Inverse Probability Weighting Difference-in-Differences (IPWDID)

Yuqin Wei **yuqin.wei@komodohealth.com**
Matthew Epland, Ph.D. **matthew.epland@komodohealth.com**
Jingyuan (Hannah) Liu **hannah.liu@komodohealth.com**
Komodo Health
90 5th Avenue, 5th Floor
New York, NY 10011, USA

Abstract

In this American Causal Inference Conference (ACIC) 2022 challenge submission, the canonical difference-in-differences (DID) estimator has been used with inverse probability weighting (IPW) and strong simplifying assumptions to produce a benchmark model of the sample average treatment effect on the treated (SATT). Despite the restrictive assumptions and simple model, satisfactory performance in both point estimate and confidence intervals was observed, ranking in the top half of the competition.

Keywords: Difference-in-Differences, DID, Inverse Probability Weighting, IPW

1. Introduction

The American Causal Inference Conference (ACIC) data challenge provides a unique opportunity for practitioners in academia and industry to test the latest causal inference techniques on a shared problem and dataset. As relative newcomers to the field, we applied the canonical difference-in-differences (DID) method with inverse probability weighting (IPW) to gain experience with causal inference problems. We hope that the IPWDID results presented here can serve as a benchmark by which novel models proposed in recent years can be compared against a well-established model.

2. Methodology and Motivation

2.1 Problem and Notation

The 2022 ACIC challenge datasets were designed to mirror data from a real large-scale intervention in the U.S. healthcare system aimed at lowering Medicare expenditures. Let $Y_{i,t}(1)$ and $Y_{i,t}(0)$ denote the potential expenditure outcomes for patient i, where the patient receives the treatment or control condition, respectively, at time t. The observed outcome for patient i is given by $Y_{i,t} = z_i\, p_t\, Y_{i,t}(1) + (1 - z_i\, p_t)\, Y_{i,t}(0)$, where z_i indicates membership in the treatment group, and p_t is an indicator for *post*. The average treatment effect on the treated (ATT) is then defined as:

$$\mathrm{ATT}\,(t) = E_{i|Z_i=1}\left(Y_{i,t}(1) - Y_{i,t}(0)\right). \tag{1}$$

Note, in the ACIC challenge the target estimand is the ATT over the observed units, *i.e.* the sample average treatment effect on the treated (SATT). However, we will continue to refer to ATT for simplicity.

2.2 Difference-in-Differences (DID)

Given the longitudinal nature of the data, one potential approach is to compare changes between cohorts before and after an intervention in a difference-in-differences (DID) analysis. When the difference between the intervention group and control group is constant over time in the absence of intervention, known as the parallel trends assumption (PTA), the DID design enables us to estimate causal effects even in the presence of time-invariant unmeasured confounding [1].

A typical DID model involves two time periods, pre- and post-intervention, however the ACIC 2022 challenge includes two years of pre-intervention data, $t = 1, 2$ and two years of post-intervention data, $t = 3, 4$. This provides us with a few options for handling the pre-intervention period, namely using a two-way fixed effect model or limiting the use of pre-intervention data to only one term [2].

For a DID model with multiple pre-intervention time periods, testing for pre-intervention differences in trends, *i.e.* a pretest, is suggested, either through testing pre-intervention regression coefficients or via a visual inspection. However, as shown in [3] conditioning the analysis on the result of a pretest can distort estimation and inference, potentially exacerbating the bias of point estimates and under-coverage of confidence intervals. In addition, it is difficult to automate the pretest for each of the 3400 realizations of the data generation process present in the challenge dataset.

A two-way fixed effect model is commonly used for DID models in cases with panel data like this one. Yet, a two-way fixed effect model may not correctly estimate ATT when there is a strong heterogeneous effect [4], or the effect is not linearly additive [5]. Methods that improve two-way fixed effect regression [2], or improve the estimate when the PTA is violated [6], have been studied in the field of economics. However, these methods are not well understood and are infrequently used in applied research.

