Efficient estimation in the sub and full populations with monotonically missing at random data*

Saraswata Chaudhuri†

Updated version: June 4, 2017

Abstract

We consider estimation of a parameter defined by moment restrictions on a generic target population, and using monotonically missing at random data. Attrition is the common cause for such data, and in such a context of a R-period-long study, the generic target is the underlying population of the sample units leaving the study in contiguous periods $a, \ldots, b$ $(a \leq b = 1, \ldots, R)$. The novelty of our paper lies in the consideration and a unified treatment of such generic targets. Semiparametric efficiency bound and the efficient influence function for the parameter are obtained, which is the contribution of the paper. A consistent, asymptotically normal and efficient GMM estimator is presented. Our results nest the well-known special cases where $(a = 1, b = R)$ or $(R = 2, a = b = 1)$, but more generally provide new insights that are worth pointing out, and enable computationally simple but efficient estimation of a rich set of parameters that could be of interest in practice.

JEL Classification: C13; C14; C31.

Keywords: Attrition; Incomplete sub-samples; Multi-phase sampling; Semiparametric efficiency; Generalized method of moments.

*This is a followup to the paper "A Note on Efficiency Gains from Multiple Incomplete Sub-samples" unlike which the focus here is on unplanned missingness of the data. This leads to substantively different results in terms of the efficiency bounds for the parameters of interest and the contribution of various sample units toward efficient estimation.

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

1 Introduction

Estimation based on monotonically missing data has received special attention in the missing data literature; see e.g. the textbooks by Little and Rubin (2002), Tsiatis (2006), etc. Since the pioneering work of Robins et al. (1994, 1995), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Holcroft et al. (1997), etc., efficient estimation under such cases has generally been considered under the missing at random, i.e., a selection on observables, assumption. The focus of efficient estimation has conventionally been on parameters defined by the joint distribution of the underlying variables in the full population, i.e., the population from which the full set of sample units were randomly drawn.

However, selection on observables creates systematic difference in the sample units differing in terms of the variables observed in them. Consider, for example, attrition in a R-period-long study/survey where the decision to leave at the end of any period depends on the variables observed until the end of the said period (we will maintain a more precisely stated assumption). Then the two underlying (sub-) populations for the units leaving at the end of the a-th and b-th periods ($a \leq b \in \{1, \ldots, R\}$) are generally different, and it is interesting, at least for descriptive purposes, to study such sub-populations.

In this paper we consider efficient estimation of parameters defined by the joint distribution of the variables in such sub-populations or unions of contiguous sub-populations (e.g. $[a, \ldots, b]$), including the full population as a special case. Another important special case of our paper is Chen et al. (2008) who consider $R = 2$ and either $a = b = 1$ or $a = 1, b = 2$. We obtain the efficient influence functions and the efficiency bounds under a unified framework. This is the main contribution of the paper. An efficient and doubly-robust GMM estimator is presented. It has standard asymptotic properties.

Our results provide new insights on the usability of certain sample units for efficient estimation. Roughly speaking, in the context of the attrition example above, we find that if interest is in the sub-populations for the units who leave in periods $[a, \ldots, b]$, then the units who left before period a are not usable. This happens precisely because we: (i) do not allow any dimension-reduction in our selection on observables assumption, and (ii) do not assume that the conditional probabilities of missingness are known. (i) and (ii) lead to a framework that is more realistic for analyzing cases of attrition. By contrast, all sample units are usable for estimation when the conditional probabilities for missingness satisfy either an extreme dimension-reduction assumption pinned to the baseline (period 1) as in Chaudhuri (2014) (exception: $a = 2, b = R$), or are assumed known as in Chaudhuri (2017).

Of course, both dimension-reduction and known conditional probabilities are immaterial in the conventional case where the full population is of interest, i.e., when $a = 1, b = R$. However, they turn out to be of major importance when interest lies in the sub-populations, as is the case in this paper.

It should be noted that while we refer to the example of attrition to fix ideas since it is a common cause of monotonically missing data and has been carefully studied under the selection on observables assumption (even in economics: see Fitzgerald et al. (1996), Nicoletti (2006), etc.), our results and their implications apply generally to efficient estimation with any kind of monotonically missing at random data. Indeed, to illustrate the efficient estimation we consider a setup for our simulation study that reflects the ordered multistage dynamic decision model such as in Heckman et al. (2016)'s Figure 1, Ding and Lehrer (2010), etc. Of course, since we do not explicitly consider the estimation of dynamic treatment effects in this paper, we do not make the subtle distinction between attrition (subsequent outcomes not observed) and dropouts (subsequent outcomes observed) as in footnote 1 of Heckman et al. (1998), and rather proceed with the notion of attrition. Nevertheless, our result on efficient estimation should be applicable if one is interested in obtaining dynamic treatment effects following Lechner and Miquel (2010) (and the references therein), who closely relate their framework to the vast literature on dynamic treatment effects in the biostatistics and epidemiology literature.

Our paper proceeds as follows. Section 2 establishes and discusses the results. Section 3 is a Monte Carlo experiment studying the efficient estimator via simulations. Section 4 concludes. Proofs of the technical results are collected in Appendix A, and expository endnotes are collected in Appendix B.

2 Efficiency bound and efficient estimation

Let $Z := (Z_1', \ldots, Z_R')'$ where Z_r is a $d_r \times 1$ random vector and $\sum_{r=1}^{R} d_r$ is finite. Following Tsiatis (2006), let C be a random variable with support $\mathcal{C} := \{1, \ldots, R\}$ and $T_C(Z)$ a transformation defined as $T_r(Z) := (Z_1', \ldots, Z_r')'$ with dimension $(\sum_{s=1}^{r} d_s) \times 1$ for $r = 1, \ldots, R$. In the context of attrition in a R-period-long study/survey, Z_r are the variables specific to period r, while $T_r(Z)$ are all the variables observed for an unit that leaves at the end of period r, i.e., with $C = r$, for $r = 1, \ldots, R$. Formally, let $O := (C, T_C'(Z))'$ denote what is observed for an unit in the sample (study/survey).

We maintain a general selection on observables assumption:

$$P(C = r|Z) \equiv P(C = r|T_R(Z)) = P(C = r|T_r(Z)) \text{ for } r = 1, \ldots, R. \tag{1}$$

This is the MAR assumption [see e.g. Robins and Rotnitzky (1995), Tsiatis (2006)] in the sense of Rubin (1976). It is important to recognize that (1) implies that for any $r = 2, \ldots, R$:

$$P(C \geq r|Z) = 1 - \sum_{j=1}^{r-1} P(C = j|T_j(Z)) = 1 - \sum_{j=1}^{r-1} P(C = j|T_{r-1}(Z)) = P(C \geq r|T_{r-1}(Z)). \tag{2}$$

Under (1), we consider sub-populations defined by $(a \leq C \leq b)$ for $a, b \in \{1, \ldots, R\}$ (i.e., in $\mathcal{C}$). If $a = b = r$ then, in the context of attrition, this is the hypothetical sub-population from which the units who left at the end of period r can be viewed as being randomly drawn. If $a < b$, then this is the sub-population for the units who left in the periods $a, a+1, \ldots, b$. Convention: those who stay until the end, leave at the end of period R. Thus, if $a = 1$ and $b = R$, then this is the full population.

The underlying distributions of Z, denote them by $F_{Z|(a \leq C \leq b)}(z)$, in these sub-populations are typically different, and we define the parameter of interest in (3) as a finite dimensional feature of $F_{Z|(a \leq C \leq b)}(z)$ as follows. Consider a function $m(Z; \beta) : \text{Support}(Z) \times \mathcal{B} \mapsto \mathbb{R}^{d_m}$, $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$ where $d_\beta \leq d_m$. For a given $a, b \in \{1, \ldots, R\}$ $(a \leq b)$, let the parameter value of interest β^0 be defined as:

$$E[m(Z; \beta) | a \leq C \leq b] = 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta^0. \tag{3}$$

Our goal is the efficient estimation of β^0. To define efficient we first obtain the efficiency bound for β^0 by maintaining the following standard assumptions [see e.g. Chaudhuri (2017) for discussion].

Assumption A

(A1) The observed sample units $\{O_i := (C_i, T'_{C_i}(Z_i))'\}_{i=1}^n$ are i.i.d. copies of $O := (C, T'_C(Z))'$.

(A2) $(P(C = r | T_R(Z)))_{r=1}^{R-1} > 0$ and $P(C = R | T_R(Z)) > \underline{p}$ almost surely in $T_R(Z)$ for a fixed $\underline{p} \in (0, 1)$.

(A3) $M := \frac{\partial}{\partial \beta'} E\left[m(Z; \beta^0) | a \leq C \leq b\right]$ is a $d_m \times d_\beta$ finite matrix of full column rank.

