IMPROVED METHODS FOR CAUSAL INFERENCE AND DATA COMBINATION

BY HENG SHU

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Zhiqiang Tan

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2015

ABSTRACT OF THE DISSERTATION

Improved Methods for Causal Inference and Data Combination

by Heng Shu

Dissertation Director: Zhiqiang Tan

In this dissertation, we develop improved estimation of average treatment effect on the treatment (ATT) which achieves double robustness, local efficiency, intrinsic efficiency and sample boundedness, using a calibrated likelihood approach. Moreover, we consider an extension of two-group causal inference problem to a general data combination problem, and develop estimators achieving desirable properties beyond double robustness and local efficiency. The proposed methods are shown, both theoretically and numerically, to be superior in robustness, efficiency or both to various existing estimators.

In the first part, we review existing estimators on average treatment effect (ATE), mainly based on Tan (2006, 2010). This review provides a useful basis for improved estimation of average treatment effect on the treated (ATT).

In the second part, we propose new methods to estimate the average treatment effect on the treated (ATT), which is of extensive interest in Econometrics, Biostatistics and other research fields. This problem seems to be often treated as a simple modification or extension of that of estimating overall average treatment effects (ATE). But the propensity score is no longer ancillary for estimation of ATT, in contrast with estimation of ATE. We study the efficient influence function and the corresponding semiparametric

variance bound for the estimation of ATT under three different assumptions: a non-parametric model, a correct propensity score model and known propensity score. Then we construct Augmented Inverse Probability Weighted (AIPW) estimators which are locally efficient and doubly robust. Furthermore, we develop calibrated regression and likelihood estimators that are not only doubly robust and locally efficient, but also intrinsically efficient and sample bounded. Two simulations and real data analysis on a job training program are provided to demonstrate the advantage of our estimators compared with existing estimators.

In the third part, we extend our methods to a general data combination problem for moment restriction models (Chen et al. 2008). Similarly, we derive augmented inverse probability weighted (AIPW) estimators that are locally efficient and doubly robust. Moreover, we develop calibrated regression and likelihood estimators which achieve double robustness, local efficiency and intrinsic efficiency. For illustration, we take the linear two-sample instrumental variable problem as an example, and derive all the relevant estimators by applying the general estimators in this specific example. Finally, a simulation study and an Econometric application on a public housing project are provided to demonstrate the superior performance of our improved estimators.

Acknowledgements

I am deeply grateful to my advisor, Professor Zhiqiang Tan. I feel very fortunate to have been working with him and learning a lot from his extensive knowledge and insights. His guidance and great ideas are like navigation light to me, always lead me out of the wrong direction or dead end. Without his supervision and support, the completion of this dissertation would have been impossible.

I also would like to thank Professor Minge Xie, Professor Ying Hung and Professor Tobias Gerhard, for their time and effort to serve on my thesis committee. At the same time, I'd like to thank the graduate director, Professor John Kolassa, who has always been patient and eager to offer guidance no matter in study and life.

Moreover, I am greatly indebted to the Department of Statistics and Biostatistics at Rutgers University for the excellent research environment and continuous financial support throughout my PhD study. The amazing professors here always astonish me by their great intelligence, inspire me by their hard work, move me by their love towards the students. I am so lucky to know lots of lovely peer and friends here. Especially I want to thank Xialu Liu and Xinyan Chen for their continuous encouragements and help. This 5-year PhD study life, with you two being my companions, is definitely one of my most rewarding stages in my life journey.

Finally, I would like to give my deepest gratitude to my parents for their unconditional love and support. I also would like to thank my husband, Yingwei Li. He is always by my side whenever I need support even we are separated by 12-hour driving distance. I can never finish my PhD thesis without his encouragement and support.

Dedication

To My Parents, Yuanhong Shu and Yulian Yang

and

To My Husband, Yingwei Li

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Causal Inference

Drawing inference about average effects of some treatments or actions has applications in various research fields, such as Biostatistics, Economics, political science, sociology and so on. For example, we would like to find out whether a new drug or medical treatment would cure/alleviate a disease effectively. In Economics, we want to set up an experiment to investigate if a training program is useful to help people find better jobs.

Considering the practical cases described above, let's study the mathematical representation. Suppose a simple random sample of n subjects is available from a population under study. The observed data consist of independent and identically distributed observations $\{(Y_i, T_i, X_i) : i = 1, ..., n\}$ of (Y, T, X), where Y is an outcome variable, T is a dichotomous treatment variable ($T = 1$ if treated or $T = 0$ otherwise), and X is a vector of measured covariates. Our objective is to evaluate the effect of the treatment on the outcome Y.

Ideally, we could design and carry out a randomized experiment to figure out the difference. That is to assign the experiment objectives into two groups randomly, then the difference between the sample average simply gives a valid estimate of the treatment effect. The purpose of randomization is to ensure that two treatment groups are comparable with regard to their pre-treatment variables or covariates.

However, in practice, we cannot employ randomized experiments due to some ethical or practical considerations. We thereby have to use observational studies to investigate the treatment or action effect. In most cases, the two groups have differences in various

characteristics due to the self-selection in observational studies. Then simple difference between treatment groups will not represent the effect purely caused by the treatment. How to draw valid inference about treatment effects from observational studies is a major challenge.

1.1.1 Neyman-Rubin Causal Model

Neyman-Rubin causal model, is originated with Neyman's (Neyman, 1923) non-parametric model for randomized experiments where each unit has two potential outcomes, one if the unit is treated and the other if untreated. Rubin (1974) developed the model into a general framework for causal inference with implications for observational studies.

Potential outcomes (Y^0, Y^1) are expressed in the form of counterfactual statements, which state what would be the response under treatment 0 or 1 respectively. Generally, we assume treatment 0 is the control group, while treatment 1 refers to the group receiving the treatment. The consistency between the observed outcome Y and the potential outcome (Y^0, Y^1) is $Y = Y^1$ if $T = 1$ and $Y = Y^0$ if $T = 0$. In another word, $Y = TY^1 + (1 - T)Y^0$.

Causal inference in this framework is a missing data problem because Y^1 and Y^0 are never both observed at the same time.

1.1.2 Treatment Effect (ATE & ATT)

There are two causal effects commonly of interest. The first one is the average treatment effect (ATE), defined as $E(Y^1 - Y^0) = \mu^1 - \mu^0$, with $\mu^t = E(Y^t)$. It is the most commonly studied in statistical and econometric literature. And the other one is the average treatment effect on the treated (ATT), defined as $E(Y^1 - Y^0|T = 1) = \nu^1 - \nu^0$ with $\nu^t = E(Y^t|T = 1)$. From the definition formula, we can see that ATE is defined as the mean difference of two potential outcomes under the active treatment and the control over the entire population, whereas ATT is defined as the mean difference of two potential outcomes over the subpopulation of individuals who received the active treatment.

It is interesting to point out that *these two concepts of effect are actually the same*

in a randomized experiment. In a randomized experiment, the following assumption always holds:

$$T \perp (Y^0, Y^1) \tag{1.1}$$

Then, we can see the definition of ATT reduces to $E(Y^1 - Y^0) = \mu^1 - \mu^0$, the ATE definition.

Sometimes, we may be more interested in the treatment effect on a special subpopulation rather than the whole population in the context of narrowly targeted programs. Similarly as argued in Heckman & Robb (1985) and Heckman et al. (1997), we are often interested in how much the people participating the program can benefit from the program when evaluating the effect of some training program.

Drawing inferences about ATE and ATT is challenging because, in reality, all but one potential outcome are missing for each subject. Nevertheless, under unconfoundedness (i.e., exogeneity) and overlap assumptions, the ATE and ATT are point identifiable from observed data (e.g., Imbens 2004). There is an extensive collection of theory and methods developed for statistical estimation of ATE and ATT under exogeneity. Let Y be an observed outcome, T a treatment indicator, and X a vector of covariates. Semiparametric efficiency bounds for estimation of both ATE and ATT are obtained by Hahn (1998), and can be seen as special cases of semiparametric theory in Robins et al. (1994) and Chen et al. (2008) for conditional mean models with missing data. Asymptotically globally efficient estimators for ATE and ATT are studied by Hahn (1998), Hirano et al. (2003), and Chen et al. (2008) among others, using nonparametric series/sieve estimation on the propensity score, $\pi(X) = P(T = 1|X)$, or the outcome regression function, $m_t(X) = E(Y|T = t, X)$, or both. But the smoothness conditions typically assumed for such methods can be problematic in practical situations with a high-dimensional covariate vector X (Robins & Ritov, 1997).

Alternatively, various methods are developed by using parametric working models on the propensity score $\pi(X)$ or the outcome regression function $m_t(X)$ or both, to achieve desirable properties such as local efficiency, double robustness, and beyond.

This line of research has been well pursued for estimation of ATE (e.g., Robins et al. 1994; Tan 2006, 2010; Cao et al. 2009). See also Kang & Schafer (2007) and its discussion. For an estimator of ATE, double robustness means that the estimator remains consistent if either the propensity score model or the outcome regression model is correctly specified. Local efficiency means that if both the propensity score model and the outcome regression model are correctly specified, then the estimator achieves the semiparametric efficiency bound, which is the same whether the propensity score is known, parametrically modeled, or completely unknown due to the ancillarity of the propensity score for estimation of ATE (Hahn, 1998).

1.2 Data Combination

Generally, Economists always need to draw some inferences regarding a population based on a large enough sample. However, most of the times, they need to combine the samples from different sources.

The reasons why they need to collect different samples could be different; one possibility is a single sample doesn't include all the relevant variables such as the estimation of average treatment effect on the treated (ATT) for program evaluation (Heckman & Robb, 1985; Imbens, 2004), or some variables in the sample are measured with error (e.g., Carroll & Wand 1991), or even if all the relevant variables could be collected from one sample, the limited sample size restricts the accuracy and efficiency of the estimators.

Without loss of generality, suppose we have two independent random samples. The first sample consists the measurements of variables (y, z), and is usually called **primary data** with limited sample size n_1. Let's call it "Data (1)", and use superscript (1) to indicate the data set. While the second one contains measurements of variables (x, z), and is generally called **auxiliary data** with large enough data size n_0. We call it "Data (0)", and use superscript (0) to indicate the data set. The variable y is only available from Data (1), and variable x is only available from Data (0), but z is available from both Data (1) and Data (0).

Since the statistical analysis of interest is considered under the distribution of the primary data, the most straightforward and easiest assumption when we combine different samples is that the auxiliary data has the same distribution as the primary data. However, the common variables across the samples may be measured with different levels of error, and most likely, the common variables may sometimes have quite different distributions across the samples. Different methods have been proposed to augment the auxiliary data in order to ensure the comparability of primary data and auxiliary data, mainly by introducing various forms of weights to balance the discrepancy between the two samples. Wooldridge (2002) introduced the inverse probability weighted M-estimators for cross-section and two-period panel data applications. Hellerstein & Imbens (1999) constructed the weights in least square by imposing the moment restrictions based on auxiliary data to be used in weighted regression analysis.

Hahn (1998) systematically studied the semiparametric efficiency bounds of ATT estimation, a specific example of two-sample combination, under the nonparametric model case and the case when propensity score model is known. Chen et al. (2008) extended the problem to a general two-sample combination framework represented by the moment restrictions and studied the semiparametric efficiency bounds of estimating parameters of interest under the nonparametric model, propensity score model is known or a correctly specified parametric propensity score model. They also propose Generalized Method of Moments (GMM) estimator achieving the variance bounds in these three cases using sieve estimation of conditional expectation. However, their procedure requires nonparametric modelling which is too challenging when the common variables z are high dimensional (Rothe & Firpo, 2013). The recent work done by Graham et al. (2015) provides a locally efficient parametric estimator under the general two-sample combination framework, and it is doubly robust under some assumption regarding the equivalence of the covariate matrix between propensity score (PS) model and outcome regression (OR) model.

1.3 Outline of Thesis

The estimation of ATE has drawn lots of attention and research, but there seems to be much less attention on locally efficient and doubly robust estimation of ATT except for Graham et al. (2015) and Zhao & Percival (2015) recently. There are two possible reasons. On one hand, ATT can often be estimated by a simple modification or extension of estimators of ATE. On the other hand, semiparametric theory for estimation of ATT is complicated by the fact that the propensity score is no longer ancillary Hahn (1998).

In Chapter 2, we review the estimation of average treatment effect (ATE), in order to get a better understanding of the difference and the relationship with ATT. We show that the semiparametric efficiency bounds of estimating ATE remain the same regardless of the information of propensity score, demonstrating the ancillary role of propensity score in estimating ATE. Then we mainly review the methods in Tan (2006, 2010), where the estimators achieve double robustness, local efficiency, intrinsic efficiency and sample boundedness. This review provides us a good basis to carry out the study of ATT estimation.

Improved estimators of average treatment effect on the treated (ATT) are proposed in Chapter 3. According to three different efficient influence functions under different model information of the propensity score, we derive an augmented inverse probability probability (AIPW) estimator which is doubly robust and locally efficient. Moreover, we develop calibrated regression and likelihood estimators that are not only locally efficient and doubly robust, but also intrinsically efficient and sample bounded. By intrinsic efficiency, this estimator achieves greater efficiency than AIPW estimators when a propensity score model is correctly specified but an outcome regression model may be misspecified. We further present data two simulation studies and an Econometric application on evaluating a job training program first studied by LaLonde (1986). All the numerical results demonstrate the advantage of the proposed methods when compared with existing methods.

In Chapter 4, we extend the methods in ATT estimation to deal with data combination problems, where different datasets need to be combined for regression analysis. We formulate the problem in the form of moment estimating equations similar to Chen et al. (2008), and derive doubly robust and locally efficient AIPW estimators directly based on the efficient influence functions. Then calibrated regression and likelihood estimator are proposed to achieve intrinsic efficiency beyond double robustness and local efficiency. Specifically, we show how to use our methods to solve the linear two-sample instrumental variable problem by applying the general estimators to this special case. Finally, we provide one simulation study and one Econometric application on a public housing project. Our improved estimators are found to perform better than existing estimators.

Chapter 2

Review of Methodologies for Estimation of Average Treatment Effect (ATE)

2.1 Set-up

As introduced in Chapter 1, the observed data consist of independent and identically distributed observations $\{(Y_i, T_i, X_i) : i = 1, ..., n\}$ of (Y, T, X), where Y is an outcome variable, T is a dichotomous treatment variable ($T = 1$ if treated or $T = 0$ otherwise), and X is a vector of measured covariates. Potential outcomes (Y^0, Y^1) represent the response under control group $T = 0$ or under active treatment $T = 1$ respectively.

2.2 Assumptions for Identification of ATE

- Unconfoundedness

$$(Y^0, Y^1) \perp T|X \tag{2.1}$$

  This assumption was first proposed in this form by Rosenbaum & Rubin (1983), who named it as "ignorable treatment assignment".

- Overlap

$$0 < P(T = 1|X) < 1 \tag{2.2}$$

  This assumption requests the population across two treatments share the same support of the pretreatment variables to avoid extrapolation. In another word, for each unit in the whole sample, it cannot enter one certain group with probability one; otherwise, it is impossible for us to find out the corresponding potential

outcome if the unit enters the opposite group.

2.3 Two Modelling Approaches

2.3.1 Outcome Regression Model

The first approach is building a regression model for the outcome regression (OR) function, $m_t(x) = E(Y|T = t, X)$

$$E(Y|T = t, X) = m_t(X; \alpha_t) = \Psi\{\alpha_t^T g_t(X)\}, \qquad t = \{0, 1\} \tag{2.3}$$

where $\Psi(\cdot)$ is an inverse link function, $g_0(X)$ and $g_1(X)$ are vectors of known functions of X, and (α_0, α_1) are vectors of unknown parameters. Let $(\hat{\alpha}_0, \hat{\alpha}_1)$ be the maximum quasi-likelihood estimates of (α_0, α_1), and let's denote $\hat{m}_t(X) = \Psi\{\hat{\alpha}_t^T g_t(X)\}$ for $t = 0, 1$.

If the model (2.3) is correctly specified for t=0 and 1, we could construct consistent estimators of μ^1 and μ^0 through

$$\hat{\mu}^1_{\text{OR}} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_1(X_i) \qquad \hat{\mu}^0_{\text{OR}} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_0(X_i) \tag{2.4}$$

Then ATE could be estimated by $\hat{\mu}^1_{\text{OR}} - \hat{\mu}^0_{\text{OR}}$.

2.3.2 Propensity Score Model

Another basic approach is to build a regression model of propensity score (PS) (Rosenbaum & Rubin, 1983), the conditional probability of receiving the treatment, $\pi(X) = P(T = 1|X)$.

$$P(T = 1|X) = \pi(X; \gamma) = \Pi\{\gamma^T f(X)\} \tag{2.5}$$

where $\Pi(\cdot)$ is an inverse link function, $f(x)$ is a vector of known functions, and γ is a vector of unknown parameters. Based on the log-likelihood function, we define the

score function of γ as

$$S_\gamma(T,X) = \left[\frac{T}{\pi(X;\gamma)} - \frac{1-T}{1-\pi(X;\gamma)}\right]\frac{\partial\pi(X;\gamma)}{\partial\gamma} \tag{2.6}$$

Generally, logistic regression is typically used: $\pi(X;\gamma) = [1+\exp\{-\gamma^T f(X)\}]^{-1}$. Then the score function reduces to $S_\gamma = \{T-\pi(X;\gamma)\}f(X)$. Let $\hat{\gamma}$ be the maximum likelihood estimator of γ and let's denote $\hat{\pi}(X) = \pi(X;\hat{\gamma})$ for simplicity. $\hat{\gamma}$ is the parameter which satisfies the score equation $\tilde{E}\{S_\gamma(T,X)\} = 0$, which for logistic regression reduces to

$$\tilde{E}\Big\{[T-\pi(X;\gamma)]f(X)\Big\} = 0 \tag{2.7}$$

where $\tilde{E}(\cdot)$ represents the simple sample average, and we will use this representation in the remainder of this thesis.

ATE can be estimated by matching, stratification, or weighting on the fitted propensity score $\hat{\pi}(X)$, and the details could be found in Imbens (2004). We mainly discuss the inverse probability weighting (IPW) estimators here.

Two standard IPW estimators of μ^t for $t = 0, 1$ are

$$\hat{\mu}^1_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^n \frac{T_i}{\hat{\pi}(X_i)}Y_i, \quad \hat{\mu}^1_{\text{IPW,ratio}} = \sum_{i=1}^n \frac{T_i}{\hat{\pi}(X_i)}Y_i \Big/ \sum_{i=1}^n \frac{T_i}{\hat{\pi}(X_i)} \tag{2.8}$$

$$\hat{\mu}^0_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^n \frac{1-T_i}{1-\hat{\pi}(X_i)}Y_i, \quad \hat{\mu}^0_{\text{IPW,ratio}} = \sum_{i=1}^n \frac{1-T_i}{1-\hat{\pi}(X_i)}Y_i \Big/ \sum_{i=1}^n \frac{1-T_i}{1-\hat{\pi}(X_i)} \tag{2.9}$$

Then ATE could be simply estimated by $\hat{\mu}^1_{\text{IPW}} - \hat{\mu}^0_{\text{IPW}}$ or $\hat{\mu}^1_{\text{IPW,ratio}} - \hat{\mu}^0_{\text{IPW,ratio}}$. If model (2.5) is correctly specified, the IPW estimators are consistent. While even mild mis-specification would lead to poor estimation. For example, the fitted propensity score are close to 0 or 1 for some observations, then the IPW estimates will be very unstable because $\hat{\pi}^{-1}(X_i)$ or $\{1-\hat{\pi}(X_i)\}^{-1}$ is large.

2.4 Semiparametric Theory in Estimation of ATE

We review the semiparametric theory for ATE estimation in Robins et al. (1994), Hahn (1998) and Chen et al. (2008). We will also review their findings on ATT estimation

later in Chapter 3. In Proposition 2.1, we first describe the semiparametric influence functions and efficiency bounds for estimating ATE in three different settings.

Proposition 2.1 *For estimation of $\mu^1 = E(Y^1)$, under the assumption "unconfondedness" and "overlap", based on three different assumptions on the propensity score listed as follows:*

(i) No information is known about the propensity score,

(ii) The propensity score $\pi(X)$ is known,

(iii) The propensity score $\pi(X)$ is unknown but assumed to belong to a correctly specified parametric family $\pi(X;\gamma)$.

In all the three assumptions above, the efficient influence function of estimating μ^1 remains the same, and it is

$$\varphi^1(Y,T,X) = \frac{T}{\pi(X)}[Y - m_1(X)] + m_1(X) - \mu^1$$

Proposition 2.2 *For estimation of $\mu^0 = E(Y^0)$, under the assumption "unconfondedness" and "overlap", based on three different assumptions on the propensity score defined in Proposition 2.1, the efficient influence function of estimating μ^0 remains the same, and it is*

$$\varphi^0(Y,T,X) = \frac{1-T}{1-\pi(X)}[Y - m_0(X)] + m_0(X) - \mu$$

Proposition 2.3 *For estimation of $\mu = E(Y^1 - Y^0)$, under the assumption "unconfondedness" and "overlap", based on three different assumptions on the propensity score defined in Proposition 2.1, the efficient influence function of estimating μ remains the same, and it is*

$$\begin{aligned}\varphi(Y,T,X) =& \frac{T}{\pi(X)}[Y - m_1(X)] - \frac{1-T}{1-\pi(X)}[Y - m_0(X)] \\ &+ m_1(X) - m_0(X) - \mu\end{aligned}$$

We can see that the efficient influence functions and semiparametric variance bounds remain the same in all three cases. This means the propensity score is ancillary in the estimation of ATE, μ; the knowledge of propensity score doesn't reduce the semiparametric variance bound.

Proposition 2.4 *Under the assumption "unconfondedness" and "overlap", the asymptotic variance bound is the same regardless of the information about the propensity score, and the bound is*

$$V = Var\{\varphi(Y,T,X)\} = E\Big[\frac{\sigma_1^2(X)}{\pi(X)} + \frac{\sigma_0^2(X)}{1-\pi(X)} + (m_1(X) - m_0(X) - \mu)^2\Big] \qquad (2.10)$$

where $\sigma_1^2(X) = Var(Y^1|X)$ *and* $\sigma_0^2(X) = Var(Y^0|X)$.

Later in Chapter 3, we will see this property doesn't hold in the estimation of average treatment effect on the treated (ATT). The semiparametric variance bound will be the lowest with the exact knowledge of the propensity score, and the bound is the highest under the nonparametric model of the propensity score.

2.5 Existing Estimators

As discussed above, the estimator $\hat{\mu}^1_{\text{OR}} - \hat{\mu}^0_{\text{OR}}$ is consistent when OR model (2.3) is correctly specified for both $t = 1$ and 0.

Moreover, $\hat{\mu}^1_{\text{IPW}} - \hat{\mu}^0_{\text{IPW}}$ or $\hat{\mu}^1_{\text{IPW,ratio}} - \hat{\mu}^0_{\text{IPW,ratio}}$ is consistent when the PS model (2.5) is correctly specified. But even slight mis-specification of the propensity score model can result in poor estimates, e.g. Kang & Schafer (2007).

It is desirable to design an estimator using both PS model (2.5) and OR model (2.3) to achieve double robustness, which means the estimator is consistent as long as only one of the two models is correctly specified.

A prototypical doubly robust estimator is the augmented inverse-probability-weighted

(AIPW) estimator which is proposed by Robins et al. (1994):

$$\hat{\mu}^1_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}(X_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\hat{\pi}(X_i)} - 1\right)\hat{m}_1(X_i) \tag{2.11}$$

$$\hat{\mu}^0_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\hat{\pi}(X_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-T_i}{1-\hat{\pi}(X_i)} - 1\right)\hat{m}_0(X_i) \tag{2.12}$$

The difference $\hat{\mu}^1_{\text{AIPW}} - \hat{\mu}^0_{\text{AIPW}}$ is exactly the same form as the efficient influence function in Proposition 2.3 with $(\pi(X), m_t(X))$ replaced by $(\hat{\pi}(X), \hat{m}_t(X))$. And this estimator is already shown to be locally efficient in Robins et al. (1994), achieving the semiparametric variance bound when both PS model (2.5) and OR model (2.3) are correctly specified.

Various doubly robust and locally efficient estimators of ATE have been proposed recently. It is interesting to pursue additional good properties beyond these two. In fact, all the locally efficient estimators are equivalent to the first order of each other when both PS and OR model are correctly specified, we need to compare their statistical properties if only one of the two models is correct.

In Tan (2006), he constructed a likelihood estimator based on a nonparametric likelihood approach, and also constructed a regression estimator which is equivalent to the first order of the likelihood estimator. The regression estimator is of a special form to achieve *"intrinsic efficiency"*, meaning if the propensity score model is correctly specified, then this estimator is asymptotically efficient among a class of augmented inverse-probability-weighted estimators that use the same fitted outcome regression (OR) function.

Robins et al. (2007) discussed the concept of population-boundedness or sample-boundedness, which means an estimator will lie within, respectively, the range of all possible values or that of observed values of the outcome. This property could prevent us from obtaining poor estimates even when the inverse probability weights are highly variable.

Tan (2010) built on the work in Tan (2006) and developed a calibrated likelihood estimator possessing all the desirable properties described above, doubly robustness,

locally efficiency, intrinsically efficiency and sample boundedness. In the following sections of this chapter, we will review this approach for the ATE estimation, since this approach provides a basis for the research in the ATT estimation in Chapter 3 and data combination problem we investigated in Chapter 4.

2.6 Regression Estimator

Based on the fitted values $(\hat{\pi}(X), \hat{m}_t(X))$, we define the regression estimator of $\mu^t = E(Y^t)$ as

$$\hat{\mu}^t_{\text{reg}} = \tilde{E}\left(\hat{\eta}_t - \tilde{\beta}_t^{\mathrm{T}}\hat{\xi}_t\right), \quad t = 0 \text{ or } 1 \tag{2.13}$$

where $\tilde{\beta}_t = \tilde{E}^{-1}(\hat{\xi}_t\hat{\zeta}_t^{\mathrm{T}})\tilde{E}(\hat{\xi}_t\hat{\eta}_t)$ with

$$\begin{aligned}
\hat{\eta}_1 &= \frac{T}{\hat{\pi}(X)}Y, & \hat{\eta}_0 &= \frac{1-T}{1-\hat{\pi}(X)}Y,\\
\hat{\xi}_1 &= \left\{\frac{T}{\hat{\pi}(X)} - 1\right\}\frac{\hat{h}(X)}{1-\hat{\pi}(X)}, & \hat{\xi}_0\,(=-\hat{\xi}_1) &= \left\{\frac{1-T}{1-\hat{\pi}(X)} - 1\right\}\frac{\hat{h}(X)}{\hat{\pi}(X)},\\
\hat{\zeta}_1 &= \frac{T}{\hat{\pi}(X)}\frac{\hat{h}(X)}{1-\hat{\pi}(X)}, & \hat{\zeta}_0 &= \frac{1-T}{1-\hat{\pi}(X)}\frac{\hat{h}(X)}{\hat{\pi}(X)},
\end{aligned}$$

and $\hat{h}(X) = \{\hat{h}_1^{\mathrm{T}}(X), \hat{h}_2^{\mathrm{T}}(X)\}^{\mathrm{T}}$, where

$$\begin{aligned}
\hat{h}_1(X) &= \left[\{1-\hat{\pi}(X)\}\hat{v}_1^{\mathrm{T}}(X), \hat{\pi}(X)\hat{v}_0^{\mathrm{T}}(X)\right]^{\mathrm{T}},\\
\hat{h}_2(X) &= \frac{\partial \pi(X;\gamma)}{\partial \gamma^{\mathrm{T}}} = \hat{\pi}(X)\{1-\hat{\pi}(X)\}f(X),\\
\hat{v}_1(X) &= \{1, \hat{m}_1(X)\}^{\mathrm{T}}, \quad \hat{v}_0(X) = \{1, \hat{m}_0(X)\}^{\mathrm{T}}.
\end{aligned} \tag{2.14}$$

where $f(X)$ is the vector of variables in propensity score model (2.5) including the constant 1.

The variables included in $\hat{h}(X)$ are designed with special considerations. We include $\hat{m}_0(X)$ and $\hat{m}_1(X)$ into $\hat{v}_0(X)$ and $\hat{v}_1(X)$ respectively to achieve double robustness and local semiparametric efficiency, in view of the form of AIPW estimator (2.11). The variables in $\hat{h}_2(X)$ are included for achieving intrinsic efficiency, which is justified in the proof of intrinsic efficiency in Tan (2006). Also, constant 1 is included in $\hat{v}_0(X)$

and $\hat{v}_1(X)$ to ensure efficiency gains over the ratio formed IPW estimator $\hat{\mu}^1_{\text{IPW,ratio}}$ and $\hat{\mu}^0_{\text{IPW,ratio}}$ under a correctly specified PS model.

When PS model is correctly specified, $\tilde{E}(\hat{\eta}_t)$ converges asymptotically to μ^t, and ξ_t converges to 0. In another word, we design $\hat{\xi}_t$ as control variates in Monte Carlo integration or auxiliary variables in survey sampling. The name "regression estimator" is used according to the Monte Carlo integration literature, such as Hammersley & Handscomb (1964) and the Survey Sampling literature, such as Cochran (1977). The effect of using control variates is variance reduction, as shown in the following proposition proved in Tan (2006).

Proposition 2.5 *Under suitable regularity conditions (see Tan (2006)), the estimator $\hat{\mu}^t_{reg}$ for μ^t has the following properties for $t = 0, 1$.*

(i) $\hat{\mu}^t_{reg}$ is locally efficient: it achieves the nonparametric efficiency bound (2.10). when both model (2.3) for the corresponding t and model (2.5) are correctly specified.

(ii) $\hat{\mu}^t_{reg}$ is doubly robust: it remains consistent when either model (2.3) for the corresponding t or model (2.5) is correctly specified.

(iii) $\hat{\mu}^t_{reg}$ is intrinsically efficient: if model (2.5) is correctly specified, then it achieves the lowest asymptotic variance among the class of estimators

$$\tilde{E}\left(\hat{\eta}_t - b_t^{\mathrm{T}}\hat{\xi}_t\right) \tag{2.15}$$

where b_t is an arbitrary vector of constants.

Based on the properties above, the estimator of ATE $\hat{\mu}^1_{\text{reg}} - \hat{\mu}^0_{\text{reg}}$ obtain the following desirable properties.

Corollary 2.6 *The estimator $\hat{\mu}^1_{reg} - \hat{\mu}^0_{reg}$ for ATE has the following properties.*

(i) $\hat{\mu}^1_{reg} - \hat{\mu}^0_{reg}$ is locally efficient: it achieves the nonparametric efficiency bound, (2.10). when both model (2.3) for $t = 0, 1$ and model (2.5) are correctly specified.

(ii) $\hat{\mu}^1_{reg} - \hat{\mu}^0_{reg}$ *is doubly robust: it remains consistent when either model (2.3) for* $t = 0, 1$ *or model (2.5) is correctly specified.*

(iii) $\hat{\mu}^1_{reg} - \hat{\mu}^0_{reg}$ *is intrinsically efficient: if model (2.5) is correctly specified, then it achieves the lowest asymptotic variance among the class of estimators*

$$\tilde{E}\left(\hat{\eta}_1 - \hat{\eta}_0 - b_0^{\mathrm{T}}\hat{\xi}_0\right)$$

where b_0 *is an arbitrary vector of constants.*

When the OR model (2.3) for $t = 0$ and 1 are both correctly specified, as discussed in Tan (2006), $\hat{\beta}_t$ converges to a constant vector β_t^* such that

$$\hat{\mu}^t_{\mathrm{reg}} = \tilde{E}(\hat{\eta}_t - \beta_t^{*\mathrm{T}}\hat{\xi}_t) + o_p(n^{-1/2}) = \hat{\mu}^t_{\mathrm{AIPW}} + o_p(n^{-1/2}) \quad \text{for} \quad t = 0, 1 \tag{2.16}$$

This is because $\hat{m}_1(X)$ is a linear combination of variables in $\hat{h}(X)/\{1 - \hat{\pi}(X)\}$ and $\hat{m}_0(X)$ is a linear combination of the variables in $\hat{h}(X)/\hat{\pi}(X)$. And we already know $\hat{\mu}^t_{\mathrm{AIPW}}$ is locally efficient estimator of μ^t from the previous discussion, so μ^t_{reg} is also locally efficient. Moreover, $\hat{\mu}^t_{\mathrm{AIPW}}$ is also known to be doubly robust. Therefore, as long as the OR model (2.3) is correctly specified, $\hat{\mu}^t_{\mathrm{reg}}$ is consistent even if we have a misspecified PS model (2.5). On the other hand, if the PS model is correct, $\tilde{E}(\hat{\xi}_t)$ converges to 0, so $\hat{\mu}^t_{\mathrm{reg}}$ converges to $\hat{\mu}^t_{\mathrm{IPW}}$, hence $\tilde{E}(\hat{\xi}_t)$ is also consistent under a correctly specified PS model (2.5). So we could conclude that $\hat{\mu}^t_{\mathrm{reg}}$ is doubly robust.

A canonical estimator of the optimal choice of b_t in minimizing the asymptotic variance of (2.15) is $\hat{\beta}_t^* = \tilde{E}(\hat{\xi}_t\hat{\xi}_t^{\mathrm{T}})^{-1}\tilde{E}(\hat{\xi}_t\hat{\eta}_t)$, which differs subtly from $\tilde{\beta}_t$. It can be proved that the corresponding estimator, $\hat{\mu}^{t*}_{\mathrm{reg}} = \tilde{E}(\hat{\eta}_t - \hat{\beta}_t^{*\mathrm{T}}\hat{\xi}_t)$, for μ^t is asymptotically equivalent to the first order to $\hat{\mu}^t_{\mathrm{reg}}$ when the PS model is correctly specified. But when PS model is misspecified, $\hat{\mu}^{t*}_{\mathrm{reg}}$ is no longer consistent, even when OR model is correctly specified. While $\hat{\mu}^t_{\mathrm{reg}}$ remains consistent even under a misspecified PS model (2.5). Therefore, this specific form of $\tilde{\beta}_t$ makes this regression estimator not only intrinsic efficiency, but also double robustness.

The IPW estimator $\hat{\mu}^t_{\mathrm{IPW}}(\hat{\pi})$ falls into the class (2.15) with $b_t = 0$ for $t = 0, 1$. The

ratio estimator $\hat{\mu}^t_{\text{IPW,ratio}}(\hat{\pi})$ doesn't obviously belongs to this class (2.15), but when PS model (2.5) is correctly specified, it could be proven to be asymptotically equivalent to the first order to $\tilde{E}\Big[\hat{\eta}_1 - [T/\hat{\pi}(X) - 1]\mu^1\Big]$ and $\tilde{E}\Big[\hat{\eta}_0 - [(1-T)/\{1-\hat{\pi}(X)\} - 1]\mu^0\Big]$ for $t = 1$ and 0 separately, which fall into the class because constant 1 is included in $\hat{h}(X)/\{1-\hat{\pi}(X)\}$ and $\hat{h}(X)/\hat{\pi}(X)$ for $t = 1$ and 0 respectively.

Therefore, by intrinsic efficiency, $\hat{\mu}^t_{\text{reg}}$ is asymptotically as least as efficient as not only $\hat{\mu}^t_{\text{AIPW}}(\hat{\pi}, \hat{m}_t)$, but also $\hat{\mu}^t_{\text{IPW}}(\hat{\pi})$ and $\hat{\mu}^t_{\text{IPW,ratio}}(\hat{\pi})$ for $t = 0, 1$, when PS model (2.5) is correctly specified.

The estimator $\hat{\mu}^t_{\text{AIPW}}(\hat{\pi}, \hat{m}_t)$ belongs to the class of estimators (2.15) since $\hat{m}_1(X)$ and $\hat{m}_0(X)$ are included in $\hat{h}(X)/\{1-\hat{\pi}(X)\}$ and $\hat{h}(X)/\hat{\pi}(X)$ respectively. Therefore, $\hat{\mu}^1_{\text{reg}} - \hat{\mu}^0_{\text{reg}}$ is asymptotically as least as efficient as $\hat{\mu}^1_{\text{AIPW}}(\hat{\pi}, \hat{m}_1) - \hat{\mu}^0_{\text{AIPW}}(\hat{\pi}, \hat{m}_0)$ for estimation of average treatment effect (ATE), when PS model (2.5) is correctly specified.

2.7 Likelihood Estimator

The regression estimator in the last section is already doubly robust, locally and intrinsically efficient. However, a common drawback of regression estimator and AIPW estimators are that they may lie outside either the sample or the population range of observed outcomes. This may be caused by the case that the fitted propensity score $\hat{\pi}(X)$ is close to 0 among the treated or close to 1 among the control group. Tan (2006) proposed a nonparametric likelihood estimator to solve this issue, but it is not doubly robust. Tan (2010) built on the previous work, and developed a calibrated likelihood estimator achieving all the good features described above.

In this section, we will review likelihood estimators of μ^t that are not only doubly robust, locally nonparametric efficient, and intrinsically efficient similarly to the regression estimators, but also sample-bounded in falling within the range of $\{Y_i : T_i = t, i = 1, \ldots, n\}$.

There are two steps to construct the desired likelihood estimators. The first step is to derive locally and intrinsically efficient, but non-doubly robust, likelihood estimators.

The nonparametric likelihood of $(X_i, T_i, Y_i)(i = 1, \ldots, n)$ is

$$\begin{aligned} L_1 \times L_2 = & \prod_{i=1}^{n} \left[\pi(X_i;\gamma)^{T_i}\{1-\pi(X_i;\gamma)\}^{1-T_i}\right] \\ & \times \prod_{i=1}^{n} \left[G_1(\{X_i, Y_i\})^{T_i} G_0(\{X_i, Y_i\})^{1-T_i}\right] \end{aligned} \tag{2.17}$$

where G_t is the joint distribution of (X, Y^t) for $t = 0, 1$. Maximizing L_1 will get the maximum likelihood estimator $\hat{\gamma}$, thereby $\hat{\pi}(X)$.