We decided to address these issues in the challenge dataset by limiting the observations to year 2 and beyond, thereby assuming the PTA only begins in year 2. This simplification is arbitrary, and causes data from year 1 to be dropped completely, but should help to reduce the bias from potentially violating the PTA, in particular when compared to models assuming the PTA holds over both years 1 and 2. In Section 3.3 we will show a sensitivity analysis using the average cost from years 1 and 2 as the pre-intervention cost to estimate the impact of neglecting year 1.

Using the canonical DID model, the DID estimator of ATT for each year $t = 3, 4$ is then

$$\hat{\tau}_t^{\text{DID}} = \overline{Y}_{1,t} - \overline{Y}_{1,2} - \left(\overline{Y}_{0,t} - \overline{Y}_{0,2}\right), \tag{2}$$

where $Y_{i,t}$ is the observed outcome for patient i in year t, $\overline{Y}_{z,t} = \frac{1}{n_z} \sum_{i|Z_i=z} Y_{i,t}$ is the estimate for Y, and n_z is the sample size of each cohort. In this framework, the PTA can be stated as:

$$E\left(Y_{i,t}(0) - Y_{i,2}(0) \mid Z_i = 1\right) = E\left(Y_{i,t}(0) - Y_{i,2}(0) \mid Z_i = 0\right). \tag{3}$$

2.3 IPWDID

The PTA may also be implausible if the pre-intervention covariates are unbalanced between the intervention and control arms [1]. We restate the PTA conditioning on covariates as:

$$E\left(Y_{i,t}(0) - Y_{i,2}(0) \mid Z_i = 1, X_i\right) = E\left(Y_{i,t}(0) - Y_{i,2}(0) \mid Z_i = 0, X_i\right). \tag{4}$$

It can be shown that a simple two-step strategy, inverse probability weighting (IPW) [7], can adjust for the covariate imbalance. First, the IPW is calculated while targeting the ATT. Second, an IPWDID estimator on the weighted samples is computed. Using the same framework as above, we then have

$$\hat{\tau}_t^{\text{IPWDID}} = \frac{1}{n} \sum_i \frac{Z_i - \hat{\pi}\left(X_i\right)}{1 - \hat{\pi}\left(X_i\right)} \left(Y_{i,t} - Y_{i,2}\right), \tag{5}$$

where $\hat{\pi}(X_i)$ is an estimator of the true propensity score $p(X) = P\left(Z_i = 1 \mid X_i\right)$.

We then construct the final ATT as a weighted average across years $t = 3, 4$ as

$$\hat{\tau}^{\text{IPWDID}} = \frac{n_3\, \hat{\tau_3}^{\text{IPWDID}} + n_4\, \hat{\tau_4}^{\text{IPWDID}}}{n_3 + n_4}, \tag{6}$$

where n_t is the number of records with non-missing values in year t.

For simplicity, we use a logistic regression model over all covariates to estimate the propensity score. As the feature definitions are unknown, we did not assume any causal relationships or preform feature engineering, beyond one-hot encoding the categorical features. Lastly, IPW weights over 5 are truncated to 5 to avoid extreme weights and improve the stability of the estimate.

2.3.1 Alternatives

Here we shall briefly discuss some of the alternatives to the IPWDID model as described above. DID regression can be used by regressing the outcome variable on all covariates, but this requires a strong assumption to be made on the functional form of the outcome model. Methods that analyze the response surface directly, such as Bayesian additive regression trees (BART) [8], may perform better, but are not considered here due to their complexity. Doubly robust (DR) estimators have been proposed [7, 9, 10] which only require either the propensity score model or the outcome model to be correctly specified in order to obtain an unbiased estimator. They are also more complex than IPWDID, and are likewise not considered. Additionally, it has been argued that when both the propensity score model and the outcome model are misspecified, there is no evidence that DR estimators perform better than IPW [11].

Finally, propensity score matching (PSM) can be used in parallel with DID, as proposed in [1]. PSM aims to reduce the bias due to confounding variables by finding matching intervention and control pairs in the data via the propensity score. We performed experiments using PSM which constructed more balanced intervention and control patient cohorts, with promising results for the DID estimand. However, due to the prohibitive computational burden of PSM, in particular when working with the patient-level data, we selected the IPWDID approach for the challenge.

