Estimation of Causal Effect in the Absence of Treatment Observability

Joon Sup Park

Abstract

This paper is to present a method to estimate causal effect when the treatment variable is unobserved. Treatment unobservability is quite a common problem in non-compliance settings, where individuals' actual treatment could be different from the treatment they were assigned. Given fairly standard assumptions, I propose a parametric method to recuperate the treatment variable from the outcome and pre-treatment variables. On the way, I present a diagnosis test that tells us when this method would deliver reliable estimates of the treatment variable and ATE. Simulation studies on R show that the "synthetic" treatment variable and ATE coincide with the true treatment variable and the true ATE fairly well. Limitations of the methods and further research questions will be discussed.

1. Introduction

In this paper, I present a method to estimate causal effect when the treatment variable is unobserved.

Most, if not all, literature on causal inference assumes the observability of treatment and bases its estimators on this assumption. Broadly speaking, we can classify two major methods in causal inference: 1)a parametric method that relies on outcome modeling and 2)a non-parametric method that relies on propensity score. The outcome modeling assumes that we have good knowledge of the functional form of how the outcome variable depends on the pre-treatment variables and the treatment variable, i.e., Y = f(X, Z) + e: here, the treatment variable Z simply enters in the regression model of the outcome. The propensity score, on the other hand, is a probability that a unit receives a treatment conditional on her pre-treatment variables, i.e., Pr(Z | X), and, thus, the treatment variable Z enters in its estimation. In sum, both methods of conducting causal inference relies on the knowledge of treatment variables.

More often than not, however, we are in situations where the treatment is not observed, and there are many ways this problem can occur. For observational studies, we can easily think of an economic shock whose propensity to occur depends on some features of an individual such as her income, job tenure, wealth, age, and education, but whose actual occurrences are not observed by bureaus of statistic. For experimental studies, we have a case of non-compliance where individuals actually take a treatment different from the assigned treatment. Thus, the problem of treatment unobservability is far from limited to pathological cases. On the contrary, it is prevalent in a lot of real problems we want to tackle.

Despite the importance and prevalence of this problem, however, there is few research papers that deal with it. The aim of this paper is to present a method of estimating causal effect in the absence of treatment observability. As stressed above, treatment observability is a crucial component in causal inference, and any method to estimate causal effect in the absence of it will have several limitations. The methods proposed in this paper are no exceptions. Accordingly, I will explore what limitations they have and discuss what conditions are needed for them to be effective.

2. Methodology

I propose a method to estimate causal effect in the absence of treatment observability. It is a parametric method that relies on knowledge of the functional form of how the outcome depends on the pre-treatment variables and the treatment variable.

I retain the following standard assumptions in causal inference literature:

Assumption I (Ignorability): $\{Y(0), Y(1)\} \perp Z \mid X, \ \forall X$

Assumption II (Overlap): $0 < \Pr(Z \mid X) < 1, \ \forall X$

The ignorability assumption implies that there is no unmeasured confounders. This means that the only variables that affect Y(0) and Y(1) are summarized in the measured confounders X's and the treatment Z. The overlap assumption implies that every unit with any confounders X's will have positive probability of receiving treatment Z=1 and no-treatment Z=0.

In addition to these standard assumptions, I also assume that the treatment variable is binary, that the treatment effect is additive, and the noise term for the outcome variable follows Gaussian distribution identically and independently

Assumption III (Binary Treatment): $Z \in \{0,1\}$

Assumption IV (Additivity of Treatment Effect): $Y = f(X) + Z\delta + \epsilon$

Assumption V (Independent Gaussian Noise): $Y = f(X) + Z\delta + \epsilon, \ \epsilon \sim_{iid} N(0, \sigma_y^2)$

I only add that the assumption III, IV, and V are also fairly standard in simpler settings of causal inference. For simplicity, also assume that we know that the treatment effect is positive.

Assumption VI (Positive Treatment Effect): $\delta > 0$

Now we introduce an assumption that we have knowledge of the functional form of how Y depends on X and Z, i.e., of $f(X)$.