When considering maximizing L_2, we choose to ignore the fact that G_t, $t = 0, 1$ will induce the same marginal distribution of X, and retain only the constraints

$$\int \hat{h}(X) dG_1 = \int \hat{h}(X) dG_0$$

where $\hat{h}(X)$ is defined exactly the same as (2.14). Furthermore, we require that G_1 is a probability measure supported on $\{(X_i, Y_i) : T_i = 1, i = 1, \ldots, n\}$ and hence $\int dG_1 = 1$. And G_0 is a nonnegative measure with support on $\{X_i : T_i = 0, i = 1, \ldots, n\}$. Maximizing L_2 subject to these constraints leads to the estimators

$$\hat{G}_1(\{X_i, Y_i\}) = \frac{n^{-1}}{\omega(X_i;\hat{\lambda})} \qquad (T_i = 1) \tag{2.18}$$

$$\hat{G}_0(\{X_i, Y_i\}) = \frac{n^{-1}}{1-\omega(X_i;\hat{\lambda})} \qquad (T_i = 0) \tag{2.19}$$

where $\omega(X, \lambda) = \hat{\pi}(X) + \lambda^{\mathrm{T}}\hat{h}(X)$, $\hat{\lambda} = \mathrm{argmax}_{\lambda}\ell(\lambda)$, and

$$\ell(\lambda) = \tilde{E}\big[T \log\{\omega(X;\lambda)\} + (1-T)\log\{1-\omega(X;\lambda)\}\big] \tag{2.20}$$

subject to $\omega(X_i;\lambda) > 0$ if $T_i = 1$ and $\omega(X_i;\lambda) < 1$ if $T_i = 0$ for $i = 1, \ldots, n$. Setting the gradient of $\ell(\lambda)$ to zero shows that $\hat{\lambda}$ is a solution to

$$\tilde{E}\left[\frac{T-\omega(X;\lambda)}{\omega(X;\lambda)\{1-\omega(X;\lambda)\}}\hat{h}(X)\right] = 0. \tag{2.21}$$

Then the resulting estimator of μ^1 and μ^0 are

$$\hat{\mu}^1_{\text{lik}} = \tilde{E}\left\{\frac{TY}{\omega(X;\hat{\lambda})}\right\} \tag{2.22}$$

$$\hat{\mu}^0_{\text{lik}} = \tilde{E}\left\{\frac{(1-T)Y}{1-\omega(X;\hat{\lambda})}\right\} \tag{2.23}$$

However, the estimators $\hat{\mu}^t_{\text{lik}}$ for $t = 0, 1$ are not doubly robust, although they are proved to be locally efficient and intrinsically efficient among the class of estimators with the form (2.15). Tan (2010) proposed calibration on the coefficients in the linear extended propensity score. The process is described as below:

For $t = 0$ or 1, partition $\hat{h}$ as $\hat{h} = (\hat{h}^{\mathrm{T}}_{1t}, \hat{h}^{\mathrm{T}}_{1(t)}, \hat{h}^{\mathrm{T}}_2)^{\mathrm{T}}$ and accordingly λ as $\lambda = (\lambda^{\mathrm{T}}_{1t}, \lambda^{\mathrm{T}}_{1(t)}, \lambda^{\mathrm{T}}_2)^{\mathrm{T}}$, where $\hat{h}_{1t} = \hat{\pi}\hat{v}_0$ or $(1-\hat{\pi})\hat{v}_1$ if $t = 0$ or 1, and $\hat{h}_{1(t)}$ consists of the elements of $\hat{h}_1$ excluding $\hat{h}_{1t}$. Moreover, let's denote $R_t = 1\{T = t\}$, $\hat{\pi}(t, X) = \hat{P}(T = t|X)$, so $\hat{\pi}(t = 1, X) = \hat{\pi}(X)$ and $\hat{\pi}(t = 0, X) = 1-\hat{\pi}(X)$. Similarly, we define $\omega(t, X; \lambda) = 1-\omega(X; \lambda)$ or $\omega(X; \lambda)$ for $t = 0$ or 1 respectively. Define $\tilde{\lambda}^t = (\tilde{\lambda}^{\mathrm{T}}_{1t}, \hat{\lambda}^{\mathrm{T}}_{1(t)}, \hat{\lambda}^{\mathrm{T}}_2)^{\mathrm{T}}$, where $\hat{\lambda}^{\mathrm{T}}_{1(t)}$ and $\hat{\lambda}_2$ are obtained from $\hat{\lambda}$, and $\tilde{\lambda}_{1t}$ is a maximizer of the function

$$\kappa_t(\lambda_{1t}) = \tilde{E}\left[R_t \frac{\log\{\omega(t, X; \lambda_{1t}, \hat{\lambda}_{1(t)}, \hat{\lambda}_2)\} - \log\{\omega(t, X; \hat{\lambda})\}}{1-\hat{\pi}(t, X)} - \lambda^{\mathrm{T}}_{1t}\hat{v}_t(X)\right],$$

subject to $\omega(t, X_i; \lambda_{1t}, \hat{\lambda}_{1(t)}, \hat{\lambda}_2) > 0$ if $T_i = t$ for $i = 1, \ldots, n$. Setting the gradient of $\kappa_t(\lambda_{1t})$ to 0 shows that $\tilde{\lambda}_{1t}$ is a solution to

$$\tilde{E}\left[\left\{\frac{R_t}{\omega(t, X; \lambda_{1t}, \hat{\lambda}_{1(t)}, \hat{\lambda}_2)} - 1\right\}\hat{v}_t(X)\right] = 0. \tag{2.24}$$

For $t = 0, 1$, the resulting estimator of μ^t is

$$\tilde{\mu}^t_{\text{lik}} \quad = \quad \tilde{E}\left\{\frac{R_t Y}{\omega(t, X; \tilde{\lambda}^t)}\right\}$$

The likelihood estimator $\tilde{\mu}^t_{\text{lik}}$ has several desirable properties as follows.

Proposition 2.7 *Under suitable regularity conditions (see Tan (2010)), the estimator $\tilde{\mu}^t_{lik}$ for μ^t has the following properties for $t = 0, 1$.*

(i) $\tilde{\mu}^t_{lik}$ *is sample-bounded: it lies within the range of* $\{Y_i : T_i = t, i = 1, \ldots, n\}$.

(ii) If model (2.5) is correctly specified, then $\tilde{\mu}^t_{lik}$ *is asymptotically equivalent, to the first order, to* $\hat{\mu}^t_{reg}$. *Hence* $\tilde{\mu}^t_{lik}$ *is intrinsically efficient among the class (2.15) and locally nonparametric efficient, similarly as* $\hat{\mu}^t_{reg}$ *in Proposition 4.3.*

(iii) $\tilde{\mu}^t_{lik}$ *is doubly robust, similarly as* $\hat{\mu}^t_{reg}$ *in Proposition 4.3.*

The sample-boundedness of $\tilde{\mu}^t_{\text{lik}}$ holds because $\omega(t, X_i; \tilde{\lambda}^t) > 0$ if $T_i = t$ for $i = 1, \ldots, n$ and $\tilde{E}\{R_t/\omega(t, X_i; \tilde{\lambda}^t)\} = 1$ by Eq. (3.7) with constant 1 is included in $\hat{v}_t(X)$. The double robustness of $\tilde{\mu}^t_{\text{lik}}$ follows mainly due to: $\tilde{E}\{R_t\hat{m}_t(X)/\omega(t, X_i; \tilde{\lambda}^t)\} = \tilde{E}\{\hat{m}_t(X)\}$ by Eq. (3.7) with $\hat{m}_t(X)$ included in $\hat{v}_t(X)$.

The implication of intrinsic efficiency for $\tilde{\mu}^t_{\text{lik}}$ is similar to that for $\hat{\mu}^t_{\text{reg}}$ as discussed in Section 2.6. If the PS model (2.5) is correctly specified while the OR model (2.3) may be misspecified, then $\tilde{\mu}^t_{\text{lik}}$ is asymptotically at least as efficient as $\hat{\mu}^t_{\text{AIPW}}(\hat{\pi}, \hat{m}_t)$ for $t = 0$ or 1, and $\tilde{\mu}^1_{\text{lik}} - \tilde{\mu}^0_{\text{lik}}$ is asymptotically at least as efficient as $\hat{\mu}^1_{\text{AIPW}} - \hat{\mu}^0_{\text{AIPW}}(\hat{\pi}, \hat{m}_0)$.

2.8 Conclusion

In this chapter, we mainly review the semiparametric efficiency theory for estimation of ATE and the existing estimators of ATE. ATE estimation is special due to the ancillary role of propensity score, that is the variance bounds remain the same regardless of the information of propensity score. We presented a doubly robust and locally efficient estimator of ATE by simply using efficient influence functions as estimating functions, and this estimator is just the AIPW estimator proposed by Robins et al. (1994). Moreover, we described the calibrated regression and likelihood estimators discussed in Tan (2006, 2010) in details. The proposed estimators of ATE achieve not only double robustness and local efficiency, but also intrinsic efficiency and sample boundedness. And this work provide us an important basis for the study of ATT estimation in Chapter 3.

Chapter 3

Improved Estimation of Average Treatment Effects on the Treated (ATT): Local Efficiency, Double Robustness, and Beyond

In this chapter, we will introduce a new approach to obtain an improved estimator of ATT. First, we derive augmented inverse probability weighted (AIPW) estimators of ATT that are locally efficient and doubly robust, by directly using efficient influence functions as estimating functions, and then to develop calibrated regression and likelihood estimators that achieve desirable properties beyond local efficiency and double robustness.

There are several interesting phenomena clarified from our work, all different from familiar results for estimation of ATE. First, there are two AIPW estimators achieving local efficiency of different types. If the propensity score and outcome regression models are correctly specified, the first estimator achieves the semiparametric efficiency bound, V_{NP}, calculated when the propensity score is unknown, whereas the second estimator achieves the semiparametric efficiency bound, $V_{\text{SP}}\,(\leq V_{\text{NP}})$, calculated under the parametric propensity score model used. These two estimators are then referred to as locally nonparametric or, respectively, semiparametric efficient.

Second, the locally nonparametric efficient estimator AIPW of ATT is doubly robust, but the locally semiparametric efficient AIPW estimator is generally not. Therefore, it is the efficient influence function calculated under the nonparametric model (i.e., when the propensity score as well as the outcome regression function is unknown) that leads to doubly robust estimation. Incidentally, it can be shown that the doubly robust estimators of ATT in Graham et al. (2015) and Zhao & Percival (2015) are also locally nonparametric efficient.

Third, due to the discrepancy between the locally nonparametric and semiparametric AIPW estimators, a direct application of the techniques in Tan (2006, 2010) and Cao et al. (2009) would fail to yield an improved estimator of ATT that is not only doubly robust and locally nonparametric efficient, but also intrinsically efficient in achieving greater efficiency than AIPW estimators when the propensity score model is correctly specified but the outcome regression model may be misspecified. We show that such improved estimators can still be developed by introducing a simple idea, namely, working with an augmented propensity score model which includes the fitted outcome regression functions as additional regressors.

To illustrate the advantage of the improved estimators, we present two simulation studies and an econometric application related to LaLonde (1986) and subsequent analyses (e.g., Dehejia & Wahba 2002; Smith & Todd 2005a). In contrast with these previous works, we compare the performance of different methods by examining not only the effect or bias estimates (where the experimental treatment or, respectively, control group is compared with a non-experimental comparison group), but also how well the differences between the effect and bias estimates agree with the benchmark estimate (where the experimental control and treatment groups are compared). The latter comparisons are relevant even if the non-experimental group might inherently differ from the cohort on which the experiment was conducted.

3.1 Setup and Classical Estimators

Here we use exactly the same setup as the ATE case. Two causal parameters commonly of interest are the average treatment effect (ATE), defined as $E(Y^1 - Y^0) = \mu^1 - \mu^0$ with $\mu^t = E(Y^t)$, and the average treatment effect on the treated (ATT), defined as $E(Y^1 - Y^0|T = 1) = \nu^1 - \nu^0$ with $\nu^t = E(Y^t|T = 1)$.

While the parameter ν^1 is directly identifiable as $E(TY)/E(T)$, a fundamental difficulty in identification of ν^0 is that Y^0 is missing for treated subjects with $T = 1$. Nevertheless, it is known (e.g., Imbens 2004) that the ν^0 and hence ATT are identifiable from observed data under the two assumptions:

(A1) Unconfoundedness for controls: $T \perp Y^0|X$, i.e., T and Y^0 are conditionally independent given X;

(A2) Weak overlap: $0 \leq P(T=1|X=x) < 1$ for all x.

Assumption (A2) allows that $P(T=1|X=x)$ is 0 for some values x, i.e., subjects with certain covariate values will always take treatment 0.

By the fact that $\nu^1 = E(TY)/E(T)$, a consistent, nonparametric estimator of ν^1 is $\hat{\nu}^1_{\rm NP} = n_1^{-1}\sum_{i=1}^n T_i Y_i$, where $n_1 = \sum_{i=1}^n T_i$ and $n_0 = n - n_1$ are the sizes of treated and untreated groups respectively in the sample. However, modeling (or dimension-reduction) assumptions, in addition to (A1)–(A2), are, in general, needed to obtain consistent estimation of ν^0 and ATT from finite samples with high-dimensional X. There are broadly two modelling approaches as follows.

One approach is to build a regression model for the outcome regression (OR) function defined in (2.3). If model (2.3) is correctly specified for $t = 0$ or 1, then a consistent estimator for ν^t is $\hat{\nu}^t_{\rm OR} = n_1^{-1}\sum_{i=1}^n T_i\,\hat{m}_t(X_i)$. The ATT can be estimated by $\hat{\nu}^1_{\rm OR} - \hat{\nu}^0_{\rm OR}$. In the special case where $\Psi(\cdot)$ is the identity link and parallel regression functions are assumed for the two treatment groups, i.e., $E(Y|T=t,X) = \alpha_{1,t} + \alpha_{(1)}^{\rm T} g_{(1)}(X)$ with $g_{(1)}(X)$ excluding 1, the ATT can be directly estimated as $\alpha_{1,1} - \alpha_{1,0}$.

An alternative approach is to build a regression model for the propensity score (PS) defined in (2.5). Let's use $\tilde{E}(\cdot)$ denote a sample average, for example, $\tilde{E}(T) = n_1/n$. Then ν^0 and ATT can be estimated by matching, stratification, or weighting on the fitted propensity score $\hat{\pi}(X)$ (e.g., Imbens 2004). We focus on inverse probability weighting (IPW), which is central to rigorous theory of statistical estimation in missing-data problems (e.g., Tsiatis 2006). Two standard IPW estimators for ν^0 are (e.g., McCaffrey et al. 2004; Abadie 2005)

$$\hat{\nu}^0_{\rm IPW}(\hat{\pi}) = \tilde{E}\left\{\frac{(1-T)\hat{\pi}(X)Y}{1-\hat{\pi}(X)}\right\}\Big/\tilde{E}(T),$$
$$\hat{\nu}^0_{\rm IPW,ratio}(\hat{\pi}) = \tilde{E}\left\{\frac{(1-T)\hat{\pi}(X)Y}{1-\hat{\pi}(X)}\right\}\Big/\tilde{E}\left\{\frac{(1-T)\hat{\pi}(X)}{1-\hat{\pi}(X)}\right\}.$$

The estimator of ATT based on $\hat{\nu}^0_{\rm IPW}(\hat{\pi})$ and $\hat{\nu}^1_{\rm NP}$ is then

$$\hat{\nu}^1_{\rm NP} - \hat{\nu}^0_{\rm IPW}(\hat{\pi}) = \tilde{E}\left\{\frac{T-\hat{\pi}(X)}{1-\hat{\pi}(X)}Y\right\}\Big/\tilde{E}(T).$$

If model (2.5) is correctly specified, then the IPW estimators are consistent. However, if model (2.5) is misspecified or even mildly so, these estimators can perform poorly, especially due to the instability of inverse weighting to fitted propensity scores $\hat{\pi}(X_i)$ near 1 for some untreated subjects (e.g., Kang & Schafer 2007).

3.2 Semiparametric Theory and AIPW Estimation

For consistency, the estimator $\hat{\nu}^0_{\rm OR}$ requires a correctly specified OR model (2.3) for $t = 0$, whereas $\hat{\nu}^0_{\rm IPW}$ and $\hat{\nu}^0_{\rm IPW,ratio}$ require a correctly specified PS model (2.5). Alternatively, it is desirable to develop estimators of ν^0 and ATT using both OR model (2.3) and PS model (2.5) to gain efficiency and robustness, similarly as in estimation of ATE. In this section, we review semiparametric theory obtained in Hahn (1998) and Chen et al. (2008), and then derive locally efficient and doubly robust estimators of ν^0 and ATT in the form of augmented IPW (AIPW) estimators.

First, Proposition 3.1 gives semiparametric influence functions and Table 3.1 with $t = 0$ gives the semiparametric efficiency bounds for estimation of ν^0 under three different settings, based on Hahn (1998) and Chen et al. (2008).

Proposition 3.1 *Let $q = E(T)$ and define*

$$\tau^0(\pi, h) = \frac{1-T}{1-\pi(X)}\pi(X)Y - \left\{\frac{1-T}{1-\pi(X)} - 1\right\}h(X).$$

The efficient influence function for estimation of ν^0 is as follows, depending on assumptions on the propensity score.

(i) The efficient influence function is

$$\varphi^0_{NP}(Y, T, X) = \left\{\tau^0(\pi, m_0) - T\nu^0\right\}\Big/q.$$

Table 3.1: Efficiency bounds for estimation of $\nu^t = E(Y^t|T=1)$

Assumption	Efficiency bound
Nonparametric model	$V^t_{\rm NP} = \text{var}\{\varphi^t_{\rm NP}(Y,T,X)\}$
Parametric PS model	$V^t_{\rm SP} = \text{var}\{\varphi^t_{\rm SP}(Y,T,X)\}$
Known PS	$V^t_{\rm SP*} = \text{var}\{\varphi^t_{\rm SP*}(Y,T,X)\}$

(ii) If the propensity score $\pi(X)$ is known, then the efficient influence function is

$$\begin{aligned}\varphi^0_{SP*}(Y,T,X) &= \left\{\tau^0(\pi,\pi m_0) - \pi(X)\nu^0\right\}\Big/ q \\ &= \varphi^0_{NP}(Y,T,X) - \{T-\pi(X)\}\frac{m_0(X)-\nu^0}{q}.\end{aligned}$$

(iii) If the propensity score $\pi(X)$ is unknown but assumed to belong to a correctly specified parametric family $\pi(X;\gamma)$, then the efficient influence function is

$$\varphi^0_{SP}(Y,T,X) = \varphi^0_{SP*}(Y,T,X) + \Pi\left[\{T-\pi(X)\}\frac{m_0(X)-\nu^0}{q}\Big| s_\gamma(T,X)\right],$$

where for two random vectors Z_1 and Z_2, $\Pi(Z_2|Z_1) = \text{cov}(Z_2,Z_1)\text{var}^{-1}(Z_1)Z_1$, i.e., the projection of Z_2 onto Z_1.

As discussed in Hahn (1998) and Chen et al. (2008), the efficiency bounds in Table 3.1 satisfy the following order: $V^0_{\rm NP} \geq V^0_{\rm SP} \geq V^0_{\rm SP*}$, with strict inequalities in general. In fact, the influence functions $\varphi^0_{\rm NP}$, $\varphi^0_{\rm SP}$, and $\varphi^0_{\rm SP*}$ can all be expressed as the following functional with suitable choices of $h(X)$:

$$\varphi^0_h(Y,T,X) = \left\{\tau^0(\pi,h) - T\nu^0\right\}\Big/ q. \tag{3.1}$$

The minimum variance of $\varphi^0_h(Y,T,X)$ over possible choices of $h(X)$ is exactly $V^0_{\rm SP*}$, corresponding to the choice $h(X) = \pi(X)m_0(X) + \{1-\pi(X)\}\nu^0$.

This ordering of efficiency bounds agrees with the usual comparison that the efficiency bound under a more restrictive model is no greater than under a less restrictive model. But this relationship differs from the result that the semiparametric efficiency

bounds for estimation of $\mu^t = E(Y^t)$ are the same whether under the nonparametric model for $\pi(X)$, or under a parametric model for $\pi(X)$, or with exact knowledge of $\pi(X)$. Conceptually, these differences reflect the fact the propensity score is ancillary for estimation of ATE, but not ancillary for estimation of ATT (Hahn, 1998).

We now derive two estimators of ν^0 that depend on both fitted outcome regression function $\hat{m}_0(X)$ and fitted propensity score $\hat{\pi}(X)$, by directly taking the efficient influence functions in Proposition 3.1 as estimating functions, with $\hat{m}_0(X)$ and $\hat{\pi}(X)$ in place of the unknown truth $m_0(X)$ and $\pi(X)$. Proposition 3.2 shows that both estimators possess local efficiency but of different types, and only one estimator is doubly robust. For clarity, the semiparametric efficiency bound V^0_{NP} under the nonparametric model is hereafter referred to as the nonparametric efficiency bound. See, for example, Newey (1990), Robins & Rotnitzky (2001), and Tsiatis (2006) for general discussions on local efficiency and double robustness.

Proposition 3.2 *Under suitable regularity conditions (see Appendix), the following results hold.*

(i) Define an estimator of ν^0 *as*

$$\hat{\nu}^0_{NP}(\hat{\pi}, \hat{m}_0) = \tilde{E}\left\{\tau^0(\hat{\pi}, \hat{m}_0)\right\} \Big/ \tilde{E}(T).$$

Then $\hat{\nu}^0_{NP}(\hat{\pi}, \hat{m}_0)$ *is locally nonparametric efficient: it achieves the nonparametric efficiency bound* V^0_{NP} *when both model (2.3) for* $t = 0$ *and model (2.5) are correctly specified. Moreover,* $\hat{\nu}^0_{NP}(\hat{\pi}, \hat{m}_0)$ *is doubly robust: it remains consistent when either model (2.3) for* $t = 0$ *or model (2.5) is correctly specified.*

(ii) Define an estimator of ν^0 *as*

$$\hat{\nu}^0_{SP}(\hat{\pi}, \hat{m}_0) = \tilde{E}\left\{\tau^0(\hat{\pi}, \hat{\pi}\hat{m}_0)\right\} \Big/ \tilde{E}\{\hat{\pi}(X)\}.$$

For logistic PS model (2.5), $\hat{\nu}^0_{SP}(\hat{\pi}, \hat{m}_0)$ *can be equivalently expressed as*

$$\hat{\nu}^0_{SP}(\hat{\pi}, \hat{m}_0) = \tilde{E}\left\{\tau^0(\hat{\pi}, \hat{\pi}\hat{m}_0)\right\} \Big/ \tilde{E}(T),$$

because $\tilde{E}(T) = \tilde{E}\{\hat{\pi}(X)\}$ *by Eq. (2.7) with* $f(X)$ *including 1. Then* $\hat{\nu}^0_{SP}(\hat{\pi}, \hat{m}_0)$ *is locally semiparametric efficient: it achieves the semiparametric efficiency bound* V^0_{SP} *when both model (2.3) for* $t = 0$ *and model (2.5) are correctly specified. But* $\hat{\nu}^0_{SP}(\hat{\pi}, \hat{m}_0)$ *is, generally, not doubly robust.*

The estimators $\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi}, \hat{m}_0)$ and, for a logistic PS model, $\hat{\nu}^0_{\mathrm{SP}}(\hat{\pi}, \hat{m}_0)$ are in the form of AIPW estimators, with the choice $h = \hat{m}_0$ or $h = \hat{\pi}\hat{m}_0$ respectively:

$$\begin{aligned}\hat{\nu}^0(\hat{\pi}, h) &= \tilde{E}\left\{\tau^0(\hat{\pi}, h)\right\} \Big/ \tilde{E}(T) \\ &= \tilde{E}\left[\frac{1-T}{1-\hat{\pi}(X)}\hat{\pi}(X)Y - \left\{\frac{1-T}{1-\hat{\pi}(X)} - 1\right\} h(X)\right] \Big/ \tilde{E}(T),\end{aligned}$$

which are defined by directly taking (3.1) as the estimating function with the fitted propensity score $\hat{\pi}(X)$ in place of the unknown truth $\pi(X)$. Setting $h(X) \equiv 0$ leads to the simple estimator $\hat{\nu}^0_{\mathrm{IPW}}$. Although AIPW estimators for $\mu^t = E(Y^t)$ have been well studied in estimation of ATE and other missing-data problems (Robins et al., 1994; Tan, 2006), the estimators $\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi}, \hat{m}_0)$ and $\hat{\nu}^0_{\mathrm{SP}}(\hat{\pi}, \hat{m}_0)$ seem to be derived for the first time using efficient influence functions from semiparametric theory.

By local semiparametric efficiency, the estimator $\hat{\nu}^0_{\mathrm{SP}}(\hat{\pi}, \hat{m}_0)$ achieves the minimum asymptotic variance among all regular estimators under PS model (2.5), including AIPW estimators $\hat{\nu}^0_h(\hat{\pi}, \hat{m}_0)$ over possible choices of $h(X)$, when both model (2.3) for $t = 0$ and model (2.5) are correctly specified. But $\hat{\nu}^0_{\mathrm{SP}}(\hat{\pi}, \hat{m}_0)$ is not doubly robust, and $\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi}, \hat{m}_0)$ is doubly robust. This situation differs from the case where among the class of AIPW estimators of μ^0, the estimator

$$\hat{\mu}^0_{\mathrm{AIPW}} = \tilde{E}\left[\frac{1-T}{1-\hat{\pi}(X)}Y - \left\{\frac{1-T}{1-\hat{\pi}(X)} - 1\right\}\hat{m}_0(X)\right],$$

is doubly robust, i.e., consistent when either OR model (2.3) for $t = 0$ or PS model (2.5) is correctly specified, and locally semiparametric or nonparametric efficient, i.e., achieving the minimum asymptotic variance among all regular estimators under parametric PS model (2.5) or, respectively, under the nonparametric model when model (2.3) for $t = 0$ and model (2.5) are correctly specified. As discussed after Proposition

3.1, the semiparametric efficient bound for estimation of μ^0 under a parametric PS model coincides with that under the nonparametric model.

Next, we present semiparametric influence functions in Proposition 3.3 and semiparametric efficiency bounds in Table 3.1 with $t = 1$ for estimation of ν^1, based on Hahn (1998) and Chen et al. (2008). Similarly as for estimation of ν^0, the efficiency bounds satisfy $V^1_{\mathrm{NP}} \geq V^1_{\mathrm{SP}} \geq V^1_{\mathrm{SP}*}$, with strict inequalities in general.

Proposition 3.3 *The efficient influence function for estimation of ν^1 is as follows, depending on assumptions on the propensity score.*

(i) The efficient influence function is

$$\varphi^1_{NP}(Y,T,X) = \big(TY - T\nu^1\big) \Big/ q.$$

(ii) If the propensity score $\pi(X)$ is known, then the efficient influence function is

$$\begin{aligned}\varphi^1_{SP*}(Y,T,X) &= \big[TY - \{T-\pi(X)\}m_1(X) - \pi(X)\nu^1\big] \Big/ q \\ &= \varphi^1_{NP}(Y,T,X) - \{T-\pi(X)\}\frac{m_1(X)-\nu^1}{q}.\end{aligned}$$

(iii) If the propensity score $\pi(X)$ is unknown but assumed to belong to a correctly specified parametric family $\pi(X;\gamma)$, then the efficient influence function is

$$\varphi^1_{SP}(Y,T,X) = \varphi^1_{SP*}(Y,T,X) + \Pi\left[\{T-\pi(X)\}\frac{m_1(X)-\nu^1}{q}\Big| s_\gamma(T,X)\right].$$

The estimator $\hat{\nu}^1_{\mathrm{NP}} = \tilde{E}(TY)/\tilde{E}(T)$ is always consistent and has the efficient influence function $\varphi^1_{\mathrm{NP}}(Y,T,X)$. Therefore, $\hat{\nu}^1_{\mathrm{NP}}$ is fully robust to model misspecification, and globally nonparametric efficient. Alternatively, taking $\varphi^1_{\mathrm{SP}*}(Y,T,X)$ as an estimating function with $\hat{m}_1(X)$ and $\hat{\pi}(X)$ in place of $m_1(X)$ and $\pi(X)$ gives an estimator of ν^1 that is locally semiparametric efficient, but not doubly robust.

Proposition 3.4 *Under suitable regularity conditions (see Appendix), the following results hold.*

(i) The estimator $\hat{\nu}^1_{NP} = \tilde{E}(TY)/\tilde{E}(T)$ *is consistent and achieves the nonparametric efficiency bound* V^1_{NP}*, independently of model (2.3) for* $t = 1$ *and model (2.5).*

(ii) Define an estimator of ν^1 *as*

$$\hat{\nu}^1_{SP}(\hat{\pi}, \hat{m}_1) = \tilde{E}\left[TY - \{T - \hat{\pi}(X)\}\hat{m}_1(X)\right]/\tilde{E}\{\hat{\pi}(X)\}.$$

For logistic PS model (2.5), $\hat{\nu}^1_{SP}(\hat{\pi}, \hat{m}_1)$ *can be equivalently expressed as*

$$\hat{\nu}^1_{SP}(\hat{\pi}, \hat{m}_1) = \tilde{E}\left[TY - \{T - \hat{\pi}(X)\}\hat{m}_1(X)\right]/\tilde{E}(T).$$

Then $\hat{\nu}^1_{SP}(\hat{\pi}, \hat{m}_1)$ *is locally semiparametric efficient: it attains the semiparametric efficiency bound* V^1_{SP} *when both model (2.3) for* $t = 1$ *and model (2.5) are correctly specified. But* $\nu^1_{SP}(\hat{\pi}, \hat{m}_1)$ *is not doubly robust.*

Finally, for estimation of ATT $= \nu^1 - \nu^0$, the efficient influence function is the difference of the efficient influence functions for estimation of ν^1 and ν^0 under each of the three settings in Propositions 3.1 and 3.3. Combining the estimators of ν^0 and ν^1 in Propositions 3.2 and 3.4 leads to the following results.

Corollary 3.5 *Under suitable regularity conditions (see Appendix), the following results hold.*

(i) The estimator $\hat{\nu}^1_{NP} - \hat{\nu}^0_{NP}(\hat{\pi}, \hat{m}_0)$ *for ATT is locally nonparametric efficient: it achieves the nonparametric efficiency bound,* $\mathrm{var}\{\varphi^1_{NP}(Y, T, X) - \varphi^0_{NP}(Y, T, X)\}$*, when both model (2.3) for* $t = 0$ *and model (2.5) are correctly specified. Moreover, this estimator is doubly robust: it remains consistent when either model (2.3) for* $t = 0$ *or model (2.5) is correctly specified.*

(ii) The estimator $\hat{\nu}^1_{SP}(\hat{\pi}, \hat{m}_0) - \hat{\nu}^0_{SP}(\hat{\pi}, \hat{m}_0)$ *for ATT is locally semiparametric efficient: it achieves the semiparametric efficiency bound,* $\mathrm{var}\{\varphi^1_{SP}(Y, T, X) - \varphi^0_{SP}(Y, T, X)\}$*, when both model (2.3) for* $t = 0, 1$ *and model (2.5) are correctly specified. But this estimator is, generally, not doubly robust.*

3.3 Improved Estimation

We develop estimators of ν^0 that are not only locally nonparametric efficient and doubly robust, but also intrinsically efficient: when the PS model (2.5) is correctly specified but the OR model (2.3) for $t = 0$ may be misspecified, these estimators achieve at least as small asymptotic variances among a class of AIPW estimators, including $\hat{\nu}^0_{\rm NP}(\hat{\pi}, \hat{m}_0)$ but only with $\hat{\pi}(X)$ replaced by the fitted value from a slightly augmented PS model as defined later in (3.2). The new estimators are then similar to $\hat{\nu}^0_{\rm NP}(\hat{\pi}, \hat{m}_0)$, in being consistent when either the PS model or the OR model is correctly specified and achieving the nonparametric efficiency bound $V^0_{\rm NP}$ when both models are correctly specified, but often achieve greater efficiency over $\hat{\nu}^0_{\rm NP}(\hat{\pi}, \hat{m}_0)$ when the PS model is correctly specified but the OR model is misspecified.

Similarly, we develop estimators of ATT that are not only locally nonparametric efficient and doubly robust, but also often provide efficiency gains over $\hat{\nu}^1_{\rm NP} - \hat{\nu}^0_{\rm NP}(\hat{\pi}, \hat{m}_0)$ when the PS model is correctly specified but the OR model is misspecified.

Before proceeding, we point out that although, by symmetry, it also seems desirable to construct estimators of ν^0 or ATT that are not only locally nonparametric efficient and doubly robust, but also achieve efficiency gains approximately over $\hat{\nu}^0_{\rm NP}(\hat{\pi}, \hat{m}_0)$ or $\hat{\nu}^1_{\rm NP} - \hat{\nu}^0_{\rm NP}(\hat{\pi}, \hat{m}_0)$ when the OR model is correctly specified but the PS model is misspecified, such estimators have not been obtained so far.

3.3.1 Regression Estimators

We derive regression estimators for ν^0 and ATT to achieve the desired properties, similarly to regression estimators for ATE (Tan, 2006) but with an important new idea as follows. For simplicity, assume in Sections 3.3.1–3.3.2 that PS model (2.5) is logistic regression. See Appendix for an extension when PS model (2.5) is non-logistic regression. Consider an augmented logistic PS model

$$\begin{aligned} P(T = 1|X) &= \pi_{\rm aug}(X; \gamma, \delta, \hat{\alpha}) \\ &= \text{expit}\left\{\gamma^{\rm T} f(X) + \delta_0\, \hat{m}_0(X) + \delta_1\, \hat{m}_1(X)\right\}, \end{aligned} \tag{3.2}$$

where $\text{expit}(c) = \{1 + \exp(-c)\}^{-1}$, $\hat{\alpha} = (\hat{\alpha}_0^{\mathrm{T}}, \hat{\alpha}_1^{\mathrm{T}})^{\mathrm{T}}$ are estimates of $\alpha = (\alpha_0^{\mathrm{T}}, \alpha_1^{\mathrm{T}})^{\mathrm{T}}$ from OR model (2.3), and $\delta = (\delta_0, \delta_1)^{\mathrm{T}}$ are unknown coefficients for additional regressors $\hat{m}_0(X)$ and $\hat{m}_1(X)$. Let $(\tilde{\gamma}, \tilde{\delta})$ be the MLE of (γ, δ) and $\tilde{\pi}(X) = \pi_{\text{aug}}(X; \tilde{\gamma}, \tilde{\delta}, \hat{\alpha})$. An important consequence of including the additional regressors is that, by Eq. (2.7), we have, in addition to $\tilde{E}[\{T - \tilde{\pi}(X)\}f(X)] = 0$,

$$\tilde{E}\left[\{T - \tilde{\pi}(X)\}\hat{m}_t(X)\right] = 0, \quad t = 0, 1. \tag{3.3}$$

For the augmented PS model, there may be linear redundancy in the variables, $\{f(X)$, $\hat{m}_0(X)$, $\hat{m}_1(X)\}$, in which case the regressors need to be redefined accordingly. For example, if all variables in $g_t(X)$ are linear combinations of $f(X)$, and $\Psi(\cdot)$ is the identity link corresponding to linear regression in (2.3), then

Condition L: $\hat{m}_t(X)$ is a linear combination of variables in $f(X)$ for $t = 0, 1$.

Under Condition L, the augmented model (3.2) reduces to the original model (2.5), and hence all the subsequent results are valid with $\tilde{\pi}(X) = \hat{\pi}(X)$.

With $\tilde{\pi}(X)$ the fitted value from the augmented PS model (3.2), we define the regression estimator of $\nu^t = E(Y^t|T = 1)$ as

$$\tilde{\nu}^t_{\text{reg}} = \tilde{E}\left(\tilde{\eta}_t - \tilde{\beta}_t^{\mathrm{T}}\tilde{\xi}_t\right)/\tilde{E}(T), \quad t = 0 \text{ or } 1,$$

where $\tilde{\beta}_t = \tilde{E}^{-1}(\tilde{\xi}_t\tilde{\zeta}_t^{\mathrm{T}})\tilde{E}(\tilde{\xi}_t\tilde{\eta}_t)$ with

$$\tilde{\eta}_1 = TY, \qquad \tilde{\eta}_0 = \frac{1-T}{1-\tilde{\pi}(X)}\tilde{\pi}(X)Y,$$

$$\tilde{\xi}_1 = \left\{\frac{T}{\tilde{\pi}(X)} - 1\right\}\frac{\tilde{h}(X)}{1-\tilde{\pi}(X)}, \qquad \tilde{\xi}_0\,(= -\tilde{\xi}_1) = \left\{\frac{1-T}{1-\tilde{\pi}(X)} - 1\right\}\frac{\tilde{h}(X)}{\tilde{\pi}(X)},$$

$$\tilde{\zeta}_1 = \frac{T}{\tilde{\pi}(X)}\frac{\tilde{h}(X)}{1-\tilde{\pi}(X)}, \qquad \tilde{\zeta}_0 = \frac{1-T}{1-\tilde{\pi}(X)}\frac{\tilde{h}(X)}{\tilde{\pi}(X)},$$

and $\tilde{h}(X) = \{\tilde{h}_1^{\mathrm{T}}, (C\tilde{h}_2)^{\mathrm{T}}\}^{\mathrm{T}}(X)$ are defined with a constant matrix C such that the variables in $\tilde{h}(X)$ are linearly independent, and

$$\begin{aligned}
\tilde{h}_1(X) &= [\{1-\tilde{\pi}(X)\}\tilde{v}_1^{\mathrm{T}}(X), \tilde{\pi}(X)\tilde{v}_0^{\mathrm{T}}(X)]^{\mathrm{T}}, \\
\tilde{h}_2(X) &= \tilde{\pi}(X)\{1-\tilde{\pi}(X)\}\left\{f_{(1)}^{\mathrm{T}}(X), \hat{m}_0(X)\right\}^{\mathrm{T}}, \\
\tilde{v}_1(X) &= \{\tilde{\pi}(X), \tilde{\pi}(X)\hat{m}_1(X)\}^{\mathrm{T}}, \quad \tilde{v}_0(X) = \{\tilde{\pi}(X), \tilde{\pi}(X)\hat{m}_0(X)\}^{\mathrm{T}}.
\end{aligned}$$

where $f_{(1)}(X)$ is the vector of nonconstant variables in $f(X)$, because $\tilde{\pi}(X)\{1-\tilde{\pi}(X)\}$ is already a component of $\{1-\tilde{\pi}(X)\}\tilde{v}_1^{\mathrm{T}}(X)$ in $\tilde{h}_1(X)$. For example, if Condition L holds for $t = 0$ or 1, then $\tilde{h}(X)$ should be specified such that one variable is removed from the vector $\tilde{\pi}(X)\{1-\tilde{\pi}(X)\}f_{(1)}(X)$ in $\tilde{h}_2(X)$.

The variables in $\tilde{h}(X)$ are included for the following considerations. The variables $\tilde{\pi}(X)\hat{m}_0(X)$ and $\tilde{\pi}(X)\hat{m}_1(X)$ are included in $\tilde{v}_0(X)$ and $\tilde{v}_1(X)$ respectively to achieve double robustness and local nonparametric efficiency, as later seen from Eq. (3.5). Moreover, the variables in $\tilde{h}_2(X)$, in addition to $\{1-\tilde{\pi}(X)\}\tilde{v}_1(X)$, are included to accommodate the variation of $(\tilde{\gamma}, \tilde{\delta})$ for achieving intrinsic efficiency, as later described in Proposition 3.6. The corresponding variables in $\tilde{\xi}_0$ or $\tilde{\xi}_1$ are exactly the scores $\{T-\tilde{\pi}(X)\}\{f^{\mathrm{T}}(X), \hat{m}_0(X), \hat{m}_1(X)\}^{\mathrm{T}}$ for the augmented PS model (3.2). Finally, $\tilde{\pi}(X)$ is included in $\tilde{v}_0(X)$ and $\tilde{v}_1(X)$ to ensure efficiency gains over the ratio estimator $\hat{\nu}^0_{\mathrm{IPW,ratio}}(\tilde{\pi})$ under a correctly specified PS model, as discussed after Corollary 3.7.

The name "regression estimator" is adopted from the literatures of survey sampling (Cochran, 1977) and Monte Carlo integration (Hammersley & Handscomb, 1964), and should be distinguished from the estimator $\hat{\nu}^t_{\mathrm{OR}}$ based on outcome regression in Section 3.1. The idea is to exploit the fact that if the PS model is correct, then $\tilde{E}(\tilde{\eta}_t)$ asymptotically has mean $E(TY^t)$ (to be estimated) and $\tilde{\xi}_t$ mean 0 (known). That is, $\tilde{\xi}_t$ serves as auxiliary variables (in the terminology of survey sampling) or control variates (in that of Monte Carlo integration). The effect of variance reduction using regression estimators is seen from in the following results.

Proposition 3.6 *Under suitable regularity conditions (see Appendix), the estimator $\tilde{\nu}^t_{reg}$ for ν^t has the following properties for $t = 0, 1$.*

(i) $\tilde{\nu}^t_{reg}$ *is locally nonparametric efficient: it achieves the nonparametric efficiency bound* V^t_{NP} *when both model (2.3) for the corresponding* t *and model (2.5) are correctly specified.*

(ii) $\tilde{\nu}^t_{reg}$ *is doubly robust: it remains consistent when either model (2.3) for the corresponding* t *or model (2.5) is correctly specified.*

(iii) $\tilde{\nu}^t_{reg}$ *is intrinsically efficient: if model (2.5) is correctly specified, then it achieves the lowest asymptotic variance among the class of estimators*

$$\tilde{E}\left(\tilde{\eta}_t - b_t^{\mathrm{T}}\tilde{\xi}_t\right)/\tilde{E}(T), \tag{3.4}$$

where b_t *is an arbitrary vector of constants.*

Corollary 3.7 *The estimator* $\tilde{\nu}^1_{reg} - \tilde{\nu}^0_{reg}$ *for ATT has the following properties.*

(i) $\tilde{\nu}^1_{reg} - \tilde{\nu}^0_{reg}$ *is locally nonparametric efficient: it achieves the nonparametric efficiency bound,* $\mathrm{var}\{\varphi^1_{NP}(Y,T,X) - \varphi^0_{NP}(Y,T,X)\}$*, when both model (2.3) for* $t = 0, 1$ *and model (2.5) are correctly specified.*

(ii) $\tilde{\nu}^1_{reg} - \tilde{\nu}^0_{reg}$ *is doubly robust: it remains consistent when either model (2.3) for* $t = 0, 1$ *or model (2.5) is correctly specified.*

(iii) $\tilde{\nu}^1_{reg} - \tilde{\nu}^0_{reg}$ *is intrinsically efficient: if model (2.5) is correctly specified, then it achieves the lowest asymptotic variance among the class of estimators*

$$\tilde{E}\left(\tilde{\eta}_1 - \tilde{\eta}_0 - b_0^{\mathrm{T}}\tilde{\xi}_0\right)/\tilde{E}(T),$$

where b_0 *is an arbitrary vector of constants.*

The use of augmented propensity scores $\tilde{\pi}(X)$ is crucial for $\tilde{\nu}^t_{\mathrm{reg}}$ to be doubly robust or, specifically, consistent under a correctly specified OR model but a misspecified PS model. [There are special cases, for example, Condition L, where $\tilde{\pi}(X)$ reduces to $\hat{\pi}(X)$.] If the OR model (2.3) for $t = 0$ or 1 is correctly specified, then, as shown in the

Appendix, the vector $\tilde{\beta}_t$ converges to a constant vector β_t^* such that

$$\tilde{\nu}^t_{\rm reg} = \tilde{E}\left(\tilde{\eta}_t - \beta_t^{*\rm T}\tilde{\xi}_t\right)/\tilde{E}(T) + o_p(n^{-1/2}) = \hat{\nu}^t_{\rm SP}(\tilde{\pi}, \hat{m}_t) + o_p(n^{-1/2}), \qquad (3.5)$$

mainly because $\tilde{\pi}(X)\hat{m}_0(X)$ is a linear combination of variables in $\tilde{h}(X)/\tilde{\pi}(X)$ and $\tilde{\pi}(X)\hat{m}_1(X)$ is a linear combination of variables in $\tilde{h}(X)/\{1-\tilde{\pi}(X)\}$. By Eq. (3.3) for the augmented PS model, $\hat{\nu}^t_{\rm SP}(\tilde{\pi}, \hat{m}_t)$ is identical to $\hat{\nu}^0_{\rm NP}(\tilde{\pi}, \hat{m}_0)$ for $t = 0$, which is doubly robust, or to $\hat{\nu}^1_{\rm NP}$ for $t = 1$, which is fully robust. Therefore, $\tilde{\nu}^t_{\rm reg}$ is consistent when the OR model (2.3) for the corresponding t is correctly specified. This result would not hold when $\tilde{\nu}^t_{\rm reg}$ were defined with $\hat{\pi}(X)$ in place of $\tilde{\pi}(X)$.

