Topics in Econometrics and Statistics Project

on

Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores

Hirano, Imbens, Ridder (2003)

Omer Erhan Erbis
Matrikelnummer: 3466411

July 2023

Contents

1 Introduction

Treatment effects have always been one of the most interesting parameters in the applied economics. Determining the policy efficiency and possible implementations being dependent on effectiveness of decisions caused treatment effects to shine in the experimental research in all areas. Trying to foresee future and learn from past to decide how to act in the present is the natural reason of popularity of treatment effects.

The importance of treatment effects brings two important consequences, namely the precision and accuracy in the estimation process. Correct estimation of treatment effects and setting up correct confidence intervals consist of the basis of the main dimension theoretical treatment effects research.

With the advancements in the literature, different identification strategies were formed and many estimators utilizing these strategies were constructed. From parametric to non-parametric families of estimators, many types of estimators were applied on treatment effects to obtain better estimators in terms of accuracy and precision. Although forming estimators with high accuracy was reached, main question remained was how to improve the precision of the estimates.

The importance of precision thus statistical efficiency of estimators became a high concern in the literature. Forming semi-parametric efficiency bounds proved useful that demonstrated the aimed value for the efficiency. Therefore the estimation strategies that prioritized the efficiency are highlighted in treatment effects literature.

Hahn (1998) has shown one particular curious result that higher knowledge regarding the design of the experiment, which was believed to be contributing to the precision, was not useful and in fact if used incorrectly lead to inefficient estimators. To further this, the estimated value of this quantity regarding the design of the experiment proved to be carrying out more important qualities, and interestingly brought an efficient estimator in contrast. Hirano et al (2003) takes the lead from Hahn (1998) and provides a different non-parametric estimator for the treatment effects in a similar spirit with less parameters to estimate that still provides an efficient estimator for the treatment effects.

2 Literature Review

Fisher's seminal "Statistical Methods for Research Workers" (1925) paper revolutionized the statistics field and enlightened the way to more robust and complex models. Although it was previously utilized, the rise of Randomized Control Trials(RCT) in statistics and econometrics literature paved up the way to correct identification on desired parameters of interest. This early assumption, although too strong to satisfy generally, provided a way and enabled researchers to come up with correct estimates.

Treatment effects, by nature, explains the causal links the world is governed, and the rational decision makers effort to understand the world brought this statistical quantity to the center of attention of econometrics. Potential outcomes framework, which provides the notation in all contemporary treatment effects research first appeared on Neyman's master thesis "On the Application of Probability Theory to Agricultural Experiments" (1923), and was later popularized by Rubin "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies" (1974).

The estimators of treatment effects aimed for both accuracy and precision. Although they managed to attain accuracy, most often than not the precision was not established well for treatment effects.

The natural idea to satisfy this was RCT setup, but since the assumption was hard to justify, researchers turned to search for milder, yet still effective assumptions to enable identification.
This brought us to Unconfoundedness assumption, an approach developed from Rubin (1974) that imposed milder assumptions on the data with better testability. Unconfoundedness was the explanation of circumstances that within strata of observed covariates, potential outcomes corresponding to treatment conditions being balanced between groups. Rosenbaum and Rubin (1983) built up on this with the introduction of balancing scores and proving that probability of treatment assignment, propensity scores, being a balancing score that eased identification process due to lessening the data availability restrictions posed by unconfoundedness. This assumption was similarly named as "strongly ignorable treatment assignment" by Rosenbaum and Rubin (1983), "conditional independence" by Lechner (1999) and "selection on observables" by Barnow, Cain and Goldberger (1971).
The literature developed on the idea of unconfoundedness assumption and with the wide use of propensity scores, treatment effect identification was still not efficient. Hahn (1999) later on proved that even the knowledge of propensity scores did not contribute to precision of some treatment effect estimations, and this started a controversy of efficiency of treatment effect estimators. Hirano et al (2003) builds up on all these discussions and later proves the findings of Hahn (1999) and suggests a new estimator for treatment effects with estimated propensity scores that are not only better but also are efficient.

3 Treatment Effects

Assume a generic form of expression for the dependent variable so that Y is defined by the functional form with a generic $g(.)$ such that:

$$Y_i = g\left(T_i, X_i, \varepsilon_i\right)$$

where:

- Y_i denoting the dependent variable
- T_i an indicator for the treatment
- X_i the attributes/confounders/covariates
- ε_i the error term

In the literature of treatment effects, one of the most used frameworks is potential outcomes framework. Introduced by Neyman (1923) and populatized by Rubin (1974), the potential outcomes partials out treatment status of units and defines the "potential" outcomes instead of realized outcomes, which one of the potential outcomes corresponds to, as:

Definition 3.1 (Potential Outcomes) *Potential outcomes for dependent variable are defined as:*

$$Y_i(T_i = t) = Y_i(t) = g(t, X_i, \varepsilon_i)$$

Potential outcomes are very useful expressing the realized outcomes and identification. With the Stable Unit Treatment Value Assumption (SUTVA) and given the attributes, the potential

outcomes capture the realized dependent variable with the additional information on the treatment status.

