A review of parametric doubly-robust estimators using missing-at-random data*

Saraswata Chaudhuri†, Hye-Young Min‡and Jean-Louis Barnwell§

Date: February 11, 2019.

Abstract

In many applications, estimation of the parameters of interest also requires modeling and estimating nuisance parameters. Misspecification of nuisance parameters in general leads to incorrect estimation of the parameters of interest. There are, however, important cases such as estimation using missing-at-random data where doubly-robust estimators exit. These estimators are based on estimating functions involving two types of nuisance parameters where misspecification of any one (but not both) of which does not lead to incorrect estimation of the parameters of interest. That is, these estimators are robust to two different types of misspecifications as long as those do not happen simultaneously. However, possibly due to the technical nature of the related literature, these estimators have not enjoyed widespread use in economics. Our goal is pedagogical, and we seek to provide a non-technical overview of these estimators to facilitate their use in economics.

JEL Classification: C12; C13; C30

Keywords: Doubly-robust estimator; Missing at random data; Neyman C-alpha

*This paper is written for the Festschrift in honour of Professor Nityananda Sarkar of the Indian Statistical Institute (ISI), Kolkata. The first author was a student at ISI Kolkata. We draw on part of an earlier (2012) version: "Doubly-Robust Parametric Estimation in Moment Conditions Models with Missing-At-Random Data" authored by S. Chaudhuri and and H-Y. Min. The third author, J-L. Barnwell, makes important contribution to the current version and fills the role of H-Y. Min who has long moved outside academia after her M.A. in Economics. The earlier version of the paper benefitted from comments of M. Caner, D. Frazier, D. Guilkey, J. Hill, M. Kejriwal, S. Park, K. Peter, B. McManus, E. Renault, E. Rose, B. Tsang, M. Wiswall and the seminar participants at Purdue and Virginia Tech.

†Department of Economics, McGill University and Cireq, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

‡Former graduate student in economics, UNC Chapel Hill.

§Department of Economics, McGill University, Canada. Email: jean-louis.barnwellmenard@mail.mcgill.ca.

1 Introduction

Doubly-robust estimators were introduced by Robins et al. (1994), Robins et al. (1995), Robins and Rotnitzky (1995), Holcroft et al. (1997), etc.; formalized in Scharfstein et al. (1999); and subsequently studied and used widely in biostatistics, epidemiology and statistics. More recently, the econometrics literature has also developed an active interest in these estimators; see, e.g., Hirano and Imbens (2001), Wooldridge (2007), Graham et al. (2012), Busso et al. (2009, 2011), Chaudhuri and Guilkey (2016), Rothe and Firpo (2016), Graham et al. (2016), Sloczynski and Wooldridge (2018), etc.

The innovations in these papers are technical in nature, and the discussion therein are specific to the scenarios of their respective concern. Unfortunately, expository articles aimed at practitioners seem missing; and perhaps as a consequence of this, the application of doubly-robust estimators is rarely found in economics, although well-known contexts — such as missing data (due to attrition, non-response, etc.), measurement error, program evaluation, etc. — for applications abound.

Our paper attempts to fill this gap by: (1) providing a non-technical overview of doubly-robust estimators, their implementation and asymptotic properties, and (2) demonstrating their nice properties even in relatively small samples using a simulation study. Our paper is not intended as a survey, nor does it provide strictly original results. Rather, its purpose is to assist the pedagogy of doubly-robust estimators by relating them to the standard pedagogy of two-step/plug-in estimators.

Our paper proceeds as follows. Section 2 defines the generic doubly-robust estimator. Section 3 defines the estimation setup adopting the broad setup of Chen et al. (2008) that covers scenarios of missing data (due to attrition, non-response, etc.), measurement error, program evaluation, etc. as special cases. Section 4 describes a basic doubly-robust estimator for estimation under this setup, with a focus on emphasizing the features that make a doubly-robust estimator special among the class of two-step/plug-in estimators that are more familiar to economists. Section 5 is a simulation study comparing doubly-robust and non-robust estimators in an instrumental variables (IV) regression with missing IV, which has received attention recently; see, e.g., Mogstad and Wiswall (2012), Chaudhuri and Guilkey (2016), Abrevaya and Donald (2017), Kennedy and Small (2018). Section 6 concludes.

2 Doubly-robust estimator: A general overview

Doubly-robust estimators for any parameter of interest θ are obtained by solving estimating equations with doubly-robust population analogs. To fix ideas, consider the observed data $O_1, \ldots, O_n$ that are

independent and identically distributed (i.i.d.) copies of the random variable $O \in \mathbb{R}^{d_o}$ which satisfies:

$$0 \;=\; E\left[\psi(O;\theta^0, h_1^0(O), h_2^0(O))\right] \text{ if and only if } \theta = \theta^0 \tag{1}$$

where θ^0 represents the true value of $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$. $h_1^0(O)$ and $h_2^0(O)$ are $d_1 \times 1$ and $d_2 \times 1$ dimensional unknown functions of O, and they represent the "true" nuisance parameters in our paper.

To focus on the main idea, we will abstract from over-identified models and consider $\psi(.)$ as a $d_\theta \times 1$ known function of its arguments specified in (1). Section 4 provide a number of common examples of the moment function $\psi(.)$ and the associated nuisance parameters $h_1^0(O)$ and $h_2^0(O)$.

(1) represents the population analog of the d_θ estimating equations for θ, and it is referred to as "doubly-robust" to the misspecification of the (true) nuisance parameters $h_1^0(O)$ and $h_2^0(O)$ if:

$$0 \;=\; E\left[\psi(O;\theta^0, h_1^*(O), h_2^0(O))\right] \tag{2}$$

$$0 \;=\; E\left[\psi(O;\theta^0, h_1^0(O), h_2^*(O))\right] \tag{3}$$

where $h_1^*(O)$ and $h_2^*(O)$ are any $d_1 \times 1$ and $d_2 \times 1$ dimensional functions of O. In words, double-robustness of the estimating function $\psi(.)$ means that the unbiasedness of the function $\psi(.)$ for 0 holds at θ^0 if either $h_1^0(O)$ or $h_2^0(O)$, but not necessarily both, is replaced by conformable functions.

Therefore, one would naturally construct an estimator $\widehat{\theta}_n$ for θ as the solution of:

$$o_p\left(\frac{1}{\sqrt{n}}\right) \;=\; \frac{1}{n}\sum_{i=1}^{n} \psi(O_i, \theta, \widehat{h}_{1,n}(O_i), \widehat{h}_{2,n}(O_i)) \tag{4}$$

where the $o_p(1/\sqrt{n})$ term in (4) is a useful generalization to accommodate for approximate solutions, while $\widehat{h}_{1,n}(O)$ and $\widehat{h}_{2,n}(O)$ are intended to be consistent estimators of $h_1^0(O)$ and $h_2^0(O)$ respectively.

In practice, however, one postulates parametric models $h_1^*(O;\beta)$ and $h_2^*(O;\gamma)$ with $\beta^* \in \mathbb{R}^{d_\beta}$ and $\gamma^* \in \mathbb{R}^{d_\gamma}$ implicitly defined such that $h_1^*(O) \equiv h_1^*(O;\beta^*)$ and $h_2^*(O) \equiv h_2^*(O;\gamma^*)$. Ideally, $h_1^*(O;\beta^*) = h_1^0(O)$ and $h_2^*(O;\gamma^*) = h_2^0(O)$, but the key point is that this is difficult to enforce in practice, and hence β^* and γ^* are generally the pseudo-true values with respect to $h_1^0(O)$ and $h_2^0(O)$ respectively. Letting $\widehat{\beta}_n$ and $\widehat{\gamma}_n$ denote the estimators of β^* and γ^* respectively obtained by methods such as least squares, maximum likelihood, etc. – see, e.g., White (1981), White (1982) – using the observed data $O_1, \ldots, O_n$, one obtains $\widehat{h}_{1,n}(O_i) = h_1^*(O_i;\widehat{\beta}_n)$ and $\widehat{h}_{2,n}(O_i) = h_2^*(O_i;\widehat{\gamma}_n)$.

Interestingly, one could show that $\widehat{\theta}_n$ is consistent for θ^0 under standard conditions by virtue of

the unbiasedness conditions (2) and (3) if either one or both of the following specifications hold:

$$h_1^*(O) = h_1^0(O) \tag{5}$$

$$h_2^*(O) = h_2^0(O). \tag{6}$$

That is, the estimator $\widehat{\theta}_n$'s consistency for θ^0 is doubly-robust to the misspecification of the nuisance parameter models due to violation of (5) or (6). More generally, $\widehat{\theta}_n$ would perhaps be consistent for some pseudo-true value $\theta^* \in \Theta$; and $\theta^* = \theta^0$ if at least one of the specifications (5) and (6) holds.

