Revision memo:
"Estimating population average treatment effects from experiments with noncompliance" (DGJCI.2018.0011)

1 Editor's comments

1.1 Definition of SATE

Following R1's guidance, and conforming to naming conventions in the recent literature on causal inference with noncompliance (Yau and Little 2001; Frumento et al. 2012), we now call this quantity the sample Complier Average Causal Effect (CACE).

1.2 DAG

1.3 No defier assumption

1.4 Assumption 7

1.5 Conceptualizing PATT-C

1.6 Theorem 1

1.7 Prediction threshold

1.8 Same W

1.9 Formal definition of PATT

In Section 4, we formally define the PATT estimator (Eq. (3))

2 Reviewer 1 (R1)'s comments

2.1 Definition of SATE

R1 points out that what we have referred to as SATE is inconsistent with how we'd usually define SATE. The quantity we're interested in estimating is the sample local average treatment effect among the compliers. This quantity is commonly referred to as the LATE in the econometrics literature (e.g., Angrist, Imbens, and Rubin 1996) (AIR). Freedman (2006) shows that the instrumental variables estimator for the LATE proposed by AIR is equivalent to scaling the ITT effect by the proportion of treated compliers in the RCT.

Following the guidance of R1, we refer to this quantity in the revised manuscript as the Complier Average Causal Effect (CACE) and define it in Eq. (4). Referring to this quantity as the CACE also conforms to recent naming conventions in the literature concerning program evaluation in the presence of noncompliance (Yau and Little 2001; Frumento et al. 2012).

2.2 Estimation of PATT

R1 asks for clarification on the assumptions needed to identify the unadjusted population (PATT) estimator and for more discussion of the estimator's performance in the simulations.

Our PATT estimator should be viewed as the unadjusted analogue to the compliance-adjusted population estimator, PATT-C. When estimating PATT, we are estimating the response curve for all RCT members, conditional on their covariates and actual treatment received. We then use the response model to estimate the outcomes of population members who received treatment, given their covariates.

Previous approaches rely on reweighting methods to estimate the ITT effect as a function of covariates in the RCT first and then project to the population. As R1 points out, these methods assume exchangeability of potential outcomes between the covariate-adjusted treated and controls in the RCT before projecting. Our approach for estimating PATT-C differs because we only want the response curve for RCT compliers and we cannot identify who among the RCT controls is a "complier"; i.e., RCT controls who would have complied had they been assigned treatment. Complier treated and complier controls aren't exchangeable by design, since we need to assume we know the compliance model. As R1 notes, we need to assume in estimating either PATT or PATT-C that the response surface is the same for compliers in the RCT and population members who received treatment. If the strong ignorability assumptions do not hold, then the potential outcomes Y_{i10} and Y_{i11} for population members who received treatment cannot be estimated using the response model.

In Section 4 of the revised manuscript, we formally define the PATT estimator (Eq. (3)) and also clarify the assumptions required for its estimation. In Section 4.2 of the revised manuscript, we explain that PATT is performing comparatively worse because it isn't adjusted for compliance and consequently performs poorly when the population compliance rate is relatively low.

2.3 Small points

Alternative complier-adjusted population estimator

R1 is correct: a more appropriate thought experiment to our proposed PATT-C is an estimator that reweights the ITT effect to the whole population and then divides by the proportion of treated compliers in the population. The problem is that we don't know the compliance rate in the population. Our approach of explicitly modeling compliance allows us to identify the likely compliers in the RCT control group, whose outcomes we model in S.3 of the estimation procedure. We revised the Introduction to include the more appropriate thought experiment and discuss the rationale for our approach.

Exclusion criteria and strong ignorability

The strong ignorability assumptions would be violated if the known exclusion criteria are correlated with unobserved factors that also determine potential outcomes. High exclusion would therefore increase the likelihood that there are unobserved differences between the RCT and target population.

We have revised the manuscript to make this point more straightforward and to provide an example from our RCT application. We also note that bias resulting from violations of

ignorability assumptions would be detected in the placebo tests. Our placebo test results show no bias in estimates of the complier-average population effects.

RCT policy motivation

We've added a line in the introduction to emphasize the policy motivation for the health insurance RCT, as suggested by R1.

HH-level treatment

We have clarified when introducing the RCT data in Section 5 that the response and complier models include household size as a covariate because lottery selection was random conditional on household size. R1 is correct that because treatment occurred at the household level, we should define the population in the empirical application as individuals grouped within households. Accordingly, in the revised manuscript we cluster standard errors at the household level.

Typos

We fixed the two typos helpfully pointed out by R1.

3 Other revisions

3.1 Simulation

We found that the confounding variable W_i^4 was missing in the response equation in both the code and the write-up of the simulation design in Section 4.

In the current manuscript, we have included W_i^4 and constant c_3 (set to 1) in the response equation in the code and write-up and re-ran the simulation. The new compliance heatmaps (Figures 2 and A1) share a common gradient scale to ensure comparability between the PATT-C, PATT, and CACE estimators. The heatmaps show the PATT-C yields lower estimation error than its unadjusted counterpart when the population compliance rate is relatively low (i.e., 80% or less). The new bar plot comparing the RMSE of estimators when varying the population compliance rate (Figure 3) shows, as expected, the estimation error of the estimators is inversely related to the population compliance rate. PATT-C outperforms its unadjusted counterpart when the compliance rate is relatively low (i.e., 80% and lower).

3.2 Changes in empirical application analysis

For consistency with the empirical analysis in Finkelstein et al. (2012), we make the following changes in the analysis of the empirical application in Section 5:

- Divide the health care use responses in the NHIS survey (12 month look-back) by two in order to make them comparable with the OHIE survey outcomes (6 month look-back). Exclude binary response ("Any ER visit") to ensure comparability.

- Report family-wise p-value (along with the per-comparison p-values) to adjust for multiple comparisons in Tables A3 and A4
- Include as covariates in the complier and response models indicator variables for survey wave (and their interactions with household size indicators) because the proportion of treated participants varies across the response survey waves.
- Include as pretreatment demographic controls:
  - Bins of number of children in household
  - Race: Asian; American Indian or Alaska Native; Other
  - Ever diagnosed with Ephysema or Chronic Bronchitis (COPD)
  - Currently living with partner or spouse
  - Currently employed or self-employed
- Weight descriptive statistics and treatment effect using survey weights. The OHIE survey weights account for the probability of being sampled and survey procedures. The NHIS survey weights adjust for the probability of selection and non-response.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91, no. 434 (June): 444–455.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, and Heidi Allen. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year." *The Quarterly Journal of Economics* 127 (3): 1057.

Freedman, David A. 2006. "Statistical Models for Causation What Inferential Leverage Do They Provide?" *Evaluation Review* 30 (6): 691–713.

Frumento, Paolo, Fabrizia Mealli, Barbara Pacini, and Donald B Rubin. 2012. "Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data." *Journal of the American Statistical Association* 107 (498): 450–466.

Yau, Linda HY, and Roderick J Little. 2001. "Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed." *Journal of the American Statistical Association* 96 (456): 1232–1244.