Propensity score analysis with complex survey data: when treatment effects are heterogeneous across strata and clusters

Trang Quynh Nguyen (tnguye28@jhu.edu), Elizabeth A. Stuart

Johns Hopkins Bloomberg School of Public Health

Society for Research on Educational Effectiveness Spring 2018 Conference
Washington DC, 2018·03·02

Outline

Outline

Big picture

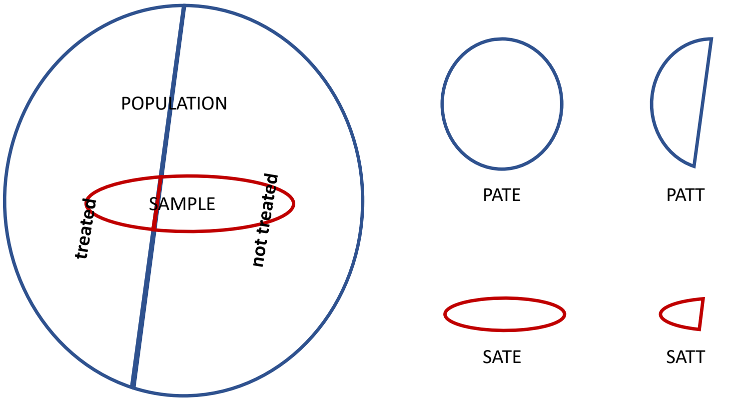
- Researchers may be interested in making causal statements about populations – relevant for policy recommendations
  - What "works" in general practice?
  - What "works" for the general population?

- Ideal: a randomized trial in a representative sample. Rare!

- Instead we have the trade-off:
  - Randomized trials: unbiased for sample, but selective populations
  - Non-experimental studies: data on broad populations, but selection bias

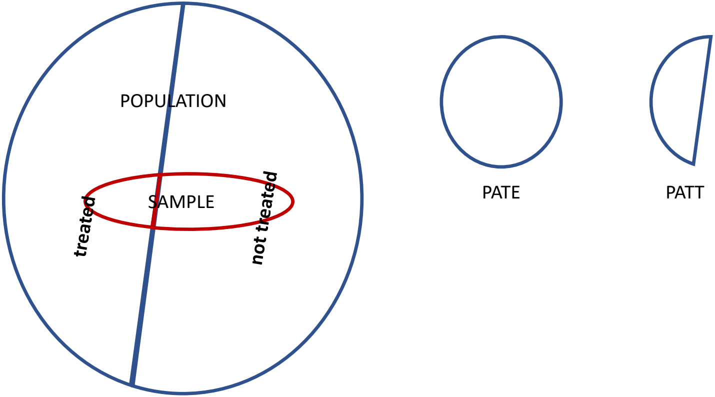
Population vs. sample effects

ATE = average treatment effect; ATT = average treatment effect on the treated



Estimating population effects

How to use a representative yet complex sample to estimate population effects?
– eg the Early Childhood Longitudinal Studies, the Education Longitudinal Study



Outline

Propensity scores (PS)

- To infer effect of treatment A (eg childcare subsidy to poor families) on outcome Y (eg first grade readiness to learn): need treated ($A = 1$) and comparison ($A = 0$) groups to be comparable
    - Not in observational studies
    - So, make them look similar on observed characteristics X – those that may confound treatment effects
    - Key assumption: no unmeasured confounders U
- PS = probability of receiving treatment, given covariates X
    - Is “balancing score”, ie given PS, distribution of X is the same between treated and comparison
    - Use the estimated PS to balance covariate distribution: matching, weighting, subclassification
- After balance obtained
    - Compare outcome between balanced treated and comparison groups
    - Or fit an outcome model (w/ covariates) to the balanced sample

PS methods and complex samples

- Using PS methods on representative population datasets <u>should</u> get us population treatment effects