2.1 Efficiency bound for estimation of β^0:

Proposition 1 is our main result. It provides the efficiency bound for β^0 along with the efficient influence function. The key quantity (and the novel contribution) for this purpose is defined as follows:

$$\begin{aligned}
\varphi(O; \beta) \;:=\; & \sum_{r=b+1}^{R} \frac{I(C \geq r)}{P(C \geq r | T_r)} \frac{P(a \leq C \leq b | T_b)}{P(a \leq C \leq b)} \left(E[m(T_R; \beta) | T_r] - E[m(T_R; \beta) | T_{r-1}]\right) \\
& + \sum_{r=a+1}^{b} \frac{I(C \geq r)}{P(C \geq r | T_r)} \frac{P(a \leq C \leq r-1 | T_{r-1})}{P(a \leq C \leq b)} \left(E[m(T_R; \beta) | T_r] - E[m(T_R; \beta) | T_{r-1}]\right) \\
& + \sum_{r=a}^{b} \frac{I(C = r)}{P(a \leq C \leq b)} E[m(T_R; \beta) | T_r].
\end{aligned} \tag{4}$$

In (4) and onward we use the following notation unless confusing. $I(C \geq R) \equiv I(C = R)$, $T_r \equiv T_r(Z)$ (note that $T_R \equiv Z$), and $m(Z; \beta) \equiv E[m(T_R; \beta) | T_R]$. If $b = R$, then the indices (e.g. in summation) running from $b + 1$ to R are void. If $a = b$ then the indices running from $a + 1$ to b are void, while those running from a to b contain only one term and it corresponds to a (equivalently b).

Proposition 1 *Let* (1), (3) *and assumption A hold. Let the* $d_m \times d_m$ *matrix* $V := Var(\varphi(O;\beta^0))$ *be finite and positive definite. Then the asymptotic variance lower bound for* $\sqrt{n}(\widehat{\beta} - \beta^0)$ *of any regular estimator* $\widehat{\beta}$ *for* β^0 *is given by* $\Omega := (M'V^{-1}M)^{-1}$. *An estimator* $\widehat{\beta}$ *whose asymptotic variance equals* Ω *has the asymptotically linear representation*

$$\sqrt{n}(\widehat{\beta} - \beta^0) = -\Omega M'V^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(O_i;\beta^0) + o_p(1). \blacksquare$$

Remark 1: The following special cases of this general result are already well-known in the literature. Taking $R = 2$ and $a = b = 1$ gives the result of Case (1) in Theorem 1 of Chen et al. (2008):

$$\varphi(O;\beta) = \frac{I(C=2)}{P(C=2|T_1)}\frac{P(C=1|T_1)}{P(C=1)}(m(T_2;\beta) - E[m(T_2;\beta)|T_1]) + \frac{I(C=1)}{P(C=1)}E[m(T_2;\beta)|T_1].$$

Taking $R = 2$, $a = 1, b = 2$ gives the result for Case (2) in Theorem 1 of Chen et al. (2008):

$$\varphi(O;\beta) = \frac{I(C=2)}{P(C=2|T_1)}\frac{P(C=1|T_1)}{P(C=1)}(m(T_2;\beta) - E[m(T_2;\beta)|T_1]) + E[m(T_2;\beta)|T_1].$$

Also see Robins et al. (1994). Taking a general R and setting $a = 1, b = R$ gives [see Appendix B.1]

$$\varphi(O;\beta) = \sum_{r=1}^{R}\frac{I(C \geq r)}{P(C \geq r|T_r)}\left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) + E[m(T_R;\beta)|T_1],$$

the well-known result for the full population under monotone missing at random data as in Robins and Rotnitzky (1995); Rotnitzky and Robins (1995), Holcroft et al. (1997). Also see Tsiatis (2006).[1]

Remark 2: If focus is on an individual sub-population, i.e., $a = b$, then for a general R, $\varphi(O;\beta)$ is

$$\sum_{r=a+1}^{R}\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(C=a|T_a)}{P(C=a)}\left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) + \frac{I(C=a)}{P(C=a)}E[m(T_R;\beta)|T_a]$$

by Proposition 1. Thus, $\varphi(O;\beta) = I(C = R)m(T_R;\beta)/P(C = R)$ in the special case when $a = b = R$.

Remark 3: Implicit in (1) and hence in Proposition 1 is a restriction on dimension-reduction: We do not allow $P(C = r|Z) = P(C = r|T_s)$ for $s < r = 2, \ldots, R-1$ since we consider these as additional information and not merely special cases of (1). As a result, crucially, the terms involving $E[m(Z;\beta)|T_r]$ for $r < a$ do not appear in the expression for $\varphi(O;\beta)$ in (4). On the other hand, under an extreme and uniform (in r) dimension-reduction assumption[2] $P(C = r|Z) = P(C = r|T_1)$

[1] While this particular expression for $\varphi(O;\beta)$ under $a = 1, b = R$ and a general R adheres to that in Proposition 6 of Chaudhuri (2017), it is straightforward to see that it boils down to the equivalent expression stated in the original references. The special case of $a = b = 1$ for a general R is also considered in Proposition 6 of Chaudhuri (2017).

[2] With $R = 3$, this reflects a common identification assumption in mediation analysis; see e.g. Imai et al. (2010). With a general R, this assumption has also been used in one of the attrition analyses in Fitzgerald et al. (1996).

that pins these conditional probabilities to the baseline (period 1), Proposition 1 in Chaudhuri (2014) establishes that the corresponding $\varphi(O;\beta)$ is (writing $E[m(Z;\beta)|T_r]$ as $m^\dagger(T_r;\beta)$ for brevity):

$$\varphi^\dagger_{[a,b]}(O;\beta) = \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r)} \frac{P(a \leq C \leq b|T_1)}{P(a \leq C \leq b)} (m^\dagger(T_r;\beta) - m^\dagger(T_{r-1};\beta)) + \frac{I(a \leq C \leq b)}{P(a \leq C \leq b)} m^\dagger(T_1;\beta), \quad (5)$$

i.e., all $E[m(Z;\beta)|T_r]$ $(r = 1,\ldots,R)$ are usable irrespective of a, b except if $a = 2, b = R$. It is also interesting to note that the terms $E[m(Z;\beta)|T_r]$ $(r = a,\ldots,b)$, i.e., those that come from the sub-populations of interest, contribute differently in this dimension-reduction case than in (the last two lines of) (4) unless $a = 1, b = R$, i.e., unless interest lies in the full population. Thus, the results in Chaudhuri (2014) are not special cases of our Proposition 1, as stated.

Remark 4: The observation from Remark 3 is in direct contrast to the case of planned missingness in the data, i.e., where $P(C = r|Z) = P(C = r|T_r)$ and the latter is known, under which Proposition 1 in Chaudhuri (2017) establishes that

$$\begin{aligned} \varphi_{[a,b]}(O;\beta)^* &= \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r)} (m^*(T_r;\beta) - m^*(T_{r-1};\beta)) + m^*(T_1;\beta), \quad (6) \\ \text{where } m^*(T_r;\beta) &:= E\left[\frac{P(a \leq C \leq b|T_R)}{P(a \leq C \leq b)} m(T_R;\beta) \middle| T_r \right]. \end{aligned}$$

Thus under planned missingness, nesting of the assumptions $\{P(C = r|Z) = P(C = r|T_s) : s \leq r\}$ is automatic, and all the $E[m(Z;\beta)|T_r]$ $(r = 1,\ldots,R)$ are always usable irrespective of whether dimension-reduction is allowed or not. As a result, while Chaudhuri (2017) also contrasts unplanned (with $P(C = r|Z) = P(C = r|T_1)$) and planned missingness, the contrast in our paper is much more prominent because we maintain $P(C = r|Z) = P(C = r|T_r)$ and do not allow for dimension-reduction.

Remark 5: Now consider a similarity with Chaudhuri (2014, 2017). It follows from (5) and (6) that

$$\varphi_{[a,b]}(O;\beta)^\dagger = \sum_{j=a}^{b} \frac{P(C = j)}{P(a \leq C \leq b)} \varphi_{[j,j]}(O;\beta)^\dagger \quad \text{and} \quad \varphi_{[a,b]}(O;\beta)^* = \sum_{j=a}^{b} \frac{P(C = j)}{P(a \leq C \leq b)} \varphi_{[j,j]}(O;\beta)^*,$$

which reflect the standard relationship among the moment restrictions in (3) that

$$E[m(Z;\beta)|a \leq C \leq b] = \sum_{j=a}^{b} \frac{P(C = j)}{P(a \leq C \leq b)} E[m(Z;\beta)|C = j].$$

It turns out that same also holds in our paper, i.e., with the $\varphi(O;\beta)$ defined in (4) [see Appendix B.2]. This is intuitive, and it presents a way of combining the efficient estimators for the individual sub-populations to obtain the efficient estimator for any (contiguous) union of such sub-populations.