Treatment effects capture the potential effect of treatment on units. The representation of treatment effects require information on the setup, that is, information regarding the relations of dependent and independent variables and treatment status. SUTVA assumptions, although hard to test, reduce the dimension of the question and taking treatment effects as external variables enables the treatment effects to be defined with potential outcomes. The treatment effects are thus defined as:

Definition 3.2 (Treatment Effects) *In a binary treatment effects setup, under SUTVA conditions, the individual treatment effects are defined as:*

$$\tau_i = Y_i(1) - Y_i(0)$$

Thus the treatment effects for unit i is can be calculated with the knowledge of potential outcomes. Since the realized outcomes are also a function of potential outcomes and partialed-out treatment effects, the following equation can be shown to hold true:

$$\begin{aligned} Y_i &= T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_t(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) \cdot T_i \\ &= Y_i(0) + \tau_i \cdot T_i \end{aligned}$$

Therefore, if the potential outcome of not receiving treatment, treatment status and the realized outcome are known, individual treatment effects can be found by this equation.

Individual treatment effects, although being important, capture only a small part of population, and most often than not a more generalized version of treatment effects are required for determining the effects of treatment. There are multiple aggregate industry-standard treatment effects to refer to aggregate effects of a treatment.

The Average Treatment Effects (ATE) define the averaged-out individual treatment effects for the whole population.

Definition 3.3 (Average Treatment Effects) *Average Treatment Effects on the whole population are defined as:*

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$$

The Conditional Average Treatment Effects (CATE) define averaged treatment effects for substrata with fixed X values and might be easier to calculate and be a more accurate measure due to fixed attributes and these loosely-fixing unobserved error terms.

Definition 3.4 *Conditional Average Treatment Effects*

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X = x]$$

Average Treatment Effects on Treated (ATT) characterize averaged-out the treatment effects of the treatment-receiver group, which is the recent focus of the treatment effects literature as with a good design of experiment, ATT estimates can closely relate to ATE.

Definition 3.5 *Average Treatment Effects on Treated*

$$\tau_{treated} = \mathbb{E}[Y_i(1) - Y_i(0)|T_i = 1]$$

Weighted Average Treatment Effects (WATE) specify the weighted average treatment effects for any population by weighting sub-populations by a function of x, represented by $g(x)$ with the knowledge of the CATE and the distribution provided.

Definition 3.6 *Weighted Average Treatment Effects*

$$\tau_{wate} = \frac{\int \mathbb{E}[Y(1) - Y(0)|X = x]g(x)dF(x)}{\int g(x)dF(x)}$$

4 Identification

The most crucial aspect of treatment effects setup is treatment assignment process. The experimental design determines the treatment assignment process and provides necessary information for identification of treatment effects. However, diagnosing the exact treatment assignment process is not possible without strict conditions on design and perfect information.
Therefore the identification of the parameter of interest, that is a form of aggregate treatment effects, is often impossible without additional unobserved information on the experiment. Thus the identification process requires certain assumptions on the data to identify the aggregate treatment effects.
Full-randomization assumption, that is, hypothesising about a randomized control trials design would be the simplest solution on experimental data for identification. However, this assumption is frequently violated due to it being too strong on the experimental design. Therefore in the research of treatment effects, a weaker version, namely the unconfoundedness assumption, is frequently imposed on the data at hand that still enables identification of treatment effects of interest. This quality on the sample is easier to satisfy with a good design on experiment.

Assumption 1 (Unconfoundedness Assumption)

$$T \perp (Y(0), Y(1))|X$$

Unconfoundedness assumption implies that the distribution of treatment assignment given covariates X are independent of potential outcomes, that is, there is no selection effect with respect to potential outcomes given covariates X. Satisfying this setup of design is more probable due to being milder on data, and design strategies can satisfy this easier, also being a bit easier to test for.
Another important statistical quantity regarding treatment assignment would be the probability of receiving a treatment. Allowing treatment assignment to be an endogenous process requires treatment assignment probabilities to be a function of covariates as well. The treatment assignment probabilities, namely propensity scores, are defined as,

Definition 4.1 (Propensity Scores) *Propensity score is defined as the probability of treatment assignment given covariates, that is:*

$$p(X_i) = P(T_i = 1|X_i)$$

Propensity scores are important quantities that given the correct functional form, the treatment assignment probabilities that are dictated by design can be identified. The power of propensity scores lies in its relation to design that dictates identification, and prior research has proven that propensity scores are indeed very efficient in identification of treatment effects. With the unconfoundedness assumption, the seminal paper of Rosenbaum and Rubin (1983) show that conditioning on the propensity scores are enough to identify the potential outcomes.

