Tutorial

MDM Medical Decision Making

Medical Decision Making
1–12

DOI: 10.1177/0272989X211027181
journals.sagepub.com/home/mdm
SAGE

Augmented Inverse Probability Weighting and the Double Robustness Property

Christoph F. Kurz

This article discusses the augmented inverse propensity weighted (AIPW) estimator as an estimator for average treatment effects. The AIPW combines both the properties of the regression-based estimator and the inverse probability weighted (IPW) estimator and is therefore a "doubly robust" method in that it requires only either the propensity or outcome model to be correctly specified but not both. Even though this estimator has been known for years, it is rarely used in practice. After explaining the estimator and proving the double robustness property, I conduct a simulation study to compare the AIPW efficiency with IPW and regression under different scenarios of misspecification. In 2 real-world examples, I provide a step-by-step guide on implementing the AIPW estimator in practice. I show that it is an easily usable method that extends the IPW to reduce variability and improve estimation accuracy.

Highlights

- Average treatment effects are often estimated by regression or inverse probability weighting methods, but both are vulnerable to bias.
- The augmented inverse probability weighted estimator is an easy-to-use method for average treatment effects that can be less biased because of the double robustness property.

Keywords
double robustness, propensity score, regression, simulation study

Date received: December 1, 2020; accepted: May 27, 2021

Estimating treatment effects is central to health economic practice to evaluate policy interventions and medical trials. Ideally, the effects caused by treatments are investigated in experiments that randomly assign individuals to treatment or control, thereby ensuring that comparable groups are compared under competing treatments, but many experiments with humans are infeasible or unethical.[1] Therefore, most analyses rely on observational data.

In this article, I discuss an estimator for average treatment effects (ATEs) known as the augmented inverse propensity weighted (AIPW) estimator. Although the AIPW has been known for 20 y, most analyses in health and social sciences rely on the traditional inverse propensity weighted (IPW) estimator or the regression estimator.[2]

The AIPW estimator involves 2 basic steps: first, fitting a propensity score model (i.e., the estimated probability of treatment assignment conditional on observed baseline characteristics), and second, fitting 2 models that

Corresponding Author
Munich School of Management and Munich Center of Health Sciences, Ludwig-Maximilians-Universität Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany
Email: kurz@bwl.lmu.de.

estimate the outcome under treatment and control conditions. Each outcome is then weighted by the propensity score from the previous step to produce a weighted average of the 2 outcome models.

I illustrate that the AIPW estimator has a property called "double robustness," meaning that it is consistent (i.e., it converges in probability to the true value of the parameter) for the ATE if either the propensity score model or the outcome model is correctly specified.[3] To demonstrate the accuracy of the AIPW estimator for the ATE, I conduct a Monte Carlo simulation study to compare it to the IPW estimator and a regression estimator and provide real-world examples on how to apply it. I show that this is an easy-to-use estimator that can be more accurate under different scenarios of misspecification (i.e., when treatment assignment or outcome model are uncertain).

Causal Effects Framework

Throughout this article, I will stick to the causal effects framework established by Neyman[4] and Rubin.[5] Suppose N individuals, indexed by $n = 1, \ldots, N$ and randomly sampled from some population, a binary treatment assignment $A \in \{0, 1\}$, $\boldsymbol{X}_n$ a set of observed covariates, and Y_n a continuous outcome. In this setting, $Y_n^{(1)}$ is the outcome that we would observe if individual n had received treatment and $Y_n^{(0)}$ otherwise. These quantities are often referred to as the "potential" outcomes, given that we never observe both $Y_n^{(1)}$ and $Y_n^{(0)}$ for the same individual:

$$Y_n^{(A_n)} = \begin{cases} Y_n^{(1)} & \text{if } A_n = 1 \\ Y_n^{(0)} & \text{if } A_n = 0. \end{cases}$$

In other words, for a binary cause with 2 causal states and associated potential outcome variables $Y_n^{(1)}$ and $Y_n^{(0)}$, a corresponding causal exposure variable, A_n, is specified that takes on 2 values: A_n is equal to 1 for members of the population who are exposed to the treatment state and equal to 0 for members of the population who are exposed to the control state. Exposure to the alternative causal states is decided through a specific mechanism, usually an individual's decision to join one state or another, an outside actor's decision to assign individuals to one state or another, a planned random allocation carried out by an investigator, or any combination of these alternatives.[6] The causal effect of the treatment can be represented by the difference in potential outcomes, $Y^{(1)} - Y^{(0)}$. We assume that the actual observed outcome is connected to the potential outcomes through

$$Y_n = Y_n^{(1)} A_n + Y_n^{(0)}(1 - A_n).$$

This equation implies that one can never observe the potential outcome under the treatment state for those observed in the control state, and one can never observe the potential outcome under the control state for those observed in the treatment state. The fact that we are missing either $Y_n^{(1)}$ or $Y_n^{(0)}$ for every observation is sometimes called the "fundamental problem of causal inference."[7,8]

The probability distribution of $Y_n^{(A_n)}$ represents how outcomes in the population would turn out if everyone received treatment ($A = 1$) or control ($A = 0$), with means $\mathbb{E}\left[Y_n^{(1)}\right]$ and $\mathbb{E}\left[Y_n^{(0)}\right]$, respectively.

