowship 218554/Z/19/Z

Abstract

I discuss the assumptions needed for identification of average treatment effects and local average treatment effects in instrumented difference-in-differences (IDID), and the possible trade-offs between assumptions of standard IV and those needed for the new proposal IDID, in one- and two-sample settings. I also discuss the interpretation of the estimands identified under monotonicity. I conclude by suggesting possible extensions to the estimation method, by outlining a strategy to use data-adaptive estimation of the nuisance parameters, based on recent developments.

KEYWORDS
causal inference, causal machine learning, difference in difference, instrumental variables

I congratulate the authors on their work, instrumented difference-in-differences (IDID), which extends difference-in-difference (DID) estimation to situations where there is unobserved confounding, but nevertheless there is a valid instrumental variable (IV) for the exposure trend, in settings with binary exposure and repeated cross-sectional data. After establishing identification of the average treatment effect (ATE) and local average treatment effects (LATE), the authors also provide us with several estimators, including a multiple-robust estimator based on semi-parametric theory and prove that this is consistent asymptotically normal under the usual regularity conditions.

Here, I discuss (i) the assumptions needed for identification and their plausibility, as well as possible trade-offs between standard IV and this new proposal and (ii) extensions to the estimation method based on recent developments.

I follow the same notation as Ye et al. (2022), where D denotes the exposure, Y the outcome, $\mathbf{X}$ the measured covariates, and Z the instrument. Let O denote $(Z, \mathbf{X}, D, Y)$, the observed data. A subscript $t \in \{0, 1\}$ will be used for D and Y to indicate the variable at that time point.

1 | ASSUMPTIONS AND INTERPRETATION OF THE ESTIMANDS

While the authors provide some insights on the interpretation of the necessary assumptions as well as their plausibility, I believe that the readers might benefit from a deeper discussion.

The authors state that the IDID method "relaxes" both the assumptions of standard IV and standard DID (i.e., parallel trends). I believe that it is more accurate to say that IDID replaces some aspects of the standard assumptions

of each by adding supplementary assumptions drawn from the other method.

Let us take standard IV identification assumptions as a starting point. IDID inherits the core IV assumptions, but applied to the trends instead of a single time point, in the following sense. Instrument relevance (Assumption 2a) says that the instrument is associated with the exposure trend, while Assumption 2b, explicitly the exclusion restriction part (ER), states that the instrument Z and the outcome trend $Y_1 - Y_0$ are conditionally independent given $\mathbf{X}$ and $D_1 - D_0$ (as can be seen in the DAG included in Ye et al. 2022). This subtle change allows for the IV to have a direct effect (not through D) on the outcome Y_1, thus violating the standard IV ER assumption. The trade-off is that to satisfy this ER assumption on the trend, the IV Z will now need to satisfy (i) $Y_1^{(11)} - Y_0^{(01)} | \mathbf{X} \sim_d Y_1^{(10)} - Y_0^{(00)} | \mathbf{X}$, and a further assumption, (ii) $Y_1^{(01)} - Y_0^{(01)} | \mathbf{X} \sim_d Y_1^{(00)} - Y_0^{(00)} | \mathbf{X}$, where following Ye et al. (2022), $\sim_d$ denotes having the same distribution, and $Y_t^{(dz)}$ denotes the potential outcome at time t if exposed to d and Z had been externally set to z.

The first part (i) equates to requiring that Z does not modify the treatment effect, as the authors mention. The assumption (ii) above is, however, a "parallel trends" type assumption: it says that the outcome trend in the untreated potential outcomes is the same on the encouraged (those where $z = 1$) and not encouraged groups ($z = 0$), respectively. This is analogous to the standard DID assumption of parallel trends, which requires the untreated potential outcomes among the exposed and the unexposed (with respect to D) to be the same.

Let us continue using standard IV as our template. Recall that the three core IV assumptions (relevance, ER, and unconfoundedness) are not sufficient to point identify a causal effect (Hernán & Robins, 2006). An extra assumption is required. In the case of standard (one-time point) IV estimation, the ATE can be point-identified by requiring that there is no effect modification by Z among the treated and untreated population. Note that this is indeed similar to assumption 2b(i). However, 2b(i) is not sufficient for point identification of the ATE in IDID. Indeed, this "no effect modification by Z" assumption arose in this setting as part of the ER for outcome trends.