Proposition 4.1 (Matching) *Rosenbaum and Rubin (1983) show that under Unconfoundedness Assumption,*

$$T \perp (Y(0), Y(1))|p(X)$$

With this proposition, the treatment effects can be identified with propensity score matching as:

$$\begin{aligned}
\tau &= \mathbb{E}[\tau_i] \\
&= \mathbb{E}[\mathbb{E}[\tau_i|p(X)]] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0)|p(X)]] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|p(X)]] - \mathbb{E}[\mathbb{E}[Y_i(0)|p(X)]] \\
&\overset{RR}{=} \mathbb{E}[\mathbb{E}[Y_i|p(X), T = 1]] - \mathbb{E}[\mathbb{E}[Y_i|p(X)T = 0]]
\end{aligned}$$

For each possible X value, it is often impossible to condition on with the data at hand to obtain the identification from limited sample. However, what matching achieves is that given the correct form of propensity scores, the identification of the treatment effects is possible which is a substantial result obtained from a limited sample.
With the proof of propensity scores satisfying the balancing score definition by Rosenbaum and Rubin, the propensity scores started to get utilized in treatment effects setup widely. Contemporary research utilizes the propensity scores and matching in inverse probability weighting estimation and doubly-robust estimation for the treatment scores to identify the treatment effects more efficiently with assumptions on the design in addition to outcome regression models.

4.1 Inverse Probability Weighting Technique

The Inverse probability weighting technique existed for long time in econometrics literature. The identification of statisctics drawn from possibly multiple populations is often dealt with weighting the data with probabilities of drawing data from each population. This approach improves the accuracy of the estimation by creating a pseudo collective population from which the data is supposedly drawn from to account for the sample distribution at hand. Two of the most well-known early examples of inverse probability weighting estimators are Horwitz-Thompson and Hajek estimators for mean of the dependent variable.
With the inverse probability weighting technique utilizing the propensity scores and matching in mind, under the Unconfoundedness Assumption, one can show that the potential outcomes can be identified in the following forms:

$$
\begin{aligned}
\mathbb{E}[Y_i(1)] &\overset{LIE}{=} \mathbb{E}[\mathbb{E}[Y_i(1)|X_i]] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[Y_i(1)|X_i]P(T_i=1|X_i)}{P(T_i=1|X_i)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[Y_{i,2}(1)|X_i]\,\mathbb{E}[T_i|X_i])}{p(X_i)}\right] \\
&\overset{UA}{=} \mathbb{E}\left[\frac{\mathbb{E}[Y_i(1)T_i|X_i]}{p(X_i)}\right] \\
&\overset{UA}{=} \mathbb{E}\left[\mathbb{E}\left[\frac{Y_i(1)T_i}{p(X_i)}|X_i\right]\right] \\
&\overset{LIE}{=} \mathbb{E}\left[\frac{Y_iT_i}{p(X_i)}\right]
\end{aligned}
$$

and noting $1-p(X_i) = 1-P(T_i=1 \mid X_i) = P(T_i=0 \mid X_i) = P(1-T_i=1 \mid X_i)$ by binary nature of treatment effects,

$$
\begin{aligned}
\mathbb{E}[Y_i(0)] &\overset{LIE}{=} \mathbb{E}[\mathbb{E}[Y_i(0)|X_i]] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[Y_i(0)|X_i]P(1-T_i=1|X_i)}{P(1-T_i=1|X_i)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[Y_{i,2}(1)|X_i]\,\mathbb{E}[1-T_i|X_i])}{1-p(X_i)}\right] \\
&\overset{UA}{=} \mathbb{E}\left[\frac{\mathbb{E}[Y_i(1)(1-T_i)|X_i]}{1-p(X_i)}\right] \\
&\overset{UA}{=} \mathbb{E}\left[\mathbb{E}\left[\frac{Y_i(1)(1-T_i)}{1-p(X_i)}|X_i\right]\right] \\
&\overset{LIE}{=} \mathbb{E}\left[\frac{Y_i(1-T_i)}{1-p(X_i)}\right]
\end{aligned}
$$

4.2 The Proposed Estimand

With the previous results from inverse probability weighting, it can be shown that the Average Treatment Effects are equivalent to:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}\left[\frac{Y_iT_i}{p(X_i)} - \frac{Y_i(1-T_i)}{1-p(X_i)}\right]
\end{aligned}
$$

