

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

Kevin Han

January 2021

## 1 Introduction

When estimating treatment effects, the golden 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, the key to validate the use of observational data is hard to verify 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 method 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 a simple estimator to estimate the average treatment effect of interest using both the observational data and the experimental data.

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 concrete solution to each extension. Section 5 compare several different methods through simulations.

## 2 Setup

Suppose we want to estimate the treatment effect of an intervention on some primary outcome  $Y^P \in \mathbb{R}$ . For each unit  $i$  in the observational study, along with the treatment assignments  $W_i$ , the outcome  $Y_i^P$ , we also observe another surrogate outcome  $Y_i^S \in \mathbb{R}$  and record a number of pre-treatment covariates  $X_i$ . Here, the surrogate outcome  $Y^S$  can be any variable that change after treatment. In this paper, we mainly discuss the case that  $Y^S$  is one-dimensional, but our method can be generalized naturally to multi-dimensional surrogate outcome. If the unconfoundedness assumption is satisfied, i.e.

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

then either IPW estimator or AIPW estimator suffices for our propose. However, there exist many settings under which researchers do not believe unconfoundedness holds, hence makes estimating treatment effect using only the observational study impossible. To this end, we assume that there is another prior study on the surrogate outcome  $Y^S$  such that unconfoundedness holds. Typically, this can be a small-scale randomized experiment on the surrogate outcome. In summary, we assume that we have two samples: the observational sample and the experimental sample. There is a  $(X_i, W_i, Y_i^S, Y_i^P)$  tuple associated with every unit  $i$  in our observational sample and a  $(X_i, W_i, Y_i^S)$  tuple associated with every unit  $i$  in the experimental sample. The experimental sample size  $N_E$  is considered to be much smaller than the observational sample size  $N_O$ . We are interested in the quantity

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

where  $G_i$  is the indicator function of which sample unit  $i$  belongs to. We note by passing that this is exactly the same setup as in Athey et al. [2020].

### 3 A simple estimator

We develop our simple estimator in this section. To be able to point-identify the ATE of  $Y^P$ , we assume the following structural model of  $Y^P$ :

$$Y_i^P = f(X_i, Y_i^S, \epsilon_i), \quad \epsilon_i \perp\!\!\!\perp X_i, Y_i^S, \quad (1)$$

i.e. all the effect of treatment on the primary outcome is mediated through the surrogate outcome. Therefore, the surrogate outcome together with pre-treatment covariates determine the primary outcome. Now,  $\tau^P$  is identifiable.

To see this, define

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

and

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

then  $\mathbb{E}[Y_i^P(w)]$  is just  $\mathbb{E}[\mu(X_i, Y_i^S(w))]$ . The joint distribution of  $X_i$  and  $Y_i^S(w)$  is identifiable from the experimental sample because of unconfoundedness. There is a concrete model that we know well:  $Y_i^P = \rho Y_i^S + f(X_i) + \epsilon_i$  where  $\epsilon_i$  is independent with  $Y_i^S$  and  $X_i$ . For such model, we can use Robinson residual-in-residual method to estimate  $\rho$  and the final estimate of the ATE would be consistent. For the general case, we can estimate the  $\tau^P$  as follows:

1. Regress  $Y^P$  on  $Y^S$  and  $X$  to obtain an estimate of  $\mu$ ,  $\hat{\mu}$ .
2. Estimate the conditional average treatment effect  $\tau(x)$  on the surrogate outcome  $Y^S$ , obtain an estimate of  $\tau$ ,  $\hat{\tau}$ .
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. Now we can estimate  $\mathbb{E}[Y_i^P(1)]$  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. Now we can estimate  $\mathbb{E}[Y_i^P(0)]$  by  $\frac{1}{N_O} \sum_{i=1}^{N_O} \hat{\mu}(X_i, \hat{Y}_i^S(0))$ .
7. The final estimate would be  $\hat{\tau}^P = \frac{1}{N_O} \sum_{i=1}^{N_O} \hat{\mu}(X_i, \hat{Y}_i^S(1)) - \frac{1}{N_O} \sum_{i=1}^{N_O} \hat{\mu}(X_i, \hat{Y}_i^S(0))$ .

With the above procedure, to estimate the ATE on the primary outcome, we only need one model for the conditional response function  $\mu$  and one model for CATE estimation. In the next section, we will discuss different variants of the procedure above in 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 will also discuss the setting where we drop the unconfoundedness assumption on experimental sample. In fact, as long as the conditional average treatment effect  $\tau$  is identifiable, unconfoundedness is not necessary.