Because calculating individual-level causal effects is usually difficult, we concentrate on estimating carefully described aggregate causal effects. Where the difference in possible outcomes is used to describe the individual-level causal effect, aggregate causal effects are usually defined as averages of these individual-level effects. The broadest possible average effect is the ATE in the population as a whole. To have a causal interpretation for the ATE,

$$\tau_{ATE} = \mathbb{E}\left[Y_n^1 - Y_n^0\right],$$

we need to require restrictions on the data-generating distribution. First, we assume that the stable unit treatment value assumption (SUTVA)[9] holds, such that the treatment status of a given individual does not affect the potential outcomes of any other individual. It is a basic assumption of causal effect stability that can often be enforced through careful study design.[10] See Morgan and Winship[6] for examples of possible violations of the SUTVA assumption.

Second, we assume exchangeability (also called "ignorability," "exogeneity," "unconfoundedness," or "selection on observables") given $\boldsymbol{X}$, that is, all causes of both the treatment and the outcome have been measured:

$$\left\{Y_n^{(1)}, Y_n^{(0)}\right\} \perp\!\!\!\perp A_n | \boldsymbol{X}_n, \quad (1)$$

where the symbol $\perp\!\!\!\perp$ denotes statistical independence. This assumption implies that the missing outcome information for the counterfactual treatment status for a given individual can be recovered using individuals with similar observed characteristics. In other words, under

Munich School of Management and Munich Center of Health Sciences, Ludwig-Maximilians-Universität Munich, Munich, Germany (CFK); Institute of Health Economics and Health Care Management, Helmholtz Zentrum München, Neuherberg, Germany. The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The author received no financial support for the research, authorship, and/or publication of this article.

exchangeability, the actual exposure does not predict the counterfactual outcome,[11] and controlling for X makes the treatment unconfounded. Exchangeability means that treatment decisions (or assignments) are not based on what the outcomes might be under various scenarios for treatment. If we undertake to assess the effect of some treatment, we need to make sure that any response differences between the treated and the control group are due to the treatment itself and not to some intrinsic differences between the groups that are unrelated to the treatment.[6,11]

An important third assumption is positivity (or overlap), that is, every individual has a positive nonzero probability of receiving treatment,

$$0 < \mathbb{P}(A_n = 1|\boldsymbol{X}_n) < 1. \tag{2}$$

Equation 1 allows for the identificaiton of treatment effect parameters by conditioning on X,

$$\mathbb{E}\left[Y^{(1)} - Y^{(0)}|X\right] = C(X) = \mathbb{E}[Y|A = 1, X] - \mathbb{E}[Y|A = 0, X],$$

and together with Equation 2 implies

$$\tau_{ATE} = \mathbb{E}[\mathbb{E}[Y|A = 1, X] - \mathbb{E}[Y|A = 0, X]]. \tag{3}$$

The violation of any of these assumptions means causal effects will not be identifiable,[7,12] meaning estimates will be biased for the "true" ATE.

Under SUTVA, the ATE represents the difference in average outcomes induced by shifting the entire population from no treatment (i.e., control) to treatment,[13] but researchers often aim to estimate ATE after accounting for a set of individuals' characteristics, X. This means that when we average over all possible values of X, we get

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y|A = 1, X]] &= \mathbb{E}\left[\mathbb{E}\left[Y^{(1)}|A = 1, X\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[Y^{(1)}|X\right]\right] \\ &= \mathbb{E}\left[Y^{(1)}\right],\end{aligned}$$

and similarly, $\mathbb{E}[\mathbb{E}[Y|A = 0, X]] = \mathbb{E}\left[Y^{(0)}\right]$.

Estimating the Treatment Effect

Traditional causal estimation relies on either fitting a model for the outcome conditional on the treatment and confounding covariates (often called modeling the response surface [RSM][i] or modeling the treatment assignment mechanism).[14] Models to fit the response surface include regression models but also more advanced models from the machine-learning literature.[15,16] According to Equation 3, if we define $\mathbb{E}[Y|A_n = 1, \boldsymbol{X}_n = x] = f(1, x)$ and $\mathbb{E}[Y|A_n = 0, \boldsymbol{X}_n = x] = f(0, x)$, we get the formula for the treatment effect:

$$\tau_{ATE} = \mathbb{E}[f(1, X) - f(0, X)],$$

and all approaches that can flexibly estimate f yield the following estimator of the ATE:

$$\hat{\tau}_{ATE}^{RSM} = \frac{1}{N}\sum_{1=n}^{N}(\hat{f}(1, X_n) - \hat{f}(0, X_n)).$$

This is the empirical analog of Equation 3, and it relies on the same assumptions (Equation 1 and Equation 2). It is also equivalent to estimating the contrast function $C(X)$ at each of the observed data points:

$$\hat{C}_{RSM}(X_n) = \hat{f}(1, X_n) - \hat{f}(0, X_n).$$

These outcome models are basically used to estimate a contrast between what would happen if every observation were put in the control group and what would happen if every observation were put in the treatment group.[17]