Remark 5: $\varphi(O;\beta)$ in (4) is doubly-robust in the sense of Scharfstein et al. (1999) [see Appendix B.3]. This leads to desirable properties for the estimator of β^0 based on the estimating function $\varphi(O;\beta)$ when the unknown nuisance functions, i.e., the conditional probabilities and expectations, are estimated parametrically or nonparametrically [see Robins et al. (1994), Robins and Ritov (1997), Bang and Robins (2005), Tan (2007), Cao et al. (2009), Rothe and Firpo (2016), etc.].

2.2 Efficient estimator of β, and estimation of its asymptotic variance:

Now we obtain the efficient GMM estimator of β^0 and describe its asymptotic properties by appealing to existing results. Since these are not our contribution, we only provide a working and informal description for practitioners by citing, for brevity, a rather incomplete set of relevant references.

Informed by Proposition 1, we treat $\varphi(O;\beta)$ as the moment vector. $\varphi(O;\beta)$ contains the unknown nuisance parameters — (i) the conditional probabilities $P(a \leq C \leq r|T_r) = \sum_{j=a}^{r} P(C=j|T_j)$ (by (1)) for $r = a,\ldots,b-1$ and $P(C \geq r|T_r)$ for $r = a+1,\ldots,R$, and (ii) the conditional expectations $E[m(T_R;\beta)|T_r]$ for $r = a,\ldots,R-1$ — and they need to be replaced by either parametric or nonparametric estimators for the GMM estimation of β^0.[3]

For notational brevity denote generically the parameters in (i) and (ii) by $p(.)$ and $q(.;\beta)$ respectively, and their estimators by $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ respectively. Note by inspecting (4) that estimation of the individual elements of $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ are only required when their arguments, i.e., the conditioning variables in the definition of the concerned individual elements of $p(.)$ and $q(.;\beta)$, are observed. Define

$$g(O_i,p(.),q(.;\beta);\beta) := \varphi(O_i;\beta), \text{ and } \bar{g}_n(\beta,p(.),q(.;\beta)) = \frac{1}{n}\sum_{i=1}^{n} g(O_i,p(.),q(.;\beta);\beta).$$

Given $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$, and a $d_m \times d_m$ weighting matrix W_n, the GMM estimator $\widehat{\beta}_n(W_n)$ is defined as:

$$\widehat{\beta}_n(W_n) \approx \arg\min_{\beta\in\mathcal{B}} \;\; \bar{g}_n(\beta,\widehat{p}(.),\widehat{q}(.;\beta))' \; W_n \; \bar{g}_n(\beta,\widehat{p}(.),\widehat{q}(.;\beta)).$$

Straightforward extension of Holcroft et al. (1997), Tsiatis (2006), etc. give consistency and asymptotic normality of $\widehat{\beta}_n(W_n)$ if $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ are parametrically estimated.[4] Remark 5 following Proposition 1 indicates that $\widehat{\beta}_n(W_n)$ is consistent for β^0 itself if the parametric model for at least either $p(.)$ or $q(.;\beta)$ is correctly specified. $\widehat{\beta}_n(W_n)$ is asymptotically efficient with asymptotic variance Ω if both parametric models are correctly specified and if the weighting matrix is efficient (i.e., if $W_n \xrightarrow{P} V^{-1}$).

[3] The other unknown nuisance parameter $1/P(a \leq C \leq b)$ appears as a common scalar-multiple in $\varphi(O;\beta)$, and hence is innocuous under assumption (A2). We ignore it in our presentation while noting that if needed, as would be the case when estimating variance, it can be replaced by an estimator $n/\sum_{i=1}^{n} I(a \leq C_i \leq b)$ without affecting anything else.

[4] Also see Theorems 3.1 and 3.2 in Chaudhuri and Min (2012), who use similar notation as in this paper.

When both these parametric models are correctly specified, the asymptotic variance of $\widehat{\beta}_n(W_n)$ can be computed in the standard way (adjusting for whether W_n is efficient) by averaging the outer product of $g(O_i, \widehat{p}(.), \widehat{q}(.; \widehat{\beta}_n(W_n)); \widehat{\beta}_n(W_n))$ and without considering that $\widehat{p}(.)$ and $\widehat{q}(.; \beta)$ are actually estimated. (The jacobian M can be estimated by numerical methods.) Otherwise, however, one should take the estimation of $\widehat{p}(.)$ and $\widehat{q}(.; \beta)$ into account (as in Theorem 6.1 of Newey and McFadden (1994)), and instead use the routine modification described in Section 6.3 of Newey and McFadden (1994).

If instead nonparametric estimators $\widehat{p}(.)$ and $\widehat{q}(.; \beta)$ are used, which is practical when the sample size is sufficiently large relative to the dimensions of $\{T_r : r = 1, \ldots, R\}$, consistency of $\widehat{\beta}_n(W_n)$ (for β^0) and its asymptotic normality follow under the conditions of Theorem 1 of Chen et al. (2003) and Theorem 4.1 of Chen (2007) respectively. Under the latter's conditions, estimation of $\widehat{p}(.)$ and $\widehat{q}(.; \beta)$ has no effect on the asymptotic variance of $\widehat{\beta}_n(W_n)$. These results have recently been established by Rothe and Firpo (2016) under weaker conditions exploiting the double-robustness noted in Remark 5. $W_n \xrightarrow{P} V^{-1}$ gives efficiency of $\widehat{\beta}_n(W_n)$. Conditions similar to Theorems 6 and 7 of Cattaneo (2010) give consistency of the asymptotic variance estimator obtained based on averaging the outer product of $g(O_i, \widehat{p}(.), \widehat{q}(.; \widehat{\beta}_n(W_n)); \widehat{\beta}_n(W_n))$ and using, if needed, a numerical derivative-based estimate of M.

3 Monte Carlo Experiment

3.1 Simulation Design

As an illustration with the R-period-long study/survey running example, we now concretely consider a setup that reflects an individual's decision to stay or leave a program (school, smoking cessation, etc.) dynamically over periods. We adhere to the selection on observables assumption in (1), and in this sense our setup is closer to that of Ding and Lehrer (2010), Lechner and Miquel (2010), etc. rather than Heckman et al. (2016) (and the similar references therein) who consider selection on unobservables.

The design and the data generating process (DGP) considered are as follows. Let

$$\begin{aligned} Y_t &= \frac{1}{2}Y_{t-1} + \frac{1}{4}Y_{t-2} + \frac{1}{4}X_t + e_t \\ X_t &= X_{t-1} + v_t \end{aligned} \tag{7}$$

for $t = 1, \ldots, T$. Take $t = -1, 0$ as the initial conditions. With the index r running from $1, \ldots, R$ and defining $R := T$, one could think of Y_r as the (potential) outcome from staying until the end of the r-th period in the program, and take X_r's as the exogenous variables in the successive periods.

Let the decision to leave the program at the end of the r-th period, conditional on staying until

then, depend on how the individual's expectation of their outcome compare with the actual outcome:

$$\begin{aligned} Y_r^* &= Y_{r-1} + u_r \\ I(C = r) &= I(Y_r^* < Y_r) \prod_{j=1}^{r-1} I(Y_j^* \geq Y_j) \end{aligned} \tag{8}$$

for $r = 1, \ldots, R-1$ where Y_r^* can be thought of as the individual's expectation, and she/he leaves at the end of r resulting in $C = r$ provided that the expectation falls short of the actual outcome, i.e., $Y_r^* < Y_r$. Our setup therefore updates the expectation of the (future potential) outcome [see (8)] faster than the outcome itself [see (7)], and thus leads to a perverse scenario under which, roughly speaking, those with lower outcome continue with the program whereas those with higher outcome leave. Naturally,

$$I(C = R) = 1 - \sum_{r=1}^{R-1} I(C = r),$$

i.e., the decision to continue until the end of the last period does not depend on the outcome or the expectation of the outcome for the last period (conditional on what has already been observed). Indeed, in setups driven by the selection on observables assumption, and as evident from (1) and its derivative (2), the last period's outcome (or its expecation/anticipation) does not influence the dropout behavior in any period. (Empirical relevance of this phenomenon depends on the context of the application, and we abstract from any such discussion in this simulation study.)

We take $R = 3$. As usual, the analyst observes C but not Y_r^*. In terms of our notation for observability, this means: $Z_1 = (Y_{-1}, Y_0, Y_1, X_{-1}, X_0, X_1)'$, $Z_2 = (Y_2, X_2)'$ and $Z_3 = (Y_3, X_3)'$.

Model errors: We assume that e_t and v_t are independent and also i.i.d. $N(0,1)$ for all t. We also assume that u_r is i.i.d. $N(0, (2.5)^2)$, and is independent of e_t and v_t for all t, r.[5] This results in roughly 50% of the individuals with $C = 1$, 26% with $C = 2$, and 24% with $C = 3$.