Assumption VII (Knowledge of f(X))

Finally, assume that the functional form is linear, i.e., $f(X) = X\beta$.

Assumption VIII (Linearity of f(X)): $Y = X\beta + Z\delta + \epsilon,\ \ \epsilon \sim_{iid} N(0, \sigma_y^2)$

Here, everything hinges on how well we reconstruct the treatment variable Z. First, we regress Y on X alone, our only variables. Then we have

$$\hat{\beta}^{(1)} = (X^T X)^{-1} X^T y = \beta + (X^T X)^{-1} X^T Z\delta + (X^T X)^{-1} X^T \epsilon$$

which is biased. If we take a difference, we get the residual

$$\begin{aligned} \hat{\epsilon}^{(1)} = y - X\hat{\beta}^{(1)} &= X\beta + Z\delta + \epsilon - X\beta - X(X^T X)^{-1} X^T Z\delta + X(X^T X)^{-1} X^T \epsilon \\ &= (I - X(X^T X)^{-1} X^T)\epsilon + (I - X(X^T X)^{-1} X^T) Z\delta \end{aligned}$$

And it is precisely the bias in the residual $\hat{\epsilon}$ that stems from $(I - X(X^T X)^{-1} X^T) Z\delta$ that we use to reconstruct Z. Since Z is binary and takes 0 or 1, we have 2 clusters within $\hat{\epsilon}^{(1)}$:

$$\hat{\epsilon}^{(1)} = \begin{cases} (I - X(X^T X)^{-1} X^T)\epsilon, & if\ Z = 0 \\ (I - X(X^T X)^{-1} X^T)\epsilon + (I - X(X^T X)^{-1} X^T) 1_n \delta, & if\ Z = 1 \end{cases}$$

We can use various probabilistic or non-probabilistic algorithms to cluster $\hat{\epsilon}^{(1)}$ into 2 clusters. As an illustrative example, I used the simplest clustering algorithm, K-means, whose results are reported in the next section. Once we have two clusters, we label the units associated with the larger-in-average cluster with the "synthetic Z" of 1, $\hat{Z}^{(1)} = 1$, and label the units associated with the smaller-in-average cluster with the "synthetic Z" of 0, $\hat{Z}^{(1)} = 0$.

The next step is to evaluate δ based on this synthetic Z, $\hat{Z}^{(1)}$. Regressing Y on X and $\hat{Z}^{(1)}$, we get the coefficients $\hat{\beta}^{(2)}$ and $\hat{\delta}^{(1)}$, and $\hat{\delta}^{(1)}$ would be our first estimate for the average treatment effect (ATE).

Note that successful clustering of $\hat{\epsilon}^{(1)} = (I - X(X^T X)^{-1} X^T)\epsilon + (I - X(X^T X)^{-1} X^T) Z\delta$ into the cluster with $(I - X(X^T X)^{-1} X^T)\epsilon$ and the cluster with $(I - X(X^T X)^{-1} X^T)\epsilon + (I - X(X^T X)^{-1} X^T) 1_n \delta$ depends on 3 terms: ϵ, $X(X^T X)^{-1} X^T$, and δ. Thus, the larger the variance of ϵ compared to δ, the more difficult it would be to successfully cluster $\hat{\epsilon}^{(1)}$ into the cluster with $(I - X(X^T X)^{-1} X^T)\epsilon$ and the cluster with $(I - X(X^T X)^{-1} X^T)\epsilon + (I - X(X^T X)^{-1} X^T) 1_n \delta$. And we can check if clustering will be successful based on the plot and histogram of $\hat{\epsilon}^{(1)}$.

