# A note on efficient estimation with monotonically missing at random data[*]

Jean-Louis Barnwell[†] and Saraswata Chaudhuri[‡]

August 6, 2021.

**Abstract**

This note focuses on efficient estimation of parameters in generic target (sub) populations defined by the missingness pattern of monotonically missing at random data. Attrition from multi-period studies typically generates such data and, in such contexts, our target parameters describe, e.g., the counterfactuals that were unrealized due to the choice of attrition in different periods. The novel features of the results on efficiency are emphasized. A standard doubly-robust estimator for the generic target parameter follows by equating to zero the sample analog of its efficient influence function that plugs in parametric or nonparametric estimators of the unknown nuisance parameters.

*Keywords:* Attrition; Efficiency; Monotonic MAR; Project STAR; Sub-populations.

[*] We thank Daniel Farewell, Erica Moodie, Fabian Lange, Francesco Amodio, Geert Ridder, Marine Carrasco, Steven Lehrer, Thierry Magnac, Tom Parker, Youngki Shin and the seminar participants at Indiana University, Sherbrooke University and ISI Kolkata, and the conference participants at the NY Camp Econometrics 2018, Cireq Econometrics 2018, the Midwest Econometrics Group 2018 for helpful comments. This note was previously circulated with the title "Efficient estimation in sub and full populations with monotonically missing at random data".

[†] Department of Economics, McGill University, Montreal, Canada. Email: jean-louis.barnwellmenard@mail.mcgill.ca.

[‡] Corresponding author. Department of Economics, McGill University and Cireq, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

# 1 Introduction

Our note is about efficient estimation with monotone missing at random data. Attrition to an absorbing state is the primary application of this estimation framework in econometrics; see, e.g., Fitzgerald et al. (1996), Abowd et al. (2001), Wooldridge (2002), Nicoletti (2006) and Wooldridge (2010), etc. Efficient estimation is important in this context because attrition, by definition, involves loss in data that makes it imperative that the available information be used optimally so that the precision of estimates is not needlessly compromised.

Our results on efficient estimation build on Robins and Rotnitzky (1992), Robins et al. (1995) and Rotnitzky and Robins (1995), extending them to sub-populations defined by the monotone pattern of missingness. To set the benchmark that the competing (regular) estimators should strive to reach, we obtain the efficiency bound for estimating parameters under the general framework of moment restrictions models. Our proposed estimator can reach this bound, and turns out to be a standard two-step estimator satisfying double-robustness with respect to the underlying nuisance parameters that can be estimated parametrically or nonparametrically; see, e.g., Robins et al. (1994), Robins and Ritov (1997), Scharfstein et al. (1999), Bang and Robins (2005), Tan (2007), Cao et al. (2009), Rothe and Firpo (2019), etc.

With respect to the vast literature on missing at random data, our contribution is the efficiency results on a variety of sub-populations defined by the monotone pattern of missingness which, in our context, reflects the attrition behavior of agents. These results are also broadly applicable beyond econometrics. For example, thanks to the equivalence theorem of Molenberghs et al. (1998), our results apply directly to the well-known pattern mixture models of Glynn et al. (1986), Little (1993, 1994), etc. Our results also provide new insights on the usability of the sample units toward efficient estimation in sub-populations. This was possible by going beyond, as in the companion paper Chaudhuri (2020), the framework of Robins et al. (1995), Rotnitzky and Robins (1995), Hahn (1998), Chen et al. (2008), etc.

Our note proceeds as follows. Section 2 lays out the framework. Section 3 presents the theory – efficiency bound, efficient influence function, estimation, etc. – by relating them with the literature. The framework in Section 2 is identical to that of Chaudhuri (2020) except in one key aspect — out of necessity, we work with unplanned incompleteness instead of planned incompleteness as in Chaudhuri (2020). We will see in Section 3 that this leads to fundamental differences in the usability of the sample units toward efficient estimation.

The proof of our main result is tedious, and hence for brevity of the note we present the proofs of all the results from Sections 2 and 3 in Supplemental Appendix A. Formal results on the asymptotic properties of the proposed estimator are presented in Supplemental Appendix B since these results for standard two-step estimators (like ours) are very well known and certainly not our contribution. They are presented only for completeness. Complementing the core theory in our note: (i) Supplemental Appendix C is a Monte Carlo experiment demonstrating excellent small-sample properties of our proposed estimator, and (ii) Supplemental Appendix D is an empirical illustration of the benefits of this estimator's precision in drawing substantive conclusions on the effect of small class size across dimensions induced by the attrition behavior of students from the widely studied, attrition-infested Project STAR.

## 2 Framework

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. Let $C$ be a random variable with support $\{1, \ldots, R\}$ 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 $r = 1, \ldots, R$. Let $O := (C, T_C'(Z))'$ denote what is observed for a unit in the sample.

In the case of attrition from a, e.g., 4 period study: $R = 4$, $Z_r$ are the variables that are observed in period $r$, and $T_r(Z)$ is the cumulative history of the variables observed until and including period $r$. If a unit leaves after period $j$ ($\in \{1, \ldots, R\}$), then its $C = j$ and we only observe $T_j(Z)$ for it. $C = R$ is the same as never leaving (some denote this as $C = \infty$).

We maintain the selection on observables, i.e., missing at random (MAR), assumption:

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

Since $T_r(Z)$ is observable when $C \geq r$, (1) imposes that the conditional hazard $P(C = r|T_R(Z), C \geq r)$ at period $r$ does not depend on the unobservables $Z_{r+1}, \ldots, Z_R$ once conditioned on the observables $T_r(Z)$. (1) is the MAR assumption in the sense of Rubin (1976).

Plausibility of MAR depends on the context. However, MAR has been widely used in studies on attrition especially if, as in our paper, the missingness is monotone, i.e., if individuals never return after leaving. By contrast, if the missingness is non-monotone, then selection on observables is unrealistic since the choice to return could depend on unobservables, i.e., on what happened when the individual was out of the study; see, e.g., Gill and Robins (1997), Gill et al. (1997), Robins and Gill (1997) and Vansteelandt et al. (2007). There is a vast literature on point or interval identification under selection on unobservables. Our note does not contribute to that literature. Instead, our note is only concerned with efficiency under selection on observables and a thorough analysis of the efficiency results.

**Nomenclature:** We refer to the underlying population of $O := (C, T'_C(Z))'$ as the full population. We refer to the partition of this full population by the values taken by $C$ as sub-populations; e.g., sub-population $r$ is the underlying population from which units with $C = r$ can be viewed as randomly drawn. There are $R$ unitary sub-populations indexed by $r = 1, \ldots, R$. Unions of unitary sub-populations form a composite sub-population, e.g., $C \in \{1, 2\}$, or the full population $C \in \{1, \ldots, R\}$. Under the selection on observables condition in (1), the unconditional distribution of $Z$ may not be the same as the distribution of $Z$ conditional on $C = r$ for $r = 1, \ldots, R$, i.e., the sub-populations are possibly heterogeneous.

In the case of attrition from a, e.g., 4 period study, the difference in the conditional and unconditional distributions of $Z$ means that, e.g., $E[Z_4]$ is not necessarily equal to $E[Z_4|C = r]$ for any $r = 1, \ldots, 4$. (Note that $Z_4$ is not observed if $C = 1, 2, 3$.) Naturally, $E[Z_4|C = r]$ may not be equal to $E[Z_4|C = s]$ for $r \neq s$ and $r, s = 1, \ldots, 4$, and this means that the timing of attrition or dropout from any study/program matters in general; see Supplemental Appendix D for a concrete empirical example. Therefore, in this context, the sub-populations are defined by the attrition categories based on the timing of attrition.

We will work with a generic target sub-population $[a, b]$ $(\equiv (a \leq C \leq b))$ for $a \leq b$ and $a, b \in \{1, \ldots, R\}$. If $a = b = r$ then this is the underlying unitary sub-population from which the units who left at the end of period $r$ can be viewed as randomly drawn. If $a < b$ then this is the composite sub-population for the units who left in the periods $a, a+1, \ldots, b$. If $a = 1$ and $b = R$ then this is the full-population. We will cover all these cases.

Denote the distribution of $Z$ in the target population by $F_{Z|(a \leq C \leq b)}(z)$. This is the weighted average of the distributions of $Z$ in sub-populations $a, \ldots, b$ with weights $P(C = j)/P(a \leq C \leq b)$ for $j = a, \ldots, b$. We will define the parameter of interest as a finite dimensional feature of $F_{Z|(a \leq C \leq b)}(z)$. Accordingly, consider a function $m(Z; \beta) : \text{Support}(Z) \times \mathcal{B} \mapsto \mathbb{R}^d$, $\beta \in \mathcal{B} \subset \mathbb{R}^d$. Then, for a given $a, b \in \{1, \ldots, R\}$ with $a \leq b$, define the parameter value of interest $\beta^0_{[a,b]}$ as:

$$E[m(Z; \beta) \mid a \leq C \leq b] = 0 \text{ for } \beta \in \mathcal{B} \text{ if and only if } \beta = \beta^0_{[a,b]}. \tag{2}$$

The observability of $m(Z; \beta)$ depends on the underlying elements of $Z$ involved in $m(Z; \beta)$. However, as shown in Lemma 1, identification of $\beta^0_{[a,b]}$ still follows by the Narain-Horvitz-Thompson-Hajek inverse probability weighting (IPW) principle by virtue of the selection on observables condition in (1).

For notational brevity, we will henceforth denote the conditional hazards as:

$$p^0_r(T_r(Z)) := P(C = r | T_r(Z), C \geq r) \text{ for } r = 1, \ldots, R-1.$$

**Lemma 1** *If $P(C = R|T_R(Z)) > 0$ almost surely then the selection on observables condition (1) implies that $E[m(Z; \beta)|a \leq C \leq b] = E\left[\omega^{IPW}_{[a,b]} m(Z; \beta)\right]$ for each $\beta \in \mathcal{B}$ where:*

$$\omega^{IPW}_{[a,b]} := \frac{I(C = R)}{\prod_{j=1}^{R-1}(1 - p^0_j(T_j(Z)))} \frac{\sum_{j=a}^{b} p^0_j(T_j(Z)) \prod_{k=1}^{j-1}(1 - p^0_k(T_k(Z))}{P(a \leq C \leq b)}.$$

*(For notational brevity we used the convention that if $a = 1$ then $\prod_{k=1}^{a-1}(1 - p^0_k(T_k(Z)) = 1$.)*

**Remark:** Lemma 1 is the basis of the well-known IPW estimation and provides a simple way of constructing a feasible quantity $\omega^{\text{IPW}}_{[a,b]} m(Z;\beta)$ whose expectation equals $E[m(Z;\beta)|a \leq C \leq b]$ in (2), and hence can be used as unbiased estimating equations for estimating $\beta^0_{[a,b]}$. Of course, $\omega^{\text{IPW}}_{[a,b]}$ contains $R-1$ unknown nuisance parameters, i.e., the unknown conditional hazards $p^0_r(T_r(Z))$ for $r = 1, \ldots, R-1$.[1] However, they can be estimated based on the observed variables since the conditioning variables $T_r(Z)$ are observed exactly when needed.

Unfortunately, such IPW estimating equations are generally not the "best" ones under the selection on observables assumptions in (1).[2] To define "best", we will obtain the semiparametric efficiency bound and the efficient influence function for the estimation of $\beta^0_{[a,b]}$ maintaining the following familiar assumptions as in Tsiatis (2006) and Chen et al. (2008).

**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) $P(C = R|T_R(Z)) \geq \underline{p}$ almost surely for some fixed $\underline{p} \in (0,1)$.

(A3) $M_{[a,b]} := E\left[ \frac{\partial}{\partial \beta'} m(Z; \beta^0_{[a,b]}) \middle| a \leq C \leq b \right]$ is a $d \times d$ finite, nonsingular matrix.

**Remark:** (A1) rules out dependence and heterogeneity across sample units when viewed as random draws from $O$. (A2) imposes the bounded away from zero condition instead of only $P(C = R|T_R(Z)) > 0$ to avoid the "limited overlap" problem; see, e.g., Khan and Tamer (2010). Note that (1) and (A2) imply that $\prod_{r=1}^{R-1}(1 - p^0_r(T_r(Z))) = P(C = R|T_R(Z)) \geq \underline{p}$. Without altering the results, the differentiability condition in (A3) can be relaxed if needed by instead assuming (A3) for $\frac{\partial}{\partial \beta'} E\left[ m(Z; \beta^0_{[a,b]}) \middle| a \leq C \leq b \right]$ following, e.g., Chen et al. (2008).

[1] To see this, take the simplest case: $a = 1, b = R$. (1) implies that $\omega^{\text{IPW}}_{[a,b]} = I(C = R)/P(C = R|T_{R-1})$ since $P(C = R|T_R) = P(C = R|T_{R-1})$; see Lemma 8 in Supplemental Appendix A for the steps. Unless $R = 2$ or $R > 2$ along with a dimension-reduction assumption in (1) such as $P(C = R|T_R) = P(C = R|T_1)$ as in Cattaneo (2010) or Chaudhuri and Guilkey (2016), the conditioning variables in $P(C = R|T_{R-1})$ are not always observable. Therefore, it is misleading to say that there is only one nuisance parameter $P(C = R|T_{R-1})$, because, without the $R-1$ conditional hazards $p^0_1(T_1), \ldots, p^0_{R-1}(T_{R-1})$ there is no way to contemplate about $P(C = R|T_{R-1})$ under the selection on observables condition in (1). Hence, it is important to remember that there are $R-1$ nuisance parameters $p^0_1(T_1), \ldots, p^0_{R-1}(T_{R-1})$ regardless of $R = 2$ or $R > 2$.

[2] See, e.g., Chaudhuri and Guilkey (2016) and Chaudhuri (2020) for more on this issue. These two papers and the current note do not contradict the important efficiency result of IPW estimating equations in Hirano et al. (2003), Chen et al. (2008), Graham (2011), etc. involving nonparametric estimator of $p^0_1(T_1(Z))$ in the special cases with $R = 2$ and $a = 1, b = 2$ or $a = b = 1$; also see footnote 14 in Supplemental Appendix D.

# 3 Theory

## 3.1 Main result: Efficient influence function and efficiency bound

The key quantity to describe the efficient influence function and the efficiency bound is:

$$
\begin{aligned}
\varphi_{[a,b]}(O;\beta) \;:=\; & \sum_{r=b+1}^{R} \frac{I(C\geq r)}{\prod_{j=1}^{r-1}(1-p_j^0(T_j))} \frac{\sum_{j=a}^{b} p_j^0(T_j)\prod_{k=1}^{j-1}(1-p_k^0(T_k))}{P(a\leq C\leq b)} \left[q_r^0(T_r;\beta)-q_{r-1}^0(T_{r-1};\beta)\right] \\
&+\sum_{r=a+1}^{b} \frac{I(C\geq r)}{\prod_{j=1}^{r-1}(1-p_j^0(T_j))} \frac{\sum_{j=a}^{r-1} p_j^0(T_j)\prod_{k=1}^{j-1}(1-p_k^0(T_k))}{P(a\leq C\leq b)} \left[q_r^0(T_r;\beta)-q_{r-1}^0(T_{r-1};\beta)\right] \\
&+\sum_{r=a}^{b} \frac{I(C=r)}{P(a\leq C\leq b)} q_r^0(T_r;\beta) \qquad (3)
\end{aligned}
$$

where, for brevity, we use the following notation in (3) and onward in the main text:

$$
q_r^0(T_r;\beta) := E[m(Z;\beta)|T_r(Z)] \equiv E[m(T_R(Z);\beta)|T_r(Z)] \;\; \text{for } r=a,\dots,R.
$$

In (3) and onward: (i) we use the convention (as in Lemma 1) that $\prod_{k=1}^{a-1}(1-p_k^0(T_k(Z))=1$ if $a=1$; (ii) we interchangeably write $I(C\geq R)$ and $I(C=R)$, and also $T_r$ and $T_r(Z)$ (where $T_R\equiv Z$); (iii) if $b=R$, then the indices (e.g., in the sums) running from $b+1$ to $R$ are void; and (iv) if $a=b$ then similar indices running from $a+1$ to $b$ are void, and those running from $a$ to $b$ contain only one term which corresponds to $a$ (equivalently, $b$).

