econometrics

MDPI

Article

Estimating Treatment Effects Using Observational Data and Experimental Data with Non-Overlapping Support

Kevin Han [1,*], Han Wu [1], Linjia Wu [2], Yu Shi [3] and Canyao Liu [3]

[1] Department of Statistics, Stanford University, Stanford, CA 94305, USA
[2] Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, USA
[3] Yale School of Management, Yale University, New Haven, CT 06511, USA
[*] Correspondence: kevinhan1995@outlook.com

Abstract: When estimating treatment effects, the gold standard is to conduct a randomized experiment and then contrast outcomes associated with the treatment group and the control group. However, in many cases, randomized experiments are either conducted with a much smaller scale compared to the size of the target population or accompanied with certain ethical issues and thus hard to implement. Therefore, researchers usually rely on observational data to study causal connections. The downside is that the unconfoundedness assumption, which is the key to validating the use of observational data, is untestable and almost always violated. Hence, any conclusion drawn from observational data should be further analyzed with great care. Given the richness of observational data and usefulness of experimental data, researchers hope to develop credible methods to combine the strength of the two. In this paper, we consider a setting where the observational data contain the outcome of interest as well as a surrogate outcome, while the experimental data contain only the surrogate outcome. We propose an easy-to-implement estimator to estimate the average treatment effect of interest using both the observational data and the experimental data.

Keywords: causal inference; treatment effects; observational studies; surrogate outcomes; unconfoundedness

check for updates

Citation: Han, Kevin, Han Wu, Linjia Wu, Yu Shi, and Canyao Liu. 2024. Estimating Treatment Effects Using Observational Data and Experimental Data with Non-Overlapping Support. *Econometrics* 12: 26. https://doi.org/10.3390/econometrics12030026

Academic Editor: Julien Chevallier

Received: 17 June 2024
Revised: 11 September 2024
Accepted: 12 September 2024
Published: 20 September 2024

1. Introduction

In the realm of causal inference, randomized experiments stand as the gold standard for estimating treatment effects (Imbens and Rubin 2015; Rubin 1974). By randomly assigning individuals to treatment and control groups, researchers can contrast outcomes to study the impact of the intervention. However, the practicality of conducting large-scale randomized experiments often falls short, either due to logistical constraints or ethical considerations. Consequently, researchers often turn to observational data to explore causal relationships, despite its inherent limitations. The cornerstone of using observational data lies in the unconfoundedness assumption (Rosenbaum and Rubin 1983), which states that all confounding factors in the experiment are adequately controlled for. However, in reality, this assumption is often untestable and frequently violated, casting doubt on the validity of conclusions drawn from observational studies.

Due to the complementary strengths of both experimental and observational data, researchers have been proposing methods to combine both experimental data and observational data to estimate treatment effects. For example, Kallus et al. (2018) estimates the bias from using observational data to estimate the average treatment effect with a linear estimator, and Rosenman et al. (2023) takes the advantage of the classic James–Stein shrinkage estimator (James and Stein 1992) to combine the estimates from observational data and experimental data while assuming unconfoundedness. In this paper, we address this challenge within a specific context, where observational data include both the outcome of interest and a surrogate outcome, while experimental data only provide the surrogate outcome. Our objective is to propose an easily implementable estimator that leverages both sources of data to estimate the average treatment effect of interest. By bridging the gap between observational and experimental data, our approach offers a robust and reliable method for treatment effect estimation.

The remainder of the paper is organized as follows. Section 2 introduces the basic setup. In Section 3, we develop our method to estimate the treatment effect of primary outcome by using information from the experimental study. In Section 4, we discuss several widely studied extensions to the basic setup and give a concrete solution to each extension. Section 5 compares several different methods through simulations.

2. Setup

Suppose we want to estimate the treatment effect of an intervention on a primary outcome $Y^P \in \mathbb{R}$. We consider leveraging the data from two distinct studies: an observational study and an experimental study. For each unit *i* in the observational study, we collect data on treatment assignments W_i, the primary outcome Y_i^P, a surrogate outcome $Y_i^S \in \mathbb{R}$, and a set of pre-treatment covariates X_i. The surrogate outcome Y^S is any variable that changes post-treatment, and while we primarily discuss the case where Y^S is one dimensional, our methodology is readily extendable to multi-dimensional surrogate outcomes.

Under the assumption of unconfoundedness, i.e.,

$$Y_i(1), Y_i(0) \perp\!\!\!\perp W_i \mid X_i,$$

either the Inverse Probability Weighting (IPW) estimator or the Augmented Inverse Probability Weighting (AIPW) estimator would suffice for estimating the treatment effect. However, there are numerous scenarios in which the assumption of unconfoundedness is not plausible. In such cases, estimating the treatment effect using only the observational data becomes infeasible.