4.3 The Proposed Average Treatment Effects Estimator

The inverse probability weighting technique provides a very nice form for the treatment effect identification. Therefore Hirano et al (2003) pursue this inverse probability weighting technique to identify the average treatment effects with the knowledge of data on dependent variable treatment assignment vector and covariates. Since the covariates enter the estimand in the form of propensity scores, if the propensity scores are estimated accurately, the treatment effects will be identified. The most controversial result Hirano et al (2003) reach is that they argue that using the estimated propensity scores instead of the true propensity score result in efficient estimation, and the knowl-

edge of the true propensity scores does not generate any additional information and in fact if used, lead to an inefficient estimator. This result previously discussed in Hahn (1998) is extremely important. Differently than Hahn's (1998) estimator, Hirano et al (2003) require only an estimation of propensity scores and of suggest a nonparametric estimator for propensity scores to estimate the treatment effects as:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Y_i T_i}{p(\hat{X}_i)} - \frac{Y_i(1-T_i)}{1-p(\hat{X}_i)}\right]$$

Before investigating the nonparametric estimator of the propensity scores, illustrating the results with a simple example they suggest should prove useful.

5 Simple Example

To illustrate the results, introduction of the suggested simple example is useful. In the binary treatment effect environment, instead of any generic function, the data at hand is assumed to be from a censored data framework, that is, realized Y_i is observed if treatment is registered to the unit i, otherwise the observed outcome is equal to 0. So the form we have for the dependent variable is:
???

$$Y_i^* = \beta_0 \cdot T_i \;\; where \;\; Y_i^* = Y_i \cdot T_i$$

The number of covariates X_i are assumed to be only one, that is $X_i = [X_i]_{1x1}$ and the X_is are assumed to be binary variables. Thus the collection of data at hand can be written as,

$$\{X_i, T_i, Y_i^*\}_{i=1}^{N}$$

Target is to estimate the population average of Y_i, that is $\beta_0 = \mathbb{E}[Y_i]$ Censored data framework, that is Y_i is only observed if $T_i = 1$.
The analog of Unconfoundedness Assumption in this framework is the Missing at Random Assumption by Rubin (1976):

Assumption 2 (Missing at Random Assumption) *Missing at Random Assumption is defined as:*

$$T \perp Y|X$$

The true selection probability is assumed to be constant at the level $p(x) = 1/2$, thus the missing data assumption translates into Missing Completely at Random assumption.
The target parameter conforms to the previously suggested form of average treatment effects by inverse probability weighting, as β_0 noting $Y(0) = 0$ can be expressed as:

$$
\begin{aligned}
\mathbb{E}\left[\frac{Y_i T_i}{\mathbb{E}[T_i|X_i]}\right] &\overset{LIE}{=} \mathbb{E}\left[\frac{\mathbb{E}\left[Y_i T_i|X_i\right]}{\mathbb{E}[T_i|X_i]}\right] \\
&\overset{MAR}{=} \mathbb{E}\left[\frac{\mathbb{E}\left[Y_i|X_i\right]\mathbb{E}\left[T_i|X_i\right]}{\mathbb{E}[T_i|X_i]}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i|X_i\right]\right] \\
&\overset{LIE}{=} \mathbb{E}[Y_i] \\
&= \beta_0
\end{aligned}
$$

The normalized variance bound for β_0 is derived from Hahn(1998) is given as:

$$V_{bound} = 2 \cdot \mathbb{E}[V(Y|X)] + V(\mathbb{E}[Y|X])$$

The estimation of $beta_0$ using this form can be done in many ways. However, there are are two natural estimators for $beta_0$ that instinctively come to existence considering the previous discussions to researchers: The true weights estimator and the estimated weights estimator.
The "true weights" estimator weights the sample with inverse true selection probability:

$$\hat{\beta}_{tw} = \sum_{i=1}^{N} \frac{Y_i T_i}{1/2}$$

The "estimated weights" estimator weights the sample with inverse of a nonparametric estimate of selection probability:

$$\hat{\beta}_{ew} = \sum_{i=1}^{N} \frac{Y_i T_i}{\hat{p}(X_i)}$$

where for $N_{T=t\,X=x} = \sum_{i=1}^{N} \mathbb{1}(T_i = t)\mathbb{1}(X_i = x)$ and estimated propensity scores defined as:

$$\hat{p}(x) = \begin{cases} N_{10}/\left(N_{00} + N_{10}\right) & \text{if } x = 0 \\ N_{11}/\left(N_{01} + N_{11}\right) & \text{if } x = 1 \end{cases}$$

Hirano et al (2003) argue that the normalized variances of these estimators are found to be:

$$
\begin{aligned}
V_{tw} &= 2 \cdot \mathbb{E}[V(Y|X)] + V(\mathbb{E}[Y|X]) + \mathbb{E}[\mathbb{E}[Y|X]^2] \\
V_{ew} &= 2 \cdot \mathbb{E}[V(Y|X)] + V(\mathbb{E}[Y|X])
\end{aligned}
$$

These calculations demonstrate that the true weights estimator has strictly larger variance than the variance bound if $\mathbb{E}[Y|X] \neq 0$. The estimated weights estimator, on the other hand, not only has a lower variance than the true weights estimator but also reaches the variance bound and proves to be fully efficient. This result confirms the idea Hahn (1998) and Hirano et al (2003) base their results, and brought a debate that continued for a period of time in treatment effects

literature.