**Proposition 2** *Let (1), (2) and assumption A hold. Let the $d\times d$ matrix $V_{[a,b]} := Var(\varphi_{[a,b]}(O;\beta_{[a,b]}^0))$ be finite and positive definite where $\beta_{[a,b]}^0$ and $\varphi_{[a,b]}(O;\beta)$ are as defined in (2) and (3) respectively. Then the asymptotic variance lower bound for $\sqrt{n}(\widehat{\beta}_{[a,b]}-\beta_{[a,b]}^0)$ of any regular estimator $\widehat{\beta}_{[a,b]}$ for $\beta_{[a,b]}^0$ is given by $\Omega_{[a,b]} := M_{[a,b]}^{-1}V_{[a,b]}M_{[a,b]}^{-1\prime}$. An estimator $\widehat{\beta}_{[a,b]}$ whose asymptotic variance equals $\Omega_{[a,b]}$ has the asymptotically linear representation:*

$$
\sqrt{n}(\widehat{\beta}_{[a,b]}-\beta_{[a,b]}^0) = -M_{[a,b]}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{[a,b]}(O_i;\beta_{[a,b]}^0)+o_p(1).
$$

## 3.2 Discussion of the main result: Proposition 2 and $\varphi_{[a,b]}(O;\beta)$

### 3.2.1 Relation with the well-known literature on missing data:

**Corollary 3** *Proposition 2 covers the well-known special cases found in the selection on observables/MAR literature. In these cases $\varphi_{[a,b]}(O;\beta)$ reduces to the following simpler forms.*

*(i) Taking $R=2$, i.e., $Z=(Z_1', Z_2')'$ with $T_1 = Z_1, T_2 = Z$, and then $a=b=1$ in (3) gives:*

$$\varphi_{[a,b]}(O;\beta) = \frac{I(C=2)}{P(C=2|T_1)}\frac{P(C=1|T_1)}{P(C=1)}(q_2^0(T_2;\beta) - q_1^0(T_1;\beta)) + \frac{I(C=1)}{P(C=1)}q_1^0(T_1;\beta).$$

*(ii) Taking $R=2$, i.e., $Z=(Z_1', Z_2')'$ with $T_1 = Z_1, T_2 = Z$, and then $a=1$ and $b=2$ in (3) gives:*

$$\varphi_{[a,b]}(O;\beta) = \frac{I(C=2)}{P(C=2|T_1)}(q_2^0(T_2;\beta) - q_1^0(T_1;\beta)) + q_1^0(T_1;\beta).$$

*(iii) Taking a general $R$ and then $a=1, b=R$ in (3) gives:*

$$\varphi_{[a,b]}(O;\beta) = \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r)}\left(q_r^0(T_r;\beta) - q_r^0(T_{r-1};\beta)]\right) + q_1^0(T_1;\beta).$$

**Remark:** Corollary 3(i) gives Case (1) in Theorem 1 of Chen et al. (2008). Corollary 3(ii) gives Case (2) in Theorem 1 of Chen et al. (2008); also see Robins et al. (1994). Corollary 3(iii) gives the result for the full-population under monotone MAR data with $R > 2$ like Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Holcroft et al. (1997), etc.[3]

### 3.2.2 Application to individual sub-populations, i.e., $a = b$:

If we focus on any generic individual sub-population, i.e., $a = b$, then it is now easy to see from (3) that for a general $R$, the corresponding $\varphi_{[a,b]}(O;\beta)$ in (3) becomes:

$$\sum_{r=a+1}^{R} \frac{I(C \geq r)}{\prod_{j=1}^{r-1}(1-p_j^0(T_j))}\frac{p_a^0(T_a)\prod_{j=1}^{a-1}(1-p_k^0(T_j))}{P(C=a)}\left(q_r^0(T_r;\beta) - q_{r-1}^0(T_{r-1};\beta)\right) + \frac{I(C=a)}{P(C=a)}q_a^0(T_a;\beta).$$

[3] An early reference for the case $a=1, b=R$ is Robins and Rotnitzky (1992) who consider treatment effect estimation under noncompliance (a third choice for therapy) similar to what attrition is in our empirical illustration in Supplemental Appendix D. Unlike us, they do not consider the analysis in sub-populations.

Thus, $\varphi_{[a,b]}(O;\beta) = I(C=R)m(T_R;\beta)/P(C=R)$ in the special case when $a=b=R$. More generally, it is interesting to note that for any individual sub-population $a(=b)$, only those sample units who did not already leave before period $a$ contribute to this efficient estimation.[4] This is in dramatic contrast to Chaudhuri (2020)' s Proposition 1 where, under the additional assumption that $p_r^0(T_r)$ is known for $r=1,\dots,R-1$, it is shown that all the sample units are always usable irrespective of the sub-population under consideration.

Hence, the selection on observables condition in (1) alone is not sufficient to make all the sample units usable. More information is required for that purpose. The known hazard condition in Chaudhuri (2020) in the context of planned incompleteness in surveys provided such information by supplementing (1). An alternative source of such information, while still allowing for unknown hazards, is the imposition of dimension-reduction on (1) to strengthen it, e.g., $P(C=r|T_R,C\geq r) = P(C=r|T_1,C\geq r)$ as in Proposition 5 in Chaudhuri (2020). See Hoonhout and Ridder (2019), Chaudhuri (2020), etc. for other insights on the information content of the selection on observables condition in similar multi-period settings.

Lastly, note that while Case (1) in Theorem 1 of Chen et al. (2008) also focuses on an individual sub-population, their choice $a=b=1$ corresponds to the first period and this means that there is no prior period of attrition. Hence, while Chen et al. (2008)'s choice is a special case of our result, restricting the focus only to that case would not allow to infer on the un-usability of sample units who left before the period of interest (see also footnote 4).

### 3.2.3 The weighted average representation:

The moment restrictions in (2) will always satisfy the standard weighted average representation (recall the definition of $F_{Z|(a\leq C\leq b)}(z)$ from Section 2) across sub-populations constituting

[4] More precisely, while the units that left before period $a$ are essential in identifying the conditional hazards $p_1^0(T_1),\dots,p_{a-1}^0(T_{a-1})$ in periods $1,\dots,a-1$, they do not contribute with the variance-reducing-augmenting terms (to the IPW) involving the conditional expectations $q_1^0(T_1;\beta),\dots,q_{a-1}^0(T_{a-1};\beta)$ of the moment vector. Hence, it is not as if we can completely ignore (drop) the periods $1,\dots,a-1$ because then we will not be able to identify these conditional hazards or terms like $\prod_{j=1}^a(1-p_j^0(T_j)) = \prod_{j=1}^a(1-P(C=j|T_j,C\geq j)) = \prod_{j=1}^a(1-P(C=j|T_a,C\geq j)) = \prod_{j=1}^a P(C>j|T_a,C\geq j) = P(C>a|T_a) = 1-P(C\leq a|T_a)$ where the second equality follows by MAR in (1) since $T_a$ is contained in $T_R$, while the fourth equality is due to the telescoping product. This intuition is obviously moot if $a=1$ as in Case 1 of Theorem 1, Chen et al. (2008).

$[a, b]$:

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

Interestingly, this weighted average representation also holds for $\varphi_{[a,b]}(O;\beta)$ defined in (3).

**Corollary 4** *For $a \leq b \in \{1, \dots, R\}$, $\varphi_{[a,b]}(O;\beta)$ in (3) has the weighted average representation:*

$$\varphi_{[a,b]}(O;\beta) = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \leq C \leq b)} \varphi_{[j,j]}(O;\beta).$$

**Remark:** This representation is intuitively appealing, and it presents a way of combining the efficient estimators for the individual sub-populations to obtain the efficient estimator for their contiguous unions. Recent examples of combining either (i) estimators or (ii) moment restrictions for similarly defined sub-populations to obtain their full-population counterparts include Dardanoni et al. (2011) and Muris (2018) for (i), and Chaudhuri and Guilkey (2016) and Abrevaya and Donald (2017) for (ii), and Chaudhuri (2020) for both (i) and (ii).

### 3.2.4 Doubly robust estimating equations:

$\varphi_{[a,b]}(O;\beta)$ depends on the $R-1$ conditional hazards $p_r^0(T_r)$ for $r = 1, \dots, R-1$ (also see footnote 1), and also on the $R-a$ conditional expectations $q_r^0(T_r;\beta)$ for $r = a, \dots, R-1$. (The $q_r^0(T_r;\beta)$'s do not appear if $a = b = R$; see Section 3.2.2.) These are nuisance parameters. The structure of $\varphi_{[a,b]}(O;\beta)$ provides "some" protection against misspecification of these two sets of nuisance parameters. Let us note why. Imitating the structure of $\varphi_{[a,b]}(O;\beta)$, define:

$$\begin{aligned} & g(O;\beta, p(Z), q(Z;\beta)) \\ := \quad & \sum_{r=b+1}^{R} \frac{I(C \geq r)}{\prod_{j=1}^{r-1}(1-p_j(T_j))} \frac{\sum_{j=a}^{b} p_j(T_j) \prod_{k=1}^{j-1}(1-p_k(T_k))}{P(a \leq C \leq b)} [q_r(T_r;\beta) - q_{r-1}(T_{r-1};\beta)] \\ & + \sum_{r=a+1}^{b} \frac{I(C \geq r)}{\prod_{j=1}^{r-1}(1-p_j(T_j))} \frac{\sum_{j=a}^{r-1} p_j(T_j) \prod_{k=1}^{j-1}(1-p_k(T_k))}{P(a \leq C \leq b)} [q_r(T_r;\beta) - q_{r-1}(T_{r-1};\beta)] \\ & + \sum_{r=a}^{b} \frac{I(C=r)}{P(a \leq C \leq b)} q_r(T_r;\beta), \end{aligned} \tag{4}$$

where $q_R(T_R;\beta) \equiv m(Z;\beta)$, and for generic functions $q_r : (T_r;\beta) \mapsto \mathbb{R}^d$ for $r = a, \ldots, R-1$, and $p_r : (T_r) \mapsto (0,1)$ for $r = 1, \ldots, R-1$, and their respective "true values" $q_r^0(T_r;\beta)$ for $r = a, \ldots, R-1$ and $p_r^0(T_r)$ for $r = 1, \ldots, R-1$, we consolidated notation by writing:

$$\begin{aligned} &q(Z;\beta) := (q_a'(T_a;\beta), \ldots, q_{R-1}'(T_{r-1};\beta))', \quad q^0(Z;\beta) := (q_a^{0'}(T_a;\beta), \ldots, q_{R-1}^{0'}(T_{r-1};\beta))', \\ &p(Z) := (p_1(T_1), p_2(T_2), \ldots, p_{R-1}(T_{r-1}))', \quad p^0(Z) := (p_1^0(T_1), p_2^0(T_2), \ldots, p_{R-1}^0(T_{r-1}))'. \end{aligned} \tag{5}$$

Compare (3) and (4) to see that $g(O;\beta,p(Z),q(Z;\beta))$ simply replaces in $\varphi_{[a,b]}(O;\beta)$ all the $p_r^0(T_r)$'s by the corresponding $p_r(T_r)$'s and all the $q_r^0(T_r;\beta)$'s by the corresponding $q_r(T_r;\beta)$'s.

**Lemma 5** *It follows from* (3) *and* (4) *that* $\varphi_{[a,b]}(O;\beta) = g(O;\beta,p^0(Z),q^0(Z;\beta))$ *for all* $O$ *and* $\beta$. *The selection on observables condition in* (1) *implies that for all* $\beta \in \mathcal{B}$*:*

$$E[g(O;\beta,p(Z),q^0(Z;\beta))] = E[g(O;\beta,p^0(Z),q(Z;\beta))] = E[m(Z;\beta)|a \leq C \leq b].$$

**Remark:** $g(O;\beta,p(Z),q(Z;\beta))$ can be used as the moment vector for the estimation of $\beta_{[a,b]}$. Lemma 5 shows that the structure of $\varphi_{[a,b]}(O;\beta)$ bestows on $g(O;\beta,p(Z),q(Z;\beta))$ protection against misspecification of the two sets of nuisance parameters $p^0(Z)$ and $q^0(Z;\beta)$. This is the "double-robustness" property in the sense of Scharfstein et al. (1999). Also see Robins et al. (1994), Robins and Ritov (1997), Bang and Robins (2005), Tan (2007), Cao et al. (2009), Rothe and Firpo (2019), etc. Specifically, Lemma 5 shows that this moment vector will not alter the global identification condition (2) for $\beta_{[a,b]}^0$ if at least one set of nuisance parameters is correctly specified in $g(O;\beta,p(Z),q(Z;\beta))$, i.e., $p(Z) = p^0(Z)$ or/and $q(Z;\beta) = q^0(Z;\beta)$.

## 3.3 Estimator of $\beta$ and a sketch of its asymptotic properties

For a given $[a,b]$, we will use an estimator of $\beta$ that solves the sample estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n} g(O_i;\beta,\widehat{p}(Z_i),\widehat{q}(Z_i;\beta)) = 0. \tag{6}$$

(6) is the sample analog of the population estimating equations $E[g(O;\beta,p(Z),q(Z;\beta))]=0$ with parametric or nonparametric estimators $\widehat{p}(Z)$ and $\widehat{q}(Z;\beta)$ in place of $p(Z)$ and $q(Z;\beta)$.

The theory behind such "two-step" estimation of $\beta$ with nuisance parameters is well known. Parametric or nonparametric estimation of the nuisance parameters is also standard. Hence we only provide here a minimal sketch of the estimation and the theory while relegating a formal presentation (including estimation of standard error) to Supplemental Appendix B.

First, consider the conditional hazards. Let the parametric model, e.g., logit/probit, for $p_r^0(T_r)$ be $p_r(T_r;\gamma_r)$ where $\gamma_r$ is a $d_{\gamma_r}\times 1$ unknown vector for $r=1,\ldots,R-1$. We obtain the quasi-maximum likelihood estimator $\widehat{\gamma}_r$ of $\gamma_r$ by solving the score equations:

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} S_r(O_i;\widehat{\gamma}_r) \quad \text{for} \quad r=1,\ldots,R-1, \quad \text{where for} \quad i=1,\ldots,n, \\ S_r(O_i;\gamma_r) &:= I(C_i\geq r)\frac{I(C_i=r)-p_r(T_{r,i};\gamma_r)}{p_r(T_{r,i};\gamma_r)(1-p_r(T_{r,i};\gamma_r))}\left\{\frac{\partial}{\partial\gamma_r}p_r(T_{r,i};\gamma_r)\right\} \end{aligned} \tag{7}$$

We obtain the parametric estimator $\widehat{p}(Z)$ as $\widehat{p}(Z)=(p_1(T_1;\widehat{\gamma}_1),\ldots,p_{R-1}(T_{R-1};\widehat{\gamma}_{R-1}))'$.

Now, consider the conditional expectations. Let the parametric model, e.g., linear model, for $q_r^0(T_r;\beta)$ be $q_r(T_r;\beta,\lambda_r(\beta))$ where $\lambda_r(\beta)$ is a $d_{\lambda_r}\times 1$ unknown vector for $r=a,\ldots,R-1$. We obtain the least squares estimator $\widehat{\lambda}_r(\beta)$ of $\lambda_r(\beta)$ as a function of $\beta$ by solving:

$$\begin{aligned} 0 = \frac{1}{n}\sum_{i=1}^{n} L_r(O_i;\beta,\widehat{\lambda}_r(\beta)) \quad \text{for} \quad r=a,\ldots,R-1, \quad \text{where for} \quad i=1,\ldots,n, \\ L_r(O_i;\beta,\lambda_r) := I(C_i=R)\left\{\frac{\partial}{\partial\lambda_r}q_r'(T_{r,i};\beta,\lambda_r)\right\}(m(T_{R,i};\beta)-q_r(T_{r,i};\beta,\lambda_r))\,. \end{aligned} \tag{8}$$

These equations are simply the first order condition of least squares in a system of equations. We obtain the parametric estimator $\widehat{q}(Z;\beta)$ as $\widehat{q}(Z;\beta)=(q_a'(T_a;\beta,\widehat{\lambda}_a(\beta)),\ldots,q_{R-1}'(T_{R-1};\beta,\widehat{\lambda}_{R-1}(\beta)))'$.

One could, in principle, also use any nonparametric estimator $\widehat{p}(Z)$ and $\widehat{q}(Z;\beta)$. Use of series/sieve estimators appears to be common, e.g., Hahn (1998), Hirano et al. (2003), Chen et al. (2005), Chen et al. (2008), Cattaneo (2010), Ackerberg et al. (2012), etc. For example, one could take: (i) $p_r(T_r;\gamma_r)$ as logit (e.g., Hirano et al. (2003)) or probit (e.g., Ackerberg et al. (2012)) with index $\xi'_{d_{\gamma_r}}(T_r)\gamma_r$ for $r = 1, \ldots, R-1$ and take (ii) $q_r(T_r; \beta, \lambda_r(\beta))$ as $\pi'_{d_{\lambda_r}}(T_r)\lambda_r(\beta)$ for $r = a, \ldots, R-1$ where $\xi_{d_{\gamma_r}}(T_r)$ and $\pi_{d_{\lambda_r}}(T_r)$ are the first $d_{\gamma_r}$ and $d_{\lambda_r}$ terms of some (possibly the same) basis function. The promise of nonparametrics materializes if $d_{\gamma_r} \to \infty$ for $r = a, \ldots, R-1$ and $d_{\lambda_r} \to \infty$ $r = 1, \ldots, R-1$ as $n \to \infty$. On the other hand, the implementation of such nonparametric estimators $\widehat{p}(Z)$ and $\widehat{q}(Z; \beta)$ is exactly the same as that of their parametric counterpart (following (7) and (8)) for a given sample size $n$.

**Assumption CH:** The conditional hazard (CH) model is correct, i.e., there exists a $\gamma^0 = (\gamma_1^{0\prime}, \ldots, \gamma_{R-1}^{0\prime})'$ with $\gamma_r^0 \in \Gamma_r$ such that $p_r(T_r; \gamma_r^0) = p_r^0(T_r)$ for $r = 1, \ldots, R-1$.

**Assumption CE:** The conditional expectation (CE) model is correct, i.e., there exists a $\lambda^0 = (\lambda_a^{0\prime}, \ldots, \lambda_{R-1}^{0\prime})'$ with $\lambda_r^0 \in \Lambda_r$ such that $q_r(T_r; \beta_{[a,b]}^0, \lambda_r^0) = q_r^0(T_r; \beta_{[a,b]}^0)$ for $r = a, \ldots, R-1$.

Assumptions CH and CE are relevant only for parametric estimators $\widehat{p}(Z)$ and $\widehat{q}(Z; \beta)$. For nonparametric estimators, on the other hand, the approximation to the truth can be made arbitrarily well (in $L_2$ or sup norm) given sufficient smoothness of $p^0(Z)$ and $q^0(Z; \beta)$ as $d_{\gamma_r} \to \infty$ for $r = a, \ldots, R-1$ and $d_{\lambda_r} \to \infty$ $r = 1, \ldots, R-1$ as $n \to \infty$.