Now, if we checked that clustering would be successful, we can reach further to obtain a better estimate of Z than $\hat{Z}^{(1)}$. In that case, we know that the synthetic Z, $\hat{Z}^{(1)}$, provides good information about Z. This implies that the estimate of the coefficients of X obtained from regressing Y on X and $\hat{Z}^{(1)}$, i.e., $\hat{\beta}^{(2)}$, would be a better estimate of β than its estimate obtained from regressing Y on X alone, i.e., $\hat{\beta}^{(1)}$. In turn, the residual $\hat{\epsilon}^{(2)} = y - X\hat{\beta}^{(2)}$ would provide more precise information of Z than $\hat{\epsilon}^{(1)} = y - X\hat{\beta}^{(1)}$. Thus, clustering $\hat{\epsilon}^{(2)}$ into two would result in finer labeling of Z and obtain a better synthetic Z, $\hat{Z}^{(2)}$. Regressing Y on X and $\hat{Z}^{(2)}$, we get the coefficients $\hat{\beta}^{(3)}$ and $\hat{\delta}^{(2)}$, and $\hat{\delta}^{(2)}$ would be a better estimate for the ATE than our first estimate, $\hat{\delta}^{(1)}$.

Then, again, we can initiate the next iteration with $\hat{\epsilon}^{(3)} = y - X\hat{\beta}^{(3)}$ and cluster $\hat{\epsilon}^{(3)}$ into two, obtaining a better synthetic Z, $\hat{Z}^{(3)}$, from which to obtain $\hat{\beta}^{(4)}$ and $\hat{\delta}^{(3)}$ by regressing Y on X and $\hat{Z}^{(3)}$, and so on. Setting a threshold value of small $t > 0$, we can iterate this procedure S times until $\|\hat{\beta}^{(S+1)} - \hat{\beta}^{(S)}\| < t$ to obtain our final estimates of Z and ATE, i.e., $\hat{Z}^{(S)}$ and $\hat{\delta}^{(S)}$.

3. Results

I conducted a simulation study to see if the parametric method proposed in Section 2.1 yields a good enough estimate of Z and ATE. For simulation, I took the following data generating process:

$$X_i = \left(X_{i,1}, X_{i,2}, \cdots, X_{i,16}\right) \sim_{iid} Multivariate\ Normal(0, \sigma_x^2 I_{16}), \qquad \forall i \in \{1, \cdots, n = 10000\}$$

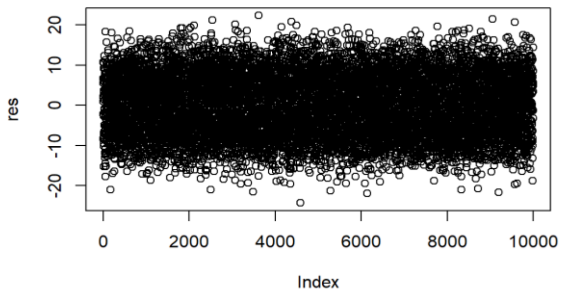
$$Z_i \sim Bernoulli(\pi_i), \qquad where\ \pi_i = \frac{\exp\{\mathrm{X_i}\theta\}}{1 + \exp\{\mathrm{X_i}\theta\}}, \qquad \forall i \in \{1, \cdots, n = 10000\}$$

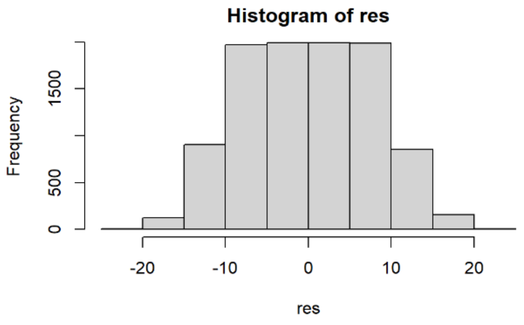
$$where\ \theta = (-1, 0.5, -0.25, -0.1, -1, 0.5, -0.25, -0.1, -1, 0.5, -0.25, -0.1, -1, 0.5, -0.25, -0.1)$$

$$Y_i = 210 + X_i^T\beta + Z_i\delta + \epsilon_i, \qquad where\ \epsilon \sim_{iid} N\left(0, \sigma_y^2\right), \qquad \forall i \in \{1, \cdots, n = 10000\}$$

$$where\ \beta = (27.4, 13.7, -10, 20, 27.4, 13.7, -10, 20, 27.4, 13.7, -10, 20, 27.4, 13.7, -10, 20)\ and\ \delta = 20$$

