Bounding Approaches for Generalization

Wendy Chan

University of Pennsylvania

Background

Experiments ensure an important aspect of internal validity through treatment randomization, but many experiments lack generalizability when the study samples are not a random subset of the population of interest (Shadish et al., 2002). Statisticians developed methods to improve generalizations using propensity scores, but these methods require *sampling ignorability.* Sampling ignorability assumes that the propensity scores contain all the covariates that explain treatment effect variation and sample selection and that every school in the sample has a "comparable" school in the population. If these core assumptions hold, bias-reduced estimates of the population average treatment effect (PATE) can be attained.

Focus of Study

In this study, we consider an alternative assumption, monotone sample selection, that *bounds* rather than point identifies the PATE. Additionally, we assess bounds under this assumption are tightened under stratification using covariates that explain treatment effect variation and sample selection. Two questions are of interest:

1. To what extent does monotone sample selection improve the precision of bounds compared to the case without assumptions?
2. To what extent does stratification on covariates improve the bounds under monotone sample selection?

Research Design

Let *P* be a population of *N* schools of which *n* are selected into the experimental sample. For a binary treatment, let W = 1 (0) indicate whether a school is assigned to treatment (control), Z = 1 (0) indicate whether the school was selected (not selected) into the sample, and Y(1), Y(0) be the potential outcomes under treatment and control, respectively. The PATE is the difference E(Y(1) – Y(0)), which is decomposed as follows:

$$E(Y(1))= E(Y(1)|W=1, Z=1)\,P(W=1, Z=1) + E(Y(1)|W=0, Z=1)\,P(W=0, Z=1) + E(Y(1)|W=1, Z=0)\,P(W=1, Z=0) + E(Y(1)|W=0, Z=0)\,P(W=0, Z=0) \quad (1)$$

$$E(Y(0)) = E(Y(0)|W=1, Z=1)\,P(W=1, Z=1) + E(Y(0)|W=0, Z=1)\,P(W=0, Z=1) + E(Y(0)|W=1, Z=0)\,P(W=1, Z=0) + E(Y(0)|W=0, Z=0)\,P(W=0, Z=0) \quad (2)$$

Without random sampling, the terms E(Y(1)|W=1, Z=0), E(Y(1)|W=0, Z=0), E(Y(0)|W=1, Z=0), E(Y(0)|W=0, Z=0) are unobservable sample counterfactuals.

If Y(1), Y(0) are bounded by the same lower and upper bound, Y^L, Y^U, the worst case bounds for the PATE are found by replacing the sample counterfactuals with Y^L (Y^U) to get the lower (upper) bound for E(Y(1)) and E(Y(0)). Bounds for the PATE are then derived by taking the difference between the lower and upper bounds of each expected potential outcome.

Monotone Sample Selection

In place of invoking sampling ignorability, we explore the plausibility and the resulting bounds under a monotonicity assumption on the unobservable sample counterfactuals. Evaluation studies have explored a similar assumption, monotone treatment selection, where schools select the treatment that yields the better potential outcome (Manski, 1990). We propose a new assumption, *monotone sample selection* (MSS), where schools that select into the sample (Z=1) over not

selecting into the sample (Z=0) do so to attain the better potential outcome. Formally, MSS is defined as follows:

$$Z = 1 \text{ implies } Y(W, Z=1) \geq Y(W, Z = 0)$$

MSS contributes identifying power by bounding the sample counterfactuals with a smaller upper bound compared to using Y^U.

Bounds with Stratification

The bounds under MSS can be tightened using stratification with covariates (Miratrix et al., 2017). The goal is to divide the schools into subgroups that are compositionally similar in covariate distribution. Here, we stratify the population using the estimated propensity scores and derive an overall bound by aggregating over the stratum-specific bounds. We stratify by the propensity scores because they are summary measures of multiple covariates. Note that we do not use propensity scores to derive point estimates in this context.

Simulation Study

We compared the worst case and MSS bounds, with and without stratification, in a simulation study with three propensity score models. Models 1, 2, and 3 each omitted a covariate *X* that impacted sample selection and the outcomes, but Model 3 included a covariate that was highly correlated with *X*. Model 1 incorporated three out of the five relevant covariates that affected sample selection and the treatment effect whereas Models 2 and 3 only incorporated two. Figure 1 presents the results for a study with $n = 100$ schools out of $N = 2000$. As shown, the bounds are tightest (smallest width) under MSS with stratification, specifically for Model 1 whose propensity score model contained most of the relevant covariates. Interestingly, the MSS bounds under stratification with Model 3 had the second smallest bound width, implying that propensity scores that include a "proxy" covariate also improve the precision in the bounds.

SimCalc

We applied the MSS framework to a cluster randomized trial on a mathematics computer aid, SimCalc (Roschelle et al., 2010). The study took place in Texas from 2008-2009 and consisted of 92 participating middle schools that were not a random subset of the population of 1,713 Texas schools in that year. The outcomes were the student gain scores in mathematics achievement, aggregated to the school level. We argue that MSS is a plausible assumption for the SimCalc study because the participating schools were recruited from Education Service Centers (ESCs) that developed strong networks with the teachers in the community. It is plausible that the sample schools were more willing to participate in ESC-supported studies.

Table 1 provides the corresponding bounds for $k = 5$ strata. Under the worst-case scenario, the average difference in gain scores (on a standardized scale) ranged from -5.13 to 5.28. Under MSS with stratification, this range shrunk to the tightest bound at [-1.06, 2.72]. Although this bound includes zero, the asymmetry of the range suggests that the impact of SimCalc was more positive.