**Proposition 6** *Consider any given target sub-population $[a, b]$ and the estimator $\widehat{\beta}$ obtained by solving the associated estimating equations in (6) for the target parameter $\beta_{[a,b]}^0$.*

*(i) Suppose that we use parametric estimators $\widehat{p}(Z)$ and $\widehat{q}(Z; \beta)$ in (6). Suppose that the standard assumptions of two-step estimation stated in Supplemental Appendix B.1 following Newey and McFadden (1994) and van der Vaart (1998) hold. Then $\sqrt{n}(\widehat{\beta} - \beta^*) \xrightarrow{d} N(0, \Upsilon_{\beta\beta})$ for some $d \times 1$ vector $\beta^* \in \mathcal{B}$ and $d \times d$ positive definite matrix $\Upsilon_{\beta\beta}$, both defined in Supplemental Appendix B.1, and satisfying $\beta^* = \beta_{[a,b]}^0$ if assumption CH or/and assumption CE holds, and $\Upsilon_{\beta\beta} = \Omega_{[a,b]}$ if both assumption CH and assumption CE hold.*

*(ii) Suppose that we use nonparametric estimators $\widehat{p}(Z)$ and $\widehat{q}(Z; \beta)$ in (6). Suppose that the standard assumptions of two-step profiled semiparametric estimation stated in Supplemental Appendix B.2 following Chen et al. (2003) hold. Then $\sqrt{n}(\widehat{\beta} - \beta_{[a,b]}^0) \xrightarrow{d} N(0, \Omega_{[a,b]})$.*

We already saw that the implementation of the efficient estimator is standard. Proposition 6 shows that its asymptotic properties are also standard. Hence we relegate its detailed discussion along with the estimation of the standard errors to Supplemental Appendix B.

To complement our theoretical discussion: (i) Supplemental Appendix C presents a Monte Carlo experiment demonstrating excellent properties of the efficient estimator in small samples, and (ii) Supplemental Appendix D illustrates the benefits of that efficiency of the proposed estimator in drawing substantive conclusions on the effect of small classes in the widely studied Project STAR. It turns out in both (i) and (ii) that the commonly used but not necessarily efficient estimators are vastly outperformed by our proposed estimator.

We now conclude with the following remark. Our note focused on efficiency when interest lies on sub-populations. We provided a (hopefully) thorough analysis of efficiency, and highlighted the novel aspects of the contribution of sample units toward efficiency and its comparison with less general setups under selection on observables. We hope that this analysis of the information content of the selection on observables condition is of interest to the reader. On the applied side, it is imperative to use efficient estimators whenever possible since attrition over multiple periods can cause severe loss in the data. Hence, we also hope that the standardness of the proposed efficient estimator is encouraging to practitioners.

# 4 Bibliography

Abowd, J. M., Crepon, B., and Kramarz, F. (2001). Moment Estimation with Attrition: An Application to Economic Models. *Journal of the American Statistical Association*, 96:1223–1231.

Abrevaya, J. and Donald, S. G. (2017). A GMM approach for dealing with missing data on regressors and instruments. *Review of Economics and Statistics*, 99:657–662.

Ackerberg, D., Chen, X., and Hahn, J. (2012). A Practical Asymptotic Variance Estimator For Two-Step Semiparametric Estimators. *The Review of Economics and Statistics*, 94:481–498.

Bang, H. and Robins, J. M. (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. (2020). On Efficiency Gains from Multiple Incomplete Subsamples. *Econometric Theory*, 36:488–525.

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

Chen, X., Hong, H., and Tamer, E. (2005). Measurement Error Models with Auxiliary Data. *Review of Economic Studies*, 72:343–366.

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.

Dardanoni, V., Modica, S., and Peracchi, F. (2011). Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics*, 162:362–368.

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.

Gill, R. and Robins, J. M. (1997). Non-Response Models For The Analysis Of Non-Monotone Ignorable Missing Data. *Statistics in Medicine*, 16:39–56.

Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997). Coarsening at Random: Characterizations, Conjectures and Counterexamples. In Lin, D. Y. and Fleming, T. R., editors, *Proceedings of The First Seattle Symposium in Biostatistics: Survival Analysis*, Lecture Notes in Statitsics, pages 255–294. New York: Springer-Verlag.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). *Selecion modeling versus mixture modeling with nonignorable nonresponses*, pages 115–142. Springer-Verlag, NY.

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

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

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.

Hoonhout, P. and Ridder, G. (2019). Nonignorable Attrition in Multi-Period Panels With Refreshment Samples. *Journal of Business and Economic Statistics*, 37:377–390.

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

Little, R. J. A. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association*, 88:125–134.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81:471–483.

Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, 52:153–161.

Muris, C. (2018). Efficient GMM estimation with incomplete data. Forthcoming in Review of Economics and Statistics.

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. M. and Gill, R. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, 16:39–56.

Robins, J. M. 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. M. and Rotnitzky, A. (1992). Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. In N. Jewell, K. D. and Farewell, V. T., editors, *AIDS Epidemiology: Methodological Issues*, pages 297–331. Birkhliuser, Boston.

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

Robins, J. 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, J. 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. (2019). Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically. *Econometric Theory*, 35: 1048–1087.

Rotnitzky, A. and Robins, J. M. (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.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Vansteelandt, S., Rotnitzky, A., and Robins, J. M. (2007). Estimation of regression models for mean of repeated outcomes under nonignorable nonmotonone nonresponse. *Biometrika*, 94:841–860.

Wooldridge, J. M. (2002). Inverse Probability Weighted M-Estimation for Sample Selection, Attrition, and Stratification. *Portuguese Economic Journal*, 1:117–139.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section & Panel Data.* MIT Press.

# Supplemental Appendix:

## A note on efficient estimation with monotonically missing at random data

Jean-Louis Barnwell[5] and Saraswata Chaudhuri[6]

## Table of Contents

[5] Department of Economics, McGill University, Montreal, Canada. Email: jean-louis.barnwellmenard@mail.mcgill.ca.

[6] Corresponding author. Department of Economics, McGill University and Cireq, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

The numbering of the equations, corollaries, lemmas and propositions in this supplemental appendix is consistent with the main text of our paper.

# A Supplemental Appendix A: Proofs of main results

The following two lemmas are useful for most of the proofs below.

**Lemma 7** *The MAR condition in (1) implies and is implied by the following condition:*

$$P(C = r|T_R) = P(C = r|T_r) \text{ for } r = 1, \ldots, R-1. \tag{9}$$

**Proof of Lemma 7:** First we show that if (1) holds then (9) also holds. Take any $r = 1, \ldots, R-1$ and note that:

$$P(C = r|T_R) = p^0_r(T_R)\prod_{k=1}^{r-1}(1-p^0_k(T_R)) = p^0_r(T_r)\prod_{k=1}^{r-1}(1-p^0_k(T_k)) = p^0_r(T_r)\prod_{k=1}^{r-1}(1-p^0_k(T_r)) = P(C = r|T_r)$$

where the second and the third equalities follow by (1). Now we show that if (9) holds then (1) also holds. Take any $r = 1, \ldots, R-1$ and note that:

$$\begin{aligned} P(C = r|T_R, C \geq r) &= \frac{P(C = r|T_R)}{P(C \geq r|T_R)} = \frac{P(C = r|T_R)}{1 - P(C \leq r-1|T_R)} = \frac{P(C = r|T_R)}{1 - \sum_{j=1}^{r-1} P(C = j|T_R)} \\ &= \frac{P(C = r|T_r)}{1 - \sum_{j=1}^{r-1} P(C = j|T_j)} = \frac{P(C = r|T_r)}{1 - \sum_{j=1}^{r-1} P(C = j|T_r)} \\ &= \frac{P(C = r|T_r)}{P(C \geq r|T_r)} = P(C = r|T_r, C \geq r) \end{aligned}$$

where the fourth and fifth equality follow by (9). ■

**Lemma 8** *The MAR condition in (1) implies that:*

$$P(C \geq r|T_j) = P(C \geq r|T_{r-1}) \text{ for } r = 1, \ldots, R-1 \text{ and } j = r, \ldots, R.$$

**Proof of Lemma 8:** Lemma 7 shows that (1) implies (9). Now, take $r = 1, \ldots, R-1$ and

$j = r, \ldots, R$ and note that:

$$P(C \geq r|T_j) = 1 - \sum_{k=1}^{r-1} P(C = k|T_j) = 1 - \sum_{k=1}^{r-1} P(C = k|T_k) = 1 - \sum_{k=1}^{r-1} P(C = k|T_{r-1}) = P(C \geq r|T_{r-1})$$

where the second and the third equalities follow by (9). ■

**Remarks:**

1. Lemma 8 implies that if $R = 2$ then $P(C = 2|T_2) = P(C = 2|T_1)$. This is the familiar form in which the MAR assumption is generally found in the econometrics literature where the focus has typically on the case of $R = 2$; also see footnote 1.

2. We introduced the notation in the above two lemmas for brevity of expressions in the proofs in this appendix. The original notation with the conditional hazards is very transparent in terms of accounting for the observability of the conditioning variables (and hence for estimation), and precisely for that reason it leads to longer expressions.

**Proof of Lemma 1:** Note that:

$$\begin{aligned}
\omega^{\text{IPW}}_{[a,b]} &:= \frac{I(C = R)}{\prod_{j=1}^{R-1}(1 - p_j^0(T_j))} \frac{\sum_{j=a}^{b} p_j^0(T_j) \prod_{k=1}^{j-1}(1 - p_k^0(T_k)}{P(a \leq C \leq b)} \\
&= \frac{I(C = R)}{\prod_{j=1}^{R-1}(1 - p_j^0(T_R))} \frac{\sum_{j=a}^{b} p_j^0(T_R) \prod_{k=1}^{j-1}(1 - p_k^0(T_R)}{P(a \leq C \leq b)} \\
&= \frac{I(C = R)}{P(C = R|T_R)} \frac{P(a \leq C \leq b|T_R)}{P(a \leq C \leq b)}
\end{aligned}$$

where the first equality follows by (1) and Lemma 7, and the second one by the standard relation between hazards and distribution functions. Therefore, since $Z \equiv T_R$, it follows by using the law of iterated expectations in the second and third equalities below, that:

$$\begin{aligned}
E\left[\omega^{\text{IPW}}_{[a,b]} m(Z;\beta)\right] &= E\left[\frac{I(C = R)}{P(C = R|T_R)} \frac{P(a \leq C \leq b|T_R)}{P(a \leq C \leq b)} m(T_R;\beta)\right] \\
&= E\left[\frac{P(a \leq C \leq b|T_R)}{P(a \leq C \leq b)} m(T_R;\beta)\right] = E\left[\frac{I(a \leq C \leq b)}{P(a \leq C \leq b)} m(T_R;\beta)\right] \\
&= E[m(Z;\beta)|a \leq C \leq b]. \blacksquare
\end{aligned}$$

**Proof of Proposition 2:** Guided by the rationale of brevity of expressions as stated in Remark 2 following Lemma 8, let us use (1) and Lemma 7 to rewrite $\varphi_{[a,b]}(O;\beta)$ in (3)

$$\begin{aligned}\varphi_{[a,b]}(O;\beta) \quad = \quad & \sum_{r=b+1}^{R} \frac{I(C\geq r)}{P(C\geq r \mid T_r)} \frac{P(a\leq C\leq b \mid T_b)}{P(a\leq C\leq b)} \left(E[m(T_R;\beta)\mid T_r] - E[m(T_R;\beta)\mid T_{r-1}]\right) \\ & + \sum_{r=a+1}^{b} \frac{I(C\geq r)}{P(C\geq r \mid T_r)} \frac{P(a\leq C\leq r-1 \mid T_{r-1})}{P(a\leq C\leq b)} \left(E[m(T_R;\beta)\mid T_r] - E[m(T_R;\beta)\mid T_{r-1}]\right) \\ & + \sum_{r=a}^{b} \frac{I(C=r)}{P(a\leq C\leq b)} E[m(T_R;\beta)\mid T_r]. \end{aligned} \tag{10}$$

We follow the two steps as in, e.g., Chen et al. (2008) in this proof. 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 and, thereby, the efficiency bound as the expectation of the outer product of the efficient influence function.

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

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

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

in terms of $(C,Z')'$. Let $\eta^0$ be the value of $\eta$ such that $f_{\eta^0}(O)$ equals the true $f(O)$ for which (2) actually holds. ($\eta^0$ is used to denote the truth in this proof but is omitted in Step 2 when it is obvious.) Then, the score function with respect to $\eta$ is, in terms of $(C,Z')'$,:

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

where $s_\eta(Z_1) := \frac{\partial}{\partial\eta}\log f_\eta(Z_1)$, $s_\eta(Z_r|Z_1,\ldots,Z_{r-1}) := \frac{\partial}{\partial\eta}\log f_\eta(Z_r|Z_1,\ \ldots,Z_{r-1})$ for $r =$

$2,\ldots,R$, and $\dot{P}_\eta(C=r|Z_1,\ldots,Z_r) := \frac{\partial}{\partial\eta}P_\eta(C=r|Z_1,\ldots,Z_r)$ for $r=1,\ldots,R$. For Step-2, it is useful to note from Lemma 8 that for any $r=2,\ldots,R$:

$$\dot{P}_\eta(C\geq r|Z) = -\dot{P}_\eta(C\leq r-1|Z_1,\ldots,Z_{r-1}) = \dot{P}_\eta(C\geq r|Z_1,\ldots,Z_{r-1}). \tag{11}$$

The tangent set is the mean square closure of all $d$ dimensional linear combinations of $S_\eta(O)$ for all such smooth parametric sub-models, and it can be generically defined as:

$$\mathcal{T} := \nu_1(Z_1) + \sum_{r=2}^{R} I(C\geq r)\nu_r(Z_1,\ldots,Z_r) + \sum_{r=1}^{R} I(C=r)\omega_r(Z_1,\ldots,Z_r), \tag{12}$$

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 $m(Z;\beta^0_{[a,b]})$ as $m$, and drop the subscript $\eta$ from all quantities (e.g., the expectations below) evaluated at $\eta^0$. Under our assumptions, (the proof of) Lemma 1 showed that the moment conditions in (2) for a given $a,b$ can be expressed as:

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

Differentiating with respect to $\eta$ under the integral gives:

$$0 = M_{[a,b]}\frac{\partial\beta^0_{[a,b]}(\eta_0)}{\partial\eta'} + E\left[m\left\{s(Z)' + \frac{\dot{P}(a\leq C\leq b|Z)'}{P(a\leq C\leq b|Z)} - \frac{\dot{P}(a\leq C\leq b)'}{P(a\leq C\leq b)}\right\}\middle| a\leq C\leq b\right]$$

where $s(Z) := s(Z_1 + \sum_{r=2}^{R} s(Z_r|Z_1,\ldots,Z_{r-1})$ and $\dot{P}(a\leq C\leq b) := \frac{\partial}{\partial\eta}P_{\eta^0}(a\leq C\leq b)$. Hence, (1) and Lemma 7, (2) and assumption (A3) give:

$$\frac{\partial\beta^0_{[a,b]}(\eta_0)}{\partial\eta'} = -M^{-1}_{[a,b]}\left\{E\left[ms(Z)'|a\leq C\leq b\right] + \sum_{r=a}^{b}E\left[m\frac{\dot{P}(C=r|Z_1,\ldots,Z_r)'}{P(a\leq C\leq b)}\right]\right\}.$$

We will establish that $-M^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ is the efficient influence function by showing that

$E[-M_{[a,b]}^{-1}\varphi_{[a,b]}(O;\beta_{[a,b]}^0)S(O)'] = \frac{\partial\beta_{[a,b]}^0(\eta_0)}{\partial\eta'}$ and that $M_{[a,b]}^{-1}\varphi_{[a,b]}(O;\beta_{[a,b]}^0) \in \mathcal{T}$ defined in (12).

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

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

where the third and fourth lines follow by Lemma 8. Hence, we subsequently obtain that:

$$\begin{aligned} B_{11} &= \sum_{r=b+1}^{R} E\left[\frac{P(a \le C \le b|T_b)}{P(a \le C \le b)} E[m|T_r] s(Z_r|T_{r-1})'\right] + 0 \\ &= E\left[\frac{I(a \le C \le b)}{P(a \le C \le b)} ms(Z_R, \ldots, Z_{b+1}|T_b)'\right] = E\left[ms(Z_R, \ldots, Z_{b+1}|T_b)'|a \le C \le b\right].(13) \end{aligned}$$

The first equality follows since 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 $k \ge r+1$: $E\left[E[m|T_r]s(Z_k|T_{k-1})'\right] = E\left[E[m|T_r]E[s(Z_k|T_{k-1})'|T_{k-1}]\right] = 0$. The second equality follows by (1) and Lemma 7 and the definition of score. The last equality is obvious.

Second, following the steps that led to the first line on the RHS of (13), we obtain that:

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

Therefore,

$$\begin{aligned} B_{21} &= \sum_{r=a+1}^{b} \sum_{k=a}^{r-1} E\left[\frac{P(C = k|T_k)}{P(a \le C \le b)} ms(Z_r|T_{r-1})'\right] \\ &= \sum_{r=a+1}^{b} \sum_{k=a}^{r-1} E\left[ms(Z_r|T_{r-1})'|C = k\right] \frac{P(C = k)}{P(a \le C \le b)} \\ &= \sum_{k=a}^{b-1} E\left[m \sum_{r=k+1}^{b} s(Z_r|T_{r-1})' \middle| C = k\right] \frac{P(C = k)}{P(a \le C \le b)} \\ &= \sum_{k=a}^{b-1} E\left[ms(Z_b, \ldots, Z_{k+1}|T_k)'|C = k\right] \frac{P(C = k)}{P(a \le C \le b)}. \qquad (14) \end{aligned}$$

The first equality follows by (1) and Lemma 7. The second equality follows by the same steps that gave the second line on the RHS of (13). The third equality follows by interchanging the order of summations (allowed here). The last equality follows by the definition of score.