To address this challenge, we introduce a secondary source of data: a prior study focusing on the surrogate outcome Y^S, where the assumption of unconfoundedness is satisfied. Typically, this prior study takes the form of a small-scale randomized experiment concerning the surrogate outcome. Therefore, we operate with two samples: an observational sample and an experimental sample. Each unit *i* in the observational sample provides a tuple (X_i, W_i, Y_i^S, Y_i^P), while each unit *i* in the experimental sample provides a tuple (X_i, W_i, Y_i^S). It is important to note that the size of the experimental sample N_E is significantly smaller than the size of the observational sample N_O.

Our primary objective is to estimate the quantity

$$\tau^P = \mathbb{E}[Y_i^P(1) - Y_i^P(0) \mid G_i = O],$$

where G_i is an indicator function denoting the sample to which unit *i* belongs. This setup is consistent with the framework presented by Athey et al. (2020).

By integrating data from both the observational and experimental studies, we aim to leverage the strengths of each approach. The observational study provides a large sample size and detailed covariate information, while the experimental study offers reliable causal inference for the surrogate outcome under the unconfoundedness assumption. This combined approach allows us to robustly estimate the treatment effect on the primary outcome Y^P, even in the presence of potential confounding factors in the observational study.

3. Method

In this section, we develop our method to estimate the average treatment effect (ATE) of Y^P. To achieve point identification of the ATE, we assume the following structural model for Y^P:

$$Y_i^P = f(X_i, Y_i^S, \epsilon_i), \qquad \epsilon_i \perp\!\!\!\perp X_i, Y_i^S, \tag{1}$$

where ϵ_i is independent of X_i and Y_i^S. Note that this structural model is general in the sense that the primary outcome can depend on the pre-treatment covariates in an arbitrary way.

We assume the errors to be exogenous. Our estimating procedure can be extended to the case of endogenous error settings like the instrumental variable setup easily. This model implies that all the effect of treatment on the primary outcome is mediated through the surrogate outcome. Consequently, the surrogate outcome, together with the pre-treatment covariates, determines the primary outcome. Under this assumption, τ^P is identifiable.

To see this, we define

$$\tau^S(x) = \mathbb{E}[Y_i^S(1) - Y_i^S(0) \mid X_i = x]$$

and

$$\mu(x,y) = \mathbb{E}\Big[Y_i^P \mid X_i = x, Y_i^S = y, G_i = O\Big].$$

Then, $\mathbb{E}\big[Y_i^P(w)\big]$ can be expressed as $\mathbb{E}[\mu(X_i, Y_i^S(w))]$. The joint distribution of X_i and $Y_i^S(w)$ is identifiable from the experimental sample due to unconfoundedness.

Consider a concrete model where $Y_i^P = \rho Y_i^S + f(X_i) + \epsilon_i$ with ϵ_i being independent of Y_i^S and X_i. For such a model, we can use the Robinson residual-in-residual method to estimate ρ, ensuring the final estimate of the ATE is consistent. For the general case, we can estimate τ^P using the following procedure:

1. Regress Y^P on Y^S and X to obtain an estimate of μ, denoted as $\hat{\mu}$.
2. Estimate the conditional average treatment effect $\tau(x)$ on the surrogate outcome Y^S, obtaining an estimate $\hat{\tau}$.
3. Define $\hat{Y}_i^S(1) = Y_i^S$ if $W_i = 1$ and $\hat{Y}_i^S(1) = Y_i^S + \hat{\tau}(X_i)$ if $W_i = 0$.
4. Estimate $\mathbb{E}\big[Y_i^P(1)\big]$ by $\frac{1}{N_O}\sum_{i=1}^{N_O} \hat{\mu}(X_i, \hat{Y}_i^S(1))$.
5. Define $\hat{Y}_i^S(0) = Y_i^S$ if $W_i = 0$ and $\hat{Y}_i^S(0) = Y_i^S - \hat{\tau}(X_i)$ if $W_i = 1$.
6. Estimate $\mathbb{E}\big[Y_i^P(0)\big]$ by $\frac{1}{N_O}\sum_{i=1}^{N_O} \hat{\mu}(X_i, \hat{Y}_i^S(0))$.
7. The final estimate is $\hat{\tau}^P = \frac{1}{N_O}\sum_{i=1}^{N_O} \hat{\mu}(X_i, \hat{Y}_i^S(1)) - \frac{1}{N_O}\sum_{i=1}^{N_O} \hat{\mu}(X_i, \hat{Y}_i^S(0))$.