Now, consider the general form of the asymptotic distribution of this doubly-robust estimator $\widehat{\theta}_n$. First, note that, under conditions as in, e.g., White (1981) and White (1982), $\widehat{\beta}_n$ and $\widehat{\gamma}_n$ have the asymptotically linear representations:

$$\sqrt{n}(\widehat{\beta}_n - \beta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L(O_i) + o_p(1) \tag{7}$$

$$\sqrt{n}(\widehat{\gamma}_n - \gamma^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S(O_i) + o_p(1) \tag{8}$$

with some influence function $L(O_i)$ and $S(O_i)$ respectively for which, in the case of two-step estimation, it is conventional to assume (see, e.g., Section 6 of Newey and McFadden (1994)) that:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\psi'(O_i), L'(O_i), S'(O_i)\right)' \xrightarrow{d} N(0, \Sigma) \tag{9}$$

for some $(d_\theta + d_\beta + d_\gamma) \times (d_\theta + d_\beta + d_\gamma)$ positive definite Σ, and where $\psi(O_i) \equiv \psi(O_i, \theta^*, h_1^*(O_i; \beta^*), h_{2,}^*(O_i; \gamma^*))$. Now, expanding (4) under standard regularity conditions (for $\psi(.)$ smooth or non-smooth in θ) as in Newey and McFadden (1994) or Chen et al. (2003) (parametric version) gives:

$$\begin{aligned} o_p(1) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(O_i) + \Psi_\theta \sqrt{n} \left(\widehat{\theta}_n - \theta^*\right) + \Psi_\beta \sqrt{n} \left(\widehat{\beta}_n - \beta^*\right) + \Psi_\gamma \sqrt{n} \left(\widehat{\gamma}_n - \gamma^*\right) \\ \Rightarrow \sqrt{n}(\widehat{\theta}_n - \theta^*) &= -\Psi_\theta^{-1} A \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i + o_p(1) \xrightarrow{d} N(0, \Psi_\theta^{-1} A \Sigma A' \Psi_\theta^{-1'}) \end{aligned} \tag{10}$$

using (7)–(9), and where for each of $\eta = \theta$, $\eta = \beta$ and $\eta = \gamma$:

$$\Psi_\eta \equiv \Psi_\eta(\theta^*, h_1^*(O), h_2^*(O)) \equiv \left\{ \frac{\partial}{\partial \eta'} E\left[\psi(O; \theta, h_1^*(O; \beta^*), h_2^*(O; \gamma^*)\right] \right\}_{\theta=\theta^*, \beta=\beta^*, \gamma=\gamma^*} \tag{11}$$

$$\text{and } A \equiv \begin{bmatrix} I_{d_\theta} & \Psi_\beta & \Psi_\gamma \end{bmatrix}. \tag{12}$$

Therefore, asymptotic normality follows in the standard way. Interestingly, when we consider (10) for specific examples in Section 4, it will become evident that $\Psi_\beta = 0$ when (3) holds, while $\Psi_\gamma = 0$ when (2) holds. In other words, estimation of one set of nuisance parameters would not affect the asymptotic variance of $\widehat{\theta}_n$ when the parametric model for the other set of nuisance parameters is correctly specified. Importantly, when both sets of nuisance parameter models are correctly specified, the asymptotic variance of $\widehat{\theta}_n$ is not at all affected by the estimation of any nuisance parameters – it is as if the true nuisance parameters have been plugged-in for the estimation of $\widehat{\theta}_n$ in (4). Thus, it will be evident that the doubly-robust estimators can enjoy the celebrated Neyman-C-alpha property.

3 Estimation framework with missing-at-random data

Now we describe an estimation framework following Chen et al. (2008) that has found wide applicability in economics, and that is useful for a self-contained illustration of the doubly-robust estimator. Generalizations of certain features of this framework have also proved useful – see, e.g., Graham et al. (2016), Chaudhuri and Guilkey (2016), Chaudhuri et al. (2018), Chaudhuri (2017), Barnwell and Chaudhuri (2018) – but the essential idea behind doubly-robust estimators remains the same.

Consider a random variable $Z = (Z_0,' Z_1')'$ where Z_0 is always observed but Z_1 can be missing. Let R be a binary variable taking value 1 if Z_1 is missing, and is 0 otherwise. The observed sample is $\{O_i \equiv (R_i, Z_{0i}', (1-R_i)Z_{1i}')\}_{i=1}^n$ that are i.i.d. copies of the random variable $O \equiv (R, Z_0', (1-R)Z_1')'$.

We will focus on a parameter value $\theta^0 \in \Theta \subset \mathbb{R}^d$ that is uniquely defined by a set of moment restrictions, and for this purpose we consider the following two distinct cases to define θ^0:

$$E[g(Z;\theta)] = 0 \text{ if and only if } \theta = \theta^0 \tag{13}$$

$$E[g(Z;\theta)|R=1] = 0 \text{ if and only if } \theta = \theta^0. \tag{14}$$

Chen et al. (2004) present five important examples in applied economics where moment restrictions like (13) or (14) arise. These include common cases such as estimation under nonclassical measurement error but with validation data (i.e., with $R = 0$), missing regressors, survey revisions, program evaluation such as estimation of the average treatment effect (on the treated, untreated), etc. We will also consider a new example – missing instrumental variables – that also fits the above framework but has only begun to receive attention very recently; see, e.g., Mogstad and Wiswall (2012), Chaudhuri and Guilkey (2016), Abrevaya and Donald (2017), Kennedy and Small (2018).

With missing observations, identification of θ^0 has traditionally been achieved by the "missing-at-random"(MAR) assumption, and an assumption on the "overlap" between the two groups $R = 0, 1$; see, among others, Rubin (1976), Rosenbaum and Rubin (1983), Robins et al. (1994), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Hahn (1998), Hirano et al. (2003), Wooldridge (2007), Chen et al. (2008), Graham (2011), Graham et al. (2012, 2016). These are stated as follows.
MAR: The following conditional independence assumption holds for the missingness indicator:

$$P(R = 1|Z_0, Z_1) = P(R = 1|Z_0). \tag{15}$$

Strict overlap: For each Z_0, the conditional probability of missingness is bounded away from 0 and 1:

$$P(R = 1|Z_0) \in (1 - \bar{p}, \bar{p}) \text{ for some } \bar{p} \text{ satisfying } 0 < \bar{p} < 1. \tag{16}$$

Now we consider examples of estimation framework – linear regression, instrumental variables regression and quantile regression – to illustrate the effect of MAR on the so-called "complete-case estimator"that uses only the non-missing observations and ignores the missing ones. This practice is common in applications. For brevity, we only consider the moment restrictions in (13) for now.
Example 1: Missing outcome variable in ordinary least squares regression

Let $Z_0 = X$ and $Z_1 = Y$, and consider a model $Y = X\theta^0 + \epsilon$ (all scalar). Let $E[X\epsilon] = E[\epsilon] = 0$. This implies that $\theta^0 = E[XY]/E[X^2]$ is defined by (13) with $g(Z; \theta) = X(Y - X\theta)$. The common practice is to use only the non-missing observations for estimation, i.e., the estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n}(1 - R_i)X_i(Y_i - X_i\widetilde{\theta}_{n,\text{Comp-Case}}) = 0;$$

and the complete-case estimator $\widetilde{\theta}_{n,\text{Comp-Case}} \xrightarrow{P} \dfrac{E[(1-R)XY]}{E[(1-R)X^2]} = \dfrac{E[(1-P(R=1|X))XY]}{E[(1-P(R=1|X))X^2]} \neq \theta^0$

generally unless the stronger assumption $E[\epsilon|X] = E[\epsilon] = 0$, i.e., $E[Y|X] = X\theta^0$, holds. On the other hand, using an inverse probability weighted (IPW) moment vector with the non-missing data $(1 - R)g(Z; \theta)/(1 - P(R = 1|X))$ solves this problem under (15) giving the estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(1 - R_i)}{(1 - P(R = 1|X_i))}X_i(Y_i - X_i\widehat{\theta}_{n,\text{IPW}}) = 0; \text{ and the estimator}$$

$$\widehat{\theta}_{n,\text{IPW}} \xrightarrow{P} E\left[\frac{1-R}{1-P(R=1|X)}XY\right] \Big/ E\left[\frac{1-R}{1-P(R=1|X)}X^2\right] = \frac{E[XY]}{E[X^2]} = \theta^0.$$

Example 2: Missing explanatory variable in ordinary least squares regression

Let $Z_0 = Y$ and $Z_1 = X$, and consider the same model as in Example 1. Now, the complete-case estimator $\widetilde{\theta}_n$ is inconsistent for θ^0 even under the much stronger condition $E[\epsilon|X] = 0$ because

$$\widetilde{\theta}_{n,\text{Comp-Case}} \xrightarrow{P} \frac{E[(1-R)XY]}{E[(1-R)X^2]} = \frac{E[(1-P(R=1|Y))XY]}{E[(1-P(R=1|Y))X^2]} \neq \theta^0.$$

On the other hand, the IPW moment vector, as before, solves the problem exactly as in Example 1, with the minor difference that now the conditional probabilities are conditional on Y (and not X).