2.4 Bootstrapping Confidence Intervals

Bootstrap sampling with 500 repetitions, per realization, is used to quantify the 90 % confidence intervals of the point estimates. Patients, or practices, are resampled with replacement to compute Hall's Basic confidence interval [12].

Two other approaches were also considered, robust sandwich estimators and hierarchical bootstrapping. Robust sandwich estimators were not selected as they tend to be over-conservative when used together with IPW [13]. Hierarchical bootstrapping [14] aims to maintain the original hierarchical data structure within each bootstrap resample. In this approach, practices are first resampled with replacement, and then patients within selected practices are resampled with replacement. However, each bootstrap resample may not have an equal sample size.

We tried to implement the hierarchical bootstrapping method in our initial challenge submission, but the resulting confidence intervals were questionable. We suspect this could be due to an implementation error, though have yet to fully explain what went wrong. Instead, due to limited time we switched to the simpler non-hierarchical bootstrapping approach described above.

2.5 Addressing Missing Data

It is important to note that all practices and patients do not have data observed across all years. A common approach to tackle the problem of missing data is to assume the patterns of missing data are at random (MAR) and apply multiple imputations (MI). Noting that more than half of the patients have at least one measurement both before and after the intervention, for simplicity and computation efficiency we assume the data are missing completely at random (MCAR) and remove records that only have cost data before or after the intervention. For example, records with year 2 and year 3 (year 4) cost data present are utilized when estimating $\hat{\tau}_3^{\text{IPWDID}}$ ($\hat{\tau}_4^{\text{IPWDID}}$). This results in 30 % of patients being dropped on average in each realization of the data.

2.6 Subgroup Analysis

The IPWDID method targets the population-level ATT and does not primarily estimate heterogeneous treatment effects. The treatment effects within subgroups need to be evaluated separately. Here we simply focus on data points within each subgroup; the propensity score estimated over all samples is used, but the IPW is re-calculated within the subgroup, and the ATT is re-evaluated using the subgroup-specific IPW. To save computation time all subgroup results are evaluated in the same bootstrap iteration as the overall result.

2.7 Patient-Level and Practice-Level Submission

Here we summarize the four submissions made by the IPWDID team over the course of the challenge. Submissions with ***** were submitted after the official challenge deadline.

2.7.1 **ipwdid_1**

ipwdid_1 is the initial submission before the deadline. It only evaluates patient-level data, and uses a hierarchical bootstrapping approach for the confidence intervals. When results

were released at ACIC 2022 the width of the confidence intervals was shown to be extremely large.

2.7.2 **ipwdid_2***

ipwdid_2* is a correction of **ipwdid_1** that maintains the original point estimates, while incorporating a corrected, and more straightforward, patient-level data resampling approach for the confidence intervals.

2.7.3 **ipwdidp_1***

ipwdidp_1* is an extension of **ipwdid_1** to the practice-level data, using the same methods. Patient-level ATT was calculated via a sample size average across practices.

2.7.4 **ipwdidp_2***

ipwdidp_2* is a sensitivity analysis as suggested by the organizers. **ipwdidp_2*** uses the average of year 1 and 2 costs as the baseline cost, but is otherwise identical to **ipwdidp_1***.

3. Results

As all four submissions used similar methods, we will elaborate the results from the **ipwdid_2*** model unless stated otherwise.

3.1 Bias and Root Mean Square Error (RMSE)

The bias and root mean square error (RMSE) for overall SATT in **ipwdid_2*** are 10.0 and 19.4, respectively. Compared to the top-performing model with a bias of 3.9 and RMSE of 11.0, the gap is moderate. While the result is not optimal, it is better than 75 % of other submissions.

When validating the model in the absence of confounders, utilizing the no confounding realizations identified by the challenge organizers post-submission, we observe a small bias of -1.7 and RMSE of 8.6. However, the bias increases with the increment of the confounding strength. The average bias across all scenarios for weak and strong confounding settings are 7.3 and 14.2, respectively. This indicates a weakness of our model when facing strong confounding.