However, modeling the response surface has a major disadvantage: it requires that the postulated model f is correct, that is, f must capture the true relationship between Y, A, and $\boldsymbol{X}$. For example, regression analysis cannot reliably adjust for differences in observed covariates when there are substantial differences in the distribution of these covariates in the 2 groups.[18] Estimation over such nonoverlapping ranges may massively underestimate the uncertainty in this estimator.[19]

For these reasons, researchers opted for methods that model treatment assignment instead of (regression) models for the outcome.[20] IPW removes confounding by creating a "pseudo-population" in which the treatment is independent of the measured confounders.[21] When we define the propensity score as the conditional probability of receiving treatment given covariate values,

$$e(X) = \mathbb{E}[A = 1|X],$$

we can rewrite the treatment effect as

$$\tau_{ATE} = \mathbb{E}\left[\frac{AY}{e(X)} - \frac{(1 - A)Y}{1 - e(X)}\right]. \tag{4}$$

See Appendix A.

Again, the IPW estimator can be obtained by the contrast function

$$\hat{C}_{IPW}(X_n) = \frac{A_n Y_n}{\hat{e}(X_n)} - \frac{(1-A_n)Y_n}{1-\hat{e}(X_n)},$$

and the IPW estimator for the ATE is

$$\hat{\tau}_{ATE}^{IPW} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{A_n Y_n}{\hat{e}(X_n)} - \frac{(1-A_n)Y_n}{1-\hat{e}(X_n)}\right).$$

The fundamental difference between modeling the response surface and approaches using propensity scores (including IPW) is that the former models the relationship between a covariate and the outcome, whereas the latter models the relationship between the covariate and the treatment assignment. The IPW estimator is constructed by estimating each individual's propensity score and then weighting the observation for that individual by the inverse of this estimated probability. That is, participants in the treatment condition receive a weight of $1/\hat{e}(X_n)$, and participants in the control condition receive a weight of $1/(1-\hat{e}(X_n))$. Because of the exchangeability assumption, we can generate an unbiased estimate of every potential outcome by reweighting each sample by the inverse probability of that sample receiving the treatment we observed.

If the propensity score is known, then this IPW estimator is unbiased. Therefore, $e(x)$ must be the true propensity score for this estimator to be consistent.[22] In addition, the IPW has poor small sample size properties when the propensity score gets close to 0 or 1. For example, a unit that receives treatment and very low propensity scores (i.e., is highly unlikely to receive the treatment based on the observed covariates) will provide extreme contributions to the estimate.

AIPW Estimator

For these reasons, the IPW estimator has been improved to combine information about the probability of treatment and predictive information about the outcome variable. Robins et al.[17] augmented the IPW by a weighted average of the outcome model, called the augmented inverse propensity weighted (AIPW) estimator:

$$\hat{C}_{AIPW}(X_n) = \left(\underbrace{\frac{A_n Y_n}{e(X_n)}}_{\text{IPW}} - \underbrace{\frac{A_n - e(X_n)}{e(X_n)} f(1, X_n)}_{\text{Augmentation}}\right) - \left(\underbrace{\frac{(1-A_n)Y_n}{1-e(X_n)}}_{\text{IPW}} - \underbrace{\frac{A_n - e(X_n)}{e(X_n)} f(0, X_n)}_{\text{Augmentation}}\right),$$

$$\hat{\tau}_{ATE}^{AIPW} = \frac{1}{N}\sum_{n=1}^{N}\hat{C}_{AIPW}(X_n).$$

This AIPW is called "doubly robust" because it is consistent as long as either the treatment assignment mechanism or the outcome model is correctly specified. If, for example, the propensity score $e(X_n)$ does a very good job of estimating whether or not the patient will receive the treatment of interest, then $A_n - e(X_n)$ will go to 0 in expectation and AIPW will simplify to the IPW estimator. In the same way, if the propensity score is inaccurate, the estimator reduces to the RSM model. See Appendix B for detailed proofs. If both RSM and propensity score are replaced with their true counterparts, the augmentation term again has expectation zero.

The AIPW is more flexible in that it does not require the same set of covariates X_n to be used in both the propensity score model and the response surface model. The only requirement is that conditional ignorability holds given X_n. In addition, $\hat{C}_{AIPW}$ can be shown to be asymptotically normally distributed, and valid standard errors can be derived by an empirical sandwich estimator[23,24] or bootstrapping,[20,25] It is easy to see that this method augments the IPW to reduce variability and improve estimate efficiency while holding the same assumptions as the IPW. The only drawback of this estimator is that one must estimate a propensity score model and 2 response surface (e.g., regression) models (1 for treatment and 1 for control).