Example 3: Missing instrumental variable in instrumental variables regression

Let $Z_0 = (Y, X)$ and $Z_1 = W$, and consider a model $Y = X\theta^0 + \epsilon$ where $E[X\epsilon] \neq 0$ and $E[\epsilon] = 0$. Let $E[W\epsilon] = 0$ and $E[WX] \neq 0$, i.e., W is the IV. Hence, $\theta^0 = E[WY]/E[WX]$ is defined by (13) with $g(Z;\theta) = W(Y - X\theta)$. Using only the non-missing observations gives the estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n}(1-R_i)W_i(Y_i - X_i\widetilde{\theta}_{n,\text{Comp-Case}}) = 0;$$

$$\text{and the estimator } \widetilde{\theta}_{n,\text{Comp-Case}} \xrightarrow{P} \frac{E[(1-R)WY]}{E[(1-R)WX]} = \frac{E[(1-P(R=1|Y,X))E[W|Y,X]Y]}{E[(1-P(R=1|Y,X))E[W|Y,X]X]} \neq \theta^0$$

generally even under the stronger assumption $E[\epsilon|W] = E[\epsilon] = 0$. As before, an IPW moment vector $(1-R)g(Z;\theta)/(1-P(R=1|Y,X))$ solves this problem under (15) giving the estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(1-R_i)}{(1-P(R=1|Y_i,X_i))}W_i(Y_i - X_i\widehat{\theta}_{n,\text{IPW}}) = 0; \quad \text{and the estimator}$$

$$\widehat{\theta}_{n,\text{IPW}} \xrightarrow{P} E\left[\frac{1-R}{1-P(R=1|Y,X)}WY\right] \Big/ E\left[\frac{1-R}{1-P(R=1|Y,X)}WX\right] = \frac{E[WY]}{E[WX]} = \theta^0.$$

Example 4: Missing explanatory variable in quantile regression

Let $Z_0 = Y$ and $Z_1 = X$. Consider the model as in Example 2, but instead of a restriction on the conditional mean of ϵ let it be on its $\tau \in (0,1)$-th conditional quantile as $P(\epsilon \leq 0|X) = \tau$. Then, θ^0 defined by (13) with $g(Z;\theta) = X(1(Y - X\theta \leq 0) - \tau)$, where $1(.)$ denotes the indicator function, typically does not have a closed form expression. However, θ^0 is such that $x\theta^0$ is the τ-th conditional quantile of Y given $X = x$. Using only the non-missing observations gives the estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n}(1-R_i)X_i(1(Y_i - X_i\widetilde{\theta}_{n,\text{Comp-Case}} \leq 0) - \tau) = o_p\left(\frac{1}{\sqrt{n}}\right)$$

where the $o_p(1/\sqrt{n})$ term on the right hand side (RHS) accommodates for non-exact solutions (see

(4)). However, unless R is independent of Y and X, the right hand side (LHS) of the estimating equations at θ cannot be unbiased or consistent for the expected moment function $E[g(Z;\theta)]$. Hence, $\widetilde{\theta}_{n,\text{Comp-Case}}$ cannot be consistent for θ^0. As before, an IPW moment vector $(1-R)g(Z;\theta)/(1-P(R=1|Y))$ solves this problem under (15) giving the estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(1-R_i)}{1-P(R=1|Y_i)}X_i(1(Y_i - X_i\widehat{\theta}_{n,\text{IPW}} \leq 0) - \tau) \quad = \quad o_p\left(\frac{1}{\sqrt{n}}\right)$$

because, the LHS at θ is now unbiased and, under standard conditions, consistent for $E[g(Z;\theta)]$.

The general principle by which the IPW moment vector solves the problem under MAR (15) in these four examples under (13) (and, similarly, also under (14)) can be summarized as follows.

Lemma 3.1 *Under* (15) *and standard existence conditions of expectations, the following identities (i) and (ii) in* θ *apply to the cases of the moment restrictions in* (13) *and* (14) *respectively:*

(i) $E\left[\frac{1-R}{1-P(R=1|Z_0)}g(Z;\theta)\right] = E[g(Z;\theta)]$

(ii) $E\left[\frac{P(R=1|Z_0)}{P(R=1)}\frac{1-R}{1-P(R=1|Z_0)}g(Z;\theta)\right] = E[g(Z;\theta)|R=1]$. ■

Weighting the moment vector for the sample units with Z fully observed (i.e., $R=0$) by suitable functions of $P(R=1|Z_0)$ restores the original moment restrictions in (13) and (14) under MAR (15). Then, estimation of θ based on estimating equations that are sample analogs of these reweighted moment vectors leads to consistency of the estimator. Such estimators are known as IPW estimators.

$P(R=1|Z_0)$ is unknown but can be estimated. (Do note that $P(R=1|Z)$ could not be estimated without (15).) The parametric IPW estimator for θ uses an estimator of $P(R=1|Z_0)$ based on a parametric model. However, the parametric IPW estimator is generally inconsistent for θ^0 if this parametric model is incorrect. Even otherwise, its asymptotic variance does not generally attain a suitable efficiency bound; see, e.g., Hahn (1998), Hirano et al. (2003), Wooldridge (2007).[1]

However, doubly-robust estimation can be applied in this context to alleviate both problems – bias and inefficiency. We discuss this now by referring to our general discussion from Section 2.

4 Doubly-Robust estimators for θ in (13) and (14) under MAR

There exists a variety of ways of constructing doubly-robust estimators in this context; see, e.g., Tan (2007), Cao et al. (2009), etc. for very innovative proposals. Here, we only describe the most basic

[1] Nonparametric estimation of $P(R=1|Z_0)$ can "technically" solve both problems – misspecification and efficiency – under suitable conditions; see, e.g., Hirano et al. (2003) and Chen et al. (2008). However, this is rarely used in practice.

form of doubly robust estimators that is due to the pioneering work of Robins et al. (1994). This should serve as a stepping stone for the interested reader into more sophisticated related methods.

The inefficiency of the parametric IPW estimator noted above leads to the natural question: What needs to be the variance for an estimator to be efficient in this context? The answer is provided by Theorem 1 of Chen et al. (2008). We restate it now in a slightly extended form. First, define:

$$\begin{aligned}\varphi(O;\theta) &= \frac{1-R}{1-P(R=1|Z_0)}\left(g(Z;\theta)-E[g(Z;\theta)|Z_0]\right)+E[g(Z;\theta)|Z_0] \\ \varphi_1(O;\theta) &= \frac{P(R=1|Z_0)}{P(R=1)}\frac{1-R}{1-P(R=1|Z_0)}\left(g(Z;\theta)-E[g(Z;\theta)|Z_0]\right)+\frac{R}{P(R=1)}E[g(Z;\theta)|Z_0],\end{aligned}$$

and the expectation of their outer products as:

$$V(\theta)=E[\varphi(O;\theta)\varphi'(O;\theta)] \text{ and } V_1(\theta)=E[\varphi_1(O;\theta)\varphi_1'(O;\theta)].$$

To accommodate for cases such as quantile regression, i.e., the $g(Z;\theta)$ in Example 4, we will allow $g(Z;\theta)$ to be non-differentiable in θ. However, $E[g(Z;\theta)]$ or $E[g(Z;\theta)|R=1]$ (as appropriate) needs to be differentiable, which is also satisfied in the quantile regression example. Accordingly, define:

$$G(\theta)=\frac{\partial}{\partial\theta'}E[g(Z;\theta)] \text{ and } G_1(\theta)=\frac{\partial}{\partial\theta'}E[g(Z;\theta)|R=1].$$

Proposition 4.1 *Assume that the MAR condition* (15) *and the strict overlap condition* (16) *hold. (i) Suppose that the moment restrictions in* (13) *hold. Let the Jacobian* $G\equiv G(\theta^0)$ *be full rank, and the variance matrix* $V\equiv V(\theta^0)$ *be positive definite. Then, the asymptotic variance lower bound for any regular estimator of* θ^0 *is* $\Omega\equiv G^{-1}VG^{-1'}$. *Any regular estimator* $\widehat{\theta}_n$ *with asymptotic variance equal to* Ω *has the asymptotically linear representation:*

$$\sqrt{n}\left(\widehat{\theta}_n-\theta^0\right)=-G^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(O_i;\theta^0)+o_p(1).$$

(ii) Suppose that the moment restrictions in (14) *hold. Let the Jacobian* $G_1\equiv G_1(\theta^0)$ *be full rank, and the variance matrix* $V_1\equiv V_1(\theta^0)$ *be positive definite. Then, the asymptotic variance lower bound for any regular estimator of* θ^0 *is* $\Omega_1\equiv G_1^{-1}V_1G_1^{-1'}$. *Any regular estimator* $\widehat{\theta}_n$ *with asymptotic variance equal to* Ω *has the asymptotically linear representation:*

$$\sqrt{n}\left(\widehat{\theta}_n-\theta^0\right)=-G_1^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_1(O_i;\theta^0)+o_p(1). \blacksquare$$

Remark: See Chen et al. (2008) for the proof. Under MAR in (15) and the overlap condition in (16), it is easy to verify that $G(\theta) = \frac{\partial}{\partial \theta'} E[\varphi(O;\theta)]$, $G_1(\theta) = \frac{\partial}{\partial \theta'} E[\varphi_1(O;\theta)]$. Therefore, the asymptotically linear representations in Proposition 4.1 suggest that an estimator $\widehat{\theta}_n$ obtained as the solution to:

$$o_p\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{n}\sum_{i=1}^{n}\varphi(O_i;\theta) \quad \text{or} \quad o_p\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{n}\sum_{i=1}^{n}\varphi_1(O_i;\theta)$$

would be efficient for θ^0 in (13) or (14) respectively. However, the problem with this is that both $\varphi(O;\theta)$ and $\varphi_1(O;\theta)$ involve unknown nuisance parameters $P(R=1|Z_0)$ and $E[g(Z;\theta)|Z_0]$ that need to be estimated (parametrically, in this paper,) for the above estimation strategy to be feasible. ■

Such estimation, however, runs the risk of misspecification of these nuisance parameters. This is where the discussion of the doubly-robust estimator from Section 2 becomes relevant. We will now revisit the general discussion from Section 2 and emphasize the simplifications and the interesting aspects of the doubly-robust estimators evidenced under our setup of estimation with MAR data.

