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

1 Reviewer: 1 (R1)

1.1 Expand set of simulations

Following R1's excellent suggestion, we expand the set of simulations to follow more of an experimental design, by varying the parameters that determine the degree of confounding with sample selection, confounding with treatment assignment, and confounding with compliance. In addition, we made the following changes to the simulations:

- Compare the PATT and PATT-C estimators against the SATE, which is just the ITT effect scaled by the compliance rate, since the SATE is a more appropriate benchmark
- Use gradient boosting to predict compliance rate in the RCT and to predict the potential outcomes of observed and predicted RCT compliers
- Vary the e_k, $k = \{1...6\}$ parameters along a grid of five standard normal deviates

In the new simulations, we find that the estimation error of PATT-C is invariant to increases in the compliance rate. In comparison, SATE performs worse when the compliance rate is low, and is also considerably more variable than both of the population estimators due to the fact that SATE is unable to account for differences in pretreatment covariates between the RCT sample and target population.

1.2 Presentation

1.2.1 Assumptions

Following R1's suggestion, we have moved Assumptions 1 – 5 from the Appendix into the main text.

1.2.2 Consistency for index i

R1 points out that the subject-level index i is inconsistently used in the original manuscript. The revised manuscript ensures that all subject-level quantities are indexed with i.

1.2.3 Define SATT

In the revised manuscript, we compare the PATT and PATT-C estimators against the SATE, which is just the ITT effect scaled by the compliance rate, since the SATE is a more common estimator.

1.2.4 NHIS sample design

R1 points out that the complex sample design of the NHIS was ignored in the application. Footnote 8 notes this as a limitation and also references a National Center for Health Statistics report on the NHIS sampling design. The revised manuscript also properly cites NHIS as a data source.

1.2.5 Interpretation of results

R1 suggests providing more context for interpreting whether the subgroup analyses presented in Section 5.4 are reasonable. We expand Section 5.4 in the revised manuscript to compare our sample subgroup estimates with those published in the online appendices of (Taubman et al. 2014) and Kowalski (2016). These studies do not perform subgroup analyses for the broader population. Similar to our sample estimates, these studies find considerable treatment effect heterogeneity in terms of gender, age, smoking status, and pre-lottery welfare participation.

In addition, R1 points out that in the Discussion of the original manuscript, the direction of the PATT estimates on ER and outpatient visits are incorrectly described. The revision removes this description as it is not relevant to the discussion on subgroup analyses.

2 Reviewer: 2 (R2)

2.1 Identifying assumptions

2.1.1 Conditioning on D_i

R2 suggests framing the ignorability Assumptions 3 and 4 on D_i rather than T_i and C_i to be more consistent with the framing of Hartman et al. (2015) and to ensure that treatment means the same thing in both the experiment and the population. The reason we decided to frame the ignorability assumptions conditional on T_i and C_i is to distinguish between noncompliers who should have received treatment (i.e., individuals with $T_i = 1$ and $D_i = 0$) from noncompliers assigned to control; i.e., individuals with $T_i = 0$ and $D_i = 0$. Conditioning on T_i and C_i is important for deriving the estimator for $\tau_{\text{PATT-C}}$ (Eq. 2).

We have made more clear our motivation for conditioning on T_i and C_i when introducing notation in Section 2.1 in the revised manuscript.

2.1.2 Y subscripts

R2 is correct that it would be more appropriate to define Y based on d rather than t. As described in the comment 2.3 below, the original manuscript confused treatment eligibility

T_i with treatment received D_i. Our estimation procedure relies on fitting a response curve to D_i in the RCT, since we cannot actually observe T_i in the population. We have revised the manuscript accordingly.

2.1.3 W_i

With regards to our use of the same W_i across all assumptions, here we are implicitly assuming that the covariates that determine sample selection also determine population treatment assignment and complier status. This is useful not only for notational ease but also reflects our modeling assumption that W_i is the same. In the revised manuscript, we discuss this choice in Sections 2.1 and 3.1.

Moreover, in Section 3.2 of the revised manuscript we describe the need for using candidate learners with built-in variable selection methods, such as the lasso, in the compliance and response model ensembles. The idea is that we input the same W_i and each candidate learner selects the most important covariates for predicting complier status or potential outcomes.