- But original PS methods assume simple random sampling
    - Many applications with complex survey data ignore survey weights (DuGoff, Schuler, & Stuart, 2014)

- PS methods for complex samples still open area of research

PSs and complex samples: survey weights

- Survey weights incorporate sampling probabilities, non-response adjustment, post-stratification
- Have received much research attention: eg Zanutto (2006), Dugoff et al. (2014), Ridgeway et al. (2015), Austin et al. (2016), Lenis et al. (2017)
- My understanding from this literature (assuming no U)
  - Use survey weights for PS model? It depends.
    - PS matching/subclassification: no need to incorporate survey weights
    - PS weighting: generally, survey-weight the PS model (more in a bit!)
  - Use survey weights for outcome model? Yes!
    - PS matching/subclassification: survey-weight the outcome model
    - PS weighting: multiply survey weights and PS weights
    - Weight transfer? If survey weights depends on A given X – yes for PS matching. I think yes for PS weighting as well.

PSs and complex samples: other design features

- Include strata, clusters as design features in survey analysis commands (eg when fitting outcome model) for appropriate variance estimation
- Strata: include stratum indicators as predictors in outcome model
- Clusters: there is a relevant literature on multilevel PS methods, motivated by clustered data (not necessarily complex surveys)
  – see Hong & Raudenbush 2006, Arpino & Mealli 2011, Kelcey 2011, Thoemmes & West 2011, Li et al. 2013
  - Treatment assignment model may be multilevel with influences by covariates at cluster/individual levels and random effects

Outline

Our motivation: concern about heterogeneity

- Strata Z:
  - Two of the reasons for using stratified sampling instead of SRS:
    - to ensure enough representation of each stratum (subpopulation)
    - to reduce variance of estimates, because within-stratum variance is believed to be smaller than total variance
  - Both imply potentially important/substantial differences across strata
  - Our concern: strata may be systematically different with respect to
    - covariate distribution
    - covariates' influence on treatment assignment, treatment prevalence
    - treatment effects, covariates' modification of treatment effects
  - An otherwise appropriate PS analysis that simply treats Z as a design feature in fitting models might be biased
- Clusters C:
  - Clusters within a stratum may also vary in the same aspects
  - Assume such variation within a stratum is random
    - same spirit with the assumption that sampling units are exchangeable

Setup: Population structure

- L strata
- M clusters, nested in strata
- N units, nested in clusters

Setup: Treatment assignment and treatment effects

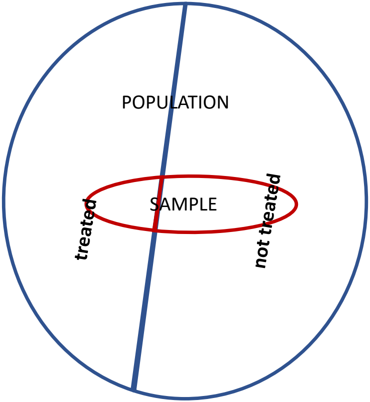
- Treatment assignment
  - True model $\mathrm{P}(A = 1 \mid X, Z, C)$
    - Assume $0 < \mathrm{P}(A = 1 \mid X, Z, C) < 1$ in the inference population
- Potential outcomes and treatment effects
  - Potential outcomes $Y(a)$, for $a = 0, 1$
  - True model $\mathrm{P}[Y(a) \mid X, Z, C]$
    - Assume no unmeasured confounders $(Y(1), Y(0)) \perp\!\!\!\perp A \mid (X, Z, C)$
  - Individual effects, $\mathrm{TE}_i = Y_i(1) - Y_i(0)$, unidentified
  - Interested in population average effect:

$$\mathrm{PATE} = \mathrm{E}[Y(1) - Y(0)] \quad \text{or} \quad \mathrm{PATT} = \mathrm{E}[Y(1) - Y(0) \mid A = 1]$$

