Estimating population average treatment effects from experiments with noncompliance

Kellie Ottoboni, Jason Poulos

April 19, 2015

Abstract

This is the abstract...

1 Introduction

Randomized control trials (RCTs) are the "gold standard" for estimating the causal effect of a treatment. However, external validity is often an issue when RCTs use samples of individuals who do not represent the population of interest. For example, RCTs in which participants volunteer to sign up for health insurance coverage may exhibit a sample population that is in poorer health than the target population. External validity is particularly relevant to policymakers who want to know how a sample average treatment effect (SATE) would generalize to the broader population. Hartman et al. propose a method of reweighting the responses of individuals in an RCT study according to the distribution of covariates in the target population in order to estimate the population average treatment effect on the treated (PATT). Under a series of assumptions, the PATT estimate is asymptotically unbiased **?**.

A prevalent issue in RCTs is noncompliance. In one-way crossover from treatment to control, individuals who are randomly assigned to the treatment group refuse the treatment. The serves to dilute the treatment effect, and the resulting intention to treat estimate is biased towards 0. We propose to extend the method of Hartman et. al. estimate the PATT in settings with RCT non-compliance. After deriving the estimator, we will apply the method to identify the effect of extending Medicare coverage to the low–income population in the U.S.

Statistical Analysis We propose to modify the estimator of PATT from **?** by allowing for the possibility of one-way crossover from treatment to control in the RCT. We make the same assumptions needed for **?**; these include the stable unit treatment value assumption and strong ignorability. To ensure identifiability, we also assume monotonicity (no defiers) and exclusion restriction as in **?**. It does not make sense to talk about compliers and non-compliers in the population, as there is no imposed treatment assignment; compliance is observable only in the RCT sample.

The estimator of PATT involves the expectation of the response of individuals in the RCT sample, conditional on their covariates, where the expectation is taken over the distribution of population covariates. To estimate the conditional expectation of responses in the RCT, we plan to use an ensemble learning method to estimate the response surface. Then, we will use the response model to estimate population members' outcomes given their covariates. These estimates will be used to estimate the PATT. TO DO: Update this paragraph

2 Data

We apply the proposed method to estimate the effect of extending Medicaid coverage on emergency–room use and other health outcomes. Medicaid is a federally-funded program, so understanding how the RCT estimates generalize to a broader population of individuals who will be covered by other Medicaid expansions is informative for public policy.

Oregon Health Insurance Experiment We draw RCT data from the Oregon Health Insurance Experiment **??**. In 2008, approximately 90,000 uninsured low-income adults participated in a lottery to receive Medicaid benefits.[1] Treatment occurred at the household level: participants selected by the lottery won the opportunity for themselves and any household member to apply for Medicaid.[2] In total, about 35,000 participants (representing about 30,000 households) were selected by the lottery; the remaining participants were not able to apply for Medicaid and served as controls in the experiment. Participants in selected households received benefits if they returned an enrollment application within 45 days of receipt. Among participants in selected households, about 60% mailed back applications and only 30% successfully enrolled.[3] The RCT data includes demographic variables such as age, gender, ethnicity, pre-existing medical conditions, education, employment, income, and insurance coverage.

The Oregon Health Study obtained data on the number of emergency room visits for every RCT participant who resides near twelve hospitals in the Portland area ($N = 24,646$). The authors find Medicaid coverage increased emergency-room use over an 18–month period by 40% relative to the control group **?**.

Observational data We have data on the target population from two surveys conducted by the Center for Disease Control: the Behavioral Risk Factors Surveillance Study (BRFSS) **?** and the National Health Interview Study (NHIS) **?**. We will restrict our analysis to adults aged 19-64 whose income is below 100% of the FPL, comparing those with Medicaid to uninsured adults. The outcomes of interest from NHIS are amount of hospital use: outpatient visits, ER visits, and inpatient hospital admissions. The BRFSS outcomes are self-reported health ratings. Both surveys record individuals' demographic covariates, which match those measured in the RCT.

3 Empirical results

TO DO: SHORT INTRO

3.1 Simulation Design

RCT eligibility, complier status, and treatment assignment in the population depend on observed covariates. The observed covariates (W_1, W_2, W_3) are multivariate normal with mean $(0.5, 1, -1)$ and covariances $\text{Cov}(W_1, W_2) = 1$ and $\text{Cov}(W_1, W_3) = \text{Cov}(W_2, W_3) = 0.5$.

[1] Eligible participants include Oregon residents (US citizens or legal immigrants) aged 19 to 64 not otherwise eligible for public insurance, who who have been without insurance for six months, and have income below the federal poverty level (FPL) and assets below $2,000.