We observed significantly poorer results from models reporting SATT for subgroups, especially small subgroups. For comparison, the bias and RMSE of **ipwdid_2*** for the subgroup are 11.1 and 36.5, respectively, while the top-performing model had a bias of 3.9 and RMSE of 15.8.

3.2 Coverage

All models except **ipwdid_1** yield a coverage probability less than 90 %. **ipwdid_1** returns an average of 92.5 % coverage, which is very close to the nominal confidence interval at 90 %. However, this is likely a coincidence, as supported by the model having a 100 % coverage in the absence of confounders. This could be due to either a coding error or a lack of

understanding of the hierarchical bootstrap strategy. As a result, the confidence interval of **ipwdid_1** is over-conservative and does not merit further interpretation.

In the corrected version **ipwdid_2***, we see that the coverage of overall SATT is 60 %. This is better than or equal to 60 % of other submissions, excluding those with obvious major issues in confidence interval estimation.

Given that the model suffers from moderate to strong bias, it is not surprising to see the coverage is below the nominal coverage, similar to all participants' results. The coverage for the overall model without confounding is 81 %, which is close to the nominal coverage. However, no model with confounders and no subgroup model reaches the same level of coverage, especially for small subgroups and scenarios with large confounding strengths.

3.3 Patient-Level vs. Practice-Level

The **ipwdidp_1*** model on practice-level data has behavior similar to **ipwdid_2*** on patient-level data, with a few noticeable differences. First, the estimates have a larger variance, as can be seen in the overall SATT estimate's relatively small bias of -2.4, and large RMSE of 21.5. Second, for **ipwdidp_1*** we observe a negative bias for scenarios when confounding is partially based on pre-intervention trends, Scenario A, and a positive bias when confounding is not based on pre-trends, Scenario B. In contrast, for **ipwdid_2*** on patient-level data the bias is systematically positive across the board. Third, the bias and RMSE from **ipwdidp_1*** is smaller for subgroup models.

Finally, a significant improvement in coverage is observed. For example, **ipwdidp_1*** recovers the 90 % confidence interval in the absence of confounders. When operating on data variations with weak and strong confounders, the model consistently yields an average coverage of 74 % and 75 % of the confidence interval, respectively. We conclude that the practice-level model performs much better for the interval estimate than the patient-level model. This indicates that our method did not fully utilize the granularity of the patient-level data.

Compared to **ipwdidp_1***, the **ipwdidp_2*** sensitivity analysis used the average of year 1 and year 2 cost as the pre-intervention cost outcomes. We consistently observed a larger positive bias, similar RMSE, and lower coverage for most of the simulation scenarios. Thus, simply adding year 1 data to the pre-intervention period does not improve the IPWDID model; a more advanced treatment of year 1 is required.

4. Discussion

With the goal of calibrating popular and conventional methods for social science and epidemiological research, we have tried a few variations of the DID estimator with IPW. Massive simplifications have been applied, and strong assumptions were made without thorough validation, yet we show that IPWDID methods are quite robust across the different simulated scenarios. IPWDID returns a sufficiently good estimate and reasonable confidence interval for inference, when the confounding is not strong. In addition, the implementation of the method is relatively straightforward in comparison to novel methods. This enables researchers with basic statistical backgrounds to effectively utilize the IPWDID method with a lower risk of implementation errors. Finally, we show the method ranks in the top half of the competition, despite so much simplification and restriction. IPWDID beats

other methods that potentially utilize the data more efficiently, either in constructing the response surface or by being doubly robust.