### 4.1 Different support of pre-treatment covariates

The first scenario that we consider is the setting in Kallus et al. [2018] where the support of pre-treatment covariates in the experimental sample is different from the support of pre-treatment covariates in the observational sample. This is usually the case in practice since the experimental sample typically comes from historical data and we cannot guarantee that the experimental study and the observational study are targeting exactly the same population. Under this setting, if we only use the experimental sample to estimate the conditional average treatment effect, we need to extrapolate on the observational sample. Such extrapolation will be more problematic if the sample size of the experimental sample is much smaller compared to the sample size of the observational sample. Therefore, for our propose, we should calibrate our conditional average treatment estimate on the experimental sample. Kallus et al. [2018] noticed that if we define  $e^E(x) = \mathbb{P}(W_i = 1|X_i = x, G_i = E)$  and  $q^E(X_i) = \frac{W_i}{e^E(X_i)} - \frac{1-W_i}{1-e^E(X_i)}$ , then

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

Define  $\omega(x)$  to be  $\mathbb{E}[Y_i|W_i = 1, X_i = x, G_i = O] - \mathbb{E}[Y_i|W_i = 0, X_i = x, G_i = O]$ , then the above observation motivates the following procedure to estimate the conditional average treatment effect of the surrogate outcome on the observational sample:

1. Run any CATE algorithm on the observational sample, obtain  $\hat{\omega}$ .

2. Solve the following optimization problem to obtain  $\hat{\theta}$ :

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

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

Now we can use the above estimate of  $\tau$  for the estimator described in Section 3. Essentially, the idea here is to use a loss function to estimate the difference between ill-posed target  $\omega$  and the true quantity of interest  $\tau$ . A more general version can be obtained if we do not fit a linear function but a non-parametric function of  $X_i$ .

## 4.2 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.

### 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:

$$\begin{aligned} Y_i^S &= \alpha^T X_i + W_i \tau + \epsilon_i, & \epsilon_i &\perp\!\!\!\perp Z_i \\ W_i &= \beta^T X_i + Z_i \gamma + \xi_i. \end{aligned}$$

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

### 4.2.2 Nonparametric IV

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

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

This is a special case of the more general nonparametric instrumental variable model [Newey and Powell, 2003, Hall and Horowitz, 2005, Horowitz, 2011]. Here, to estimate  $\tau$ , 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,

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

so

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

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

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

for any  $u \in [0, 1]$  where the expectation on 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

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

then we arrive at the following operator equation

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

We can estimate  $\tau + g$  using Hall-Horowitz estimator. Similarly, we have another operator equation where we only have  $g$  by conditioning on  $W = 0$ . With that equation, we are able to estimate  $g$ . Then we can estimate  $\tau$  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] aims to address the general nonparametric IV problem while we only care about  $\tau(x)$ .

With our structural model assumption, Athey et al. [2019] propose the Generalized Random Forests (GRF) to estimate the conditional average treatment effect  $\tau$ . We recommend to use GRF for estimating  $\tau$ . In fact, one advantage of using GRF is that it can be generalized to the setting where  $W$  is no longer binary but a real number.

### 4.3 IV setting with different support of pre-treatment covariates

In this section, we combine our two considerations above. We want to address the setting where we have different support of pre-treatment covariates and a nonparametric instrumental variable model for the experimental sample. We first note that if we let

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

Then

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

Therefore, we can write

$$\tau(x) = \arg \min_{\tau: \mathcal{X} \rightarrow \mathbb{R}} \mathbb{E}[(\tau(x)[\gamma(x) - e(x)\pi(x)] - [m(x) - \mu(x)\pi(x)])^2].$$

It is possible to directly estimate  $\tau$  with the above loss function but we found that it does not work well when we have multi-dimensional pre-treatment covariates as we need to estimate many nuisance parts and the errors may aggregate. However, this loss defining property of  $\tau$  motivates the following procedure (which we abbreviate by Kallus IV):

1. Run any CATE estimation algorithm  $\mathcal{Q}$  on  $\{X_i, W_i, Y_i^S\}_{i=1}^m$  to get an estimate  $\hat{\omega}$ .
2. Solve the following optimization algorithm on the experimental sample:

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

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

Essentially we are adapting the procedure in Kallus et al. [2018] with a different objective function when estimating  $\theta$ . Similar to our remark in the unconfounded case, we can actually fit a non-parametric function of  $X_i$  instead of a linear function. However, we found that this will give us rather unstable estimates when we have many covariates.

## 5 Simulations

In the previous sections, we outlined a procedure to estimate the average treatment effect of the primary outcome given prior information in the experimental sample and considered three scenarios in which we can utilize our procedure described in Section 3. In this section, we compare several estimators through simulations. In particular, we hope to compare our procedure

with the canonical imputation estimator in Athey et al. [2020] when we have an unconfounded experimental sample.

We consider two settings: there is no confounding in the experimental sample (i.e., we have either a randomized experiment or an unconfounded experiment) and there is confounding (we assume a nonparametric IV model for the experimental sample). For each setting, we consider two subcases: 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 the support of the pre-treatment covariates in the experimental sample is not the same as the support of pre-treatment covariates in the observational sample (but they do overlap). When there is no confounding, we compare three estimators: the imputation estimator in Athey et al. [2020], our estimator with  $\tau(x)$  estimated by generalized random forest and our estimator with  $\tau(x)$  estimated by the approach in Kallus et al. [2018]. When there is confounding, both the imputation estimator and the approach in Kallus et al. [2018] are no longer valid as they require the experimental sample to be unconfounded. Hence, we will compare two estimators: our estimator with  $\tau(x)$  estimated by generalized random forest and our estimator with  $\tau(x)$  estimated by Kallus IV.