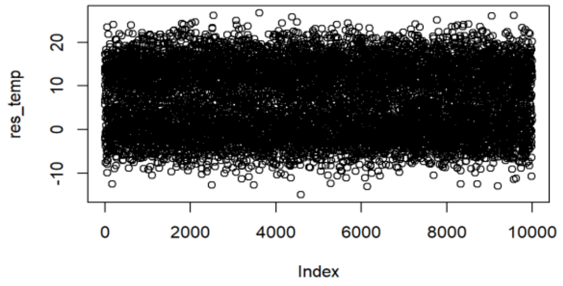
For the first set of simulation, I set $\sigma_x^2 = 1$ and $\sigma_y^2 = 1$. The following is the plot and histogram for the residuals $\hat{\epsilon}^{(1)}$ after regressing Y on X alone, the first step of our parametric method:

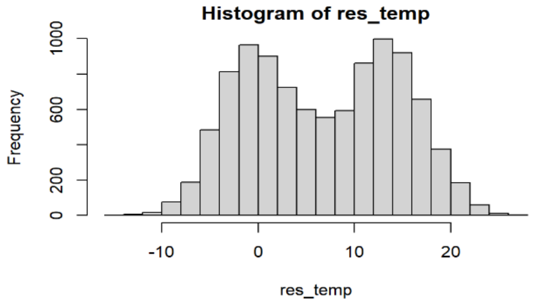




This would be a borderline case where clustering the residuals into two may prove successful. Indeed, the percentage of the cases where first synthetic Z, $\hat{Z}^{(1)}$, equals the true Z was 87.97%. The first estimate of the ATE, $\hat{\delta}^{(1)}$, however, was 13.50, vastly underestimating the true ATE $\delta = 20$.

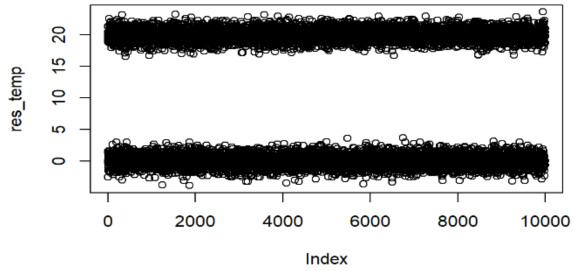
However, already in the 2nd step of our methods, i.e., obtaining the residuals via $\hat{\epsilon}^{(2)} = Y - X\hat{\beta}^{(2)}$, we have the following plot and histogram, which looks much more separable:

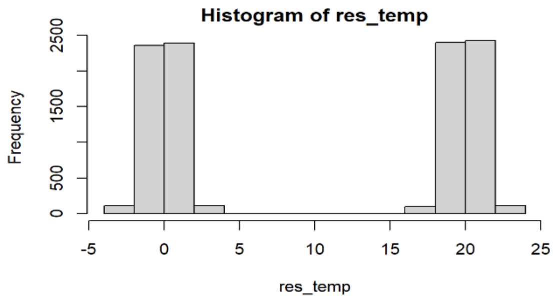




The percentage of the cases where second synthetic Z, $\hat{Z}^{(2)}$, equals the true Z rose up to 93.88%. The second estimate of the ATE, $\hat{\delta}^{(2)}$, too, rose to 14.94, a better estimate of the true ATE $\delta = 20$ than $\hat{\delta}^{(1)}$.

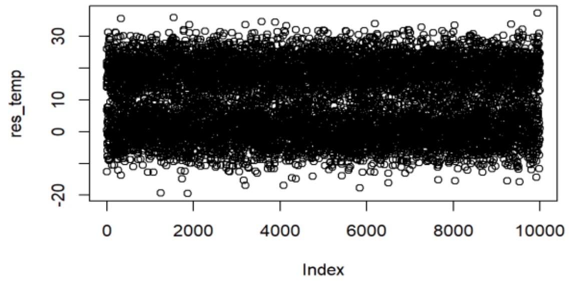
After iterating the procedure S times until $\left\|\hat{\beta}^{(S+1)} - \beta^{(S)}\right\| < 0.1$, we have the following plot and histogram for the S'th step residual $\hat{\epsilon}^{(S)} = Y - X\hat{\beta}^{(S)}$, which looks fully separable into two clusters.