With this procedure, we can estimate the ATE on the primary outcome using a single model for the conditional response function μ and one model for the conditional average treatment effect (CATE) estimation. In the following sections, we will discuss various adaptations of this procedure for different scenarios.

4. Applications

In the previous section, we develop a general procedure to combine both the experimental sample and the observational sample. It relies on first estimating the conditional average treatment effect on the surrogate outcome and then correcting the surrogate outcomes in the observational sample. Estimating the conditional average treatment effect (CATE) is usually a case-by-case problem and involves different estimation methods for different settings. In this section, we discuss four settings where we can apply the estimator in Section 3 with different versions of step 2. We also discuss the setting where we drop the unconfoundedness assumption on the experimental sample. In fact, as long as the conditional average treatment effect τ is identifiable, unconfoundedness is not necessary.

4.1. Covariate Support Mismatch between Samples

The first scenario we consider aligns with the setting discussed in Kallus et al. (2018), where the support of pre-treatment covariates in the experimental sample differs from that in the observational sample. We tackle this scenario by adding a calibration step on top of estimating procedure. Such a situation often arises in practice because the experimental sample typically derives from historical data, making it unlikely that the experimental and observational studies target identical populations. Under these circumstances, using only the experimental sample to estimate the conditional average treatment effect (CATE) requires extrapolation to the observational sample. Such extrapolation becomes particularly problematic when the experimental sample size is much smaller than that of the observational sample. Therefore, it is essential to calibrate our CATE estimates on the experimental sample to avoid potential biases.

Kallus et al. (2018) observed that if we define $e^E(x) = \mathbb{P}(W_i = 1 \mid X_i = x, G_i = E)$ and $q^E(X_i) = \frac{W_i}{e^E(X_i)} - \frac{1-W_i}{1-e^E(X_i)}$, then

$$\mathbb{E}[q^E(X_i)Y_i \mid X_i] = \tau(X_i).$$

We now define $\omega(x)$ as

$$\omega(x) = \mathbb{E}[Y_i \mid W_i = 1, X_i = x, G_i = O] - \mathbb{E}[Y_i \mid W_i = 0, X_i = x, G_i = O],$$

then the above observation motivates the following procedure to estimate the CATE of the surrogate outcome on the observational sample:

1. Apply any CATE estimation algorithm on the observational sample to obtain an estimate $\hat{\omega}$.
2. Solve the following optimization problem to obtain $\hat{\theta}$:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N_E} \left(q^E(X_i)Y_i - \hat{\omega}(X_i) - \theta^T X_i \right)^2.$$

3. Define $\hat{\tau}(x) = \hat{\theta}^T x + \hat{\omega}(x)$.

Using the above estimate of τ, we can proceed with the estimator described in Section 3. We can view the above steps as performing additional calibrations. The core idea is to leverage a loss function to estimate the difference between the ill-posed target ω and the true quantity of interest τ. A more general approach can be achieved by fitting a non-parametric function of X_i instead of a linear function.

In summary, this procedure helps to mitigate the issues arising from covariate support mismatch between the experimental and observational samples. By calibrating the CATE estimates from the experimental sample with information from the observational sample, we improve the robustness and reliability of our treatment effect estimates.

4.2. Instrumental Variable (IV) Setting in the Experimental Sample

In this section, we drop our unconfoundedness assumption on the experimental sample and consider the instrumental variable setting which is widely studied in econometrics literature. Note that without the unconfoundedness assumption, point identification is limited to a few specific settings like instrumental variables. Future work could include extending our estimating procedures to the setting where only observational data are available and incorporate our approach with the existing literature on estimating conditional average treatment effects with observational data (Wang et al. 2022; Wu and Yang 2022; Xie et al. 2012).

4.2.1. Constant Effect

We start with the simplest instrumental variable setting where the effect is constant. In particular, we consider a setting where in the experimental sample, we have an instrumental variable Z with the following structural model:

$$Y_i^S = \alpha^T X_i + W_i\tau + \epsilon_i, \qquad \epsilon_i \perp\!\!\!\perp Z_i$$

$$W_i = \beta^T X_i + Z_i\gamma + \xi_i.$$

Such a model is introduced in almost every econometrics textbook, for example, in Angrist and Pischke (2009). It can be seen easily that the parameter τ is exactly the conditional average treatment effect of Y^S. It is well known that we can then estimate τ by two-stage least squares (2SLS) in the instrumental variable literature.