The IDID settings require a stronger assumption to point identify the ATE, namely assumption 2c, *"no unmeasured common effect modifier"*. This assumption has been proposed in standard IV settings as an alternative to "no effect modification by Z"(Cui & Tchetgen, 2021). It is this alternative assumption (2c) that can be replaced by the monotonicity assumption, $D_t^{(1)} \geq D_t^{(0)}$ with probability 1, where $D_t^{(z)}$ denotes the potential exposure for time $t \in \{0, 1\}$, leading to identification of the LATE. Note that unlike relevance (which needs to hold for the exposure trend), the monotonicity assumption needs to hold at both time points, and therefore, it could be considered a stronger set of assumptions than those required for standard IV.

1.1 | Interpretation of the LATE estimand

The identified LATE can be interpreted as the causal effect in the "compliers" stratum, $D_t^{(1)} - D_t^{(0)} = 1$, that is, those who receive $D_t = 1$ when $Z = 1$ but not otherwise at both time points, only when the IV is causally related to the exposure (Swanson & Hernán, 2018). Establishing whether the relationship between Z and the exposure trend is causal would typically vary from application to application. Thus, the interpretation of the LATE will depend on the type of "encouragement" instrument used when applying IDID.

Even with a causal IV, LATE is controversial in clinical and epidemiological applications, as the compliers stratum always remains unobserved.

1.2 | Extra assumptions for the two-sample estimator

Following Ye et al. (2022), for $C \in \{D, Y\}$, let $\mu_C(t, z, \mathbf{x}) = E[C|T = t, Z = z, \mathbf{X} = \mathbf{x}]$ and $\delta_C(\mathbf{X}) = \mu_C(1, 1, \mathbf{X}) - \mu_C(0, 1, \mathbf{X}) - \mu_C(1, 0, \mathbf{X}) + \mu_C(0, 0, \mathbf{X})$, and let $\mu_C(t, z)$ and δ_C denote the analogous quantities without observed covariates.

Suppose that we have two cross-sectional samples, where we have only measured either the exposure or the outcome, correspondingly denoted by b and a (for before and after). Just like in standard two-sample IV, we can use (Z_b, D_b, X_b) to estimate the relationship between the exposure and the instrument, $D \sim Z$ (or in this case the exposure trend), and use the second sample (Z_a, X_a, Y_a) to estimate the relationship between the outcome and the instrument, $Y \sim Z$ (or in our settings, the outcome trend). Now, the ATE is identified by a two-sample Wald estimand $\beta_0 = \frac{\delta_{Y_a}}{\delta_{D_b}}$, so long as $E(Y_a|T_a, Z_a) = E(Y_b|T_b, Z_b)$, and $E(D_a|T_a, Z_a) = E(D_b|T_b, Z_b)$.

Such "structural stability" assumptions rule out covariate shifts across the two samples for the outcome and the exposure in the strata defined by Z, and therefore seem implausible, especially in situations where we seek to apply them, where we do not have access to the same individuals in the two time periods. It would be of interest to explore relaxing this, and allow covariate shifts for the two

time periods, perhaps by adapting techniques developed by Nie et al. (2019).

Moreover, using two-sample Wald estimand in conjunction with monotonicity has implications for what the estimand corresponds to. By analogy to the standard IV (Zhao et al., 2019), we can see that for identification of the LATE, we also need to assume "structural invariance" for the compliers class at time t, that is, $P(D_t^{(1)} - D_t^{(0)} = 1|T_a, Z_a) = P(D_t^{(1)} - D_t^{(0)} = 1|T_b, Z_b)$. Without this assumption, in general the estimand corresponding to the two sample Wald estimator will be a scaling of the LATE in the outcome sample a, β^a_{late}, which is the one we are really interested in. Just as in standard IV, the scaling will be the ratio of the trend in proportions of compliers in the two samples $\frac{P(D_1^{(1)}-D_1^{(0)}=1|T_a,Z_a)-P(D_0^{(1)}-D_0^{(0)}=1|T_a,Z_a))}{P(D_1^{(1)}-D_1^{(0)}=1|T_b,Z_b)-P(D_0^{(1)}-D_0^{(0)}=1|T_b,Z_b))}$. For this to have the same sign as β^a_{late} we also need to assume that this scaling factor is positive.