The estimator $\tilde{\nu}^t_{\rm reg}$ is locally nonparametric efficient, similarly as $\hat{\nu}^0_{\rm NP}(\tilde{\pi}, \hat{m}_0)$ or $\hat{\nu}^1_{\rm NP}$. In fact, $\tilde{\nu}^t_{\rm reg}$ is generally not locally semiparametric efficient with respect to PS model (2.5), but locally semiparametric efficient with respect to PS model (3.2) in the following sense: $\tilde{\nu}^t_{\rm reg}$ achieves the semiparametric efficiency bounded calculated under model (3.2), when both model (2.3) and model (2.5) are correctly specified. When model (2.5) holds, the efficiency bound $V^t_{\rm SP}$ under model (3.2) coincides with the nonparametric efficiency bound $V^t_{\rm NP}$, because $\{T-\pi(X)\}\{m_t(X)-\nu^t\}$ is a linear combination of the score function, which contains $\{T-\pi(X)\}\{1, m_0(X), m_1(X)\}^{\rm T}$ under model (3.2) as shown in Appendix I. On the other hand, $\tilde{\nu}^t_{\rm reg}$ with $\tilde{\pi}(X)$ replaced by $\hat{\pi}(X)$ throughout would be locally semiparametric efficient with respect to original PS model (2.5), but generally not doubly robust, similarly as $\hat{\nu}^t_{\rm SP}(\tilde{\pi}, \hat{m}_t)$.

A classical estimator of the optimal choice of b_t in minimizing the asymptotic variance of (3.4) is $\hat{\beta}_t = \tilde{E}(\tilde{\xi}_t\tilde{\xi}_t^{\rm T})^{-1}\tilde{E}(\tilde{\xi}_t\tilde{\eta}_t)$, which differs from $\tilde{\beta}_t$ in a subtle manner. It can be shown that the corresponding estimator, $\hat{\nu}^t_{\rm reg} = \tilde{E}(\tilde{\eta}_t - \hat{\beta}_t^{\rm T}\tilde{\xi}_t)/\tilde{E}(T)$, for ν^t is asymptotically equivalent to the first order to $\tilde{\nu}^t_{\rm reg}$ when the PS model is correctly specified. But $\hat{\nu}^t_{\rm reg}$, unlike $\tilde{\nu}^t_{\rm reg}$, is generally inconsistent for ν^t, even when the OR model is correctly specified and the PS model may be misspecified. The particular form of $\tilde{\beta}_t$, although seems ad hoc in the above definition, can also be derived through empirical efficiency maximization (Rubin & van der Laan, 2008; Tan, 2008) and design-optimal regression estimation in Poisson sampling (Tan, 2013). See further discussion related

to calibration estimation after Proposition 3.8.

By intrinsic efficiency, if the PS model is correctly specified, then $\tilde{\nu}^0_{\text{reg}}$ is asymptotically at least as efficient as not only $\hat{\nu}^0_{\text{NP}}(\tilde{\pi}, \hat{m}_0)$, but also $\hat{\nu}^0_{\text{IPW}}(\tilde{\pi})$ and $\hat{\nu}^0_{\text{IPW,ratio}}(\tilde{\pi})$. The estimator $\hat{\nu}^0_{\text{NP}}(\tilde{\pi}, \hat{m}_0)$, defined as $\hat{\nu}^0_{\text{NP}}(\hat{\pi}, \hat{m}_0)$ with $\hat{\pi}(X)$ replaced by $\tilde{\pi}(X)$, remains locally nonparametric efficient and doubly robust, and falls in the class (3.4) for $t = 0$ because $\hat{m}_0(X)$ is a linear combination of the variables, $\tilde{\pi}(X)\hat{m}_0(X)$ and $\{1 - \tilde{\pi}(X)\}\hat{m}_0(X)$, included in $\tilde{h}(X)/\tilde{\pi}(X)$. Moreover, the simple estimator $\hat{\nu}^0_{\text{IPW}}(\tilde{\pi})$ based on $\tilde{\pi}(X)$ also falls in the class (3.4) for $t = 0$, with $b_0 = 0$. The ratio estimator $\hat{\nu}^0_{\text{IPW,ratio}}(\tilde{\pi})$ does not directly fall in the class (3.4), but can be shown to be asymptotically equivalent to the first order, under a correctly specified PS model, to $\tilde{E}(\hat{\eta}_0 - [(1 - T)/\{1 - \tilde{\pi}(X)\} - 1]\nu^0)/\tilde{E}(T)$, which falls in class (3.4) for $t = 0$ becauase 1 is a linear combination of the variables, $\tilde{\pi}(X)$ and $1 - \tilde{\pi}(X)$, in $\tilde{h}(X)/\tilde{\pi}(X)$.

The estimator $\hat{\nu}^1_{\text{NP}} = \tilde{E}(\hat{\eta}_1)$ falls in the class (3.4) for $t = 1$, with $b_1 = 0$. Therefore, the estimator $\tilde{\nu}^1_{\text{reg}} - \tilde{\nu}^0_{\text{reg}}$ for ATT is asymptotically at least as efficient as $\hat{\nu}^1_{\text{NP}} - \hat{\nu}^0_{\text{NP}}(\tilde{\pi}, \hat{m}_0)$ when the PS model is correctly specified, even though both estimators are locally nonparametric efficient and doubly robust.

A technical complication of using augmented propensity scores $\tilde{\pi}(X)$ is that $\tilde{\nu}^0_{\text{reg}}$ may not, in general, be intrinsically efficient, when compared to the class of estimators (3.4) with $\tilde{\pi}(X)$ replaced by $\hat{\pi}(X)$ in $\tilde{\eta}_0$ and $\tilde{\xi}_0$. [Nevertheless, such intrinsic efficiency holds in the special case where the OR model (2.3) for $t = 0$ is linear regression with all variables in $g_0(X)$ also included in $f(X)$.] Particularly, if the PS model (2.5) is correctly specified, then $\tilde{\nu}^0_{\text{reg}}$ may not be as efficient as $\hat{\nu}^0_{\text{NP}}(\hat{\pi}, \hat{m}_0)$ based on $\hat{\pi}(X)$ even though $\tilde{\nu}^0_{\text{reg}}$ is proven to be asymptotically at least as efficient as $\hat{\nu}^0_{\text{NP}}(\tilde{\pi}, \hat{m}_0)$ based on $\tilde{\pi}(X)$ and, when the OR model (2.3) for $t = 0$ is also correctly specified, asymptotically equivalent to $\hat{\nu}^0_{\text{NP}}(\hat{\pi}, \hat{m}_0)$ and $\hat{\nu}^0_{\text{NP}}(\tilde{\pi}, \hat{m}_0)$. However, the increase in the asymptotic variance of $\hat{\nu}^0_{\text{NP}}(\tilde{\pi}, \hat{m}_0)$ over that of $\hat{\nu}^0_{\text{NP}}(\hat{\pi}, \hat{m}_0)$ is usually small, caused by the use of a slightly augmented PS model (3.2). The estimator $\tilde{\nu}^0_{\text{reg}}$ may still often achieve efficiency gains over $\hat{\nu}^0_{\text{NP}}(\hat{\pi}, \hat{m}_0)$ when the PS model is correctly specified but the OR model is misspecified, as shown in our simulation studies.

3.3.2 Likelihood Estimators

A practical limitation of the regression estimators as well as AIPW estimators is that they may lie outside either the sample or the population range of observed outcomes. For example, $\tilde{\nu}^t_{\text{reg}}$ may take values outside the interval $(0, 1)$ for binary outcomes. Such behavior may occur due to the presence of fitted propensity scores $\tilde{\pi}(X_i)$ near 1 or, equivalently, large inverse weights $\{1 - \tilde{\pi}(X_i)\}^{-1}$ among the untreated. In this section, we derive likelihood estimators for ν^t that are not only doubly robust, locally nonparametric efficient, and intrinsically efficient similarly to the regression estimators, but also sample-bounded in falling within the range of $\{Y_i : T_i = t, i = 1, \ldots, n\}$. These likelihood estimators are therefore much less sensitive to large inverse weights than the regression and AIPW estimators.

There are two steps in constructing the desired likelihood estimators, similarly as for ATE estimation in Tan (2010) but using the fitted propensity scores $\tilde{\pi}(X)$ from augmented PS model (3.2). First, we derive intrinsically efficient, but non-doubly robust, likelihood estimators by the approach of empirical likelihood (Owen, 2001) taking $\tilde{\eta}_t - \nu^t T$ and $\tilde{\xi}_t$ as asymptotically unbiased estimating functions or, equivalently, the approach of nonparametric likelihood (Tan, 2006, 2010). Specifically, our approach is to maximize the log empirical likelihood, $\sum_{i=1}^n \log p_i$, subject to the constraints

$$\sum_{i=1}^n p_i \tilde{\xi}_{1,i} = 0 \quad \text{and} \quad \sum_{i=1}^n p_i(\tilde{\eta}_{t,i} - \nu^t T_i) = 0 \text{ for } t = 0, 1,$$

where p_i is a nonnegative weight assigned to (Y_i, T_i, X_i) for $i = 1, \ldots, n$ with $\sum_{i=1}^n p_i = 1$. We show in the Appendix that the resulting estimates of ν^0 and ν^1 are

$$\begin{aligned}
\hat{\nu}^0_{\text{lik}} &= \tilde{E}\left\{\frac{(1-T)\tilde{\pi}(X)Y}{1-\omega(X;\hat{\lambda})}\right\} \Big/ \tilde{E}\left\{\frac{(1-T)\tilde{\pi}(X)}{1-\omega(X;\hat{\lambda})}\right\}, \\
\hat{\nu}^1_{\text{lik}} &= \tilde{E}\left\{\frac{T\tilde{\pi}(X)Y}{\omega(X;\hat{\lambda})}\right\} \Big/ \tilde{E}\left\{\frac{T\tilde{\pi}(X)}{\omega(X;\hat{\lambda})}\right\},
\end{aligned}$$

where $\omega(X;\lambda) = \tilde{\pi}(X) + \lambda^{\mathrm{T}}\tilde{h}(X)$ and $\hat{\lambda}$ is a maximizer of the function

$$\ell(\lambda) = \tilde{E}[T\log\omega(X;\lambda) + (1-T)\log\{1-\omega(X;\lambda)\}],$$

subject to $\omega(X_i;\lambda) > 0$ if $T_i = 1$ and $\omega(X_i;\lambda) < 1$ if $T_i = 0$ for $i = 1,\ldots,n$. Setting the gradient of $\ell(\lambda)$ to zero shows that $\hat{\lambda}$ is a solution to

$$\tilde{E}\left[\frac{T-\omega(X;\lambda)}{\omega(X;\lambda)\{1-\omega(X;\lambda)\}}\tilde{h}(X)\right] = 0. \tag{3.6}$$

Because $\tilde{\pi}(X)$ is a linear combination of variables in $\tilde{h}(X)$, it follows from Eq. (3.6) that the two denominators, $\tilde{E}[(1-T)\tilde{\pi}(X)/\{1-\omega(X;\hat{\lambda})\}]$ and $\tilde{E}[T\tilde{\pi}(X)/\omega(X;\hat{\lambda})]$, in the definitions of $\hat{\nu}^0_{\mathrm{lik}}$ and $\hat{\nu}^1_{\mathrm{lik}}$ are equal to each other.

The estimator $\hat{\nu}^t_{\mathrm{lik}}$ can be shown to be intrinsically efficient among the class of estimators (3.4) and locally nonparametric efficient, but generally not doubly robust. We introduce the following modified likelihood estimators, to achieve double robustness but without affecting the first-order asymptotic behavior.

For $t = 0$ or 1, partition $\tilde{h}$ as $\tilde{h} = \{\tilde{h}^{\mathrm{T}}_{1t}, \tilde{h}^{\mathrm{T}}_{1(t)}, (C\tilde{h}_2)^{\mathrm{T}}\}^{\mathrm{T}}$ for a constant matrix C and accordingly λ as $\lambda = (\lambda^{\mathrm{T}}_{1t}, \lambda^{\mathrm{T}}_{1(t)}, \lambda^{\mathrm{T}}_2)^{\mathrm{T}}$, where $\tilde{h}_{1t} = \tilde{\pi}\tilde{v}_0$ or $(1-\tilde{\pi})\tilde{v}_1$ if $t = 0$ or 1, and $\tilde{h}_{1(t)}$ consists of the elements of $\tilde{h}_1$ excluding $\tilde{h}_{1t}$. Moreover, write $R_t = 0$ or 1, $\tilde{\pi}(t,X) = 1-\tilde{\pi}(X)$ or $\tilde{\pi}(X)$, and $\omega(t,X;\lambda) = 1-\omega(X;\lambda)$ or $\omega(X;\lambda)$ respectively for $t = 0$ or 1. Define $\tilde{\lambda}^t = (\tilde{\lambda}^{\mathrm{T}}_{1t}, \hat{\lambda}^{\mathrm{T}}_{1(t)}, \hat{\lambda}^{\mathrm{T}}_2)^{\mathrm{T}}$, where $\hat{\lambda}_{1(t)}$ and $\hat{\lambda}_2$ are obtained from $\hat{\lambda}$, and $\tilde{\lambda}_{1t}$ is a maximizer of the function

$$\kappa_t(\lambda_{1t}) = \tilde{E}\left[R_t\frac{\log\{\omega(t,X;\lambda_{1t},\hat{\lambda}_{1(t)},\hat{\lambda}_2)\} - \log\{\omega(t,X;\hat{\lambda})\}}{1-\tilde{\pi}(t,X)} - \lambda^{\mathrm{T}}_{1t}v_t(X)\right],$$

subject to $\omega(t,X_i;\lambda_{1t},\hat{\lambda}_{1(t)},\hat{\lambda}_2) > 0$ if $T_i = t$ for $i = 1,\ldots,n$. Setting the gradient of $\kappa_t(\lambda_{1t})$ to 0 shows that $\tilde{\lambda}_{1t}$ is a solution to

$$\tilde{E}\left[\left\{\frac{R_t}{\omega(t,X;\lambda_{1t},\hat{\lambda}_{1(t)},\hat{\lambda}_2)} - 1\right\}\tilde{v}_t(X)\right] = 0. \tag{3.7}$$

For $t = 0, 1$, the resulting estimator of ν^t is

$$\tilde{\nu}^t_{\text{lik}} \;=\; \tilde{E}\left\{\frac{R_t\tilde{\pi}(X)Y}{\omega(t,X;\tilde{\lambda}^t)}\right\}\Big/\tilde{E}\left\{\frac{R_t\tilde{\pi}(X)}{\omega(t,X;\tilde{\lambda}^t)}\right\} = \tilde{E}\left\{\frac{R_t\tilde{\pi}(X)Y}{\omega(t,X;\tilde{\lambda}^t)}\right\}\Big/\tilde{E}(T),$$

where the second equation holds due to Eq. (3.7) with $\tilde{\pi}(X)$ included in $\tilde{v}_0(X)$ and $\tilde{v}_1(X)$, and $\tilde{E}\{T - \tilde{\pi}(X)\} = 0$ by the score equation for model (3.2). The likelihood estimator $\tilde{\nu}^t_{\text{lik}}$ has several desirable properties as follows.

Proposition 3.8 *Under suitable regularity conditions (see Appendix), the estimator $\tilde{\nu}^t_{lik}$ for ν^t has the following properties for $t = 0, 1$.*

(i) $\tilde{\nu}^t_{lik}$ is sample-bounded: it lies within the range of $\{Y_i : T_i = t, i = 1, \ldots, n\}$.

(ii) If model (2.5) is correctly specified, then $\tilde{\nu}^t_{lik}$ is asymptotically equivalent, to the first order, to $\tilde{\nu}^t_{reg}$. Hence $\tilde{\nu}^t_{lik}$ is intrinsically efficient among the class (3.4) and locally nonparametric efficient, similarly as $\tilde{\nu}^t_{reg}$ in Proposition 3.6.

(iii) $\tilde{\nu}^t_{lik}$ is doubly robust, similarly as $\tilde{\nu}^t_{reg}$ in Proposition 3.6.

The sample-boundedness of $\tilde{\nu}^t_{\text{lik}}$ holds because $\omega(t, X_i; \tilde{\lambda}^t) > 0$ if $T_i = t$ for $i = 1, \ldots, n$ and $\tilde{E}\{R_t\tilde{\pi}(X)/\omega(t, X_i; \tilde{\lambda}^t)\} = \tilde{E}\{\tilde{\pi}(X)\} = \tilde{E}(T)$ by Eq. (3.7). The double robustness of $\tilde{\nu}^t_{\text{lik}}$ follows mainly for two reasons: $\tilde{E}\{R_t\tilde{\pi}(X)\hat{m}_t(X)/\omega(t, X_i; \tilde{\lambda}^t)\} = \tilde{E}\{\tilde{\pi}(X)\hat{m}_t(X)\}$ by Eq. (3.7) with $\tilde{\pi}(X)\hat{m}_t(X)$ included in $\tilde{v}_t(X)$, and $\tilde{E}\{\tilde{\pi}(X)\hat{m}_t(X)\} = \tilde{E}\{T\hat{m}_t(X)\}$ by Eq. (3.3) for the augmented PS model (3.2).

Eq. (3.7), which underlies both sample-boundedness and double robustness as discussed above, can be connected to calibration estimation using auxiliary information in survey sampling (Deville & Sarndal, 1992; Tan, 2013). In fact, the inverse weighted average of $\tilde{v}_t(X) = \tilde{\pi}(X)\{1, \hat{m}_t(X)\}^{\text{T}}$ is matched (or calibrated) with the simple sample average of $\tilde{v}_t(X)$. This is equivalent to saying that if Y is replaced by $\hat{m}_t(X)$, then the numerator in the definition of $\tilde{\nu}^t_{\text{lik}}$ yields exactly $\tilde{E}\{\tilde{\pi}(X)\hat{m}_t(X)\}$. A similar property holds for $\tilde{\nu}^t_{\text{reg}}$: if Y is replaced by $\hat{m}_t(X)$, then the numerator in the definition of $\tilde{\nu}^t_{\text{reg}}$ yields exactly $\tilde{E}\{\tilde{\pi}(X)\hat{m}_t(X)\}$. By this relationship, $\tilde{\nu}^t_{\text{reg}}$ and $\tilde{\nu}^t_{\text{lik}}$ can be referred to as calibrated regression and likelihood estimators.

The implication of intrinsic efficiency for $\tilde{\nu}^t_{\text{lik}}$ is similar to that for $\tilde{\nu}^t_{\text{reg}}$ as discussed in Section 3.3.1. If the PS model (2.5) is correctly specified while the OR model (2.3) may be misspecified, then $\tilde{\nu}^0_{\text{lik}}$ is asymptotically at least as efficient as $\hat{\nu}^0_{\text{NP}}(\tilde{\pi}, \hat{m}_0)$, and $\tilde{\nu}^1_{\text{lik}} - \tilde{\nu}^0_{\text{lik}}$ is asymptotically at least as efficient as $\hat{\nu}^1_{\text{NP}} - \hat{\nu}^0_{\text{NP}}(\tilde{\pi}, \hat{m}_0)$.

3.4 Extensions and Comparisons

To possibly enhance numerical stability and finite-sample performance, we suggest the following versions of $\tilde{\nu}^t_{\text{reg}}$ and $\tilde{\nu}^t_{\text{lik}}$ with simplifications of $\tilde{\pi}(X)$ and $\tilde{h}(X)$:

(i) Consider an augmented logistic PS model in place of (3.2):

$$\begin{aligned} P(T=1|X) &= \pi_{\text{aug2}}(X; \gamma_0, \delta, \hat{\alpha}, \hat{\gamma}) \\ &= \text{expit}\left[\text{logit}\{\hat{\pi}(X)\} + \gamma_0 + \delta_0\, \hat{m}_0(X) + \delta_1\, \hat{m}_1(X)\right], \end{aligned} \tag{3.8}$$

where $\text{logit}(\hat{\pi}) = \log\{\hat{\pi}/(1-\hat{\pi})\}$ is included as an offset, and γ_0 and $\delta = (\delta_0, \delta_1)^{\text{T}}$ are unknown coefficients. Let $(\tilde{\gamma}_0, \tilde{\delta})$ be the MLE of (γ_0, δ), and redefine $\tilde{\pi}(X) = \pi_{\text{aug2}}(X; \tilde{\gamma}_0, \tilde{\delta}, \hat{\alpha}, \hat{\gamma})$. This augmented model (3.8) is meaningful even when the original model (2.5) is non-logistic regression or when $\hat{\gamma}$ is obtained by non-maximum likelihood estimation, for example, penalized estimation.

(ii) Redefine $\tilde{h}(X) = \tilde{h}_1(X)$, that is, with $\tilde{h}_2(X)$ removed. Then $\tilde{\beta}_t$ is defined by projection of $\tilde{\eta}_t$ on a lower-dimensional vector $\tilde{\xi}_t$, and $\hat{\lambda}$ is defined by solving a lower-dimensional optimization problem. The dimension reduction may improve numerical stability and finite-sample performance of $\tilde{\nu}^t_{\text{reg}}$ and $\tilde{\nu}^t_{\text{lik}}$.

For concreteness, the resulting estimators $\tilde{\nu}^t_{\text{reg}}$ and $\tilde{\nu}^t_{\text{lik}}$ are denoted by $\tilde{\nu}^t_{\text{reg2}}$ and $\tilde{\nu}^t_{\text{lik2}}$ respectively. These simplified estimators can be shown to remain locally nonparametric efficient and doubly robust as in Propositions 3.6 and 3.8; they are generally not intrinsically efficient, but are expected to asymptotically nearly as efficient as $\tilde{\nu}^t_{\text{reg}}$ and $\tilde{\nu}^t_{\text{lik}}$ when the PS model (2.5) is correctly specified. Informally, $\tilde{\nu}^t_{\text{reg2}}$ and $\tilde{\nu}^t_{\text{lik2}}$ would be intrinsically efficient if $\hat{\pi}(X) = \pi(X; \hat{\gamma})$ were replaced, in model (3.8) and the definition of $\tilde{\pi}(X)$, by $\pi(X; \gamma^*)$ with γ^* the limit of $\hat{\gamma}$ in probability.

While $\tilde{h}_2(X)$ can be removed from $\tilde{h}(X)$ for dimension reduction, we point out that $\tilde{h}_1(X)$ can be extended to include additional functions of X for achieving calibration on those variables in addition to $\tilde{v}_t(X)$. Specifically, let $c_t(X)$ be a vector of known but possibly data-dependent functions of X *including* 1, for example, $g_t(X)$ in the OR model (2.3) for $t = 0, 1$. Redefine the augmented PS model (3.8) as

$$\begin{aligned} P(T = 1|X) &= \pi_{\text{aug2}}(X; \gamma_0, \delta, \hat{\gamma}) \\ &= \text{expit}\left[\text{logit}\{\hat{\pi}(X)\} + \gamma_0 + \delta_0^{\mathrm{T}} c_{0(1)}(X) + \delta_1^{\mathrm{T}} c_{1(1)}(X)\right], \end{aligned} \tag{3.9}$$

where γ_0 and $\delta = (\delta_0^{\mathrm{T}}, \delta_1^{\mathrm{T}})$ are unknown coefficients, and $c_{0(1)}(X)$ or $c_{1(1)}$ is the vector of nonconstant variables in $c_0(X)$ or $c_1(X)$ respectively. Redefine $\tilde{\pi}(X) = \pi_{\text{aug2}}(X; \tilde{\gamma}_0, \tilde{\delta}, \hat{\gamma})$ with $(\tilde{\gamma}_0, \tilde{\delta})$ the MLE of (γ_0, δ) for model (3.9), and redefine $\tilde{h}(X) = \tilde{h}_1(X)$ with $\tilde{v}_t(X) = \tilde{\pi}(X)c_t^{\mathrm{T}}(X)$ for $t = 0, 1$. Then Eq. (3.7) in conjunction with the score equation for model (3.8) leads to calibration equations

$$\tilde{E}\left\{\frac{(1-T)\tilde{\pi}(X)}{1-\omega(X;\tilde{\lambda}^0)}c_0(X)\right\} = \tilde{E}\{\tilde{\pi}(X)c_0(X)\} = \tilde{E}\{Tc_0(X)\}, \tag{3.10}$$

$$\tilde{E}\left\{\frac{T\tilde{\pi}(X)}{\omega(X;\tilde{\lambda}^1)}c_1(X)\right\} = \tilde{E}\{\tilde{\pi}(X)c_1(X)\} = \tilde{E}\{Tc_1(X)\}. \tag{3.11}$$

By the discussion after (3.5), the resulting estimators $\tilde{\nu}^t_{\text{reg2}}$ and $\tilde{\nu}^t_{\text{lik2}}$ are doubly robust and locally nonparametric efficient in the case where

Condition R: $\hat{m}_t(X)$ is a linear combination of $c_t(X)$ for $t = 0, 1$.

This condition is satisfied when $c_t(X)$ contains all variables in $g_t(X)$ including 1, and $\Psi(\cdot)$ is the identity link in the OR model (2.3).

In the rest of this section, we provide comparisons between our calibrated regression and likelihood methods and several related methods for estimating ATT. The estimators of ν^0 in Qin & Zhang (2008) and Graham et al. (2015) are in the form

$$\frac{1}{n_1}\sum_{i=1}^{n}\frac{(1-T_i)\hat{\pi}(X_i)}{w_i}Y_i,$$

where $\{w_i > 0 : T_i = 0, i = 1, \ldots, n\}$ are derived such that, similarly to (3.10)–(3.11),

$$\sum_{i=1}^{n} \frac{(1 - T_i)\hat{\pi}(X_i)}{w_i} c_0(X_i) = \sum_{i=1}^{n} \hat{\pi}(X_i) c_0(X_i).$$

Qin & Zhang (2008) studied asymptotic behavior of their estimator under a correctly specified PS model, but did not investigate local efficiency or double robustness and hence did not address the question of how $c_0(X)$ should be specified to gain efficiency or robustness over non-augmented IPW estimators. For our current setting, Graham et al. (2015) showed that their estimator is locally semiparametric efficient with respect to PS model (2.5) under condition R, and doubly robust under

Condition R^+: condition R holds, PS model (2.5) is logistic regression, and

$$c_0(X) = c_1(X) = f(X).$$

These results can be related to our results as follows.

First, similarly as discussed after Proposition 3.6, the semiparametric efficiency bound V^t_{SP} with respect to model (2.5) coincides with the nonparametric efficiency bound V^t_{NP} when model (2.5) is logistic regression and $\{T - \pi(X)\}\{m_t(X) - \nu^t\}$ is a linear combination of $\{T - \pi(X)\}f(X)$. Therefore, under Condition R^+, the estimator of Graham et al. (2015) is doubly robust and locally nonparametric as well as semiparamtric efficient. Second, if Condition R^+ holds, then Condition L holds and hence $\tilde{\pi}(X)$ reduces to $\hat{\pi}(X)$. In this case, our estimators $\tilde{\nu}^t_{\mathrm{reg}}$ and $\tilde{\nu}^t_{\mathrm{lik}}$, while using $\hat{\pi}(X)$ directly, are not only doubly robust and locally nonparametric efficient, but also intrinsically efficient among the class of estimator (3.4) with $\tilde{\pi}(X)$ the same as $\hat{\pi}(X)$. The estimator of Graham et al. (2015) can be shown to be asymptotically equivalent, to the first order, to some estimator in class (3.4) under a correctly specified PS model (2.5). Therefore, under Condition R^+, our estimators are proved to be asymptotically at least as efficient as the estimator of Graham et al. (2015) when the PS model (2.5) is correctly specified but the OR model (2.3) is misspecified. Finally, our work handles the general case where PS model (2.5) is non-logistic regression, and leads to both AIPW estimators that are doubly robust and locally nonparametric efficient, but also improved estimators that further achieve intrinsic efficiency.

If PS model (2.5) is logistic regression, then the methods of Hainmueller (2012) and Imai & Ratkovic (2014) seem to use the same estimator of ν^0,

$$\hat{\nu}^0_{\mathrm{HIR}} = \frac{1}{n_1}\sum_{i=1}^{n}(1-T_i)r(X_i;\breve{\gamma})Y_i = \frac{\sum_{i=1}^{n}(1-T_i)r(X_i;\breve{\gamma})Y_i}{\sum_{i=1}^{n}(1-T_i)r(X_i;\breve{\gamma})},$$

where $r(X;\gamma) = \pi(X;\gamma)/\{1-\pi(X;\gamma)\} = \exp\{\gamma^{\mathrm{T}} f(X)\}$ and $\breve{\gamma}$ is determined from the balancing equation similar to Eq. (3.10)–(3.11),

$$\sum_{i=1}^{n}(1-T_i)r(X_i;\gamma)f(X_i) = \sum_{i=1}^{n} T_i f(X_i). \tag{3.12}$$

Eq. (3.12) differs from balancing equations used for ATE estimation in Imai & Ratkovic (2014). The two expressions of $\hat{\nu}^0_{\mathrm{HIR}}$ follow from the fact that $\sum_{i=1}^{n}(1-T_i)r(X_i;\gamma) = n_1$ by Eq. (3.12) with $f(X)$ including a constant. That is, $\hat{\nu}^0_{\mathrm{HIR}}$ can be seen as standard IPW estimators: $\hat{\nu}^0_{\mathrm{HIR}} = \hat{\nu}^0_{\mathrm{IPW}}(\breve{\pi}) = \hat{\nu}^0_{\mathrm{IPW,ratio}}(\breve{\pi})$, where $\breve{\pi}(X) = \pi(X;\breve{\gamma})$ is substituted for $\hat{\pi}(X) = \pi(X;\hat{\gamma})$ with the MLE $\hat{\gamma}$. Under Condition L, the estimator $\hat{\nu}^0_{\mathrm{HIR}}$ can be shown to be doubly robust and locally nonparametric efficient (Zhao & Percival, 2015). However, $\hat{\nu}^0_{\mathrm{HIR}}$ is not intrinsically efficient and hence, similarly to the estimator of Graham et al. (2015), not as efficient as our estimators $\tilde{\nu}^0_{\mathrm{reg}}$ and $\tilde{\nu}^0_{\mathrm{lik}}$ when the PS model (2.5) is correctly specified but the OR model (2.3) is misspecified.

3.5 Simulation studies

We conducted two simulation studies to compare the proposed and existing estimators. We present in Section (3.9) the results under the simulation settings of Kang & Schafer (2007) and McCaffrey et al. (2007). Here we present the results under the simulation settings of Qin & Zhang (2008) and Graham et al. (2015).

The simulation setting of Qin & Zhang (2008) is originally designed in the context of difference-in-differences estimation, but can be equivalently recast for estimation of ATT as shown in Graham et al. (2015). Specifically, suppose that the covariate vector, $X = (X_1, X_2)$, is generated as

$$X_1 \sim N(0,1), \quad X_2|X_1 \sim N(1+0.6X_1, 1).$$

The true propensity score is generated as a logistic regression function

$$\pi(X) = P(T=1|X) = \text{expit}(\gamma_0^* + \gamma_1^* X_1 + \gamma_2^* X_2),$$

where $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.1, 0.1)$, $(1.0, 0.2, 0.2)$, or $(1.0, 0.5, 0.5)$, corresponding to increasing selection bias into treatment. The potential outcomes (Y^1, Y^0) are generated (regardless of T for exogenity) as

$$Y^1|X,T \sim N\{m_1(X), X_2^2\}, \quad Y^0|X,T \sim N\{m_0(X), X_2^2\},$$

where $m_0(X)$ and $m_1(X)$ are set in two possible ways:

(i) LIN-OR: $m_1(X) = 2 + 2X_1 + 2X_2$, $m_0(X) = 2X_1 + 2X_2$,

(ii) QUA-OR: $m_1(X) = 2 + 2X_1^2 + 3X_2^2 - X_2$, $m_0(X) = 2X_1^2 + 3X_2^2 - X_2$.

It is easily shown that the true value of ATT is always 2, because the regression functions $m_0(X)$ and $m_1(X)$ are parallel to each other.

For estimation of ATT, consider an outcome regression model (2.3) with the identity link $\Psi(\cdot)$ and the regressor vector $g_0(X) = g_1(X) = (1, X_1, X_2)^{\mathrm{T}}$ or $(1, X_1^2, X_2^2)^{\mathrm{T}}$, corresponding to a linear or quadratic OR model. Under the LIN-OR setting, the linear or quadratic OR model is, respectively, correctly specified or misspecified. Under the QUA-OR setting, both of the OR models are misspecified, but the quadratic OR model is misspecified to a lesser degree. Similarly, consider a propensity score model (2.5) with the logistic link $\Pi(\cdot)$ and the regressor vector $f(X) = (1, X_1, X_2)^{\mathrm{T}}$ or $(1, X_1^2, X_2^2)^{\mathrm{T}}$, corresponding to a logistic linear or quadratic PS model, which is, respectively, correctly specified or misspecified.

We implemented the following estimators of ATT:

- (OR) $\hat{\mu}^1_{\mathrm{OR}} - \hat{\mu}^1_{\mathrm{OR}}$;
- (IPW.r) $\hat{\mu}^1_{\mathrm{NP}} - \hat{\mu}^0_{\mathrm{IPW,ratio}}(\hat{\pi})$;

- (AIPW) $\hat{\mu}^1_{\rm NP} - \hat{\mu}^0_{\rm NP}(\hat{\pi}, \hat{m}_0)$;
- (LIK) $\tilde{\mu}^1_{\rm lik} - \tilde{\mu}^0_{\rm lik}$, (LIK2) $\tilde{\mu}^1_{\rm lik2} - \tilde{\mu}^0_{\rm lik2}$;
- (HIR) $\hat{\mu}^1_{\rm NP} - \hat{\mu}^0_{\rm IPW}(\breve{\pi})$, (AIPW.HIR) $\hat{\mu}^1_{\rm NP} - \hat{\mu}^0_{\rm NP}(\breve{\pi}, \hat{m}_0)$.

Table 3.2: Qin–Zhang simulation results with $(\gamma^*_1, \gamma^*_2, \gamma^*_3) = (1.0, 0.2, 0.2)$

Models	OR	IPW.r	AIPW	LIK	LIK2	HIR	AIPW.HIR	EL	AST
				Data generated under LIN-OR setting					
linear PS,	0.0120	0.0147	0.0125	0.0118	0.0120	0.0123	0.0123	0.0031	-0.0065
linear OR	(0.0175)	(0.0358)	(0.0201)	(0.0209)	(0.0208)	(0.0200)	(0.0200)	(0.0275)	(0.0261)
linear PS,	0.7170	0.0147	0.0139	0.0168	0.0132	0.0123	0.0122	-0.0009	-0.0039
quadratic OR	(0.0767)	(0.0358)	(0.0500)	(0.0221)	(0.0225)	(0.0200)	(0.0431)	(0.0306)	(0.0371)
quadratic PS,	0.0120	0.6655	0.0106	0.0125	0.0105	0.7501	0.0114	···	···
linear OR	(0.0175)	(0.0878)	(0.0269)	(0.0212)	(0.0221)	(0.0756)	(0.0244)	···	···
quadratic PS,	0.7170	0.6655	0.7644	0.7023	0.7120	0.7501	0.7501	···	···
quadratic OR	(0.0767)	(0.0878)	(0.0828)	(0.0746)	(0.0746)	(0.0756)	(0.0756)	···	···
				Data generated under QUA-OR setting					
linear PS,	0.7028	0.0414	0.0500	0.0471	0.0522	0.0553	0.0553	0.0477	0.0787
linear OR	(0.4176)	(0.6407)	(0.5201)	(0.0796)	(0.0946)	(0.3683)	(0.3683)	(0.1227)	(0.3620)
linear PS,	-0.1473	0.0414	0.0142	0.0120	0.0138	0.0553	0.0144	0.0028	0.0078
quadratic OR	(0.0238)	(0.6407)	(0.0216)	(0.0224)	(0.0221)	(0.3683)	(0.0223)	(0.0309)	(0.0218)
quadratic PS,	0.7028	-0.4155	-0.6468	0.0549	0.1256	-0.1554	-0.4657	···	···
linear OR	(0.4176)	(0.6381)	(0.6286)	(0.0485)	(0.1044)	(0.0249)	(0.1021)	···	···
quadratic PS,	-0.1473	-0.4155	-0.1599	-0.1465	-0.1493	-0.1554	-0.1554	···	···
quadratic OR	(0.0238)	(0.6381)	(0.0263)	(0.0272)	(0.0258)	(0.0249)	(0.0249)	···	···

Note: In the upper rows are the Monte Carlo biases (= means−2), and in the brackets are the corresponding Monte Carlo variances. EL: Qin & Zhang (2008) and AST: Graham et al. (2015).

Table 3.2 and Figures 3.1-3.2 present the results for these estimators, from 1000 Monte Carlo samples of size $n = 1000$, under the PS setting with moderate selection bias, $(\gamma^*_1, \gamma^*_2, \gamma^*_3) = (1.0, 0.2, 0.2)$. In addition, results are reproduced under the same setting for two estimators in Qin & Zhang (2008) and Graham et al. (2015). See Section (3.9) for the results under the other two PS settings.

The following remarks can be drawn on the comparisons of various estimators. First, the OR estimator is approximately unbiased only when the OR model used is correctly specified (i.e., linear OR model under LIN-OR setting).

Second, the IPW.ratio estimators are approximately unbiased only when the PS model used is correctly specified (i.e., linear PS model), but they have large variances

with noticeably outlying values.

Third, the HIR estimator is approximately unbiased when the PS model is correctly specified, but becomes biased when the PS model is misspecified and even when the OR model is correctly specified (for example, quadratic PS model and linear OR model under LIN-OR setting). The HIR estimator is not doubly robust, because Condition L is not satisfied in this situation.

Fourth, the four estimators, AIPW, LIK, LIK2, and AIPW.HIR are doubly robust: they are approximately unbiased when either the PS model is correctly specified (i.e., linear PS model) or the OR model is correctly specified (i.e., linear OR model under LIN-OR setting). In accordance with local efficiency, these estimators have similar variances to each other when both the PS and OR models are correctly specified. But LIK and LIK2 have smaller variances, sometimes substantially so, than AIPW and AIPW.HIR estimators when the PS model is correctly specified but the OR model is misspecified. For example, for linear PS model and linear OR model under QUA-OR setting, the variance of LIK is smaller than that of AIPW by a factor of $0.52/0.08 = 6.5$ and that of AIPW.HIR by a factor of $0.37/0.08 \approx 4.6$. Such differences are supported by our theoretical results on intrinsic efficiency.

Fifth, in contrast with AIPW and AIPW.HIR, the LIK estimator appears to be approximately unbiased when the quadratic PS model and linear OR model are used under the QUA-OR setting (hence both PS and OR models are misspecified). This behavior is not indicated by general theory, but can be explained by the fact that even though the PS model (2.5) is misspecified, the augmented PS model (3.2) happens to be correctly specified in this case: $\{\hat{m}_0(X), \hat{m}_1(X)\}$ provide exactly the correct regressors (X_1, X_2) up to linear transformation.