To define β^0, we take the moment function in (3) as $m(Z; \beta) = Y_3 - \beta$, and accordingly consider the six different target (sub-) populations characterized by $(1 \leq C \leq 3)$, $(C = 1)$, $(C = 2)$, $(C = 3)$, $(1 \leq C \leq 2)$ and $(2 \leq C \leq 3)$ respectively, giving six different parameter of interest. For concreteness, the first parameter denotes the average period 3 outcome for all the units in the study, the second one is the average period 3 outcome of those who left after the first period had they continued until the end of period 3, and so on for the other parameters, while the last one is the average period 3

[5]The larger variance for u_r is assumed in an effort to avoid the adverse consequences of the so-called limited overlap problem [recall assumption A2; see Khan and Tamer (2010), Chaudhuri and Hill (2016), etc.]. This effort was to an extent successful since, although Figure 3 indicates that the estimated $P(C = 3|T_3)$ (and, for completeness, also $P(C = 2|T_2)$ and $P(C = 1|T_1)$) could be arbitrarily close to zero with large probability, the discussion in Appendix B.5 indicates that this matters for our results in a serious way only for very small size ($n = 100$) and that too, quite infrequently. Further support of this ad-hoc claim can be found from the extreme values of the estimators reported in Tables 1 and 2.

outcome of those who left at the end of period 2 or 3 had those who left after period 2 continued until the end of period 3. For simplicity we compute the true value of these parameters numerically by generating data from the above DGP with sample size 10 million, estimating the mean of Y_3 for each (sub)-population (feasible with the generated data), and then averaging each mean over 10,000 Monte Carlo trials. Accordingly the six different β^0's are 1, 1.1709, .9617, .6858, 1.0994 and .8291 respectively. There is error involved in this approximation but as evident from Table 4 and Figure 1 in Appendix B.4, the error is rather of a small order to seriously affect our subsequent analysis.[6]

The study of the efficient estimation under this setup is now conducted for sample sizes $n = 100, 200, 500$ and 1000. All results reported below are based on 10,000 Monte Carlo trials.

3.2 Simulation Results

Efficient GMM estimation is conducted following Section 2.2 but ignoring the weighting matrix W_n since $d_m = d_\beta = 1$. The nuisance parameters, i.e., the conditional expectations (of Y_3) and conditional probabilities are estimated by least squares and probit regressions respectively. For both we specify the regression function as linear in the conditioning variables (no interactions).

Note that the parameter β for the sub-population $C = 3$ is not interesting in the sense that the complete case (GMM) estimator $\sum_{i=1}^n I(C_i = 3)Y_{3i}/\sum_{j=1}^n I(C_j = 3)$ already estimates it consistently, and by virtue of Proposition 1 (Remark 1), efficiently. Therefore, we do not consider this parameter further. (See Table 3 for the performance of this estimator and the t-test based on it for all parameters.)

For the other five parameters, the results for the efficient GMM estimator is presented in Table 1. For each β, and for each of the 10,000 Monte Carlo trials we obtain the estimate, subtract it from the corresponding "true value" in Table 4, and report the mean, median, minimum and maximum of this difference (as Bias, MedBias, MinBias, MaxBias) over the 10,00 trials. The mean of the absolute value of this difference is reported as ABias, the Monte Carlo standard deviation as Std, the mean of the standard deviation based on the asymptotic variance formula as AStd, and the empirical size of the t-test based on Std and the asymptotic variance formula as Size and ASize respectively. The efficient estimator performs quite well in all aspects even for relatively smaller sample sizes.

To put this performance into context, although it might not be fair, we report in Table 2 the same (except AStd and Asize) for the Horvitz and Thompson (1952)-type IPW estimator

$$\sum_{i=1}^n \frac{I(C_i = 3)}{\widehat{P}(C = 3|T_{3i})} \frac{\widehat{P}(a \leq C_i \leq b|T_{bi})}{\sum_{j=1}^n I(a \leq C_j \leq b)} Y_{3i}.$$

[6]For further appreciation of the selection involved, we also report in Figure 2 [see Appendix.4] the kernel estimator of the mean of Y_1, Y_2 and Y_3 conditional on the estimated $P(C = 1|T_1)$, $P(C = 2|T_2)$ and $P(C = 3|T_3)$ respectively, exactly under the same DGP and with sample size 10 million but based on the results from the first Monte Carlo trial.

[Also see Appendix B.5.] Naturally, this estimator also does not suffer from Bias by virtue of (1). Its efficiency loss with respect to the efficient estimator is evident by comparing the corresponding Std, and this is expected from Proposition 1. The loss can vary from 8.5% to 50% (with the median and average case both being 27%) across the numbers reported under Std in Tables 1 and 2. Interestingly (and perhaps often overlooked), while Proposition 1 is only about the (asymptotic) variance, comparing the other characteristics in Tables 1 and 2 clearly demonstrates that the efficient estimator is much less dispersed than the IPW estimator, in all aspects and for all sample sizes considered in our paper.

4 Conclusion

This paper establishes the efficiency bound for parameters that are features of the joint distribution of a set of variables in sub-populations and a broad class of unions of sub-populations defined by the observability of the concerned variables. The novelty of the paper lies in the consideration of the sub-populations. It is maintained that the observability of variables follows a monotone pattern and the selection involved is the so-called selection on observables. This result should be useful, for example, for a variety of analysis using data sets that suffer from attrition.

The efficient influence function, whose variance gives the efficiency bound, also displays important characteristics that are not reflected in the analysis of the special cases considered so far in the literature. We note these characteristics and describe their novel and interesting implications. On the other hand, we also note that our result generalizes the well-known existing results in the literature. Furthermore, the double-robustness of the moment vector, that we eventually use for estimation, leads to desirable properties of the simple GMM estimator described in the paper. The simulation experiment also confirms good properties of this estimator in samples of relatively small size.

References

Bang, H. and Robins, J. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61:962–972.

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. (2014). A Note on Efficiency Gains from Multiple Incomplete Subsamples. Mimeo.

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

Chaudhuri, S. and Hill, J. B. (2016). Heavy Tail Robust Estimation and Inference for Average Treatment Effect. Technical report, University of North Carolina, Chapel Hill.

Chaudhuri, S. and Min, H. (2012). Doubly-Robust Parametric Estimation in Moment Conditions Models with Missing Data. Mimeo.

Chen, X. (2007). Large Sample Sieve Estimation Of Semi-Nonparametric Models. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume VIB, chapter 76, pages 5550–5632. Elsevier Science Publisher.

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.

Ding, W. and Lehrer, S. (2010). Estimating Treatment Effects from Contaminated Multiperiod Education Experiments: The Dynamic Impacts of Class Size Reductions. *Review of Economics and Statistics*, 92:31–42.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1996). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. In *Working Paper Series*. NBER.

Heckman, J., Humphries, J., and Veramendi, G. (2016). Dynamic treatment effects. *Jorunal of Econometrics*, 191:276–292.

Heckman, J., Smith, J., and Taber, C. (1998). Accounting for Dropouts in Evaluations of Social Programs. *Review of Economics and Statistics*, LXXX:1–14.

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.

Horvitz, D. and Thompson, D. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of American Statistical Association*, 47:663–685.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25:51–71.

Khan, S. and Tamer, E. (2010). Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica*, 78:2021–2042.

Lechner, M. and Miquel, R. (2010). Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions. *Empirical Economics*, 39:111–137.

Little, R, J. A. and Rubin, D. D. (2002). *Statistical Analysis with Missing Data*. Wiley - Interscience.

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.

Nicoletti, C. (2006). Nonresponse in dynamic panel data models. *Journal of Econometrics*, 132:461–489.

Robins, J. and Ritov, Y. (1997). Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theroy for Semi-Parametric Models. *Statistics in Medicine*, 16:285–319.

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.

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.