Setup: Sample participation

- Multi-stage probability sampling
  - Clusters are sampled within strata
    - Sampling probabilities may depend on stratum and cluster
  - Units are sampled within sampled clusters
    - Usually units within a cluster are sampled with equal probability
- Non-response
  - May depend on factors/characteristics W at cluster or unit level
  - Surveys often adjust for non-response
- Sample participation S requires being sampled and responding
  - True model $\mathrm{P}(S = 1 \mid Z, C, W)$
  - Survey weights are estimates of $1/\mathrm{P}(S = 1 \mid Z = Z_i, C = C_i, W = W_i)$

Weights for estimating population effects



To estimate PATE, need to weight sample treated and sample comparison groups to the population w.r.t. variables that influence $Y_i(a)$ (or TE_i)

Weights for estimating population effects

- The weights that do this are the inverse of

$$P(S = 1, A = A_i \mid X = X_i, Z = Z_i, C = C_i)$$

  - Case 1: if sampling happened after treatment assigned, factor

$$= P(S = 1 \mid A = A_i, X_i, Z_i, C_i)P(A = A_i \mid X_i, Z_i, C_i)$$

  - Case 2: if treatment assigned after sample assembled, factor

$$= P(S = 1 \mid X_i, Z_i, C_i)P(A = A_i \mid S = 1, X_i, Z_i, C_i)$$

- First piece: taken care of by survey weights, assuming $(A, X) \subset W$ or $X \subset W$
- Second piece: population PS in case 1, sample PS in case 2

PSs need to be estimated

- Assume first case, need to estimate population PS, $\mathrm{P}(A = 1|X, Z, C)$
- Survey weights help us use the sample to estimate population PS
- If sample size of each cluster is large, can estimate within each cluster
- If not, need to use some model, eg common logit, probit
- Consider Z first (assuming number of strata not large):
  - ignore strata – not very good
  - stratum indicators – better
  - stratified by stratum – probably best
- Consider C (assuming a lot of clusters):
  - use multilevel modeling – probably best
  - ignore clusters – maybe not bad in some cases

Outline

Simulations to date

- For each scenario, generate 100 populations

- For each population, draw 10,000 samples

Population structure

stratum	number of clusters	cluster size
1	90	6000
2	60	6000
3	70	4000
4	80	4000
5	200	2000
6	150	2000

Covariate distribution

- binary X_1: prevalence varies
  - systematically across strata: .55, .35, .3, .7, .4, .6
  - randomly across clusters: deviations = beta(2,2) recentered and scaled to range $(-.05, .05)$
- continuous X_2:

$$X_{2i} = X_{1i} + U_c^{X_2} + \epsilon_i^{x_2}, \quad U_c^{X_2} \sim N(0, .2), \quad \epsilon_i^{x_2} \sim N(0, 1)$$

Treatment assignment

$$\begin{aligned}\text{logit}[\text{P}(A=1|X,Z,C)] = & [-.5 + (.3)\mathbb{1}\{Z=1,2\} - (.2)\mathbb{1}\{Z=5,6\} + U_c^{A1}] + \\ & [1 + (.5)\mathbb{1}\{Z=1,2\} - (.5)\mathbb{1}\{Z=5,6\} + U_c^{AX_1}]X_1 + \\ & [.5 + (.2)\mathbb{1}\{Z=1,2\} - (.2)\mathbb{1}\{Z=5,6\} + U_c^{AX_2}]X_2 + \end{aligned}$$

- Scenarios vary in the inclusion or exclusion of
  - strata main and interaction effects
  - random cluster effects (normal or recentered gamma)