Accordingly, we will define $\psi(.)$ in (1) informed by the structure of $\varphi(.)$ and $\varphi_1(.)$ in the cases of the moment restrictions (13) and (14) respectively. Under both cases (13) and (14), the nuisance parameters $E[g(Z;\theta^0)|Z_0]$ and $P(R=1|Z_0)$ will play the role of $h_1^0(O)$ and $h_2^0(O)$ respectively:

$$E[g(Z;\theta^0)|Z_0] \equiv h_1^0(O) \equiv h_1^0(Z_0) \quad \text{and} \quad P(R=1|Z_0) \equiv h_2^0(O) \equiv h_2^0(Z_0). \tag{17}$$

Lemma 4.2 *Let the MAR condition* (15) *and the strict overlap condition* (16) *hold along with the representation of the nuisance parameters in* (17)*.*
(i) Consider the moment restrictions in (13)*. Then,* $\psi(O;\theta,h_1^0(Z_0),h_2^0(Z_0))$ *defined as:*

$$\psi(O;\theta,h_1^0(Z_0),h_2^0(Z_0)) = \frac{1-R}{1-h_2^0(Z_0)}\left(g(Z;\theta)-h_1^0(Z_0)\right)+h_1^0(Z_0)$$

guided by $\varphi(O;\theta)$*, satisfies the condition in* (1) *and the double-robustness conditions* (2) *and* (3)*.*
(ii) Consider the moment restrictions in (14)*. Then,* $\psi(O;\theta,h_1^0(Z_0),h_2^0(Z_0))$ *defined as:*

$$\psi(O;\theta,h_1^0(Z_0),h_2^0(Z_0)) = \frac{h_2^0(Z_0)}{P(R=1)}\frac{1-R}{1-h_2^0(Z_0)}\left(g(Z;\theta)-h_1^0(Z_0)\right)+\frac{R}{P(R=1)}h_1^0(Z_0)$$

guided by $\varphi_1(O;\theta)$*, satisfies the condition in* (1) *and the double-robustness conditions* (2) *and* (3)*.* ■

Remark: The important feature that enables the double-robustness conditions in Lemma 4.2 under

the MAR condition (15) is the structure of $h_1^0(O)$ and $h_2^0(O)$ in (17), namely that both are indeed functions of only Z_0 under the setup of Sections 3-5. To emphasize this, we will henceforth write them as $h_1^0(Z_0)$ and $h_2^0(Z_0)$ respectively as was done in (17) and Lemma 4.2; and also write similarly for the related quantities $h_1^*(Z_0)$ and $h_2^*(Z_0)$. This feature leads to a further useful simplification in the general results on the asymptotic distribution, more specifically, the asymptotic variance as stated in (10), (11) and (12) in Section 2. We now make this explicit in Lemma 4.3 below. ■

Let $h_1^*(Z_0, \beta)$ and $h_2^*(Z_0; \gamma)$ be the postulated parametric models, possibly misspecified, for $h_1^0(Z_0)$ and $h_2^0(Z_0)$ respectively. Let β^* and γ^* be the pseudo-true values for β and γ, i.e., $h_1^*(Z_0, \beta^*) = h_1^*(Z_0)$ and $h_2^*(Z_0; \gamma) = h_2^*(Z_0)$. By definition, the functions $h_1^*(Z_0; \beta)$ and $h_2^*(Z_0; \gamma)$ are known up to β and γ. Let θ^* be the pseudo-true value for θ with $\theta^* = \theta^0$ if either or both of the following hold:

$$\Delta_1(Z_0) \equiv h_1^0(Z_0) - h_1^*(Z; \beta^*) = 0 \tag{18}$$

$$\Delta_2(Z_0) \equiv h_2^0(Z_0) - h_2^*(Z; \gamma^*) = 0. \tag{19}$$

Of course, if neither (18) nor (19) holds, then $\theta^* \neq \theta^0$, and it is generally not possible to characterize the pseudo-true value θ^*, if it exists, in a meaningful way. Hence, in the sequel, we will assume:

Assumption M:

At least one of (18) and (19) holds, i.e., the pseudo-true value θ^* is the true value θ^0.

Lemma 4.3 *Let the MAR condition* (15)*, the strict overlap condition* (16)*, and Assumption M hold. (i) Consider the moment restrictions in* (13) *and define* $\psi(O; \theta, h_1^*(Z_0; \beta), h_2^*(Z_0; \gamma))$ *as:*

$$\psi(O; \theta, h_1^*(Z_0; \beta), h_2^*(Z_0; \gamma)) = \frac{1 - R}{1 - h_2^*(Z_0; \gamma)} \left(g(Z; \theta) - h_1^*(Z_0; \beta)\right) + h_1^*(Z_0; \beta).$$

Then, the derivatives of the expectations, i.e., Ψ_β *and* Ψ_γ *defined in* (11)*, but with* $\theta^* = \theta^0$*, are:*

$$\begin{aligned}
\Psi_\beta \equiv \frac{\partial}{\partial \beta'} E\left[\psi(O; \theta^0, h_1^*(Z_0; \beta^*), h_2^*(Z_0; \gamma^*))\right] &= E\left[\frac{\Delta_2(Z_0)}{1 - h_2^*(Z_0; \gamma^*)} \left\{\frac{\partial}{\partial \beta'} h_1^*(Z_0; \beta^*)\right\}\right] \\
\Psi_\gamma \equiv \frac{\partial}{\partial \gamma'} E\left[\psi(O; \theta^0, h_1^*(Z_0; \beta^*), h_2^*(Z_0; \gamma^*))\right] &= -E\left[\frac{1 - h_2^0(Z_0)}{(1 - h_2^*(Z_0; \gamma^*))^2} \Delta_1(Z_0) \left\{\frac{\partial}{\partial \gamma'} h_2^*(Z_0; \gamma^*)\right\}\right].
\end{aligned}$$

(ii) Consider the moment restrictions in (14) *and define* $\psi(O; \theta, h_1^*(Z_0; \beta), h_2^*(Z_0; \gamma))$ *as:*

$$\psi(O; \theta, h_1^*(Z_0; \beta), h_2^*(Z_0; \gamma)) = \frac{h_2^*(Z_0)}{P(R = 1)} \frac{1 - R}{1 - h_2^*(Z_0; \gamma)} \left(g(Z; \theta) - h_1^*(Z_0; \beta)\right) + \frac{R}{P(R = 1)} h_1^*(Z_0; \beta).$$

Then, the derivatives of the expectations, i.e., Ψ_β and Ψ_γ defined in (11)*, but with $\theta^* = \theta^0$, are:*

$$\begin{aligned}\Psi_\beta &= E\left[\frac{\Delta_2(Z_0)}{P(R=1)(1-h_2^*(Z_0;\gamma^*))}\left\{\frac{\partial}{\partial\beta'}h_1^*(Z_0;\beta^*)\right\}\right] \\ \Psi_\gamma &= -E\left[\frac{1-h_2^0(Z_0)}{P(R=1)}\frac{h_2^*(Z_0;\gamma^*)^2}{(1-h_2^*(Z_0;\gamma^*))^2}\Delta_1(Z_0)\left\{\frac{\partial}{\partial\gamma'}h_2^*(Z_0;\gamma^*)\right\}\right]. \ \blacksquare\end{aligned}$$

Remarks: Lemma 4.3 helps to make precise the Neyman-C-alpha property of doubly-robust estimators, a key property that we discuss now. First, when (18) ($\Delta_1(Z_0) = 0$) holds, i.e., the model for $h_1^0(Z_0)$ is correctly specified, then the estimation of γ in the other nuisance parameter model $h_2^*(Z_0;\gamma)$ does not affect the asymptotic variance of the estimator for the parameter of interest θ. This is seen by considering (10), (11) and (12) and noting that $\Psi_\gamma = 0$ by Lemma 4.3, and hence:

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta^0) &= -\Psi_\theta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(O_i) \ - \ \Psi_\theta^{-1}\Psi_\beta\sqrt{n}\left(\widehat{\beta}_n - \beta^*\right) \ + \ o_p(1) \\ &\xrightarrow{d} N\left(0, \Psi_\theta^{-1}\left[\begin{array}{ccc} I_{d_\theta} & \Psi_\beta & 0 \end{array}\right]\Sigma\left[\begin{array}{ccc} I_{d_\theta} & \Psi_\beta & 0 \end{array}\right]'\Psi_\theta^{-1'}\right). \qquad (20)\end{aligned}$$

Similarly, when (19) ($\Delta_2(Z_0) = 0$) is satisfied, i.e., the model for $h_2^0(Z_0)$ is correctly specified, then the estimation of β in the other nuisance parameter model $h_1^*(Z_0;\beta)$ does not affect the asymptotic variance of $\widehat{\theta}_n$ because, considering (10), (11) and (12), then Lemma 4.3 implies that $\Psi_\beta = 0$, giving:

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta^0) &= -\Psi_\theta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(O_i) \ - \ \Psi_\theta^{-1}\Psi_\gamma\sqrt{n}\left(\widehat{\gamma}_n - \gamma^*\right) \ + \ o_p(1) \\ &\xrightarrow{d} N\left(0, \Psi_\theta^{-1}\left[\begin{array}{ccc} I_{d_\theta} & 0 & \Psi_\gamma \end{array}\right]\Sigma\left[\begin{array}{ccc} I_{d_\theta} & 0 & \Psi_\gamma \end{array}\right]'\Psi_\theta^{-1'}\right). \qquad (21)\end{aligned}$$

Finally, when both (18) and (19) ($\Delta_1(Z_0) = 0$ and $\Delta_2(Z_0) = 0$) hold, i.e., both nuisance parameter models are correctly specified, then the estimation of neither β nor γ has any effect on the asymptotic variance of $\widehat{\theta}_n$ because, now Lemma 4.3 implies that $\Psi_\beta = 0$ and $\Psi_\gamma = 0$, and hence:

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta^0) &= -\Psi_\theta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(O_i) \ + \ o_p(1) \xrightarrow{d} N\left(0, \Psi_\theta^{-1}\left[\begin{array}{ccc} I_{d_\theta} & 0 & 0 \end{array}\right]\Sigma\left[\begin{array}{ccc} I_{d_\theta} & 0 & 0 \end{array}\right]'\Psi_\theta^{-1'}\right) \\ &\sim \begin{cases} N(0,\Omega) & \text{under the moment restrictions in (13)} \\ N(0,\Omega_1) & \text{under the moment restrictions in (14).} \end{cases} \qquad (22)\end{aligned}$$

That is, $\widehat{\theta}_n$ is efficient according to Proposition 4.1.[2] Nonparametric estimation of the nuisance

[2]This follows under the moment restrictions in (13) and (14) respectively by the definition of $\psi(O)$ that links it to

parameters $h_1^0(Z_0)$ and $h_2^0(Z_0)$ as in, e.g., Cattaneo (2010) and Rothe and Firpo (2016), skips the first two observations (20) and (21), and directly arrives at observations similar to (22). However, given the practical relevance of parametric estimation, especially with real-life data, it is also important to appreciate the first two observations. Collectively, these three observations (20)-(22) describe the Neyman-C-alpha property of the doubly-robust estimation under the setup of Sections 3 and 4.[3] ■

Let us now describe the implementation of the doubly-robust estimators with the help of the four examples introduced in Section 3. Since the nuisance parameter $h_2^0(Z_0) \equiv P(R=1|Z_0)$ can be modeled as $h_2^*(Z_0;\gamma)$ without consideration of the moment function $g(Z;\theta)$, let us consider its estimation first by (quasi)-maximum likelihood, under standard conditions as in White (1982), as:

$$\widehat{h}_{2,n}(Z_{0,i}) \equiv h_2^*(Z_{0,i};\widehat{\gamma}_n) \text{ where}$$
$$\widehat{\gamma}_n \equiv \arg\max_{\gamma} \frac{1}{n}\sum_{i=1}^{n} \left[R_i \log\left(h_2^*(Z_{0,i};\gamma)\right) + (1-R_i)\log\left(1-h_2^*(Z_{0,i};\gamma)\right)\right]. \tag{23}$$

This allows for fully general parametric specification for $h_2^*(Z_0;\gamma)$. In practice, $h_2^*(Z_0;\gamma)$ is typically a logit or probit specification with an index $r'(Z_0)\gamma$ where $r(Z_0)$ is a $d_\gamma \times 1$ vector of chosen functions of Z_0 such as powers with interactions of the elements, or discretization of Z_0.

Now, consider the estimation of the nuisance parameter $h_2^0(Z_0) \equiv E[g(Z;\theta^0)|Z_0]$. This would typically be estimated by (non-) linear least squares under standard conditions as in White (1981):

$$\widehat{h}_{1,n}(Z_{0,i};\theta) \equiv h_1^*(Z_{0,i};\widehat{\beta}_n(\theta)) \text{ where}$$
$$\widehat{\beta}_n(\theta) \equiv \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(1-R_i)\left(g(Z_{0,i},Z_{1,i};\theta) - h_1^*(Z_{0,i};\beta)\right)'\left(g(Z_{0,i},Z_{1,i};\theta) - h_1^*(Z_{0,i};\beta)\right) \tag{24}$$

for any θ. While this general expression is stated in a profiled form with respect to θ, inspection of the four examples suggests that this would not always be necessary. To see this, first note that:

Example 1: $E[g(Z;\theta)|Z_0] = E[X(Y-X\theta)|X] = XE[Y|X] - X^2\theta$

Example 2: $E[g(Z;\theta)|Z_0] = E[X(Y-X\theta)|Y] = E[X|Y]Y - E[X^2|Y]\theta$

Example 3: $E[g(Z;\theta)|Z_0] = E[W(Y-X\theta)|Y,X] = E[W|Y,X]Y - E[W|Y,X]X\theta$

Example 4: $E[g(Z;\theta)|Z_0] = E[X(1(Y-X\theta\leq 0)-\tau)|Y] = E[X1(Y-X\theta\leq 0)|Y] - E[X|Y]\tau.$

$\varphi(O;\theta)$ and $\varphi(O;\theta)$, and hence results in $\Psi_\theta = G$ and G_1, and the $d_\theta \times d_\theta$ upper diagonal block of Σ as V and V_1.

[3] See Prokhorov and Schmidt (2009) for general discussion of the influence of the estimation of nuisance parameters in missing data models on the asymptotic variance of the estimator for the parameter of interest.

Postulating a (possibly misspecified) parametric model $h_1^*(Z_0;\beta)$ in Examples 1, 2 and 3 is a standard task because of the linearity in θ, since this essentially involves modeling $E[Y|X]$, $E[X|Y]$ and $E[X^2|Y]$, and $E[W|Y,X]$ respectively, which when combined with the respective complete expressions above gives the parametric models $h_1^*(Z_0;\beta)$'s. Because of this linearity in θ, the doubly-robust estimator $\widehat{\theta}_n$ in (4) for the parameter of interest θ has closed form expressions in Examples 1, 2 and 3 under the moment restrictions in (13) and (14). Specifically, writing the estimated conditional expectations generically as $\widehat{E}_n[.|Z_0]$ to avoid notational clutter, while recognizing how they constitute the estimated $h_1^*(Z_0;\beta)$, i.e., $\widehat{h}_{2,n}(Z_0)$, one can obtain the following closed form and computationally convenient expressions for $\widehat{\theta}_n$ by simple algebra. (Their extension to vector-valued $g(Z;\theta)$ is obvious.)

Example 1: The doubly-robust estimators $\widehat{\theta}_n$ for θ^0's defined in (13) and (14) are respectively:

$$\begin{aligned}
\widehat{\theta}_n &= \frac{\sum_{i=1}^{n} \frac{1-R_i}{1-\widehat{h}_{2,n}(X_i)} X_i(Y_i - \widehat{E}_n[Y|X_i]) + \sum_{i=1}^{n} X_i \widehat{E}_n[Y|X_i]}{\sum_{i=1}^{n} X_i^2}, \\
\widehat{\theta}_n &= \frac{\sum_{i=1}^{n} \widehat{h}_{2,n}(X_i) \frac{1-R_i}{1-\widehat{h}_{2,n}(X_i)} X_i(Y_i - \widehat{E}_n[Y|X_i]) + \sum_{i=1}^{n} R_i X_i \widehat{E}_n[Y|X_i]}{\sum_{i=1}^{n} R_i X_i^2}.
\end{aligned}$$

Example 2: The doubly-robust estimators $\widehat{\theta}_n$ for θ^0's defined in (13) and (14) are respectively:

$$\begin{aligned}
\widehat{\theta}_n &= \frac{\sum_{i=1}^{n} \frac{1-R_i}{1-\widehat{h}_{2,n}(Y_i)} (X_i - \widehat{E}_n[X|Y_i])Y_i + \sum_{i=1}^{n} \widehat{E}_n[X|Y_i]Y_i}{\sum_{i=1}^{n} \frac{1-R_i}{1-\widehat{h}_{2,n}(Y_i)} (X_i^2 - \widehat{E}_n[X^2|Y_i]) + \sum_{i=1}^{n} \widehat{E}_n[X^2|Y_i]}, \\
\widehat{\theta}_n &= \frac{\sum_{i=1}^{n} \widehat{h}_{2,n}(Y_i) \frac{1-R_i}{1-\widehat{h}_{2,n}(Y_i)} (X_i - \widehat{E}_n[X|Y_i])Y_i + \sum_{i=1}^{n} R_i \widehat{E}_n[X|Y_i]Y_i}{\sum_{i=1}^{n} \widehat{h}_{2,n}(Y_i) \frac{1-R_i}{1-\widehat{h}_{2,n}(Y_i)} (X_i^2 - \widehat{E}_n[X^2|Y_i]) + \sum_{i=1}^{n} R_i \widehat{E}_n[X^2|Y_i]}.
\end{aligned}$$

Example 3: The doubly-robust estimators $\widehat{\theta}_n$ for θ^0's defined in (13) and (14) are respectively:

$$\widehat{\theta}_n = \frac{\sum_{i=1}^{n} \frac{1-R_i}{1-\widehat{h}_{2,n}(Y_i,X_i)} (W_i - \widehat{E}_n[W|Y_i,X_i])Y_i + \sum_{i=1}^{n} \widehat{E}_n[W|Y_i,X_i]Y_i}{\sum_{i=1}^{n} \frac{1-R_i}{1-\widehat{h}_{2,n}(Y_i,X_i)} (W_i - \widehat{E}_n[W|Y_i,X_i])X_i + \sum_{i=1}^{n} \widehat{E}_n[W|Y_i,X_i]X_i},$$

$$\widehat{\theta}_n = \frac{\sum_{i=1}^{n} \widehat{h}_{2,n}(Y_i, X_i) \frac{1-R_i}{1-\widehat{h}_{2,n}(Y_i, X_i)} (W_i - \widehat{E}_n[W|Y_i, X_i]) Y_i + \sum_{i=1}^{n} R_i \widehat{E}_n[W|Y_i, X_i] Y_i}{\sum_{i=1}^{n} \widehat{h}_{2,n}(Y_i, X_i) \frac{1-R_i}{1-\widehat{h}_{2,n}(Y_i, X_i)} (W_i - \widehat{E}_n[W|Y_i, X_i]) X_i + \sum_{i=1}^{n} R_i \widehat{E}_n[W|Y_i, X_i] X_i}.$$