Finally, we compare our likelihood estimators with the estimators in Qin & Zhang (2008) and Graham et al. (2015) when the PS model is correctly specified (i.e., linear PS model). Results for a misspecified PS model were not available in these previous simulation studies. Similarly as in the comparisons with AIPW and AIPW.HIR, our likelihood estimators have smaller variances than those in Qin & Zhang and Graham et al. when the PS model is correctly specified but the OR model is misspecified. For

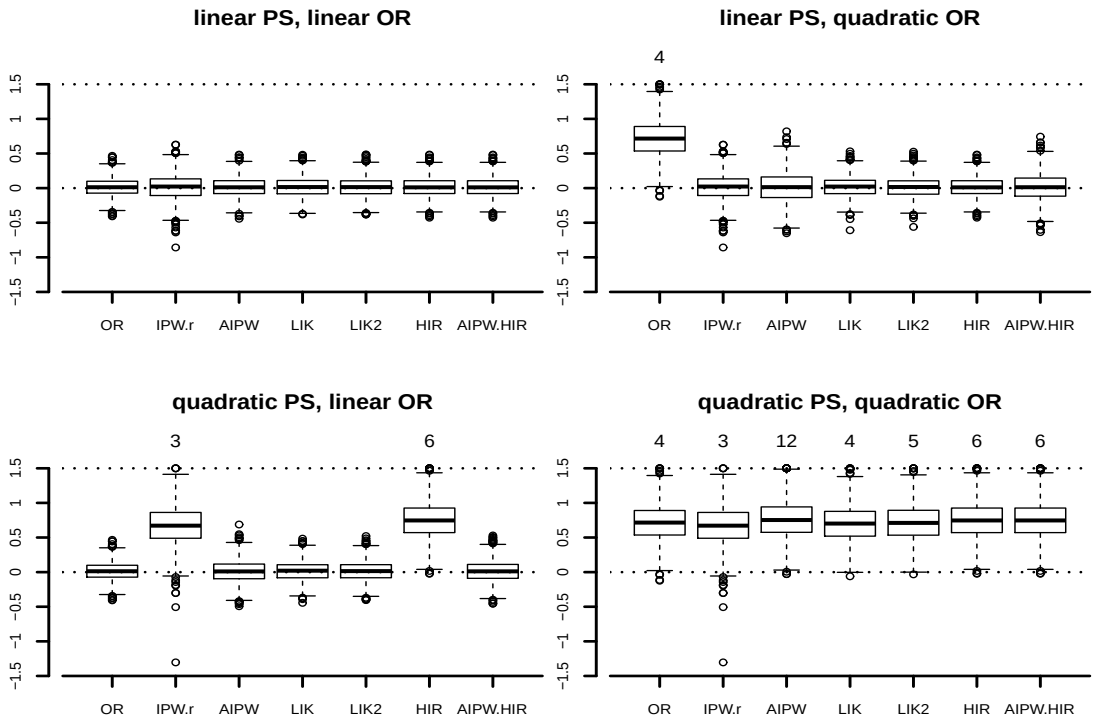

Figure 3.1: Boxplots of estimates minus the truth under LIN-OR setting with $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.2, 0.2)$. All values are censored within the range of y-axis, and the number of values that lie outside the range are indicated next to the lower and upper limits of y-axis.

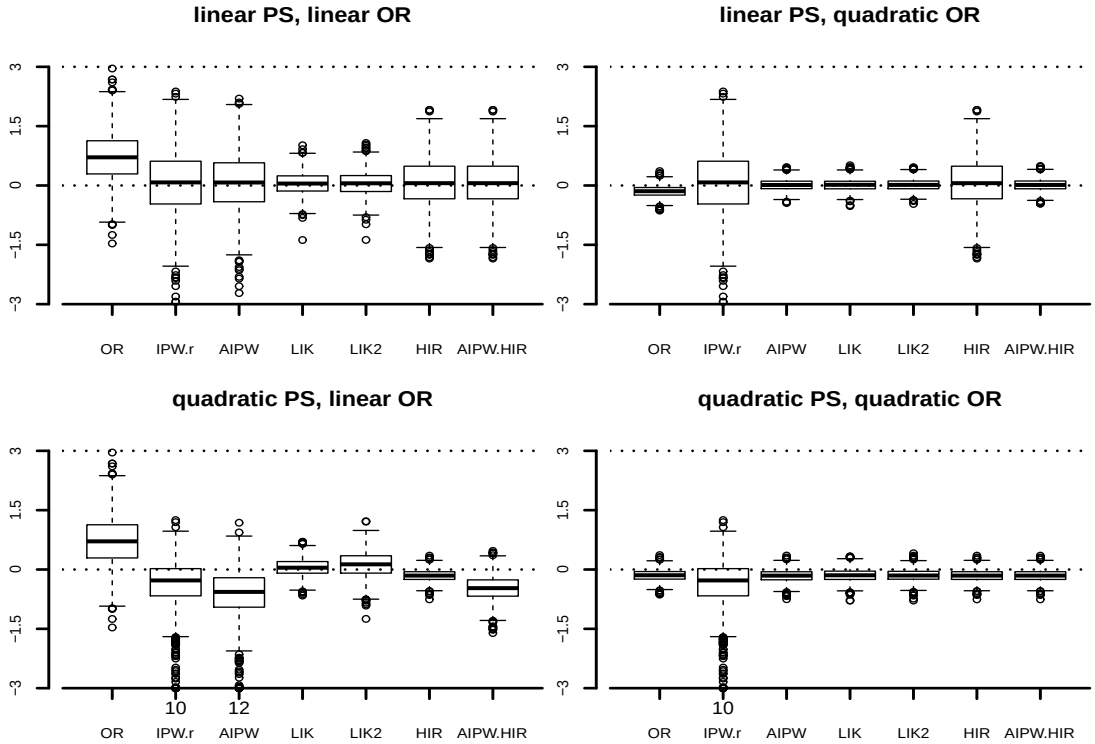

Figure 3.2: Boxplots of estimates minus the truth under QUA-OR setting with $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.2, 0.2)$

example, for linear PS model and linear OR model under QUA-OR setting, the variance of LIK is smaller than that of Qin & Zhang by a factor of $0.12/0.08 = 1.5$ and that of Graham et al. by a factor of $0.36/0.08 = 4.5$. Another interesting observation is that when the OR model is also correctly specified or approximately so, our likelihood estimators and Graham et al. have similar variances, but smaller than that of Qin & Zhang estimator, indicating a lack of local efficiency for the latter estimator. For example, the factor of efficiency gain is $.031/0.022 \approx 1.4$ for linear PS model and quadratic OR model under the QUA-OR setting.

3.6 Analysis of LaLonde data

NSW ("National Supported Work Demonstration") is a randomized job training program implemented in 1970s to provide work experience for individuals who had economic and social disadvantages. The randomized experiment provides benchmark estimates of average treatment effects. To study econometric methods for program evaluation with non-experimental data, LaLonde (1986) constructed an observational study by replacing the data from the experimental control group with survey data from either Current Population Survey (CPS) or the Panel Study of Income Dynamics (PSID). The question of interest is how well the experimental benchmark estimates of average treatment effects can be recovered by econometric methods when applied to such composite observational studies. LaLonde (1986) showed that many commonly used methods failed to replicate the experimental results.

Analysis of LaLonde's composite data has since been extensively discussed in the evaluation and causal inference literature. Dehejia & Wahba (1999, 2002) obtained effect estimates that have low biases from the experimental benchmark, while applying propensity score matching methods to a particular subsample of LaLonde's original data. Smith & Todd (2005a) raised the criticism that the propensity score matching estimates are highly sensitive to both the analysis sample used and the specification of propensity score models. They calculated direct estimates of the bias by applying matching to the experimental control group and a non-experimental comparison group (either CPS or PSID), whereas LaLonde and Dehejia & Wahba calculated the bias by

applying matching to the experimental treatment group and a non-experimental comparison group and then comparing the resulting estimate to the experimental benchmark. See Diamond & Sekhon (2013), Hainmueller (2012), and Imai & Ratkovic (2014), among others, for more recent analyses.

We investigate the performances of the proposed and existing estimators for analyzing LaLonde's original composite data. Specifically, we apply various estimators of ATT as listed in Section 3.5 in the following analyses:

- Analysis (i): NSW experimental treatment group is combined with either CPS or PSID non-experimental comparison group for effect estimation or, equivalently, for bias estimation by subtracting the experimental benchmark from all effect estimates;

- Analysis (ii): NSW experimental control group is combined with either CPS or PSID non-experimental comparison group for bias estimation.

For each application, we consider two possible PS models and two possible OR models, as specified in Table 3.3. The quadratic PS model differs, only by a few terms, from the PS model obtained in an iterative model-building approach by Dehejia & Wahba (2002) for analyzing NSW+CPS or NSW+PSID composite data.

For propensity score estimation, we use either the experimental treatment group in (i) or experimental control group in (ii) as treated observations ($T = 1$) and the non-experimental comparison group as untreated observations ($T = 0$). This strategy is in line with LaLonde (1986) and Dehejia & Wahba (1999, 2002), but differs from Smith & Todd (2005a) and Imai & Ratkovic (2014). In the latter articles, both the experimental treatment and control groups are used as treated observations ($T = 1$) when estimating propensity scores, but then either the experimental treatment or control group is used in, respectively, effect or bias estimation. This scheme does not mimic the practical situation of econometric analysis where a single dataset is used, and may not even be desirable as discussed in Dehejia (2005b).

Before turning to our results, we provide some remarks to explain how the relative performances of estimators will be assessed from such empirical results. First,

as discussed in Dehejia (2005a) in response to Smith & Todd (2005a), applications of propensity score methods should involve searching for a propensity score model that leads to balance of covariates between treatment groups. The approach suggested in Rosenbaum & Rubin (1984) and Dehejia & Wahba (1999, 2002) is conceptually useful but leaves open the issue of how PS models can actually be built to achieve covariate balance. Alternatively, simple PS models such as in Table 3.3 may often be used in applied research. Second, Smith & Todd (2005b) presented additional analyses in response to Dehejia (2005a) to argue that the low-bias matching estimates in Dehejia & Wahba (1999, 2002) are sensitive not only in regard to the sample and propensity score specification as shown in Smith & Todd (2005a), but also, among other factors, to whether the propensity score and subsequently the bias are estimated using the experimental treatment or control group, as in Analyses (i) and (ii) described above. Third, a criterion typically used in previous analyses of LaLonde data is that the bias estimates should be close to 0 for a good method. But the true bias can be 0 only when the exogeneity assumption (A1) holds on the composite sample, i.e., potential outcomes are influenced by the measured covariates in the experimental sample in the same way as in the comparison sample (CPS or PSID). Nevertheless, the difference between the two bias estimates from Analyses (i) and (ii), as examined in Smith & Todd (2005b), can be shown to be 0 (up to random variation) even when the exogeneity assumption (A1) fails on the composite sample. See "Violation of the exogeneity assumption" in Appendix for details. By all the preceding considerations, we will assess the relative performances of estimators mainly in terms of how close the two bias estimates from Analyses (i) and (ii) are to each other, depending on PS and OR models used.

Table 3.4 and Figure 3.3 present the results from Analyses (i) and (ii) for various estimators as listed in Section 3.5, based on 500 bootstrap samples of the NSW+PSID composite data. See Section (3.9) for the results on the NSW+CPS composite data, where the relative performances of estimators are more similar to each other than on the NSW+PSID composite data.

Among all estimators studied, the IPW.ratio estimator yields point estimates of effect closest to the experimental benchmark $886 and estimates of bias closest to 0 from

Table 3.3: PS and OR models for LaLonde data

Name	Regressors $f(X)$ in PS model or $g(X)$ in OR model
Linear	$(1, age, school, black, hisp, married, nodegr, re74, re75, u74, u75)$
Quadratic	$(1, age, school, black, hisp, married, nodegr, re74, re75, u74, u75, age^2, school^2, re74^2, re75^2)$

Note: The variables are defined as in Table 2 of Dehejia & Wahba (2002). The PS model is $T|X \sim f(X)$ with logistic link. The OR model is $Y|(T = t, X) \sim t + g(X)$ with identity link.

Table 3.4: Bootstrap results from Analyses (i) and (ii) on NSW+PSID composite data

		OR	IPW.ratio	AIPW	LIK2	HIR	AIPW.HIR
Linear PS, Linear OR	Treatment Effect	-1690	901	1109	555	475	475
		(650)	(781)	(852)	(616)	(598)	(598)
	Evaluation Bias	-2941	-6	337	-211	-118	-118
		(636)	(764)	(815)	(523)	(496)	(496)
	Difference	1251	907	772	765	594	594
		(590)	(669)	(757)	(563)	(549)	(549)
Linear PS, Quadratic OR	Treatment Effect	-1577	901	613	378	475	441
		(803)	(781)	(729)	(613)	(598)	(601)
	Evaluation Bias	-2674	-6	-7	-365	-118	-177
		(807)	(764)	(653)	(529)	(496)	(501)
	Difference	1096	907	620	743	594	618
		(610)	(669)	(641)	(560)	(549)	(553)
Quadratic PS, Linear OR	Treatment Effect	-1690	901	1216	573	393	477
		(650)	(799)	(896)	(623)	(606)	(601)
	Evaluation Bias	-2941	-9	451	-236	-254	-142
		(636)	(791)	(862)	(537)	(505)	(498)
	Difference	1251	910	765	809	647	618
		(590)	(685)	(804)	(560)	(556)	(547)
Quadratic PS, Quadratic OR	Treatment Effect	-1577	901	571	332	393	393
		(803)	(799)	(742)	(618)	(606)	(606)
	Evaluation Bias	-2674	-9	-27	-429	-254	-254
		(807)	(791)	(666)	(531)	(505)	(505)
	Difference	1096	910	598	761	647	647
		(610)	(685)	(651)	(560)	(556)	(556)

Note: In the upper rows are the bootstrap means, and in the brackets are the corresponding bootstrap standard errors. Treatment Effect is obtained from Analysis (i), and Evaluation Bias from Analysis (ii). The difference is to be compared with the experimental benchmark \$886 with standard error \$488. To tackle numerical non-convergence when computing estimates during bootstrapping, the following procedure is used. We performed Principle Component Analysis to the regressors from the composite data, NSW (treatment+control) + PSID, and dropped principle components whose sample variances are less than $(0.3)^2$ of the component with the largest sample variance. Then we resampled the entire composite dataset and conducted Analyses (i) and (ii) on each bootstrap sample.

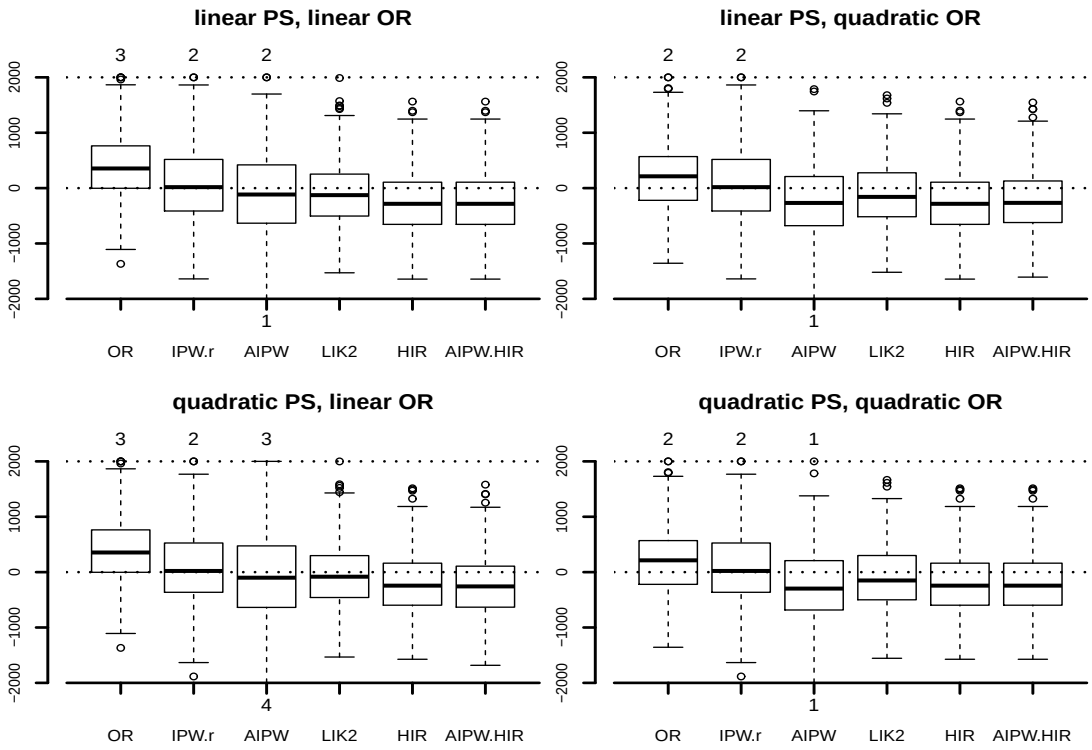

Figure 3.3: Bootstrap boxplots of differences of bias estimates from Analyses (i) and (ii) on NSW+PSID composite data. All values are censored within the range of y-axis, with number of values laying outside indicated next to the lower and upper limits of y-axis.

Analyses (i) and (ii), using either the linear or quadratic PS model. But the bootstrap variances for IPW.ratio are among the highest for all estimators studied. Although such point estimates of effect are much closer to the experimental benchmark than various previously obtained estimates on LaLonde NSW+PSID data (e.g., Diamond & Sekhon 2013; Imai & Ratkovic 2014), these results may not present real evidence for any advantage of IPW.ratio for reasons discussed above.

In terms of how close the difference between effect and bias estimates is to the experimental benchmark (i.e., how close the two bias estimates are close to each other) from Analyses (i) and (ii), the estimators IPW.ratio, AIPW, and LIK2 yield the most accurate point estimates among all estimators studied, regardless of PS and OR models used. But the bootstrap variances for LIK2 are much smaller than those of IPW.rato and AIPW. As explained above, these results present strong evidence for the advantage of the proposed estimator LIK2.

3.7 Conclusion

We study the problem of estimating ATTs from observational data and make the following contributions. In spite of non-ancillarity of the propensity score, we show how efficient influence functions from semiparametric theory can be harnessed to derive AIPW estimators that are locally efficient and doubly robust. Furthermore, we develop calibrated regression and likelihood estimators that achieve desirable properties in efficiency and boundedness beyond local efficiency and double robustness. From two simulation studies and reanalysis of LaLonde (1986) data, the proposed methods perform overall the best compared with various existing methods.

The ideas developed in this article can be extended in various directions. For example, it is interesting to consider marginal and nested structural models for ATTs in subpopulations, i.e., $E(Y^1 - Y^0|T = 1, V)$ with some selected covariates V, and develop calibrated regression and likelihood estimators. Moreover, as seen from Graham et al. (2015), estimation of ATT can be put in a broader class of data combination problems. The methods developed here can be extended in that direction.

3.8 Appendix

Throughout, we make the following assumptions regarding the estimators $\hat{\alpha}_t$ for OR model (2.3), $\hat{\gamma}$ for PS model (2.5), and $(\tilde{\gamma}, \tilde{\delta})$ for augmented PS model (3.2), allowing for possible model misspecification (e.g., White 1982).

(C1) Assume that $\hat{\alpha}_t$ converges to a constant α_t^* such that $\hat{\alpha}_t - \alpha_t^* = O_p(n^{-1/2})$ for $t = 0, 1$. Write $m_t^*(X) = m_t(X; \alpha_t^*)$. If model (2.3) is correctly specified, then $m_t^*(X) = m_t(X)$. In general, $m_t^*(X)$ and $m_t(X)$ may differ from each other.

(C2) Assume that $\hat{\gamma}$ converges to a constant γ^* such that

$$\hat{\gamma} - \gamma^* = V^{-1} \tilde{E} \{s_{\gamma^*}(T, X)\} + o_p(n^{-1/2}),$$

where $E\{s_{\gamma^*}(T, X)\} = 0$, and the matrix $V = -E\{\partial s_\gamma(T, X)/\partial \gamma^{\mathrm{T}}\}|_{\gamma=\gamma^*}$ is non-singular. Write $\pi^*(X) = \pi(X; \gamma^*)$. If model (2.5) is correctly specified, then

$\pi^*(X) = \pi(X)$ and $V = \mathrm{var}\{s_{\gamma^*}(T, X)\}$. In general, $\pi^*(X)$ and $\pi(X)$ may differ from each other.

(C3) For augmented PS model (3.2), define

$$s^\dagger(T, X; \gamma, \delta, \alpha) = \{T - \pi_{\mathrm{aug}}(X; \gamma, \delta, \alpha)\}\{f^{\mathrm{T}}(X), m_0(X; \alpha_0), m_1(X; \alpha_1)\}^{\mathrm{T}}.$$

Assume that $(\tilde{\gamma}, \tilde{\delta})$ converges to a constant $(\gamma^\dagger, \delta^*)$ such that

$$\begin{pmatrix} \tilde{\gamma} - \gamma^\dagger \\ \tilde{\delta} - \delta^* \end{pmatrix} = V^{\dagger -1}\, \tilde{E}\left\{ s^\dagger(T, X; \gamma^\dagger, \delta^*, \hat{\alpha}) \right\} + o_p(n^{-1/2}),$$

where $E\{s^\dagger(T, X; \gamma^\dagger, \delta^*, \alpha^*)\} = 0$, and the matrix $V^\dagger = -E\{\partial s^\dagger(T, X; \gamma, \delta, \alpha^*)/\partial(\gamma^{\mathrm{T}}, \delta^{\mathrm{T}})\}|_{(\gamma,\delta)=(\gamma^\dagger,\delta^*)}$ is nonsingular. Write $\pi^\dagger(X) = \pi_{\mathrm{aug}}(X; \gamma^\dagger, \delta^*, \alpha^*)$. If model (2.5) is correctly specified, then $(\gamma^\dagger, \delta^*) = (\gamma^*, 0)$, $\pi^\dagger(X) = \pi(X)$, $V^\dagger = \mathrm{var}\{s^\dagger(T, X; \gamma^*, 0, \alpha^*)\}$, and the asymptotic expansion for $(\tilde{\gamma}, \tilde{\delta})$ reduces to

$$\begin{pmatrix} \tilde{\gamma} - \gamma^* \\ \tilde{\delta} \end{pmatrix} = V^{\dagger -1}\, \tilde{E}\left\{ s^\dagger_{(\gamma^*,0)}(T, X) \right\} + o_p(n^{-1/2}),$$

where $s^\dagger_{(\gamma^*,0)}(T, X) = s^\dagger(T, X; \gamma^*, 0, \alpha^*)$.

In addition, we assume that the following regularity conditions hold (e.g., Robins et al. 1994, Appendix B).

(C4) $E\{(Y^t)^2\} < \infty$ and $E\{m_t^{*2}(X)\} < \infty$ for $t = 0, 1$.

(C5) There exists $\epsilon > 0$ such that $\pi^*(x) \le 1 - \epsilon$ and $\pi^\dagger(x) \le 1 - \epsilon$ for all x.

(C6) There exists a neighborhood $N_{1,t}$ of α_t^* such that $E\{\sup_{\alpha_t \in N_{1,t}} \|\partial m_t(X; \alpha_t)/\partial \alpha_t\|^2\} < \infty$ for $t = 0, 1$, where $\|A\| = (\sum_{ij} A_{ij}^2)^{1/2}$ for any matrix with element A_{ij}.

(C7) There exists a neighborhood N_2 of γ^* such that $E\{\sup_{\gamma \in N_2} \|\partial \pi(X; \gamma)/\partial \gamma\|^2\} < \infty$ and $E\{\sup_{\gamma \in N_2} \|\partial^2 \pi(X; \gamma)/\partial \gamma \partial \gamma^{\mathrm{T}}\|^2\} < \infty$.

(C8) There exists a neighborhood N_3 of $(\gamma^*, \delta^*, \alpha^*)$ such that $E\{\sup_{\theta\in N_3} \|\partial\pi_{\text{aug}}(X;\theta)/\partial\theta\|^2\} < \infty$ and $E\{\sup_{\theta\in N_3} \|\partial^2\pi_{\text{aug}}(X;\theta)/\partial\theta\partial\theta^{\mathrm{T}}\|^2\} < \infty$, with $\theta = (\gamma^{\mathrm{T}}, \delta^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}$.

We provide the following lemma on asymptotic expansions of AIPW estimators.

Lemma 3.9 *Assume that $E\{h^2(X)\} < \infty$. If the PS model (2.5) is correctly specified, then the following results hold.*

(i) $\hat{\nu}^0(\hat{\pi}, h)$ admits the asymptotic expansion,

$$\hat{\nu}^0(\hat{\pi}, h) - \nu^0 = q^{-1}\tilde{E}\Big(\phi_h^0(Y,T,X) - T\nu^0 - \Pi\{\phi_h^0(Y,T,X)|s_{\gamma^*}(T,X)\} + \Pi\left[\{T-\pi(X)\}m_0(X)|s_{\gamma^*}(T,X)\right]\Big) + o_p(n^{-1/2}),$$

where $\phi_h^0(Y,T,X) = [(1-T)/\{1-\pi(X)\}]\pi(X)Y - [(1-T)/\{1-\pi(X)\}-1]h(X)$.

(ii) Define $\hat{\nu}^1(\hat{\pi}, h) = \tilde{E}[TY - \{T-\hat{\pi}(X)\}h(X)]/\tilde{E}(T)$. Then $\hat{\nu}^1(\hat{\pi}, h)$ admits the asymptotic expansion,

$$\begin{aligned}
&\hat{\nu}^1(\hat{\pi}, h) - \nu^1 \\
=&q^{-1}\tilde{E}\Big(\phi_h^1(Y,T,X) - T\nu^1 + \Pi\left[\{T-\pi(X)\}h(X)|s_{\gamma^*}(T,X)\right]\Big) + o_p(n^{-1/2}), \\
=&q^{-1}\tilde{E}\Big(\phi_h^1(Y,T,X) - T\nu^1 - \Pi\{\phi_h^1(Y,T,X)|s_{\gamma^*}(T,X)\} \\
&\quad + \Pi\left[\{T-\pi(X)\}m_1(X)|s_{\gamma^*}(T,X)\right]\Big) + o_p(n^{-1/2}),
\end{aligned}$$

where $\phi_h^1(Y,T,X) = TY - \{T-\pi(X)\}h(X)$.

Proof of Lemma 4.5. By direct calculation and Slutsky theorem, we have

$$\hat{\nu}^0(\hat{\pi}, h) - \nu^0 = q^{-1}\tilde{E}\left[\frac{1-T}{1-\hat{\pi}(X)}\hat{\pi}(X)Y - \left\{\frac{1-T}{1-\hat{\pi}(X)} - 1\right\}h(X) - T\nu^0\right] + o_p(n^{-1/2}).$$

By a Taylor expansion for $\hat{\gamma}$ about γ^* and direct calculation, we have

$$\begin{aligned}
&\tilde{E}\left[\frac{1-T}{1-\hat{\pi}(X)}\pi(X)Y-\left\{\frac{1-T}{1-\hat{\pi}(X)}-1\right\}h(X)\right]\\
=&\tilde{E}\left[\frac{1-T}{1-\pi(X)}\pi(X)Y-\left\{\frac{1-T}{1-\pi(X)}-1\right\}h(X)\right]\\
&\quad+E\left[\frac{1-T}{\{1-\pi(X)\}^2}\frac{\partial\pi(X;\gamma^*)}{\partial\gamma}\{\pi(X)Y-h(X)\}\right](\hat{\gamma}-\gamma^*)+o_p(n^{-1/2})\\
=&\tilde{E}\Big(\phi_h^0(Y,T,X)-\Pi\{\phi_h^0(Y,T,X)|s_{\gamma^*}(T,X)\}\Big)+o_p(n^{-1/2}).
\end{aligned}$$

By similar arguments, we have

$$\begin{aligned}
&\tilde{E}\left[\frac{1-T}{1-\hat{\pi}(X)}\{\hat{\pi}(X)-\pi(X)\}Y\right]\\
=&E\left[\frac{1-T}{1-\pi(X)}\frac{\partial\pi(X;\gamma^*)}{\partial\gamma}Y\right](\hat{\gamma}-\gamma^*)+o_p(n^{-1/2})\\
=&\tilde{E}\Big(\Pi\left[\{T-\pi(X)\}m_0(X)|s_{\gamma^*}(T,X)\right]\Big)+o_p(n^{-1/2}).
\end{aligned}$$

Combining the preceding three expansions gives the desired expansion for $\hat{\nu}^0(\hat{\pi},h)$. Similarly, the expansion for $\hat{\nu}^1(\hat{\pi},h)$ can be shown. □

Proofs of Propositions 3.2 & 3.4

First, we show the local nonparametric efficiency of $\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi},\hat{m}_0)$. If both model (2.3) for $t=0$ and model (2.5) are correctly specified, then by Slutsky theorem,

$$\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi},\hat{m}_0)=\tilde{E}\left[\frac{1-T}{1-\hat{\pi}(X)}\hat{\pi}(X)Y-\left\{\frac{1-T}{1-\hat{\pi}(X)}-1\right\}m_0(X)\right]\Big/\tilde{E}(T)+o_p(n^{-1/2}).$$

The leading term can be reexpressed as

$$\tilde{E}\left[\frac{1-T}{1-\hat{\pi}(X)}\hat{\pi}(X)\{Y-m_0(X)\}+Tm_0(X)\right]\Big/\tilde{E}(T)$$

and, by Slutsky theorem, approximated by

$$\tilde{E}\left[\frac{1-T}{1-\pi(X)}\pi(X)\{Y-m_0(X)\}+Tm_0(X)\right]\Big/\tilde{E}(T)+o_p(n^{-1/2}),$$

which gives the desired result. Alternatively, the result follows from Lemma 4.5(i) with $h(X) = m_0(X)$ and the fact that $\phi^0_{m_0}(Y,T,X) = \phi^0_{\pi m_0}(Y,T,X) + \{T - \pi(X)\}m_0(X)$ and hence $\Pi\{\phi^0_{m_0}(Y,T,X)|s_{\gamma^*}(T,X)\} = \Pi[\{T - \pi(X)\}m_0(X)|s_{\gamma^*}(T,X)\}$.

Second, we show the double robustness of $\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi}, \hat{m}_0)$. If PS model (2.5) is correctly specified, then $\tilde{E}([(1-T)/\{1-\hat{\pi}(X)\}-1]\hat{m}_0(X)) = \tilde{E}([(1-T)/\{1-\pi(X)\}-1]m^*_0(X)) + O_p(n^{-1/2}) = O_p(n^{-1/2})$ and hence

$$\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi}, \hat{m}_0) = \tilde{E}\left\{\frac{1-T}{1-\hat{\pi}(X)}\hat{\pi}(X)Y\right\}\Big/\tilde{E}(T) + O_p(n^{-1/2}) = \nu^0 + O_p(n^{-1/2}).$$

On the other hand, $\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi}, \hat{m}_0)$ can be reexpressed as

$$\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi}, \hat{m}_0) = \tilde{E}\left[\frac{1-T}{1-\hat{\pi}(X)}\hat{\pi}(X)\{Y - \hat{m}_0(X)\} + T\hat{m}_0(X)\right]\Big/\tilde{E}(T).$$

If OR model (2.3) for $t = 0$ is correctly specified, then $\tilde{E}([(1-T)/\{1-\hat{\pi}(X)\}]\hat{\pi}(X)\{Y - \hat{m}_0(X)\}) = \tilde{E}([(1-T)/\{1-\pi^*(X)\}]\pi^*(X)\{Y - m_0(X)\}) + O_p(n^{-1/2}) = O_p(n^{-1/2})$ and hence

$$\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi}, \hat{m}_0) = \tilde{E}\left\{T\hat{m}_0(X)\right\}\Big/\tilde{E}(T) = \nu^0 + O_p(n^{-1/2}).$$

Third, we show the local semiparametric efficiency of $\hat{\nu}^0_{\mathrm{SP}}(\hat{\pi}, \hat{m}_0)$. If both model (2.3) for $t = 0$ and model (2.5) are correctly specified, then

$$\begin{aligned}
&\hat{\nu}^0_{\mathrm{SP}}(\hat{\pi}, \hat{m}_0) - \nu^0 \\
=&q^{-1}\tilde{E}\left[\frac{1-T}{1-\hat{\pi}(X)}\hat{\pi}(X)Y - \left\{\frac{1-T}{1-\hat{\pi}(X)} - 1\right\}\hat{\pi}(X)m_0(X) - \hat{\pi}(X)\nu^0\right] + o_p(n^{-1/2}) \\
=&q^{-1}\tilde{E}\left[\frac{1-T}{1-\hat{\pi}(X)}\hat{\pi}(X)(Y - \nu^0) - \left\{\frac{1-T}{1-\hat{\pi}(X)} - 1\right\}\pi(X)\{m_0(X) - \nu^0\}\right] + o_p(n^{-1/2}).
\end{aligned}$$

by direct calculation and Slutsky theorem. Applying, to the above, Lemma 4.5(i) with Y replaced by $Y-\nu^0$ and $h(X)=\pi(X)\{m_0(X)-\nu^0\}$ yields

$$\hat{\nu}^0_{\mathrm{SP}}(\hat{\pi},\hat{m}_0)-\nu^0=q^{-1}\tilde{E}\Big(\phi^0_h(Y-\nu^0,T,X)-\Pi\{\phi^0_h(Y-\nu^0,T,X)|s_{\gamma^*}(T,X)\} \\ +\Pi\left[\{T-\pi(X)\}\{m_0(X)-\nu^0\}|s_{\gamma^*}(T,X)\right]\Big)+o_p(n^{-1/2}),$$

The desired results follows because $\phi^0_h(Y-\nu^0,T,X)=\tau^0(\pi,\pi m_0)-\pi(X)\nu^0$ by direct calculation, and the variable $\phi^0_h(Y-\nu^0,T,X)$ is uncorrelated with the score $s_{\gamma^*}(T,X)$ and hence $\Pi\{\phi^0_h(Y-\nu^0,T,X)|s_{\gamma^*}(T,X)\}=0$.

Finally, we show the local semiparametric efficiency of $\hat{\nu}^1_{\mathrm{SP}}(\hat{\pi},\hat{m}_1)$. If both model (2.3) for $t=1$ and model (2.5) are correctly specified, then

$$\hat{\nu}^1_{\mathrm{SP}}(\hat{\pi},\hat{m}_1)-\nu^1=q^{-1}\tilde{E}\left[TY-\{T-\hat{\pi}(X)\}m_1(X)-\hat{\pi}(X)\nu^1\right]+o_p(n^{-1/2}) \\ =q^{-1}\tilde{E}\left[T(Y-\nu^1)-\{T-\hat{\pi}(X)\}\{m_1(X)-\nu^1\}\right]+o_p(n^{-1/2}),$$

by direct calculation and Slutsky theorem. Applying, to the above, Lemma 4.5(ii) with Y replaced by $Y-\nu^1$ and $h(X)=m_1(X)-\nu^1$ yields

$$\hat{\nu}^1_{\mathrm{SP}}(\hat{\pi},\hat{m}_1)-\nu^1=q^{-1}\tilde{E}\Big(\phi^1_h(Y-\nu^1,T,X) \\ +\Pi\left[\{T-\pi(X)\}\{m_1(X)-\nu^1\}|s_{\gamma^*}(T,X)\right]\Big)+o_p(n^{-1/2}).$$

The desired result follows because $\phi^1_h(Y-\nu^1,T,X)=TY-\{T-\pi(X)\}m_1(X)-\pi(X)\nu^1$ by direct calculation. □

Proof of Proposition 3.6

First, it is straightforward to show that $\tilde{\beta}_t=\beta^*_t+o_p(1)$, where $\beta^*_t=E^{-1}(\xi^*_t\zeta^{*\mathrm{T}}_t)E(\xi^*_t\eta^*_t)$ and η^*_t, ξ^*_t, ζ^*_t, and $h^*(X)$ are defined as $\tilde{\eta}_t$, $\tilde{\xi}_t$, $\tilde{\zeta}_t$, and $\tilde{h}(X)$ respectively but with $\pi^\dagger(X)$ and $m^*_t(X)$ in place of $\tilde{\pi}(X)$ and $\hat{m}_t(X)$ throughout.

Second, we show the local nonparametric efficiency and double robustness of $\tilde{\nu}^t_{\mathrm{reg}}$. By the discussion in Section 3.3.1, it suffices to show that if the OR model (2.3) for $t=0$ or 1

is correctly specified, then asymptotic expansion (3.5) holds for the corresponding t. By construction, $\tilde{\pi}(X)\tilde{m}_0(X)$ is a linear combination of $\tilde{h}(X)/\tilde{\pi}(X)$, that is, $\tilde{\pi}(X)\tilde{m}_0(X) = c_0^{\mathrm{T}}\tilde{h}(X)/\tilde{\pi}(X)$ for some constant vector c_0. Then $\pi^{\dagger}(X)m_0^*(X) = c_0^{\mathrm{T}}h^*(X)/\pi^{\dagger}(X)$ also holds for the same vector c_0. If model (2.3) for $t = 0$ holds, then $m^*(x) = m_0(X)$ and hence $\pi^{\dagger}(X)m_0(X) = c_0^{\mathrm{T}}h^*(X)/\pi^{\dagger}(X)$. By direct calculation, we have

$$\beta_0^* = E^{-1}\left\{\xi_0^*\frac{1-T}{1-\pi^{\dagger}(X)}\frac{h^{*\mathrm{T}}(X)}{\pi^{\dagger}(X)}\right\}E\left\{\xi_0^*\frac{1-T}{1-\pi^{\dagger}(X)}\pi^{\dagger}(X)m_0(X)\right\} = c_0.$$

and hence asymptotic expansion (3.5) holds for $t = 0$. Similarly, because $\tilde{\pi}(X)\tilde{m}_1(X)$ is a linear combination of $\tilde{h}(X)/\{1-\tilde{\pi}(X)\}$, it can be shown that if the OR model (2.3) for $t = 1$ is correctly specified, then expansion (3.5) holds for $t = 1$.

Third, we show the intrinsic efficiency of $\tilde{\nu}^0_{\mathrm{reg}}$ among the class of estimators (3.4) for $t = 0$, denoted by $\tilde{\nu}^0(b_0)$. By direct calculation and Slutsky theorem, we have

$$\begin{aligned}\tilde{\nu}^0(b_0) - \nu^0 &= q^{-1}\tilde{E}(\tilde{\eta}_0 - b_0^{\mathrm{T}}\tilde{\xi}_0 - \nu^0 T) + o_p(n^{-1/2}) \\ &= q^{-1}\tilde{E}\left[\tilde{\eta}_0 - b_0^{\mathrm{T}}\left\{\frac{1-T}{1-\tilde{\pi}(X)} - 1\right\}\frac{h^*(X)}{\pi(X)} - \nu^0 T\right] + o_p(n^{-1/2}).\end{aligned}$$

If PS model (2.5) is correctly specified, then applying, to the above, Lemma 4.5(i) with $\hat{\pi}(X)$ replaced by $\tilde{\pi}(X)$ and $h(X) = b_0^{\mathrm{T}}h^*(X)/\pi(X)$ yields

$$\begin{aligned}\tilde{\nu}^0(b_0) - \nu^0 = q^{-1}\tilde{E}\Big(&\eta_0^* - b_0^{\mathrm{T}}\xi_0^* - \pi(X)\nu^0 - \Pi\{\eta_0^* - b_0^{\mathrm{T}}\xi_0^*|s^{\dagger}_{(\gamma^*,0)}(T,X)\} \\ &+ \Pi\left[\{T-\pi(X)\}\{m_0(X)-\nu^0\}|s^{\dagger}_{(\gamma^*,0)}(T,X)\right]\Big) + o_p(n^{-1/2}).\end{aligned}$$

where $\phi_h^0(Y,T,X) = \eta_0^* - b_0^{\mathrm{T}}\xi_0^*$ and $T\nu^0$ is decomposed as $\pi(X)\nu^0 + \{T-\pi(X)\}\nu^0 = \pi(X)\nu^0 + \Pi[\{T-\pi(X)\}\nu^0|s^{\dagger}_{(\gamma^*,0)}(T,X)]$ because $T-\pi(X)$ is contained in $s^{\dagger}_{(\gamma^*,0)}(T,X)$. The first term inside $\tilde{E}()$ above, $\eta_0^* - \pi(X)\nu^0 - b_0^{\mathrm{T}}\xi_0^* - \Pi\{\eta_0^* - b_0^{\mathrm{T}}\xi_0^*|s^{\dagger}_{(\gamma^*,0)}(T,X)\}$, is uncorrelated with the second term, $\Pi[\{T-\pi(X)\}\{m_0(X)-\nu^0\}|s^{\dagger}_{(\gamma^*,0)}(T,X)]$, which is independent of b_0. Moreover, the first term can be expressed as $\eta_0^* - \pi(X)\nu^0 - a_0^{\mathrm{T}}\xi_0^*$ for some constant vector a_0, because, by construction, each variable in $s^{\dagger}_{(\gamma^*,0)}(T,X)$ is a linear combination of varibles in ξ_0^*. By combining these two facts, we see that the

asymptotic variance of $\tilde{\nu}^0(b_0)$ is minimized when a_0 is equal to

$$\text{var}^{-1}(\xi_0^*)\text{cov}\left\{\xi_0^*,\, \eta_0^* - \pi(X)\nu^0\right\} = E^{-1}(\xi_0^*\zeta_0^{*\text{T}})E(\xi_0^*\eta_0^*) = \beta_0^*.$$

But to make a_0 equal to β_0^*, it suffices to set $b_0 = \beta_0^*$, because $\eta_0^* - \beta_0^{*\text{T}}\xi_0^*$ is uncorrelated with $s^{\dagger}_{(\gamma^*,0)}(T,X)$ and hence $\Pi\{\eta_0^* - \beta_0^{*\text{T}}\xi_0^*|s^{\dagger}_{(\gamma^*,0)}(T,X)\} = 0$. If PS model (2.5) is correctly specified, then $\tilde{\nu}^0_{\text{reg}} = \tilde{\nu}^0(\beta_0^*) + o_p(n^{-1/2})$. Therefore, $\tilde{\nu}^0_{\text{reg}}$ is intrinsically efficient among the class of estimators $\tilde{\nu}^0(b_0)$.

