Econometric Theory, **36**, 2020, 488–525.
doi:10.1017/S0266466619000239

ON EFFICIENCY GAINS FROM MULTIPLE INCOMPLETE SUBSAMPLES

SARASWATA CHAUDHURI
McGill University

Cost-effective survey methods such as multi(R)-phase sampling typically generate samples that are collections of monotonic subsamples, i.e., the variables observed for the units in subsample r are also observed for the units in subsample $r+1$ for $r = 1, \ldots, R-1$. These subsamples represent subpopulations that can be systematically different if the selection of a unit in each phase of sampling depends on the observed variables for that unit from past phases. Our article is about optimally combining all the subsamples for the efficient estimation of a finite dimensional parameter defined by moment restrictions on a generic target population that is an arbitrary union of these subpopulations. Only the R-th subsample is assumed to contain all the variables that are arguments of the moment function. Semiparametric efficiency bounds for estimation are obtained under a unified framework, allowing for full generality of the selection on observables in the sampling design. Contribution of each subsample toward efficient estimation is analyzed; and this turns out to differ fundamentally from that in setups where the same collection of subsamples is instead generated unplanned by unknown sampling. Uniquely, our setup enables all the subsamples to contribute to the efficient estimation for all the target populations, which we show is not possible in other setups. Efficient estimation is standard. Simulation evidence of substantive efficiency gains from using all the subsamples is provided for all the targets.

1. INTRODUCTION

Planned incompleteness in the data can be useful when conducting surveys under budget constraints. The basic idea behind planned incompleteness is that: when it is expensive to collect all the variables for all the units in the sample, the next best

I am very much grateful to the editor P.C.B. Phillips, the co-editor P. Guggenberger and three anonymous referees for their detailed insightful comments. The article was circulated before as "A Note on Efficiency Gains from Multiple Incomplete Subsamples" but the title was modified at the suggestion of the editor. Previous versions of the article, some of which are available on the author's webpage, benefitted from the helpful comments of A. Prokhorov, C. Muris, D. Guilkey, D. Frazier, E. Renault, F. Lange, J. Hill, J. Haushofer, J. MacKinnon, J. Wooldridge, M. Carrasco, M. Chemin, P. Saha Chaudhuri, S.J. Lee, and V. Zinde-Walsh, the seminar participants at Brown, Concordia, McGill (Econ and Biostat), Queen's, U. Canterbury, U. Montreal, U. New South Wales, UNC Chapel Hill, U. Sydney, West Virginia University and the Midwest Econometrics Group meetings (2013). Address correspondence to Saraswata Chaudhuri, Department of Economics, McGill University, Montreal, Canada; e-mail: saraswata.chaudhuri@mcgill.ca.

 © Cambridge University Press 2019

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

alternative could be to collect the less expensive variables for all the units in the sample but the more expensive variables only for a subset of these units.

A variable may be more expensive to collect for numerous reasons; e.g., a correct measurement may be expensive, it may require intensive follow-ups, and it may require tracking of or offering incentives to respondents. In all such cases, planned incompleteness cuts the cost of surveys by generating a sample in which only a subset of the units contains all the intended variables, while the rest contains various collections of only the less expensive variables. This happens by plan, i.e., sampling design, and, thus, the targeted use of survey resources eliminates or at least reduces unplanned nonresponse or mismeasurement that could have otherwise complicated subsequent analyses of the data.[1]

The idea of planned incompleteness is not new, and is more frequently employed in fields of research where the use of primary data is more prevalent than what has been typical in economics. The multiple-matrix sampling of Shoemaker (1973), the split-questionnaire design of Raghunathan and Grizzle (1995), the partial questionnaire design of Wacholder, Carroll, Pee, and Gail (1994), the many-measurement-design, and the multiforms surveys discussed by Graham, Hofer, and MacKinnon (1996) and Graham, Taylor, Olchowski, and Cumsille (2006) are standard examples of planned incompleteness designs that are widely used in biostatistics, epidemiology, psychology, and statistics to reduce the burden on the respondents and the cost of surveys.