Conclusions

While MSS, like sampling ignorability, cannot be validated empirically, we argue that its plausibility may be suggested by factors that motivate an experimental study. For generalization studies, the simulation and empirical results illustrate that MSS contributes identifying power and these bounds are improved when the population and sample schools are stratified on

covariates that are relevant for predicting treatment variation and sample selection. Furthermore, the simulation results suggest that even if the propensity score models omitted important covariates, including a proxy covariate would still yield precision gains.

References

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, *80*(2), 319-323.

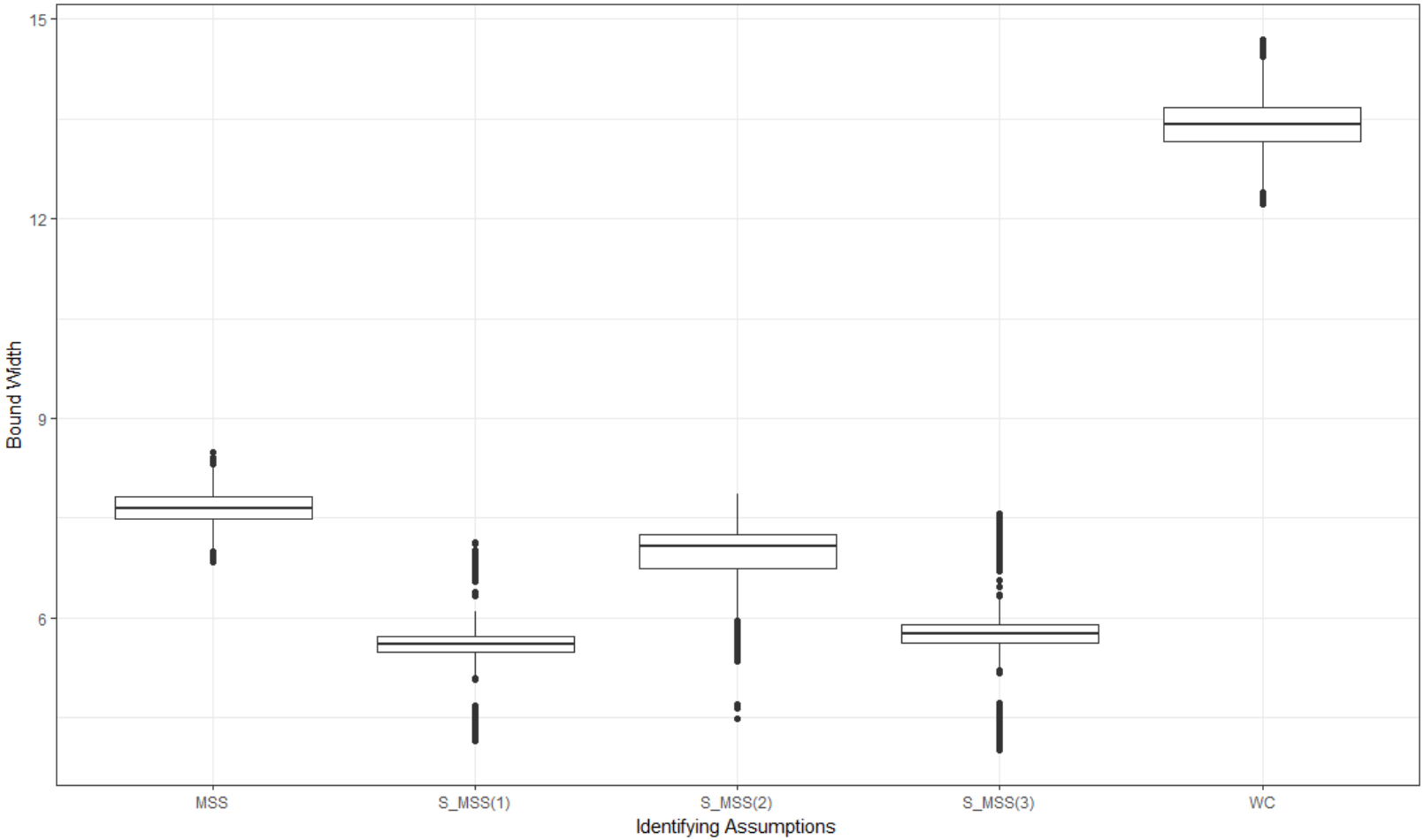
Miratrix, L., Furey, J., Feller, A., Grindal, T., & Page, L.C. (2017). *Bounding, an accessible method for estimating principal causal effects, examined and explained.* arXiv:1701.03139

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., ... & Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal*, *47*(4), 833-878.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Tables and Figures

Figure 1. Comparison of Bound Widths under MSS



Note: MSS refers to the bounds under monotone sample selection. S_MSS(1), S_MSS(2), S_MSS(3) refer to the bounds under MSS with stratification for propensity score models 1, 2, and 3, respectively. WC refers to the worst-case bounds without assumptions.

Table 1. Bounds for the PATE for SimCalc

	Worst Case	Bound Width	MSS	Bound Width
Unweighted	[-5.13, 5.28]	10.41	[-1.15, 2.67]	3.81
Stratum 1	[-4.51, 4.88]	9.40	[-0.94, 2.40]	3.34
Stratum 2	[-5.08, 5.27]	10.36	[-1.19, 2.91]	4.10
Stratum 3	[-5.28, 5.37]	10.65	[-1.40, 2.93]	4.33
Stratum 4	[-5.37, 5.43]	10.81	[-0.92, 2.69]	3.60
Stratum 5	[-5.39, 5.45]	10.84	[-0.87, 2.68]	3.55
Stratification	[-5.13, 5.28]	10.41	[-1.06, 2.72]	3.78

Note: Unweighted refers to the case without stratification.