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

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

5 Tables for the main text

Target for β	$n = 100$								
	Bias	MedBias	MinBias	MaxBias	ABias	Std	AStd	Size	ASize
$1 \leq C \leq 3$	-.0037	-.0102	-1.4862	2.5896	.2544	.3254	.2512	5.3	13.8
$C = 1$	-.0041	-.0149	-2.6509	3.3278	.3177	.4092	.3336	5.1	11.8
$C = 2$	-.0068	-.0132	-2.4474	2.2725	.3484	.4501	.3986	5.3	8.2
$1 \leq C \leq 2$	-.0056	-.0147	-1.9473	2.8004	.2896	.3727	.2893	5.1	13.4
$2 \leq C \leq 3$	-.0034	-.0050	-1.6282	1.4262	.2492	.3187	.2803	5.1	8.5
	$n = 200$								
	Bias	MedBias	MinBias	MaxBias	ABias	Std	AStd	Size	ASize
$1 \leq C \leq 3$	-.0005	-.0042	-.9553	1.0449	.1650	.2078	.1884	5.1	8.1
$C = 1$	-.0009	-.0077	-1.7499	1.6389	.2015	.2551	.2436	5.0	6.4
$C = 2$	.0006	-.0021	-1.7145	3.1583	.2211	.2832	.2816	4.9	4.9
$1 \leq C \leq 2$	-.0009	-.0065	-1.2391	1.3064	.1836	.2321	.2139	4.9	7.5
$2 \leq C \leq 3$	.0001	-.0013	-.9836	1.7208	.1659	.2101	.2045	5.0	5.7
	$n = 500$								
	Bias	MedBias	MinBias	MaxBias	ABias	Std	AStd	Size	ASize
$1 \leq C \leq 3$	-.0020	-.0039	-.6830	.4939	.1009	.1269	.1208	5.0	6.7
$C = 1$	-.0024	-.0045	-.5654	.6081	.1225	.1540	.1546	5.0	5.3
$C = 2$	-.0015	-.0026	-1.6263	.7290	.1313	.1668	.1753	5.1	4.1
$1 \leq C \leq 2$	-.0023	-.0040	-.8461	.5669	.1103	.1392	.1358	5.2	6.0
$2 \leq C \leq 3$	-.0014	-.0016	-.9141	.4623	.1024	.1288	.1299	4.8	4.8
	$n = 1000$								
	Bias	MedBias	MinBias	MaxBias	ABias	Std	AStd	Size	ASize
$1 \leq C \leq 3$	.0007	.0005	-.3095	.3960	.0702	.0880	.0858	5.0	6.0
$C = 1$	.0007	-.0001	-.3861	.4707	.0849	.1066	.1096	5.1	4.4
$C = 2$	.0023	.0012	-.4008	.4773	.0902	.1126	.1232	4.7	3.0
$1 \leq C \leq 2$	.0011	.0009	-.3326	.4204	.0763	.0957	.0961	5.3	5.1
$2 \leq C \leq 3$	.0009	.0009	-.3358	.4034	.0712	.0892	.0920	5.1	4.4

Table 1: All results for the efficient estimator are reported based on 10,000 Monte Carlo trials. Bias, MedBias, MinBias, MaxBias and ABias stand for the mean, median, minimum, maximum and mean absolute value respectively of the difference between the efficient estimator and the corresponding β^0. Std and AStd stand for the standard deviation based on Monte Carlo and the (average from the) asymptotic variance formula respectively. Size and ASize stand for the empirical size of the asymptotic 5% two-sided t-test using the Monte Carlo and asymptotic standard deviation respectively.

Target for β	$n = 100$						
	Bias	MedBias	MinBias	MaxBias	ABias	Std	Size
$1 \leq C \leq 3$	-.0449	-.0823	-1.5733	7.9008	.3080	.4278	3.7
$C = 1$	-.0593	-.1386	-2.2333	16.3040	.4159	.6143	3.2
$C = 2$	-.0579	-.1168	-4.5318	6.3902	.3987	.5611	4.1
$1 \leq C \leq 2$	-.0589	-.1184	-2.0413	10.6195	.3653	.5187	3.5
$2 \leq C \leq 3$	-.0305	-.0469	-1.9653	3.8259	.2743	.3678	4.2
	$n = 200$						
	Bias	MedBias	MinBias	MaxBias	ABias	Std	Size
$1 \leq C \leq 3$	-.0143	-.0338	-1.2238	3.4379	.2006	.2701	4.0
$C = 1$	-.0209	-.0606	-1.1614	4.6535	.2697	.3698	4.1
$C = 2$	-.0141	-.0421	-2.0402	10.4536	.2611	.3839	3.4
$1 \leq C \leq 2$	-.0189	-.0505	-1.4550	4.4394	.2345	.3223	3.8
$2 \leq C \leq 3$	-.0075	-.0174	-1.2727	6.0802	.1840	.2527	3.6
	$n = 500$						
	Bias	MedBias	MinBias	MaxBias	ABias	Std	Size
$1 \leq C \leq 3$	-.0044	-.0088	-.5559	1.0035	.1192	.1527	4.6
$C = 1$	-.0060	-.0213	-.6293	1.3960	.1610	.2084	4.5
$C = 2$	-.0041	-.0180	-.9816	2.4238	.1542	.2016	4.7
$1 \leq C \leq 2$	-.0056	-.0152	-.6295	1.2874	.1377	.1779	4.7
$2 \leq C \leq 3$	-.0027	-.0061	-.5498	1.3350	.1116	.1417	4.8
	$n = 1000$						
	Bias	MedBias	MinBias	MaxBias	ABias	Std	Size
$1 \leq C \leq 3$	-.0004	-.0030	-.3264	.6055	.0820	.1040	4.9
$C = 1$	-.0011	-.0094	-.4356	1.0855	.1112	.1426	4.6
$C = 2$	.0012	-.0040	-.4346	1.3595	.1045	.1335	4.7
$1 \leq C \leq 2$	-.0004	-.0063	-.3835	.8609	.0941	.1201	4.9
$2 \leq C \leq 3$	.0003	-.0016	-.3567	.6662	.0768	.0968	5.2

Table 2: All results for the IPW estimator are reported based on 10,000 Monte Carlo trials. Bias, MedBias, MinBias, MaxBias, IQRBias and ABias stand for the mean, median, minimum, maximum and mean absolute value respectively of the difference between the IPW estimator and the corresponding β^0. Std stands for the standard deviation based on Monte Carlo. Size stands for the empirical size of the asymptotic 5% two-sided t-test using the Monte Carlo standard deviation.

Target	$n = 100$		$n = 200$		$n = 500$		$n = 1000$	
Population	Std = .3140		Std = .2207		Std = .1386		Std = .0994	
for β	Bias	Size	Bias	Size	Bias	Size	Bias	Size
$1 \leq C \leq 3$	-.3148	16.8	-.3151	29.8	-.3155	62.5	-.3148	88.7
$C = 1$	-.4857	33.5	-.4860	59.1	-.4864	93.8	-.4857	99.8
$C = 2$	-.2765	14.1	-.2768	24.1	-.2772	51.9	-.2765	79.7
$C = 3$	-.0006	5.3	-.0009	5.2	-.0013	5.2	-.0006	4.8
$1 \leq C \leq 2$	-.4142	25.7	-.4145	46.8	-.4149	84.9	-.4142	98.7
$2 \leq C \leq 3$	-.1439	7.7	-.1442	9.9	-.1446	18.1	-.1439	30.8

Table 3: All results for the complete-case estimator are reported based on 10,000 Monte Carlo trials. Bias stands for the mean bias. The only representative population for the complete-case estimator is $C = 3$. Std stands for the Monte Carlo standard deviation, and since the estimator is identical for all (sub-) populations (and hence the bias for the other targets), there is only a single Std reported for each sample size n. Size stands for the empirical size of the asymptotic 5% two-sided t-test.

Appendix A: Proof of the main results

Notation: f and F denote the density and distribution functions, and the concerned random variables are specified inside parentheses. Their conditional counterparts are denoted accordingly using the obvious notation. $L_0^2(F)$ denotes the space of mean-zero, square integrable functions with respect to F.

Proof of Proposition 1: We follow the three steps in Chen et al. (2008). Step 1 characterizes the tangent set for all regular parametric sub-models satisfying the semiparametric assumptions on the observed data. Step 2 obtains the efficient influence function for a given rotation of $m(Z;\beta)$. Step 3 obtains the optimal rotation and, thereby, gives the efficiency bound as the expectation of the outer product of the efficient influence function.

STEP - 1: Consider a regular parametric sub-model indexed by a parameter θ for the distribution of the observed data $O = (C', T_C'(Z))'$. The log of the distribution, in terms of $(C, Z')'$, is

$$\log f_\theta(O) = \log f_\theta(Z_1) + \sum_{r=2}^{R} I(C \geq r) \log f_\theta(Z_r|Z_1, \ldots, Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \log P_\theta(C = r|Z_1, \ldots, Z_r).$$

Let θ_0 be the unique value of θ such that $f_{\theta_0}(O)$ equals the true $f(O)$, and accordingly for all the quantities. The score function with respect to θ can be written in terms of $(C, Z')'$ as

$$S_\theta(O) = s_\theta(Z_1) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_r|Z_1, \ldots, Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \frac{\dot{P}_\theta(C = r|Z_1, \ldots, Z_r)}{P_\theta(C = r|Z_1, \ldots, Z_r)}$$

where $s_\theta(Z_1) := \frac{\partial}{\partial\theta} \log f_\theta(Z_1)$, $s_\theta(Z_r|Z_1, \ldots, Z_{r-1}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_r|Z_1, \ldots, Z_{r-1})$ for $r = 2, \ldots, R$, and $\dot{P}_\theta(C = r|Z_1, \ldots, Z_r) := \frac{\partial}{\partial\theta} P_\theta(C = r|Z_1, \ldots, Z_r)$ for $r = 1, \ldots, R$. For Step-2, it is useful to note from (1) and (2) that for any $r = 2, \ldots, R$:

$$\dot{P}_\theta(C \geq r|Z) = -\dot{P}_\theta(C \leq r - 1|Z_1, \ldots, Z_{r-1}) = \dot{P}_\theta(C \geq r|Z_1, \ldots, Z_{r-1}). \tag{9}$$

The tangent set is the mean square closure of all d_β dimensional linear combinations of $S_\theta(O)$ for all such smooth parametric sub-models, and it takes the form

$$\mathcal{T} := \nu_1(Z_1) + \sum_{r=2}^{R} I(C \geq r) \nu_r(Z_1, \ldots, Z_r) + \sum_{r=1}^{R} I(C = r) \omega_r(Z_1, \ldots, Z_r), \tag{10}$$

where $\nu_1(Z_1) \in L_0^2(F(Z_1))$ and $\nu_r(Z_1, \ldots, Z_r) \in L_0^2(F(Z_r|Z_1, \ldots, Z_{r-1}))$ for $r = 2, \ldots, R$, and $\omega_r(Z_1, \ldots, Z_r)$ is any square integrable function of $Z_1, \ldots, Z_r$ for $r = 1, \ldots, R$.