Third, we consider $B_{31}$ and note that using the definition of score in the first line below and the same argument as before in the second (last) line below give:

$$\begin{aligned} B_{31} &= \sum_{r=a}^{b}\sum_{k=1}^{r} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]s(Z_k|T_{k-1})'\right] = \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]s(T_r)'\right] \\ &= \sum_{r=a}^{b} E\left[ms(T_r)'|C=r\right]\frac{P(C=r)}{P(a\leq C\leq b)}. \qquad (15) \end{aligned}$$

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

$$\begin{aligned} B_{12} &= \sum_{r=b+1}^{R}\sum_{k=r}^{R} E\left[\frac{I(C=k)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right] \\ &= \sum_{r=b+1}^{R} E\left[\frac{1}{P(C\geq r|T_r)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\sum_{k=r}^{R}\dot{P}(C=k|T_k)'\right] \\ &= \sum_{r=b+1}^{R} E\left[\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\frac{\dot{P}(C\geq r|T_{r-1})'}{P(C\geq r|T_{r-1})}\right] \\ &= 0. \qquad (16) \end{aligned}$$

The second equality follows by (1) and Lemma 7. The third equality follows line by Lemma 7, Lemma 8 and (11). The fourth (last) equality 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 $B_{12}$ above) we obtain:

$$B_{22} = 0. \qquad (17)$$

Lastly, as before, note that:

$$B_{32} = \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(C=r|T_r)}\frac{E[m|T_r]\dot{P}(C=r|T_r)'}{P(a\leq C\leq b)}\right] = E\left[m\sum_{r=a}^{b}\frac{\dot{P}(C=r|T_r)'}{P(a\leq C\leq b)}\right]. \qquad (18)$$

Therefore, (13)-(18) imply that $E[-M^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})S(O)'] = \frac{\partial\beta^0_{[a,b]}(\eta_0)}{\partial\eta'}$. Finally, by matching the first set of terms in $-M^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ (i.e., those that correspond to line one in (10)) 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 (10)) 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 (10)) 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 $-M^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ is the efficient influence function. The expectation of the outer-product of $-M^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ gives the efficiency bound: $M^{-1}_{[a,b]}E\left[\varphi_{[a,b]}(O;\beta^0_{[a,b]})\varphi'_{[a,b]}(O;\beta^0_{[a,b]})\right]M^{-1'}_{[a,b]} = M^{-1}_{[a,b]}V_{[a,b]}M^{-1'}_{[a,b]}$. ■

**Proof of Corollary 3:** (i) Taking $R=2$ and $a=b=1$ in (10) (equivalently in (3)) gives:

$$\begin{aligned}\varphi_{[a,b]}(O;\beta) &= \frac{I(C=2)}{P(C=2|T_2)}\frac{P(C=1|T_1)}{P(C=1)}(q^0_2(T_2;\beta)-q^0_1(T_1;\beta))+\frac{I(C=1)}{P(C=1)}q^0_1(T_1;\beta)\\ &= \frac{I(C=2)}{P(C=2|T_1)}\frac{P(C=1|T_1)}{P(C=1)}(q^0_2(T_2;\beta)-q^0_1(T_1;\beta))+\frac{I(C=1)}{P(C=1)}q^0_1(T_1;\beta)\end{aligned}$$

where, since $R=2$, the second line used Lemma 8 to write $P(C=2|T_2)$ as $P(C=2|T_1)$.

(ii) Follows exactly similarly as (i).

(iii) Take $a=1, b=R$ in (3) (equivalently and more specifically in (10)), write $q^0_r$ for $q^0_r(T_r;\beta)$ for brevity, and use Lemma 8 for the third equality below to get the result:

$$\begin{aligned}\varphi_{[a,b]}(O;\beta) &= \sum_{r=2}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}P(C\leq r-1|T_{r-1})\left(q^0_r-q^0_{r-1}\right)+\sum_{r=1}^{R}I(C=r)q^0_r\\ &= \sum_{r=2}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}\left(q^0_r-q^0_{r-1}\right)\\ &\quad -\sum_{r=2}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}P(C\geq r|T_{r-1})\left(q^0_r-q^0_{r-1}\right)+\sum_{r=1}^{R}I(C=r)q^0_r\\ &= \sum_{r=2}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}\left(q^0_r-q^0_{r-1}\right)-\sum_{r=2}^{R}I(C\geq r)\left(q^0_r-q^0_{r-1}\right)+\sum_{r=1}^{R}I(C=r)q^0_r\\ &= \sum_{r=2}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}\left(q^0_r-q^0_{r-1}\right)-\left\{\sum_{r=2}^{R}I(C=r)q^0_r-I(C\geq 2)q^0_1\right\}+\sum_{r=1}^{R}I(C=r)q^0_r\\ &= \sum_{r=2}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}\left(q^0_r-q^0_{r-1}\right)+q^0_1.\ ■\end{aligned}$$

**Proof of Corollary 4:** Let $q^0_r := q^0_r(T_r;\beta)$ and $\omega_j := P(C=j)/P(a\leq C\leq b)$. Then:

$$\sum_{j=a}^{b}\omega_j\varphi_{[j,j]}(O;\beta)=\sum_{j=a}^{b}\left\{\sum_{r=j+1}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}\left(q^0_r-q^0_{r-1}\right)+\frac{I(C=j)}{P(a\leq C\leq b)}q^0_j\right\}$$

where we use for brevity the expression (10) that is equivalent to the expression (3). Hence:

$$\begin{aligned}\sum_{j=a}^{b}\omega_j\varphi_{[j,j]}(O;\beta) &= \sum_{r=b+1}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(q_r^0-q_{r-1}^0\right)+\sum_{r=a}^{b}\frac{I(C=r)}{P(a\leq C\leq b)}q_r^0\\ &\quad+\sum_{j=a}^{b}\sum_{r=j+1}^{b}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}\left(q_r^0-q_{r-1}^0\right)\end{aligned}$$

where the first term follows by (1) and Lemma 7. Match the first two terms with the terms on lines one and three of (10). Hence the demonstration will be complete if the third term in the above display is equal to the term on the second line of (10). This follows by interchanging the order of the summations (allowed here) and noting that the third term in the above display is:

$$\begin{aligned}&\sum_{j=a}^{b}\sum_{r=j+1}^{b}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}\left(q_r^0-q_{r-1}^0\right)\\ =\;&\sum_{r=a+1}^{b}\sum_{j=a}^{r-1}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}\left(q_r^0-q_{r-1}^0\right)\\ =\;&\sum_{r=a+1}^{b}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)}\left(q_r^0-q_{r-1}^0\right).\end{aligned}$$

The last equality follows by (1) and Lemma 7. This is the term on the second line of (10). ■

**Proof of Lemma 5:** The first part, i.e., $\varphi_{[a,b]}(O;\beta)=g(O;\beta,p^0(Z),q^0(Z;\beta))$ for all $O$ and $\beta$, follows from a comparison of (3) and (4). Now consider the second part. First, note that:

$$\begin{aligned}&E\left[g(O;\beta,p(Z),q^0(Z;\beta))\right]\\ =\;&\sum_{r=b+1}^{R}E\left[\frac{I(C\geq r)}{\prod_{j=1}^{r-1}(1-p_j(T_j))}\frac{\sum_{j=a}^{b}p_j(T_j)\prod_{k=1}^{j-1}(1-p_k(T_k))}{P(a\leq C\leq b)}\left[q_r^0(T_r;\beta)-q_{r-1}^0(T_{r-1};\beta)\right]\right]\\ &+\sum_{r=a+1}^{b}E\left[\frac{I(C\geq r)}{\prod_{j=1}^{r-1}(1-p_j(T_j))}\frac{\sum_{j=a}^{r-1}p_j(T_j)\prod_{k=1}^{j-1}(1-p_k(T_k))}{P(a\leq C\leq b)}\left[q_r^0(T_r;\beta)-q_{r-1}^0(T_{r-1};\beta)\right]\right]\\ &+\sum_{r=a}^{b}E\left[\frac{I(C=r)}{P(a\leq C\leq b)}q_r^0(T_r;\beta)\right].\qquad(19)\end{aligned}$$

Consider the $r-$th term on the RHS of the first line in (19). Using (1) along with Lemma 7 and the law of iterated expectations we see that, since $r = b+1, \dots R$, we will have:

$$\begin{aligned}
& E\left[E\left[\frac{1-I(C\leq r-1)}{\prod_{j=1}^{r-1}(1-p_j(T_j))}\frac{\sum_{j=a}^{b}p_j(T_j)\prod_{k=1}^{j-1}(1-p_k(T_k))}{P(a\leq C\leq b)}\left[q_r^0(T_r;\beta)-q_{r-1}^0(T_{r-1};\beta)\right]\;\middle|\; T_{r-1}\right]\right] \\
= \;& E\left[(\text{some function of } T_{r-1})\times(\text{some function of } T_b)\times E\left[q_r^0(T_r;\beta)-q_{r-1}^0(T_{r-1};\beta)\mid T_{r-1}\right]\right] \\
= \;& E\left[(\text{some function of } T_{r-1})\times(\text{some function of } T_b)\times 0\right] \;=\; 0.
\end{aligned}$$

Now consider the $r-$th term on the RHS of the second line in (19) to see that since $r = a+1, \dots b$, we will have exactly similarly as in the above display:

$$E\left[\frac{I(C\geq r)}{\prod_{j=1}^{r-1}(1-p_j(T_j))}\frac{\sum_{j=a}^{r-1}p_j(T_j)\prod_{k=1}^{j-1}(1-p_k(T_k))}{P(a\leq C\leq b)}\left[q_r^0(T_r;\beta)-q_{r-1}^0(T_{r-1};\beta)\right]\right]=0.$$

Using the above two displays, we obtain from (19) that:

$$\begin{aligned}
E\left[g(O;\beta,p(Z),q^0(Z;\beta))\right] &= \sum_{r=a}^{b}E\left[\frac{I(C=r)}{P(a\leq C\leq b)}q_r^0(T_r;\beta)\right]=\sum_{r=a}^{b}E\left[\frac{I(C=r)}{P(a\leq C\leq b)}m(Z;\beta)\right] \\
&= E[m(Z;\beta)|a\leq C\leq b] \qquad (20)
\end{aligned}$$

where the second equality follows by using (1) jointly with Lemma 7 that gives (9).

Finally we will show that $E\left[g(O;\beta,p^0(Z),q(Z;\beta))\right] = E[m(Z;\beta)|a \leq C \leq b]$. The notation will be less messy if, similar to the equivalent representation (10) of (3), we work with the following analogous equivalent representation of $g(O;\beta,p^0(Z),q(Z;\beta))$ in (4):

$$\begin{aligned}
&\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left[\frac{I(C\geq R)}{P(C\geq R|T_R)}m(Z;\beta)\right. \\
&\qquad\qquad\left.+\sum_{r=b+1}^{R-1}\left(\frac{I(C\geq r)}{P(C\geq r|T_r)}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1})}\right)q(T_r;\beta)\right] \\
&+\sum_{r=a}^{b}\left\{\left(\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1})}\frac{P(a\leq C\leq r|T_r)}{P(a\leq C\leq b)}\right)\right. \\
&\qquad\qquad\left.+\frac{I(C=r)}{P(a\leq C\leq b)}\right\}q(T_r;\beta). \qquad (21)
\end{aligned}$$

(Remarks: No generality is lost by this analogous expression of $g(O;\beta,p^0(Z),q(Z;\beta))$ because we are working with $p(Z)=p^0(Z)$. Also, in terms of the notation in this equation, $\{a\leq C\leq a-1\}$ should be considered a null event.)

First, note that the proof of Lemma 1 gives:

$$E\left[\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\frac{I(C\geq R)}{P(C\geq R|T_R)}m(Z;\beta)\right]=E[m(Z;\beta)|a\leq C\leq b].$$

Second, by using (1) and Lemma 7 note that for each $r=b+1\ldots,R-1$:

$$E\left[\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(\frac{I(C\geq r)}{P(C\geq r|T_r)}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1})}\right)q(T_r;\beta)\right]=0.$$

Lastly, following similar steps as above we note that for each $r=a\ldots,b$:

$$\begin{aligned}
&E\left[\left\{\left(\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1})}\frac{P(a\leq C\leq r|T_r)}{P(a\leq C\leq b)}\right)\right.\right.\\
&\qquad\left.\left.+\frac{I(C=r)}{P(a\leq C\leq b)}\right\}q(T_r;\beta)\right]\\
=\;&E\left[\left\{\left(\frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)}-\frac{P(a\leq C\leq r|T_r)}{P(a\leq C\leq b)}\right)+\frac{P(C=r|T_r)}{P(a\leq C\leq b)}\right\}q(T_r;\beta)\right]\\
=\;&E\left[\left\{-\frac{P(C=r|T_r)}{P(a\leq C\leq b)}+\frac{P(C=r|T_r)}{P(a\leq C\leq b)}\right\}q(T_r;\beta)\right]\\
=\;&0.
\end{aligned}$$

(21) and the above three displays give $E\left[g(O;\beta,p^0(Z),q(Z;\beta))\right]=E[m(Z;\beta)|a\leq C\leq b]$. ■

**Proof of Proposition 6:** We use Lemma 5 to systematically develop the proof of part (i) along with auxiliary results in Supplemental Appendix B.1 such that it follows directly from Theorem 5.9 in van der Vaart (1998) and Theorem 3.4 in Newey and McFadden (1994). Similarly, we develop the proof of part (ii) in Supplemental Appendix B.2 Chen et al. (2003).■

# B Supplemental Appendix B: $\widehat{\beta}$ based on equation (6)

## B.1 $\widehat{\beta}$ in (6) using parametric estimators of $p^0(Z)$ and $q^0(z;\beta)$

### B.1.1 Asymptotic properties of $\widehat{\beta}$:

With all the components of estimation in place in Section 3.3 and the idea clear, it is convenient to represent the concerned estimators as the solution of stacked estimating equations, i.e., as simple Z-estimators avoiding the notation related to profiled estimation as in (6). Accordingly, let $\gamma := (\gamma_1', \dots, \gamma_{R-1}')'$, $p(Z;\gamma) := (p_1(T_1;\gamma_1), \dots, p_{R-1}(T_{R-1};\gamma_{R-1}))'$, $\lambda := (\lambda_a', \dots, \lambda_{R-1}')'$ and $q(Z;\beta,\lambda) := (q_a'(T_a;\beta,\lambda_a), \dots, q_{R-1}'(T_{R-1};\beta,\lambda_{R-1}))'$ (i.e., without profiled representation), and define:

$$\psi(O_i;\beta,\gamma,\lambda) := \begin{bmatrix} g(O_i;\beta,p(Z_i;\gamma),q(Z_i;\beta,\lambda)) \\ S(O_i;\gamma) \\ L(O_i;\beta,\lambda) \end{bmatrix} \quad \text{where}$$

$S(O_i;\gamma) := \left[S_1'(O_i;\gamma_1), \dots, S_{R-1}'(O_i;\gamma_{R-1})\right]'$ and $L(O_i;\beta,\lambda) := \left[L_a'(O_i;\beta,\lambda_a), \dots, L_{R-1}'(O_i;\beta,\lambda_{R-1})\right]'$.

Then, note that the same $\widehat{\beta}, \widehat{\gamma} := (\widehat{\gamma}_1', \dots, \widehat{\gamma}_{R-1}')'$ and $\widehat{\lambda} := (\widehat{\lambda}_a'(\widehat{\beta}), \dots, \widehat{\lambda}_{R-1}'(\widehat{\beta}))'$ that solve, respectively, (6), the equations in (7) (for $r = 1, \dots, R-1$), and the equations in (8) with $\beta = \widehat{\beta}$ (for $r = a, \dots, R-1$), will also solve the much more familiar looking set of equations:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \psi(O_i;\beta,\gamma,\lambda). \tag{22}$$

Let $\theta := (\beta', \gamma', \lambda')' \in \Theta \subseteq \mathbb{R}^{d_\theta}$ where $d_\theta = d + \sum_{r=1}^{R-1} d_{\gamma_r} + \sum_{r=a}^{R-1} d_{\lambda_r}$. $d_\theta$ is also the number of estimating equations in the stacked system in (22). $\widehat{\theta} := (\widehat{\beta}', \widehat{\gamma}', \widehat{\lambda}')'$ is a solution of (22).

The representation in (22) allows us to obtain the following asymptotic results under the standard framework of Newey and McFadden (1994) or Chapter 5 of van der Vaart (1998). Consequently, we will state the general results from these references as propositions and then highlight the specialities specific to our framework as corollaries and associated remarks.

**Proposition 9** *Let* $\theta^* := (\beta^{*'}, \gamma^{*'}, \lambda^{*'})' \in \Theta$ *satisfy* $\inf_{\theta:\|\theta-\theta^*\|\geq\epsilon} \|E[\psi(O;\theta)]\| > 0 = E[\psi(O;\theta^*)]$ *for every* $\epsilon > 0$. *Let* $\sup_{\theta\in\Theta} \|n^{-1}\sum_{i=1}^n \psi(O_i;\theta) - E[\psi(O;\theta)]\| = o_p(1)$. *Then* $\widehat{\theta} \xrightarrow{p} \theta^*$.

**Assumption CH:** The conditional hazard (CH) model is correct, i.e., there exists a $\gamma^0 = (\gamma_1^{0'}, \ldots, \gamma_{R-1}^{0'})'$ with $\gamma_r^0 \in \Gamma_r$ such that $p_r(T_r;\gamma_r^0) = p_r^0(T_r)$ for $r = 1, \ldots, R-1$.

**Assumption CE:** The conditional expectation (CE) model is correct, i.e., there exists a $\lambda^0 = (\lambda_a^{0'}, \ldots, \lambda_{R-1}^{0'})'$ with $\lambda_r^0 \in \Lambda_r$ such that $q_r(T_r;\beta_{[a,b]}^0, \lambda_r^0) = q_r^0(T_r;\beta_{[a,b]}^0)$ for $r = a, \ldots, R-1$.

**Corollary 10** *Let* (1), (2) *and the conditions of Proposition 9 hold. If assumption CH holds, then* $\gamma^* = \gamma^0$ *and hence* $\beta^* = \beta_{[a,b]}^0$. *If assumption CE holds, then* $\lambda^* = \lambda^0 := \lambda^0(\beta_{[a,b]}^0)$ *and hence* $\beta^* = \beta_{[a,b]}^0$. *If neither of assumptions CH or CE holds then, in general,* $\beta^* \neq \beta_{[a,b]}^0$.