Monte Carlo Simulation

In a simulation study, I compared the performance of the 3 ATE estimators described above: 1) the doubly robust AIPW estimator, 2) the IPW estimator, and 3) the ATE estimator based on RSM using regression. All the data sets in the simulation feature 5 variables (similar to Glynn and Quinn[20]): 3 covariates X_1, X_2, X_3, binary treatment status A, and continuous outcome Y. I drew X_1, X_2, and X_3 from standard normal distributions and treatment status A from a Bernoulli distribution, where the probabilities of $A = 1$ were dependent on the realized X_1, X_2 through the standard normal distribution function $\Phi(\cdot)$, $\Pr(A = 1|X) = \Phi(X_1 + X_2 + 0.4X_1X_2)$. The interaction term adds a small degree of confounding. After X_1, X_2, X_3, and A were generated, I drew the outcome variable Y from a normal distribution with a mean that depends on X_2, X_3, and A and a constant variance of one, $Y \sim \mathcal{N}(1 \cdot A + X_2 + X_3, 1)$. Using this specification, the true treatment effect is exactly 1. Note that the treatment assignment depends on X_1 and X_2, whereas the outcome depends on X_2, X_3, and A. This means that adjusting for X_2 is sufficient to produce a consistent estimate of the ATE of A on Y.[26]

I defined 3 settings to compare the accuracy of AIPW, IPW, and RSM to estimate the treatment effect: 1) both

propensity score and response surface are correctly specified; 2) the propensity score is correct, but the response surface is not; and 3) the propensity score is incorrect but the response surface is correctly specified. The incorrect propensity score is a logistic regression estimate of $A \sim X_1$ (leaving out the confounder X_2), whereas the incorrect response surface model for the outcome estimated $Y \sim X_3$ (again leaving out X_2) using linear regression. I further defined a setting with low number of observations ($N = 300$) and one with a high number of observations ($N = 5000$). All Monte Carlo experiments were repeated 1000 times.

Figure 1 presents the results of the Monte Carlo simulation study. The general pattern is that only the AIPW provides unbiased estimates across all settings. In the setting in which both propensity score and outcome response surface model are consistent, all 3 estimators, AIPW, IPW, and RSM, estimate the ATE very accurately. Here, RSM provides the least variation around the true effect, followed by AIPW. In the setting in which the response surface is incorrectly specified, RSM estimates are severly biased, as expected. AIPW and IPW are unbiased here. In the third setting with incorrectly specified propensity score, IPW is biased while AIPW and RSM remain intact. In all settings, we see a finite sample size effect as N increases from 300 to 5000, producing more accurate ATE estimates.

Application

In this section, I will walk through 2 examples on how to calculate the AIPW in a real-world setting. The first example uses data from the RAND Health Insurance Experiment (HIE). Because these data are based on a randomized experiment, they avoid many problems such as the exchangeability assumptions that observational data have. However, this simple example nicely illustrates that under such "perfect" conditions, AIPW, IPW, and RSM produce similar results. The second example uses data from the National Health Interview Survey (NHIS) to show how the AIPW can be robust to misspecification.

Both examples are implemented in the R programming language,[27] but the AIPW estimator is also available in, for example, the `teffects aipw` function in STATA and as an option for the `causaltrt` function in SAS. For simplicity, these examples include a continuous outcome (see Chernozhukov et al.[28] for a discussion of binary outcomes). The R code is also available as an online supplement that additionally illustrates how to calculate bootstrap confidence intervals.

The RAND HIE

In this example, I calculate the AIPW estimator using data from the RAND HIE. This US study ran from 1974 to 1982 and measured health care costs, among other outcomes, of people randomly assigned to 1 of 14 different health insurance plans. The HIE was motivated primarily by an interest to assess the impact of health insurance on health care costs and health. In a very ambitious attempt, the investigators tried to answer whether free medical care led to better health and lower costs than insurance plans that require the patient to shoulder part of the cost. Participants did not have to pay insurance premiums, although there were a number of cost-sharing clauses in the policies, resulting in significant variations in the amount of insurance they provided. The most generous plan offered free comprehensive treatment, whereas 6 "catastrophic coverage" plans at the other end of the insurance spectrum required families to pay 95% of their health care costs. All other plans were in between these extremes with different amounts of co-payment and coverage. Because of the many small treatment groups spread over 14 insurance plans, most analyses start by grouping subjects who were assigned to similar HIE plans together.[29]

Here, I provide a very simplified analysis to measure the causal effect of a "free" plan versus a "catastrophic" coverage plan on total medical costs. As illustrated, the outcome of interest is total medical spending over 1 y, and I include 5 covariates in the analysis (age, sex, race, education, and number of chronic diseases). First, I load the HIE data (available in the R package sampleSelection) and clean the data set:

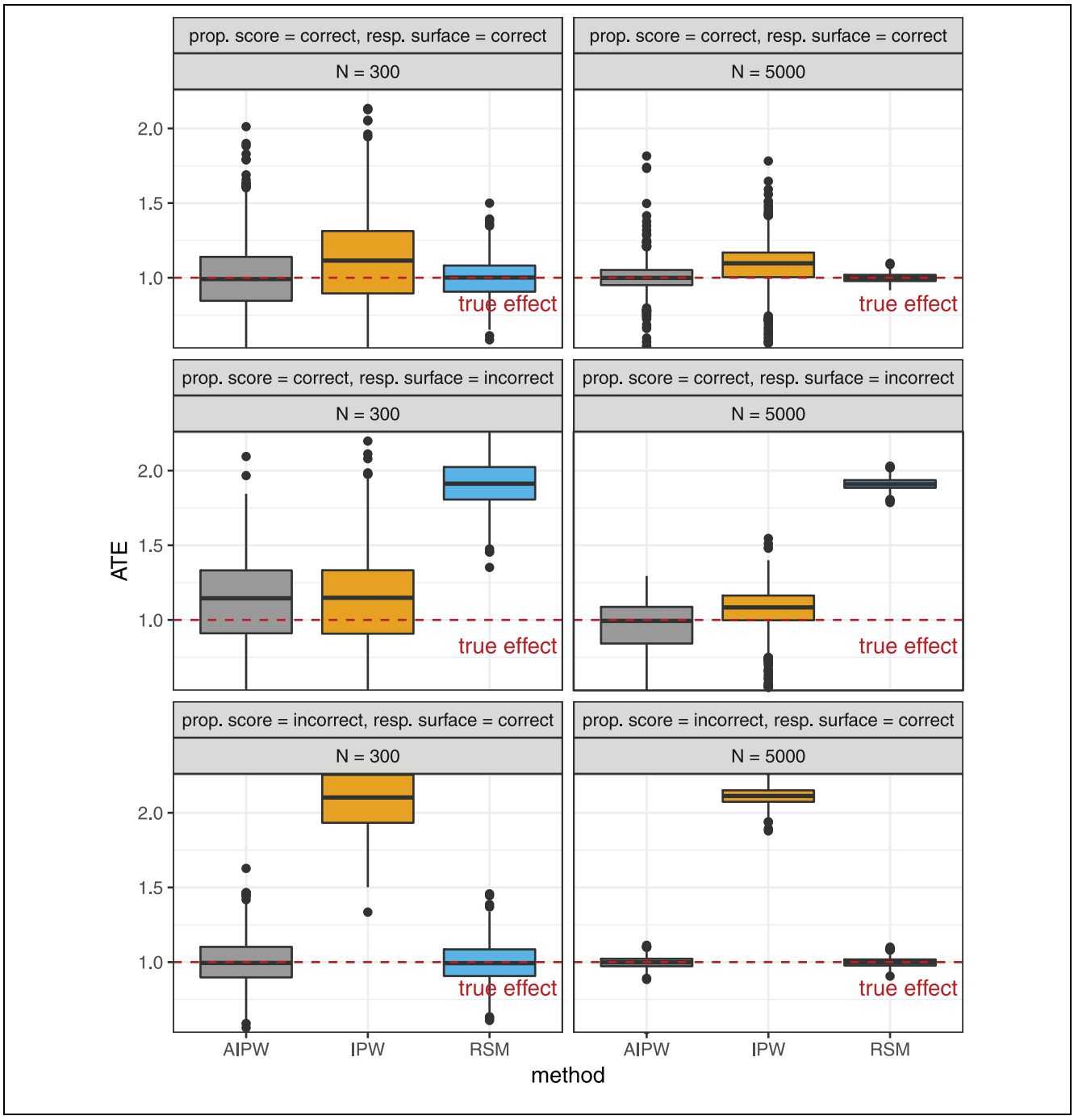

Figure 1 Box plot results of the simulation study to compare the accuracy of augmented inverse propensity weighted, inverse probability weighted, and response surface estimators for the average treatment effect, based on 1000 repetitions in each panel. The dashed red line marks the true treatment effect of 1.0.

```
library(tidyverse)
data("RandHIE", package = "sampleSelection")
rand <- RandHIE %>%
  mutate(plantype = case_when(plan %in% c(2,3,4,14,15,16) ~ "Catastrophic",
                              plan == 11 ~ "Free") %>% as_factor) %>%
  # select only first year observation for each person
  group_by(zper) %>% slice(1) %>%
  select(xage,      # age in years
         female,    # 1 if person is female
         black,     # 1 if race of household head is black
         educdec,   # education of household head in years
         disea,     # number of chronic diseases
         meddol,    # all covered medical expenses
         plantype) %>%
  # remove observations with missing values
  drop_na()
```

The next step is to calculate the propensity score $e(X)$ and the response surface model, together with $f(1,X)$ and $f(0,X)$. For simplicity, I use the same covariates in both propensity model and regression.

```
# calculate the propensity score
ps <- predict(glm(plantype ~ xage + female + black + educdec + disea, family = "
    binomial", data = rand), type = "response") %>% as.numeric

# calc Ey
m <- glm(meddol ~ plantype + xage + female + black + educdec + disea, family = "
    gaussian", data = rand)
# calc ey1
ey1 <- predict(m, newdata = rand %>% mutate(plantype="Free"), type = "response")
    %>% as.numeric
# calc ey0
ey0 <- predict(m, newdata = rand %>% mutate(plantype="Catastrophic"), type = "
    response") %>% as.numeric
```