Finally, we show the intrinsic efficiency of $\tilde{\nu}^1_{\text{reg}}$ among the class of estimators (3.4) for $t = 1$, denoted by $\tilde{\nu}^1(b_1)$. By direct calculation and Slutsky theorem, we have

$$\begin{aligned}\tilde{\nu}^1(b_1) - \nu^1 &= q^{-1}\tilde{E}(\tilde{\eta}_1 - b_1^{\text{T}}\tilde{\xi}_1 - \nu^1 T) + o_p(n^{-1/2}) \\ &= q^{-1}\tilde{E}\left[\tilde{\eta}_1 - b_1^{\text{T}}\{T - \tilde{\pi}(X)\}\frac{h^*(X)}{\pi(X)\{1-\pi(X)\}} - \nu^1 T\right] + o_p(n^{-1/2}).\end{aligned}$$

If PS model (2.5) is correctly specified, then applying, to the above, Lemma 4.5(i) with $\hat{\pi}(X)$ replaced by $\tilde{\pi}(X)$ and $h(X) = b_1^{\text{T}}h^*(X)/[\pi(X)\{1-\pi(X)\}]$ yields

$$\begin{aligned}\tilde{\nu}^1(b_1) - \nu^1 = q^{-1}\tilde{E}\Big(&\eta_1^* - b_1^{\text{T}}\xi_1^* - \pi(X)\nu^1 - \Pi\{\eta_1^* - b_1^{\text{T}}\xi_1^*|s^{\dagger}_{(\gamma^*,0)}(T,X)\} \\ &+ \Pi\left[\{T-\pi(X)\}\{m_1(X) - \nu^1\}|s^{\dagger}_{(\gamma^*,0)}(T,X)\right]\Big) + o_p(n^{-1/2}),\end{aligned}$$

where $\phi_h^1(Y,T,X) = \eta_1^* - b_1^{\text{T}}\xi_1^*$. The intrinsic efficiency of $\tilde{\nu}^0_{\text{reg}}$ can be similarly obtained as above for the intrinsic efficiency of $\tilde{\nu}^1_{\text{reg}}$.

Derivation of empirical likelihood estimates

The empirical likelihood estimate of ν^t is $\hat{\nu}^t_{\text{lik}} = \sum_{i=1}^n \hat{p}_i\tilde{\eta}_{t,i}/\sum_{i=1}^n \hat{p}_i T_i$, where $(\hat{p}_1,\ldots,\hat{p}_n)$ are obtained from the constrained maximization problem:

$$\begin{aligned}\max_{p_1\geq 0,\ldots,p_n\geq 0} \quad & \sum_{i=1}^n \log p_i \\ \text{subject to} \quad & \sum_{i=1}^n p_i = 1 \text{ and } \sum_{i=1}^n p_i\tilde{\xi}_{1,i} = 0.\end{aligned}$$

By standard calculation (Qin & Lawless 1994), we have

$$\hat{p}_i = \frac{n^{-1}}{1 + \hat{\lambda}^{\mathrm{T}} \tilde{\xi}_{1,i}},$$

where $\hat{\lambda}$ is a maximizer of the function

$$\ell_{\mathrm{EL}}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \lambda^{\mathrm{T}} \tilde{\xi}_{1,i}\right).$$

Write $\tilde{\pi}_i = \tilde{\pi}(X_i)$, $\tilde{h}_i = \tilde{h}(X_i)$, and $\omega_i = \omega(X_i; \lambda)$ for $i = 1, \ldots, n$. By direct calculation, $\ell_{\mathrm{EL}}(\lambda)$ can be reexpressed as

$$\begin{aligned}
\ell_{\mathrm{EL}}(\lambda) &= \frac{1}{n} \sum_{i=1}^{n} \log\left\{1 + \lambda^{\mathrm{T}} \frac{T_i - \tilde{\pi}_i}{\tilde{\pi}_i(1 - \tilde{\pi}_i)} \tilde{h}_i\right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{T_i \log\left(1 + \lambda^{\mathrm{T}} \frac{\tilde{h}_i}{\tilde{\pi}_i}\right) + (1 - T_i) \log\left(1 - \lambda^{\mathrm{T}} \frac{\tilde{h}_i}{1 - \tilde{\pi}_i}\right)\right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \{T_i \log \omega_i + (1 - T_i) \log(1 - \omega_i)\} - \frac{1}{n} \sum_{i=1}^{n} \{T_i \log \tilde{\pi}_i + (1 - T_i) \log(1 - \tilde{\pi}_i)\},
\end{aligned}$$

which equals $\ell(\lambda)$ up to an additive constant. Therefore, $\hat{\lambda}$ is a maximizer of $\ell(\lambda)$. The desired expressions for $\hat{\nu}^1_{\mathrm{lik}}$ and $\hat{\nu}^0_{\mathrm{lik}}$ hold because, by direct calculation,

$$\begin{aligned}
\sum_{i=1}^{n} \hat{p}_i \tilde{\eta}_{1,i} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{\eta}_{1,i}}{1 + \hat{\lambda}^{\mathrm{T}} \tilde{\xi}_{1,i}} = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i}{1 + \hat{\lambda}^{\mathrm{T}} \frac{\tilde{h}_i}{\tilde{\pi}_i}} = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i \tilde{\pi}_i Y_i}{\hat{\omega}_i}, \\
\sum_{i=1}^{n} \hat{p}_i \tilde{\eta}_{0,i} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{\eta}_{0,i}}{1 + \hat{\lambda}^{\mathrm{T}} \tilde{\xi}_{1,i}} = \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - T_i) \frac{\tilde{\pi}_i}{1 - \tilde{\pi}_i} Y_i}{1 - \hat{\lambda}^{\mathrm{T}} \frac{\tilde{h}_i}{1 - \tilde{\pi}_i}} = \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - T_i) \tilde{\pi}_i Y_i}{1 - \hat{\omega}_i},
\end{aligned}$$

where $\hat{\omega}_i = \omega(X_i; \hat{\lambda})$ for $i = 1, \ldots, n$. □

Proof of Proposition 3.8

We need only to show that if model (2.5) is correctly specified, then $\tilde{\nu}^t_{\mathrm{lik}}$ is asymptotically equivalent, to the first order, to $\tilde{\nu}^t_{\mathrm{reg}}$ for $t = 0, 1$. By direct calculation and Slutsky

theorem, we have

$$\tilde{\nu}^0_{\text{lik}} - \nu^0 = q^{-1}\tilde{E}\left[\frac{(1-T)\tilde{\pi}(X)Y}{1-\omega(X;\tilde{\lambda}^0)} - T\nu^0\right] + o_p(n^{-1/2}).$$

If model (2.5) is correctly specified, then

$$\tilde{E}\left[\frac{(1-T)\tilde{\pi}(X)Y}{1-\omega(X;\tilde{\lambda}^0)}\right] = \tilde{E}\left[\frac{(1-T)\tilde{\pi}(X)Y}{1-\omega(X;\hat{\lambda})}\right] + o_p(n^{-1/2}),$$

by a Taylor expansion for $\tilde{\lambda}^0$ about $\hat{\lambda}$ and the fact that $\tilde{E}([(1-T)/\{1-\omega(X;\hat{\lambda})\} - 1]\tilde{\pi}(X)) = o_p(n^{-1/2})$, similarly as in the asymptotic expansion of the calibrated likelihood estimator in Tan (2010). Moreover, if model (2.5) is correctly specified, then $\hat{\lambda}$ converges to 0 in probability and

$$\tilde{E}\left[\frac{(1-T)\tilde{\pi}(X)Y}{1-\omega(X;\hat{\lambda})}\right] = \tilde{E}\left(\tilde{\eta}_0 - \beta_0^{*\mathrm{T}}\tilde{\xi}_0\right) + o_p(n^{-1/2}).$$

by a Taylor expansion for $\hat{\lambda}$ about 0, similarly as in the asymptotic expansion of the non-calibrated likelihood estimator in Tan (2010). The desired result for $\tilde{\nu}^0_{\text{lik}}$ then follows from the preceding expansions. Similarly, the result for $\tilde{\nu}^1_{\text{lik}}$ can be shown. □

Extension with non-logistic PS model

We discuss an extension of the regression and likelihood estimators $\tilde{\nu}^t_{\text{reg}}$ and $\tilde{\nu}^t_{\text{lik}}$ when the PS model (2.5) is non-logistic regression. Consider an augmented PS model

$$\begin{aligned} P(T=1|X) &= \pi_{\text{aug}}(X;\gamma,\gamma_0,\delta,\hat{\alpha}) \\ &= \Pi\left\{\gamma^{\mathrm{T}}f(X) + \gamma_0\hat{\rho}^{-1}(X) + \delta_0\,\hat{\rho}^{-1}(X)\hat{m}_0(X) + \delta_1\,\hat{\rho}^{-1}(X)\hat{m}_1(X)\right\}, \end{aligned}$$

where $\hat{\rho}(X) = \rho(X;\hat{\gamma})$ and $\rho(X;\gamma) = \Pi'\{\gamma^{\mathrm{T}}f(X)\}/[\pi(X;\gamma)\{1-\pi(X;\gamma)\}]$, which reduces to a constant 1 for logistic regression. Let $(\tilde{\gamma},\tilde{\gamma}_0,\tilde{\delta})$ be the estimates of (γ,γ_0,δ) solving the estimating equations

$$\tilde{E}\left[\{T - \pi_{\text{aug}}(X;\gamma,\gamma_0,\delta,\hat{\alpha})\}\{\hat{\rho}(X)f^{\mathrm{T}}(X), 1, \hat{m}_0(X), \hat{m}_1(X)\}^{\mathrm{T}}\right] = 0.$$

Let $\tilde{\pi}(X) = \pi_{\rm aug}(X; \tilde{\gamma}, \tilde{\gamma}_0, \tilde{\delta}, \hat{\alpha})$, and define the estimators $\tilde{\nu}^t_{\rm reg}$ and $\tilde{\nu}^t_{\rm lik}$ same as before, except that $\tilde{h}(X)$ is defined with

$$\tilde{h}_2(X) = \tilde{\pi}(X)\{1 - \tilde{\pi}(X)\}\{\hat{\rho}(X) f^{\rm T}(X), \hat{m}_0(X)\}^{\rm T}.$$

Then Propositions 4.3 and 4.4 can be shown to hold as before.

Particularly, to establish intrinsic efficiency, it can be shown that if PS model (2.5) is correctly specified, then the estimates $(\tilde{\gamma}, \tilde{\gamma}_0, \tilde{\delta})$ are asymptotically equivalent to the first order to the MLE of $(\gamma, \gamma_0, \delta)$ from the following "model,"

$$\begin{aligned} &P(T = 1|X) = \pi^*_{\rm aug}(X; \gamma, \gamma_0, \delta, \hat{\alpha}) \\ &= \Pi\left\{\gamma^{\rm T} f(X) + \gamma_0 \rho^{*-1}(X) + \delta_0\, \rho^{*-1}(X) m^*_0(X) + \delta_1\, \rho^{*-1}(X) m^*_1(X)\right\}, \end{aligned}$$

where $\rho^*(X) = \rho(X; \gamma^*)$. That is, the random variation in $\hat{\rho}(X)$, $\hat{m}_0(X)$, and $\hat{m}_1(X)$ does not affect the asymptotic behavior of $(\tilde{\gamma}, \tilde{\gamma}_0, \tilde{\delta})$ to the first order. The proofs of Propositions 4.3 and 4.4 can be completed similarly as before.

Violation of the exogeneity assumption

We present large-sample limits for estimators of ATT when the exogeneity assumption (A1) may be violated, i.e., T and Y^0 may not be conditionally independent given X. Similar results are known for estimators of ATE under possible violation of exogeneity assumptions (e.g., Robins 1999; Tan 2006). We mainly use these results to justify how various estimators are compared in our analysis of LaLonde data in Section 3.6, although the results can be broadly used.

Suppose that the exogeneity ssumption (A1) may be violated. The following results can be shown by similar calculations as under Assumption (A1).

(i) If the OR model (2.3) is correctly specified for $t = 0$, then $\hat{\nu}^0_{\rm OR}$, $\hat{\nu}^0_{\rm NP}(\hat{\pi}, \hat{m}_0)$, $\tilde{\nu}^0_{\rm reg}$, and $\tilde{\nu}^0_{\rm lik}$ converge in probability as $n \to \infty$ to $E\{Tm_0(X)\}/E(T)$, which reduces to $E(Y^0|T = 1)$ when Assumption (A1) holds but not generally so. Moreover, if the OR model (2.3) is correctly specified for $t = 1$, then $\tilde{\nu}^1_{\rm reg}$ and $\tilde{\nu}^1_{\rm lik}$ converge in

probability as $n \to \infty$ to $E(Y|T=1)$.

(ii) If the PS model (2.5) is correctly specified, then $\hat{\nu}^0_{\text{IPW}}(\hat{\pi})$, $\hat{\nu}^0_{\text{NP}}(\hat{\pi}, \hat{m}_0)$, $\tilde{\nu}^0_{\text{reg}}$, and $\tilde{\nu}^0_{\text{lik}}$ converge in probability as $n \to \infty$ to $E\{Tm_0(X)\}/E(T)$.

In the context of LaLonde analysis, let T be the indicator for the NSW cohort, i.e., $T = 1$ for the NSW treatment group in Analysis (i) or NSW control group in Analysis (ii) and $T = 0$ for the comparison group, and let D be the indicator for job training, i.e., $D = 1$ for the NSW treatment group and $D = 0$ for the NSW control group and the comparison group. Define Y^{11} as the potential outcome that would be observed if an individual was selected into NSW cohort and assigned to treatment, Y^{01} as the potential outcome that would be observed if an individual was selected into NSW cohort and assigned to control, and Y^{00} as the potential outcome that would be observed if an individual was selected into the comparison cohort and hence no job training. It is not necessary that $Y^{01} \equiv Y^{00}$, which would rule out any placebo effect such that earnings could be affected by merely participating in the NSW experiment. The exogeneity assumption (A1), $T \perp Y^{00}|X$, means that the NSW and comparison cohorts would have similar distributions of of earnings, at each covariate level x, if both placed in the comparison cohort and not assigned to job training. This assumption is implicitly made in all previous studies starting from LaLonde (1986), but can potentially be violated.

Because the NSW treatment and control groups are randomized, the difference

$$E(Y^{11}|T=1) - E(Y^{01}|T=1)$$

is the experimental benchmark. For Analysis (i) with NSW treatment group combined with a comparison group, a valid ATT estimator should be close to $E(Y^{11}|T=1) - E\{Tm_0(X)\}/E(T)$, and the corresponding bias be close to

$$\begin{aligned} &E(Y^{11}|T=1) - E\{Tm_0(X)\}/E(T) - \{E(Y^{11}|T=1) - E(Y^{01}|T=1)\} \\ &= E(Y^{01}|T=1) - E\{Tm_0(X)\}/E(T), \end{aligned}$$

where $m_0(X) = E(Y^{00}|T = 0, X)$. For Analysis (ii) with NSW control group combined with a comparison group, a valid ATT estimator should be close to

$$E(Y^{01}|T = 1) - E\{Tm_0(X)\}/E(T).$$

Therefore, the two bias estimates separately from Analyses (i) and (ii) should be close to each other for a good method, even when the exogeneity assumption (A1) is violated. This relationship forms the basis in our assessment of relative performances of various estimators of ATT in Section 3.6.

3.9 Additional Simulation Results

3.9.1 Qin–Zhang Simulation

Table 3.5 and Figures 3.4-3.5 present the results from 1000 Monte Carlo samples of size $n = 1000$, under the PS setting with small selection bias, $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.1, 0.1)$. Table 3.6 and Figures 3.6-3.7 present the results from 1000 Monte Carlo samples of size $n = 1000$, under the PS setting with large selection bias, $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.5, 0.5)$.

The relative performances of the estimators under study are similar to those under the PS setting with large selection bias, $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.2, 0.2)$. In particular, efficiency gains of the calibrated likelihood estimators over the doubly robust estimators, AIPW and AIPW.HIR, remain considerable across these settings, when the PS model is correctly specified but the OR model is misspecified.

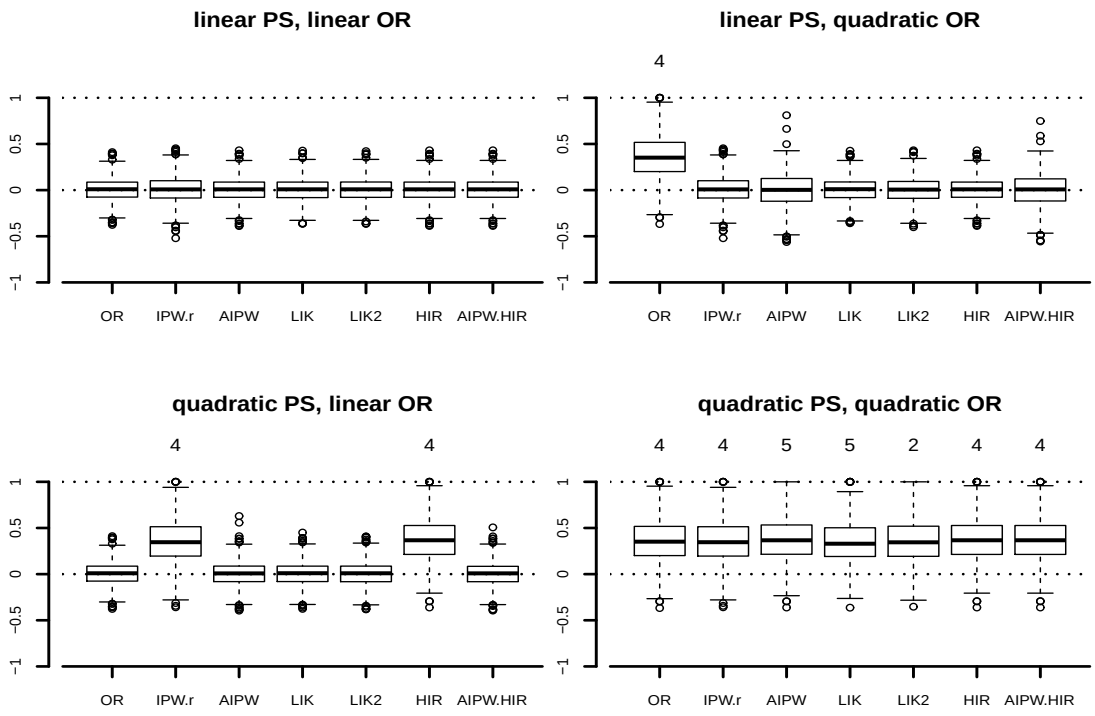

Figure 3.4: Boxplots of estimates minus the truth under LIN-OR setting with $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.1, 0.1)$.

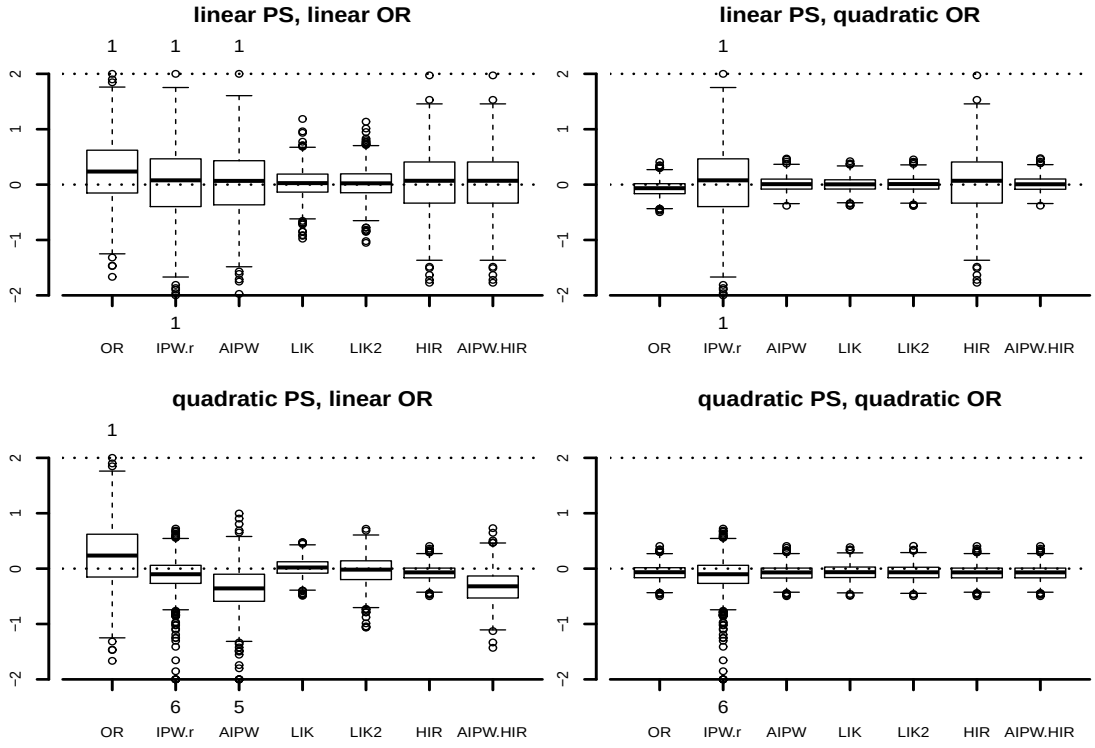

Figure 3.5: Boxplots of estimates minus the truth under QUA-OR setting with $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.1, 0.1)$.

Table 3.5: Qin–Zhang simulation results with $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.1, 0.1)$

Models	OR	IPW.r	AIPW	LIK	LIK2	HIR	AIPW.HIR	EL	AST
Data generated under LIN-OR setting									
linear PS,	0.0070	0.0076	0.0069	0.0062	0.0066	0.0069	0.0069	0.0038	-0.0004
linear OR	(0.0147)	(0.0200)	(0.0153)	(0.0157)	(0.0156)	(0.0153)	(0.0153)	(0.0204)	(0.0154)
linear PS,	0.3551	0.0076	0.0032	0.0072	0.0040	0.0069	0.0032	0.0040	-0.0083
quadratic OR	(0.0562)	(0.0200)	(0.0320)	(0.0163)	(0.0185)	(0.0153)	(0.0312)	(0.0241)	(0.0285)
quadratic PS,	0.0070	0.3488	0.0062	0.0072	0.0070	0.3687	0.0063	···	···
linear OR	(0.0147)	(0.0553)	(0.0176)	(0.0160)	(0.0167)	(0.0557)	(0.0171)	···	···
quadratic PS,	0.3551	0.3488	0.3721	0.3428	0.3516	0.3687	0.3687	···	···
quadratic OR	(0.0562)	(0.0553)	(0.0576)	(0.0544)	(0.0541)	(0.0557)	(0.0557)	···	···
Data generated under QUA-OR setting									
linear PS,	0.2235	0.0275	0.0291	0.0233	0.0249	0.0302	0.0302	0.0347	0.0009
linear OR	(0.3152)	(0.4034)	(0.3335)	(0.0647)	(0.0730)	(0.2999)	(0.2999)	(0.1561)	(0.3050)
linear PS,	-0.0690	0.0275	0.0094	0.0071	0.0084	0.0302	0.0094	0.0029	-0.0011
quadratic OR	(0.0190)	(0.4034)	(0.0173)	(0.0162)	(0.0170)	(0.2999)	(0.0178)	(0.0226)	(0.0168)
quadratic PS,	0.2235	-0.1398	-0.3619	0.0214	-0.0387	-0.0731	-0.3250	···	···
linear OR	(0.3152)	(0.1555)	(0.1949)	(0.0241)	(0.0635)	(0.0191)	(0.0906)	···	···
quadratic PS,	-0.0690	-0.1398	-0.0742	-0.0672	-0.0698	-0.0731	-0.0731	···	···
quadratic OR	(0.0190)	(0.1555)	(0.0193)	(0.0198)	(0.0195)	(0.0191)	(0.0191)	···	···

Table 3.6: Qin–Zhang simulation results with $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.5, 0.5)$

Models	OR	IPW.r	AIPW	LIK	LIK2	HIR	AIPW.HIR	EL	AST
Data generated under LIN-OR setting									
linear PS,	0.0089	0.0323	0.0107	0.0009	0.0030	0.0109	0.0109	0.0051	0.0024
linear OR	(0.0280)	(0.2078)	(0.0608)	(0.0733)	(0.0698)	(0.0547)	(0.0547)	(0.0900)	(0.0537)
linear PS,	1.8926	0.0323	0.0471	0.0527	0.0665	0.0109	0.0294	-0.0089	0.0244
quadratic OR	(0.1748)	(0.2078)	(0.2414)	(0.0642)	(0.0663)	(0.0547)	(0.0998)	(0.1103)	(0.1015)
quadratic PS,	0.0089	1.3964	0.0262	0.0026	0.0059	1.8722	0.0169	···	···
linear OR	(0.0280)	(0.9500)	(0.3731)	(0.0739)	(0.0770)	(0.1931)	(0.0731)	···	···
quadratic PS,	1.8926	1.3964	1.8918	1.8529	1.8459	1.8722	1.8722	···	···
quadratic OR	(0.1748)	(0.9500)	(0.4227)	(0.2195)	(0.2220)	(0.1931)	(0.1931)	···	···
Data generated under QUA-OR setting									
linear PS,	3.2822	0.1296	0.1560	0.3212	0.3819	0.2428	0.2428	0.1969	0.1943
linear OR	(0.9469)	(3.0404)	(3.7185)	(0.4148)	(0.5712)	(0.9017)	(0.9017)	(0.2647)	(0.7010)
linear PS,	-0.4663	0.1296	0.0077	0.0091	0.0061	0.2428	0.0156	0.0075	0.0095
quadratic OR	(0.0593)	(3.0404)	(0.0657)	(0.0796)	(0.0798)	(0.9017)	(0.0603)	(0.1026)	(0.0549)
quadratic PS,	3.2822	-1.9909	-1.9277	0.3801	0.9864	-0.4403	0.1483	···	···
linear OR	(0.9469)	(13.0100)	(34.4996)	(0.3366)	(0.3682)	(0.0742)	(0.3204)	···	···
quadratic PS,	-0.4663	-1.9909	-0.4319	-0.4754	-0.4449	-0.4403	-0.4403	···	···
quadratic OR	(0.0593)	(13.0100)	(0.1954)	(0.0918)	(0.0858)	(0.0742)	(0.0742)	···	···

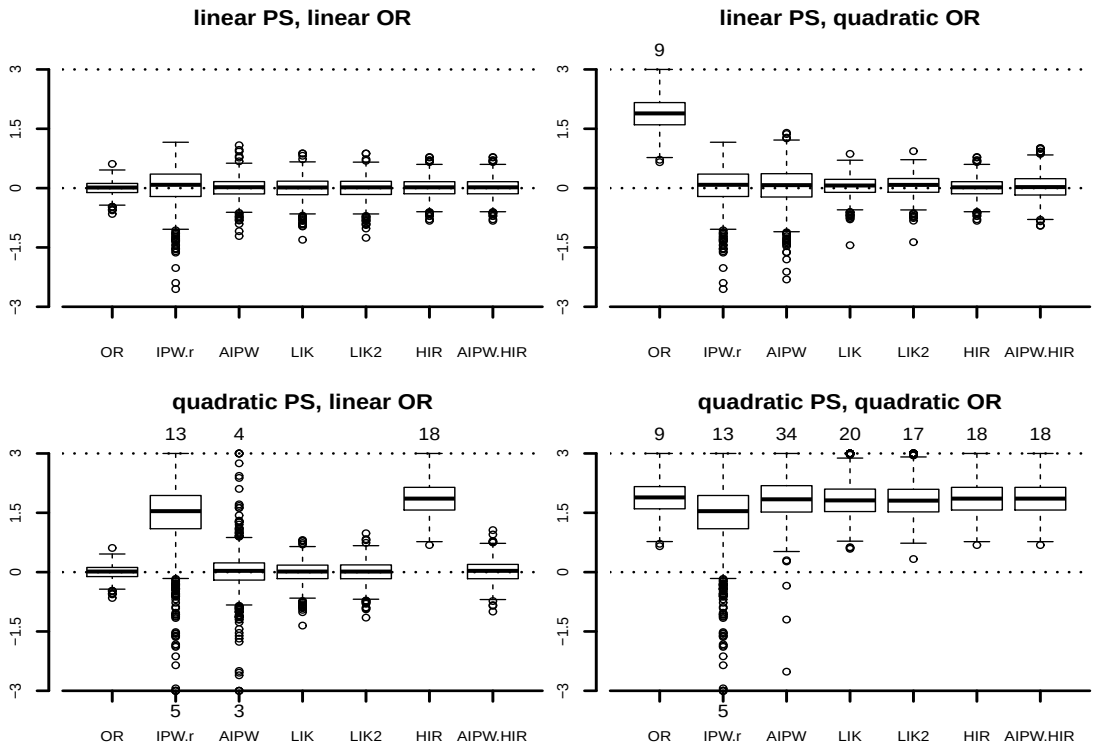

Figure 3.6: Boxplots of estimates minus the truth under LIN-OR setting with $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.5, 0.5)$.

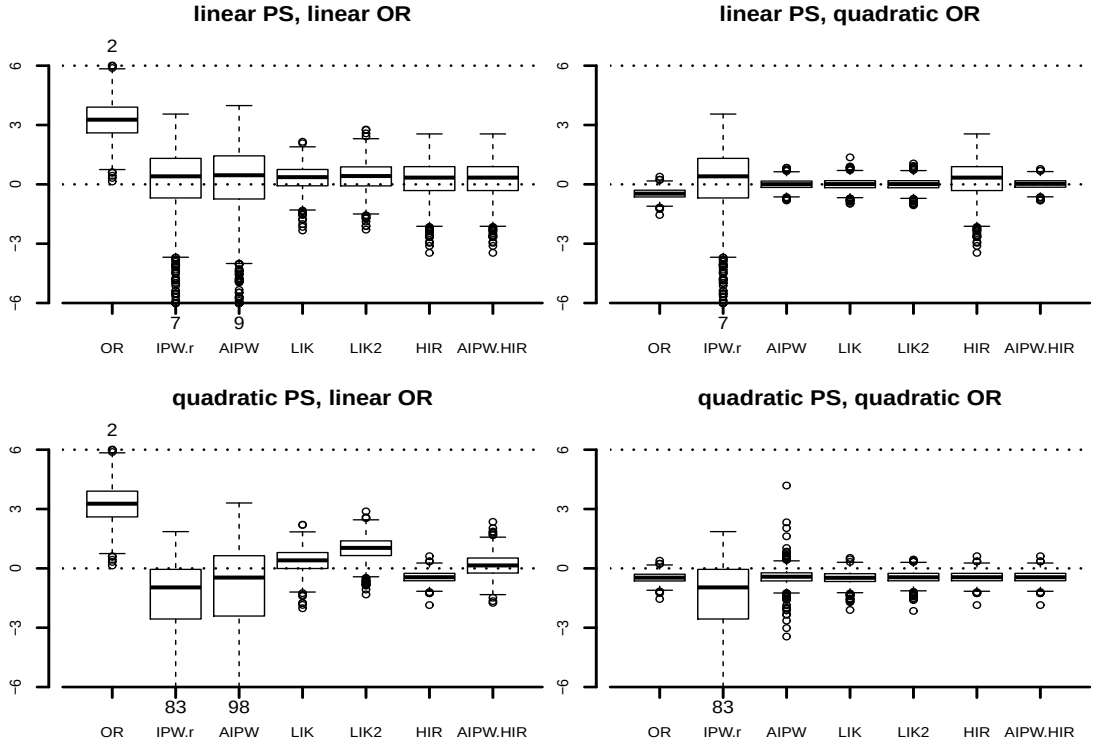

Figure 3.7: Boxplots of estimates minus the truth under QUA-OR setting with $(\gamma_1^*, \gamma_2^*, \gamma_3^*) = (1.0, 0.5, 0.5)$.

3.9.2 Kang–Schafer Simulation

In addition to the simulation study with the design of Qin & Zhang (2008), we also conducted a simulation study with the design of Kang & Schafer (2007) and a modified design defined in McCaffrey et al. (2007).

In Kang & Schafer (2007), the data are generated as $z = (z_1, z_2, z_3, z_4)^{\mathrm{T}}$, $y = 210 + 27.4z_1 + 13.7z_2 + 13.7z_3 + 13.7z_4 + \epsilon$, and $T = 1\{U \leqslant \mathrm{expit}(-z_1 + 0.5z_2 - 0.25z_3 - 0.1z_4)\}$, where $(z_1, z_2, z_3, z_4, \epsilon, U)$ are mutually independent, $(z_1, z_2, z_3, z_4, \epsilon)$ are marginally normally distributed with mean 0 and variance 1 and U is uniformly distributed on $(0, 1)$. Let $x = (x_1, x_2, x_3, x_4)^{\mathrm{T}}$, $x_1 = \exp(0.5z_1)$, $x_2 = z_2/\{1+\exp(z_1)\}+10$, $x_3 = (0.04z_1z_3 + 0.6)^3$, and $x_4 = (z_2 + z_4 + 20)^2$.

Two OR models (2.3) are specified with the identity link $\Psi(\cdot)$ and the regressor vector $g_0(z) = g_1(z) = (1, z_1, z_2, z_3, z_4)^{\mathrm{T}}$ or $(1, x_1, x_2, x_3, x_4)^{\mathrm{T}}$, corresponding to a correctly specified or misspecified OR model (denoted by OR z or OR x). Similarly, two PS models (2.5) are specified with the logistic link $\Pi(\cdot)$ and the regressor vector $f(z) = (1, z_1, z_2, z_3, z_4)^{\mathrm{T}}$ or $(1, x_1, x_2, x_3, x_4)^{\mathrm{T}}$, corresponding to a correctly specified or misspecified PS model (denoted by PS z or PS x).

The modified design in McCaffrey et al. (2007) is defined the same as above, except that an interaction term is added when generating the response, $y = 210 + 27.4z_1 + 13.7z_2 + 13.7z_3 + 13.7z_4 + 20z_1z_2 + \epsilon$. Three possible OR models (2.3) are specified with the identity link $\Psi(\cdot)$ and the regressor vector $g_0(z) = g_1(z) = (1, z_1, z_2, z_3, z_4, z_1z_2)^{\mathrm{T}}$, $(1, z_1, z_2, z_3, z_4)^{\mathrm{T}}$ or $(1, x_1, x_2, x_3, x_4)^{\mathrm{T}}$, corresponding to a correctly specified, slightly misspecified, or misspecified OR model (denoted by OR z2, OR z, or OR x). Two possible PS models (2.5) are specified the same as above.

For these two designs, Table 3.7 and Figure 3.8-3.9 present the results for various estimators from 5000 Monte Carlo sample with size $n = 1000$. The true value of ATT is easily shown to be always 0.

The relative performances of the estimators under study are overall similar to those found in the Qin–Zhang simulation study. A seemingly unexpected phenomenon, in view of intrinsic efficiency of LIK, is that the HIR and AIPW.HIR estimators have

smaller variances than LIK and LIK2 estimators in the Kang–Schafer design when PS z and OR x models (which are correctly specified and misspecified respectively) are used. But this difference can be explained as follows. In this case, because the true $m_0(X)$ is a linear combination of $f(X)$ used, the HIR estimator can be shown to achieve the nonparametric efficiency bound by similar arguments as in the proof of local nonparametric efficiency of $\hat{\nu}^0_{\mathrm{NP}}(\hat{\pi}, \hat{m}_0)$. This can also be seen numerically from Monte Carlo standard errors. The estimator AIPW.HIR (which is doubly robust) has a moderately inflated from that of HIR (which is non-doubly robust) and hence smaller than those of LIK and LIK2. This phenomenon depends on the particular way in which the Kang–Schafer design is defined; it does not occur in the McCaffrey-et-al design when PS z and OR z or OR x models are used.

Table 3.7: Kang–Schafer and McCaffrey-et-al simulation results

Models	OR	IPW.ratio	AIPW	LIK	LIK2	HIR	AIPW.HIR
			Kang–Schafer design				
PS z, OR z	-0.00021 (0.07881)	-0.15529 (2.29565)	0.00009 (0.08899)	0.00038 (0.09119)	0.00038 (0.09014)	-0.00001 (0.08815)	-0.00001 (0.08815)
PS x, OR z	-0.00021 (0.07881)	-7.19115 (1.76502)	-0.00027 (0.08423)	0.00019 (0.09345)	-0.00015 (0.09431)	-4.43418 (1.03883)	-0.00035 (0.08538)
PS z OR x	-9.94070 (1.53143)	-0.15529 (2.29565)	-0.28659 (2.51075)	-0.34182 (0.98906)	-0.41053 (1.19155)	-0.00001 (0.08815)	-0.25263 (0.78813)
PS x, OR x	-9.94070 (1.53143)	-7.19115 (1.76502)	-6.15538 (1.70318)	-4.80166 (1.54315)	-5.60558 (1.56532)	-4.43418 (1.03883)	-4.43418 (1.03883)
			McCaffrey-et-al design (with interaction)				
PS z, OR z2	-0.00001 (0.08057)	-0.25727 (3.74352)	0.00018 (0.08915)	0.00031 (0.09289)	0.00030 (0.09261)	-0.26256 (1.90160)	1e-6 (0.08844)
PS x, OR z2	-0.00001 (0.08057)	-5.36684 (2.74036)	-0.00024 (0.08382)	-0.00048 (0.09746)	-0.00090 (0.09833)	-2.68423 (1.69244)	-0.00039 (0.08490)
PS z, OR z	-6.45619 (1.92221)	-0.25727 (3.74352)	-0.16704 (3.44095)	-0.29360 (1.43785)	-0.41220 (1.94467)	-0.26256 (1.90160)	-0.26256 (1.90160)
PS x, OR z	-6.45619 (1.92221)	-5.36684 (2.74036)	0.79425 (2.42957)	-0.25981 (1.31109)	0.81471 (1.48044)	-2.68423 (1.69244)	1.16681 (1.06564)
PS z, OR x	-10.47822 (2.17768)	-0.25727 (3.74352)	-0.38755 (4.11415)	-0.35741 (1.22501)	-0.59751 (1.92895)	-0.26256 (1.90160)	-0.47727 (1.89506)
PS x, OR x	-10.47822 (2.17768)	-5.36684 (2.74036)	-4.36236 (2.86418)	-2.06547 (1.84705)	-3.12139 (2.21782)	-2.68423 (1.69244)	-2.68423 (1.69244)

Note: In the upper rows are the Monte Carlo means, and in the brackets are the corresponding Monte Carlo variances.