Unfortunately, however, when the moment restrictions in (13) and (14) are nonlinear in θ, a closed form expression for the estimator of θ does not typically exist even without missing data. Naturally, one cannot then expect a closed form expression with missing data. A popular (if slightly extreme) example of such moment restrictions is quantile regression, i.e., our Example 4. It is also under this example that the profiled estimator $\widehat{\beta}_n(\theta)$ in the parametric model $h_1^*(Z_0; \beta)$ becomes necessary. In this case, the doubly-robust estimator $\widehat{\theta}_n$ solves the following estimating equation:

Example 4: The doubly-robust estimators $\widehat{\theta}_n$ for θ^0's defined in (13) and (14) solve respectively:

$$\begin{aligned}
o_p\left(\frac{1}{\sqrt{n}}\right) &= \frac{1}{n}\sum_{i=1}^{n} \frac{(1-R_i)}{1-\widehat{h}_{2,n}(Y_i)} X_i (1(Y_i - X_i\widehat{\theta}_n \leq 0) - \tau) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left(1 - \frac{(1-R_i)}{1-\widehat{h}_{2,n}(Y_i)}\right) \widehat{E}_n\left[X(1(Y_i - X\widehat{\theta}_n \leq 0) - \tau)|Y_i\right], \\
o_p\left(\frac{1}{\sqrt{n}}\right) &= \frac{1}{n}\sum_{i=1}^{n} \widehat{h}_{2,n}(Y_i) \frac{(1-R_i)}{1-\widehat{h}_{2,n}(Y_i)} X_i (1(Y_i - X_i\widehat{\theta}_n \leq 0) - \tau) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left(R_i - \widehat{h}_{2,n}(Y_i)\frac{(1-R_i)}{1-\widehat{h}_{2,n}(Y_i)}\right) \widehat{E}_n\left[X(1(Y_i - X\widehat{\theta}_n \leq 0) - \tau)|Y_i\right].
\end{aligned}$$

These are estimating equations that augment those for the IPW estimator in Section 3 with the last set of terms (those on the respective second lines) in both the above equations. This only adds to the computational burden, and does not cause problem with the theory of the asymptotic properties of $\widehat{\theta}_n$. Similarly, the non-differentiability of this specific $g(Z; \theta)$ with respect to θ is also a computational complexity that, evidently, even the IPW estimator defined in Section 3 (Example 4) cannot avoid.

The asymptotic theory in such cases is standard. Interested readers can, for example, use a parametric nuisance parameter version of Theorem 2 in Chen et al. (2003) to obtain the conditions that are sufficient for our general discussion in Section 2 (specifically, (10), (11) and (12)) to hold.

Finally, it should be noted that estimation of the asymptotic variance of $\widehat{\theta}_n$ can proceed in the standard way by computing the sample analogs of the relevant components of the asymptotic variance for two-step/plug-in estimators using the generic expressions stated in our Section 2. Lemma 4.3 and the remark following it, see, in particular, (20), (21) and (22), specify the (desired) probability limit of such estimators under various scenarios of misspecification of the nuisance parameter models.

5 Finite-sample behavior: Monte-Carlo experiment with MAR IV

We conduct a simulation experiment based on our Example 3 and demonstrate the inconsistency for θ^0 due to ignoring missing observations when the instrumental variable is MAR, the correction for it due to the IPW estimator, and finally the further improvement due to the doubly-robust estimator.

5.1 Simulation design and the estimators

Consider a random sample $\{Y_i, X_i, W_i\}_{i=1}^n$ drawn from the following model: $Y = X\theta^0 + (u + v)$, $X = W + v$ where $(u, v) \sim N(0, I_2)$ are the model errors. We consider two designs that differ in the specification for the unconditional distribution of the instrument W – Design-I: $W \sim N(0, 1)$ and Design-II: $W \sim Bin(1, .5)$. In both cases W is independent of u, v. The parameter of interest is θ. Its value θ^0 $(= -1)$ is defined by the moment restriction (13) with $g(Z; \theta) = W(Y - X\theta)$.

The full-observation IV estimator $\widetilde{\theta}_{\text{n,Full-Obsn.}} = \sum_{i=1}^n W_i Y_i / \sum_{i=1}^n W_i X_i$ is consistent for θ^0. This estimator sets the benchmark for bias and variance.

Now we construct the MAR-sample by drawing a random sample $\{R_i\}_{i=1}^n$ from R where

$$R|Y, X, W \sim R|Y, X \sim Bin\left(1, P(R = 1|Z_0) = \frac{1}{4} + \frac{1}{\pi}\arctan(Y^2)\right),$$

and then by deleting the W_i's corresponding to $R_i = 1$. The choice of $P(R = 1|Z_0)$ is influenced by the strict overlap assumption in (16). (Being overly cautious, we use the additive factor $1/4$ instead of $1/2$ unlike the standard Cauchy distribution function; it does not change our conclusions.)

With missing W_i's, the full-observation estimator, our benchmark, is no longer feasible. We consider the following feasible parametric estimators based on the available observations as alternatives to the benchmark: the complete-case estimator, the IPW and the doubly-robust estimators that, in the present context, were discussed under Example 3 in Sections 3 (for the former two) and 4.

We estimate the nuisance parameters $E[g(Z; \theta)|Z_0]$ and $P(R = 1|Z_0)$ based on parametric models.

For $E[g(Z; \theta)|Z_0] = E[W|Y, X](Y - X\theta)$, our postulated model is $h_1(Z_0; \beta, \theta) = t(Z_0)(Y - X\theta)\beta$ where $t(Z_0) = [1, Y, X]'$. This model is correct for Design-I, i.e., Assumption M holds for (18).

For $P(R = 1|Z_0)$, we consider three options: (i) the infeasible true $h_2^*(Z_0; \gamma^*) = P(R = 1|Z_0)$, (ii) a correct $h_2^*(Z_0; \gamma) = \frac{1}{4} + \frac{1}{\pi}\arctan(\gamma \times Y^2)$ with $\gamma^* = 1$, and (iii) two commonly used but wrong models $h_2^*(Z_0; \gamma) = \exp(r'(Z_0)\gamma)/[1 + \exp(r'(Z_0)\gamma)]$ where $r(Z_0) = [1, Y, X]'$ (call it PM-1:h_2) and $r(Z_0) = [1, Y, X, Y^2, X^2, YX]'$ (call it PM-2:h_2). Assumption M holds for (19) under (i) and (ii).

5.2 Summary of results

Based on 10000 simulations we estimate the mean bias, absolute bias, standard deviation and inter-quartile range of all the estimators for sample sizes $n = 500$ and 1000, and report them in Tables 1 (Design-1, $n = 500$), 2 (Design-1, $n = 1000$), 3 (Design-2, $n = 500$) and 4 (Design-2, $n = 1000$).

Parametric IV-Estimators	Mean Bias	Abs. Bias	Stdev	IQR
Full-Obsn	-0.0016	0.0504	0.0632	0.0847
Comp-Case	0.1953	0.1968	0.0870	0.1140
Model for $P(R = 1\|Z_0)$: option (i)				
IPW	-0.0022	0.0817	0.1035	0.1357
doubly-robust	-0.0017	0.0607	0.0766	0.1021
Model for $P(R = 1\|Z_0)$: option (ii)				
IPW	-0.0012	0.0817	0.1034	0.1353
doubly-robust	-0.0017	0.0607	0.0766	0.1022
Model for $P(R = 1\|Z_0)$: option (iii), PM-1:h_2				
IPW	0.1951	0.1967	0.0875	0.1145
doubly-robust	-0.0019	0.0596	0.0753	0.1006
Model for $P(R = 1\|Z_0)$: option (iii), PM-2:h_2				
IPW	-0.1288	0.1443	0.1291	0.1663
doubly-robust	-0.0014	0.0718	0.0914	0.1180

Table 1: Design-I (MAR instrument $W \sim N(0, 1)$): Sample size $n = 500$. Results are based on averages over 10000 replications in Matlab with seed 0 for random number generation.

Parametric IV-Estimators	Mean Bias	Abs. Bias	Stdev	IQR
Full-Obsn	-0.0011	0.0357	0.0449	0.0601
Comp-Case	0.1965	0.1966	0.0612	0.0816
Model for $P(R = 1\|Z_0)$: option (i)				
IPW	-0.0016	0.0581	0.0731	0.0982
doubly-robust	-0.0015	0.0431	0.0544	0.0727
Model for $P(R = 1\|Z_0)$: option (ii)				
IPW	-0.0011	0.0581	0.0731	0.0980
doubly-robust	-0.0015	0.0431	0.0544	0.0729
Model for $P(R = 1\|Z_0)$: option (iii), PM-1:h_2				
IPW	0.1963	0.1964	0.0614	0.0820
doubly-robust	-0.0013	0.0422	0.0533	0.0715
Model for $P(R = 1\|Z_0)$: option (iii), PM-2:h_2				
IPW	-0.1384	0.1416	0.0930	0.1227
doubly-robust	-0.0011	0.0515	0.0652	0.0857

Table 2: Design-I (MAR instrument $W \sim N(0, 1)$): Sample size $n = 1000$. Results are based on averages over 10000 replications in Matlab with seed 0 for random number generation.