Finally, I can easily calculate the AIPW using a custom function:

```
# aipw estimator
aipw <- function(a, y, ps, ey1, ey0) {
  mean( ( (a*y)/ps - (1-a)*y/(1-ps) ) - (a-ps)/(ps*(1-ps)) * ( (1-ps)*ey1 + ps*
      ey0 ) )
}
aipw(a = as.numeric(rand$plantype)-1, ps = ps, y = rand$meddol, ey1 = ey1, ey0 =
    ey0)
> 78.31
```

This results in an estimate of US $78.31 increased medical spending for participants in the "free" health insurance plan over a year. For comparison, I calculate the IPW estimator and the response surface model based on linear regression:

```
# ipw estimator
ipw <- function(a, y, ps) {
  mean(a*y/ps - (1-a)*y/(1-ps))
}
ipw(a = as.numeric(rand$plantype)-1, y = rand$meddol, ps = ps)
> 78.18

# regression estimate
mean(ey1 - ey0)
> 77.52
```

I get estimates of 78.18 for IPW and 77.52 for the response surface model. All results are very close, and one can be somewhat sure that both the propensity model and outcome response surface model are correctly specified. This is expected because this example is a randomized experiment in which the risk of misspecification is low.

NHIS 2009

The NHIS is an annual survey of the US population with detailed information on health and health insurance. The research question in this example is whether having health insurance is associated with better health. For that, I used an index that ranges from 1 to 5, where 5 indicates *excellent health* and 1 *poor health* as the outcome of interest. This index is based on self-reports. For simplicity, I restrict to a sample of married 2009 NHIS respondents between 26 and 59 y old who may or may not be insured. Available covariates include age, sex, education, family size, employment status, and household income. The first step is loading and cleaning the data:

```
data("NHIS2009", package = "masteringmetrics")
# install with
# devtools::install_github("jrnold/masteringmetrics", subdir = "masteringmetrics
    ")

nhis <- NHIS2009 %>%
  # only include married adults between 26 and 59 in age and
 # remove single households
 filter(between(age, 26, 59),
        marradult, adltempl >= 1) %>%
 select(age,          # age
        fml,          # female yes/no
        yedu,         # years of education
        famsize,      # family size
        empl,         # employment status
        inc,          # houshold income
        health,       # health status
        uninsured)    # insurance status
```

I then calculate the propensity score. As in the previous example, I include all available covariates in the propensity score model:

```
# calculate the propensity score
ps <- predict(glm(uninsured-1 ~ age + fml + yedu + famsize + empl + inc, family =
     "binomial", data = nhis), type = "response") %>% as.numeric
```

Then I calculate the regression estimates,

```
# calc Ey
m <- glm(health ~ uninsured + age + fml + yedu + famsize + empl + inc, family = "
    gaussian", data = nhis)
# calc ey1
ey1 <- predict(m, newdata = nhis %>% mutate(uninsured=1), type = "response") %>%
    as.numeric
# calc ey0
ey0 <- predict(m, newdata = nhis %>% mutate(uninsured=0), type = "response") %>%
    as.numeric
```

and lastly the AIPW estimator using the previously defined function:

```
aipw(a = nhis$uninsured-1, ps = ps, y = nhis$health, ey1 = ey1, ey0 = ey0)
> -0.073
```

I get an estimate of -0.073, meaning that the absence of health insurance has a slightly negative effect on self-reported health. Compare this with the response surface estimator using regression:

```
# regression estimate
mean(ey1 - ey0)
> -0.044
```

The regression estimate of -0.044 is slightly higher than the AIPW but still very close. Finally, I calculate the IPW for comparison:

```
ipw(a = nhis$uninsured-1, y = nhis$health, ps = ps)
> -0.543
```

The IPW estimate of -0.543 is significantly lower than the previous 2 estimates of AIPW and RSM. What does this mean? It could hint that the IPW estimator is actually biased in this case. The IPW relies solely on the propensity score. Note that to calculate the propensity score, I regressed insurance status on age, sex, education, family size, employment status, and income. The reason this is problematic is because employment status might be reversely causated with insurance; having insurance can be a consequence of being employed. In addition, timing is important. Control covariates should generally be measured before the treatment variable (here, insurance status), so they cannot be changed by the treatment.[29] This is unproblematic for age, sex, and probably education, but income and employment status may be on direct or intermediate causal pathways of insurance status. This is less of an issue in the RSM model with health status as the outcome because the health variable is determined at the time of the survey. This example shows that misspecifications are often difficult to detect, but the AIPW estimate is consistent in such a case.

Discussion and Conclusion

In this article, I have illustrated the AIPW estimator, an extension of the traditional IPW estimator. The most interesting property is that AIPW is "doubly robust" in that it will be consistent for the ATE whenever either the propensity model is correctly specified or the outcome response surface models are correctly specified. This has the advantage that the AIPW can provide more reliable results in a complicated real-world setting where treatment assignment process and outcome model are uncertain.