4.2.2. Non-Parametric IV

Now, we consider a more general instrumental variable setting. Specifically, we consider the following model:

$$Y_i^S = \tau(X_i)W_i + g(X_i) + \epsilon_i, \qquad \epsilon_i \perp\!\!\!\perp Z_i,$$

where the effect is a function of the covariates rather than a constant. This is in fact a special case of the more general non-parametric instrumental variable model (Hall and Horowitz 2005; Horowitz 2011; Newey and Powell 2003). To estimate τ, we can follow Hall and Horowitz (2005). First, note that

$$\begin{aligned}\mathbb{E}[Y|W=1,Z=z] &= \mathbb{E}[\tau(X)|W=1,Z=z] + \mathbb{E}[g(X)|W=1,Z=z] \\ &= \int_0^1 (\tau(x)+g(x)) f_{X|W=1,Z}(x,z)dx \\ &= \int_0^1 (\tau(x)+g(x)) \frac{f_{XZ|W=1(x,z)}}{f_{Z|W=1}(z)} dx\end{aligned}$$

Therefore,

$$\mathbb{E}[Y|W=1,Z=z] f_{Z|W=1}(z) = \int_0^1 (\tau(x)+g(x)) f_{XZ|W=1}(x,z)dx.$$

And hence,

$$\mathbb{E}[Y|W=1,Z=z] f_{Z|W=1}(z) f_{XZ|W=1}(u,z) = \int_0^1 (\tau(x)+g(x)) f_{XZ|W=1}(x,z) f_{XZ|W=1}(u,z)dx. \tag{2}$$

If we define

$$t(x,u) = \int_0^1 f_{XZ|W=1}(x,z) f_{XZ|W=1}(u,z)dz$$

and integrate both sides of (2) with respect to z, then we have

$$\mathbb{E}[Y f_{XZ|W=1}(u,Z)] = \int_0^1 (\tau(x)+g(x)) t(x,u)dx$$

for any $u \in [0,1]$, where the expectation on the left-hand side is taken with respect to the conditional joint distribution $(Y,Z|W=1)$. If we define

$$(Th)(u) = \int_0^1 h(x)t(x,u)dx$$

and

$$r(u) = \mathbb{E}[Y f_{XZ|W=1}(u,Z)]$$

then we arrive at the following operator equation

$$r(u) = (T(\tau+g))(u).$$

We can estimate $\tau + g$ using the Hall–Horowitz estimator. Similarly, we have another operator equation, where only g is involved by conditioning on $W = 0$. With that equation, we are able to estimate g. Then, we can estimate τ by taking the difference.

Hall and Horowitz (2005) give good theoretical properties of this method. However, it involves estimating density functions, which is unstable in practice. In fact, Hall and Horowitz (2005) aim to address the general non-parametric IV problem, while we only care about $\tau(x)$.

While our structural model assumption represents a specific case within the broader framework of non-parametric instrumental variable (IV) models, we can leverage alternative methods for more general applications. Specifically, the generalized random forest (GRF) methodology, proposed by Athey et al. (2019), offers a flexible and computationally efficient approach for estimating the conditional average treatment effect (CATE), especially under our structural model assumption.

The GRF framework extends the traditional random forest algorithm to accommodate the estimation of heterogeneous treatment effects (more broadly, any quantity of interest identified as the solution to a set of local moment equations Athey et al. (2019)).

We recommend the use of GRF for estimating τ for the following two reasons:

1. Flexibility: GRF is capable of modeling complex, non-linear relationships between the covariates and the treatment effect, which is often essential in practical applications where such relationships are not adequately captured by parametric models.
2. Generalizability: One notable advantage of GRF is its ability to generalize beyond binary treatment variables. As discussed in Athey et al. (2019), GRF can be extended to settings where the treatment variable W is a real-valued continuous variable.

4.3. Instrumental Variable Setting with Different Support of Pre-Treatment Covariates

In this section, we address the scenario where we have different support of pre-treatment covariates and a non-parametric instrumental variable (IV) model for the experimental sample. This approach is particularly relevant when considering complex experimental designs with multi-dimensional covariates.