2.1.4 Assumption 2

R2 is correct that while Assumption 2 conditions on W, we use propensity to comply in the estimation. In S.1 of the estimation procedure it becomes clear why we write the assumption this way. Assumption 2 implies that $P(C_i = 1|S_i = 1, T_i = 1, W_i) = P(C_i = 1|S_i = 1, T_i = 0, W_i)$. We estimate the first term to get to the second term. In addition, it would be confusing to have a propensity score here because there are different propensities: propensity to comply given W_i, propensity to be included in the RCT, propensity to receive treatment in the observational sample.

We've included in the revised manuscript when stating Assumption 2 a justification for conditioning on W and it's usefulness in the estimation strategy.

R2 is right in pointing out that the original manuscript doesn't make clear where the assumption is used in the proof. The revised manuscript makes clear – as suggested by R2 – that the last line of the proof follows because S.1 allows us to use "complier controls."

2.1.5 Placebo test

R2 helpfully suggests that we conduct placebo tests to provide evidence supporting the identifying assumptions. R2 is correct that a placebo test for Assumption 2 is not possible because we never observe whether RCT controls would actually take-up treatment if assigned.

However, we are able to compare the observed responses of RCT compliers and the responses of population "compliers," adjusted by the covariate distribution of RCT compliers. Significant difference between the mean outcomes of these groups would indicate that either Assumption 1 (for $d = 1$), or Assumptions 3 and 4 are violated. Section 5.3.1 of the revised manuscript further describes the placebo tests.

2.1.6 Violation of no defiers assumption

R2 justifiably asks for more discussion regarding how the violation of Assumption 6 would impact our empirical findings. Section 5.3.2 discusses two sources of bias arising from the presence of defiers: the proportion of defiers in the study and the difference in the average causal effects of treatment received for compliers and defiers.

In the OHIE, the proportion of defiers is relatively small. We argue that the difference in average causal effects of treatment received for compliers and defiers would have to be considerably large in order for the bias to meaningfully alter the interpretation of the empirical results.

2.2 Prediction threshold and modeling assumptions

We use a standard prediction threshold of 50% to classify compliers, so that complier status C_i is a binary variable. This is necessary for S.3 in the estimation procedure, where we subset to observed compliers assigned to treatment and predicted compliers assigned to control.

We describe the prediction threshold and discuss additional modeling assumptions in Section 3.1 of the revised manuscript.

2.3 Interpretation of PATT

R2 is correct in stating that the comparison between PATT and the Hartman et al. (2015) estimator is not an appropriate comparison. The (unadjusted) PATT estimator in our study is the population-average causal effect of taking up treatment, adjusted according to the covariate distribution of RCT participants. In contrast, the Hartman et al. (2015) estimator is the ITT estimator reweighted according the to covariate distribution of the population. We clarify the distinction in Section 4 of the revised manuscript. We have also corrected the manuscript and to make clear that the response curve is fitted to treatment received D_i rather than treatment eligibility T_i, since we cannot actually observe T_i in the population.

Assumption 1 would indeed be violated if the Hartman et al. (2015) estimator were applied in a setting with noncompliance.

2.4 Relationship to AIR approach

R2 points out that when estimating the average treatment effect on treated compliers in a randomized trial, we usually divide the ITT effect by the compliance rate under the assumptions outlined in Angrist, Imbens, and Rubin (1996), and in this approach we don't need to identify individuals in the control group who are compliers.

We include a discussion of the need to identify compliers — rather than weight the unadjusted PATT estimate by the population compliance rate — in Section 1 of the revised manuscript. We decided to take this approach because the compliance rate is likely to differ between the sample and population, as well as across subgroups. Moreover, our approach allows us to to decompose PATT estimates by covariate group.

3 Reviewer: 3 (R3)

3.1 Generalizability and originality

R3 expressed concern that the compliance and generalizability aspects of the paper are not sufficiently connected to warrant an original contribution. We make two original contributions to the literature on extrapolating RCT results to populations: first, we define the assumptions necessary to identify complier-average causal effects for the target population. The need for this contribution is acknowledged in the discussion of Hartman et al. (2015).