Potential outcomes and treatment effects

$$
\begin{aligned}
Y(0) = {} & U_c^{Y_0} + \\
& X_1 + \\
& X_2 + \\
& \epsilon^{Y_0} \\
Y(1) = {} & U_c^{Y_0} + [(2)\mathbb{1}\{Z = 1,2\} - (2)\mathbb{1}\{Z = 5,6\} + U_c^{\text{TE}}] + \\
& X_1 + [1 + (.5)\mathbb{1}\{Z = 1,2\} - (.5)\mathbb{1}\{Z = 5,6\} + U_c^{\text{TE}X_1}]X_1 + \\
& X_2 + [1 + (.5)\mathbb{1}\{Z = 1,2\} - (.5)\mathbb{1}\{Z = 5,6\} + U_c^{\text{TE}X_2}]X_2 + \\
& \epsilon^{Y_1} \\
\text{TE} = {} & [(2)\mathbb{1}\{Z = 1,2\} - (2)\mathbb{1}\{Z = 5,6\} + U_c^{\text{TE}}] + \\
& [1 + (.5)\mathbb{1}\{Z = 1,2\} - (.5)\mathbb{1}\{Z = 5,6\} + U_c^{\text{TE}X_1}]X_1 + \\
& [1 + (.5)\mathbb{1}\{Z = 1,2\} - (.5)\mathbb{1}\{Z = 5,6\} + U_c^{\text{TE}X_2}]X_2 + \\
& \epsilon^{Y_1} - \epsilon^{Y_0}
\end{aligned}
$$

$\epsilon^{Y_1}, \epsilon^{Y_0} \sim N(0,1)$. Random cluster effects are normal or recentered gamma.

Sample participation

- In all scenarios, S depends on Z and C via sampling design
  - base scenario: sample 10 clusters per stratum, 100 units per cluster
- Variation due to non-response
  - S does not depend on X or A (base scenario)
  - S depends on binary X_1
  - S depends on A
- Such dependence is captured in survey weights

Methods implemented

- So far, use one-level models, ignoring clusters
- 3 methods w.r.t. strata
  - Naive: ignore strata in both PS and outcome models
  - Strata as covariates: include strum indicators in PS and outcome models
  - Stratified analysis: fit PS model, balance covariates, and fit outcome model in each stratum separately and then combine
- All models fit using `survey` package, with strata, clusters and weights as design features

Results

- Variation in model for sample participation does not matter
  - Not surprising as we have correct survey weights
- Random cluster effects of all kinds only increase variance and do not affect bias
  - Because our outcome model is linear – biases in weights lead to biases contributed by individuals to the PATE that average to zero
  - May not be the case with a nonlinear outcome model
  - Then might want to use a multilevel model to better estimate the PSs
  - Also, a multilevel outcome model may help reduce variance

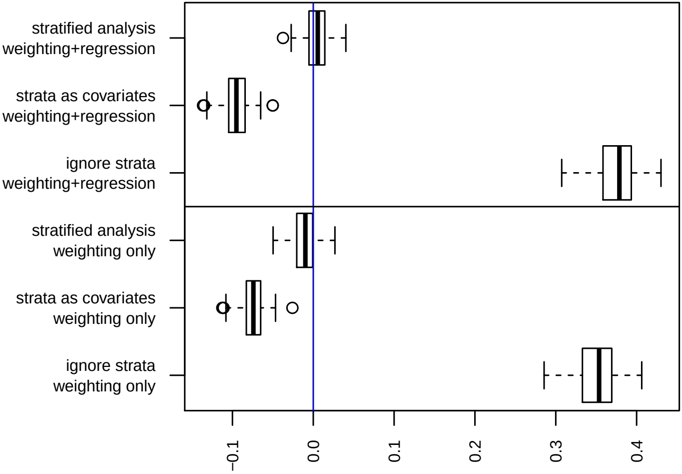
Results

- When treatment effects vary across strata, the naive method is biased
  - Because naive method does not balance Z
  - Should also be problematic when Z is a confounder but not an effect modifier (we didn't have such scenario though)

- When covariates' influence on treatment assignment also varies across strata, the strata-as-covariates method is also biased, but stratified analysis remains unbiased