To formalize our setup, we define the following conditional expectations:

$$\begin{aligned}
\mu(x) &= \mathbb{E}[Y \mid X = x], \\
\pi(x) &= \mathbb{E}[Z \mid X = x], \\
e(x) &= \mathbb{E}[W \mid X = x], \\
m(x) &= \mathbb{E}[YZ \mid X = x], \\
\gamma(x) &= \mathbb{E}[WZ \mid X = x].
\end{aligned}$$

Given these definitions, it follows that

$$\tau(x)[\gamma(x) - e(x)\pi(x)] - [m(x) - \mu(x)\pi(x)] = 0.$$

Thus, we can write the parameter of interest $\tau(x)$ as the solution to the following minimization problem:

$$\tau(x) = \underset{\tau:\mathcal{X}\to\mathbb{R}}{\arg\min}\, \mathbb{E}\Big[(\tau(x)[\gamma(x) - e(x)\pi(x)] - [m(x) - \mu(x)\pi(x)])^2\Big].$$

The direct estimation of $\tau(x)$ using the above loss function is possible; however, it proves to be inefficient in practice, particularly when dealing with multi-dimensional pre-treatment covariates. This inefficiency arises from the need to estimate numerous nuisance parameters, leading to error accumulation and reduced robustness.

Inspired by the loss-defining property of $\tau(x)$, we propose an alternative estimation procedure, which we term as the Kallus IV method, adapted from Kallus et al. (2018). The procedure is as follows:

1. Apply any conditional average treatment effect (CATE) estimation algorithm, denoted by $\mathcal{Q}$, to the set $\{X_i, W_i, Y_i^S\}_{i=1}^m$ to obtain an initial estimate $\hat{\omega}(x)$.
2. Solve the following optimization problem on the experimental sample to refine the estimate:

$$\hat{\theta} = \underset{\theta}{\arg\min} \sum_{i=1}^{n} \Big([\hat{m}(x_i) - \hat{\mu}(x_i)\hat{\pi}(x_i)] - \Big(\theta^T x_i + \hat{\omega}(x_i)\Big)[\hat{\gamma}(x_i) - \hat{e}(x_i)\hat{\pi}(x_i)] \Big)^2.$$

3. Use the combined estimate $\hat{\omega}(x) + \hat{\theta}^T x$ as the final estimate of the CATE on the surrogate.

This procedure leverages the initial non-parametric estimate and refines it using an optimization framework that adjusts for the instrumental variables' influence. While the optimization step is currently formulated linearly in θ, it is worth noting that a non-parametric function of X_i could be fitted instead. However, empirical results indicate that non-parametric adjustments may lead to unstable estimates when the dimensionality of covariates is high, emphasizing the trade-off between flexibility and stability.

The Kallus IV method thus offers a robust approach to estimate the CATE in the presence of multi-dimensional covariates and instrumental variables.

5. Simulations

In the previous sections, we outlined a procedure to estimate the average treatment effect of the primary outcome using prior information in the experimental sample. We considered three scenarios in which our procedure, as described in Section 3, can be utilized. In this section, we compare several estimators through a series of simulations. Our primary objective is to compare our proposed procedure with the canonical imputation estimator presented by Athey et al. (2020), particularly in cases where we have an unconfounded experimental sample.

We consider two primary settings in our analysis: one where there is no confounding in the experimental sample (i.e., we have either a randomized experiment or an unconfounded experiment) and another where there is confounding (assuming a non-parametric instrumental variable (IV) model for the experimental sample). For each of these settings, we further divide our analysis into two subcases: (1) the support of the pre-treatment covariates in the experimental sample is the same as the support of pre-treatment covariates in the observational sample, and (2) the support of the pre-treatment covariates in the experimental sample is not the same as the support in the observational sample, though they do overlap.

When there is no confounding, we compare three estimators: (1) the imputation estimator as described by Athey et al. (2020), (2) our estimator with $\tau(x)$ estimated using a generalized random forest, and (3) our estimator with $\tau(x)$ estimated using the approach by Kallus et al. (2018). In the presence of confounding, both the imputation estimator and the approach by Kallus et al. (2018) become invalid, as they require the experimental sample to be unconfounded. Therefore, in these scenarios, we compare two estimators: (1) our estimator with $\tau(x)$ estimated by a generalized random forest and (2) our estimator with $\tau(x)$ estimated using the Kallus IV approach.

The simulations are designed to provide a robust comparison of these estimators under varying conditions of confounding and covariate support. By doing so, we aim to identify the strengths and limitations of each method, particularly focusing on the performance of our proposed estimator in different scenarios.

We work with the following data generating mechanism:

$$X_i \sim \mathcal{N}(0, I_{p\times p}),$$

$$\epsilon_i \sim \mathcal{N}(0,1),$$

$$Z_i \sim Binom(1/3),$$

$$Q_i \sim Binom(1/(1+e^{-\omega\epsilon_i})),$$

$$W_i = Z_i \wedge Q_i,$$

$$Y_i^S = \mu(X_i) + (W_i - 1/2)\tau(X_i) + \epsilon_i.$$

and

$$Y_i^P = \sum_{j=1}^{\kappa} X_i^{(j)} + (X_i^{(p)})^2 + 2Y_i^S + (X_i^{(p-2)} + X_i^{(p-1)}X_i^{(p-3)})Y_i^S + \xi_i$$

i.e., $Y^P = f(Y^S, X, \xi)$, where ξ is independent noise. The data generation mechanism above is the same as in the appendix of Athey et al. (2019).