STEP - 2: For brevity we write $m(Z;\beta^0)$ as m, and drop the subscript θ from all quantities evaluated at θ^0. The moment conditions in (3) for a given a, b are equivalent to the requirement that for any $d_\beta \times d_m$ matrix A, the following just-identified system of moment conditions holds:

$$AE[m|a \leq C \leq b] = AE\left[\frac{P(a \leq C \leq b|Z)}{P(a \leq C \leq b)}\frac{I(C=R)}{P(C=R|Z)}m\right] = 0.$$

where the first equality follows from (1). Differentiating with respect to θ under the integral, we obtain

$$0 = AM\frac{\partial\beta^0(\theta_0)}{\partial\theta'} + AE\left[m\left\{s(Z)' + \frac{\dot{P}(a \leq C \leq b|Z)'}{P(a \leq C \leq b|Z)} - \frac{\dot{P}(a \leq C \leq b)'}{P(a \leq C \leq b)}\right\}\middle| a \leq C \leq b\right]$$

where $s(Z) := s(Z_1 + \sum_{r=2}^{R} s(Z_r|Z_1,\ldots,Z_{r-1})$ and $\dot{P}(a \leq C \leq b) := \frac{\partial}{\partial\theta}P_{\theta^0}(a \leq C \leq b)$. Taking a full row rank A along with (1), (3) and assumption (A3) gives

$$\frac{\partial\beta^0(\theta_0)}{\partial\theta'} = -(AM)^{-1}A\left\{E\left[ms(Z)'|a \leq C \leq b\right] + \sum_{r=a}^{b} E\left[m\frac{\dot{P}(C=r|Z_1,\ldots,Z_r)'}{P(a \leq C \leq b)}\right]\right\}$$

Now we establish that for the given A, $-(AM)^{-1}A\varphi(O;\beta^0)$ is the efficient influence function by showing that $E[-(AM)^{-1}A\varphi(O;\beta^0)S(O)'] = \frac{\partial\beta^0(\theta_0)}{\partial\theta'}$ and that $(AM)^{-1}A\varphi(O;\beta^0) \in \mathcal{T}$ defined in (10).

For this purpose, note by using (4) (and switching to the notation T_r for $(Z_1,\ldots,Z_r)$ when it helps brevity) that we can write $E[\varphi(O;\beta^0)S(O)'] = \sum_{i=1}^{3}\sum_{j=1}^{2} B_{ij}$ where

$$\begin{aligned}
B_{11} &:= \sum_{r=b+1}^{R} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)D'\right],\\
B_{12} &:= \sum_{r=b+1}^{R} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
B_{21} &:= \sum_{r=a+1}^{b} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)D'\right],\\
B_{22} &:= \sum_{r=a+1}^{b} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
B_{31} &:= \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a \leq C \leq b)}E[m|T_r]D'\right],\\
B_{32} &:= \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a \leq C \leq b)}E[m|T_r]\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
D &:= s(Z_1) + \sum_{k=2}^{R} I(C \geq k)s(Z_k|T_{k-1}).
\end{aligned}$$

As noted above Proposition 1, we proceed with the understanding that if $b = R$ then $B_{11} = B_{12} = 0$,

and if $a = b$ then $B_{21} = B_{22} = 0$. Also, for notational brevity define T_0 as any constant, so that $s(Z_1) \equiv s(Z_1|T_0)$. First, note that

$$\begin{aligned}
B_{11} &= \sum_{r=b+1}^{R}\sum_{k=1}^{r} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&\quad + \sum_{r=b+1}^{R}\sum_{k=r+1}^{R} E\left[\frac{I(C \geq k)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&= \sum_{r=b+1}^{R}\sum_{k=1}^{r} E\left[\frac{P(C \geq r|T_{r-1})}{P(C \geq r|T_{r-1})}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&\quad + \sum_{r=b+1}^{R}\sum_{k=r+1}^{R} E\left[\frac{P(C \geq k|T_{k-1})}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&= \sum_{r=b+1}^{R} E\left[\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}E[m|T_r]s(Z_r|T_{r-1})'\right] + 0 \\
&= E\left[\frac{I(a \leq C \leq b)}{P(a \leq C \leq b)}ms(Z_R, \ldots, Z_{b+1}|T_b)'\right] = E\left[ms(Z_R, \ldots, Z_{b+1}|T_b)'|a \leq C \leq b\right] \qquad (11)
\end{aligned}$$

where the third and fourth lines follow by (2), the fifth line follows by noting that for all $k = 1, \ldots, r-1$: $E\left[(E[m|T_r] - E[m|T_{r-1}])s(Z_k|T_{k-1})'\right] = E\left[E[(E[m|T_r] - E[m|T_{r-1}])s(Z_k|T_{k-1})'|T_{r-1}]\right] = 0$ while for $k \geq r+1$: $E\left[E[m|T_r]s(Z_k|T_{k-1})'\right] = E\left[E[m|T_r]E[s(Z_k|T_{k-1})'|T_{k-1}]\right] = 0$, and the sixth (last) line follows by (1) and the definition of score. Second, it now follows that

$$B_{21} = \sum_{r=a+1}^{b} E\left[\frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)}E[m|T_{r-1}]s(Z_r|T_{r-1})'\right]$$

exactly following the steps that lead to the fifth line in the expression for B_{11} in (11) above. Therefore,

$$\begin{aligned}
B_{21} &= \sum_{r=a+1}^{b}\sum_{k=a}^{r-1} E\left[\frac{P(C = k|T_k)}{P(a \leq C \leq b)}ms(Z_r|T_{r-1})'\right] \\
&= \sum_{r=a+1}^{b}\sum_{k=a}^{r-1} E\left[ms(Z_r|T_{r-1})'|C = k\right]\frac{P(C = k)}{P(a \leq C \leq b)} \\
&= \sum_{k=a}^{b-1} E\left[m\sum_{r=k+1}^{b} s(Z_r|T_{r-1})'\middle| C = k\right]\frac{P(C = k)}{P(a \leq C \leq b)} \\
&= \sum_{k=a}^{b-1} E\left[ms(Z_b, \ldots, Z_{k+1}|T_k)'|C = k\right]\frac{P(C = k)}{P(a \leq C \leq b)} \qquad (12)
\end{aligned}$$

where the first line follows by (1), the second line follows by the same steps that gave the sixth line in (11), the third line follows by interchanging the order of summations (which is allowed here), and the fourth (last) line follows by the definition of score. Third, we consider B_{31} and note that using the

definition of score in the first line and the same argument as before in the second (last) line below:

$$\begin{aligned} B_{31} &= \sum_{r=a}^{b}\sum_{k=1}^{r} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]s(Z_k)|T_{k-1})'\right] = \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]s(T_r)'\right] \\ &= \sum_{r=a}^{b} E\left[ms(T_r)'|C=r\right]\frac{P(C=r)}{P(a\leq C\leq b)}. \end{aligned} \tag{13}$$

Now we consider the terms B_{12}, B_{22} and B_{32} respectively. Accordingly, first note that

$$\begin{aligned} B_{12} &= \sum_{r=b+1}^{R}\sum_{k=r}^{R} E\left[\frac{I(C=k)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right] \\ &= \sum_{r=b+1}^{R} E\left[\frac{1}{P(C\geq r|T_r)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\sum_{k=r}^{R}\dot{P}(C=k|T_k)'\right] \\ &= \sum_{r=b+1}^{R} E\left[\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\frac{\dot{P}(C\geq r|T_{r-1})'}{P(C\geq r|T_{r-1})}\right] \\ &= 0 \end{aligned} \tag{14}$$

where the second line follows by (1), the third follows line by (1), (2) and (9), and the fourth (last) line follows by taking expectation conditional on T_{r-1} for the r-th term inside the summation. Exactly following the same steps as in the above (recall the analogy with B_{11} and B_{12}) we obtain

$$B_{22} = 0. \tag{15}$$

Lastly, as before, note that

$$B_{32} = \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(C=r|T_r)}\frac{E[m|T_r]\dot{P}(C=r|T_r)'}{P(a\leq C\leq b)}\right] = E\left[m\sum_{r=a}^{b}\frac{\dot{P}(C=r|T_r)'}{P(a\leq C\leq b)}\right]. \tag{16}$$