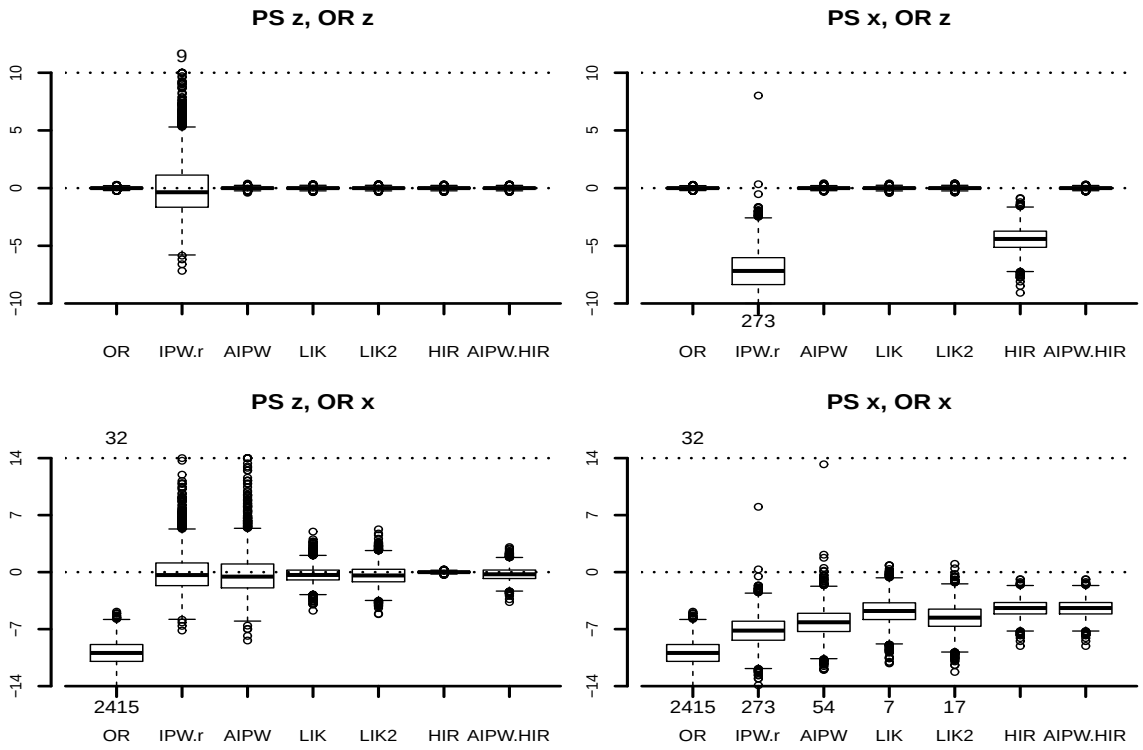

Figure 3.8: Boxplots of estimates under the Kang–Schafer design. The values are censored within the range of the y-axis, and the number of values that lie outside the range are indicated next to the lower and upper limits of the y-axis.

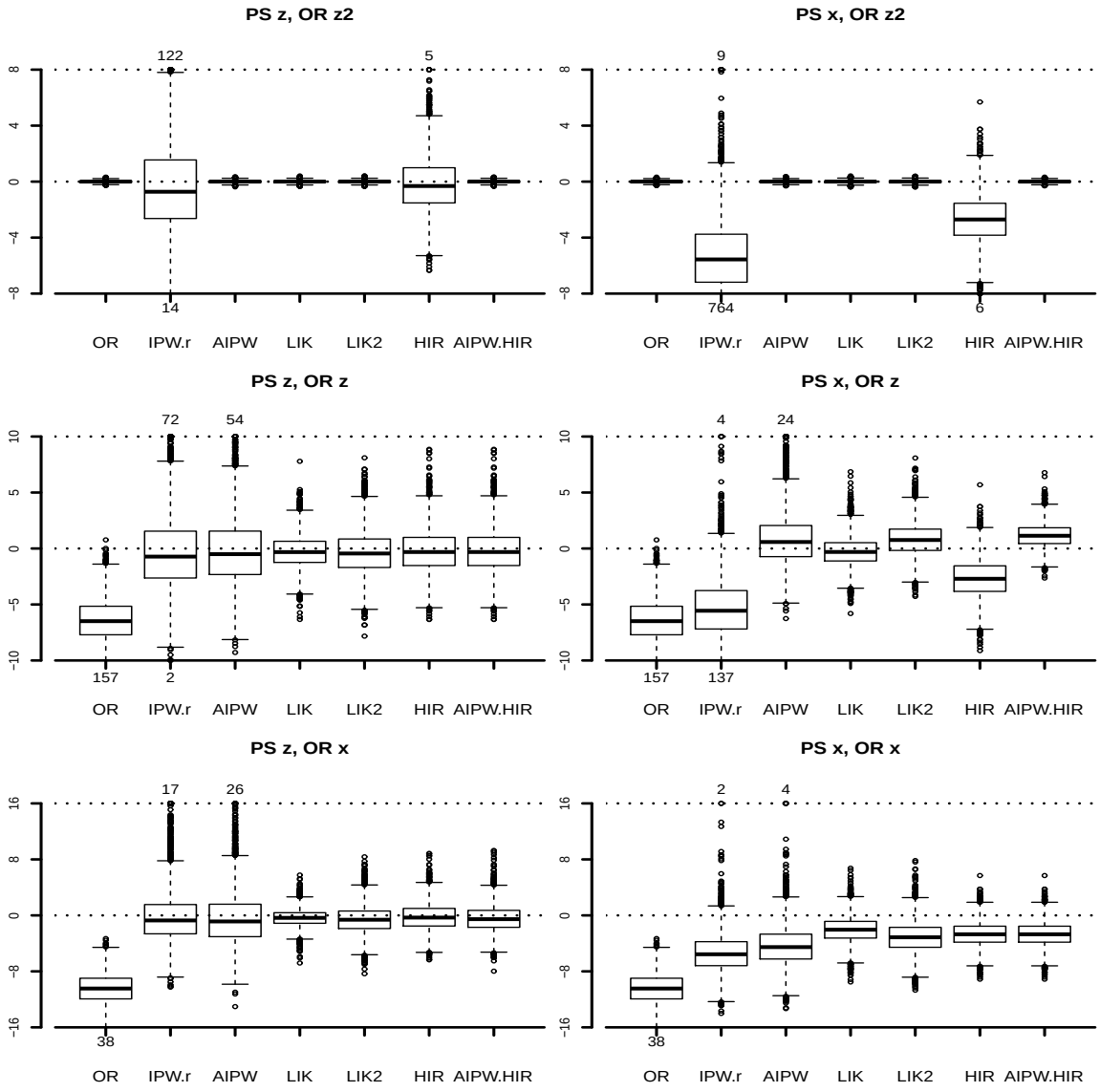

Figure 3.9: Boxplots of estimates under the McCaffrey-et-al design (with interaction).

3.9.3 Lalonde Analysis

Table 3.8 and Figure 3.10 present the results from Analyses (i) and (ii) for various estimators as listed in Section 3.5, based on 500 bootstrap samples of the NSW+CPS composite data. There are much smaller differences between the performances of the estimators than when the NSW+PSID composite data are analyzed. Another feature worthy of note is that none of the estimators lead to effect estimates close to the experimental benchmark $886 or bias estimates close to 0, even though the differences between effect and bias estimates are all roughly close to $886.

Table 3.8: Bootstrap results from Analyses (i) and (ii) on NSW+CPS composite data

		OR	IPW.ratio	AIPW	LIK2	HIR	AIPW.HIR
Linear PS, Linear OR	Treatment Effect	-800	-451	-308	-380	-503	-388
		(475)	(518)	(526)	(518)	(520)	(520)
	Evaluation Bias	-1709	-1336	-1333	-1364	-1414	-1413
		(374)	(414)	(428)	(420)	(422)	(422)
	Difference	803	885	903	880	910	910
		(527)	(518)	(532)	(526)	(531)	(531)
Linear PS, Quadratic OR	Treatment Effect	-800	-451	-308	-380	-503	-388
		(475)	(518)	(522)	(516)	(520)	(517)
	Evaluation Bias	-1611	-1336	-1196	-1254	-1414	-1295
		(379)	(414)	(436)	(430)	(422)	(429)
	Difference	811	885	888	874	910	907
		(527)	(518)	(529)	(523)	(531)	(529)
Quadratic PS, Linear OR	Treatment Effect	-906	-427	-421	-424	-465	-465
		(475)	(561)	(561)	(561)	(557)	(557)
	Evaluation Bias	-1709	-1207	-1335	-1297	-1383	-1383
		(374)	(529)	(533)	(514)	(507)	(507)
	Difference	803	780	914	873	919	919
		(527)	(547)	(544)	(538)	(532)	(532)
Quadratic PS, Quadratic OR	Treatment Effect	-800	-427	-465	-438	-432	-465
		(475)	(561)	(557)	(563)	(557)	(557)
	Evaluation Bias	-1611	-1207	-1383	-1364	-1313	-1383
		(379)	(529)	(507)	(535)	(514)	(507)
	Difference	811	780	919	926	881	919
		(527)	(547)	(532)	(543)	(535)	(532)

Note: In the upper rows are the bootstrap means, and in the brackets are the corresponding bootstrap standard errors. Treatment Effect is obtained from Analysis (i), and Evaluation Bias from Analysis (ii). The difference is to be compared with the experimental benchmark $886 with standard error $488. There was no issue of non-convergence when computing estimates during bootstrapping, and hence Principle Component Analysis is not needed.

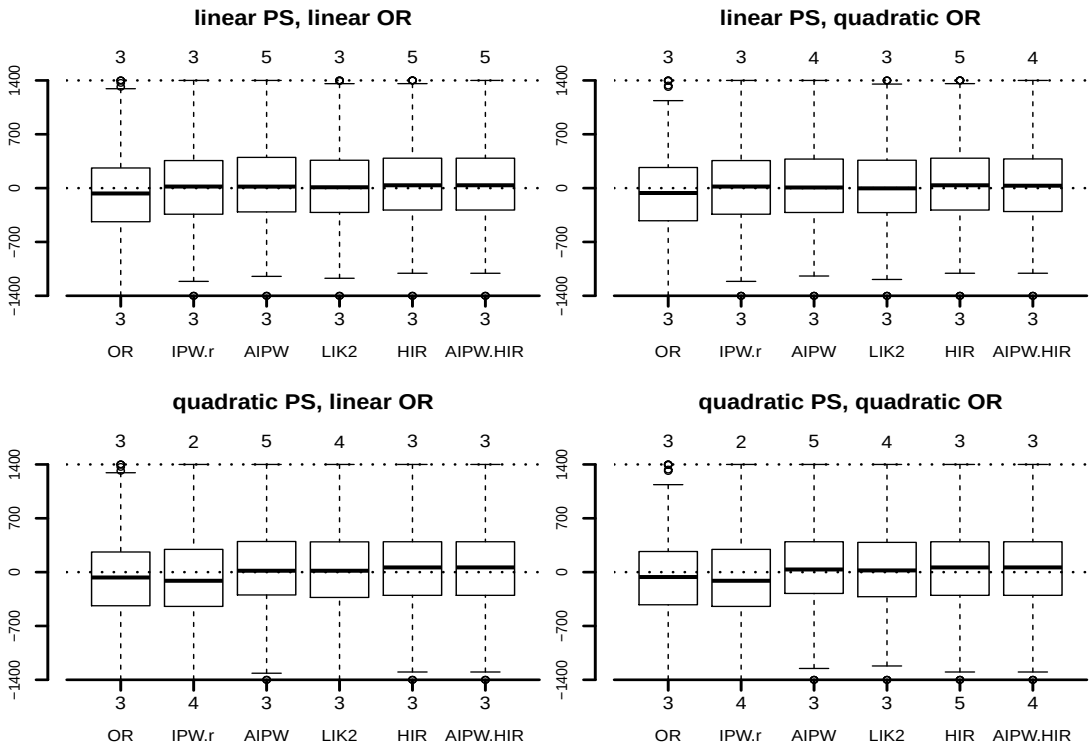

Figure 3.10: Bootstrap boxplots of differences of bias estimates from Analyses (i) and (ii) on NSW+CPS composite data.

Chapter 4

Improved Methods using Data Combination for Moment Restriction Models

In this chapter, we develop improved methods using data combination for moment restriction models. First, we examine semiparametric theory following Chen et al. (2008), and then derive augmented inverse probability weighted (AIPW) estimators that are locally efficient and doubly robust. Furthermore, we develop calibrated regression and likelihood estimators which achieve double robustness, local efficiency and desirable properties beyond.

Specifically, the proposed estimators are *doubly robust*, i.e., remain consistent as long as either a propensity score (PS) model or an outcome regression (OR) model is correct. The estimator proposed by Graham et al. (2015) is doubly robust only under the assumption that PS model and OR model share the same vector of regressors. We exploited the idea of augmenting the propensity score model with additional regressors related to fitted values from the outcome regression model. Then double robustness is obtained without the restrictive assumption in Graham et al. (2015). Second, the proposed estimators are *locally efficient*, i.e., attain the nonparametric variance bound when propensity score and outcome regression model are both correctly specified. And they are *intrinsically efficient* in achieving greater efficiency than AIPW estimators when the propensity score model is correctly specified but the outcome regression model may be misspecified.

For illustration, we take the linear two-sample instrumental variable problem as an example, and derive all the relevant estimators in details. Compared with the classical Two-Sample Instrumental Variable (TSIV) estimator (Angrist & Krueger, 1992), our

estimators could still generate consistent estimators when the two samples differ in distribution. The conventional Two-Sample Two-Stage Least Squares (TS2SLS) estimator (Bjorklund & Jantti 1997) is only consistent when OR model is correctly specified, while our estimator is doubly robust, i.e., also consistent as long as PS model is correctly specified even if OR model is misspecified. Moreover, our estimator is designed to achieve greater efficiency than AIPW estimator which is doubly robust and locally efficient, when PS model is correctly specified but the OR model may be misspecified. We present a simulation study and an Economic application on a public housing project, and provide numerical comparisons with existing methods.

4.1 Moment Restriction Models with Auxiliary Data

Let (x, z) be a random vector drawn from a population we want to investigate. But x is missing from the *primary data*, the population we are interested. So we collect *auxiliary data* which contains the measurements of (x, z), but this data may be drawn from a different population. z is a common variable across the two samples, but it may have different distributions in the two samples. In this chapter, let's use the superscript (1) to denote the *primary data* and superscript (0) to denote the *auxiliary data*.

We are interested in the estimation of parameters θ defined in terms of nonlinear moment conditions

$$E^{(1)}\Phi(x, z; \theta) = 0 \tag{4.1}$$

where $\Phi(x, z; \theta)$ are $k \times 1$ vectors, and the unknown parameter θ of interest is also a $k \times 1$ vector. And we refer to this case "just-determined". $E^{(1)}$ refers to the expectation taken with respect to the population of the primary data. So the underlying difficulty is x is missing in the primary data we are interested in.

Hahn (1998) and Chen et al. (2008) show that θ could be identified when (i) the conditional distribution of x given z is the same across two samples, ie: $F^{(1)}(x|z) = F^{(0)}(x|z)$, (ii) for the common variable z, the support in primary data are contained within the support in auxiliary data, which is an important assumption ensuring we

could apply the relationship between x and z determined from auxiliary data on the primary data without extrapolation.

If we merge the two samples, and add an indicator variable t. For the observations in Data (1), the primary data, we have $t = 1$, otherwise $t = 0$. Then we construct a merged sample with sample size $n = n_0 + n_1$.

$$\Big\{(t_i, z_i, (1 - t_i)x_i)^{\mathrm{T}}\Big\}_{i=1}^{n}$$

The data $(z_i, x_i, t_i)_{i=1}^{n}$ are now conceptualized as an i.i.d. sample from (z, x, t). We have transformed the two-sample problem to one merged-sample framework. Eq. (4.1) could be represented by

$$E\Big[\Phi(x, z; \theta)\Big|t = 1\Big] = 0 \tag{4.2}$$

Before we discuss the modeling approaches of this merged-sample problem, it is helpful to emphasize the underlying assumptions.

The first assumption is what we mentioned in previous sections: $F^{(1)}(x|z) = F^{(0)}(x|z)$. And it is interesting to find out that this is actually equivalent to the well-known assumption in causal model: "Unconfoundedness".

$$x \perp t \,|\, z \tag{4.3}$$

This conditional independence assumption has been widely used in Econometrics and Statistics to achieve identification with missing data, for example Little & Rubin (2002), Robins & Rotnitzky (1995), and Heckman et al. (1999).

The other assumption is called "overlap":

$$0 \leq P(t = 1|z) < 1 \tag{4.4}$$

This assumption can also be stated in another way: the support of z in the auxiliary data is at least as large as that in the primary data.

4.1.1 Basic Approaches of Estimation

In addition to these two assumptions, we still need modeling assumptions to demonstrate the relationship between t or x with the common variable z. Generally, there are two modeling approaches.

The first one is to construct a regression model to capture the relationship between the functions on the left-hand side of moment equation (4.2) and the common variables z. Let's call them outcome regression (OR) function, $q_0(z;\alpha_0)$, where α_0 depends on the value of θ.

$$q_0(z;\alpha_0) = E\big[\Phi(x,z;\theta)\big|z\big] = \Psi\{g_0(z)\alpha_0\} \tag{4.5}$$

where $\Psi(\cdot)$ is an inverse link function, $g_0(z)$ is a $k \times p$ matrix of known function of z, and α_0 is a $p \times 1$ vector of unknown parameters. Generally, maximum quasi-likelihood estimate of α_0 is solved through

$$\sum_{t_i=0} \Big\{ g_0(z_i)^{\mathrm{T}} \big[\Phi(x_i,z_i;\theta) - q_0(z_i;\alpha_0)\big] \Big\} = 0 \tag{4.6}$$

On the other hand, we have another equation based on (4.2),

$$\frac{1}{n_1}\sum_{i=1}^{n} t_i q_0(z_i;\alpha_0) = 0 \tag{4.7}$$

The estimators of θ and α_0 could be solved based on (4.6) together with (4.7). Let's denote them as $\hat{\theta}_{\mathrm{OR}}$ and $\hat{\alpha}_0$, and we use $\hat{q}_0(z)$ to represent $q_0(z;\hat{\alpha}_0)$ based on $(\hat{\theta}_{\mathrm{OR}},\hat{\alpha}_0)$ for convenience. If model (4.5) is correctly specified, then $\hat{\theta}_{\mathrm{OR}}$ is a consistent estimator.

The other approach is to build a regression model to predict the conditional probability that certain observation belongs to the primary sample given the common variable z. Usually we call this probability propensity score (PS), whose essential role is emphasized by Rosenbaum & Rubin (1983), and it is denoted as $\pi(z)$ in this chapter.

$$P(t=1|z) = E(t|z) = \pi(z;\gamma) = \Pi\{\gamma^{\mathrm{T}} f(z)\} \tag{4.8}$$

where $\Pi(\cdot)$ is an inverse link function. $f(z)$ is a vector of known function *including 1*, and γ is a vector of unknown parameters. The score function of γ is:

$$S_\gamma(t,z) = \left[\frac{t}{\pi(z;\gamma)} - \frac{1-t}{1-\pi(z;\gamma)}\right]\frac{\partial \pi(z;\gamma)}{\partial \gamma}$$

Usually, logistic regression is used to fit propensity score:

$$\pi(z;\gamma) = \frac{\exp[\gamma^{\mathrm{T}} f(z)]}{1+\exp[\gamma^{\mathrm{T}} f(z)]} \tag{4.9}$$

Let $\hat{\gamma}$ be the maximum likelihood estimator (MLE) of γ that solves $\tilde{E}\left[S_\gamma(t,z)\right] = 0$, which in logistic regression case reduces to

$$\tilde{E}\big[t - \pi(z;\gamma)\big] f(z) = 0 \tag{4.10}$$

where $\tilde{E}(\cdot)$ denotes simple sample average of the whole merged sample, and we will use this symbol representing the same idea in the rest of this chapter. And we use $\hat{\pi}(z)$ instead of $\pi(z;\hat{\gamma})$ for convenience.

We could simply obtain an estimation $\hat{\theta}_{\mathrm{IPW}}$ using the estimated $\hat{\pi}(z)$ above through

$$\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\hat{\pi}(z)\Phi(x,z;\theta)\right\} = 0 \tag{4.11}$$

Generally, $\hat{\theta}_{\mathrm{IPW}}$ is called Inverse Probability Weighted (IPW) estimator. Since the fitted propensity score appears in the denominator, $\hat{\theta}_{\mathrm{IPW}}$ can be very sensitive to the specification of PS model (4.8). The estimator $\hat{\theta}_{\mathrm{IPW}}$ is consistent only when PS model (4.8) is correctly specified.

4.1.2 Semiparametric Efficiency Theory and AIPW Estimator

As discussed in Section 4.1.1, $\hat{\theta}_{\mathrm{OR}}$ is consistent only when the OR model (4.5) is correctly specified, and the consistency of $\hat{\theta}_{\mathrm{IPW}}$ also depends on the correct specification of PS model (4.8). Alternatively, we could try to use both PS (4.8) and OR (4.5) model, in order to gain efficiency and robustness, similarly as in estimation of ATE and ATT.

Chen et al. (2008) studied the semiparametric efficiency theory of data combination problem defined through general moment conditions. Based on their findings, we present the semiparametric efficient influence function and the variance bound for our problem defined as (4.1).

Proposition 4.1 gives the semiparametric influence functions and we also list the semiparametric efficiency bounds for estimation of θ under three different settings in Table 4.1.

Proposition 4.1 *Let $p = E(T)$ and define*

$$\Gamma_\theta = \frac{\partial}{\partial \theta^{\mathrm{T}}} E\left[\Phi(x, z; \theta) | t = 1\right] \tag{4.12}$$

$$q_0(z) = E[\Phi(x, z; \theta) | z] \tag{4.13}$$

$$F(t, x, z) = \frac{1 - t}{p} \frac{\pi(z)}{1 - \pi(z)} [\Phi(x, z; \theta) - q_0(z)] \tag{4.14}$$

The efficient influence function for estimation of θ is as follows, depending on assumptions on the propensity score.

(i) The efficient influence function is

$$\varphi_{NP}(t, x, z) = -\Gamma_\theta^{-1} \times F_{NP}(t, x, z)$$

where

$$F_{NP}(t, x, z) = F(t, x, z) + \frac{t q_0(z)}{p}$$

(ii) If the propensity score $\pi(z)$ is known, then the efficient influence function is

$$\varphi_{SP*}(t, x, z) = -\Gamma_\theta^{-1} \times F_{SP*}(t, x, z)$$

where

$$F_{SP*}(t, x, z) = F(t, x, z) + \frac{\pi(z) q_0(z)}{p}$$

Table 4.1: Efficiency bounds for estimation of θ

Assumption	Efficiency bound
Nonparametric model	$V_{\mathrm{NP}} = \mathrm{var}\{\varphi_{\mathrm{NP}}(t,x,z)\}$
Parametric PS model	$V_{\mathrm{SP}} = \mathrm{var}\{\varphi_{\mathrm{SP}}(t,x,z)\}$
Known $\pi(z)$	$V_{\mathrm{SP}*} = \mathrm{var}\{\varphi_{\mathrm{SP}*}(t,x,z)\}$

(iii) If the propensity score $\pi(z)$ is unknown but assumed to belong to a correctly specified parametric family $\pi(z;\gamma)$, then the efficient influence function is

$$\varphi_{SP}(t,x,z) = -\Gamma_\theta^{-1} \times F_{SP}(t,x,z)$$

where

$$F_{SP}(t,x,z) = F_{SP*}(t,x,z) + \Pi\left[\{t-\pi(z)\}\frac{q_0(z)}{p}\Big| S_\gamma(t,z)\right]$$

where for two random vectors Z_1 and Z_2, $\Pi(Z_2|Z_1) = \mathrm{cov}(Z_2,Z_1)\mathrm{var}^{-1}(Z_1)Z_1$, i.e., the projection of Z_2 onto Z_1.

From the discussion in Chen et al. (2008) and Hahn (1998), the variance bounds in Table 4.1 satisfy the order that $V_{\mathrm{NP}} \geq V_{\mathrm{SP}} \geq V_{\mathrm{SP}*}$, with strict inequalities in general. This ordering of efficiency bounds agrees with the usual comparison that the efficiency bound under a more restrictive model is no greater than under a less restrictive model.

We now derive two estimators of θ based on both outcome regression function $q_0(z)$ and propensity score $\pi(z)$, by directly taking the efficient influence functions in Proposition 4.1 as estimating functions, with the truth $q_0(z)$ and $\pi(z)$ replaced by the estimated function $\hat{q}_0(z)$ and $\hat{\pi}(z)$. Proposition 4.2 shows that both estimators possess local efficiency but of different types, and only one estimator is doubly robust. For clarity, the semiparametric efficiency bound V_{NP} under the nonparametric model is also called the *nonparametric efficiency bound*, and the semiparametric efficiency bound V_{SP} under the nonparametric model is also called the *semiparametric efficiency bound*.

Proposition 4.2 *Under suitable regularity conditions (see Appendix), the following results hold.*

(i) Define an estimator of θ as the solution of the following equation

$$\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\hat{\pi}(z)\big[\Phi(x,z;\theta)-\hat{q}_0(z)\big]+t\hat{q}_0(z)\right\}\Big/\tilde{E}(t)=0 \tag{4.15}$$

Let's call the estimation $\hat{\theta}_{NP}$.

Then $\hat{\theta}_{NP}$ is locally nonparametric efficient: it achieves the nonparametric efficiency bound V_{NP} when both model (4.5) and model (4.8) are correctly specified. Moreover, $\hat{\theta}_{NP}$ is doubly robust: it remains consistent when either model (4.5) or model (4.8) is correctly specified.

(ii) Define an estimator of θ as the solution of the following equation

$$\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\hat{\pi}(z)\big[\Phi(x,z;\theta)-\hat{q}_0(z)\big]+\hat{\pi}(z)\hat{q}_0(z)\right\}\Big/\tilde{E}(t)=0 \tag{4.16}$$

Let's call the estimation $\hat{\theta}_{SP}$.

Then $\hat{\theta}_{SP}$ is locally semiparametric efficient: it achieves the semiparametric efficiency bound V_{SP} when both model (4.5) and model (4.8) are correctly specified. But $\hat{\theta}_{SP}$ is, generally, not doubly robust.

Actually, both $\hat{\theta}_{\mathrm{NP}}$ and $\hat{\theta}_{\mathrm{SP}}$ are in the form of AIPW estimators, since they could be represented by the solution of the following equation with $h(z)=\hat{q}_0(z)$ or $h(z)=\hat{\pi}(z)\hat{q}_0(z)$ respectively.

$$\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\hat{\pi}(z)\Phi(x,z;\theta)-\Big[\frac{1-t}{1-\hat{\pi}(z)}-1\Big]h(z)\right\}\Big/\tilde{E}(t)=0 \tag{4.17}$$

Setting $h(z)\equiv 0$ leads to the IPW estimator $\hat{\theta}_{\mathrm{IPW}}$. According to local semiparametric efficiency, $\hat{\theta}_{\mathrm{SP}}$ achieves the minimum variance among all the regular estimators under correctly specified parametric PS model, when both PS model and OR model are correctly specified. However, $\hat{\theta}_{\mathrm{SP}}$ is not doubly robust, but $\hat{\theta}_{\mathrm{NP}}$ is, and we will refer to $\hat{\theta}_{\mathrm{NP}}$ as the AIPW estimator in the later sections of this chapter.

4.1.3 Improved Estimation

As discussed above, the two AIPW estimators $\hat{\theta}_{\mathrm{NP}}$ and $\hat{\theta}_{\mathrm{SP}}$ locally achieve different efficiency bounds, and among them $\hat{\theta}_{\mathrm{NP}}$ is doubly robust. In Section 4.1.3, we will introduce certain estimators which are not only doubly robust and locally nonparametric efficient, but also intrinsically efficient: as long as PS model (4.8) is correctly specified, the estimator will attain the smallest asymptotic variance among a class of AIPW estimators but with $\hat{\pi}(z)$ replaced by another fitted propensity score called augmented propensity score defined later in (4.18).

Regression Estimators

We derive regression estimators, similar to the regression estimator of ATE in Tan (2006), but with an important new idea. Suppose we use logistic regression to fit the propensity score model (4.8), now let's consider an augmented logistic propensity score model

$$\begin{aligned} P(t=1|z) &= \pi_{\mathrm{aug}}(z;\gamma,\delta,\hat{\alpha}_0) \\ &= \mathrm{expit}\left\{\gamma^{\mathrm{T}} f(z) + \delta^{\mathrm{T}} \hat{q}_0(z)\right\}, \end{aligned} \tag{4.18}$$

where $\mathrm{expit}(c) = \{1+\exp(-c)\}^{-1}$, $\hat{\alpha}_0$ are estimates of α_0 based on Eqs. (4.6) and (4.7), and δ are unknown coefficients for additional regressors $\hat{q}_0(z)$ based on $\hat{\alpha}_0$. Let $(\tilde{\gamma},\tilde{\delta})$ be the MLE of (γ,δ) and $\tilde{\pi}(z) = \pi_{\mathrm{aug}}(z;\tilde{\gamma},\tilde{\delta},\hat{\alpha}_0)$. An important consequence of including the additional regressors is that, by Eq. (4.10), we have, in addition to $\tilde{E}[\{t-\tilde{\pi}(z)\}f(z)] = 0$,

$$\tilde{E}\Big[\{t-\tilde{\pi}(z)\}\hat{q}_0(z)\Big] = 0 \tag{4.19}$$

For the augmented PS model, there may be linear redundancy in the variables, $\{f(z), \hat{q}_0(z)\}$, in which case the regressors need to be redefined accordingly.

We define the regression estimator $\tilde{\theta}_{\text{reg}}$ is the solution to

$$\tilde{E}[\tilde{\tau}_{\text{reg}}(\theta)] = 0 \tag{4.20}$$

with

$$\tilde{\tau}_{\text{reg}}(\theta) = \tilde{\tau}_{\text{init}}(\theta) - \tilde{\beta}^{\text{T}}(\theta)\tilde{\xi} \quad \text{with} \quad \tilde{\beta}(\theta) = \tilde{E}^{-1}[\tilde{\xi}\tilde{\zeta}^{\text{T}}]\tilde{E}[\tilde{\xi}\tilde{\tau}_{\text{init}}^{\text{T}}(\theta)]$$

where

$$\begin{aligned}
\tilde{\tau}_{\text{init}}(\theta) &= \frac{1-t}{1-\tilde{\pi}(z)}\tilde{\pi}(z)\Phi(x,z;\theta) \\
\tilde{\xi} &= \left(\frac{1-t}{1-\tilde{\pi}(z)} - 1\right)\frac{\tilde{h}(z)}{\tilde{\pi}(z)} \\
\tilde{\zeta} &= \frac{1-t}{1-\tilde{\pi}(z)}\frac{\tilde{h}(z)}{\tilde{\pi}(z)}
\end{aligned}$$

and $\tilde{h}(z) = \{\tilde{h}_1^{\text{T}}(z), \tilde{h}_2^{\text{T}}(z)\}^{\text{T}}$, where we assumed the variables in $\tilde{h}(z)$ are linearly independent.

$$\begin{aligned}
\tilde{h}_1(z) &= \tilde{\pi}(z)\tilde{v}_0(z) \qquad \tilde{v}_0(z) = \{\tilde{\pi}(z), \tilde{\pi}(z)\hat{q}_0^{\text{T}}(z)\}^{\text{T}} \\
\tilde{h}_2(z) &= \tilde{\pi}(z)(1-\tilde{\pi}(z))\{f^{\text{T}}(z), \hat{q}_0^{\text{T}}(z)\}^{\text{T}}
\end{aligned}$$

The different roles of the variables in $\tilde{h}(z)$ can be explained as follows. The variables $\tilde{\pi}(z)\hat{q}_0(z)$ are included in $\tilde{v}_0(z)$ to achieve double robustness, as discussed later through Eq. (4.22). Moreover, the variables in $\tilde{h}_2(z)$ are included to formally achieve intrinsic efficiency as described later in Proposition 4.3.

It is interesting to find out that $\tilde{\xi}$ has mean 0 when PS model is correct. So actually here $\tilde{\xi}$ serves as auxiliary variables (in the terminology of survey sampling) or control variates (in that of Monte Carlo integration). The effect of variance reduction using regression estimators is seen from in the following results.

Proposition 4.3 *The estimator $\tilde{\theta}_{reg}$, which solves $\tilde{E}[\tilde{\tau}_{reg}(\theta)] = 0$, has the following properties.*

(i) $\tilde{\theta}_{reg}$ *is locally nonparametric efficient: it achieves the nonparametric efficiency bound,* V_{NP}*, when both model (4.5) and model (4.8) are correctly specified.*

(ii) $\tilde{\theta}_{reg}$ *is doubly robust: it remains consistent when either model (4.5) or model (4.8) is correctly specified.*

(iii) $\tilde{\theta}_{reg}$ *is intrinsically efficient: if model (4.8) is correctly specified, then it achieves the lowest asymptotic variance among the class of estimators of* θ *that are solutions to* k *estimating equations of the form*

$$\tilde{E}\left[\tilde{\tau}_{init}(\theta) - b^{\mathrm{T}}\tilde{\xi}\right] = 0 \tag{4.21}$$

where b *is a* $dim(h) \times k$ *matrix.*

Previously, we mentioned that $\tilde{\pi}(z)\hat{q}_0(z)$ are included in $\tilde{v}_0(z)$ to achieve double robustness. Here, we would like to emphasize the importance of using the augmented propensity score $\tilde{\pi}(z)$, which makes this estimator still consistent under a correctly specified OR model but a misspecified PS model. If the OR model (4.5) is correctly specified, the estimating Eq. (4.20) is asymptotically equivalent to the first order of

$$\tilde{E}\left\{\frac{1-t}{1-\tilde{\pi}(z)}\tilde{\pi}(z)[\Phi(x,z;\theta) - \hat{q}_0(z)] + \tilde{\pi}(z)\hat{q}_0(z)\right\} = 0 \tag{4.22}$$

which has exactly the same form as Eq. (4.16) but with $\hat{\pi}(z)$ replaced by the augmented propensity score $\tilde{\pi}(z)$. This is because we carefully design $\tilde{h}(z)$ so that $\tilde{\pi}(z)\hat{q}_0(z)$ is a linear combination of variables in $\tilde{h}(z)/\tilde{\pi}(z)$. Let's use $\tilde{\theta}_{\mathrm{SP}}$ to represent the solution to (4.22). Note that because we use augmented PS model $\tilde{\pi}(z)$, we have Eq. (4.18) holds, then (4.22) will be identical to the equation

$$\tilde{E}\left\{\frac{1-t}{1-\tilde{\pi}(z)}\tilde{\pi}(z)[\Phi(x,z;\theta) - \hat{q}_0(z)] + t\hat{q}_0(z)\right\} = 0 \tag{4.23}$$

whose solution $\tilde{\theta}_{\mathrm{NP}}$ is doubly robust as $\hat{\theta}_{\mathrm{NP}}$ based on $\hat{\pi}(z)$. Therefore, $\tilde{\theta}_{\mathrm{reg}}$ is consistent when OR model (4.5) is correctly specified even when the PS model (4.8) is misspecified.

Also based on the asymptotic equivalence between $\tilde{\theta}_{\mathrm{reg}}$ and $\tilde{\theta}_{\mathrm{NP}}$ when OR model

(4.5) is correctly specified, we know $\tilde{\theta}_{\text{reg}}$ is locally nonparametric efficient, similarly as $\tilde{\theta}_{\text{NP}}$. In fact, $\tilde{\theta}_{\text{reg}}$ is generally not locally semiparametric efficient with respect to PS model (4.8), but locally semiparametric efficient with respect to PS model (4.18) in the following sense: $\tilde{\theta}_{\text{reg}}$ achieves the semiparametric efficiency bounded calculated under model (4.18), when both model (4.5) and model (4.8) are correctly specified. When model (4.8) holds, the efficiency bound V_{SP} under model (4.18) coincides with the nonparametric efficiency bound V_{NP}, because $\{t-\pi(z)\}q_0(z)$ is just one component of the score function of model (4.18). On the other hand, $\tilde{\theta}_{\text{reg}}$ with $\tilde{\pi}(z)$ replaced by $\hat{\pi}(z)$ throughout would be locally semiparametric efficient with respect to original PS model (4.8), but generally not doubly robust, similarly as $\hat{\theta}_{\text{SP}}$.

Following the approach of ATE estimation in Tan (2006), we did not apply $\hat{\beta}(\theta) = \tilde{E}[\tilde{\xi}\tilde{\xi}^{\text{T}}]^{-1}\tilde{E}[\tilde{\xi}\tilde{\tau}_{\text{init}}(\theta)]$, the classical optimal choice of b for minimizing the asymptotic variance of (4.21). Although the estimator $\hat{\theta}_{\text{reg}}$, which solves the equation $\tilde{E}[\tilde{\tau}_{\text{init}}(\theta) - \hat{\beta}^{\text{T}}(\theta)\tilde{\xi}] = 0$, is asymptotically equivalent to the first order to $\tilde{\theta}_{\text{reg}}$ when the PS model is correctly specified, it is generally inconsistent estimator of θ, when OR model is correctly specified and PS model may be misspecified. Nevertheless, $\tilde{\theta}_{\text{reg}}$ remains consistent in this case, so it is a doubly robust estimator of θ.

Due to intrinsic efficiency, when PS model is correctly specified, $\tilde{\theta}_{\text{reg}}$ is asymptotically at least as efficient as not only $\tilde{\theta}_{\text{NP}}(=\tilde{\theta}_{\text{SP}})$, but also $\tilde{\theta}_{\text{IPW}}$ solved through Eq. (4.11) with $\hat{\pi}(z)$ replaced by $\tilde{\pi}(z)$. As discussed above, $\tilde{\theta}_{\text{NP}}$ is the solution to Eq. (4.23) constructed by replacing $\hat{\pi}(z)$ in Eq. (4.15) by $\tilde{\pi}(z)$, and $\tilde{\theta}_{\text{NP}}$ remains locally nonparametric efficient and doubly robust, and falls in the class of (4.21) since $\frac{1-t}{1-\tilde{\pi}(z)}\tilde{\pi}(z)\hat{q}_0(z) - t\hat{q}_0(z)$ is a linear combination $(\frac{1-t}{1-\tilde{\pi}(z)} - 1)\tilde{\pi}(z)\hat{q}_0(z)$ and $(t-\tilde{\pi}(z))\hat{q}_0(z)$ and both of them are components of $\tilde{\xi}$. The estimator $\tilde{\theta}_{\text{IPW}}$ based on augmented propensity score $\tilde{\pi}(z)$ directly falls in the class (4.21) with $b = 0$.

Generally, we can not claim $\tilde{\theta}_{\text{reg}}$ is intrinsically efficient within the class of estimators (4.21) specified by $\hat{\pi}(z)$ instead of $\tilde{\pi}(z)$. However, $\tilde{\theta}_{\text{reg}}$ may still often achieve efficiency gains over $\hat{\theta}_{\text{NP}}$ when PS model (4.8) model is correctly specified but OR model (4.5) is misspecified, showed clearly in our simulation studies.

Likelihood Estimators

The regression estimator reduce the estimation variance by introducing the control variate $\tilde{\xi}$. However, a common feature of regression estimator $\tilde{\theta}_{\text{reg}}$ and the IPW estimator $\hat{\theta}_{\text{IPW}}$ is that they both have the inverse weights $\{1-\tilde{\pi}(z)\}^{-1}$ or $\{1-\hat{\pi}(z)\}^{-1}$ based on propensity score and augmented propensity score respectively. While the presence of large propensity score (close to 1) among the observations in auxiliary data will often lead to large variance in the estimation results.

In this section, we will derive the likelihood estimators of θ which is also doubly robust, locally nonparametrically efficient and intrinsically efficient similarly to the regression estimator $\tilde{\theta}_{\text{reg}}$. But this estimator will be less sensitive to large inverse weights than the regression estimators and IPW estimators through introducing the extended propensity score $\omega(z;\lambda)$ discussed in details below.

We need two steps to derive the likelihood estimators achieving all the desirable properties. First, similar to the approach used in ATE estimation of Tan (2010) but with the augmented PS model $\tilde{\pi}(z)$, we could obtain a locally nonparametric efficient, intrinsically efficient, but non-doubly robust estimator by applying the empirical likelihood approach proposed by Owen (2001). Alternatively, the approach of nonparametric likelihood in Tan (2006, 2010) could be applied, which would lead to same results.

Specifically, under this exact-identification case, our approach is to maximize the log empirical likelihood $\sum_{i=1}^{n}\log p_i$ subject to the constraints

$$\sum_{i=1}^{n} p_i = 1 \tag{4.24}$$

$$\sum_{i=1}^{n} p_i\tilde{\xi}_i = 0 \tag{4.25}$$

where p_i is a nonnegative weight assigned to (t_i, x_i, z_i) for $i = 1, ..., n$.