Now, we can adjust several parameters in the data generation mechanism to satisfy different conditions:

1. Presence of Confounding: We vary ω to be either 0 or 1. If $\omega = 0$, there is no confounding; otherwise, there is confounding, and we are in the non-parametric instrumental variable (IV) model.
2. Sparsity of the Signal: We set κ_τ to be either 2 or 4 to vary the sparsity of the signal.
3. Additivity of the Signal: When true, $\tau(x) = \sum_{j=1}^{\kappa_\tau} \max\{0, x_j\}$; when false, $\tau(x) = \max\{0, \sum_{j=1}^{\kappa_\tau} x_j\}$.
4. Presence of Nuisance Terms: When true, $\mu(x) = 3\max\{0, x_5\} + 3\max\{0, x_6\}$ or $\mu(x) = 3\max\{0, x_5 + x_6\}$ depending on the additive signal condition; when false, $\mu(x) = 0$.
5. Identical Support: When true, we assume the distribution of the covariates in the experimental sample and that in the observational sample are the same; when false, $X_i \sim \mathcal{N}([1, \cdots, 1]^T, I_{p\times p})$ in the observational sample.

In our setup, we fix the dimension of X_i (p) to be 10, the experimental sample size (n) to be 300, and the observational sample size (m) to be 1000. We are particularly interested in the treatment effect on Y^P.

To evaluate different methods, we compare their performance based on the mean squared error (MSE). To calculate MSE, we use the Monte Carlo method to estimate the true value of the average treatment effect (ATE) and generate 200 realizations. This approach allows us to robustly assess the accuracy and reliability of the various methods under different conditions.

Tables 1 and 2 present the simulation results. We observe that when the support of the pre-treatment covariates is identical, the generalized random forest (GRF) method outperforms the other two methods, regardless of the presence of confounding. This outcome is expected, as identical support implies no need for extrapolation, rendering the improvements from the Kallus method minimal. Conversely, when the support of pre-treatment covariates differs, both the Kallus and Kallus IV methods demonstrate competitive performance. Notably, in the presence of confounding, the Kallus IV method surpasses the GRF method in terms of performance.

To further explore the scenario of differing supports, we modify the previous setting slightly. Specifically, we now assume that when the support of the pre-treatment covariates is not identical, the support of the pre-treatment covariates in the experimental sample is contained within the support of the pre-treatment covariates in the observational sample, rather than merely overlapping. Specifically, we have the following:

5a Identical support: When true, we assume the distribution of the covariates in the experimental sample and that in the observational sample are the same: $X_i^{(j)} \sim \text{Uniform}(-1, 1)$. When false, $X_i^{(j)} \sim \text{Uniform}(-1, 1)$ in the experimental sample and $X_i \sim \mathcal{N}(0, I_{p\times p})$ in the observational sample.

Table 1. Simulation results for $\omega = 0$.

ω	κ_τ	Additivity	Nuisance	Identical Support	MC Estimate	GRF	Imputation	Kallus	Winner
0	2	Yes	Yes	Yes	1.62	0.19	1.02	0.34	GRF
0	2	Yes	No	Yes	1.58	0.12	0.22	0.24	GRF
0	4	No	Yes	Yes	2.10	0.22	1.13	0.55	GRF
0	4	No	No	Yes	2.10	0.14	0.26	0.41	GRF
0	2	Yes	Yes	No	8.73	30.91	43.38	72.83	GRF
0	2	Yes	No	No	8.67	27.28	36.00	6.45	Kallus
0	4	No	Yes	No	8.11	18.56	29.07	51.25	GRF
0	4	No	No	No	8.11	15.98	23.29	7.05	Kallus

Table 2. Simulation results for $\omega = 1$.