The simulation results conform with the theoretical discussion. Ignoring the MAR instruments leads to bias in the complete-case estimator and the bias does not vanish as sample size increases. This is different from Mogstad and Wiswall (2012) and Abrevaya and Donald (2017). The bias is

Parametric IV-Estimators	Mean Bias	Abs. Bias	Stdev	IQR
Full-Obsn	-0.0048	0.0728	0.0915	0.1219
Comp-Case	0.2438	0.2462	0.1090	0.1420
Model for $P(R=1\|Z_0)$: option (i)				
IPW	-0.0067	0.1123	0.1428	0.1853
doubly-robust	-0.0049	0.0861	0.1090	0.1456
Model for $P(R=1\|Z_0)$: option (ii)				
IPW	-0.0052	0.1119	0.1423	0.1856
doubly-robust	-0.0048	0.0861	0.1090	0.1455
Model for $P(R=1\|Z_0)$: option (iii), PM-1:h_2				
IPW	0.1361	0.1487	0.1062	0.1408
doubly-robust	0.0026	0.0844	0.1065	0.1412
Model for $P(R=1\|Z_0)$: option (iii), PM-2:h_2				
IPW	-0.1578	0.1824	0.1715	0.2216
doubly-robust	0.0139	0.0939	0.1186	0.1523

Table 3: Design-II (MAR instrument $W \sim Bin(1,.5)$): Sample size $n = 500$. Results are based on averages over 10000 replications in Matlab with seed 0 for random number generation.

eliminated by the IPW and the doubly-robust estimators using the infeasible true $P(R=1|Z_0)$ or the estimated correct model, i.e., using option (i) or (ii), both of which are likely infeasible in practice.

Parametric IV-Estimators	Mean Bias	Abs. Bias	Stdev	IQR
Full-Obsn	-0.0026	0.0504	0.0632	0.0849
Comp-Case	0.2465	0.2466	0.0750	0.1007
Model for $P(R=1\|Z_0)$: option (i)				
IPW	-0.0041	0.0776	0.0982	0.1289
doubly-robust	-0.0033	0.0600	0.0754	0.1011
Model for $P(R=1\|Z_0)$: option (ii)				
IPW	-0.0033	0.0774	0.0980	0.1292
doubly-robust	-0.0033	0.0600	0.0754	0.1011
Model for $P(R=1\|Z_0)$: option (iii), PM-1:h_2				
IPW	0.1381	0.1405	0.0732	0.0983
doubly-robust	0.0041	0.0587	0.0735	0.0985
Model for $P(R=1\|Z_0)$: option (iii), PM-2:h_2				
IPW	-0.1671	0.1728	0.1199	0.1587
doubly-robust	0.0180	0.0671	0.0825	0.1092

Table 4: Design-II (MAR instrument $W \sim Bin(1,.5)$): Sample size $n = 1000$. Results are based on averages over 10000 replications in Matlab with seed 0 for random number generation.

The IPW estimator with the wrong model for $P(R=1|Z_0)$, i.e., option (iii) PPM-1:h_2, or the mildly wrong model, i.e., option (iii) PPM-2:h_2, is badly biased. However, under both these realistic scenarios involving option (iii), the doubly-robust estimator continues to avoid this bias problem whenever the parametric model for $E[g(Z;\theta)|Z_0]$ is correct (i.e., in Design-I). This is why the double-robustness property is attractive. In fact, even with the mildly wrong model for $E[g(Z;\theta)|Z_0]$ (i.e.,

in Design-II), there is almost no bias in the doubly-robust estimator.

As expected under our setup, the doubly-robust estimator is more (often much more) precise than the IPW estimator in all cases considered here.[4] In all aspects – mean bias, absolute bias, standard deviation and interquartile range – the doubly-robust estimator is closer to the infeasible full-observation IV estimator and hence seems desirable in terms of its behavior in finite samples. The same conclusion holds in the unreported simulation results under similar data generating processes.

6 Conclusion

The primary goal of our paper was to provide a non-technical overview of the so-called doubly-robust estimators that have caught the interest of econometricians in recent years. These estimators originated from the biostatistics literature, and there are a variety of doubly-robust estimators available for use; see, among others, Tan (2007) and Cao et al. (2009). However, the underlying mechanism for and the consequences of double-robustness are the same for all such estimators. Hence, we hope that our focus on the original formulation of doubly-robust estimators in Robins et al. (1994) should provide the stepping stone to the reader for exploring the other more sophisticated estimators.

Our review focused on presenting the doubly-robust estimator as a special case of the widely used two-step/plug-in estimators in econometrics where the first step estimates the nuisance parameters and the second step estimates the parameters of interest using the first step estimates. The discussion on double-robustness to misspecification is perhaps best appreciated that way. The additional benefit of the doubly-robust estimator that is evident under the setup of estimation with MAR data is the efficiency property that results from the correct specification of both nuisance parameters models. This means that in such cases the estimation of the nuisance parameters has no effect on the asymptotic variance of the estimator for the parameter of interest. However, we emphasized that the general result/intuition is that the estimation of one set of nuisance parameters does not affect the asymptotic variance when the other set of nuisance parameters is correctly specified. This property in the general case and, more conventionally, in the special case of no misspecification is due to the Neyman-C-alpha structure of the estimating function of the doubly-robust estimators.

Our small-scale simulation experiment demonstrates that the doubly-robust estimator can be much more preferable to the better-known IPW estimator even in small samples in terms of bias,

[4]The apparently puzzling phenomenon giving a smaller standard deviation of the IPW estimator under option (ii) than under the infeasible option (i) is already a well known fact [see, for e.g., Wooldridge (2007) and Graham (2011)].

variance and other properties. Moreover, as evident from the four common examples, the doubly-robust estimator is computationally not too demanding relative to the IPW estimator – closed form expression for both estimators exists in three of these four examples. It is, therefore, our hope that this demonstration of computational ease and desirable finite-sample properties, and the presentation of the asymptotic properties of the doubly-robust estimators by linking them to the well known two-step/plug-in estimators would encourage their adoption in empirical economics research.

References

Abrevaya, J. and Donald, S. G. (2017). A GMM approach for dealing with missing data on regressors and instruments. Forthcoming in Review of Economics and Statistics.

Barnwell, J. L. and Chaudhuri, S. (2018). Efficient estimation in sub and full populations with monotonically missing at random data. Technical report, McGill University.

Busso, M., DiNardo, J., and McCrary, J. (2009). Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects. Mimeo.

Busso, M., DiNardo, J., and McCrary, J. (2011). New Evidence on the finite Sample Properties of Propensity Score Reweighting and Matching Estimators. Mimeo.

Cao, W., Tsiatis, 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.

Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155:138–154.

Chaudhuri, S. (2017). A Note on Efficiency Gains from Multiple Incomplete Subsamples. Mimeo.

Chaudhuri, S., Frazier, D., and Renault, E. (2018). Indirect inference with endogenously missing exogenous variables. *Journal of Econometrics*, 205: 55–75.

Chaudhuri, S. and Guilkey, D. K. (2016). GMM with Multiple Missing Variables. *Journal of Applied Econometrics*, 31:678–706.

Chen, X., Hong, H., and Tarozzi, A. (2004). Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects. Mimeo.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric Efficiency in GMM Models with Auxiliary Data. *Annals of Statistics*, 36:808–843.

Chen, X., Linton, O., and van Keilegom, I. (2003). Estimation of Semiparametric Models when the Criteria Function is not Smooth. *Econometrica*, 71:1591–1608.

Graham, B. S. (2011). Efficiency Bounds for Missing Data Models with Semiparametric Restrictions. *Econometrica*, 79:437 – 452.

Graham, B. S., Pinto, C., and Egel, D. (2012). Inverse Probability Tilting for Moment Condition Models with Missing Data. *Review of Economic Studies*, 79:1053 – 1079.

Graham, B. S., Pinto, C. C. D. X., and Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting. *Journal of Business and Economic Statistics*, 34:288–301.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66:315–331.

Hirano, K. and Imbens, G. (2001). Estimation of Causal Effects using Propensity Score Weighting : An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology*, 2:259–278.

Hirano, K., Imbens, G., and Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores. *Econometrica*, 71:1161–1189.

Holcroft, C., Rotnitzky, A., and Robins, J. M. (1997). Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates. *Journal of Statistical Planning and Inference*, 65:349–374.

Kennedy, E. H. and Small, D. S. (2018). Paradoxes in instrumental variable studies with missing data and one-sided noncompliance. arXiv1705.00506.

Mogstad, M. and Wiswall, M. (2012). Instrumental variables estimation with partially missing instruments. *Economics Letters*, 114:186–189.

Newey, W. K. and McFadden, D. L. (1994). Large Sample Estimation and Hypothesis Testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume IV, chapter 36, pages 2212–2245. Elsevier Science Publisher.

Prokhorov, A. and Schmidt, P. (2009). GMM redundancy results for general missing data problems. *Journal of Econometrics*, 151:47–55.

Robins, J. and Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of American Statistical Association*, 90:122–129.

Robins, M., Rotnitzky, A., and Zhao, L. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of American Statistical Association*, 427:846–866.

Robins, M., Rotnitzky, A., and Zhao, L. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of American Statistical Association*, 429:106–121.

Rosenbaum, P. and Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70:41–55.

Rothe, C. and Firpo, S. (2016). Semiparametric Estimation and Inference Using doubly-Robust Moment Conditions. Mimeo.

Rotnitzky, A. and Robins, J. (1995). Semiparametric Regression Estimation in the Presence of Dependent Censoring. *Biometrika*, 82:805–820.

Rubin, D. (1976). Inference and Missing Data. *Biometrika*, 63:581–592.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94:1096–1146.

Sloczynski, T. and Wooldridge, J. M. (2018). A General Double Robustness Result for Estimating Average Treatment Effects. *Econometric Theory*, 34: 112–133.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22:560–568.