6 Main Result

After the illustration of simple example and returning to the general setting, the proposed Average Treatment Effects Estimator by Hirano et al (2003) is:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{Y_i \cdot T_i}{\hat{p}\left(X_i\right)} - \frac{Y_i \cdot (1 - T_i)}{1 - \hat{p}\left(X_i\right)} \right)$$

where the nonparametric propensity score estimate is found by the proposed Maximum Likelihood - Series Logit Estimator by Hirano et al (2003). The proposed Series Logit Estimator has the following tuning parameters and representations:

- r: dimension of X
- K: maximum power level
- $\lambda = (\lambda_1, \ldots, \lambda_r)'$ be an r-dimensional vector of nonnegative integers (power holders)
- $(\lambda(k))_{k=1}^{\infty}$ sequence of power holders
- $r_{kK}(x) = x^{\lambda(k)} = \prod_{j=1}^{r} x_j^{\lambda(k)_j}$: terms of power series
- $R^K(x) = (r_{1K}(x), r_{2K}(x), \ldots, r_{KK}(x))'$: power series
- $L(.) =:$ logistic CDF

The SLE for $p^*(x)$ is defined by $\hat{p}(x) = L\left(R^K(x)'\hat{\pi}_K\right)$ with

$$\hat{\pi}_K = \arg\max_{\pi} \sum_{i=1}^{N} \left(T_i \cdot \ln L\left(R^K\left(X_i\right)'\pi\right) + (1 - T_i) \cdot \ln\left(1 - L\left(R^K\left(X_i\right)'\pi\right)\right)\right)$$

where $p^*(x)$ is the pseudo-true value for propensity scores formed from the sample that dictates the pseudo-distribution of treatment assignment.
These results lead to the proposed efficient estimator under certain restrictions on data to guarantee for a maximizer to exist that could be found. These restrictions are given in the following assumptions:

Assumption 3 (Distribution of X) *(i) the support of X is a cartesian product of compact sets for each dimension with a density bounded away from zero.*

Assumption 4 (Regularity Conditions on Potential Outcomes) *Potential outcomes $Y(0), Y(1)$ exist ($Y(0), Y(1) < \infty$) and expected values of potential outcomes are C^1 on support of X.*

Assumption 5 (Treatment Assignment Probabilities) *Propensity score $p^*(X_i)$ is C^s where $s \geq 7r$ with $p^*(X_i)$ being bounded away from 0 and 1 on the whole support of X_i.*

Assumption 6 (Number of Terms in Estimator) *Number of terms in the Series Logit Estimator $K = N^v$ is bounded between numbers satisfying the following conditions:*

$$\frac{1}{4(\frac{s}{r}-1)} < v < 1$$

Under all these assumptions, the main result of Hirano et al (2003) can be stated as followed:

Proposition 6.1 (Efficient Estimation of Treatment Effects) *Under regularity and smoothness conditions given in the assumptions above to ensure feasibility of the estimator, the proposed estimator has the following qualities:*

1. $\hat{\tau} \xrightarrow{p} \tau^*$
2. $\sqrt{N}\left(\hat{\tau} - \tau^*\right) \xrightarrow{d} \mathcal{N}(0, V)$
3. $\hat{\tau}$ *reaches the semiparametric efficiency bound.*

Hirano et al (2003) show that the proposed estimator can be represented as asymptotically linear to:

$$\hat{\tau} = \tau^* + \frac{1}{N}\sum_{i=1}^{N}\left(\psi\left(Y_i, T_i, X_i, \tau^*, p^*\left(X_i\right)\right) + \alpha\left(T_i, X_i\right)\right) + o_p(1/\sqrt{N})$$

where

$$\psi(y, t, x, \tau, p(x)) = \frac{y \cdot t}{p(x)} - \frac{y \cdot (1-t)}{1-p(x)} - \tau$$

$$\alpha(t, x) = -\left(\frac{\mu_1(x)}{p^*(x)} + \frac{\mu_0(x)}{1-p^*(x)}\right) \cdot (t - p^*(x))$$

Thus the asymptotic variance of the estimator V can be found as:

$$V = \mathbb{E}[(\Psi + \alpha)^2]$$

and a consistent estimator for the variance can be found with imputing estimates of $\tau^*, p^*(X_i)$ and another SLE estimate of α for Ψ, α. The proposed efficient estimator of variance of Hirano et al (2003) is:

$$\hat{V} = \frac{1}{N}\sum_{i}[(\hat{\Psi} + \hat{\alpha})^2]$$

and they provide that

Proposition 6.2 (Consistency of Variance Estimator) *Under regularity and smoothness conditions given in the assumptions above, the following is true:*

$$\hat{V} \xrightarrow{p} V$$

Therefore Hirano et al (2003) provide an alternative non-parametric estimator for treatment effects that is efficient, and they also provide an estimator for the variance that converges to true variance asymptotically. They later extend their results to weighted average treatment effects, but this project focuses on estimation of average treatment effects therefore is considered out of scope.