ω	κ_τ	Additivity	Nuisance	Identical Support	MC Estimate	GRF	Kallus IV	Winner
1	2	Yes	Yes	Yes	1.63	0.46	0.65	GRF
1	2	Yes	No	Yes	1.55	0.26	0.80	GRF
1	4	No	Yes	Yes	2.12	0.49	0.64	GRF
1	4	No	No	Yes	2.11	0.30	0.51	GRF
1	2	Yes	Yes	No	8.73	31.60	28.27	Kallus IV
1	2	Yes	No	No	8.72	27.84	10.80	Kallus IV
1	2	No	Yes	No	6.35	15.00	31.79	GRF
1	2	No	No	No	6.33	12.64	11.01	Kallus IV
1	4	No	Yes	No	8.11	17.93	32.65	GRF
1	4	No	No	No	8.09	15.35	16.72	GRF
1	4	Yes	No	No	17.30	109.78	28.89	Kallus IV
1	4	Yes	Yes	No	17.38	114.74	42.00	Kallus IV

Table 3 shows the simulation results. We see that similar to the simulation results in the previous two tables, Kallus/Kallus IV performs better than GRF when we have different support.

Table 3. Simulation results, inclusion of the support.

ω	κ_τ	Additivity	Nuisance	Identical Support	MC Estimate	GRF	Kallus/Kallus IV	Winner
0	2	Yes	Yes	No	1.61	0.60	0.30	Kallus
0	2	Yes	No	No	1.60	0.58	0.29	Kallus
0	4	No	Yes	No	2.11	1.12	0.54	Kallus
0	4	No	No	No	2.10	1.16	0.45	Kallus
1	2	Yes	Yes	No	1.60	0.77	0.71	Kallus IV
1	2	Yes	No	No	1.60	0.70	0.68	Kallus IV
1	2	No	Yes	No	1.34	0.61	0.83	GRF
1	2	No	No	No	1.35	0.58	0.71	GRF
1	4	No	Yes	No	2.10	1.21	0.66	Kallus IV
1	4	No	No	No	2.08	1.24	0.57	Kallus IV
1	4	Yes	No	No	3.21	2.37	0.60	Kallus IV
1	4	Yes	Yes	No	3.23	2.26	0.54	Kallus IV

6. A Real Data Example

In this section, we investigate the performance of our procedure on a real dataset. We provided several applications in Section 4 and simulation studies in the previous section. In this section, we use a real world example to demonstrate the robustness of our procedure on real data. We utilize the famous Tennessee STAR study (Achilles et al. 2008). The Tennessee Student/Teacher Achievement Ratio (STAR) study was a large-scale, longitudinal educational experiment conducted in the late 1980s to examine the effects of class size on student performance. In this study, over 7000 students from kindergarten to third grade across 79 schools were randomly assigned to one of three types of classrooms: small classes (13–17 students), regular-sized classes (22–25 students), or regular-sized classes with a teacher's aide. The goal of the study was to assess whether smaller class sizes would lead to improved academic outcomes, such as higher test scores and long-term achievement. This dataset is also used in Kallus et al. (2018) and Athey et al. (2020). We use it in a different manner. Specifically, we select the following covariates for each student: gender, race, birth month, birthday, birth year, free lunch given or not, teacher id, and student home location. We focus on two outcomes: average grade in year 1 and average grade in year 3. We remove all the records with missing outcome variables. Now, in this study, the treatment is whether or not the student is in a small class (treatment) or regular class (control). After cleaning the data, we have a dataset with 2498 units, 9 covariates, 1 treatment variable and 2 outcome variables. We use the method in Athey et al. (2020) to generate a large population, which we view as the ground truth. We call this ground truth dataset $\mathcal{D}_{gt}$. To assess different methods, we perform the following:

1. Use $\mathcal{D}_{gt}$ to calculate the average treatment effect of average grade in year 3. This estimate τ_{gt} will be viewed as the ground truth.
2. Repeat the following steps 500 times.

3. Sample n_{exp} rural or inner-city students together with all the covariates except the student location covariate, treatment variable and average grade in year 1. This is our experimental sample $\mathcal{D}_E$.
4. Sample $n_{\text{obs}}/4$ rural or inner-city students in control group that are not sampled in experimental sample, sample $n_{\text{obs}}/4$ rural or inner-city students in treatment group whose year 1 average grade is in the lower half among treated rural or inner-city students, sample $n_{\text{obs}}/4$ urban or suburban students in control group, and finally sample $n_{\text{obs}}/4$ urban or suburban students in treatment group whose year 1 average grade is in the lower half among treated urban or suburban students. This is our observational sample (which is confounded because we remove students with higher scores selectively from the population) $\mathcal{D}_O$.
5. Use different methods to estimate τ_{gt} based on $\mathcal{D}_E$ and $\mathcal{D}_O$.
6. Compare based on mean squared error (MSE).