Finally, assumption 2d, that the CATE is constant in time, also seems implausible in this two-sample cross-sectional setting, even when the study period spans only a short time, as it is more likely that the two samples correspond to slightly different populations. This assumption seems more plausible in one-sample, longitudinal settings, where the same individuals are followed up in time.

2 | DATA-ADAPTIVE ESTIMATION

After establishing identification, the authors propose several estimators. As well as the Wald estimator, analogous to the Wald estimator in standard IV, the authors derive several estimators that target the estimand ψ_0 resulting from a projection of the true CATE $\beta_0(v)$ function onto a parametric working model $\beta(v;\psi)$.

Here, I primarily discuss the so-called "multiply robust estimator" ψ_{mr}. This is an estimating equations estimator, based on the efficient influence function, EIF, $\varphi(O, \psi_0)$ of the (projection) estimand ψ_0.

As in Ye et al. (2022), for $C \in \{D, Y\}$, denote by $\pi(t,z,\mathbf{x}) = P(T=t, Z=z|\mathbf{X}=\mathbf{x})$, $m_{CZ}(\mathbf{x}) = \mu_C(0,1,\mathbf{x}) - \mu_C(0,0,\mathbf{x})$, $m_{CT}(\mathbf{x}) = \mu_C(1,0,\mathbf{x}) - \mu_C(0,0,\mathbf{x})$, $\mathbf{\Delta}_C = (\mu_C(0,0,\mathbf{x}), m_{CZ}(\mathbf{x}), m_{CT}(\mathbf{x}))$ and $\delta = \frac{\delta_Y}{\delta_D}$.

The authors prove that the estimator $\widehat{\psi}_{mr}$ is multiple robust, and is consistent and asymptotically normal (CAN), under an appropriate Donsker condition (Assumption 3) and either: (i) models for $\delta(\mathbf{x})$, $\mathbf{\Delta}_D(\mathbf{x})$ and $\mathbf{\Delta}_Y(\mathbf{x})$ are correct; or (ii) models for $\pi(t,z,\mathbf{x})$ and $\delta_D(\mathbf{x})$ are correct; or (iii) models for $\pi(t,z,\mathbf{x})$ and $\delta(\mathbf{x})$ are correct. While the multiple-robust property means that not all the nuisance models have to be correctly specified, in practice, most parametric models are probably wrong, so a multiple robust property is of limited practical utility.

Nevertheless, recent advances in semiparametric efficiency theory have shown that EIF-based estimators can converge at fast parametric rates to the true ψ_0 and thus be asymptotically normal, even when the nuisance functionals are estimated non-parametrically at slower rates, for example, via flexible data-adaptive (machine learning) methods.

Since ψ_{mr} is an EIF-based estimating equation estimator, it is suitable for using data-adaptive methods for nuisance parameter estimation. As discussed by Ye et al. (2022), under empirical process conditions, for example, Donsker class assumptions (which can be avoided via sample splitting, see below), the error term is (to first-order approximations) the product of the errors of the nuisance models (Theorem 2 of Ye et al. 2022). This allows us to use flexible, machine learning plug-in estimators for the nuisance functionals, which typically converge at slower rates. Then, as long as each data-adaptive estimator converges to their respective truth (denoted by a subscript 0) at a sufficiently fast rate such that the condition of Theorem 2 holds, that is,

$$\left\|\widehat{\delta} - \delta_0\right\|_2 \left(\left\|\widehat{\pi} - \pi_0\right\|_2 + \left\|\widehat{\delta}_D - \delta_{D0}\right\|_2\right) + \left\|\widehat{\pi} - \pi_0\right\|_2 \left(\left\|\widehat{\mathbf{\Delta}}_Y - \mathbf{\Delta}_{Y0}\right\|_2 + \left\|\widehat{\mathbf{\Delta}}_D - \mathbf{\Delta}_{D0}\right\|_2\right) = o_p\left(n^{-1/2}\right),$$

then the estimator that results from plugging in these data-adaptive nuisance estimators is CAN and Equation (5) of Ye et al. (2022) holds. The variance can be obtained based on the variance of the EIF φ.

I remark that, in general, using machine learning plug-in nuisance estimators on estimators based on inverse probability weighting or "outcome regression" leads to biased estimators, because of the slower convergence rates. Moreover, constructing confidence intervals with valid coverage is difficult. It is important to note that generic nonparametric bootstrap arguments are no longer justified in conjunction with data-adaptive plug-in estimators for nuisance parameters (Bickel et al., 1997; Coyle & van der Laan, 2018).