We work with the following data generating mechanism:

$$\begin{aligned} X_i &\sim \mathcal{N}(0, I_{p \times p}), & \epsilon_i &\sim \mathcal{N}(0, 1), & Z_i &\sim \text{Binom}(1/3), \\ Q_i &\sim \text{Binom}(1/(1 + e^{-\omega \epsilon_i})), & W_i &= Z_i \wedge Q_i, \\ Y_i^S &= \mu(X_i) + (W_i - 1/2)\tau(X_i) + \epsilon_i. \end{aligned}$$

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  $\xi$  is independent noise. This is the same setting as in the appendix of Athey et al. [2019].

Now, we can adjust several parameters in the data generating mechanism to satisfy different conditions.

1. **Presence of confounding:** we vary  $\omega$  to be either 0 or 1. If  $\omega = 0$ , there is no confounding, otherwise there is confounding and we are in the nonparametric IV model.
2. **Sparsity of the signal:**  $\kappa_\tau \in \{2, 4\}$ .
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,  $X_i \sim \mathcal{N}([1, \dots, 1]^T, I_{p \times p})$  in the observational sample.

$\omega$	$\kappa_\tau$	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 1: Simulation results for  $\omega = 0$

Here, 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 interested in the treatment effect on  $Y^P$ . We compare different methods based on mean squared error (MSE). To calculate MSE, we use Monte Carlo method to estimate the true value of ATE and generate 200 realizations.

Table 1 and 2 show the simulation results. We see that when we have identical support of pre-treatment covariates, GRF performs better than the other two methods regardless of confounding issue. This makes sense since when the support does not change, we do not actually need to extrapolate, hence the Kallus method won't improve much. When the support is different, generally Kallus and Kallus IV are also competitive. In fact, when there is confounding, Kallus IV performs better than GRF.

To further investigate the case of different support, we change the above setting slightly. Now we assume that when the support is not identical, the support of pre-treatment covariates of the experimental sample will be contained in the support of pre-treatment covariates of the observational sample (instead of just overlap). Specifically,

- 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 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.

## 6 A real data example

In this section, we investigate the performance of our procedure on a real dataset. We utilize the famous Tennessee STAR study [Achilles et al., 2008]. 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, student home location. We focus on two outcomes: average grade in



$\omega$	$\kappa_\tau$	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 2: Simulation results for  $\omega = 1$

$\omega$	$\kappa_\tau$	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

Table 3: Simulation results, inclusion of the support

$n_{\text{exp}}$	$n_{\text{obs}}$	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 4: STAR study simulation

$n_{\text{exp}}$	$n_{\text{obs}}$	GRF	Imputation	AIPW	$\tau$
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

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

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 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 do the following:

1. Use  $\mathcal{D}_{gt}$  to calculate the average treatment effect of average grade in year 3. This estimate  $\tau_{gt}$  will be viewed as the ground truth.
2. Repeat the following steps 500 times.
3. Sample  $n_{\text{exp}}$  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 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  $\tau_{gt}$  based on  $\mathcal{D}_E$  and  $\mathcal{D}_O$ .
6. Compare based on mean squared error (MSE).

We will only compare GRF and imputation estimator 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 all 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.

## 7 Conclusion

In this paper, we proposed a simple procedure to estimate the average treatment effect of the primary outcome in observational study by utilizing an experimental study for the surrogate outcome. We showed that our procedure can be applied in many settings so long as we can estimate the conditional average treatment effect of the surrogate outcome. We compared several methods through simulations and showed that our procedure gives better estimate in terms of mean square error than the canonical imputation estimator in Athey et al. [2020].

## 8 Acknowledgement

The author thanks Kevin Guo and Guido Imbens for valuable discussions.

## References

- C. Achilles, H. P. Bain, F. Bellott, J. Boyd-Zaharias, J. Finn, J. Folger, J. Johnston, and E. Word. Tennessee’s Student Teacher Achievement Ratio (STAR) project, 2008. URL <https://doi.org/10.7910/DVN/SIWH9F>.
- J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Number 8769 in Economics Books. Princeton University Press, October 2009. URL <https://ideas.repec.org/b/pup/pbooks/8769.html>.
- S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 04 2019. doi: 10.1214/18-AOS1709. URL <https://doi.org/10.1214/18-AOS1709>.
- S. Athey, R. Chetty, and G. Imbens. Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes. *arXiv e-prints*, art. arXiv:2006.09676, 2020.
- S. Athey, G. Imbens, J. Metzger, and E. Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations, 2020.

- P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics*, 33(6):2904–2929, 12 2005. doi: 10.1214/009053605000000714. URL <https://doi.org/10.1214/009053605000000714>.
- J. L. Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394, 2011. doi: <https://doi.org/10.3982/ECTA8662>. URL <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA8662>.
- N. Kallus, A. Manas Puli, and U. Shalit. Removing Hidden Confounding by Experimental Grounding. *arXiv e-prints*, art. arXiv:1810.11646, 2018.
- W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003. doi: <https://doi.org/10.1111/1468-0262.00459>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00459>.