We will only compare the GRF and imputation estimators, as the Kallus method involves estimating the coefficient of a linear function of the covariates but we only have categorical variables. We also include the mean squared error of the AIPW estimator (notice that AIPW estimator requires the sample to be unconfounded) on observational sample. Table 4 gives the results. We see that in general, the GRF estimator outperforms the imputation estimator, and these two estimators both outperform the AIPW estimator significantly. In particular, as Table 5 shows, the empirical mean of AIPW estimates is actually a negative number (and the true treatment effect is a positive number) and is far from the true treatment effect.

Table 4. STAR study simulation.

n_{exp}	n_{obs}	GRF	Imputation	AIPW
300	1000	7.08	13.19	167.52
200	1500	9.36	12.76	167.43
500	2000	4.54	7.43	166.08

Table 5. STAR study simulation, empirical mean, and true treatment effect.

n_{exp}	n_{obs}	GRF	Imputation	AIPW	τ
300	1000	6.64	7.90	−5.21	7.62
200	1500	6.89	8.21	−5.24	7.62
500	2000	6.70	8.06	−5.21	7.62

7. Conclusions

In this paper, we proposed a straightforward procedure to estimate the average treatment effect (ATE) of the primary outcome in an observational study by leveraging an experimental study for the surrogate outcome. We demonstrated that our procedure is applicable in various settings, provided that the conditional average treatment effect (CATE) of the surrogate outcome can be accurately estimated. Through a series of simulations, we compared several methods and showed that our procedure produces a more precise estimate, in terms of mean square error (MSE), than the canonical imputation estimator proposed by Athey et al. (2020).

Our method's robustness was examined across different scenarios, including settings with and without confounding, as well as cases with identical and varying supports of pre-treatment covariates between experimental and observational samples. Furthermore, in our simulation study, we extended our discussions to scenarios where the support of pre-treatment covariates in the experimental sample is contained within the support of pre-treatment covariates in the observational sample. This setting provided additional insights into the estimators' performance under more structured support conditions, further demonstrating the effectiveness of our proposed procedure.

Author Contributions: Methodology, K.H.; simulation studies, K.H., H.W., L.W., C.L.; writing—original draft preparation, K.H.; writing—review and editing, K.H., H.W., L.W., Y.S., C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The STAR study data can be found at https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10766 (accessed on 1 September 2021).

Conflicts of Interest: The authors declare no conflicts of interest.

References

Achilles, Charles M., Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. 2008. Tennessee's Student Teacher Achievement Ratio (STAR) Project. Available online: https://doi.org/10.7910/DVN/SIWH9F (accessed on 1 September 2021).

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Number 8769 in Economics Books. Princeton: Princeton University Press.

Athey, Susan, Guido Imbens, Jonas Metzger, and Evan Munro. 2020. Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics* 240: 105076. [CrossRef]

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. Generalized random forests. *Annals of Statistics* 47: 1148–78. [CrossRef]

Athey, Susan, Raj Chetty, and Guido Imbens. 2020. Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes. *arXiv*, arXiv:2006.09676.

Hall, Peter, and Joel L. Horowitz. 2005. Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics 33*: 2904–29. [CrossRef]

Horowitz, Joel L. 2011. Applied nonparametric instrumental variables estimation. *Econometrica 79*: 347–94. [CrossRef]

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

James, William, and Charles Stein. 1992. Estimation with quadratic loss. In *Breakthroughs in Statistics: Foundations and Basic Theory*. New York: Springer, pp. 443–60.

Kallus, Nathan, Aahlad Manas Puli, and Uri Shalit. 2018. Removing hidden confounding by experimental grounding. Paper presented at the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Montréal, QC, Canada, December 2–8. Red Hook: Curran Associates Inc., pp. 10911–20.

Newey, Whitney K., and James L. Powell. 2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71: 1565–78. [CrossRef]

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55. [CrossRef]

Rosenman, Evan T. R., Guillaume Basse, Art B. Owen, and Mike Baiocchi. 2023. Combining observational and experimental datasets using shrinkage estimators. *Biometrics* 79: 2961–73. [CrossRef] [PubMed]

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688. [CrossRef]

Wang, Guihua, Jun Li, and Wallace J. Hopp. 2022. An instrumental variable forest approach for detecting heterogeneous treatment effects in observational studies. *Management Science* 68: 3399–18. [CrossRef]

Wu, Lili, and Shu Yang. 2022. Integrative *r*-learner of heterogeneous treatment effects combining experimental and observational studies. Paper presented at the First Conference on Causal Learning and Reasoning, Eureka, CA, USA, April 11–13. PMLR 177:904–926.

Xie, Yu, Jennie E. Brand, and Ben Jann. 2012. Estimating heterogeneous treatment effects with observational data. *Sociological Methodology* 42: 314–47. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.