Therefore, (11)-(16) imply that $E[-(AM)^{-1}A\varphi(O;\beta^0)S(O)'] = \frac{\partial\beta^0(\theta_0)}{\partial\theta'}$. Finally, by matching the first set of terms in $-(AM)^{-1}A\varphi(O;\beta^0)$ (i.e., those that correspond to line one in (4)) to the terms corresponding to $\nu_{b+1}(Z_1,\ldots,Z_{b+1}),\ldots,\nu_R(Z_1,\ldots,Z_R)$ in $\mathcal{T}$; the second set of terms (i.e., those that correspond to line two in (4)) to the terms corresponding to $\nu_a(Z_1,\ldots,Z_a),\ldots,\nu_b(Z_1,\ldots,Z_b)$ in $\mathcal{T}$; and the third set of terms (i.e., those that correspond to line three in (4)) to the terms corresponding to $\omega_a(Z_1,\ldots,Z_a),\ldots,\omega_b(Z_1,\ldots,Z_b)$ in $\mathcal{T}$; while matching zeros with the remaining terms in $\mathcal{T}$, it follows that $-(AM)^{-1}A\varphi(O;\beta^0)$ is the efficient influence function given A.

STEP - 3: Standard arguments give that $A_* := \arg\inf_A Var((AM)^{-1}A\varphi(O;\beta^0)) = M'V^{-1}$. Thus $(A_*M)^{-1}A_*\varphi(O;\beta^0)$ is the efficient influence function, and Ω is the efficiency bound. ■

Appendix B: Further details for some statements from the main text

B.1 The expression of $\varphi(O;\beta)$ in (4) when $a=1, b=R$:

$$
\begin{aligned}
\varphi(O;\beta) &= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} P(C\leq r-1|T_{r-1}) \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad + \sum_{r=1}^{R} I(C=r) E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad - \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} P(C\geq r|T_{r-1}) \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad + \sum_{r=1}^{R} I(C=r) E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad - \sum_{r=2}^{R} I(C\geq r) \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \text{ [by using (2)]} \\
&\quad + \sum_{r=1}^{R} I(C=r) E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad - \left\{ \sum_{r=2}^{R} I(C=r) E[m(T_R;\beta)|T_r] + I(C\geq 2) E[m(T_R;\beta)|T_1] \right\} + \sum_{r=1}^{R} I(C=r) E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) + E[m(T_R;\beta)|T_1].
\end{aligned}
$$

B.2 The weighted average of sub-population $\varphi(O;\beta)$'s as discussed in Remark 5:

Write $E[m(T_R;\beta)|T_r]$ as q_r for all r for brevity. Using the expression for $\varphi(O;\beta)$ when $a=b=j$ (as stated in Remark 1), the weighted average (with weights $P(C=j)/P(a\leq C\leq b)$) of them becomes

$$
\begin{aligned}
&\sum_{j=a}^{b} \left\{ \sum_{r=j+1}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \frac{P(C=j|T_j)}{P(a\leq C\leq b)} (q_r - q_{r-1}) + \frac{I(C=j)}{P(a\leq C\leq b)} q_j \right\} \\
= &\sum_{r=b+1}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)} (q_r - q_{r-1}) \sum_{r=a}^{b} \frac{I(C=r)}{P(a\leq C\leq b)} q_r \\
&+ \sum_{j=a}^{b} \sum_{r=j+1}^{b} \frac{I(C\geq r)}{P(C\geq r|T_r)} \frac{P(C=j|T_j)}{P(a\leq C\leq b)} (q_r - q_{r-1})
\end{aligned}
$$

where the first term follows by (1). Matching the first two terms with the terms on line one and three

of (4), the demonstration will be complete if the third term is equal to the term on the second line of (4). This follows by interchanging the order of the summation (which is allowed here) and noting that

$$\sum_{j=a}^{b}\sum_{r=j+1}^{b}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}(q_r-q_{r-1}) = \sum_{r=a+1}^{b}\sum_{j=a}^{r-1}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}(q_r-q_{r-1})$$
$$= \sum_{r=a+1}^{b}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)}(q_r-q_{r-1}),$$

where the last line follows by (1). The final expression is equal to the term on the second line of (4). ■

B.3 Double-robustness of the expression of $\varphi(O;\beta)$ in (4):

To see this, first replace the unknown $P(a\leq C\leq r|T_r)$ for $r=a+1,\ldots,b$ and $1/P(C\geq r|T_r)$ for $r=a+1,\ldots,R$ in (4) by any integrable functions of T_r, and then note that $E[\varphi(O;\beta^0)]=0$ by (3) (applied to the last line of (4)). On the other hand, rearranging the terms in (4) gives $\varphi(O;\beta)$ alternatively as:

$$\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left[\frac{I(C\geq r)}{P(C\geq r|T_r)}m(T_R;\beta)+\sum_{r=b+1}^{R-1}\left(\frac{I(C\geq r)}{P(C\geq r|T_r)}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1})}\right)E[m(T_R;\beta)|T_r]\right]$$
$$+\sum_{r=a}^{b}\left\{\left(\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1})}\frac{P(a\leq C\leq r|T_r)}{P(a\leq C\leq b)}\right)\right.$$
$$\left.+\frac{I(C=r)}{P(a\leq C\leq b)}\right\}E[m(T_R;\beta)|T_r]$$

where $\{a\leq C\leq a-1\}$ is a null event. Now replacing the unknown $E[m(T_R;\beta)|T_r]$ by any $d_m\times 1$ integrable functions of T_r for $r=1,\ldots,R-1$, it follows by (1) and (3) that $E[\varphi(O;\beta^0)]=0$. ■

B.4 The targets in our Monte Carlo experiment

Target Population for β	Descriptive Statistics					
	Mean	Std $=10^{-3}\times$	Median	IQR	Min	Max
$1\leq C\leq 3$	1	.4860	1	.0007	.9982	1.0017
$C=1$	1.1709	.6841	1.1709	.0009	1.1682	1.1735
$C=2$	.9617	.9430	.9617	.0013	.9581	.9648
$C=3$	.6858	.9769	.6858	.0013	.6817	.6895
$1\leq C\leq 2$	1.0994	.5536	1.0994	.0007	1.0975	1.1012
$2\leq C\leq 3$	.8291	.6800	.8291	.0009	.8265	.8316

Table 4: The true parameter value β^0 is approximated (column 2) for different target populations (column 1) based on averaging over 10,000 Monte Carlo trials the target-sample means obtained by using the same DGP and with sample size 10 million. Descriptive statistics such as standard deviation (Std), interquartile range (IQR), minimum (Min) and maximum (Max) of the estimator (see columns 3-7) indicate that any approximation error is unlikely to distort the results of the subsequent Monte Carlo experiments conducted using sample sizes 100, 200, 500 and 1000 respectively.

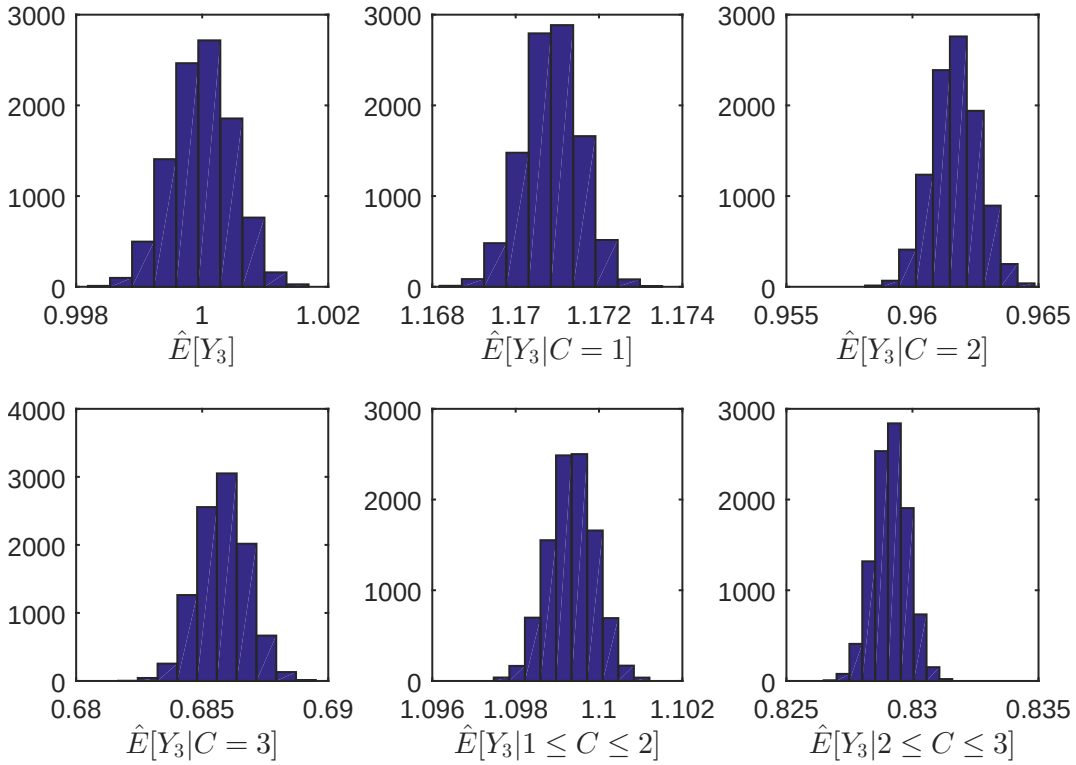

Figure 1: Histograms for the estimators whose descriptive statistics were reported in Table 4.