**Remarks:** Proposition 9 is Theorem 5.9 in van der Vaart (1998). Corollary 10 uses the result of Lemma 5 to show that the probability limit $\beta^*$ of $\widehat{\beta}$ is actually the parameter value of interest $\beta_{[a,b]}^0$ if either the conditional hazard or the conditional expectation model is correct.

**Proposition 11** *Let: (i)* $\widehat{\theta} \xrightarrow{p} \theta^* \in interior(\Theta)$, *(ii)* $\psi(O;\theta)$ *be continuously differentiable in* $\theta$ *in an open neighborhood* $\mathcal{N}$ *of* $\theta^*$, *(iii)* $E[\sup_{\theta\in\mathcal{N}} \|\partial\psi(O;\theta)/\partial\theta'\|] < \infty$, *(iv)* $\Psi(\theta) := E[\partial\psi(O;\theta)/\partial\theta']$ *be nonsingular at* $\theta = \theta^*$, *and (v)* $n^{-1/2}\sum_{i=1}^n \psi(O_i;\theta^*) \xrightarrow{d} N(0, \Sigma(\theta^*))$ *where* $\Sigma(\theta) := Var(\psi(O;\theta))$. *Define* $\Upsilon(\theta) := \Psi^{-1}(\theta)\Sigma(\theta)\Psi^{-1'}(\theta)$ *when it exists. Then:*

$$\sqrt{n}(\widehat{\theta} - \theta^*) = -\Psi^{-1}(\theta^*)\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(O_i;\theta^*) + o_p(1) \xrightarrow{d} N(0, \Upsilon(\theta^*)).$$

**Corollary 12** *Let* (1), (2), *assumptions CH and CE, and the conditions of Proposition 11 hold. Then,* $\widehat{\beta}$ *is the efficient estimator in the sense of Proposition 2, i.e.,*

$$\sqrt{n}(\widehat{\beta} - \beta_{[a,b]}^0) = -M_{[a,b]}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_{[a,b]}(O_i;\beta_{[a,b]}^0) + o_p(1) \xrightarrow{d} N\left(0, \Omega_{[a,b]}\right).$$

**Remarks:** Proposition 11 is Theorem 3.4 in Newey and McFadden (1994). The asymptotic variance of $\widehat{\beta}$ is $\Upsilon_{\beta\beta}(\theta^*)$ where $\Upsilon_{\beta\beta}(\theta)$ is the $d \times d$ upper-left (northwest) block of the matrix $\Upsilon(\theta)$ when the latter exists. Corollary 12 demonstrates that when both nuisance parameter models are correct, then $\Upsilon_{\beta\beta}(\theta^*) = \Omega_{[a,b]}$, i.e., the asymptotic variance reaches the efficiency bound in Proposition 2 and $\widehat{\beta}$ is indeed the asymptotically unbiased efficient estimator.

Note that, we avoided the dependence on $\theta^*$ in the notation used in Proposition 6 in Section 3.3. This was because $\theta^*$ was not introduced in the main text for brevity. To avoid confusion, it may be useful to write with a bit of abuse of notation $\Upsilon_{\beta\beta}(\theta^*)$ as $\Upsilon_{\beta\beta}$.

As evident from Corollary 10 and Proposition 11, $\widehat{\beta}$ remains asymptotically unbiased for $\beta^0_{[a,b]}$ (but not efficient) even if only one of assumptions CH and CE holds. Proposition 11 can be used for asymptotic inference on $\beta^0_{[a,b]}$ in all such cases. While the stacked representation in (22) makes estimation of $\Upsilon_{\beta\beta}(\theta^*)$ straightforward (see below), it is still instructive to study the structure of $\Upsilon_{\beta\beta}(\theta^*)$ when either assumption CH or assumption CE (but not both) holds. These structures and more are obtained as byproducts in the proof of Corollary 12 and are described for the interested reader in the remark below that proof in Appendix B.3.

### B.1.2 Estimation of the asymptotic variance of $\widehat{\beta}$:

The standard estimator for $\Upsilon_{\beta\beta}(\theta^*)$ is $\widehat{\Upsilon}_{\beta\beta}(\widehat{\theta})$, the $d \times d$ upper-left (northwest) block of:

$$\widehat{\Upsilon}(\widehat{\theta}) = \widehat{\Psi}^{-1}(\widehat{\theta})\widehat{\Sigma}(\widehat{\theta})\widehat{\Psi}^{-1'}(\widehat{\theta}) \text{ where } \widehat{\Psi}(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta'}\psi(O_i;\widehat{\theta}) \text{ and } \widehat{\Sigma}(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\psi(O_i;\widehat{\theta})\psi'(O_i;\widehat{\theta}).$$

**Proposition 13** *Let the conditions of Proposition 11 hold. Let* $E[\sup_{\theta\in\mathcal{N}}\|\psi(O;\theta)\|^2] < \infty$ *in an open neighborhood* $\mathcal{N}$ *of* $\theta^*$. *Then* $\widehat{\Upsilon}(\widehat{\theta}) \xrightarrow{p} \Upsilon(\theta^*)$ *and hence* $\widehat{\Upsilon}_{\beta\beta}(\widehat{\theta}) \xrightarrow{p} \Upsilon_{\beta\beta}(\theta^*)$ .

**Remark:** Proposition 13 is Theorem 4.5 of Newey and McFadden (1994). Taking numerical derivatives instead of analytical derivatives of the first $d$ rows of $\psi(O;\theta)$ with respect to $\gamma$ can sometimes be useful in practice, and this also applies in the case of IPW. Using bootstrap to directly estimate $\Upsilon(\theta^*)$ or $\Upsilon_{\beta\beta}(\theta^*)$ also seems to be a convenient common practice.

## B.2 $\widehat{\beta}$ in (6) using nonparametric estimators of $p^0(Z)$ and $q^0(z;\beta)$

If the sample size is not small then one can, in theory, (asymptotically) avoid the problems related to misspecification by instead using nonparametric estimators $\widehat{p}(Z)$ and $\widehat{q}(Z;\beta)$. This is not difficult in practice and we described it for series/sieve estimators in the main text. One can also use other methods, e.g., kernel, local polynomial, etc. We will abstract from such specific nonparametric estimators and instead consider generic estimators $\widehat{p}(Z)$ and $\widehat{q}(Z;\beta)$ satisfying standard restrictions on convergence, and, for completeness, describe the asymptotic properties of the resulting $\widehat{\beta}$ following Chen et al. (2003).

Preliminaries: Recalling the definitions of the $g(.)$ function in (4), the nuisance parameters in (5), and the estimating equations for $\beta$ in (6), we will follow Chen et al. (2003) and write:

$$G_n(\beta,p,q) := \frac{1}{n}\sum_{i=1}^{n} g(O_i;\beta,p(Z_i),q(Z_i,\beta)) \text{ and } G(\beta,p,q) := E[g(O;\beta,,p(Z),q(Z,\beta))].$$

We will follow this convention with notation also when the true values $p^0(Z)$ and $q^0(Z;\beta)$ or the estimators $\widehat{p}(Z)$ and $\widehat{q}(Z;\beta)$ are plugged in for the generic functions $p(Z)$ and $q(Z;\beta)$. Recall that $p(Z) = (p_1(T_1), p_2(T_2), \ldots, p_{R-1}(T_{r-1}))$ and $q(Z;\beta) := (q_a'(T_a;\beta), \ldots, q_{R-1}'(T_{r-1};\beta))'$. If not confusing, we will drop the arguments and instead write $p$ for $p(Z)$, $q$ for $q(X;\beta)$, etc. Let $p_r(T_r) \in \mathcal{P}_r \in (0,1)$ and $q_r(T_r;\beta) \in \mathcal{Q}_r$ for all $\beta \in \mathcal{B}$ where $\mathcal{P}_r$ for $r = 1,\ldots,R-1$ and $\mathcal{Q}_r$ for $r = a,\ldots,R-1$ are vector spaces of continuous functions of $Z$. Let $\mathcal{P} = \mathcal{P}_1 \times \ldots \times \mathcal{P}_{R-1}$ and $\mathcal{Q} = \mathcal{Q}_a \times \ldots \times \mathcal{Q}_{R-1}$ be endowed with the sup-norm metric denoted as follows: $\|p\|_\infty = \sup_{z\in\mathcal{S}_z}\|p(z)\|$, $\|q(.;\beta)\|_\infty = \sup_{z\in\mathcal{S}_z}\|q(z;\beta)\|$ and $\|q\|_\infty = \sup_{\beta,\mathcal{B},z\in\mathcal{S}_z,}\|q(z;\beta)\|$ where $\mathcal{S}_z$ denotes the support of $Z$ and, as in Section 3.4.2, $\|.\|$ denotes the Euclidean norm. For any $\delta > 0$, let $\mathcal{B}_\delta := \{\beta \in \mathcal{B} \ : \ \|\beta - \beta^0_{[a,b]}\| \le \delta\}$, $\mathcal{P}_\delta := \{p \in \mathcal{P} \ : \ \|p - p^0\|_\infty \le \delta\}$, and $\mathcal{Q}_\delta := \{q \in \mathcal{Q} \ : \ \|q - q^0\|_\infty \le \delta\}$. Finally, following Chen et al. (2003), we slightly generalize the definition of $\widehat{\beta}$ as an approximate solution of (6) by defining it as:

$$\|G_n(\widehat{\beta},\widehat{p},\widehat{q})\| \le \inf_{\beta\in\mathcal{B}}\|G_n(\beta,\widehat{p},\widehat{q})\| + o_p(n^{-1/2}). \tag{23}$$

**Proposition 14** *Let (1), (2) and assumptions A1 and A2 hold. Let:*

*(i) uniformly in $\beta \in \mathcal{B}$, $G(\beta, p, q)$ is continuous with respect to $\|.\|_\infty$ in $p, q$ at $p = p^0, q = q^0$,*

*(ii) $\|\widehat{p} - p^0\|_\infty = o_p(1)$ and $\|\widehat{q} - q^0\|_\infty = o_p(1)$,*

*(iii) for all sequence of positive numbers $\{\delta_n\}$ with $\delta_n = o(1)$,*

$$\sup_{\beta \in \mathcal{B}, p \in \mathcal{P}_{\delta_n}, q \in \mathcal{Q}_{\delta_n}} \frac{\|G_n(\beta, p, q) - G(\beta, p, q)\|}{1 + \|G_n(\beta, p, q)\| + \|G(\beta, p, q)\|} = o_p(1).$$

*Then $\widehat{\beta}$ defined in (23) satisfies: $\widehat{\beta} = \beta^0_{[a,b]} + o_p(1)$.*

**Proposition 15** *Let (1), (2) and assumption A hold. Let:*

*(i) $\widehat{\beta}$ defined in (23) satisfies: $\widehat{\beta} = \beta^0_{[a,b]} + o_p(1)$,*

*(ii) $\partial G(\beta, p^0, q^0)/\partial \beta'$ exists for $\beta \in \mathcal{B}_\delta$ for some $\delta > 0$ and is continuous at $\beta = \beta^0_{[a,b]}$,*

*(iii) $\|G(\beta, p, q) - G(\beta, p^0, q^0)\| \leq c \times \{\|p - p^0\|^2_\infty + \|q - q^0\|^2_\infty\}$ for a constant $c \geq 0$, for all $\beta \in \mathcal{B}_{\delta_n}$, $p \in \mathcal{P}_{\delta_n}$ and $q \in \mathcal{Q}_{\delta_n}$ with a positive sequence $\delta_n = o(1)$,*

*(iv) $\widehat{p} \in \mathcal{P}$ and $\widehat{q} \in \mathcal{Q}$ with probability tending to one; and $\|\widehat{p} - p^0\|_\infty = o_p(n^{-1/4})$ and $\|\widehat{q} - q^0\|_\infty = o_p(n^{-1/4})$,*

*(v) for all sequence of positive numbers $\{\delta_n\}$ with $\delta_n = o(1)$,*

$$\sup_{\beta \in \mathcal{B}_{\delta_n}, p \in \mathcal{P}_{\delta_n}, q \in \mathcal{Q}_{\delta_n}} \frac{\sqrt{n}\|G_n(\beta, p, q) - G(\beta, p, q) - G_n(\beta^0_{[a,b]}, p^0, q^0)\|}{1 + \sqrt{n}\|G_n(\beta, p, q)\| + \sqrt{n}\|G(\beta, p, q)\|} = o_p(1).$$

*(vi) $V_{[a,b]} = Var\left(\varphi_{[a,b]}(O; \beta^0_{[a,b]})\right)$ exists where $\varphi_{[a,b]}(O; \beta^0_{[a,b]})$ is defined in (3).*

*Then $\widehat{\beta}$ defined in (23) is asymptotically efficient in the sense of Proposition 2 and satisfies:*

$$\sqrt{n}\left(\widehat{\beta} - \beta^0_{[a,b]}\right) = -M^{-1}_{[a,b]} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{[a,b]}(O_i; \beta^0_{[a,b]}) + o_p(1) \xrightarrow{d} N\left(0, \Omega_{[a,b]} = M^{-1}_{[a,b]} V_{[a,b]} M^{-1\prime}_{[a,b]}\right).$$

**Remarks:** The high level conditions in Propositions 14 and 15 are from Theorems 1 and 2 of Chen et al. (2003) who also discuss the primitive sufficient conditions for them. Our proof of these propositions below involves showing that the remaining conditions in Theorems 1 and 2 in Chen et al. (2003) are also satisfied under our setup thanks to (1) and (2). It is possible to somewhat weaken the assumption on the rates of convergence in condition (iv) of Proposition 15; see, e.g., Remark 2(iii) in Chen et al. (2003) and condition (4.14)'(ii) in Chen (2007). This has been implemented exploiting the structure of the estimating equations under our setup in, e.g., Cattaneo (2010) and Rothe and Firpo (2019), and hence is not pursued here.

We also do not formally describe the estimation of asymptotic variance of $\widehat{\beta}$ for brevity, but we note that this can be obtained exactly as in Theorems 6 and 7 in Cattaneo (2010). However, based on our limited experience, we wish to remark that the promise of nonparametrics in attaining the efficiency bound may not materialize in small samples. (By this we mean that in simulations under various setups we found the Monte Carlo variance to be systematically larger than the estimated efficiency bound, although this problem is much worse for IPW estimators than the estimator in (6).) Hence it is our opinion that, if possible (e.g. when using series/sieve estimators), it is safer to estimate the asymptotic variance using the parametric formula in Supplemental Appendix B.1.2; also see Ackerberg et al. (2012). That way, the estimated variance well better reflect the true variability of the estimator; and if the sample size is genuinely large then this estimated variance will anyway come close (in probability) to the efficiency bound as the promise of nonparametrics materializes.

### B.3 Proofs of the results in Supplemental Appendix B.1-B.2

**Proof of Proposition 9:** This is Theorem 5.9 in van der Vaart (1998). ■

**Proof of Corollary 10:** The proof follows from Proposition 9 by using Lemma 5.

First, take the case where assumption CH holds. Hence, the population version of the estimating equations in (22) for $\beta$ is $E[g(O;\beta,p^0(Z),q(Z;\beta,\lambda))]=0$ for some function $q(Z;\beta,\lambda)$ that may not be $q^0(Z;\beta)$ defined in (5). However, Lemma 5 and equation (2) jointly imply that $\beta^0_{[a,b]}$ is still the unique solution for $\beta$ that solves these equations. Hence, in this case $\beta^*=\beta^0_{[a,b]}$ and $\gamma^*=\gamma^0$ along with some $\lambda^*$ constitute the well-separated $\theta^*$

defined in Proposition 9. Hence $\widehat{\beta} \xrightarrow{p} \beta^0_{[a,b]}$ under the conditions of Proposition 9.

Now, take the case where assumption CE holds. Hence, the population version of the estimating equations in (22) for $\beta$ is $E[g(O;\beta,p(Z),q^0(Z;\beta))] = 0$ for some function $p(Z)$ that may not be $p^0(Z)$ defined in (5). This follows because $\lambda^0(\beta)$ satisfying $q(Z;\beta^0_{[a,b]},\lambda^0(\beta^0_{[a,b]})) = q^0(Z;\beta^0_{[a,b]})$ solves at each $\beta$ the population version of the estimating equations in (22) for $\lambda$ under assumption CE. However, Lemma 5 and equation (2) jointly imply that $\beta^0_{[a,b]}$ is the unique solution for $\beta$ that solves these equations. Hence, in this case $\beta^* = \beta^0_{[a,b]}$ and $\lambda^* = \lambda^0(\beta^0_{[a,b]})$ along with some $\gamma^*$ constitute the well-separated $\theta^*$ defined in Proposition 9. Hence $\widehat{\beta} \xrightarrow{p} \beta^0_{[a,b]}$ under the conditions of Proposition 9. ■

**Proof of Proposition 11:** This is Theorem 3.4 in Newey and McFadden (1994). ■

**Proof of Corollary 12:** The key to this proof is to utilize Lemma 5 to impose structure on the matrix $\Psi(\theta^*)$ defined in the statement of Proposition 11. Partitioning the rows of $\Psi(\theta)$ (when it exists) according to the dimension of the vectors $g(.)$, $S(.)$ and $L(.)$ in (22) and the columns of $\Psi(\theta)$ according to the dimension of the vectors $\beta$, $\gamma$ and $\lambda$, we write:

$$\Psi(\theta) = \begin{bmatrix} \Psi_{g,\beta}(\theta) & \Psi_{g,\gamma}(\theta) & \Psi_{g,\lambda}(\theta) \\ \Psi_{S,\beta}(\theta) & \Psi_{S,\gamma}(\theta) & \Psi_{S,\lambda}(\theta) \\ \Psi_{L,\beta}(\theta) & \Psi_{L,\gamma}(\theta) & \Psi_{L,\lambda}(\theta) \end{bmatrix} = \begin{bmatrix} \Psi_{g,\beta}(\theta) & \Psi_{g,\gamma}(\theta) & \Psi_{g,\lambda}(\theta) \\ 0 & \Psi_{S,\gamma}(\theta) & 0 \\ \Psi_{L,\beta}(\theta) & 0 & \Psi_{L,\lambda}(\theta) \end{bmatrix}$$

where the second equality follows by inspection. Therefore, using the formula for partitioned inverse, the asymptotically linear representation of $\widehat{\theta}$ in Proposition 11 implies that:

$$\begin{aligned} \sqrt{n}(\widehat{\beta} - \beta^*) &= -[\Psi_{g,\beta}(\theta^*) - \Psi_{g,\lambda}(\theta^*)\Psi^{-1}_{L,\lambda}(\theta^*)\Psi_{L,\beta}(\theta^*)]^{-1} \\ &\quad \times \left[\; I_d, \quad -\Psi_{g,\gamma}(\theta^*)\Psi^{-1}_{S,\gamma}(\theta^*), \quad -\Psi_{g,\lambda}(\theta^*)\Psi^{-1}_{L,\lambda}(\theta^*) \;\right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(O_i;\theta^*) + o_p(1) \quad (24) \end{aligned}$$

This is the general result. We will now analyze the structure in (24) under the scenarios where assumption CH or CE holds or both hold. It is useful to keep in mind that $\Psi(\theta^*) = \partial E[\psi(O;\theta^*)]/\partial\theta'$ under our condition (iii) by Lemma 3.6 of Newey and McFadden (1994).