Finally, I would like to discuss the Donsker condition on the class of functions that contains the estimated EIF. To understand why this is often required, we need to take a step back and briefly discuss the error term between a typical plug-in estimator $\psi(\widehat{P}_n)$, an estimator that replaces P_0 with $\widehat{P}_n$, where the sub-index n denotes the sample size of

the data, and the value of the estimand ψ at P_0, the true data distribution. This is characterized by how ψ changes when the data distribution changes from P_0 to a different distribution $\widetilde{P}$ in a small neighborhood. This change is described by the so-called von Misses expansion, a functional-version of the Taylor expansion, with the EIF φ playing the role of the usual derivative. Using this expansion, the error term of many plug-in estimators, can be decomposed into

$$\psi(\widehat{P}_n) - \psi(P_0) = \frac{1}{n}\sum_{i=1}^{n}\varphi(O_i, P_0) - \frac{1}{n}\sum_{i=1}^{n}\varphi(O_i, \widehat{P}_n)$$
$$+(P_n - P_0)\{\varphi(O, \widehat{P}_n) - \varphi(O, P_0)\} + R_2,$$

where R_2 is a second-order term, and P_n denotes the empirical distribution function. The first term is well understood and converges to a normal, mean zero variable. The second term is known as the drift or plug-in bias term. This term is zero by construction in estimating equation estimators (see, e.g., Hines et al. 2022). The third term is known as the empirical process term.

Donsker conditions are typically required to control the asymptotic behavior of the empirical process term. This assumption can be relaxed by adopting sample splitting, or cross-fitting, as done in the de-biased machine learning and cross-validated TMLE literature (Chernozhukov et al., 2018; Zheng & van der Laan, 2011). While cross-fitting can be used in conjunction with parametric nuisance models to avoid assuming Donsker conditions, the use of sample splitting or cross-fitting is preferable to Donsker conditions when using machine learning nuisance parameter estimation, as certain data-adaptive methods (e.g., random forests) may give rise to plug-in influence function based estimators which do not satisfy the Donsker condition (Chernozhukov et al., 2018).

ACKNOWLEDGMENTS

DiazOrdaz thanks the co-editor for the invitation to discuss this paper. DiazOrdaz is funded by a Royal Society Wellcome Trust Sir Henry Dale Fellowship 218554/Z/19/Z.

ORCID

Karla DiazOrdaz https://orcid.org/0000-0003-3155-1561

REFERENCES

Bickel, P.J., Götze, F. & van Zwet, W.R. (1997) Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica*, 7(1), 1–31.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.

Coyle, J. & van der Laan, M. J. (2018) *Targeted Bootstrap*. Cham: Springer International Publishing, pp. 523–539.

Cui, Y. & Tchetgen, E. T. (2021) A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *Journal of the American Statistical Association*, 116(533), 162–173. PMID: 33994604.

Hernán, M. A. & Robins, J. M. (2006) Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4), 360–372.

Hines, O., Dukes, O., Diaz-Ordaz, K. & Vansteelandt, S. (2022) Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76, 1–13.

Nie, X., Lu, C. & Wager, S. (2019) Nonparametric heterogeneous treatment effect estimation in repeated cross sectional designs. *arXiv:1905.11622*.

Swanson, S. A. & Hernán, M. A. (2018) The challenging interpretation of instrumental variable estimates under monotonicity. *International Journal of Epidemiology*, 47(4), 1289–1297.

Ye, T., Ertefaie, A., Flory, J., Hennessy, S. & Small, D. S. (2022) Instrumented difference-in-differences. *Biometrics*.

Zhao, Q., Wang, J., Spiller, W., Bowden, J. & Small, D. S. (2019) Two-sample instrumental variable analyses using heterogeneous samples. *Statistical Science*, 34(2), 317–333.

Zheng, W. & van der Laan, M. J. (2011) Cross-validated targeted minimum-loss-based estimation. *Targeted learning*. New York, NY: Springer, pp. 459–474.

How to cite this article: DiazOrdaz, K. (2022) Discussion on: Instrumented difference-in-differences, by Ting Ye, Ashkan Ertefaie, James Flory, Sean Hennessy and Dylan S. Small. *Biometrics*, 1–4. https://doi.org/10.1111/biom.13785