[2] Since randomization is applied on the household level, we will need to account for the number of each participant's household members when specifying the probability of treatment assignment.

[3] About half of the returned applications were deemed ineligible, primarily due to failure to demonstrate income below the FPL. Enrolled participants were required to recertify their eligibility status every six months.

The equation for selection into studies is

$$S = \mathbb{I}(e_2 + g_1 W_1 + g_2 W_2 + g_3 W_3 + R > 0)$$

where R is standard normal. e_2 controls the fraction of the population eligible for the RCT. We set g_1, g_2, and g_3 to be $0.5, 0.25$, and 0.75, respectively. The compliance indicator is determined by

$$C = \mathbb{I}(e_3 + h_2 W_2 + h_3 W_3 + Q > 0)$$

where Q is standard normal. e_3 controls the fraction of compliers in the population. We set h_2 and h_3 to 0.5. In the population (individuals with $S = 0$), treatment is assigned by

$$T = \mathbb{I}(e_1 + f_1 W_1 + f_2 W_2 + V > 0)$$

Varying e_1 controls the fraction eligible for treatment in the population. V is standard normal. We set f_1 to 0.25 and f_2 to 0.75. For individuals in the RCT ($S = 1$), treatment assignment is a sample from a Bernoulli distribution. We set treatment received D according to T and C: $D = T$ if $C = 1$ and $D = 0$ if $C = 0$. Finally, the response Y is determined by

$$Y = a + bD + c_1 W_1 + c_2 W_2 + dU$$

We assume that the treatment effect b is heterogeneous depending on W_1: $b = 1$ if $W_1 > 0.75$ and $b = -1$ if $W_1 \leq 0.75$. We set a, c_1, and d to 1 and c_2 to 2. U is standard normal and $U, V, R, Q, (W_1, W_2, W_3)$ are mutually independent.

We generate a population of 30,000 individuals and randomly sample 5,000. Those among the 5,000 who are eligible for the RCT ($S = 1$) are selected. Similarly, we sample 5,000 individuals from the population and select those who are not eligible for the RCT ($S = 0$); these are a sample of the "target population". We set each individual's treatment received D according to their treatment assignment and complier status and observe their responses Y. In the assigned-treatment RCT group ($S = 1, T = 1$), we fit a logistic regression to compliance status using the covariates. With this model, we predict who in the control group ($S = 1, T = 0$) has $C = 1$, since this is unobservable. These individuals *would have* complied had they been assigned to the treatment group. We do the same for the target population control group.

For this group of observed compliers to treatment and predicted compliers from the control group of the RCT, we estimate the response curve using a random forest with features (W_1, W_2, W_3) and D. Then population local average treatment effect on the treated is estimated according to Theorem 1.

3.2 Results

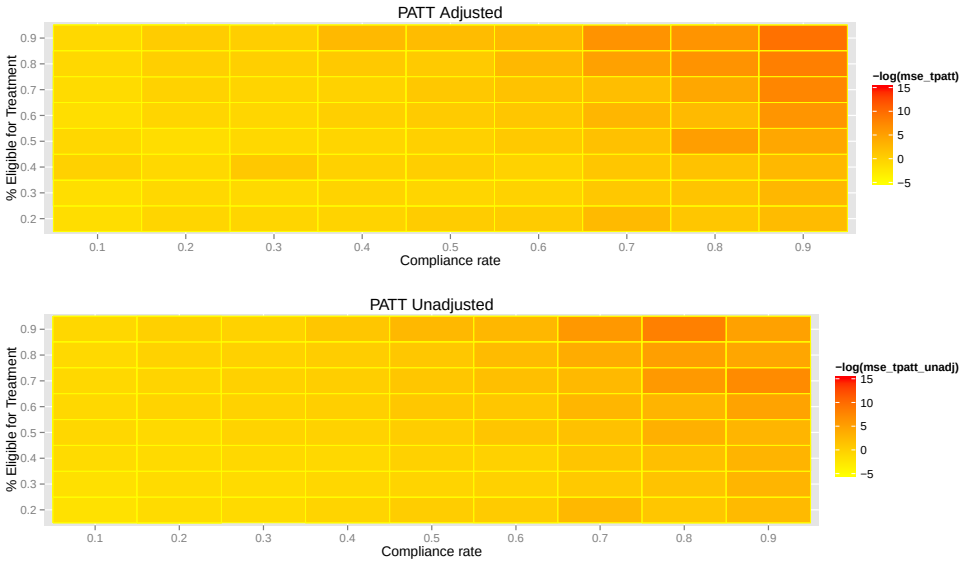

Figure 1: Simulated mean squared error, transformed by $-\log$. Darker tiles correspond to lower mean squared error.