Therefore, when assumption CH holds we have by Corollary 10 and Lemma 5 that $\Psi_{g,\lambda}(\theta^*) = 0$, and hence Corollary 10 and (24) imply that:

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta} - \beta^0_{[a,b]}\right) &= -\Psi^{-1}_{g,\beta}(\theta^*)\left[\begin{array}{ccc} I_d, & -\Psi_{g,\gamma}(\theta^*)\Psi^{-1}_{S,\gamma}(\theta^*), & 0 \end{array}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(O_i;\theta^*) + o_p(1) \\
&= -\Psi^{-1}_{g,\beta}(\theta^*)\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(O_i;\beta^0_{[a,b]}, p^0(Z_i), q(Z_i;\beta^0_{[a,b]},\lambda^*))\right. \\
&\qquad \left. - \ \Psi_{g,\gamma}(\theta^*)\Psi^{-1}_{S,\gamma}(\theta^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S(O_i;\gamma^0)\right\} + o_p(1) \qquad (25)
\end{aligned}$$

where $\theta^*$ on the RHS of (25) is now $\theta^* = (\beta^{0'}_{[a,b]}, \gamma^0, \lambda^{*'})'$ for some $\lambda^*$.

Similarly, when assumption CE holds we have by Corollary 10 and Lemma 5 that $\Psi_{g,\gamma}(\theta^*) = 0$, and hence Corollary 10 and (24) imply that:

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta} - \beta^0_{[a,b]}\right) &= -\Psi^{-1}_{g,\beta}(\theta^*)\left[\begin{array}{ccc} I_d, & 0, & -\Psi_{g,\lambda}(\theta^*)\Psi^{-1}_{L,\lambda}(\theta^*) \end{array}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(O_i;\theta^*) + o_p(1) \\
&= -\Psi^{-1}_{g,\beta}(\theta^*)\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(O_i;\beta^0_{[a,b]}, p(Z_i;\gamma^*), q^0(Z_i;\beta^0_{[a,b]}))\right. \\
&\qquad \left. - \ \Psi_{g,\lambda}(\theta^*)\Psi^{-1}_{L,\lambda}(\theta^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} L(O_i;\beta^0_{[a,b]}, \lambda^0(\beta^0_{[a,b]}))\right\} + o_p(1) \qquad (26)
\end{aligned}$$

where $\theta^*$ on the RHS of (26) is now $\theta^* = (\beta^{0'}_{[a,b]}, \gamma^*, \lambda^{0'}(\beta^0_{[a,b]}))'$ for some $\gamma^*$. Note that, to avoid any confusion, we follow the statement of Corollary 10 here and write the explicit form $\lambda^0 := \lambda^0(\beta^0_{[a,b]})$ instead of simply writing $\lambda^0$ as in assumption CE that is focused at $\beta^0_{[a,b]}$.

Finally, when both assumptions CH and CE hold we have by Corollary 10 and Lemma 5 that $\Psi_{g,\gamma}(\theta^*) = 0$ and $\Psi_{g,\lambda}(\theta^*) = 0$, and hence Corollary 10 and (24) imply that:

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta} - \beta^0_{[a,b]}\right) &= -\Psi^{-1}_{g,\beta}(\theta^*)\left[\begin{array}{ccc} I_d, & 0, & 0 \end{array}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(O_i;\theta^*) + o_p(1) \\
&= -\Psi^{-1}_{g,\beta}(\theta^0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(O_i;\beta^0_{[a,b]}, p^0(Z_i), q^0(Z_i;\beta^0_{[a,b]})) + o_p(1) \\
&= -M^{-1}_{[a,b]}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{[a,b]}(O_i;\beta^0_{[a,b]}) + o_p(1) \qquad (27)
\end{aligned}$$

where the second line on the RHS recognizes that $\theta^* = \theta^0$ $(= (\beta^{0\prime}_{[a,b]}, \gamma^{0\prime}, \lambda^{0\prime}(\beta^0_{[a,b]}))')$ now giving $p(Z;\gamma^0) = p^0(Z)$ and $q(Z;\beta^0_{[a,b]}, \lambda^0(\beta^0_{[a,b]})) = q^0(Z;\beta^0_{[a,b]})$. Using this, the third (last) line recognizes that $g(O;\beta^0_{[a,b]}, p^0(Z), q^0(Z;\beta^0_{[a,b]})) = \varphi_{[a,b]}(O;\beta^0_{[a,b]})$ and hence $\Psi_{g,\beta}(\theta^0) = M_{[a,b]}$. Hence, here we have the asymptotic variance of $\widehat{\beta}$ as $\Omega_{[a,b]}$ defined in Proposition 2. ■

**Remarks:** As stated in the statement of the corollary, it is clear from (27) that estimation of the nuisance parameters $\gamma$ and $\lambda$ for the conditional hazard and conditional expectation models does not have any effect on the asymptotic variance of $\widehat{\beta}$ if both assumptions CH and CE hold. Therefore, $\widehat{\beta}$ is efficient in this case where there is no parametric misspecification of these nuisance parameters.

However, as evident from the general result in (24), the estimation of the nuisance parameters $\gamma$ and $\lambda$ generally affects the asymptotic variance of $\widehat{\beta}$ when neither assumption CH nor assumption CE holds, and in that case the probability limit of $\widehat{\beta}$ is generally not $\beta^0_{[a,b]}$. Nevertheless, when assumption CH holds, (25) reveals that the estimation of $\lambda$ for the conditional expectations, i.e, the other set of nuisance parameters, does not affect the asymptotic variance of $\widehat{\beta}$ because the sampling variability of the estimating equations for $\lambda$ does not enter the influence function for $\widehat{\beta}$. Similarly, when assumption CE holds, (26) reveals that the estimation of $\gamma$ for the conditional hazards, i.e, the other set of nuisance parameters, does not affect the asymptotic variance of $\widehat{\beta}$ because the sampling variability of the estimating equations for $\gamma$ does not enter the influence function for $\widehat{\beta}$. These nice features related to the nuisance parameters are an artifact of the double-robustness property that ensured earlier in Corollary 10 that $\widehat{\beta} \xrightarrow{P} \beta^0_{[a,b]}$ in both these cases.

**Proof of Proposition 13:** This is Theorem 4.5 of Newey and McFadden (1994).

**Remark:** While the stacked representation in (22) makes the estimation of asymptotic variance a routine job, it is useful to remember the above remark on this asymptotic variance.

**Proof of Proposition 14:** The proof follows by verifying the conditions of Theorem 1 in Chen et al. (2003). Our conditions (i), (ii), (iii) and (iv) are, respectively, conditions (1.1), (1.3), (1.4) and (1.5) of Chen et al. (2003). Under (1) and (2), our Lemma 5 implies that $G(\beta, p^0, q^0) = 0$ if and only if $\beta = \beta^0_{[a,b]}$. This implies that not only is $G(\beta^0, p^0, q^0) = 0$ but also condition (1.2) of Chen et al. (2003) holds. Hence the final result follows. ■

**Proof of Proposition 15:** The proof follows by verifying the conditions of Theorem 2 in Chen et al. (2003). Our conditions (i), (ii), (iv) and (v) are, respectively, conditions (2.1), (2.2)(i), (2.4) and (2.5) of Chen et al. (2003). Our Lemma 5 implies that $\partial G(\beta, p^0, q^0)/\partial \beta' = M_{[a,b]}$ and hence assumption (A3) now implies condition (2.2) (ii) of Chen et al. (2003). Again, our Lemma 5 implies that the pathwise derivative of $G(\beta, p, q)$ with respect to $p$ and $q$ at any $\beta \in \mathcal{B}$ and $p = p^0, q = q^0$ exists in all direction and is indeed 0. With this is mind, our condition (iii) implies condition (2.3)(i) of Chen et al. (2003). On the other hand, the zero pathwise derivative at any $\beta \in \mathcal{B}$ implies that condition (2.3)(ii) of Chen et al. (2003) automatically holds since it makes the LHS of (2.3)(ii) 0. Finally, (2) and Lemma 5 imply that $E[g(O; \beta^0_{[a,b]}, p^0, q^0)] = 0$ where $g(O; \beta^0_{[a,b]}, p^0, q^0) = \varphi_{[a,b]}(O; \beta^0)$. Therefore, we now have by assumption (A1) and condition (vi) that condition (2.6) of Chen et al. (2003) holds with the important additional point that the estimation of the nuisance parameters has no effect on the asymptotic variance. These verifications imply that the final result follows. ■

# C Supplemental Appendix C: Monte Carlo experiment

We will now study the small-sample properties of our proposed estimator and inference based on it. For reference, we also present the same small-sample properties of the IPW estimator: (i) the Narain (1951), Horvitz and Thompson (1952) version that is commonly used in economics in similar contexts of attrition (e.g., Fitzgerald et al. (1996), Abowd et al. (2001), Wooldridge (2002), Nicoletti (2006), Wooldridge (2010), etc.), and (ii) the Hajek (1971) version that is often recommended (e.g., Hirano and Imbens (2001), Lunceford and Davidian (2004), Busso et al. (2009, 2014), etc.). The estimands considered here are similar to the motivating estimand in Section 2 and can be viewed as the components of the treatment effect estimands considered in our empirical illustration in Supplemental Appendix D.

## C.1 Simulation design

We will consider a setup reflecting the individual's decision to stay or leave dynamically over periods from programs (e.g., smoking cessation, weight loss), school, job, marriage, experiments, surveys, market, etc. We model this decision to leave after any period as a simple comparison between the individual's expectation of the outcome and their actual outcome after that period. Accordingly, we will consider an $R$-period program where $Y_r$ is the outcome from staying until the end of the $r$-th period for $r = 1, \ldots, R$ in the program. We will assume that this outcome is generated as follows. For $t = 1, \ldots, T$, let:

$$Y_t = \frac{1}{2}Y_{t-1} + \frac{1}{4}Y_{t-2} + \frac{1}{4}X_t + e_t, \quad \text{where} \quad X_t = X_{t-1} + v_t. \tag{28}$$

$e_t$ and $v_t$ are the model errors.[7] Take $X_0, Y_{-1}, Y_0$ independently $N(1, 1)$ as the initial state. Our analysis below is not conditional on the initial state, but this could be done. We will take $R = T = 3$, and let $X_r$ be the other observed variables for the $r$-th period for $r = 1, \ldots, R$.

Let the individual's expectation for the outcome in the $r$-th period be $Y_r^*$. Suppose that the individual decides to leave the program at the end of the $r$-th period, conditional on staying until then, if and only if the actual outcome exceeds the expectation, i.e., $Y_r^* < Y_r$.[8] In other words, let:

$$\begin{aligned} I(C = r) = &\ I(Y_r^* < Y_r) \prod_{j=1}^{r-1} I(Y_j^* \geq Y_j) \ \text{ for } r = 1, \ldots, R-1, \text{ while} \\ I(C = R) = &\ 1 - \sum_{r=1}^{R-1} I(C = r). \end{aligned} \tag{29}$$

The researcher observes $C$ but not $Y_r^*$. This means that $Z_1 = (Y_{-1}, Y_0, Y_1, X_{-1}, X_0, X_1)'$,

[7] Estimation of regression coefficients in the case of attrition under some form of MAR in a dynamic panel data models with fixed effects has been studied; see, e.g., Abrevaya (2019) and the references therein.

[8] Depending on the context, the decision to leave if $Y_r^* > Y_r$ might be a more sensible modeling choice.

$Z_2 = (Y_2, X_2)'$ and $Z_3 = (Y_3, X_3)'$ in our notation. So, the observables are $T_1 = Z_1$, $T_2 = (Z_1', Z_2')'$ and $T_3 = (Z_1', Z_2', Z_3')'$ for those with $C = 1$, $C = 2$ and $C = 3$ respectively.

Our distributional assumptions on the data generating process (DGP) are as follows. $e_t$ and $v_t$ are i.i.d. $N(0, 1)$ for all $t$. $u_r := Y_r^* - Y_r$ is i.i.d. $N(0, (2.5)^2)$ for all $r$. MAR in (1) is imposed by maintaining that $e_t, v_t, u_r, X_0, Y_{-1}, Y_0$ are mutually independent for all $t, r$. This results in roughly 50% of the individuals with $C = 1$, 26% with $C = 2$, and 24% with $C = 3$.

To define $\beta^0_{[a,b]}$, we take the moment function in (2) as $m(Z; \beta) = Y_3 - \beta$, and consider the six different targets $[a, b] = [1, 3], [1, 1], [2, 2, ], [3, 3], [1, 2]$ and $[2, 3]$, giving six different parameters of interest. 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, and then averaging each mean over 10,000 Monte Carlo trials. Accordingly, the six different "true values", i.e., $\beta^0_{[a,b]}$'s are: 1, 1.1709, .9617, .6858, 1.0994 and .8291 respectively. As evident from Table 1, the error in this approximation is of a rather small order to seriously affect our subsequent analysis below that is conducted with far smaller (than 10 million) sample size.

| Target Population | Descriptive Statistics | | | | | |
|---|---|---|---|---|---|---|
| $[a, b]$ for $\beta$ | Mean | $(10^{-3}\times)$ Std | Median | IQR | Min | Max |
| $[1, 3]$ | 1 | .4860 | 1 | .0007 | .9982 | 1.0017 |
| $[1, 1]$ | 1.1709 | .6841 | 1.1709 | .0009 | 1.1682 | 1.1735 |
| $[2, 2]$ | .9617 | .9430 | .9617 | .0013 | .9581 | .9648 |
| $[3, 3]$ | .6858 | .9769 | .6858 | .0013 | .6817 | .6895 |
| $[1, 2]$ | 1.0994 | .5536 | 1.0994 | .0007 | 1.0975 | 1.1012 |
| $[2, 3]$ | .8291 | .6800 | .8291 | .0009 | .8265 | .8316 |

Table 1: $\beta^0_{[a,b]}$ 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. Columns 3-7 list the standard deviation (Std), interquartile range (IQR), minimum (Min) and maximum (Max) of the estimator.

## C.2 Simulation results

We report all the simulation results based on 10,000 Monte Carlo trials. We consider sample sizes $n = 100, 200$ and 500.

The target parameter $\beta^0_{[3,3]}$ is not interesting since the so-called complete-case estimator $\sum_{i=1}^n I(C_i = 3)Y_{3i}/\sum_{j=1}^n I(C_j = 3)$ is the efficient estimator for $\beta^0_{[3,3]}$ (see Section 3.2.2). This estimator is neither efficient nor consistent for the other targets. Table 2 summarizes the performance of the complete-case estimator and, as expected, this performance is poor and misleading for all the target sub-populations except for the one with $[a, b] = [3, 3]$.

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

Table 2: 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 $[a, b] = [3, 3]$. Std stands for the Monte Carlo standard deviation, and since the estimator is identical for all sub-populations, 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.

So, we focus on the five remaining target parameters $\beta^0_{[a,b]}$. We compute our proposed estimator, denoted as EFF henceforth, following the descriptions in Section 3.3. To estimate the nuisance parameters, we postulate probit models for the conditional hazards and linear model for the conditional expectations. For both, we specify the respective index function as linear in the associated conditioning variables $T_1$, $T_2$, etc. and do not include interactions.

We report in Table 3: (i) the mean bias, referred to as Bias; (ii)-(iii) the average of the estimated standard deviation, referred to as AS Std and MC Std, based on the asymptotic variance formula (in Supplemental Appendix B.1.2) and Monte Carlo respectively; and (iv)-(v) the empirical size, referred to as AS Size and MC Size, of the two-sided 5% test of $H_0 : \beta_{[a,b]} = \beta^0_{[a,b]}$ based on t-ratios using the estimated standard deviation based on the asymptotic variance formula (in Supplemental Appendix B.1.2) and MC Std respectively. Our proposed estimator, EFF, performs quite well in all these aspects and for all the target sub-populations even when the sample size is relatively small.