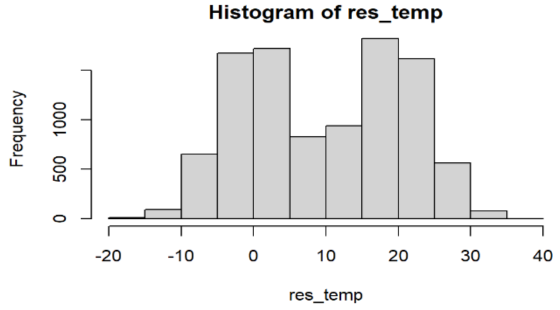




The percentage of the cases where S'th synthetic Z, $\hat{Z}^{(S)}$, equals the true Z rose up to 100%. The S'th estimate of the ATE, $\hat{\delta}^{(S)}$, was 20.002, a very good estimate of the true ATE $\delta = 20$.

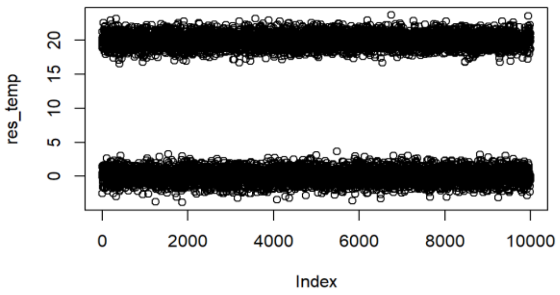
For the second set of simulation, I set $\sigma_x^2 = 1$ and $\sigma_y^2 = 5$. After iterating the procedure S times until $\left\|\hat{\beta}^{(S+1)} - \hat{\beta}^{(S)}\right\| < 0.1$, we have the following plot and histogram for the S'th step residual $\hat{\epsilon}^{(S)} = Y - X\hat{\beta}^{(S)}$, which looks not as separable as the S'th iterated results of the first set of simulation:

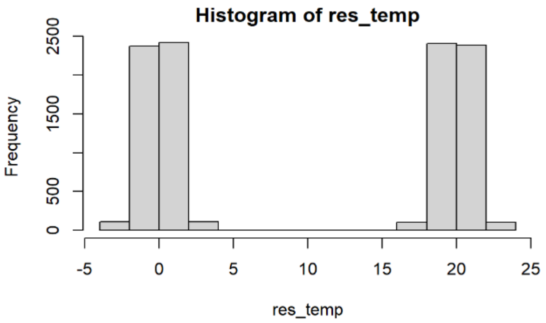




The percentage of the cases where S'th synthetic Z, $\hat{Z}^{(S)}$, equals the true Z is 97.37%. The S'th estimate of the ATE, $\hat{\delta}^{(S)}$, was 19.45, still a very good estimate of the true ATE $\delta = 20$.

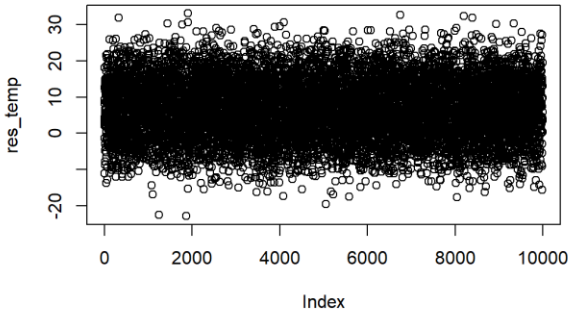
For the third set of simulation, I set $\sigma_x^2 = 5$ and $\sigma_y^2 = 1$. After iterating the procedure S times until $\|\hat{\beta}^{(S+1)} - \hat{\beta}^{(S)}\| < 0.1$, we have the following plot and histogram for the S'th step residual $\hat{\epsilon}^{(S)} = Y - X\hat{\beta}^{(S)}$, which looks as separable as the S'th iterated results of the first set of simulation:

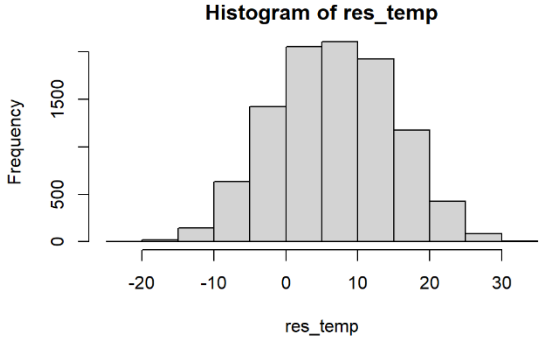




The percentage of the cases where S'th synthetic Z, $\hat{Z}^{(S)}$, equals the true Z is 100%. The S'th estimate of the ATE, $\hat{\delta}^{(S)}$, was 19.99, again a very good estimate of the true ATE $\delta = 20$.

For the final set of simulation, I set $\sigma_x^2 = 5$ and $\sigma_y^2 = 5$. After iterating the procedure S times until $\|\hat{\beta}^{(S+1)} - \hat{\beta}^{(S)}\| < 0.1$, we have the following plot and histogram for the S'th step residual $\hat{\epsilon}^{(S)} = Y - X\hat{\beta}^{(S)}$, which does not look very separable even after many iterations:





The percentage of the cases where S'th synthetic Z, $\hat{Z}^{(S)}$, equals the true Z is only 76.06%. The S'th estimate of the ATE, $\hat{\delta}^{(S)}$, was 13.54, not so good an estimate of the true ATE $\delta = 20$ given the number of iterations.

4. Discussion

As already noted in Section 2.1, successful clustering of $\hat{\epsilon}^{(1)} = (I - X(X^TX)^{-1}X^T)\epsilon + (I - X(X^TX)^{-1}X^T)Z\delta$ into the cluster with ϵ and the cluster with $(I - X(X^TX)^{-1}X^T)\epsilon + (I - X(X^TX)^{-1}X^T)1_n\delta$ would depend on 3 terms: ϵ, $X(X^TX)^{-1}X^T$, and δ. Thus, we concluded that the larger the variance of ϵ compared to δ, the more difficult it would be to successfully cluster $\hat{\epsilon}^{(1)}$ into the cluster with $(I - X(X^TX)^{-1}X^T)\epsilon$ and the cluster with $(I - X(X^TX)^{-1}X^T)\epsilon + (I - X(X^TX)^{-1}X^T)1_n\delta$.

The simulation results confirm this expectation. Moreover, they show that the larger the variance of ϵ, i.e., σ_y^2, compared to δ, the more difficult it is to successfully cluster not only $\hat{\epsilon}^{(1)}$ but successive $\hat{\epsilon}^{(s)}$'s for $s \in \{2, 3, \cdots, S\}$. Also, they show that the larger the variance of X's, i.e., σ_x^2 relative to δ, the more difficult it is to successfully cluster not only $\hat{\epsilon}^{(s)}$'s for $s \in \{1, 2, \cdots, S\}$, although to a less degree than the size of σ_y^2 relative to δ.

The results show that, in confirmation to the expectation, one needs to check the plot and histogram of $\hat{\epsilon}^{(1)}$ in judging the applicability of the proposed method, as they summarize the relative size of σ_y^2 and σ_x^2 compared to δ. Albeit informal, this informs us the conditions upon which the proposed method would be applicable. In the same token, further research is warranted in formally analyzing the conditions.

Another limit of this paper is the absence of exposition for the uncertainty quantification. Again, further research is warranted in exposing the variance of the proposed method.

5. Reference

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3), 399-424.

Chernozhukov, V., Chetverikov, D., Duflo, E., Hansen, C., Demirer, M., Newey, W., Robins, James. (2016). Double/De-biased Machine Learning for Causal and Treatment Effects. ArXiv.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 185-203.

Rubin D. (1979). Using multivariate matched sampling and regression adjustment to control bias in

observational studies. *Journal of the American Statistical Association*; 74:318-324.

Rosenbaum P, Rubin D. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*; 70(1):41-55.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1.