7 Monte Carlo Simulations

To illustrate the results Hirano et al (2003) provide, Monte Carlo simulations are considered as one of the best tools available to econometricians. In this approach, from a Data Generating Process the sample data will be simulated high number of times and estimators for the treatment effects are going to be found. Then the distribution of these treatment effects are going to be provided to show the efficiency of the proposed estimator.

The estimator Hirano et al (2003) provides requires a high degree of smoothness on propensity scores. Furthermore, inclusion of any number of power terms of covariates increases the data availability requirement tremendously. $K = N^v$ and $v < \frac{1}{9}$ conditions due to assumptions proposed lead to the data requirement being $N \approx 20000$ even for a simple linear approximation of propensity scores for $r = 2$ under all smoothness conditions for propensity scores. In addition, the estimator being in the form of unconstrained Maximum Likelihood Estimator is concerning due to objective being finding a global maximizer. The available unconstrained maximization methods depending on intensive heuristic algorithms such as Broyden–Fletcher–Goldfarb–Shanno algorithm, Berndt–Hall–Hall–Hausman algorithm, Davidon–Fletcher–Powell algorithm or Newton-Raphson algorithm render the calculations highly improbable without substantial amount of computing power. Thus in these simulations, the Monte Carlo simulations of the simple example will be provided, and the treatment effects estimates with simulated treatment effect distribution variances will be provided.

7.1 Simulations for Simple Example

The simple example follows the censored data framework, where the observed dependent variable is 0 if no treatment is introduced to the unit, otherwise the treated outcome value is observed. Covariates are unit binary values, treatment is binary, and the parameter of interest is expected value of true dependent variable.

The parameters of the simulation will be set as followed:

- n_sims = 1000 (number of Monte Carlo simulations)
- n = 1000
- r = 1
- $\pi = \frac{1}{2}$
-

$$\hat{p}(x) = \begin{cases} N_{10}/\left(N_{00} + N_{10}\right) & \text{if } x = 0 \\ N_{11}/\left(N_{01} + N_{11}\right) & \text{if } x = 1 \end{cases}$$

and the DGP will be:

- $X_i \sim B(n, p)$ where $B(n, p)$ is Binomial distribution with n trials and success probability of p.
- $u_i \sim U(0, 1)$ where $U(0, 1)$ is uniform distribution in the interval $[0, 1]$.
- $T = (pi >= u)$ is the binary value of the condition inside parantheses.
- $\beta_0 = 10$.
- $Y_i \sim N(10, 1)$ where $N(\beta_0, 1)$ is the normal distribution with $\mu = 10$ and $Var_Y = 1$.
- The data at hand is,

$$\{X_i, T_i, Y_i \cdot T_i\}_{i=1}^{N}$$

and the model is provided as,

$$\beta_0 = \mathbb{E}[Y] = \mathbb{E}\left[\frac{Y_i T_i}{\mathbb{E}[T_i|X_i]}\right]$$

Therefore the models to estimate the $beta_0$ will be:

$$\hat{\beta}_{tw} = \sum_{i=1}^{N} \frac{Y_i T_i}{1/2}$$

$$\hat{\beta}_{ew} = \sum_{i=1}^{N} \frac{Y_i T_i}{\hat{p}(X_i)}$$

The Monte Carlo simulations can be performed with both classic and GMM estimation strategies as suggested in Hirano et al (2003). The full code and graphs can be accessed in the Github repository.