We acknowledge the following limitations of the IPWDID method as implemented, and outline some potential improvements. First, we made additional simplifying assumptions not required by the IPWDID framework that could be removed with a slightly more complex analysis. For instance, the year 1 cost data and patients with missing data could be incorporated into the model to improve data usage efficiency. Second, the propensity scores could be better evaluated with a model other than logistic regression. Third, in our experiments the practice-level results appear no worse, or even better, than the patients-level results regarding both RMSE and coverage. This somewhat justifies the study design decision to focus on aggregated datasets when using IPWDID, where the data abstraction is sufficient, instead of working on the most comprehensive dataset. However, this also indicates that patient-level variation has not been well captured by our model. A more complex model that better explores the patient-level information might be preferred.

Finally, as shown by the challenge organizers, our method tends to chase noise from small subgroups. We suspect this is mainly because the method mechanically subsets the data and re-evaluates the result within each subset. A higher variance was observed when only information from the subset was used. For the same reason, heterogeneity was studied disjointly by putting subgroup analyses together, instead of tackling it simultaneously in a single model. Methods that more naturally handle heterogeneity in a single model may be preferred. Nevertheless, we would like to emphasize that subsetting data is a typical approach used in applied research. For this particular dataset, we would suggest not using IPWDID on subgroups if effect heterogeneity is the primary study objective.

While our implementation of IPWDID may lack novelty, we are happy to share our results with the community as a benchmark by which to compare the many other excellent submissions. The R code underlying this submission is available at github.com/mepland/acic_causality_challenge_2022. Meanwhile, we are somewhat surprised to see our approach outperform many other methods, including BART, targeted maximum likelihood estimation (TMLE), and DR based submissions. We feel it is important to highlight that a novel method may not work as expected if the underlying assumptions are not met, the method is not well-understood, the method is incorrectly implemented, or if the method is not suitable for the use case. Sufficient support needs to be provided from statisticians to applied researchers in other disciplines before such novel methods are adopted. Otherwise, conventional methods, such as IPWDID, may be a safer, yet still performant, choice.

Acknowledgments

We would like to acknowledge the support of this project from Komodo Health, including the necessary computational resources, as well as thank Mariel Finucane and Dan Thal for organizing this years challenge.

References

[1] A. Abadie, *Semiparametric Difference-in-Differences Estimators*, The Review of Economic Studies **72** (2005) 1–19.

[2] N. Egami and S. Yamauchi, *Using Multiple Pretreatment Periods to Improve Difference-in-Differences and Staggered Adoption Designs*, Political Analysis (2022) 1–18.

[3] J. Roth, *Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends*, American Economic Review: Insights **4** (2022) 305–22.

[4] C. de Chaisemartin and X. D'Haultfœuille, *Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects*, American Economic Review **110** (2020) 2964–96.

[5] K. Imai and I. S. Kim, *On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data*, Political Analysis **29** (2021) 405–415.

[6] A. Rambachan and J. Roth, *An Honest Approach to Parallel Trends*, 2019. `https://jonathandroth.github.io/assets/files/HonestParallelTrends_Main.pdf`.

[7] P. H. Sant'Anna and J. Zhao, *Doubly robust difference-in-differences estimators*, Journal of Econometrics **219** (2020) 101–122.

[8] J. L. Hill, *Bayesian Nonparametric Modeling for Causal Inference*, Journal of Computational and Graphical Statistics **20** (2011) 217–240.

[9] J. D. Y. Kang and J. L. Schafer, *Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data*, Statistical Science **22** (2007) 523–539, `http://www.jstor.org/stable/27645858`.

[10] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian, *Doubly Robust Estimation of Causal Effects*, American Journal of Epidemiology **173** (2011) 761–767.

[11] M. S. Ali, D. Prieto-Alhambra, L. C. Lopes, D. Ramos, N. Bispo, M. Y. Ichihara, J. M. Pescarini, E. Williamson, R. L. Fiaccone, M. L. Barreto, and L. Smeeth, *Propensity Score Methods in Health Technology Assessment: Principles, Extended Applications, and Recent Advances*, Frontiers in Pharmacology **10** (2019).

[12] P. Hall, *Theoretical Comparison of Bootstrap Confidence Intervals*, The Annals of Statistics **16** (1988) 927–953.

[13] P. C. Austin, *Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis*, Statistics in Medicine **35** (2016) 5642–5655.

[14] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.