White, H. (1981). Consequence and Detection of Misspecified Nonlinear Regression Models. *Journal of the American Statistical Association*, 76:419–433.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50:1–25.

Wooldridge, J. (2007). Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics*, 141(2):1281–1301.

7 Technical Appendix

Proof of Lemma 3.1: Consider (i) and note that:

$$E\left[\frac{1-R}{1-P(R=1|Z_0)}g(Z;\theta)\right] = E\left[\frac{1-E[R|Z]}{1-P(R=1|Z_0)}g(Z;\theta)\right] = E\left[\frac{1-P(R=1|Z_0)}{1-P(R=1|Z_0)}g(Z;\theta)\right] = E[g(Z;\theta)]$$

where the first equality holds by the law of iterated expectations (LIE) and the second equality by the MAR condition (15). Now consider (ii), and note that:

$$\begin{aligned}
E\left[\frac{P(R=1|Z_0)}{P(R=1)}\frac{1-R}{1-P(R=1|Z_0)}g(Z;\theta)\right] &= E\left[\frac{P(R=1|Z_0)}{P(R=1)}\frac{1-P(R=1|Z_0)}{1-P(R=1|Z_0)}g(Z;\theta)\right] \\
&= E\left[\frac{P(R=1|Z_0)}{P(R=1)}\frac{1-P(R=1|Z_0)}{1-P(R=1|Z_0)}g(Z;\theta)\right] \\
&= E\left[\frac{P(R=1|Z_0)}{P(R=1)}g(Z;\theta)\right] \\
&= E\left[\frac{R}{P(R=1)}g(Z;\theta)\right] \\
&= E[g(Z;\theta)|R=1]
\end{aligned}$$

where the first equality uses LIE and MAR (15), and the fourth equality uses MAR (15) and LIE (in that order). ■

Proof of Lemma 4.2: Consider (i) and note that the analog of (1) is satisfied because:

$$E\left[\psi(O;\theta^0,h_1^0(Z_0),h_2^0(Z_0))\right] = E\left[\frac{1-P(R=1|Z_0)}{1-P(R=1|Z_0)}\left\{E[g(Z;\theta^0)|Z_0]-E[g(Z;\theta^0)|Z_0]\right\}\right]+E[g(Z;\theta^0)] = 0$$

by LIE and, respectively, using the MAR condition (15) and the moment restrictions (13) for the first and second terms in the first equality. Similarly, the analogs of (2) and (3) are satisfied because:

$$\begin{aligned}
E\left[\psi(O;\theta^0,h_1^*(Z_0),h_2^0(Z_0))\right] &= E\left[\frac{1-P(R=1|Z_0)}{1-P(R=1|Z_0)}E[g(Z;\theta^0)|Z_0]\right] + E\left[\left(1-\frac{1-P(R=1|Z_0)}{1-P(R=1|Z_0)}\right)h_1^*(Z_0)\right] \\
&= 0, \\
E\left[\psi(O;\theta^0,h_1^0(Z_0),h_2^*(Z_0))\right] &= E\left[\frac{1-P(R=1|Z_0)}{1-h_2^*(Z_0)}\left\{E[g(Z;\theta^0)|Z_0]-E[g(Z;\theta^0)|Z_0]\right\}\right] + E[g(Z;\theta^0)] \\
&= 0.
\end{aligned}$$

This happens by LIE, MAR (15) and the moment restrictions (13) in the first equality in the first relationship; and by LIE and, respectively, using the MAR condition (15) and the moment restrictions (13) for the first and second terms in the first equality in the second relationship.

Now consider (ii) and note that the analog of (1) is satisfied because:

$$\begin{aligned} E\left[\psi(O;\theta^0,h_1^0(Z_0),h_2^0(Z_0))\right] &= E\left[\frac{P(R=1|Z_0)}{P(R=1)}\frac{1-P(R=1|Z_0)}{1-P(R=1|Z_0)}\left\{E[g(Z;\theta^0)|Z_0]-E[g(Z;\theta^0)|Z_0]\right\}\right] \\ &\quad + E\left[\frac{R}{P(R=1)}E[g(Z;\theta^0)|Z_0]\right] \\ &= 0 + E\left[\frac{P(R=1|Z_0)}{P(R=1)}E[g(Z;\theta^0)|Z_0]\right] \\ &= E\left[\frac{R}{P(R=1)}g(Z;\theta^0)\right] = E[g(Z;\theta^0)|R=1] = 0 \end{aligned}$$

where the second equality follows by LIE, the third one by MAR (15), the fourth one by LIE and, the last one by the moment restrictions (14). Similarly, the analogs of (2) and (3) are satisfied because:

$$\begin{aligned} E\left[\psi(O;\theta^0,h_1^*(Z_0),h_2^0(Z_0))\right] &= E\left[\frac{P(R=1|Z_0)}{P(R=1)}\frac{1-P(R=1|Z_0)}{1-P(R=1|Z_0)}E[g(Z;\theta^0)|Z_0]\right] \\ &\quad + E\left[\left(\frac{R}{P(R=1)}-\frac{P(R=1|Z_0)}{P(R=1)}\frac{1-P(R=1|Z_0)}{1-P(R=1|Z_0)}\right)h_1^*(Z_0)\right] \\ &= E\left[\frac{P(R=1|Z_0)}{P(R=1)}E[g(Z;\theta^0)|Z_0]\right] + E\left[\left(\frac{R}{P(R=1)}-\frac{P(R=1|Z_0)}{P(R=1)}\right)h_1^*(Z_0)\right] \\ &= E[g(Z;\theta^0)|R=1] = 0, \\ E\left[\psi(O;\theta^0,h_1^0(Z_0),h_2^*(Z_0))\right] &= E\left[\frac{h_2^*(Z_0)}{P(R=1)}\frac{1-P(R=1|Z_0)}{1-h_2^*(Z_0)}\left\{E[g(Z;\theta^0)|Z_0]-E[g(Z;\theta^0)|Z_0]\right\}\right] \\ &\quad + E\left[\frac{R}{P(R=1)}E[g(Z;\theta^0)|Z_0]\right] \\ &= 0 + E[g(Z;\theta^0)|R=1] = 0. \end{aligned}$$

The steps in the above derivations are self-explanatory since they use LIE, MAR (15) and the moment restrictions (13) and (14) exactly in the same way as in the other derivations so far. ■

Proof of Lemma 4.3: (i) Using the same arguments as in the previous proofs, we see that:

$$\begin{aligned} \Psi_\beta &= E\left[\left(1-\frac{1-h_1^0(Z_0)}{1-h_2^*(Z_0;\gamma^*)}\right)\left\{\frac{\partial}{\partial\beta'}h_1^*(Z_0;\beta^*)\right\}\right] = E\left[\frac{\Delta_2(Z_0)}{1-h_2^*(Z_0;\gamma^*)}\left\{\frac{\partial}{\partial\beta'}h_1^*(Z_0;\beta^*)\right\}\right] \\ \Psi_\gamma &= -E\left[\frac{1-h_2^0(Z_0)}{(1-h_2^*(Z_0;\gamma^*))^2}\left[\left(g(Z;\theta^0)-h_1^0(Z_0)\right)+\left(h_1^0(Z_0)-h_1^*(Z_0;\beta^*)\right)\right]\left\{\frac{\partial}{\partial\gamma'}h_2^*(Z_0;\gamma^*)\right\}\right] \\ &= -E\left[\frac{1-h_2^0(Z_0)}{(1-h_2^*(Z_0;\gamma^*))^2}\Delta_1(Z_0)\left\{\frac{\partial}{\partial\gamma'}h_2^*(Z_0;\gamma^*)\right\}\right] \end{aligned}$$

under MAR (15). The second equality in the first relationship follows by the definition of $\Delta_2(Z_0)$; and the second equality in the second relationship follows by LIE by recognizing that $h_1^0(Z_0) = E[g(Z;\theta^0)|Z_0]$, and also by the definition of $\Delta_1(Z_0)$.

(ii) Using the analogous arguments from the proof of (i), we see that:

$$\begin{aligned}
\Psi_\beta &= E\left[\left(\frac{h_2^0(Z_0)}{P(R=1)} - \frac{h_2^*(Z_0;\gamma^*)}{P(R=1)}\frac{1-h_2^0(Z_0)}{1-h_2^*(Z_0;\gamma^*)}\right)\left\{\frac{\partial}{\partial\beta'}h_1^*(Z_0;\beta^*)\right\}\right] \\
&= E\left[\frac{\Delta_2(Z_0)}{P(R=1)(1-h_2^*(Z_0;\gamma^*))}\left\{\frac{\partial}{\partial\beta'}h_1^*(Z_0;\beta^*)\right\}\right] \\
\Psi_\gamma &= E\left[\left(\frac{h_2^*(Z;\gamma^*)}{P(R=1)}\frac{1-h_2^0(Z_0)}{1-h_2^*(Z;\gamma^*)} - \frac{h_2^*(Z_0;\gamma^*)}{P(R=1)}\frac{1-h_2^0(Z_0)}{(1-h_2^*(Z_0;\gamma^*))^2}\right)\Delta_1(Z_0)\left\{\frac{\partial}{\partial\gamma'}h_2^*(Z_0;\gamma^*)\right\}\right] \\
&= -E\left[\frac{1-h_2^0(Z_0)}{P(R=1)}\frac{h_2^*(Z_0;\gamma^*)^2}{(1-h_2^*(Z_0;\gamma^*))^2}\Delta_1(Z_0)\left\{\frac{\partial}{\partial\gamma'}h_2^*(Z_0;\gamma^*)\right\}\right]. \ \blacksquare
\end{aligned}$$