A simulation showed that the AIPW performs about as well as the IPW or RSW under fully correct specifications. However, the AIPW estimator is more accurate when either IPW or RSM is misspecified. This is by no means a comprehensive simulation study, but it seems

reasonable that applied researchers should consider the AIPW estimator for the ATE when the specification is partially deficient.

Two examples using real data showed that the AIPW is easy to calculate in practice and can be more robust when the propensity score model may be misspecified. However, it is often difficult to specify the 2 key pieces of the AIPW estimator: the propensity score model and the response surface model. For the propensity model, it is generally wise to choose adjustment covariates that both remove bias and produce maximal overlap between the distributions of the estimated propensity scores for the treated and control units.[20] Because the AIPW estimator weights observations in accordance with their observed similary, the propensity score distributions do not need to be perfectly congruent but sufficient overlap is important.[30] Conversely, the set of adjustment covariates for the response surface model should be sufficient to control bias and minimize residual variance. In simulations, Glynn and Quinn[20] found that using a minimally sufficient set of adjustment covariates for the propensity score model and a maximally sufficient set of adjustment covariates for the response surface model can result in lower sampling variability for the AIPW estimator.

Of course, in many situations, the causal process of the data generation is unclear. In such settings, the proper specification decisions are much less clear, and one could use, for example, causal diagrams[31] to identify sets of covariates that can be adjusted for to remove bias. However, in situations in which the researcher has a plausible idea of both contextual factors that motivate treatment assignment and important adjustment covariates for the outcome, this can lead to very sophisticated AIPW estimators. For example, in an analysis by Scott,[30] the author was able to derive key factors for treatment assignment from the literature and at the same time could choose completely different covariates for the outcome model.

It is important to note that even though the AIPW estimator has many advantages, it can still be biased because of unobserved factors that affect both treatment and outcome. In situations in which both propensity score model and response surface model for the outcome are incorrectly specified, its performance might be poor. These limitations, however, are universal drawbacks of using observational data and affect all methods estimating treatment effects.

There is some evidence that the doubly robust estimator can be less efficient than the maximum likelihood estimator with a correctly specified model.[23] Thus, there is a tradeoff to consider between potentially reducing bias at the expense of precision. It is generally advisable to compare effect estimates of several models to rule out possible biases and misspecifications. Still, the AIPW estimator comes at little additional cost but with much greater reliability. Recent research explores how doubly robust effect estimates can be combined with machine learning. Such approaches allow machine learning to be used to weaken parametric modeling assumptions.[28,32,33]

Appendix A

Proof for Equation 4

$$
\begin{aligned}
&\mathbb{E}\left[Y_n^1 - Y_n^0 | \boldsymbol{X}_n = x\right] = \\
&= \mathbb{E}\left[\frac{Y_n^1}{e(x)} e(x) - \frac{Y_n^0}{1-e(x)}(1-e(x)) | \boldsymbol{X}_n = x\right] \\
&= \mathbb{E}\left[\frac{Y_n^1}{e(x)} \mathbb{E}[A|X=x] - \frac{Y_n^0}{1-e(x)}(1-\mathbb{E}[A|X=x]) | \boldsymbol{X}_n = x\right] \\
&= \mathbb{E}\left[\frac{Y_n^1}{e(x)} \mathbb{E}\left[A|Y_n^1, X=x\right] - \frac{Y_n^0}{1-e(x)}(1-\mathbb{E}\left[A|Y_n^0, X=x\right]) | \boldsymbol{X}_n = x\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{Y_n^1 A}{e(x)} | Y_n^1, X=x\right] - \mathbb{E}\left[\frac{Y_n^0(1-A_n)}{1-e(x)} | Y_n^0, X=x\right] | \boldsymbol{X}_n = x\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{A(Y_n^1 A + Y_n^0 \overbrace{A(1-A)}^{=0})}{e(x)} | Y_n^1, X=x\right] - \mathbb{E}\left[\frac{\overbrace{Y_n^1 A - Y_n^1 A^2}^{=0} + Y_n^0 - Y_n^0 A \overbrace{-Y_n^0 A + Y_n^0 A^2}^{=0}}{1-e(x)} | Y_n^0, X=x\right] | \boldsymbol{X}_n = x\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{A(Y_n^1 A + Y_n^0(1-A))}{e(x)} | Y_n^1, X=x\right] - \mathbb{E}\left[\frac{(1-A)(Y_n^1 A + Y_n^0(1-A))}{1-e(x)} | Y_n^0, X=x\right] | \boldsymbol{X}_n = x\right] \\
&= \mathbb{E}\left[\frac{AY}{e(x)} - \frac{(1-A)Y}{1-e(x)} | \boldsymbol{X} = x\right].
\end{aligned}
$$

Appendix B

Double Robustness of the AIPW Estimator

Only the $Y^{(1)}$ case is shown, but it is an analog for $Y^{(0)}$.