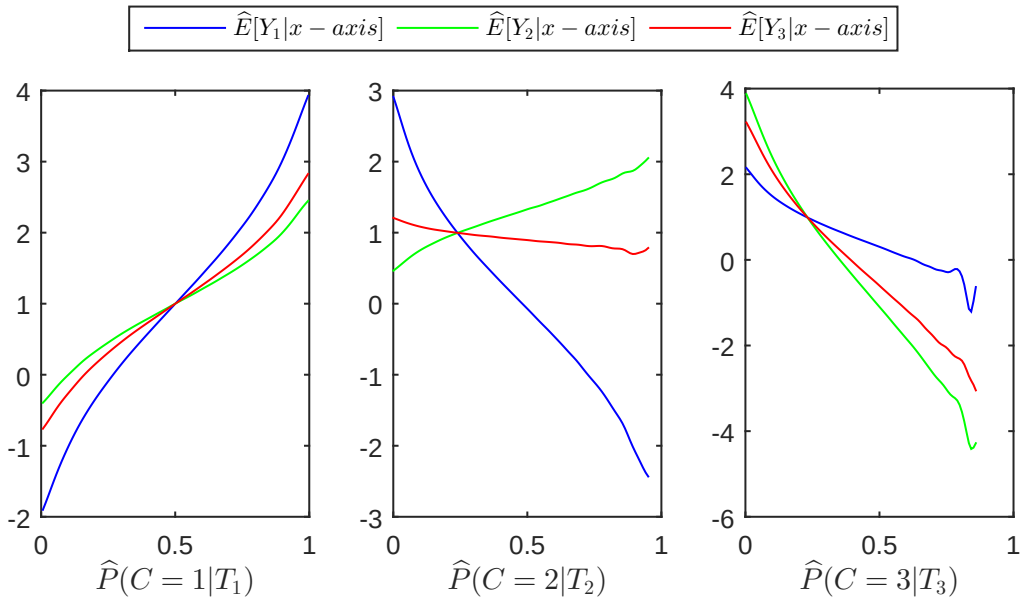

Figure 2: Each kernel estimator is obtained using bandwidth .01. The results are based on exactly the same specification as in Appendix B.5 but only with the first (among the 10,000) Monte Carlo trial. The rather curious feature of a decreasing $E[Y_3|P(C=3|T_3)]$ in the third plot is an artifact of: (i) the relation in (2), which is a derivative of(1), (ii) the positive correlation between the expected outcome and realized outcome, and (iii) the positive correlation among the potential outcomes. (ii) and (iii) are actually reflected in all three plots, i.e., those who left early would have done better, not necessarily with respect to what they did otherwise (which is unknown here since we work with attrition rather than dropout) but with respect to those who stayed. Also see the discussion below (8). For a less nuanced notion of the involved selection, see $\widehat{E}[Y_3|C=j]$ for $j=1,2,3$ in Table 4.

B.5 A note for Tables 1 and 2:

For sample size $n = 100$, there were 12 occurrences in total from 10,000 Monte Carlo trials where the estimated $P(C = 3|T_3)$ was zero for all practical purposes. Naturally, this blows up both the IPW and the efficient estimators. The results are reported in the tables ignoring these 12 occurrences that, with our random number generator, happens in trial numbers: 217, 612, 2995, 3056, 3688, 5657, 6169, 6262, 7434, 7897, 9158 and 9646. A systematic way of dealing with this problem is not pursued here since, fortunately, the estimates in our simulation study are otherwise not affected by this problem.

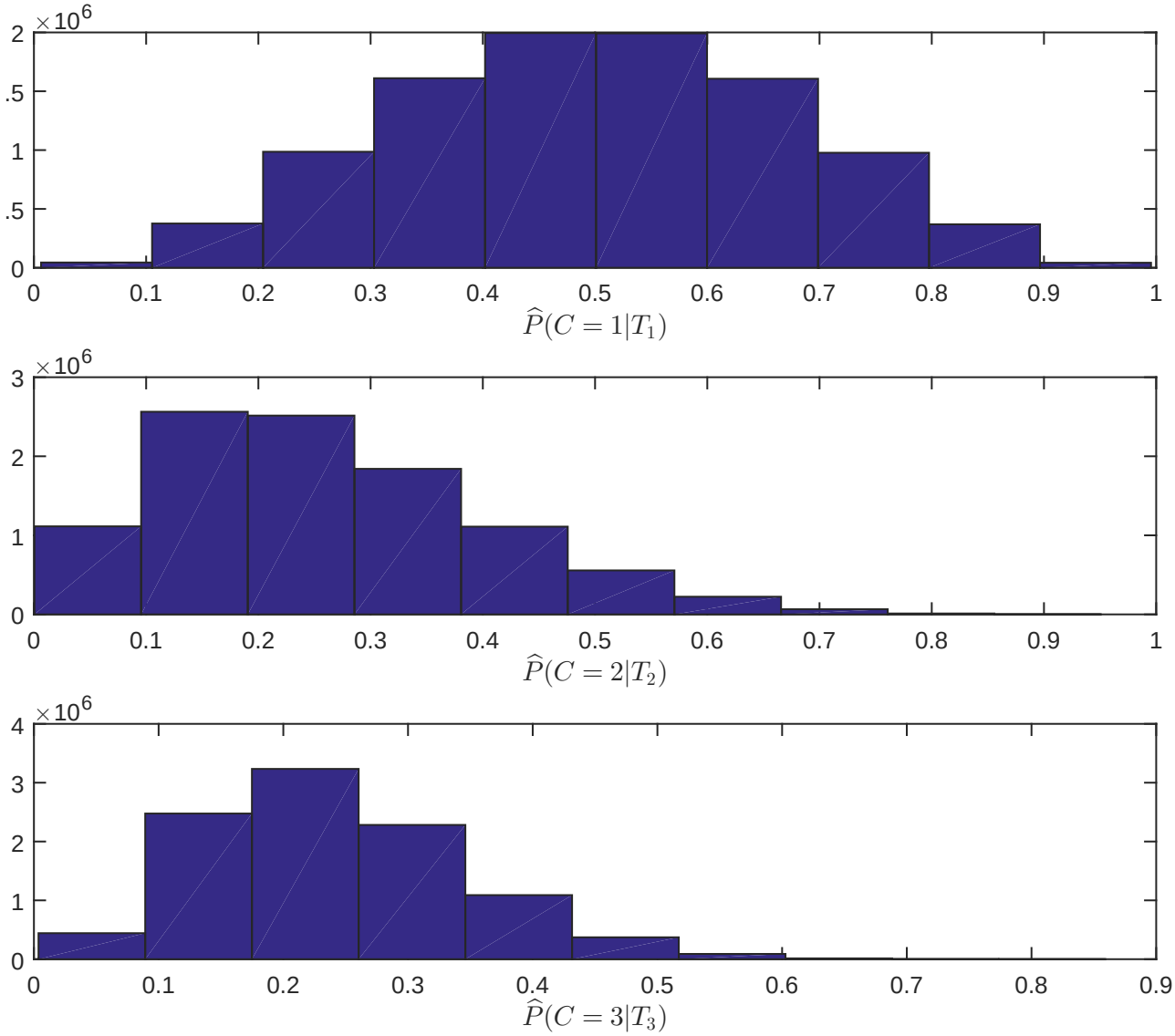

Figure 3: Histograms of $\widehat{P}(C = 1|T_1)$, $\widehat{P}(C = 2|T_2)$ and $\widehat{P}(C = 3|T_3)$ obtained from exactly the same specification as in Appendix B.5 but only with the first (among the 10,000) Monte Carlo trial. Consult this in reference to footnote 5 (page 9).