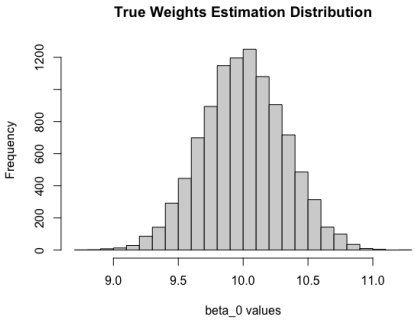

Figure 1: Classic True Weights

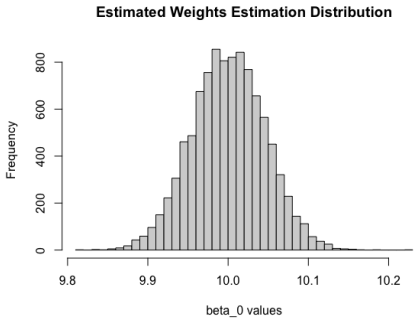

Figure 2: Classic Estimated Weights

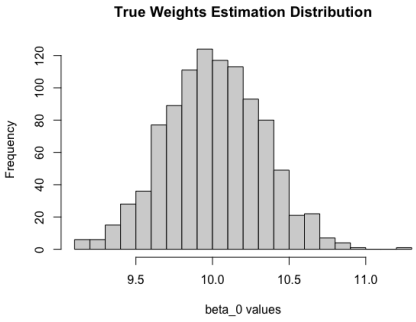

Figure 3: GMM True Weights

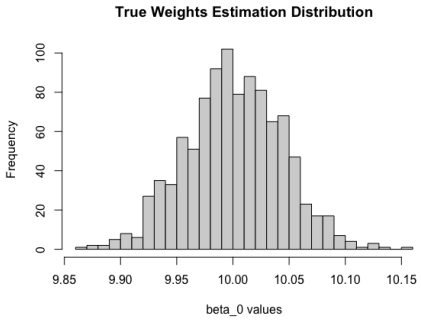

Figure 4: GMM Estimated Weights

	mean	variance
classic_tw	10.00347	0.101902687
classic_ew	10.00047	0.002080783
gmm_tw	10.009554762	0.099702628
gmm_tw.1	10.001113690	0.001863491

Table 1: Estimated β_0 Distribution Parameters

Irrespective of estimation strategies, the simulation results provide that both of the estimators lead up to consistent estimates for β_0. Since these estimators are natural estimators with sample analogs of expected values, the results coincide with expectations.
On the other hand, the variances of the distributions of β_0 vary greatly in terms of variance. The distribution of estimated value for $\hat{\beta}_{tw}$ is way more spread than the distributiin of $\hat{\beta}_{ew}$. This result is the shocking result that Hirano et al (2003) pointed out, that is the true weights estimator, although utilizes more data by inputting the correct value of propensity scores, fails to become efficient and even has larger variance than estimated weights estimator. The estimated weights estimator on the other hand has a lower variance than true weights estimator. The pseudo-distribution estimated weights estimator creates to base the samples seems to add to the value in terms of efficiency to the estimator and is preferred to direct impute of true value.

7.2 Simulation Methodology Main Result

Although it is improbable at the moment due to lack of computing power, a simulation methodology with basis code for simulations are provided in the Github repository in beta version for the full main result of Hirano et al (2003).

8 Conclusion

Treatment effects are very important statistics that decides on policies and effectiveness of decisions in any scope in life. Their importance brings high attention to the estimation of treatment effects. Efficient estimation of treatment effects is thus very important to form better confidence intervals to decide for the significance of possible or past treatments. With the help of the propensity scores, current literature allows the treatment effects to be presented in a way that enables identification with milder conditions imposed on data.
Even though it is believed that more information contributes to better estimators, with all the literature showing this holding true, simple imputing of true propensity scores in treatment effects estimation results in inefficient estimators. This controversial effect is believed to originate from imperfect samples and the imputing method's quality of ignoring results the imperfection of the sample brings. Therefore instead of utilizing the true propensity scores, Hirano et al (2003) have shown that using the non-parametrically estimated propensity scores for treatment effects estimation improves the estimation, and even leads up to an efficient estimation of treatment effects.

9 Appendix

The calculations for the asymptotic variances of the proposed true weights estimator and estimated weights estimator for the simple example is as followed:
Noting that $p(X) = \mathbb{E}[T|X] = 1/2$,
The asymptotic variance of true weights estimator is:

$$
\begin{aligned}
V_{tw} &= V\left(\frac{YT}{\frac{1}{2}}\right) \\
&= 4\,\mathbb{E}\left[(YT - \mathbb{E}[YT])^2\right] \\
&= 4\,\mathbb{E}\left[Y^2T^2 - \mathbb{E}[YT]^2\right] \\
&= 4\,\mathbb{E}\left[Y^2T^2 - \mathbb{E}\left[\mathbb{E}[YT|X]\right]^2\right] \\
&= 4\,\mathbb{E}\left[Y^2T^2 - \mathbb{E}\left[\mathbb{E}[Y|X]\cdot\mathbb{E}[T|X]\right]^2\right] \\
&= 4\,\mathbb{E}\left[Y^2T^2 - \mathbb{E}\left[\frac{1}{2}\,\mathbb{E}[Y|X]\right]^2\right] \\
&= 4\,\mathbb{E}\left[\mathbb{E}[Y^2T^2|X]\right] - 4\,\mathbb{E}\left[\mathbb{E}\left[\frac{1}{2}\,\mathbb{E}[Y|X]\right]^2\right] \\
&= 4\,\mathbb{E}\left[\mathbb{E}[Y^2|X]\cdot\mathbb{E}[T^2|X]\right] - 4\,\mathbb{E}\left[\mathbb{E}\left[\frac{1}{2}\,\mathbb{E}[Y|X]\right]^2\right] \\
&= 4\,\mathbb{E}\left[\mathbb{E}[Y^2|X]\frac{1}{2}\right] - 4\,\mathbb{E}\left[\mathbb{E}\left[\frac{1}{2}\,\mathbb{E}[Y|X]\right]^2\right] \\
&= \mathbb{E}\left[2\,\mathbb{E}[Y^2|X] - \mathbb{E}\left[\mathbb{E}[Y|X]\right]^2\right] \\
&= \mathbb{E}\left[2\,\mathbb{E}[Y^2|X] - 2\,\mathbb{E}[Y|X]^2 + \mathbb{E}[Y|X]^2 - \mathbb{E}\left[\mathbb{E}[Y|X]\right]^2 + \mathbb{E}[Y|X]^2\right] \\
&= 2(\mathbb{E}\left[\mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2\right) + \mathbb{E}[\mathbb{E}[Y|X]^2 - \mathbb{E}\left[\mathbb{E}[Y|X]\right]^2] + \mathbb{E}[\mathbb{E}[Y|X]^2]] \\
&= 2\cdot\mathbb{E}[V(Y|X)] + V(\mathbb{E}[Y|X]) + \mathbb{E}[\mathbb{E}[Y|X]^2]
\end{aligned}
$$

The asymptotic variance of estimated weights estimator by variance decomposition / law of total variance is equal to:

$$
\begin{aligned}
V_{ew} &= V\left(\frac{YT}{\hat{p}(X)}\right) \\
&= V\left(\mathbb{E}\left[\frac{YT}{\hat{p}(X)} \mid X\right]\right) + \mathbb{E}\left[V\left(\frac{YT}{\hat{p}(X)} \mid X\right)\right] \\
&= V\left(\frac{1}{\hat{p}(X)}\,\mathbb{E}\left[Y \mid X\right]\mathbb{E}\left[T \mid X\right]\right) + \mathbb{E}\left[V\left(\frac{YT}{\hat{p}(X)} \mid X\right)\right] \\
&= V\left(\frac{1}{\hat{p}(X)}\,\mathbb{E}\left[Y \mid X\right]\mathbb{E}\left[T \mid X\right]\right) + \mathbb{E}\left[\frac{1}{\hat{p}^2(X)}V\left(YT \mid X\right)\right] \\
&= V\left(\frac{1}{\hat{p}(X)}\,\mathbb{E}\left[Y \mid X\right]\mathbb{E}\left[T \mid X\right]\right) + \mathbb{E}\left[\frac{1}{\hat{p}^2(X)}\left(V(Y \mid X)V(T \mid X) + V(Y \mid X)\,\mathbb{E}[T \mid X]^2 + V(T \mid X)\,\mathbb{E}[Y \mid X]^2\right)\right] \\
&= 2\cdot\mathbb{E}[V(Y|X)] + V(\mathbb{E}[Y|X])
\end{aligned}
$$

10 References

- Keisuke Hirano, Guido W. Imbens and Geert Ridder, 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," Econometrica, Econometric Society, vol. 71(4), pages 1161-1189, July.
- Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. Econometrica, 66(2), 315–331. https://doi.org/10.2307/2998560
- Heckman, James J., et al. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." The Review of Economic Studies, vol. 64, no. 4, 1997, pp. 605–54. JSTOR, https://doi.org/10.2307/2971733.
- Ashenfelter, Orley. "Estimating the Effect of Training Programs on Earnings." The Review of Economics and Statistics, vol. 60, no. 1, 1978, pp. 47–57. JSTOR, https://doi.org/10.2307/1924332.
- Splawa-Neyman, Jerzy, et al. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." Statistical Science, vol. 5, no. 4, 1990, pp. 465–72. JSTOR, http://www.jstor.org/stable/2245382.
- Fisher, R.A. (1925). Statistical methods for research workers (11th ed. rev.). Oliver and Boyd: Edinburgh.
- Heckman, James. "Varieties of Selection Bias." The American Economic Review, vol. 80, no. 2, 1990, pp. 313–18. JSTOR, http://www.jstor.org/stable/2006591
- Rosenbaum, Paul R., and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika, vol. 70, no. 1, 1983, pp. 41–55. JSTOR, https://doi.org/10.2307/2335942. Accessed 3 Sept. 2023.
- Lechner, Michael, Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption (September 1999).
- Barnow, B. S., G. G. Cain and A. S. Goldberger 1980 "Issues in the analysis of selectivity bias." Pp. 43-59 in E. W. Stromsdorfer and G. Farkas (eds.), Evaluation Studies Review Annual, Vol. 5. Beverly Hills: Sage Publi- cations.
- https://github.com/ccfang2/dimada