By the calculation process in Qin & Lawless (1994), we get an estimation of p_i.

$$\hat{p}_i = \frac{n^{-1}}{1+\hat{\lambda}^{\mathrm{T}}\tilde{\xi}_i} \tag{4.26}$$

where $\hat{\lambda}$ is a maximizer of the function $\ell_{\mathrm{EL}}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\lambda^{\mathrm{T}}\xi_i)$.

So finally we could solve for θ through Eq. (4.27) below.

$$\sum_{i=1}^{n}\hat{p}_i\left[\frac{1-t_i}{1-\tilde{\pi}(z_i)}\tilde{\pi}(z_i)\Phi(x_i,z_i;\theta)\right]=0 \tag{4.27}$$

And we show in Appendix, after we substitute $\hat{p}_i$ with (4.26), Eq. (4.27) could be represented in another form:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-t_i}{1-\omega(z_i;\hat{\lambda})}\tilde{\pi}(z_i)\Phi(x_i,z_i;\theta)\right]=0 \tag{4.28}$$

where $\omega(z;\lambda)=\tilde{\pi}(z)+\lambda^{\mathrm{T}}\tilde{h}(z)$ and $\hat{\lambda}$ is a maximizer of the function

$$\ell(\lambda)=\tilde{E}\Big[t\log\omega(z;\lambda)+(1-t)\log\{1-\omega(z;\lambda)\}\Big],$$

subject to $\omega(z_i;\lambda)>0$ if $t_i=1$ and $\omega(z_i;\lambda)<1$ if $t_i=0$ for $i=1,\ldots,n$. Setting the gradient of $\ell(\lambda)$ to zero shows that $\hat{\lambda}$ is a solution to

$$\tilde{E}\left[\frac{t-\omega(z;\lambda)}{\omega(z;\lambda)\{1-\omega(z;\lambda)\}}\tilde{h}(z)\right]=0. \tag{4.29}$$

The solution to Eq. (4.28) can be shown to be intrinsically efficient among the class of estimators (4.21) and locally nonparametric efficient, but generally not doubly robust. Similar to the method in Tan (2010), we introduce the following modified likelihood estimators, to achieve double robustness but without affecting the first-order asymptotic behavior. Here we only need to calibrate the estimation based on the auxiliary data part, since the measurements relevant to θ are only in the auxiliary data with the indicator $t=0$.

Partition $\tilde{h}$ as $\tilde{h}=(\tilde{h}_1^{\mathrm{T}},\tilde{h}_2^{\mathrm{T}})^{\mathrm{T}}$ and accordingly λ as $\lambda=(\lambda_1^{\mathrm{T}},\,\lambda_2^{\mathrm{T}})^{\mathrm{T}}$. Define $\tilde{\lambda}=(\tilde{\lambda}_1^{\mathrm{T}},\hat{\lambda}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\hat{\lambda}_2$ are obtained from $\hat{\lambda}$, and $\tilde{\lambda}_1$ is a maximizer of the function

$$\kappa(\lambda_1)=\tilde{E}\left[(1-t)\frac{\log\{1-\omega(z;\lambda_1,\hat{\lambda}_2)\}-\log\{1-\omega(z;\hat{\lambda})\}}{\tilde{\pi}(z)}-\lambda_1^{\mathrm{T}}\tilde{v}_0(z)\right],$$

subject to $\omega(z_i;\lambda_1,\hat{\lambda}_2)<1$ if $t_i=0$ for $i=1,\ldots,n$. Setting the gradient of $\kappa(\lambda_1)$ to 0 shows that $\tilde{\lambda}_1$ is a solution to

$$\tilde{E}\left[\left\{\frac{1-t}{1-\omega(z;\lambda_1,\hat{\lambda}_2)}-1\right\}\tilde{v}_0(z)\right]=0. \tag{4.30}$$

Then we could get another estimator $\tilde{\theta}_{\text{lik}}$ by solving the equation below

$$\tilde{E}\left\{\frac{(1-t)\tilde{\pi}(z)\Phi(x,z;\theta)}{1-\omega(z;\tilde{\lambda})}\right\}=0 \tag{4.31}$$

The likelihood estimator $\tilde{\theta}_{\text{lik}}$ has several desirable properties as follows.

Proposition 4.4 *Under suitable regularity conditions (see Appendix), the estimator $\tilde{\theta}_{lik}$ for θ*

(i) If model (4.8) is correctly specified, then $\tilde{\theta}_{lik}$ is asymptotically equivalent, to the first order, to $\tilde{\theta}_{reg}$. Hence $\tilde{\theta}_{lik}$ is intrinsically efficient among the class (4.21) and locally nonparametric efficient, similarly as $\tilde{\theta}_{reg}$ in Proposition 4.3.

(ii) $\tilde{\theta}_{lik}$ is doubly robust, similarly as $\tilde{\theta}_{reg}$ in Proposition 4.3.

Again, the double robustness of $\tilde{\theta}_{\text{lik}}$ is contributed by two factors. The first one is $\tilde{E}\{(1-t)\tilde{\pi}(z)\hat{q}_0(z)/[1-\omega(z;\tilde{\lambda}_1,\hat{\lambda}_2)]\}=\tilde{E}\{\tilde{\pi}(z)\hat{q}_0(z)\}$ by Eq. (4.30) with $\tilde{\pi}(z)\hat{q}_0(z)$ included in $\tilde{v}_0(z)$, and the other important factor is application of augmented PS model, which brings $\tilde{E}\{\tilde{\pi}(z)\hat{q}_0(z)\}=\tilde{E}\{t\hat{q}_0(z)\}$ by Eq. (4.19). We also include $\tilde{\pi}(z)$ inside $\tilde{v}_0(z)$ to obtain $\tilde{E}\{(1-t)\tilde{\pi}(z)/[1-\omega(z;\tilde{\lambda}_1,\hat{\lambda}_2)]\}=\tilde{E}\{\tilde{\pi}(z)\}=\tilde{E}(t)$

Also the intrinsic efficiency makes $\tilde{\theta}_{\text{lik}}$ superior to the doubly robust AIPW estimator $\tilde{\theta}_{\text{NP}}$, the solution to (4.23), because $\tilde{\theta}_{\text{lik}}$ is asymptotically at least as efficient as $\tilde{\theta}_{\text{NP}}$ as long as PS model (4.8) is correctly specified.

4.2 Two-Sample data combination

Based on improved estimators discussed above under the moment restriction models with auxiliary data, we could apply the same method to deal with more complicated

cases introduced by Graham et al. (2015), where we have an additional variable y which is observed in *primary data* as well as z. In other words, (y, z) are measured from *primary data*, and (x, z) are measured from *auxiliary data*. The moment restriction model could be represented in the form of two separate parts:

$$E^{(1)}\big[\Phi_1(y, z; \theta) - \Phi_0(x, z; \theta)\big] = 0 \tag{4.32}$$

where $\Phi_1(\cdot)$ is a function of y and z only, and Φ_0 is a function of x and z only. Note that θ is identifiable only when functions could be written in this separable form. Since we could obtain the measurements of (y, z) under the population (1) directly from primary data, we could easily estimate $E^{(1)}\Phi_1(y, z; \theta)$ by the simple average of the functions from primary data. The difficulty of this problem is to identify $E^{(1)}\Phi_0(x, z; \theta)$.

A simple example of this case is the estimation of average treatment effect on the treated (ATT). In our notation, y represents the potential outcome of an individual under active treatment, while x denotes the opposite potential outcome in control group, and z here refers to the pretreatment covariates (Hahn 1998). Moreover, we are also given two samples; one random sample from the treatment group with the measurements of (y, z), and the other is from the control group with (x, z) recorded. For ATT estimation, the estimation of $\nu^1 = E^{(1)}(y)$ could be easily obtained by the observation in the primary data, while the challenge is to estimate ν^0 in the equation

$$E^{(1)}[x - \nu_0] = 0$$

where ν^0 is actually the θ we are interested in (4.1).

Another example would the two-sample instrumental variable problem, with the purpose of investigating the causality of some variable on the outcome when regression estimates don't always provide a consistent estimation due to the possibility of endogenous regressors. Suppose y is the response variable with a linear relationship with x and z_0^c, where x is a scalar, and z_0^c is a subvector of $z = (z_0, z_0^c)$. z_0 is also a scalar, and it is the instrumental variable of x and hence not in a direct relationship with the response variable y. The vector z are common to the two samples. For the purpose

of distinction, let's refer to $z^{(1)}$ as the z in Data (1), and $z^{(0)}$ as the z in Data (0). This set-up is first studied by Klevmarken (1982). In our notation, he considered (for convenience, we assume all the variables have been centered without loss of generality, so no intercept is needed):

$$y = \beta_0 x + \beta^{\mathrm{T}} z_0^c + \varepsilon \tag{4.33}$$

The interest of problem is to estimate the marginal effect of x on y. He assumes x is endogenous and all the common variables z are exogenous and that z contains all exogenous variables. And the relationship between x and the exogenous common variables are

$$x = \vartheta^{\mathrm{T}}(z_0, z_0^{c\mathrm{T}})^{\mathrm{T}} + \epsilon \tag{4.34}$$

This problem is revisited by Angrist & Krueger (1992) also assuming all the common variables are exogenous.

There exist a common assumption for Klevmarken (1982) and Angrist & Krueger (1992): the two independent samples are drawn from the same population with x missing in the primary data and y missing in the auxiliary data. This drawback severely limits the application of this estimator in real problems, and we will develop some estimators to overcome this difficulty in this chapter.

Before developing our methods, let's go through two most classic estimators: *Two-Sample Instrumental Variable (TSIV)* (Angrist & Krueger, 1992) and *Two-Sample Two-Stage Least Squares (TS2SLS)* (Bjorklund & Jantti, 1997) under the setting (4.33).

4.2.1 Existing Estimators

Angrist & Krueger (1992) showed that under certain conditions, consistent estimator still exists if two samples are drawn from the same joint distribution. Y and $Z_0^{(1)}$ are both $n_1 \times 1$ vectors representing the data drawn from Data (1), and X and $Z_0^{(0)}$ are $n_0 \times 1$ vectors representing the data coming from Data (0). This *Two-Sample Instrumental*

Variable (TSIV) estimator could be represented as follows.

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{pmatrix}_{\text{TSIV}} = \left(Z^{(0)\text{T}} X/n_0, \quad Z^{(1)\text{T}} Z_0^{c(1)}/n_1 \right)^{-1} \left(Z^{(1)\text{T}} Y/n_1 \right) \tag{4.35}$$

There is another estimator based on the idea of imputation, generally called *Two-Sample Two-Stage Least Squares (TS2SLS)* ((Bjorklund & Jantti, 1997)). It actually could be regarded as imputation estimator.

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{pmatrix}_{\text{TS2SLS}} = (\hat{W}^{(1)\text{T}} \hat{W}^{(1)})^{-1} \hat{W}^{(1)\text{T}} Y \tag{4.36}$$

where $\hat{W}^{(1)} = (\hat{X}^{(1)}, Z_0^{c(1)})$, $\hat{X}^{(1)} = Z^{(1)}(Z^{(0)\text{T}} Z^{(0)})^{-1} Z^{(0)\text{T}} X$.

Inoue & Solon (2010) compared these two estimators, and illustrate the relations and differences between them very well.

4.2.2 Another Representation

Based on the linear relationship of (4.33), we could construct the following estimating equation to solve the parameter of interest, (β_0, β), using the representation form like (4.1).

$$E^{(1)}\Big[(y - \beta_0 x - \beta^{\text{T}} z_0^c) z \Big] = 0 \tag{4.37}$$

where $E^{(1)}$ is the expectation taken with respect to the joint distribution of Data (1). Actually (4.37) could be written in the following form:

$$E^{(1)}(yz) - E^{(1)}(z z_0^{c\text{T}})\beta - E^{(1)}(xz)\beta_0 = 0 \tag{4.38}$$

Then the parameters of interest could be explicitly represented by:

$$\begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} = \left[E^{(1)}(xz), \quad E^{(1)}(zz_0^{c\mathrm{T}}) \right]^{-1} E^{(1)}(yz) \tag{4.39}$$

Since the measurements of (y, z) could be directly obtained from Data (1), it is easy to estimate $E^{(1)}(yz)$ and $E^{(1)}(zz_0^{c\mathrm{T}})$ through directly taking the sample average of the measurements in the primary data. This simple average estimator is always consistent estimation of $E^{(1)}(yz)$ and $E^{(1)}(zz_0^{c\mathrm{T}})$, and it is also locally efficient, achieving the semiparametric variance bound without any knowledge of the propensity score, which has been discussed in the semiparametric efficiency theory of ATT estimation in Chapter 3.

Then the difficulty of the problem is to estimate $E^{(1)}(xz)$, because x is missing from primary data. Comparing (4.35) and (4.36) with the original estimating function (4.39) shows that, in order to estimate $E^{(1)}(xz)$, TSIV approach only combines the measurements of x from auxiliary data and the measurements of z from primary data, which would be inconsistent if the distribution of x differ among the two samples. On the other hand, the TS2SLS approach utilizes the imputation of x based on the least square estimates from the auxiliary data. This estimator is consistent when the imputation is consistent estimates of x, but its efficiency has not been studied. And TS2SLS estimator can be shown to be consistent when the two samples have the same distribution, even if the imputation regression model is misspecified.

In fact, the estimation of $E^{(1)}(xz)$ also falls into the general framework of (4.1), where $\Phi(x, z; \theta) = xz - \theta$ under this notation. Therefore, let's apply our method discussed in Section 4.1 to this specific problem.

4.3 Two-Sample Instrumental Variable

As discussed in Section 4.2.2, in order to solve β_0 and β, we need to estimate three conditional expectations.

$$\mu_1 = E(yz|t=1) \qquad \mu_2 = E(zz_0^{c\mathrm{T}}|t=1) \qquad \mu_3 = E(xz|t=1) \tag{4.40}$$

Following the general framework in Section 4.1, μ_1 and μ_3 are both $k \times 1$ vectors, where k is the dimension of variable z. μ_2 is a $k \times (k-1)$ matrix, since z_0 is a scalar.

Because $E(yz|t=1) = E(tyz)/E(t)$, $E(zz_0^{c\mathrm{T}}|t=1) = E(tzz_0^{c\mathrm{T}})/E(t)$ and the measurements of (y, z) are easily obtained from the primary data indicated by $t = 1$, we could easily get the estimators of μ_1 and μ_2.

$$\begin{aligned} \hat{\mu}_1 &= \tilde{E}(tyz)/\tilde{E}(t) \\ \hat{\mu}_2 &= \tilde{E}(tzz_0^{c\mathrm{T}})/\tilde{E}(t) \end{aligned} \tag{4.41}$$

Actually, from the semiparametric theory for ATT estimation in Chapter 3, we know $\hat{\mu}_1$ and $\hat{\mu}_2$ are fully robust (always consistent) estimators of μ_1 and μ_2. Moreover, they are locally nonparametric efficient, achieving the semiparametric variance bound of estimating μ_1 and μ_2 under nonparametric model.

For the estimation of $\mu_3 = E(xz|t=1)$, we could fit this problem into the general framework discussed in previous sections by defining $\Phi(x, z; \theta) = xz - \theta$, where θ is a unknown $k \times 1$ parameter with the same dimension as z. *And θ defined in the general framework is just μ_3 we would like to find out.*

Following the general framework, we also define propensity score model and outcome regression model in this special case. For PS model, let's use logistic regression defined exactly the same as the general approach (4.9). Plug $\Phi(x, z; \theta) = xz - \theta$ into the IPW estimating equation (4.11) under the general framework, then we could obtain IPW estimator of $\theta(\mu_3)$ based on the fitted propensity score $\hat{\pi}(z)$:

$$\hat{\mu}_{3,\mathrm{IPW}} = \hat{\theta}_{\mathrm{IPW}} = \tilde{E}\left\{\frac{(1-t)\hat{\pi}(z)xz}{1-\hat{\pi}(z)}\right\} \Big/ \tilde{E}\left\{\frac{(1-t)\hat{\pi}(z)}{1-\hat{\pi}(z)}\right\}, \tag{4.42}$$

which is consistent when propensity score model (4.9) is correctly specified.

Now let's consider building an outcome regression model for estimating $E[\Phi(x, z; \theta)|z]$ in the case when $\Phi(x, z; \theta) = xz - \theta$. Since $E(xz|z)$ could be written as $E(x|z)z$, we could first build a model for $E(x|z)$.

$$m_0(z) = E(x|z) = \Psi\{\alpha_0^{\mathrm{T}} g(z)\} \tag{4.43}$$

where $\Psi(\cdot)$ is an inverse link function, $g(z)$ is a vector of known functions of z *including 1*, and α_0 is a vector of unknown parameters. let $\hat{\alpha}_0$ be the maximum quasi-likelihood estimate of α_0, and let $\hat{m}_0(z) = m_0(z; \hat{\alpha}_0)$. Then the fitted OR model is $\hat{q}_0(z) = \hat{m}_0(z)z - \theta$. Substituting this representation into (4.7), we could obtain the estimation of θ based on the OR model:

$$\hat{\mu}_{3,\mathrm{OR}} = \hat{\theta}_{\mathrm{OR}} = \tilde{E}\big\{tz\hat{m}_0(z)\big\}/\tilde{E}(t), \tag{4.44}$$

which is consistent only when outcome regression model is correctly specified.

It is interesting to notice the similarity and difference between the OR estimators here and the TS2SLS estimators (4.36). After plugging in $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\mu}_{3,\mathrm{OR}}$, we could obtain the estimator of β_0 and β based on (4.39). The estimates computed through this OR approach appear very similar to the TS2SLS estimates, but with a subtle difference. The TS2SLS estimator solves the problem through the estimating equation

$$E^{(1)}\left\{\left[y - \beta_0\hat{m}_0(z) - \beta^{\mathrm{T}} z_0^c\right]\begin{pmatrix}\hat{m}_0(z)\\ z_0^c\end{pmatrix}\right\} = 0 \tag{4.45}$$

While our OR approach solves another similar estimating equation

$$E^{(1)}\left\{\left[y - \beta_0\hat{m}_0(z) - \beta^{\mathrm{T}} z_0^c\right]\begin{pmatrix}z_0\\ z_0^c\end{pmatrix}\right\} = 0 \tag{4.46}$$

Therefore, the two equations generally lead to different estimators of β_0 and β. When $\hat{m}_0(z)$ is a linear combination of $z = (z_0, z_0^c)$, the two estimating equations above are

equivalent to each other, and will lead to the same estimations of β_0 and β.

Similarly, we could derive the doubly robust AIPW estimator $\hat{\theta}_{\text{NP}}$ defined through Eq. (4.15) in Proposition 4.2. By plugging $\Phi(x, z; \theta) = xz - \theta$ and $\hat{q}_0(z) = \hat{m}_0(z)z - \theta$ into Eq. (4.15), we easily get the AIPW estimator of $\theta(\mu_3)$ here:

$$\hat{\mu}_{3,\text{AIPW}} = \hat{\theta}_{\text{AIPW}} = \tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\hat{\pi}(z)xz - \left(\frac{1-t}{1-\hat{\pi}(z)} - 1\right)\hat{m}_0(z)z\right\}\Big/\tilde{E}(t) \tag{4.47}$$

4.3.1 Regression Estimator

Follow the setup in the general case, plug in $\Phi(x, z; \theta) = xz - \theta$ and $\hat{q}_0(z) = \hat{m}_0(z)z - \theta$ into the general setup. Meanwhile, in order to make the variables in $\tilde{h}(z)$ are linearly independent, we could do some simplification here.

Generally, $\tilde{v}_0(z)$ is defined as $\{\tilde{\pi}(z), \tilde{\pi}(z)\hat{q}_0^{\text{T}}(z)\}^{\text{T}}$. After we plug in $\hat{q}_0(z)$, $\tilde{v}_0(z) = \{\tilde{\pi}(z), \tilde{\pi}(z)\hat{m}_0(z)z^{\text{T}} - \tilde{\pi}(z)\theta^{\text{T}}\}^{\text{T}}$. Since θ is a constant, the representation of $\tilde{v}_0(z)$ could be simplified to $\tilde{v}_0(z) = \{\tilde{\pi}(z), \tilde{\pi}(z)\hat{m}_0(z)z^{\text{T}}\}^{\text{T}}$. Similarly, we could perform the same simplification to the $\hat{q}_0(z)$ in the augmented PS model, since $f(z)$ already contains the constant term. Hence $\tilde{\xi}$, $\tilde{\zeta}$ and $\tilde{h}(z)$ are redefined as following:

$$\tilde{\xi} = \left(\frac{1-t}{1-\tilde{\pi}(z)} - 1\right)\frac{\tilde{h}(z)}{\tilde{\pi}(z)} \qquad \tilde{\zeta} = \frac{1-t}{1-\tilde{\pi}(z)}\frac{\tilde{h}(z)}{\tilde{\pi}(z)} \tag{4.48}$$

with $\tilde{h}(z) = (\tilde{h}_1^T(z), \tilde{h}_2^{\text{T}}(z))^{\text{T}}$, where

$$\tilde{h}_1(z) = \tilde{\pi}(z)\tilde{v}_0(z) \qquad \tilde{h}_2(z) = \tilde{\pi}(z)(1-\tilde{\pi}(z))\{f^{\text{T}}(z), \hat{m}_0(z)z^{\text{T}}\}^{\text{T}}$$

On the other hand, now we have

$$\tilde{\tau}_{\text{init}}(\theta) = \frac{1-t}{1-\tilde{\pi}(z)}\tilde{\pi}(z)[xz - \theta]$$

When we plug all the representations in this specific case, the regression estimator $\tilde{\theta}_{\text{reg}}$

could be denoted by

$$\tilde{\theta}_{\text{reg}} = \frac{\tilde{E}(\tilde{\eta}) - \tilde{\beta}^{\mathrm{T}}\tilde{E}(\tilde{\xi})}{\tilde{E}(\tilde{\rho}) - \tilde{\kappa}^{\mathrm{T}}\tilde{E}(\tilde{\xi})} \tag{4.49}$$

where

$$\tilde{\eta} = \frac{1-t}{1-\tilde{\pi}(z)}\tilde{\pi}(z)xz \qquad \tilde{\rho} = \frac{1-t}{1-\tilde{\pi}(z)}\tilde{\pi}(z) \tag{4.50}$$

$$\tilde{\beta} = \tilde{E}^{-1}[\tilde{\xi}\tilde{\zeta}^{\mathrm{T}}]\tilde{E}[\tilde{\xi}\tilde{\eta}^{\mathrm{T}}] \qquad \tilde{\kappa} = \tilde{E}^{-1}[\tilde{\xi}\tilde{\zeta}^{\mathrm{T}}]\tilde{E}[\tilde{\xi}\tilde{\rho}^{\mathrm{T}}] \tag{4.51}$$

Furthermore, it is easy to see that $\tilde{\kappa} = (1, 0, 0, \cdots)^{\mathrm{T}}$, then the denominator of (4.49) above is equivalent to $\tilde{E}(t)$. So finally the regression estimate of θ is

$$\tilde{\theta}_{\text{reg}} = \tilde{E}\left(\tilde{\eta} - \tilde{\beta}^{\mathrm{T}}\tilde{\xi}\right) \Big/ \tilde{E}(t) \tag{4.52}$$

4.3.2 Likelihood Estimator

Similar to the regression estimator in this specific setup, let's use the new simplified version of the constraints.

$$\tilde{h}(z) = (\tilde{h}_1^{\mathrm{T}}(z), \tilde{h}_2^{\mathrm{T}}(z))^{\mathrm{T}} \qquad \tilde{v}_0(z) = \{\tilde{\pi}(z), \tilde{\pi}(z)\hat{m}_0(z)z^{\mathrm{T}}\}^{\mathrm{T}}$$

$$\tilde{h}_1(z) = \tilde{\pi}(z)\tilde{v}_0(z) \qquad \tilde{h}_2(z) = \tilde{\pi}(z)(1-\tilde{\pi}(z))\{f^{\mathrm{T}}(z), \hat{m}_0(z)z^{\mathrm{T}}\}^{\mathrm{T}}$$

Following the two steps in the general case, first solve $\hat{\lambda}$ which maximizes

$$\ell(\lambda) = \tilde{E}\Big[t\log\omega(z;\lambda) + (1-t)\log\{1-\omega(z;\lambda)\}\Big],$$

where $\omega(z;\lambda) = \tilde{\pi}(z) + \lambda^{\mathrm{T}}\tilde{h}(z)$.

Secondly, define $\tilde{\lambda} = (\tilde{\lambda}_1^{\mathrm{T}}, \hat{\lambda}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\hat{\lambda}_2$ are obtained from $\hat{\lambda}$, and $\tilde{\lambda}_1$ is a maximizer of the function

$$\kappa(\lambda_1) = \tilde{E}\left[(1-t)\frac{\log\{1-\omega(z;\lambda_1,\hat{\lambda}_2)\} - \log\{1-\omega(z;\hat{\lambda})\}}{\tilde{\pi}(z)} - \lambda_1^{\mathrm{T}}\tilde{v}_0(z)\right],$$

subject to $\omega(z_i;\lambda_1,\hat{\lambda}_2)<1$ if $t_i=0$ for $i=1,\ldots,n$. Setting the gradient of $\kappa(\lambda_1)$ to 0 shows that $\tilde{\lambda}_1$ is a solution to

$$\tilde{E}\left[\left\{\frac{1-t}{1-\omega(z;\lambda_1,\hat{\lambda}_2)}-1\right\}\tilde{v}_0(z)\right]=0. \tag{4.53}$$

After we obtain the calibrated extended propensity score $\omega(z;\tilde{\lambda}_1,\hat{\lambda}_2)$, just plug in $\Phi(x,z;\theta)=xz-\theta$ into Eq. (4.31), we could solve the doubly robust likelihood estimator of θ of this specific problem:

$$\tilde{\theta}_{\text{lik}}=\tilde{E}\left\{\frac{1-t}{1-\omega(z;\tilde{\lambda})}\tilde{\pi}(z)xz\right\}\Big/\tilde{E}\left\{\frac{1-t}{1-\omega(z;\tilde{\lambda})}\tilde{\pi}(z)\right\} \tag{4.54}$$

$$=\tilde{E}\left\{\frac{1-t}{1-\omega(z;\tilde{\lambda})}\tilde{\pi}(z)xz\right\}\Big/\tilde{E}(t) \tag{4.55}$$

where the second equality holds because of Eq. (4.53) with $\tilde{\pi}(z)$ included in $\tilde{v}_0(z)$ and $\tilde{E}[t-\tilde{\pi}(z)]=0$ by the score equation of the propensity score in logistic regression form.

4.3.3 Estimation of β_0

After obtaining doubly robust, locally nonparametric efficient and intrinsically efficient estimator of $\theta=E(xz|t=1)$ based on the likelihood estimation or regression estimation talked above, we could easily solve the estimating equation to get the estimator of β_0, the causal effect of x on y.

$$\begin{pmatrix}\tilde{\beta}_0\\ \tilde{\beta}\end{pmatrix}=\begin{bmatrix}\tilde{\theta}, & \hat{\mu}_2\end{bmatrix}^{-1}\hat{\mu}_1 \tag{4.56}$$

where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the sample average defined in (4.41), and $\tilde{\theta}$ could be $\tilde{\theta}_{\text{reg}}$ or $\tilde{\theta}_{\text{lik}}$.

Since $\hat{\mu}_1$ and $\hat{\mu}_2$ are doubly robust and locally nonparametric efficient estimator of $E^{(1)}(zz_0^{c\text{T}})$ and $E^{(1)}(yz)$ separately, and $\tilde{\theta}_{\text{reg}}$ or $\tilde{\theta}_{\text{lik}}$ is also doubly robust and locally nonparametric efficient estimator of $E^{(1)}(xz)$, we could conclude that our estimator $\tilde{\beta}_0$ obtained from Eq. (4.56) is also doubly robust and locally nonparametric efficient.

4.4 Simulation Studies

In order to assess the performance of our doubly robust likelihood estimator to solve the two-sample combination problem, we design the following simulation set-up under the two sample instrumental variable framework and compare with the some existing classical estimators. In total, we have five random variables, (y, x, z_0, z_1, z_2), where y is the response variable, z_0 is instrumental variable, and $z_0^c = (z_1, z_2)$ are other covariants in the regression model. Our target is to find out the marginal effect of x on y. We have two data sets, the primary dataset, Data (1), which contains (y, z_0, z_1, z_2), and the auxiliary dataset, Data (0), which is constituted of (x, z_0, z_1, z_2). When we merge the two data sets into one, we need one indicator variable t, and let $t = 1$ refers to Data (1), and $t = 0$ refers to Data (0).

With the purpose to check the robustness of the estimators also recover the most general practical situations, we assume the common variables (z_0, z_1, z_2) have different joint distribution across the two samples. We assume (z_0, z_1, z_2) are mutually independent, and they follow $N(1, 1)$ marginally in Data (1), and are $N(0, 1)$ marginally distributed in Data (0). According to Qin (1998, 1999), based on the density ratio in the two samples, we could easily derive the true underlying propensity score in this setting

$$P(t = 1|z) = \text{expit}\left\{-1.5 - \log 10 + z_0 + z_1 + z_2\right\} \tag{4.57}$$

Let $y = 0.5x - 0.4z_1 + 0.5z_2 + u$, where u is the error. However, x and u are correlated, so OLS estimator will be biased. Suppose $x = z_0 + 0.6z_1 - 0.5z_2 + v$, where v is also the error term. In addition, we assume

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\right]$$

It is easy to see that the *true marginal causal effect of x on y is 0.5* .

Let's assume $w_0 = \exp(-0.5z_0) + 5$, $w_1 = z_1/\{1 + 0.1\exp(z_0)\} + 10$, and $w_2 =$

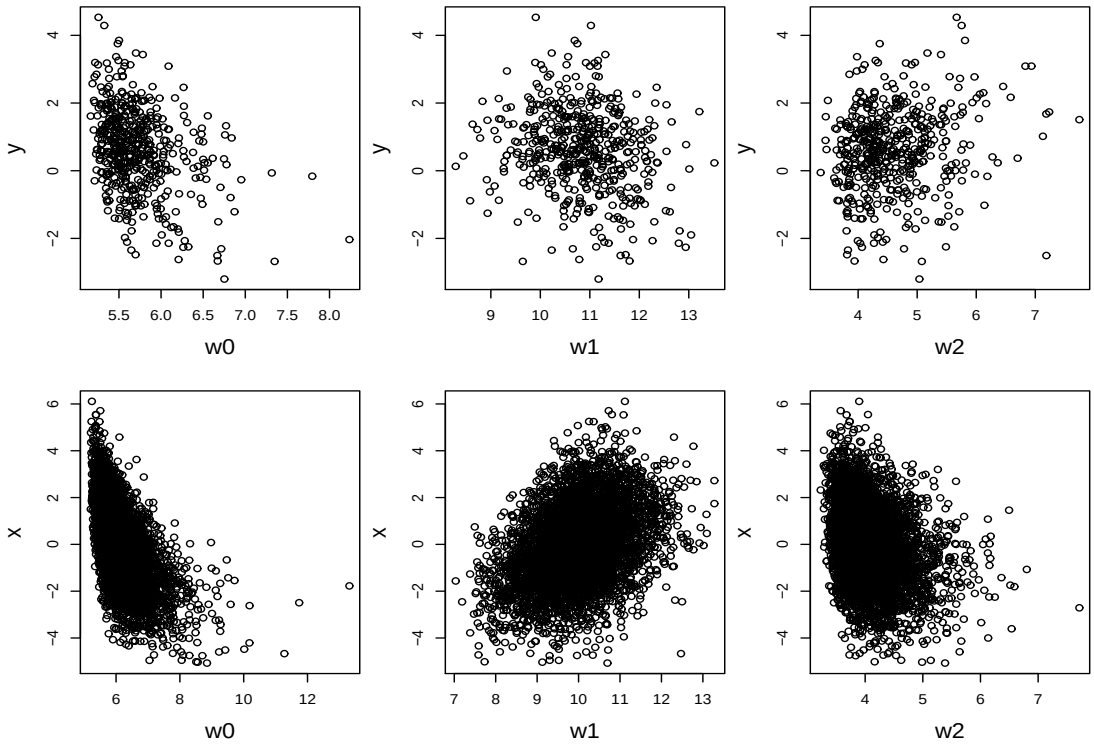

Figure 4.1: Scatterplots of y and x vs. misspecified variables in two samples

$\exp(0.4z_2) + 3$. We construct OR model with the identity link $\Psi(\cdot)$ based on $g_0(z) = (1, z_0, z_1, z_2)$ or $g_0(z) = (1, w_0, w_1, w_2)$. Based on the true data setting, they correspond to correct and misspecified OR model separately. Similarly, we build the PS model with the logistic link $\Pi(\cdot)$ based on $f(z) = (1, z_0, z_1, z_2)$ or $f(z) = (1, w_0, w_1, w_2)$. Based on the true propensity score (4.57), they correspond to correct and misspecified PS model separately.

In general, Data (1) is the primary data set with limited sample size, while Data (0) is the auxiliary data with enough data. Here we assume the sample size of Data (1) is $n_1 = 500$, and the sample size of Data (0) is $n_0 = 5000$, ten times of the size of Data (1). We show the scatterplots of response y and the misspecified variables (w_0, w_1, w_2) in Data (1) in the upper windows of Figure 4.1 and scatterplots of regressor x and the misspecified variables (w_0, w_1, w_2) in Data (0) in the lower windows of Figure 4.1. According to the plots, it is reasonable for us to build linear OR model to predict conditional mean of x and a linear regression model for the response y using the misspecified

Table 4.2: Estimators of β_0

	TSIV	TS2SLS	OR	IPW	AIPW	LIK
Estimator	$\hat{\beta}_{0,\text{TSIV}}$	$\hat{\beta}_{0,\text{TS2SLS}}$	$\tilde{\beta}_0(\hat{\theta}_{\text{OR}})$	$\tilde{\beta}_0(\hat{\theta}_{\text{IPW}})$	$\tilde{\beta}_0(\hat{\theta}_{\text{NP}})$	$\tilde{\beta}_0(\tilde{\theta}_{\text{lik}})$

variables as the regressors.

We implement the following estimators of the coefficient β_0 listed in Table 4.2. $\tilde{\beta}_0(\hat{\theta}_{\text{OR}})$, $\tilde{\beta}_0(\hat{\theta}_{\text{IPW.ratio}})$, $\tilde{\beta}_0(\hat{\theta}_{\text{NP}})$ and $\tilde{\beta}_0(\tilde{\theta}_{\text{lik}})$ are the solution of β_0 to Eq. (4.56) by replacing $\tilde{\theta}$ into $\hat{\theta}_{\text{OR}}$, $\hat{\theta}_{\text{IPW}}$, $\hat{\theta}_{\text{NP}}$ and $\tilde{\theta}_{\text{lik}}$ defined in Section 4.3, respectively.

Table 4.3 lists the bias and standard error of different estimators to solve two-sample instrument variable problem under four model specification scenarios, based on 1000 Monte Carlo samples of size $n = 5500$ with $n_1 = 5000$ and $n_0 = 500$. Figure 4.2 shows the boxplots of the gap between estimators and the true value from 1000 Monte Carlo samples. The realizations of each estimator are censored within the range of the y-axis, and the number of realizations that lie outside the range are marked next to the lower limits and upper limits of the frame for each estimator.

The TSIV estimator $\hat{\beta}_{0,\text{TSIV}}$ proposed by Angrist & Krueger (1992) doesn't show any difference under the four different scenarios since it doesn't depend on PS model or OR model. From the boxplots, we can see that $\hat{\beta}_{0,\text{TSIV}}$ has dramatic bias in all the scenarios which is not hard to explain since the common variables z in our setting have different distribution between two samples. For TSIV estimator (Angrist & Krueger 1992) to be consistent, different samples need to be drawn from the same population.

As a method depending on OR model, both TS2SLS estimator $\hat{\beta}_{0,\text{TS2SLS}}$ and OR estimator $\tilde{\beta}_0(\hat{\theta}_{\text{OR}})$ are approximately unbiased when OR model is correctly specified. By the discussion of comparing OR estimator and TS2SLS estimator based on (4.45) and (4.46), we know these two will generate the same estimates when OR model is correctly specified in our setting, which agrees with the numerical results in the table and figure. There exist difference between $\hat{\beta}_{0,\text{TS2SLS}}$ and $\tilde{\beta}_0(\hat{\theta}_{\text{OR}})$ when OR model is misspecified, since $\hat{m}_0$ is no longer a linear combination of z, but w. Both of them become biased when OR model is misspecified.

Table 4.3: Estimates (Bias and Standard Error) of β_0

	TSIV	TS2SLS	OR	IPW	AIPW	LIK
Correct PS, Correct OR	0.18859	-0.00069	-0.00070	0.01129	-0.00007	-0.00087
	(0.08008)	(0.04393)	(0.04393)	(0.09537)	(0.05701)	(0.05195)
Correct PS, Misspecified OR	0.18859	0.19009	0.45474	0.01129	0.00752	0.01129
	(0.08008)	(0.07664)	(0.09285)	(0.09537)	(0.09559)	(0.06403)
Misspecified PS, Correct OR	0.18859	-0.00069	-0.00070	0.40183	-0.00107	-0.00105
	(0.08008)	(0.04393)	(0.04393)	(1.51650)	(0.15194)	(0.05230)
Misspecified PS, Misspecified OR	0.18859	0.19009	0.45474	0.40183	0.13473	0.06672
	(0.08008)	(0.07664)	(0.09285)	(1.51650)	(1.43629)	(0.06671)

We show the numerical results of different estimators under four scenarios where the outcome regression and/or propensity score models are misspecified. Each cell gives the bias (upper) and standard error (lower) of the point estimators. The simulation is based on 1000 Monte Carlo samples with size $n = 5500$ with $n_1 = 500$ and $n_0 = 5000$.

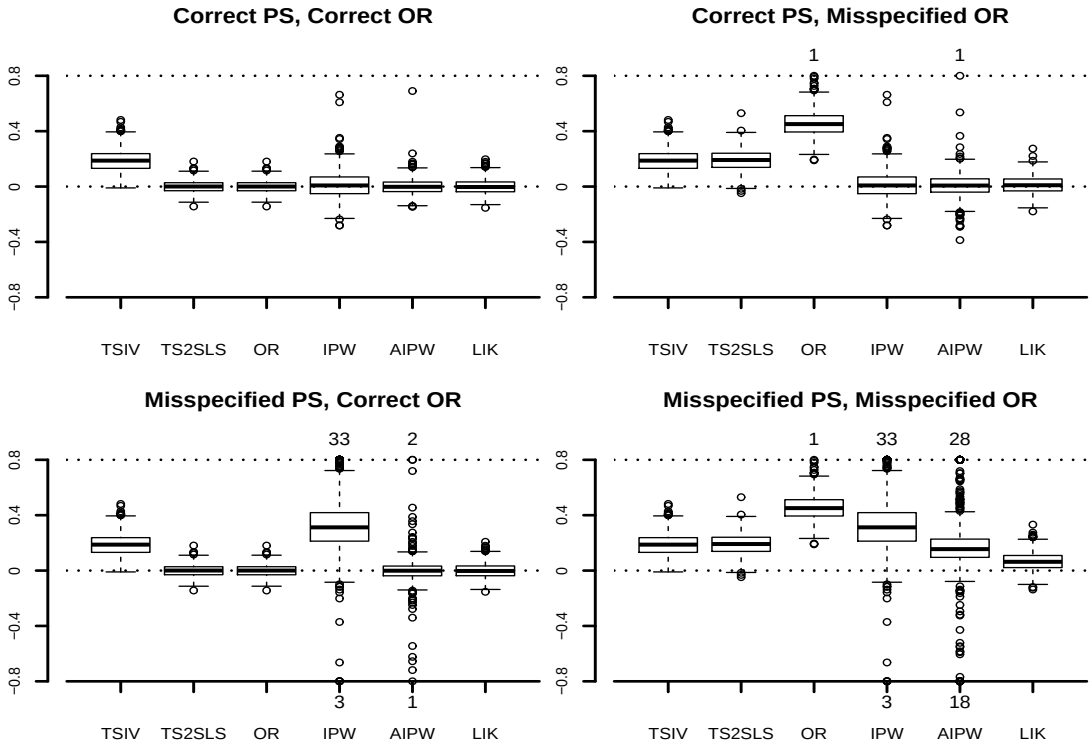

Figure 4.2: Boxplots of Estimators

The IPW estimator $\tilde{\beta}_0(\hat{\theta}_{\text{IPW}})$ is consistent only when PS model is correctly specified, but it has relatively large variance no matter PS model is correctly specified or misspecified.