To put the performance of EFF into context, we also report in Table 3 the corresponding quantities for the Narain-Horvitz-Thompson version and Hajek version of the IPW estimator that we refer to as naive IPW (NIPW) and IPW respectively. (Often NIPW is referred to as IPW in the literature.) We write down here the explicit form of NIPW and IPW:

$$NIPW_{[a,b]} = \sum_{i=1}^{n} \widehat{\omega}^{\text{IPW}}_{[a,b],i}\, Y_{3,i}\,/n \quad \text{and} \quad IPW_{[a,b]} = NIPW_{[a,b]} \Bigg/ \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\omega}^{\text{IPW}}_{[a,b],i}\right)$$

where $\widehat{\omega}^{\text{IPW}}_{[a,b],i}$ is estimator of the weight from Lemma 1 for the $i$-th observation and is obtained using the components of the estimator $\widehat{p}(Z_i)$ of the conditional hazards.

Let us now present a comparative discussion of the performance of the EFF, NIPW and IPW estimators based on the simulation results in Table 3. First, we note that Bias is small for all the estimators, but it is the smallest for EFF. The small bias should not be surprising here because our probit specification for the conditional hazard nests the truth. Hence, EFF, IPW and NIPW are all known to be both consistent and asymptotically unbiased.

As for the other characteristics, note that AS Std is the feasible measure of variability of the estimators while MC Std is infeasible but a more accurate measure of variability that can serve as the benchmark measure. Accordingly, AS Size is the empirical size of the feasible test while MC Size is that of the infeasible test based on the infeasible MC Std.

Consider MC Std. EFF has the smallest MC Std, and often by a large margin. MC Std of NIPW is the worst. IPW improves upon it but still is much larger than that of EFF.

Now, consider AS Std. This is based on a feasible measure of variability since AS Std is the average over the trials of the square root of the estimated asymptotic variance. To compute the asymptotic variance of the estimators in each trial, we always take into account that the nuisance parameters are unknown and estimated; see Supplemental Appendix B.1.2.[9]

[9] NIPW and IPW have the same asymptotic variance under assumption CH, which holds here, since $\sum_{j=1}^{n} \widehat{\omega}^{\text{IPW}}_{[a,b],j}/n$ in the denominator of IPW converges in probability to one. However, $\sum_{j=1}^{n} \widehat{\omega}^{\text{IPW}}_{[a,b],j}/n$ is often very different from 1 in finite samples. Hence we do not impose this restriction when computing the As Std for IPW. Specifically, in the delta-method computation involved here, we also take into account the estimation of this denominator of IPW. This is why the AS Stds for IPW are smaller than that of NIPW.

| | $n = 100$ | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target | Bias | | | AS Std | | | MC Std | | | AS Size | | | MC Size | | |
| $[a, b]$ for $\beta$ | EFF | NIPW | IPW | EFF | NIPW | IPW | EFF | NIPW | IPW | EFF | NIPW | IPW | EFF | NIPW | IPW |
| $[1, 1]$ | -.004 | -.059 | -.053 | .334 | .372 | .341 | .409 | .614 | .443 | 11.8 | 23.1 | 14.6 | 5.1 | 3.2 | 5.3 |
| $[2, 2]$ | -.007 | -.058 | -.032 | .399 | .368 | .367 | .450 | .561 | .475 | 8.2 | 17.9 | 13.4 | 5.3 | 4.1 | 5.2 |
| $[1, 2]$ | -.006 | -.059 | -.042 | .289 | .323 | .317 | .373 | .519 | .406 | 13.4 | 21.6 | 14.0 | 5.1 | 3.5 | 5.2 |
| $[2, 3]$ | -.003 | -.030 | -.018 | .280 | .261 | .267 | .319 | .368 | .338 | 8.5 | 15.1 | 12.6 | 5.1 | 4.2 | 5.0 |
| $[1, 3]$ | -.004 | -.045 | -.033 | .251 | .274 | .276 | .325 | .428 | .356 | 13.8 | 19.4 | 13.8 | 5.3 | 3.7 | 5.1 |
| | $n = 200$ | | | | | | | | | | | | | | |
| Target | Bias | | | AS Std | | | MC Std | | | AS Size | | | MC Size | | |
| $[a, b]$ for $\beta$ | EFF | NIPW | IPW | EFF | NIPW | IPW | EFF | NIPW | IPW | EFF | NIPW | IPW | EFF | NIPW | IPW |
| $[1, 1]$ | -.001 | -.021 | -.021 | .244 | .250 | .251 | .255 | .370 | .298 | 6.4 | 15.4 | 10.1 | 5.0 | 4.1 | 5.2 |
| $[2, 2]$ | .001 | -.014 | -.011 | .282 | .274 | .267 | .283 | .384 | .313 | 4.9 | 10.9 | 8.7 | 4.9 | 3.4 | 4.9 |
| $[1, 2]$ | -.001 | -.019 | -.017 | .214 | .226 | .228 | .232 | .322 | .268 | 7.5 | 14.1 | 9.6 | 4.9 | 3.8 | 4.9 |
| $[2, 3]$ | .000 | -.007 | -.006 | .205 | .198 | .197 | .210 | .253 | .227 | 5.7 | 9.5 | 8.6 | 5.0 | 3.6 | 4.7 |
| $[1, 3]$ | -.001 | -.014 | -.012 | .188 | .200 | .201 | .208 | .270 | .236 | 8.1 | 12.5 | 9.6 | 5.1 | 4.0 | 4.9 |
| | $n = 500$ | | | | | | | | | | | | | | |
| Target | Bias | | | AS Std | | | MC Std | | | AS Size | | | MC Size | | |
| $[a, b]$ for $\beta$ | EFF | NIPW | IPW | EFF | NIPW | IPW | EFF | NIPW | IPW | EFF | NIPW | IPW | EFF | NIPW | IPW |
| $[1, 1]$ | -.002 | -.006 | -.008 | .155 | .165 | .165 | .154 | .208 | .181 | 5.3 | 11.1 | 7.6 | 5.0 | 4.5 | 5.1 |
| $[2, 2]$ | -.002 | -.004 | -.004 | .175 | .171 | .171 | .167 | .202 | .186 | 4.1 | 8.3 | 6.8 | 5.1 | 4.7 | 5.3 |
| $[1, 2]$ | -.002 | -.006 | -.006 | .136 | .146 | .146 | .139 | .178 | .160 | 6.0 | 10.1 | 7.2 | 5.2 | 4.7 | 5.1 |
| $[2, 3]$ | -.001 | -.003 | -.003 | .130 | .128 | .128 | .129 | .142 | .137 | 4.8 | 7.1 | 6.5 | 4.8 | 4.8 | 5.1 |
| $[1, 3]$ | -.002 | -.004 | -.005 | .121 | .131 | .131 | .127 | .153 | .142 | 6.7 | 8.9 | 7.0 | 5.0 | 4.6 | 4.9 |

Table 3: Results for the Efficient (EFF), NIPW and IPW estimators are reported based on 10,000 Monte Carlo trials. Bias stands for the average of the difference between, respectively, the EFF, NIPW and IPW estimators and the corresponding true value $\beta^0_{[a,b]}$. AS Std and MC Std stand for the standard deviation based on the asymptotic variance formula and Monte Carlo respectively. AS Size and MC Size stand for the empirical size of the asymptotic 5% two-sided test of $H_0 : \beta_{[a,b]} = \beta^0_{[a,b]}$ based on t-ratios using the asymptotic standard deviation and MC Std respectively.

AS Std turns out to be a very poor measure of the true variability of NIPW and IPW for which it is generally much smaller than the benchmark measure, MC Std. Therefore, AS Std, which is based on the estimated asymptotic variance, paints a misleadingly optimistic picture of the precision of NIPW and IPW. This is unfortunate because, in practice, one can only estimate the asymptotic variance but not the Monte Carlo variance with a given data.

By contrast, the AS Std and MC Std of EFF are closer and they are essentially the same when $n = 200, 500$. Therefore, fortunately, AS Std turns out to be a reliable measure for the true variability of EFF. As a consequence, the empirical size of the 5% feasible test, i.e., AS Size, is much closer to the 5% level for EFF than it is for IPW and even more so when compared to the AS Size of NIPW.[10]

In summary, NIPW performs the worst in terms of all characteristics. IPW improves over NIPW in all aspects confirming the findings of, e.g., Hirano and Imbens (2001), Lunceford and Davidian (2004), Busso et al. (2014), etc. (Hence we do not consider NIPW elsewhere in our paper.) However, the true variability of both NIPW and IPW is underestimated by their corresponding AS Std. As a result (recall that Bias is negligible here), the AS Size of NIPW and IPW shows that feasible inference based on them can lead to bad over-rejection of the truth even in quite large samples ($n = 500$). On the other hand, EFF seems to be largely immune to these problems and leads to smaller bias and variability and better inference.

# D Supplemental Appendix D: Empirical Illustration

For completeness, we wish to provide here a short illustration of the benefits of the efficiency gains due to the proposed efficient estimator (EFF) over the more conventional IPW estimator in drawing substantive conclusion on the effect of small class size in the Project STAR. (As in the Monte Carlo, NIPW performs worse than IPW and so we only include the latter.)

[10] This relatively worse performance of AS Size of NIPW and IPW is related to the observation that AS Stds for NIPW and IPW are much smaller than their MC Stds. To see this, note from Table 3 that MC Size for EFF, NIPW and IPW are all quite close to the 5% level, which, in turn, indicates that any problem related to AS Size is due to how accurately the true variability of these estimators are measured by their respective asymptotic variance formula. NIPW and IPW suffer from this inaccuracy, but EFF does not.

## D.1 Background

Tennessee's Student/Teacher Achievement Ratio experiment, also known as Project STAR, has been extensively used to study the effect of small class size on future outcomes for the students; see, e.g., Hanushek (1999), Krueger (1999), Krueger and Whitmore (2001), Ding and Lehrer (2010), Chetty et al. (2011), etc. In Project STAR, students enrolling in grade K of 79 participating schools in the 1985-1986 school year were randomly assigned to three types of classes: small classes (13-17 students per teacher), regular classes (22-25 students per teacher), and regular classes with a full-time teacher's aide (22-25 students per teacher). The literature on Project STAR typically does not differentiate between the latter two class types, and we will follow that here and refer to them jointly as "not-small" classes.

We use the well-known and publicly available Project STAR data (Achilles et al. (2008)) containing characteristics of the schools, the teachers, demographic and socioeconomic characteristics of the students and their standardized math and reading scores from grade K to grade 3 or to a lower grade until which they stayed with a Project STAR school.

Many students — 917 out of 1900 (48.3%) from small classes and 2139 out of 4425 (48.3%) from not-small classes — entering Project STAR schools in grade K did not stay until the project ended, i.e., until the end of grade 3. (All details are available from the authors.)

We study this attrition to obtain a more complete picture of the effect of small classes on students' scores during this four-year-long experiment, i.e., in grades K-3. To this end, it is useful to first ask which effects an "ideal" Project STAR experiment would have generated with the subjects/students entering grade K in 1985 if there was no subsequent attrition or other implementation-related compromises; see, e.g., Hanushek (1999). The answer is that, since there was no protocol to randomly switch the class types of students after each grade, an "ideal" Project STAR experiment would have generated in grades K, 1, 2 and 3 the effect of continued presence in small classes with respect to continued presence in not-small classes.

We first define these effects from the "ideal" experiment. We view attrition as a mitigating action by students in response to the treatment (class type) that they perceived as unhelpful to them. To gain a better understanding of this mitigating action we then decompose these effects by the attrition behavior of students from small and not-small classes. For brevity, we present only the results for standardized reading scores. Results for standardized math scores and percentile reading, math and average scores (defined in Krueger (1999)) are similar.[11]

## D.2 Results

Let $Y^{\mathrm{s}}(\text{grade } j \text{ read})$ be the potential grade $j$ reading score of a student *had* (s)he stayed in the small class at least until the end of grade $j$ for $j = K, 1, 2, 3$ after being initially randomized to a small class in grade K. Similarly, with superscript "ns" denoting not-small, define the potential scores $Y^{\mathrm{ns}}(\text{grade } j \text{ read})$ for $j = K, 1, 2, 3$. These scores are not observed for a student in grade $j$ if the student left the participating school before grade $j = 1, 2, 3$.

As noted above, we focus on two treatment regimes — a continued presence in small classes and a continued presence in not-small classes over the four years of Project STAR. Denote the average difference between the outcomes of these two regimes at each grade $j = K, 1, 2, 3$ as:

$$\mu_j^{\mathrm{read}} := E[Y^{\mathrm{s}}(\text{grade } j \text{ read}) - Y^{\mathrm{ns}}(\text{grade } j \text{ read})].$$

**Evolution of the effect of small classes:**

First, consider the trajectory of $\mu_j^{\mathrm{read}}$ for $j = K, 1, 2, 3$ to see how the effect of the small-class regime with respect to the not-small class regime evolved over continued presence in these regimes. Their EFF and IPW estimates are plotted in Figure 1(a).[12] The EFF and IPW

[11] We ignore the compromises other than attrition to the experiment, e.g., students who enrolled after grade K or the few students (2.1%–5.7% in the respective grades) who switched their assigned class types.

[12] We obtain these estimates following Section 3.3 using parametric models specified for the conditional hazards and conditional expectations. The conditional hazard of leaving small (respectively, not-small) classes after grade j (= K, 1, 2) is modeled as logit with a linear index of a constant, dummies for race, sex, types of school (inner city, urban and rural), the share of students on free-lunch in school, dummies for all grades (present and past) where the student was on free lunch, where the student's teacher had bachelor's degree, and the difference in each of the past grades between the student's standardized math and reading scores from, respectively, the average standardized math and average standardized reading scores in small classes and also in not-small classes in their school. The differences between the student's and the average scores are continuous variables, and we also include their quadratic and cubic terms in the index. The conditional expectations of the grade $j$ $(= 1, 2, 3)$ scores in small (respectively, not-small) classes are modeled linearly with exactly the same variables as above. All the estimation results are available from us.

estimates of the trajectory are quite similar. Consistent with the literature, we observe that the initial effect $\mu_K^{\text{read}}$ is very large compared to the "value added" (e.g., $\mu_j^{\text{read}} - \mu_K^{\text{read}}$ for $j = 1, 2, 3$) in the subsequent grades 1, 2 and 3. However, our value added estimates are not as pessimistic as, e.g., Hanushek (1999)'s that led him to question the justification of the huge cost of prolonged operation of small classes, but are more in line with Krueger (1999).

We conjecture that it is the correction for attrition that makes our estimates less pessimistic than Hanushek (1999)'s. To follow up on this, as proxies to Hanushek (1999)'s annual and 4-year samples respectively, we also plot in Figure 1(a) the "In grade" and "Never left" estimates of the trajectory. They are obtained by averaging the observed score of, respectively, the students who took the tests at the end of the respective grades and the students who continued in Project STAR until the end of grade 3. Note that, Never left actually estimates $\nu_{j,3}^{\text{read}}$ while In grade estimates $\nu_{j,j}^{\text{read}}$ for $j = K, 1, 2, 3$ where:

$$\nu_{j,l}^{\text{read}} := E[Y^{\text{s}}(\text{grade } j \text{ read})|B_l^{\text{s}}] - E[Y^{\text{ns}}(\text{grade } j \text{ read})|B_l^{\text{ns}}] \quad \text{for} \quad j, l = K, 1, 2, 3$$

and $B_j^{\text{s}}$ (respectively, $B_j^{\text{ns}}$) is the event that a student assigned to small (respectively, not-small) class in grade K does not leave Project STAR until the end of grade $j$ for $j = K, 1, 2, 3$. $\nu_{K,K}^{\text{read}} = \mu_K^{\text{read}}$ but in general $\nu_{j,l}^{\text{read}} \neq \mu_j^{\text{read}}$ for $j = 1, 2, 3$ and $l = K, 1, 2, 3$ unless suitable mean independence assumptions hold or, by happenstance, the biases for small and not-small classes cancel out, i.e., $E[Y^{\text{s}}(\text{grade } j \text{ read})|B_l^{\text{s}}] - E[Y^{\text{s}}(\text{grade } j \text{ read})] = E[Y^{\text{ns}}(\text{grade } j \text{ read})|B_l^{\text{ns}}] - E[Y^{\text{ns}}(\text{grade } j \text{ read})]$. Both In grade and Never left estimates reveal that without correction for attrition the value added estimates would indeed be lower.

### Does attrition matter?

But, beyond this visual inspection, does the correction for attrition matter statistically as well? More precisely, since we observed that the attrition-corrected estimates (EFF and IPW) are larger than the attrition-uncorrected estimates (In grade, which is typically favored to Never left), it is natural to ask if this is entirely due to sampling variation or is there systematic evidence for this in the population. That is, one would want to test the null hypothesis $H_{0,j} : \mu_j^{\text{read}} = \nu_{j,j}^{\text{read}}$ against the alternative $H_{1,j} : \mu_j^{\text{read}} > \nu_{j,j}^{\text{read}}$ for $j = 1, 2, 3$.

We find the p-values for the tests for $H_{0,1}$, $H_{0,2}$ and $H_{0,3}$ (corresponding to grades 1, 2 and 3) to be 28.5%, 14.0% and 13.9% respectively when the EFF estimates for $\mu_j^{\text{read}}$ are used. The p-values are 26.8%, 42.7% and 35.3% when the IPW estimates for $\mu_j^{\text{read}}$ are used.

The EFF p-values for $H_{0,2}$ and $H_{0,3}$ are small and not sufficient in practice to take for granted the reliability of the attrition-uncorrected In grade estimates for the true effect $\mu_2^{\text{read}}$ and $\mu_3^{\text{read}}$.[13] On the other hand, the IPW p-values are quite a bit larger for $H_{0,2}$ and $H_{0,3}$. It is however not prudent (and possibly misleading) to take $H_{0,2}$ and $H_{0,3}$ for granted because, as we will see below, the large IPW p-values are entirely due to the imprecise nature of the IPW estimates. By contrast, our proposed EFF estimates help to avoid the misleading confidence in $H_{0,2}$ and $H_{0,3}$ and point toward the possibility that attrition does matter here.