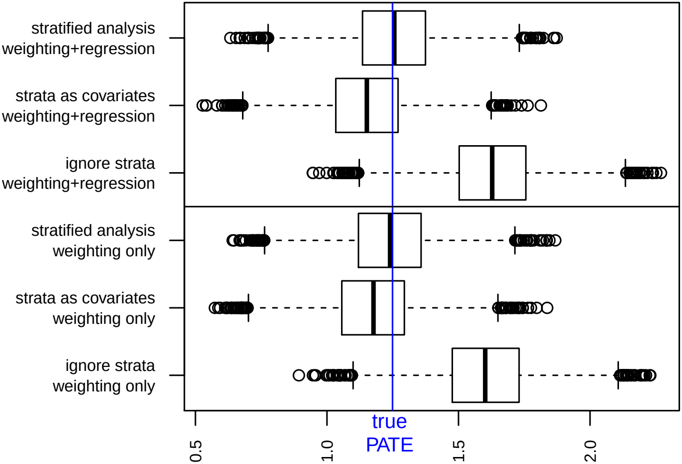
Results

Bias for 100 populations from one scenario
with all cluster- and strata-associated heterogeneity



Results

Estimates on 10,000 samples
drawn from one of those populations



Outline

Recommendations: how to handle strata

- When strata are suspected to vary with respect to either treatment effect or treatment assignment model, they should be incorporated in the analysis

- If strata are suspected to interact with covariates in influencing treatment assignment, stratified analysis is preferred

Recommendations: weights when using PS weighting

- Multiply weights: survey weight × PS weight

- Decide whether PS weight should be based on population PS or sample PS – depends on what the survey weight captures

$$\text{PATE-weight}_i = [\text{P}(S=1, A=A_i \mid X=X_i, Z=Z_i, C=C_i)]^{-1}$$

$$= \begin{cases} \underbrace{[\text{P}(S=1 \mid A_i, X_i, Z_i, C_i)]^{-1}}_{\text{does survey weight capture this?}} \times \underbrace{[\text{P}(A=A_i \mid X_i, Z_i, C_i)]^{-1}}_{\text{population PS}} & \text{case 1} \\ \underbrace{[\text{P}(S=1 \mid X_i, Z_i, C_i)]^{-1}}_{\text{or does it capture this?}} \times \underbrace{[\text{P}(A=A_i \mid S=1, X_i, Z_i, C_i)]^{-1}}_{\text{sample PS}} & \text{case 2} \end{cases}$$

References: survey weights in PS analysis

- Austin PC, Jembere N, Chiu M. (2016). Propensity score matching and complex surveys. *Stat Methods Med Res*. doi:10.1177/0962280216658920.
- Dugoff EH, Schuler MS, Stuart EA. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Serv Res*. 49:284–303.
- Lenis D, Nguyen TQ, Dong N, Stuart EA. (2017). It's all about balance: propensity score matching in the context of complex survey data. *Biostatistics*. doi:10.1093/biostatistics/kxx063
- Ridgeway G, Kovalchik SA, Griffin BA, Kabeto MU. (2015). Propensity score analysis with survey weighted data. *J Causal Inference*. 3:237–49.
- Zanutto EL. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *J Data Sci*. 4:67–91.

References: PS and multilevel data

- Arpino B, Mealli F. (2011). The specification of the propensity score in multilevel observational studies. *Comput Stat Data Anal*. 55:1770–80.
- Hong G, Raudenbush SW. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J Am Stat Assoc*. 101:901–10.
- Kelcey B. (2011). Assessing the Effects of Teachers' Reading Knowledge on Students' Achievement Using Multilevel Propensity Score Stratification. *Educ Eval Policy Anal*. 33:458–82.
- Li F, Zaslavsky AM, Landrum MB. (2013). Propensity score weighting with multilevel data. *Stat Med*. 32:3373–87.
- Thoemmes FJ, West SG. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behav Res*. 46:514–43.