Second, we propose a procedure to estimate this quantity with few additional modeling assumptions. Our estimation procedure is novel in that we are estimating the response surface for RCT compliers and using the predicted values of the response surface model to estimate the potential outcomes of population members who received treatment. We describe in Section 1 of the revision how this approach differs from reweighting methods that use propensity scores to adjust the RCT data (e.g., Stuart et al. 2011). Our estimation strategy is also novel in that we are actually predicting which of the RCT controls would have complied to treatment had they been treated. We argue in the revised Section 1 that this estimation step is important because the compliance rate is likely to differ between the sample and population, as well as across pretreatment covariate group.

3.2 Abstract

R3 points out that it the ultimate inferential goal of the paper – i.e., being able to extrapolate RCT sample estimates to a broader population of interest — is unclear in the abstract. We have rewritten the abstract to make it more clear that we are interested in population-level treatment effect estimates.

3.3 DAG

In Section 2.1 of the revised manuscript, we underscore the importance of W_i in the DAG, as suggested by R3.

R3 asks whether there could there be a role for variables that are causes of S_i but have no direct effect on T_i. This question is similar to R2's question about whether W_i is the same across all identifying assumptions (2.1.3). We have decided to keep W_i the same both for

ease of exposition and because we use the same covariate sets in our estimation procedure — a modeling assumption described in Section 3 of the revised manuscript.

3.4 Clarification re: proof

The "by (k)" comments in the Proof of Theorem 1 are supposed to be numbered assumptions. We have corrected the comments in the Proof accordingly.

Following the suggestion of R3, we have added to the discussion of the Proof the intuitive explanation that conditioning on W_i makes sample selection ignorable under Assumption 3.

3.5 Plausibility of strong ignorability assumptions

We've revised Section 2.1 to include more discussion of potential violations of the strong ignorability assumptions, and Section 5.3 discusses whether these assumptions are plausible for our empirical application. Footnote 3 makes clear why we don't have companion assumptions for non-compliers.

3.6 Clarification re: estimation procedure

3.6.1 Prediction threshold

Similar to R2's comment (2.2), R3 asks for clarification on which prediction threshold we use to classify compliers. The description of the estimation procedure in Section 3 of the revised manuscript clarifies this point.

3.6.2 Relationship to reweighting methods

The original manuscript states that the simulation results shows that a reweighting approach is needed to extrapolate RCT results to a population. This is confusing, because our estimation strategy does not use a reweighting approach, such as assigning individuals in the RCT and population a weight according the inverse propensity score (Stuart et al. 2011). It instead estimates the response surface for RCT compliers and extrapolates the response surface to treated individuals in the population. Each method accomplishes the same goal of adjusting the RCT data to a population, either by using inverse propensity score weights or the predicted values from a response surface model.

We differentiate reweighting methods and response surface approach in Section 1 and also generalize the statement at the end of Section 4.2 of the revised manuscript to include both reweighting and response surface methods.

3.6.3 Description of predictive algorithms

Following R3's suggests, we introduce Section 3.2 in the revised manuscript that introduces the predictive algorithms and the ensemble method. Additionally, we describe the method of evaluating the predictive accuracy of the ensemble.

3.7 Grammatical

We thank R3 for noticing grammatical errors and we have corrected them in the revised manuscript.

References

Hartman, Erin, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. 2015. "From SATE to PATT: Combining experimental with observational studies to estimate population treatment effects." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 10:1111.

Kowalski, Amanda E. 2016. *Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments.* Working Paper, Working Paper Series 22363. National Bureau of Economic Research, June. doi:10.3386/w22363. http://www.nber.org/papers/w22363.

Stuart, Elizabeth A, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. 2011. "The use of propensity scores to assess the generalizability of results from randomized trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174 (2): 369–386.

Taubman, Sarah, Heidi Allen, Bill Wright, Katherine Baicker, and Amy Finkelstein. 2014. "Medicaid Increases Emergency Department Use: Evidence from Oregon's Health Insurance Experiment." *Science* 343, no. 6168 (January): 263–268.