### Do attrition-corrected estimates give substantive conclusions on the effects?

Attrition-correction will be of limited use to practitioners if it does not lead to precisely estimated (zero or non-zero) effects. To explore if that is the case here, we plot in Figure 1(b) the 90%, 95% and 99% two-sided confidence intervals around the EFF and IPW estimates for $\mu_K^{\text{read}}, \mu_1^{\text{read}}, \mu_2^{\text{read}}$ and $\mu_3^{\text{read}}$. We observe that while the EFF and IPW estimates are similarly precise for $\mu_1^{\text{read}}$ (and identical by definition for $\mu_K^{\text{read}}$), the EFF estimates for $\mu_2^{\text{read}}$ and $\mu_3^{\text{read}}$ are much more precise than their IPW estimates that seem to be very uninformative.[14] These IPW estimates fail to reject a zero or negative value of $\mu_2^{\text{read}}$ or $\mu_3^{\text{read}}$ even at (much above) the 10% level. That is, IPW cannot rule out statistically that a continued presence for more than two years in small classes as opposed to not-small classes can, on average, *even hurt the students.* (This is not a case of a precisely estimated zero or negative effect which, on

[13] These are all Hausman-type tests. Typically one demands a much larger p-value (e.g., think of Hausman test for OLS versus IV regressions) to be able to have confidence in the null hypothesis for Hausman tests.

[14] The case of $\mu_1^{\text{read}}$ corresponds to $R = 2$ (since the outcome of reading score in grade 1 happens in period 2) and $a = 1, b = 2$, and hence the rich parametric specification of the conditional hazard (see footnote 12) suggests that IPW estimates could also be efficient; see, e.g., Hirano et al. (2003), Graham (2011), etc.

the other hand, would have been informative.) The practical implication of not being able to rule out such negative evidence against small classes is serious because operationalizing small classes is an expensive policy proposition to begin with; see, e.g., Hanushek (1999).

By contrast, no such problem arises with EFF. EFF tries to optimally use the information available in the selection on observables condition. This is important since attrition over multiple periods naturally affects the later periods worse and thereby leaves little room for wasting information like IPW. Consequently, in spite of using similar assumptions as IPW (see footnote 1), the EFF intervals are subsets of the IPW intervals, and a zero or negative effect of small classes is easily rejected by EFF even at the 1% level at every grade K–3.

**Attrition as a mitigating action against unhelpful class type assignment:**

Students were randomly assigned to small and not-small classes when they enrolled in a Project STAR school in grade K. Many students did not score well in their randomly assigned class type. Leaving the Project STAR school was an important course of mitigating action available to these students. If attrition in Project STAR was primarily due to this mitigating action then, given the initial random assignment, we would expect to see that students who stayed scored better than what students who left would have scored had they stayed instead.

This is exactly what we observe in our estimates for each grade 1, 2 and 3. For brevity, we report here only the results for grade 3 since it is the terminal period of the experiment, and compare those who never left with each of the other attrition categories. Table 4 reports the EFF and IPW estimates of $\alpha_3^{\text{s,read}} - \alpha_j^{\text{s,read}}$ and $\alpha_3^{\text{ns,read}} - \alpha_j^{\text{ns,read}}$ for $j = K, 1, 2$ where:[15]

$$\alpha_j^{\text{t,read}} := E[Y^{\text{t}}(\text{grade 3 read}) \mid \text{left small class after grade } j] \quad \text{for } t = \text{s, ns, and } j = K, 1, 2, 3.$$

EFF and IPW estimates are very similar, but EFF is much more precise than IPW. Consequently, EFF always confirms at the conventional levels of significance the intuition

[15] $\alpha_K^{\text{s,read}} = E[Y^{\text{s}}(\text{grade 3 read})|B_K^{\text{s}}, \bar{B}_1^{\text{s}}]$, $\alpha_j^{\text{s,read}} = E[Y^{\text{s}}(\text{grade 3 read})|B_j^{\text{s}}, \bar{B}_{j+1}^{\text{s}}]$ for $j = 1, 2$, and $\alpha_3^{\text{s,read}} = E[Y^{\text{s}}(\text{grade 3 read})|B_3^{\text{s}}]$ in terms of the events $B_l^{\text{s}}$ for $l = K, 1, 2, 3$ defined earlier and where $\bar{B}_l^{\text{s}}$ denotes the complement of the event $B_l^{\text{s}}$ for $l = 1, 2, 3$. Results for all contrasts of the $\alpha$'s are available from the authors.

that students who stayed scored better on average than what students who left would have scored had they stayed instead. By contrast, IPW sometimes fails to confirm this standard intuition behind the choice to leave. These failures happen not because the IPW estimates are different from the EFF estimates (which they are not) but because they are very imprecise.

|  | $\alpha_3^{\text{s,read}} - \alpha_j^{\text{s,read}}$ |  | $\alpha_3^{\text{ns,read}} - \alpha_j^{\text{ns,read}}$ |  |
|---|---|---|---|---|
| $j$ | EFF | IPW | EFF | IPW |
| K | $0.39^{**}$ | 0.34 | $0.48^{***}$ | $0.48^{*}$ |
|  | (0.19) | (0.46) | (0.05) | (0.33) |
| 1 | $0.45^{**}$ | 0.48 | $0.64^{***}$ | $0.63^{**}$ |
|  | (0.25) | (0.46) | (0.11) | (0.33) |
| 2 | $0.51^{***}$ | $0.47^{***}$ | $0.46^{***}$ | $0.46^{*}$ |
|  | (0.12) | (0.19) | (0.08) | (0.32) |

Table 4: EFF and IPW estimates and standard errors (in parentheses) for $\alpha_3^{\text{t,read}} - \alpha_j^{\text{t,read}}$ for $t = s, ns$ and $j = K, 1, 2$. $^{*}$, $^{**}$ and $^{***}$ signify if the null that the parameter is zero is rejected against the alternative that it is greater than zero at the 10%, 5% and 1% level respectively.

The parameters reported in Table 4 are related to the following decompositions of the effect $\mu_3^{\text{read}}$ of small classes by attrition categories (noted in Supplemental Appendix D.1):

$$\mu_3^{\text{read}} = \sum_{j=K,1,2,3} \mu_{3,j,*}^{\text{read}} \times P\left(\text{left small class after grade } j\right) = \sum_{j=K,1,2,3} \mu_{3,*,j}^{\text{read}} \times P\left(\text{left not-small class after grade } j\right)$$

based on the attrition from small and not-small classes respectively, where for $j = K, 1, 2, 3$:

$$\mu_{3,j,*}^{\text{read}} = E[Y^{\text{s}}(\text{grade } j \text{ read})| \text{ left small class after grade } j] \; - \; E[Y^{\text{ns}}(\text{grade } j \text{ read})]$$
$$\mu_{3,*,j}^{\text{read}} = E[Y^{\text{s}}(\text{grade 3 read})] \; - \; E[Y^{\text{ns}}(\text{grade 3 read})| \text{ left not-small class after grade } j];$$

and note that $\alpha_3^{\text{s,read}} - \alpha_j^{\text{s,read}} = \mu_{3,3,*}^{\text{read}} \; - \; \mu_{3,j,*}^{\text{read}}$ while $\alpha_3^{\text{ns,read}} - \alpha_j^{\text{ns,read}} = -(\mu_{3,*,3}^{\text{read}} \; - \; \mu_{3,*,j}^{\text{read}})$.

EFF and IPW estimates of these two decompositions, along with the 90%, 95% and 99% two-sided confidence intervals, are reported in Figures 1 (c)-(d) showing the relative contribution of each attrition category from small and not-small classes respectively toward the overall effect. Given the large number of students who left, it is important to understand what the effect would have been with respect to students leaving at various junctures of the experiment. $\mu_{3,*,j}^{\text{read}}$ and $\mu_{3,j,*}^{\text{read}}$ for $j = K, 1, 2, 3$ are those effects on grade 3 reading scores.

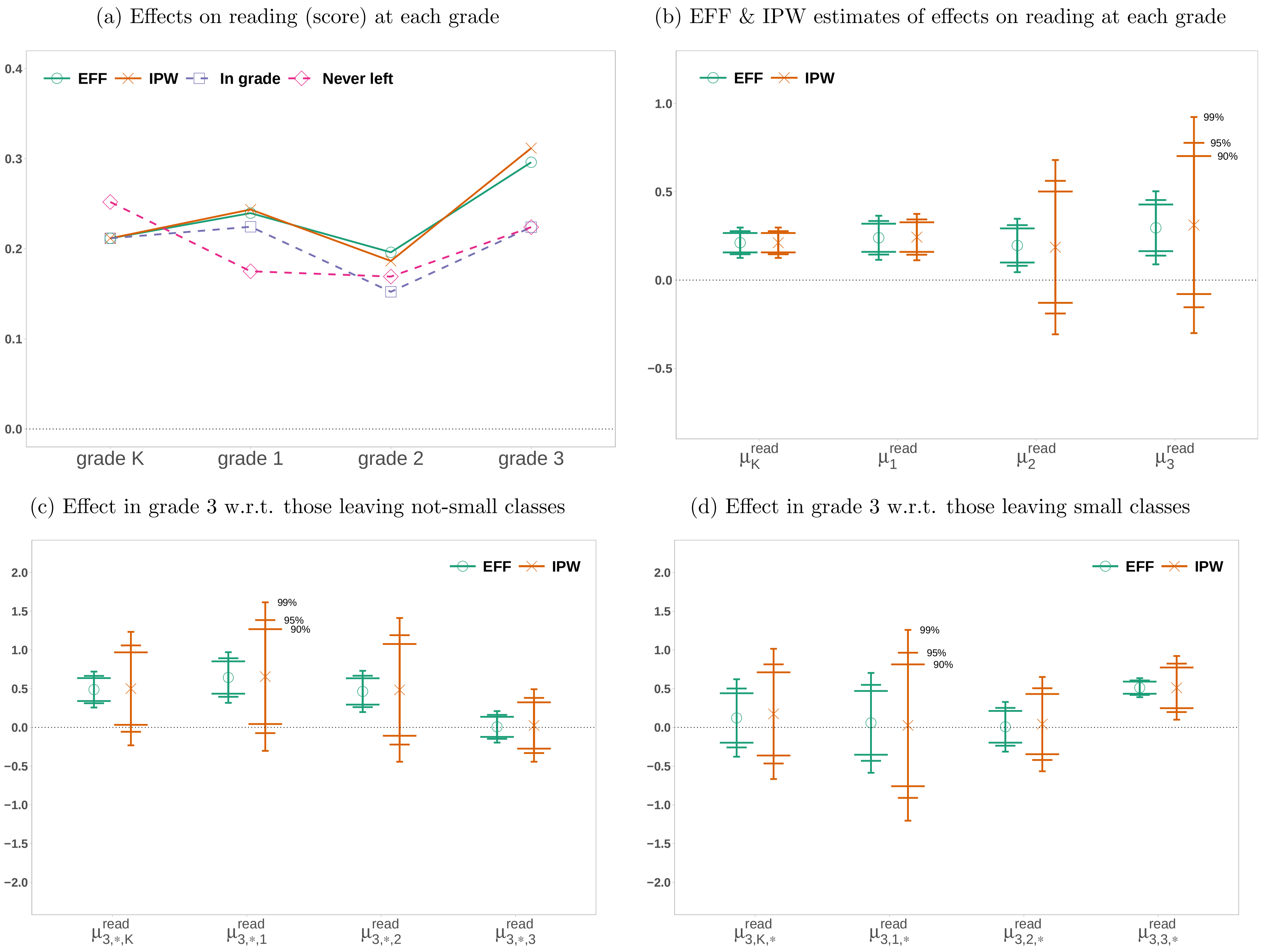
These decompositions reveal interesting patterns telling us which group of students (by their attrition category) are driving the overall effect of small classes in the terminal period grade 3, and by how much. Unfortunately, the imprecision of the IPW estimates prevents them from confirming these patterns statistically at the conventional levels of significance. On the other hand, while the EFF estimates of the decompositions are very similar to the IPW estimates, once again the precision of EFF helps to confirm these interesting patterns statistically and underscores the importance of efficient estimation in this analysis.[16]

Lastly, note that these EFF-based inferences are precise because the sub-population-specific components of the effects are estimated much more precisely by EFF. Table 5 reports the results for EFF and IPW estimation of a subset of such components. These parameters are similar to the estimands in the Monte Carlo experiment in Supplemental Appendix C.

| Left STAR school | Randomized to small class | | Randomized to not-small class | |
|---|---|---|---|---|
| at the end of grade | EFF | IPW | EFF | IPW |
| (a) K | 0.05 | 0.10 | -0.27 | -0.26 |
| | (0.19) | (0.29) | (0.05) | (0.22) |
| (b) 1 | -0.01 | -0.04 | -0.42 | -0.42 |
| | (0.25) | (0.45) | (0.10) | (0.33) |
| (c) 2 | -0.07 | -0.03 | -0.24 | -0.24 |
| | (0.12) | (0.18) | (0.07) | (0.31) |
| (d) 3 (never left) | 0.44 | 0.44 | 0.22 | 0.22 |
| | (0.04) | (0.04) | (0.03) | (0.03) |

Table 5: EFF and IPW estimates of expected (counterfactual) reading scores in grade 3 by the student's attrition period are presented under the class types to which they were initially randomized. The estimates are expressed as standardized deviation from the overall mean reading score of students in grade 3. Standard deviations are presented in parentheses. All the results in this empirical illustration are based on such parameters and the standard errors of those results were computed by noting that the estimates in Table 5 across the two class types are independent but are correlated within class types. Results for math, percentile, etc. scores are similar and the complete set of results is available from the authors.

[16] A definitive study of these patterns is beyond the scope of the present empirical illustration. However, thanks to the relationship between these decomposition-parameters and the parameters in Table 4, the results on the EFF estimates in Table 4 reject the null hypotheses $\mu^{\text{read}}_{3,*,3} = \mu^{\text{read}}_{3,*,j}$ and $\mu^{\text{read}}_{3,3,*} = \mu^{\text{read}}_{3,j,*}$ against the alternative hypotheses $\mu^{\text{read}}_{3,*,3} < \mu^{\text{read}}_{3,*,j}$ and $\mu^{\text{read}}_{3,3,*} > \mu^{\text{read}}_{3,j,*}$ respectively for $j = K, 1, 2$. As we noted in Table 4, by virtue of its precision, EFF confirms this evidence more strongly statistically. Not reported in Table 4 is that EFF also rejects the null hypotheses $\mu^{\text{read}}_{3,*,2} = \mu^{\text{read}}_{3,*,1}$ against the alternative $\mu^{\text{read}}_{3,*,2} < \mu^{\text{read}}_{3,*,1}$ at the 5% level, while, due to its lack of precision, IPW does so only at the 26% level.


(a) Effects on reading (score) at each grade

(b) EFF & IPW estimates of effects on reading at each grade

(c) Effect in grade 3 w.r.t. those leaving not-small classes

(d) Effect in grade 3 w.r.t. those leaving small classes

Figure 1: **(a)** EFF, IPW, In grade and Never left estimates of effect on reading score at each grade. **(b)** EFF and IPW estimates and confidence intervals (90%, 95%, 99%) of $\mu_K^{\text{read}}, \mu_1^{\text{read}}, \mu_2^{\text{read}}, \mu_3^{\text{read}}$. 90%, 95%, 99% EFF and IPW confidence intervals for the decomposition of $\mu_3^{\text{read}}$ by comparing: **(c)** the entirety of small classes with different attrition categories from not-small classes, and **(d)** the different attrition categories from small classes with the entirety of not-small classes.

As we have noted throughout, attrition in multi-period studies often entails substantial loss in data and, consequently, it is imperative that all the information contained in the identifying assumptions (in our paper, the MAR assumption in (1)) be utilized as efficiently as possible to estimate the parameters of interest precisely. Our theoretical results were about such estimation, and the current section illustrated its benefit based on the widely used Project STAR data. Given that the underlying computations for efficient estimation are quite standard, we hope that the potential benefit of being able to draw substantive conclusions encourages practitioners to consider efficient estimation under similar contexts.

# E Bibliography

Abowd, J. M., Crepon, B., and Kramarz, F. (2001). Moment Estimation with Attrition: An Application to Economic Models. *Journal of the American Statistical Association*, 96:1223–1231.

Abrevaya, J. (2019). Missing dependent variables in fixed-effects models. *Journal of Econometrics*, 211:151–165.

Achilles, C., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., and Word, E. (2008). Tennessee's Student Teacher Achievement Ratio (STAR) project.

Ackerberg, D., Chen, X., and Hahn, J. (2012). A Practical Asymptotic Variance Estimator For Two-Step Semiparametric Estimators. *The Review of Economics and Statistics*, 94:481–498.

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. (2014). New Evidence on the finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *Review of Economics and Statistics*, 96:885–897.

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

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.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126:1593–1660.

Ding, W. and Lehrer, S. F. (2010). Estimating treatment effects from contaminated multiperiod education experiments: the dynamic impacts of class size reductions. *The 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.

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

Hajek, J. (1971). Comment on a paper by d. basu. In Godambe, V. R. and Sprott, D. A., editors, *Foundations of Statistical Inference*, page 236. Holt, Rinehert and Winston, Toronto.

Hanushek, E. A. (1999). Some Findings from an Independent Investigation of the Tennessee

STAR Experiment and from Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis*, 21:143–63.

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.

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

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics*, 114:497–532.

Krueger, A. B. and Whitmore, D. M. (2001). The effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR. *The Economic Journal*, 111:1–28.

Lunceford, J. and Davidian, M. (2004). Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects : A Comparative Study. *Statistics in Medicine*, 23:2937–2960.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of Indian Soc. Agricultural Statistics*, 3:169–174.

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.

Rothe, C. and Firpo, S. (2019). Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically. *Econometric Theory*, 35: 1048–1087.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wooldridge, J. M. (2002). Inverse Probability Weighted M-Estimation for Sample Selection, Attrition, and Stratification. *Portuguese Economic Journal*, 1:117–139.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section & Panel Data*. MIT Press.