Covariate Adjustment with Non-Experimental Units Using `propertee`

Josh Wasserman, Ben B. Hansen
University of Michigan - Ann Arbor, Department of Statistics

Why use `propertee`?

`Propertee` provides researchers with multiple datasets–distinct or overlapping–an R package with which they can learn a covariate adjustment model of their choice with the fitting algorithm of their choice on one dataset, while propagating estimation error to an intention-to-treat (ITT) effect estimated with comprehensible causal interpretations on another.

Small experimental or matched samples often force researchers to adjust for only a select group of confounders when performing covariate adjustment. When available, larger observational datasets can be used to augment these models.

District	School	Match Group	ADSY
COTULLA	ENCINAL EL	1	1
CROWLEY	MEADOWCREEK EL	1	0
GARLAND	WILLIAMS EL	2	1
EDGEWOOD	H B GONZALEZ EL	2	0
⋮	⋮	⋮	⋮

Table 1: Matched units for the ADSY study described in the next section. Treated schools are in orange, controls are in lilac.

District	School	Grade	Match Group	ADSY	Lag 1 Meets Grade Level %	50% Meet Grade Level?
COTULLA	ENCINAL EL	3	1	1	.33	1
COTULLA	ENCINAL EL	4	1	1	.18	1
COTULLA	ENCINAL EL	5	1	1	.53	1
COTULLA	NEWMAN MI	6	N/A	0	.66	1
COTULLA	NEWMAN MI	7	N/A	0	.62	1
COTULLA	NEWMAN MI	8	N/A	0	.65	1
CROSBYTON	CROSBYTON	3	191	0	.52	1
CROSBYTON	CROSBYTON	4	191	0	.4	0
CROSBYTON	CROSBYTON	5	191	0	.44	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2: Covariate and outcome data for the ADSY study. All rows can be used to fit a prognostic regression of the binary grade-level achievement outcome on the lag 1 rate of students who met grade level, but only matched units are used to estimate the ITT effect.

Logistic regression on the combined dataset, however, can yield inconsistent effect estimates when covariates such as fixed effects for matched pair or small-strata designs are included [1].

Separating covariate adjustment estimation from ITT effect estimation, on the other hand, allows for flexibility. Practitioners can generate prognostic scores from a covariate adjustment model whose coefficient estimates are inconsistent or biased, then offset their ITT effect estimate with them without jeopardizing the consistency of their impact estimate. By producing sandwich standard errors that propagate error from the first-stage regression and attend to the experimental design, `propertee` offers asymptotically valid cluster- and heteroskedasticity-robust inference.

`propertee` supports flexible covariate adjustment modeling by propagating uncertainty from a broad prognostic regression fit to ITT effect estimation on a narrower sample.

Using `propertee` for Education Research

In an attempt to combat learning loss exacerbated by COVID-19's uprooting of traditional academic milieu, the Texas state legislature began offering schools financial support for up to to 30 additional school days under an initiative called the Additional Days School Year (ADSY) program. At a broad level, stakeholders are interested in whether ADSY helps revert the trend of increasing numbers of students' test scores failing to meet grade level. Suppose a researcher aims to analyze the effect of the program by first matching schools that receive funds to one or many schools that do not, then estimating the ITT effect on treated schools' binary outcome of whether 50% of its students achieved grade-level test scores. Their analysis of the effect of the treatment on the treated (ETT) may proceed as follows:

Fit a logistic regression model to Table 2, which contains data about suspected confounders such as prior achievement for all schools in the state.

Generate ETT weights using Table 1, which contains information about the matched units.

Offset the matched units' outcomes in Table 2 by their estimated prognostic scores and estimate the ETT on the marginal scale from the matched units only.

`Propertee` retrieves all necessary information from standard arguments to R's `lm` function, shown in the code below:

```
matched_study <- obs_study(
  ADSY ~ unit_id(School) + block(Group),
  data = table_1)
prognostic_model <- glm(
  meets_grade_level_indicator ~
  lag_1_proficiency_rate,
  data = table_2,
  family = binomial)
itt_effect_model <- as.lmitt(
  lm(meets_grade_level_indicator ~ a.(),
     data = table_2[matched_units, ],
     weights = ett(matched_study),
     offset = cov_adj(prognostic_model)),
  design = matched_study)
```

How does `propertee` propagate uncertainty?

`Propertee` views covariate adjustment and ITT effect modeling as a system of stacked estimating equations (theory is detailed in [2]). The former estimates confounding effects β using a sample $\mathcal{C}$ that may overlap with the sample used to estimate the latter, $\mathcal{Q}$. Outcomes in $\mathcal{Q}$ are offset by prognostic scores from the first-stage model. These residuals are used in the second-stage regression to estimate the causal estimand τ.

Letting $\phi(\cdot;\beta)$ denote the covariance adjustment regression function applied to the $n_{\mathcal{C}}$ observations in $\mathcal{C}$ and $\psi(\cdot;\tau,\beta)$ denote the ITT effect regression function applied to the n observations in the union of both samples (unmatched control units receive weights of 0 in this regression), with $\tilde{Y}_i$ representing data vectors that include pre-treatment covariates, treatment assignment indicators, and outcomes, we apply the following theorem from [2] and [3]:

Theorem 1

Assuming $\hat\beta$ and $\hat\tau$ are M-estimators for estimating equations with a unique and exact solution, under certain regularity conditions:

$$\sqrt{n}\begin{pmatrix}\hat\beta\\ \hat\tau\end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix}\beta\\ \tau\end{pmatrix}, A^{-1}BA^{-T}\right) \tag{1}$$

where

$$A = \begin{pmatrix} n_{\mathcal{C}}^{-1}\sum_{i:i\in\mathcal{C}}\frac{\partial}{\partial\beta}\phi(\tilde{Y}_i;\beta) & 0 \\ n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\psi(\tilde{Y}_i;\tau,\beta) & n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\tau}\psi(\tilde{Y}_i;\tau,\beta)\end{pmatrix}$$

$$B = \begin{pmatrix} n_{\mathcal{C}}^{-1}\sum_{i:i\in\mathcal{C}}Cov(\phi(\tilde{Y}_i;\beta)) & n_{\mathcal{C}}^{-1}n^{-1}\sum_{i:i\in\mathcal{C}}Cov(\phi(\tilde{Y}_i;\beta),\psi(\tilde{Y}_i;\tau,\beta)) \\ B_{12}^{T} & n^{-1}\sum_{i=1}^{n}Cov(\psi(\tilde{Y}_i;\tau,\beta))\end{pmatrix}$$

References and Acknowledgements

[1] Alan Agresti. *Categorical Data Analysis*. 2nd ed. John Wiley & Sons, Inc., 2002. ISBN: 978-0-471-24968-9.

[2] Raymond J. Carroll et al. "Appendix A: Background Material". In: *Measurement Error in Nonlinear Models : A Modern Perspective*. 2nd ed. Vol. 105. CRC Press LLC, June 2006.

[3] Leonard A. Stefanski and Dennis D. Boos. "The Calculus of M-Estimation". In: *The American Statistician* 56.1 (Feb. 2002), p. 29.

We thank Josh Errickson and Mark Fredrickson for their work on `propertee`. The research reported here was supported by the Institute of Education Sciences (IES), U.S. Department of Education (USDOE), through Grant R305D210029. The opinions expressed are those of the authors and do not represent views of the IES or the USDOE.