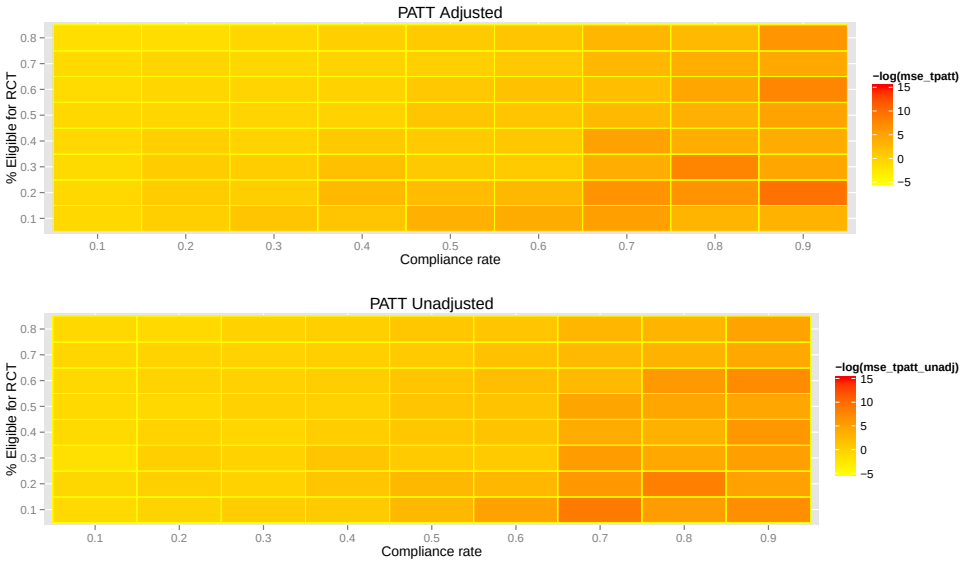

Figure 2: default

Figure 3: Simulated mean squared error, transformed by $-\log$. Darker tiles correspond to lower mean squared error.

4 Discussion

The treatment effect of Medicaid applies to uninsured adults with income below the FPL who express interest in health insurance coverage. The sample population differs in several dimensions from the target population of individuals who will be covered by other Medicaid expansions, such as the Affordable Care Act expansion to cover all adults up to 138% of the FPL. For instance, the RCT participants are disproportionately white urban–dwellers **?**. The RCT participants volunteered for the study and therefore may be in poorer health compared to the population.

TO DO: Update tense The proposed method allows us to decompose both SATE and PATT estimates by subgroup according to covariates common to both RCT and observational datasets (e.g, demographic variables, pre–existing conditions, and insurance coverage). Since the RCT participants are predominately white and volunteered for the study, we expect to find substantial differences between sample and population estimates in terms of ethnicity and pre–existing condition subgroups. Overall, we expect the PATT estimate to be lower than the SATE estimate due to the adverse selection problem in the RCT.

5 Numbering assumptions

This latex template has macros for numbering assumptions. Below are examples if you want to use them.

Assumption 1. *No interference between units:* $Y_{i\boldsymbol{Z}}$ *varies with* Z_i*, but does not vary with other elements of* $\boldsymbol{Z}$*.*

Assumption 2. *Random treatment assignment:* $\mathrm{P}(Z_i|Y_{i\boldsymbol{Z}}) = \mathrm{P}(Z_i)$ *for all* i*.*

Randomization p **value** Given assumptions **??-??**, I use Monte Carlo sampling from a randomization distribution of $\boldsymbol{\delta}$ to estimate the exact two–sided p value:

References

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, June 1996.

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, Katherine Baicker, Oregon Health Study Group, et al. The oregon health insurance experiment: Evidence from the first year. *The quarterly journal of economics*, 127(3):1057, 2012.

Centers for Disease Control and Prevention. Behavioral risk factor surveillance system. `http://www.cdc.gov/brfss/index.html`. April 2015.

National Center for Health Statistics. National health interview survey. `http://www.cdc.gov/nchs/nhis.htm`. April 2015.

Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S. Sekhon. From sate to patt: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society, Series A*, (to appear).

Sarah Taubman, Heidi Allen, Bill Wright, Katherine Baicker, Amy Finkelstein, and the Oregon Health Study Group. Medicaid increases emergency department use: Evidence from oregon's health insurance experiment. *Science*, 343(6168):263–268, January 2014.

Appendix