Both AIPW estimator $\tilde{\beta}_0(\hat{\theta}_{\text{NP}})$ and LIK estimator $\tilde{\beta}_0(\tilde{\theta}_{\text{lik}})$ are doubly robust and locally efficient, which is also reflected in our simulation: they are approximately unbiased when either PS model or OR model is correctly specified, and they have similar standard error to each other when both PS model and OR model are correctly specified, since the two estimators both achieve the semiparametric efficiency bound under nonparametric model assumption V_{NP} when two models are correctly specified. It is interesting to notice that when PS model is correctly specified but OR model is misspecified, our LIK estimator $\tilde{\beta}_0(\tilde{\theta}_{\text{lik}})$ has smaller standard error than AIPW estimator $\tilde{\beta}_0(\hat{\theta}_{\text{NP}})$, mainly because we use $\tilde{\theta}_{\text{lik}}$ which is intrinsically efficient for calculating $\tilde{\beta}_0(\tilde{\theta}_{\text{lik}})$.

Also even when both PS model and OR model are misspecified, our LIK estimator $\tilde{\beta}_0(\tilde{\theta}_{\text{lik}})$ performs the best among all with the smallest bias and standard error.

4.5 Re-assess the Outcome of Public Housing Projects

In order to improve the quality of housing and prospects of children in poor families, the Federal Government provide substantial housing subsidies on the public housing projects for the low-income families since 1937. However, with the increase of the number of household assisted, the dissatisfaction to the public housing also grow rapidly, largely in response to the rising cost of public housing and the high rates of crime, unemployment and school failure among public housing residents. But actually there is little evidence on the bad impact of public housing on the kids. Currie & Yelowitz (2000) worked on this topic to investigate the true outcome of public housing project on the living quality and children's education attainment, by combining information from several data sources.

They would like to find out the effect of public housing project on three outcomes separately (y in our notation): two direct measures of housing quality (overcrowding and density), as well as grade repetition, a measure of children's educational attainment.

For convenience and simplicity of comparison, *we only take "overcrowdedness" as the outcome y we are interested in our analysis.* They define the family is overcrowded if it has fewer than three living/bedrooms. That means if the family has less than three living/bedrooms, $y = 1$, otherwise $y = 0$. In the linear model of overcrowdedness, they include project participation as x, as well as additional exogenous explanatory variables as z_0^c such as household head's gender, age, race, education, marital status and the number of boys in the family and so on.

As an initial analysis, Currie & Yelowitz estimated ordinary least squares (OLS) regression of the effects of project participation on the outcomes, based on the data from 1992 and 1993 waves of the Survey of Income and Program Participation (SIPP). The OLS results show that living in a project house will bring poorer outcomes. However, it is very likely that OLS estimate is biased by selection. All the families eligible for registering the project are the household with income at or below 50% of the area median, so they are selected to be disadvantaged. The inferior living quality and kid's weak study resources accompanied with the household in the project may only have correlation with the public housing project, but not caused by it. And there may be some relevant explanatory variables unobserved in the data or omitted from the model. Therefore, instrumental variable methods are necessary in digging out the true causal effect.

The instrumental variable z_0 used by Currie & Yelowitz is the indicator variable whether the household is arranged to a larger house in a project due to the sex composition of the children in the household. Based on the Department of Housing and Urban Development (HUD)'s rules, boys and girls cannot be required to share one bedroom except very young children. As a result, the family with two boys or two girls will be entitled to a two-bedroom apartment, while the family with a boy and a girl will be eligible for a three-bedroom apartment. In order to focus on the effects of sex composition and abstract from any effects due to the number of children, *they restrict the analysis to families with exactly two children under 18 in the household.* On the other hands, in order for z_0, sex composition/extra room, to be a valid instrument for x, project participation, z_0 should influence x, but have no direct effect on the outcome

variable y, overcrowding, except through x. The benefit of extra room will be very likely to attract family with a boy and a girl to participate the housing project. And indeed in the first stage regression of TS2SLS approach Currie & Yelowitz (2000) used, which will be described in details below, it is shown that adding an extra bedroom increases the likelihood of project participation by 24%. On the other hand, Currie & Yelowitz (2000) pointed out that there is little reason to expect sex composition will affect overcrowding, at least in the way of their definition of overcrowding. Since they define "overcrowdedness", $y = 1$ if the family has less than two bedrooms. Then for y changing from 1 to 0, family with two kids will seek two or more bedroom instead of one bedroom. But kids of opposite sex would not play role in this change, since the two kids still need to share the bedroom when there are only two bedrooms in total. Therefore, sex composition/extra room is a valid instrument for studying the effect of project participation on overcrowding.

However, the SIPP sample is too small to be reliable for using instrumental variables methods, so Currie & Yelowitz used the two-sample two-stage least squares (TS2SLS) approach on two different data sets: 1990 to 1995 waves of the March Current Population Survey (CPS) data with sample size $n_0 = 21718$ and 1990 Census data with sample size $n_1 = 279129$.

The CPS data is the *auxiliary data* which contains the indicator variable x of housing project participation, the instrumental variable z_0 equal to one if the family has a boy and a girl and equal to zero if they have two boys or two girls, as well as the exogenous explanatory variables z_0^c which relate to overcrowding, such as household head's gender, age, race, education, marital status and the number of boys in the family, etc (all listed in Figure 4.3). As the *primary data*, the Census data is composed of the outcome y indicating overcrowding and variables (z_0, z_0^c) defined exactly the same as CPS data.

Figure 4.3 shows the error bar (one standard error) of sample means of all the common variables (z_0, z_0^c) among two samples, including the instrumental variable z_0 "extra" and other exogenous variables z_0^c. The figures in the first two rows show all the binary variables and one categorical variables ("boys") in two samples denoted by (1) and (0) separately. The binary variables contain the information of household head's

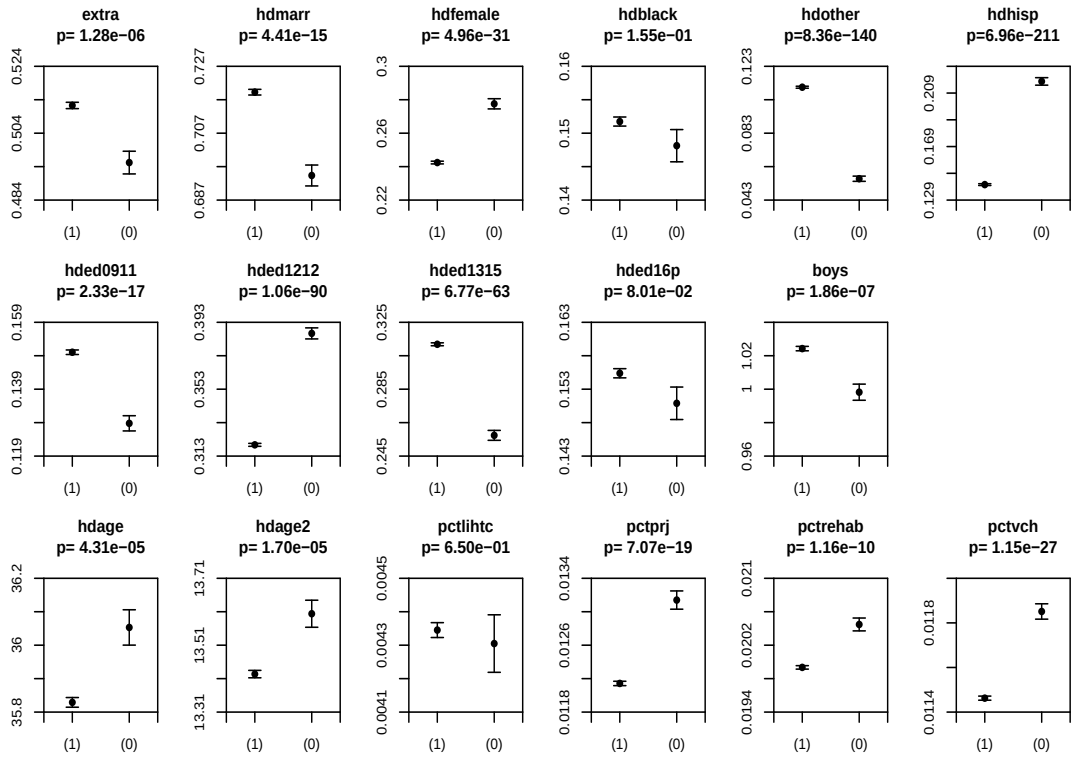

Figure 4.3: Distribution of sample average of all the common variables between two samples

The figures in the first two rows show the error bars (one standard error) of sample averages of binary variables and categorical variable (boys) in two samples. The third line shows the error bars (one standard error) of sample means of continuous variables across two samples. The p-values in the titles are p-values of two-sample proportion test for binary variables, and are p-values of two-sample t-test for continuous variables (including boys). In the bottom of each frame, label "(1)" denotes the 1990 Census data (primary data), and label "(0)" represents the 1990-1995 waves of March Current Population Survey (CPS) data (auxiliary data). "extra" is the instrumental variable z_0 indicating whether the family is entitled to an extra room due to the sex composition of the children in the household.

marriage status, gender, race, education ("hdmarr","hdfemale","hdblack", etc.). In the third row, we show the plots of all the continuous variables. The continuous variables include "age of household head"("hdage") and its squared value, "the percentage of households in projects or other subsidized housing"("pctprj") and so on. For binary variables, we compute the p-value of two-sample proportion test to test whether two samples have equal means. And for continuous variables (including boys), we compute the p-value of two-sample t-test to test equal means. All the p-values are listed in the title of Figure 4.3. Except "hdblack", "hded16p" and "pctlihtc", all the variables have significantly different means across two samples with $\alpha = 0.05$. Therefore, the two samples are actually drawn from different populations. According to this feature, we know TSIV estimator proposed by Angrist & Krueger (1992) will be biased based on

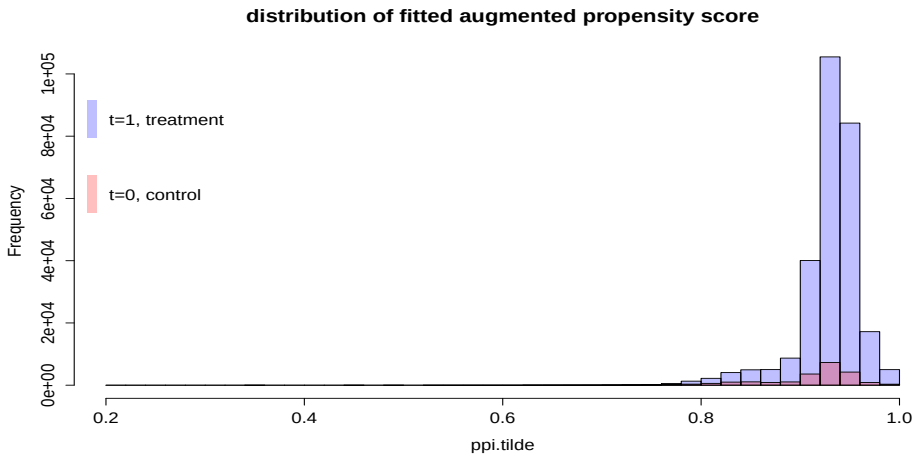

Figure 4.4: Distribution of fitted augmented propensity score

Table 4.4: Estimates of public housing project's influence on family's overcrowdedness

TSIV	Currie & Yelowitz	TS2SLS	OR	IPW	AIPW	LIK
-0.2345	-0.1595	-0.1595	-0.1595	-0.2640	-0.2574	-0.1865
(0.2487)	(0.0624)	(0.1014)	(0.1014)	(0.8679)	(0.9880)	(0.1895)

This table lists the different estimate of the causal effect of public housing project participation to family's overcrowdedness. Each cell gives the point estimate of the effect (upper) and standard error (lower) of the point estimator. The standard error is obtained based on 200 bootstrap samples.

the discussion in the simulation example.

Using the same data as Currie & Yelowitz (2000) for analysis on overcrowdedness, we estimate the effect of project participation using the existing methods and our proposed approach. For our likelihood estimator, we got an estimate of OR function $\hat{m}_0(z)$ using identity link with regressors $g_0(z) = (z_0, z_0^{c\mathrm{T}})^{\mathrm{T}}$, and we also estimate a augmented logistic PS model with regressors $f(z) = (z_0, z_0^{c\mathrm{T}})^{\mathrm{T}}$ and the augmented part $\hat{m}_0(z)z$. Figure 4.4 shows the distribution of estimated $\tilde{\pi}(z)$ across the two samples.

Table 4.4 lists the estimates using different approaches defined exactly the same as the ones in the simulation studies. In order to compare with the results obtained in Currie & Yelowitz (2000), we also include the TS2SLS estimation but with the standard errors obtained from bootstrap using 200 bootstrap samples.

We obtain exactly the same TS2SLS point estimator as Currie & Yelowitz (2000), -0.1595, showing the households in public housing projects are less likely to be overcrowded. But our bootstrap standard error is relatively larger than the result of Currie

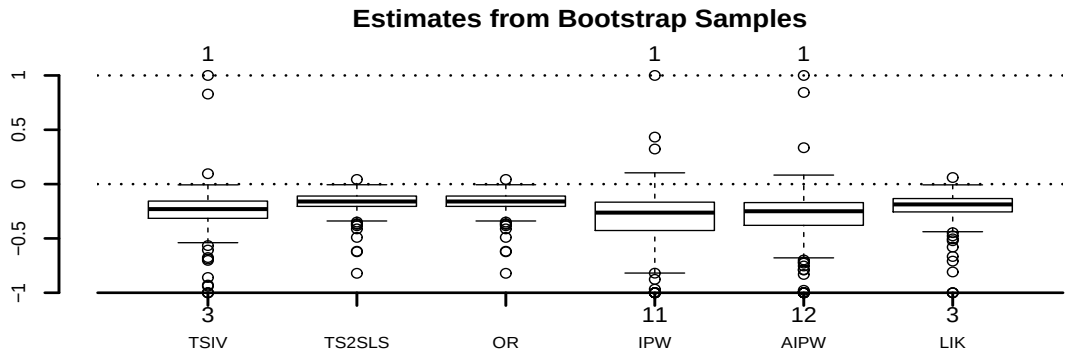

Figure 4.5: Estimates of 200 Bootstrap Samples

& Yelowitz (2000), where they claimed the standard error had been corrected to account for the fact that a predicted value of project participation is used in the OLS in the second stage. Our bootstrap standard error of TS2SLS estimate makes the housing project's functioning in reducing family's overcrowdedness slightly below 90% level of confidence. And because of the same reasons stated in Section 4.3, OR estimate leads to the same results as TS2SLS.

The TSIV, IPW and AIPW estimates all give quite large standard errors, which make the point estimates lack of stability and confidence. Also being doubly robust and locally efficient like the AIPW estimator, our LIK estimation generates much smaller standard error than AIPW estimator. The point estimate of LIK is close to TS2SLS, but is associated with a larger standard error such that the effect of housing project is not statistically significant at 90% level. Currie & Yelowitz (2000) conclude that the households in projects are less likely to suffer from overcrowded. Here the results from our improved estimator (LIK) show that there exist less strong evidence than in Currie & Yelowitz (2000), that public housing project could alleviate the overcrowdedness of household.

4.6 Conclusion

Collecting and combining samples from different sources is a very frequent and conventional approach used by economists. However, when we combine these samples, blind assumption of same distribution across two samples are very likely to generate estimates

lack of accuracy and efficiency. As a classical way to balance the discrepancy between two samples, the weighted estimator based on various kinds of fitted propensity score have huge variance.

In this chapter, we study estimation for moment restriction models with auxiliary data and then a two-sample combination problem. By directly utilizing the efficient influence function from semiparametric efficiency theory, we derive the AIPW estimator that is doubly robust and locally efficient. Beyond that, we also propose regression estimator and calibrated likelihood estimators that achieve greater efficiency than local efficiency when the propensity score is correctly specified and also keep the double robustness. Moreover, we apply this general framework to the specific two-sample linear instrumental variable problem, and develop the estimators correspondingly. The simulation study and the reanalysis of public housing project's outcome, demonstrate better performance of our proposed methods when compared to existing estimators.

4.7 Appendix

We provide the following lemma on asymptotic expansions of AIPW estimators.

Lemma 4.5 *Assume that* $E\{q_0^2(z)\} < \infty$. *If the PS model (4.8) is correctly specified, then the following asymptotic expansion holds.*

$$\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\hat{\pi}(z)[\Phi(x,z;\theta)-q_0(z)]\right\} = \tilde{E}\left\{\frac{1-t}{1-\pi(z)}\pi(z)[\Phi(x,z;\theta)-q_0(z)]\right\} + o_p(n^{-1/2})$$

Proof of Lemma 4.5.

$$\begin{aligned}
&\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\hat{\pi}(z)[\Phi(x,z;\theta)-q_0(z)]\right\} \\
=&\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\pi(z)[\Phi(x,z;\theta)-q_0(z)]\right\} + \tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\big[\hat{\pi}(z)-\pi(z)\big]\big[\Phi(x,z;\theta)-q_0(z)\big]\right\}
\end{aligned}$$

When PS model (4.8) is correctly specified, the first term could be represented as

$$\begin{aligned}
&\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\pi(z)[\Phi(x,z;\theta)-q_0(z)]\right\}\\
=&\tilde{E}\left\{\frac{1-t}{1-\pi(z)}\pi(z)[\Phi(x,z;\theta)-q_0(z)]\right\}\\
&+E\left\{\frac{1-t}{[1-\pi(z)]^2}\frac{\partial\pi(z;\gamma)}{\partial\gamma}\pi(z)\big[\Phi(x,z;\theta)-q_0(z)\big]\right\}(\hat{\gamma}-\gamma)+o_p(n^{-1/2})\\
=&\tilde{E}\left\{\frac{1-t}{1-\pi(z)}\pi(z)[\Phi(x,z;\theta)-q_0(z)]\right\}+o_p(n^{-1/2})
\end{aligned}$$

The second term above disappear because the expectation equals to 0 due to $E\{\Phi(x,z;\theta)|z\}=q_0(z)$. Similarly, the expectation term below also equals to 0

$$\begin{aligned}
&\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}(z)}\big[\hat{\pi}(z)-\pi(z)\big]\big[\Phi(x,z;\theta)-q_0(z)\big]\right\}\\
=&E\left\{\frac{1-t}{1-\pi(z)}\frac{\partial\pi(z;\gamma)}{\partial\gamma}\big[\Phi(x,z;\theta)-q_0(z)\big]\right\}(\hat{\gamma}-\gamma)+o_p(n^{-1/2})\\
=&o_p(n^{-1/2})
\end{aligned}$$

By combining these two terms, the asymptotic expansion in Lemma 4.5 holds. □

Proofs of Propositions 4.2

First, let's prove the local nonparametric efficiency of $\hat{\theta}_{\mathrm{NP}}$.

For convenience, we write $\hat{\pi}=\hat{\pi}(z)$, $\Phi(\theta)=\Phi(x,z;\theta)$, $q_0=q_0(z)$. Suppose $\hat{\theta}_{\mathrm{NP}}$ converges to θ^* such that $\hat{\theta}_{\mathrm{NP}}-\theta^*=O_p(n^{-1/2})$. If OR model (4.5) is correctly specified, and also by Slutsky Theorem, the estimating Eq. (4.15) is asymptotically equal to

$$\frac{1}{p}\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}}\hat{\pi}\Phi(\hat{\theta}_{\mathrm{NP}})\right\}=\frac{1}{p}\tilde{E}\left\{\big[\frac{1-t}{1-\hat{\pi}}\hat{\pi}-t\big]q_0\right\}+o_p(n^{-1/2})$$

Furthermore, if PS model (4.8) is also correctly specified, the estimating equation (4.15) is unbiased estimating equation, then $\theta^*=\theta_0$, the true parameter, and the we can further expand the equation as

$$\frac{1}{p}E\left\{\frac{1-t}{1-\hat{\pi}}\hat{\pi}\frac{\partial\Phi(\theta_0)}{\partial\theta^{\mathrm{T}}}\right\}(\hat{\theta}_{\mathrm{NP}}-\theta_0)=-\frac{1}{p}\left\{\tilde{E}\Big[\frac{1-t}{1-\hat{\pi}}\hat{\pi}\big[\Phi(\theta_0)-q_0\big]\Big]+\tilde{E}(tq_0)\right\}+o_p(n^{-1/2})$$

And we know when PS model (4.8) is correctly specified, $E\left\{\frac{1-t}{1-\hat{\pi}}\hat{\pi}\frac{\partial\Phi(\theta_0)}{\partial\theta^{\mathrm{T}}}\right\}$ converges to $E\left\{\frac{1-t}{1-\pi}\pi\frac{\partial\Phi(\theta_0)}{\partial\theta^{\mathrm{T}}}\right\}$ in probability, so based on Slutsky Theorem and Lemma 4.5, we have

$$\hat{\theta}_{\mathrm{NP}}-\theta_0=-\left\{\frac{\partial}{\partial\theta^{\mathrm{T}}}E\big[\Phi(\theta_0)|t=1\big]\right\}^{-1}\left\{\frac{1}{p}\tilde{E}\Big\{\frac{1-t}{1-\pi}\pi\big[\Phi(\theta_0)-q_0\big]\Big\}+\frac{1}{p}\tilde{E}(tq_0)\right\}+o_p(n^{-1/2})$$

Therefore, $\hat{\theta}_{\mathrm{NP}}$ achieves the nonparametric variance bound V_{NP} when both PS model and OR model are correctly specified.

Second, we show the double robustness of $\hat{\theta}_{\mathrm{NP}}$. If PS model (4.8) is correctly specified,

$$\tilde{E}\left[\big(t-\frac{1-t}{1-\hat{\pi}}\hat{\pi}\big)\hat{q}_0\right]=\tilde{E}\left[\big(t-\frac{1-t}{1-\pi}\pi\big)q_0^*\right]+O_p(n^{-1/2})=O_p(n^{-1/2})$$

so the estimating equation (4.15) could be represented as

$$\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}}\hat{\pi}\Phi(\theta)\right\}\Big/\tilde{E}(t)+O_p(n^{-1/2})=0$$

which makes $\hat{\theta}_{\mathrm{NP}}$ consistent as $\hat{\theta}_{\mathrm{IPW}}$.

If OR model (4.5) is correctly specified,

$$\tilde{E}\left\{\frac{1-t}{1-\hat{\pi}}\hat{\pi}\big[\Phi(\theta)-\hat{q}_0\big]\right\}=\tilde{E}\left\{\frac{1-t}{1-\pi^*}\pi^*\big[\Phi(\theta)-q_0\big]\right\}+O_p(n^{-1/2})=O_p(n^{-1/2})$$

so the estimating equation (4.15) could be represented as

$$\tilde{E}(t\hat{q}_0)\big/\tilde{E}(t)+O_p(n^{-1/2})=0$$

which makes $\hat{\theta}_{\mathrm{NP}}$ consistent as $\hat{\theta}_{\mathrm{OR}}$.

Third, let's prove the local semiparametric efficiency of $\hat{\theta}_{\mathrm{SP}}$. Similarly, based on correct specified OR model (4.5) and PS model (4.8), and using Slutsky Theorem,

estimating equation (4.16) could be asymptotically written as

$$\hat{\theta}_{\mathrm{SP}} - \theta_0 = -\left\{ \frac{\partial}{\partial \theta^{\mathrm{T}}} E\big[\Phi(\theta_0)|t=1\big] \right\}^{-1} \left\{ \frac{1}{p}\tilde{E}\Big\{ \frac{1-t}{1-\pi}\pi\big[\Phi(\theta_0) - q_0\big] \Big\} + \frac{1}{p}\tilde{E}(\hat{\pi} q_0) \right\} + o_p(n^{-1/2})$$

It is important to notice that

$$\begin{aligned} \tilde{E}(\hat{\pi} q_0) =& \tilde{E}(\pi q_0) + E\left\{ \frac{\partial \pi(\gamma)}{\partial \gamma^{\mathrm{T}}} q_0 \right\} (\hat{\gamma} - \gamma) + o_p(n^{-1/2}) \\ =& \tilde{E}(\pi q_0) + \Pi\Big[(t-\pi) q_0 | S_\gamma \Big] + o_p(n^{-1/2}) \end{aligned}$$

Based on these two facts, $\hat{\theta}_{\mathrm{SP}}$ achieves the semiparametric variance bound V_{SP} when both PS model and OR model are correctly specified. □

Proof of Proposition 4.3

First, it is straightforward that $\tilde{\beta}(\theta) = \beta^*(\theta) + o_p(1)$, where $\beta^*(\theta) = E^{-1}[\xi^* \zeta^{*\mathrm{T}}] E[\xi^* \tau^*_{\mathrm{init}}(\theta)]$ and $\tau^*_{\mathrm{init}}(\theta)$, ξ^*, ζ^*, and $h^*(z)$ are defined as $\tilde{\tau}_{\mathrm{init}}(\theta)$, $\tilde{\xi}$, $\tilde{\zeta}$, and $\tilde{h}(z)$ respectively but with $\pi^\dagger(z)$ and $q_0^*(z)$ in place of $\tilde{\pi}(z)$ and $\hat{q}_0(z)$ throughout.

Secondly, let's prove the local nonparametric efficiency and double robustness of $\tilde{\theta}_{\mathrm{reg}}$. These two properties hold automatically if we could prove the asymptotic equivalence in first order to Eq. (4.22). By design, $\tilde{\pi}(z)\hat{q}_0(z)$ is a linear combination of variables in $\tilde{h}(z)/\tilde{\pi}(z)$, i.e., $\tilde{\pi}(z)\hat{q}_0(z) = c_0^{\mathrm{T}} \tilde{h}(z)/\tilde{\pi}(z)$ for some constant c_0. Then of course we have $\pi^\dagger(z) q_0^*(z) = c_0^{\mathrm{T}} h^*(z)/\pi^\dagger(z)$. If OR model (4.5) is correctly specified, $q_0^*(z) = q_0(z)$, and then $\pi^\dagger(z) q_0(z) = c_0^{\mathrm{T}} h^*(z)/\pi^\dagger(z)$. Therefore, we have

$$\beta^* = E^{-1}\left\{ \xi^* \frac{1-t}{1-\pi^\dagger(z)} \frac{h^{*\mathrm{T}}(z)}{\pi^\dagger(z)} \right\} E\left\{ \xi^* \frac{1-t}{1-\pi^\dagger(z)} \pi^\dagger(z) q_0(z) \right\} = c_0$$

Hence, the estimating equation (4.20) is asymptotically equivalent, to the first order, to (4.22), which is equivalent to Eq. (3.1), due to the application of augmented PS model. The solution $\tilde{\theta}_{\mathrm{NP}}$ to Eq. (3.1) could be proved to be doubly robust and locally nonparametric efficient similar to the proof of $\hat{\theta}_{\mathrm{NP}}$ in previous section. So $\tilde{\theta}_{\mathrm{reg}}$ is consistent

when OR model (4.5) is correctly specified, and it is locally nonparametric efficient. On the other hand, if PS model (4.8) is correctly specified, $\tilde{E}(\tilde{\xi})$ could be regarded as $O_p(n^{-1/2})$, so $\tilde{\theta}_{\text{reg}}$ is also consistent when only PS model (4.8) is correct. That is, $\tilde{\theta}_{\text{reg}}$ is doubly robust.

Third, let's show the intrinsic efficiency of $\tilde{\theta}_{\text{reg}}$ among the class of estimators which are the solution to (4.21), and let's denote them as $\tilde{\theta}(b)$. And assume $\hat{\alpha}_0$ converges to α_0^*, $(\tilde{\gamma}, \tilde{\delta})$ converges to $(\gamma^\dagger, \delta^*)$ and $\tilde{\theta}(b)$ converges to θ^* for certain value of b. Similar to the expansion in the proof of Proposition 4.2, by direct calculation and Slutsky theorem, when PS model is correctly specified, actually $\tilde{\theta}(b)$ is an unbiased estimate, so $\theta^* = \theta_0$, the true value, and $(\gamma^\dagger, \delta^*) = (\gamma^*, 0)$, $\pi^\dagger = \pi$, and we have

$$\tilde{\theta}(b) - \theta_0 = -\left\{\frac{\partial}{\partial \theta^{\mathrm{T}}} E\big[\Phi(\theta_0)|t=1\big]\right\}^{-1} \left\{\frac{1}{p}\tilde{E}\Big[\frac{1-t}{1-\tilde{\pi}}\tilde{\pi}\Phi(\theta_0) - b^{\mathrm{T}}\big(\frac{1-t}{1-\tilde{\pi}} - 1\big)\frac{h^*}{\pi}\Big]\right\} + o_p(n^{-1/2})$$

If PS model is correctly specified, we could expand the function above further with applying Lemma 4.5 in Chapter 3 by replacing $\hat{\pi}$ with $\tilde{\pi}$ and Y with $\Phi(\theta_0)$ and $h = b^{\mathrm{T}}h^*\pi$, then we can get

$$\begin{aligned}\tilde{\theta}(b) - \theta_0 = &-\left\{\frac{\partial}{\partial \theta^{\mathrm{T}}} E\big[\Phi(\theta_0)|t=1\big]\right\}^{-1} \\ &\left\{\frac{1}{p}\tilde{E}\Big[\tau^*_{\text{init}}(\theta_0) - b^{\mathrm{T}}\xi^* - \Pi\big\{\tau^*_{\text{init}}(\theta_0) - b^{\mathrm{T}}\xi^*|S^\dagger_{(\gamma^*,0)}\big\} + \Pi\big\{(t-\pi)q_0|S^\dagger_{(\gamma^*,0)}\big\}\Big]\right\} + o_p(n^{-1/2})\end{aligned}$$

where $\tau^*_{\text{init}}(\theta_0) = \frac{1-t}{1-\pi}\pi\Phi(\theta_0)$, $\xi^* = (\frac{1-t}{1-\pi} - 1)\frac{h^*}{\pi}$. $\{\frac{\partial}{\partial \theta^{\mathrm{T}}} E\big[\Phi(\theta_0)|t=1\big]\}^{-1}$ is a constant, and inside $\tilde{E}()$, $\tau^*_{\text{init}}(\theta_0) - b^{\mathrm{T}}\xi^* - \Pi\big\{\tau^*_{\text{init}}(\theta_0) - b^{\mathrm{T}}\xi^*|S^\dagger_{(\gamma^*,0)}\big\}$ is uncorrelated with the remaining term $\Pi\big\{(t-\pi)q_0|S^\dagger_{(\gamma^*,0)}\big\}$, which is independent of b. Moreover, the former one can be expressed as $\tau^*_{\text{init}}(\theta_0) - a^{\mathrm{T}}\xi^*$ for some constant vector a, since each variable in $S^\dagger_{(\gamma^*,0)}$ is a linear combination of variables in ξ^* by construction. By combining these two facts, the asymptotic variance of $\tilde{\theta}(b)$ is minimized when a is equal to

$$\text{var}^{-1}(\xi^*)\text{cov}\left\{\xi^*, \tau^*_{\text{init}}(\theta_0)\right\} = E^{-1}(\xi^*\zeta^{*\mathrm{T}})E(\xi^*\tau^*_{\text{init}}(\theta_0)) = \beta^*.$$

But to make a equal to β^*, it suffices to set $b = \beta^*$, because $\tau^*_{\text{init}}(\theta_0) - \beta^{*\mathrm{T}}\xi^*$ is uncorrelated with $S^\dagger_{(\gamma^*,0)}$ and hence $\Pi\{\tau^*_{\text{init}}(\theta_0) - \beta^{*\mathrm{T}}\xi^*|S^\dagger_{(\gamma^*,0)}\} = 0$. If PS model (4.8) is

correctly specified, then $\tilde{\theta}_{\text{reg}} = \tilde{\theta}(\beta^*) + o_p(n^{-1/2})$. Therefore, $\tilde{\theta}_{\text{reg}}$ is intrinsically efficient among the class of estimators $\tilde{\theta}(b)$. □

Derivation of empirical likelihood estimates

Similar to the derivation of empirical likelihood estimates of Appendix I in ATT paper(Shu 2015), here we also have $\ell_{\text{EL}}(\lambda)$ could be reexpressed in another form based on $\xi_i = \frac{t_i - \tilde{\pi}_i}{\tilde{\pi}_i(1-\tilde{\pi}_i)}\tilde{h}_i$, Write $\tilde{\pi}_i = \tilde{\pi}(z_i)$, $\tilde{h}_i = \tilde{h}(z_i)$, and $\omega_i = \omega(z_i;\lambda) = \tilde{\pi}_i + \hat{\lambda}^{\mathrm{T}}\tilde{h}_i$ for $i = 1, \ldots, n$, then

$$\begin{aligned}
\ell_{\text{EL}}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n} \log\left\{1 + \lambda^{\mathrm{T}}\frac{t_i - \tilde{\pi}_i}{\tilde{\pi}_i(1-\tilde{\pi}_i)}\tilde{h}_i\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{t_i \log\left(1 + \lambda^{\mathrm{T}}\frac{\tilde{h}_i}{\tilde{\pi}_i}\right) + (1-t_i)\log\left(1 - \lambda^{\mathrm{T}}\frac{\tilde{h}_i}{1-\tilde{\pi}_i}\right)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{t_i \log\omega_i + (1-t_i)\log(1-\omega_i)\right\} - \frac{1}{n}\sum_{i=1}^{n}\left\{t_i\log\tilde{\pi}_i + (1-t_i)\log(1-\tilde{\pi}_i)\right\},
\end{aligned}$$

That is $l(\lambda)$ plus a constant term. Therefore, $\hat{\lambda}$ is a maximizer of $l(\lambda) = \tilde{E}[t\log\omega(z;\lambda) + (1-t)\log\{1-\omega(z;\lambda)\}]$.

And the left-handside of estimating equation (4.27) could be represented in a similar form when we write $\Phi_i(\theta) = \Phi(x_i, z_i; \theta)$:

$$\begin{aligned}
\sum_{i=1}^{n}\hat{p}_i\left[\frac{1-t_i}{1-\tilde{\pi}_i}\tilde{\pi}_i\Phi_i(\theta)\right] &= \frac{1}{n}\sum_{i=1}^{n}\frac{1-t_i}{1+\hat{\lambda}^{\mathrm{T}}\tilde{\xi}_i}\frac{\tilde{\pi}_i}{1-\tilde{\pi}_i}\Phi_i(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{1-t_i}{1-\hat{\lambda}^{\mathrm{T}}\frac{\tilde{h}_i}{1-\tilde{\pi}_i}}\frac{\tilde{\pi}_i}{1-\tilde{\pi}_i}\Phi_i(\theta) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{(1-t_i)}{1-\hat{\omega}_i}\tilde{\pi}_i\Phi_i(\theta)
\end{aligned}$$

where $\hat{\omega}_i = \omega(z_i;\hat{\lambda})$ for $i = 1, \ldots, n$, and so the final estimating equation is Eq. (4.28). □

Proof of Proposition 4.4

We need only to show that if PS model (4.8) is correctly specified, then $\tilde{\theta}_{\text{lik}}$ is asymptotically equivalent, to the first order, to $\tilde{\theta}_{\text{reg}}$. We can prove it if the estimating equation (4.31) is asymptotically equivalent to the first order, to Eq. (4.20).

Similar to the asymptotic expansion of the calibrated likelihood estimator in Tan (2010), if PS model (4.8) is correctly specified, the left-handside of Eq. (4.31) could be written as

$$\tilde{E}\left[\frac{(1-t)\tilde{\pi}(z)\Phi(x,z;\theta)}{1-\omega(z;\tilde{\lambda})}\right] = \tilde{E}\left[\frac{(1-t)\tilde{\pi}(z)\Phi(x,z;\theta)}{1-\omega(z;\hat{\lambda})}\right] + o_p(n^{-1/2}),$$

by Taylor expansion for $\tilde{\lambda}$ about $\hat{\lambda}$ and the fact that $\tilde{E}([(1-t)/\{1-\omega(z;\hat{\lambda})\}-1]\tilde{\pi}(z)) = o_p(n^{-1/2})$.

Moreover, just like the asymptotic expansion of the non-calibrated likelihood estimator in Tan (2010), we could prove if PS model (4.8) is correctly specified, then $\hat{\lambda}$ converges to 0 in probability and

$$\tilde{E}\left[\frac{(1-t)\tilde{\pi}(z)\Phi(x,z;\theta)}{1-\omega(z;\hat{\lambda})}\right] = \tilde{E}\left(\tilde{\tau}_{\text{init}}(\theta) - \beta^{*\text{T}}(\theta)\tilde{\xi}\right) + o_p(n^{-1/2}).$$

by a Taylor expansion for $\hat{\lambda}$ about 0. Based on all the expansions above, $\tilde{\theta}_{\text{lik}}$ attains the desired properties. □

Bibliography

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, 72:1–19.

Angrist, J. D. and Krueger, A. B. (1992). The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87:328–336.

Bjorklund, A. and Jantti, M. (1997). Intergenerational income mobility in sweden compared to the united states. *American Economic Review*, 87:1009–1018.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96:723–734.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:573–585.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36:808–843.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, 3 edition.

Currie, J. and Yelowitz, A. (2000). Are public housing projects good for kids? *Journal of Public Economics*, 75:99–124.

Dehejia, R. (2005a). Practical propensity score matching: A reply to Smith and Todd. *The Review of Economics and Statistics*, 125:355–364.

Dehejia, R. (2005b). Does matching overcome lalondes critique of nonexperimental estimators? A postscript. Manuscript.

Dehejia, R. and Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94:1053–1062.

Dehejia, R. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84:151–161.

Deville, J.-C. and Sarndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics*, 95:932–945.

Graham, B. S., de Xavier Pinto, C. C., and Egel, D. (2015). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *Journal of Business and Economic Statistics*. to appear.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20:25–46.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*. Methuen.

Heckman, J. J., LaLonde, R. J., and Smith, J. A. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies*, 64:605–654.

Heckman, J. J., LaLonde, R. J., and Smith, J. A. (1999). The economics and econometrics of active labor market programs. *Handbook of Labor Economics 3A(O. Ashenfelter and D. Card, eds.)*, pages 1865–2097.

Heckman, J. J. and Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In Heckman, J. J. and Singer, B., editors, *Longitudinal Analysis of Labor Market Data*, pages 156–246. Cambridge University Press, New York.

Hellerstein, J. K. and Imbens, G. (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Econometric and Statistics*, 81:1–14.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71:1161–1189.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society*, 76:243–263.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86:4–29.

Inoue, A. and Solon, G. (2010). Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92:557–561.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussions). *Statistical Science*, 22:523–539.

Klevmarken, W. A. (1982). Missing variables and two-stage least-squares estimation from more than one data-set. *Business and Economic Statistics Section*, pages 156–161.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76:604–620.

Little, R. J. A. and Rubin, D. B. (2002). Statistical analysis with missing data. *2nd ed. Wiley*, Hoboken:NJ. MR1925014.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9:403–425.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22:540–543.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5:99–135.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–472.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85:619–630.

Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *The Annals of Statistics*, 27:1368–1384.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22:300–325.

Qin, J. and Zhang, B. (2008). Empirical-likelihood-based difference-in-differences estimators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:329–349.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese*, 121:151–179.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16:285–319.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129.

Robins, J. M. and Rotnitzky, A. (2001). Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer'. *Statistica Sinica*, 11:920–936.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.

Robins, J. M., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science*, 22:544C559.

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

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524.

Rothe, C. and Firpo, S. (2013). Semiparametric estimation and inference using doubly robust moment conditions. IZA Discussion Papers 7564, Institute for the Study of Labor (IZA).

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Statistics*, 66:688–701.

Rubin, D. B. and van der Laan, M. J. (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *International Journal of Biostatistics*, 4:1557–4679.

Smith, J. and Todd, P. (2005a). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125:305–353.

Smith, J. and Todd, P. (2005b). Rejoinder. *Journal of Econometrics*, 125:365–375.

Tan, Z. (2006). A distributonal approach for causal inference using propensity score. *Journal of the American Statistical Association*, 101:1619–1637.

Tan, Z. (2008). Improved local efficiency and double robustness, comment on 'empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis' by Rubin and van der Laan. *International Journal of Biostatistics*, 4:Article 10.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97:661–682.

Tan, Z. (2013). Simple design-efficient calibration estimators for rejective and high-entropy sampling. *Biometrika*, 100:399–415.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.

Wooldridge, J. M. (2002). Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1:117–139.

Zhao, Q. and Percival, D. (2015). Primal-dual covariate balance and minimal double robustness via entropy balancing. Manuscript.