Scenario 1: The propensity score model is correct, $e(X) = \mathbb{E}[A|X] = \mathbb{E}\big[A|Y^{(1)}, X\big]$, but the response surface model is incorrect, $f(1,X) \neq \mathbb{E}[Y|A = 1, X]$.

$$\begin{aligned}
&\mathbb{E}[\frac{A - e(X)}{e(X)} Y^{(1)} - f(1,X)] \\
&\quad = \mathbb{E}[\mathbb{E}[\frac{A - e(X)}{e(X)}(Y^{(1)} - f(1,X))|Y^{(1)}, X]] \\
&\quad = \mathbb{E}[(Y^{(1)} - f(1,X))\mathbb{E}[\frac{A - e(X)}{e(X)}|Y^{(1)}, X]] \\
&\quad = \mathbb{E}[(Y^{(1)} - f(1,X))\frac{\mathbb{E}[A|Y^{(1)}, X] - e(X)}{e(X)}] \\
&\quad = \mathbb{E}[(Y^{(1)} - f(1,X))\frac{\mathbb{E}[A|X] - e(X)}{e(X)}] \\
&\quad = \mathbb{E}[(Y^{(1)} - f(1,X))\frac{\overbrace{e(X) - e(X)}^{=0}}{e(X)}] \\
&\quad = 0.
\end{aligned}$$

Scenario 2: The propensity score model is incorrect, $e(X) \neq \mathbb{E}[A|X] = \mathbb{E}\big[A|Y^{(1)}, X\big]$, but the response surface model is correct, $f(1,X) = \mathbb{E}[Y|A = 1, X]$.

$$\begin{aligned}
&\mathbb{E}[\frac{A - e(X)}{e(X)} Y^{(1)} - \mathbb{E}[Y|A = 1, X]] \\
&\quad = \mathbb{E}[\mathbb{E}[\frac{A - e(X)}{e(X)}(Y^{(1)} - \mathbb{E}[Y|A = 1, X])|A, X]] \\
&\quad = \mathbb{E}[\frac{A - e(X)}{e(X)}\mathbb{E}[Y^{(1)} - \mathbb{E}[Y|A = 1, X]|A, X]] \\
&\quad = \mathbb{E}[\frac{A - e(X)}{e(X)}(\mathbb{E}[Y^{(1)}|A, X] - \mathbb{E}[Y|A = 1, X])] \\
&\quad = \mathbb{E}[\frac{A - e(X)}{e(X)}\overbrace{(\mathbb{E}[Y^{(1)}|X] - \mathbb{E}[Y^{(1)}|X])}^{=0}] \\
&\quad = 0.
\end{aligned}$$

Note

i. In the following, I will stick to the RSM terminology.

References

1. Rosenbaum PR. Observational studies. In: *Observational Studies*. New York: Springer; 2002. p 1–17.
2. Lu CY. Observational studies: a review of study designs, challenges and strategies to reduce confounding. *Int J Clin Pract*. 2009;63(5):691–7.
3. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. In: *Proceedings of the American Statistical Association*. Alexandria (VA): American Statistical Association; 1999. p 6–10.
4. Neyman J. On the application of probability theory to agricultural experiments: essay on principles. Section 9. *Stat Sci*. 1923:465–72.
5. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688.
6. Morgan SL, Winship C. *Counterfactuals and Causal Inference*. Cambridge (UK): Cambridge University Press; 2015.
7. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945–60.
8. Gelman A. Causality and statistical learning. *Am J Sociol*. 2011;117(3):955–66.
9. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91(434):444–55.
10. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge (UK): Cambridge University Press; 2015.
11. Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Commun Health*. 2004;58(4):265–71.
12. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge (UK): Cambridge University Press; 2006.
13. Abadie A, Cattaneo MD. Econometric methods for program evaluation. *Annu Rev Econ*. 2018;10:465–503.
14. Austin PC, Laupacis A. A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: a review. *Int J Biostat*. 2011;7(1):1–32.
15. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci U S A*. 2016;113(27):7353–60.
16. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat*. 2011;20(1):217–40.
17. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc*. 1995;90(429):106–21.
18. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcomes Res Methodol*. 2001;2(3–4):169–88.
19. King G, Zeng L. The dangers of extreme counterfactuals. *Polit Anal*. 2006;14(2):131–59.
20. Glynn AN, Quinn KM. An introduction to the augmented inverse propensity weighted estimator. *Polit Anal*. 2010;18(1):36–56.
21. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003;71(4):1161–89.

22. Tsiatis A. *Semiparametric Theory and Missing Data*. New York: Springer Science & Business Media; 2007.
23. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19): 2937–960.
24. Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22(4): 523–39.
25. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86(1):4–29.
26. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37–48.
27. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna (Austria); R Core Team; 2020. Available from: https://www.R-project.org/
28. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*. 2018;21(1):C1–68.
29. Angrist JD, Pischke JS. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton (NJ): Princeton University Press; 2014.
30. Scott T. Does collaboration make any difference? Linking collaborative governance to environmental outcomes. *J Policy Anal Manage*. 2015;34(3):537–66.
31. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669–88.
32. Hernán MA, Robins JM. *Causal Inference*. Boca Raton (FL): CRC Press; 2020.
33. Van Der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostatist*. 2006;2(1).