Survey-cost is of concern to economists too as the collection of primary data under tight budget gets more common. Consequently, similar sampling designs have also been used, perhaps on an ad hoc basis, in the laboratory and field experiments in economics, where one often needs to contract the sampling frame in each phase of the experiment depending on the budget left and the survey time line; see, e.g., Holt and Laury (2002), Thornton (2008), Ashraf, Berry, and Shapiro (2010), Ashraf, Field, and Lee (2014), Beaman, Karlan, Thusbaert, and Udry (2015), and Beegle, Weerdt, Friedman, and Gibson (2012), etc. (Appendix A.1 discusses these references and others, and sketches an example to show how the analyses therein may benefit from the results in our article.)

However, planned incompleteness involves loss of information in the data since not all the variables are observed for all the sample units. Hence, it is imperative that any estimator using such data be as precise as possible. Unfortunately, efficient estimation using such data has rarely been considered.[2]

Our article seeks to address this issue of efficient estimation using planned incomplete data by taking the sampling design as given. We focus on monotonic multiphase samplings and, for full flexibility in the design of the multiple

[1] See Carroll, Ruppert, and Stefanski (1995), Little and Rubin (2002), etc. for methods of dealing with mismeasured or missing data.

[2] An exception is Chatterjee and Li (2010) who consider efficiency under the specific type of planned incompleteness design known as the partial questionnaire designs of Wacholder et al. (1994) [also see Chaudhuri and Guilkey (2016)].

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

phases, we maintain a general selection on observables, i.e., the missing at random (MAR), assumption. Close attention is paid to special cases of MAR.

To fix ideas, consider the prototypical multi(R)-phase sampling that, along with its variations, falls under the premise of our article. Suppose that a researcher intends to collect R sets of variables $Z_{(1)}, \ldots, Z_{(R)}$. In phase one, she collects $Z_{(1)}$ for a random selection of units. Then, recursively, at each phase $r = 2, \ldots, R$, she collects $Z_{(r)}$ for a subset of the units from phase $r-1$, selecting the subset randomly with or without regard to the already available information on $Z_{(1)}, \ldots, Z_{(r-1)}$.[3] The resulting sample consists of R groups of units such that the r-th group contains only $Z_{(1)}, \ldots, Z_{(r)}$ (these are the units followed until phase r but dropped after that) where $r = 1, \ldots, R$. We refer to these R groups as R subsamples. Only the R-th subsample is complete in the sense that it contains all the intended variables $Z_{(1)}, \ldots, Z_{(R)}$. These subsamples are monotonic, i.e., variables observed for the units in subsample r are also observed for the units in subsample $r+1$ for $r = 1, \ldots, R-1$.

Note that the underlying populations for the subsamples, call them subpopulations, can differ systematically in terms of the joint distribution of the intended variables $Z_{(1)}, \ldots, Z_{(R)}$ if the selection of the units for *any* phase takes into account any information that is already available by then.

In this article, we consider efficient estimation of a finite dimensional parameter defined by generic moment restrictions on the joint distribution of $Z_{(1)}, \ldots, Z_{(R)}$ in a generic target population that is an arbitrary union of these subpopulations. Special cases of the target include the subpopulations and the full population (union of all subpopulations). We provide a unified presentation of all cases.

The main contribution of our article is twofold. First, we obtain the efficient influence function and efficiency bound for the generic parameter of interest and analyze them closely to make explicit the role of each subsample toward efficient estimation. Second, we express the problem of efficient estimation in an alternative, equivalent way in the spirit of Brown and Newey (1998) and Graham (2011) rendering the actual efficient estimation a simple special case of Chamberlain (1992), Ai and Chen (2012), etc. Both contributions are driven by the two key features of our article—planned incompleteness and monotonicity—and, throughout, we emphasize the novelty of their implications with respect to the literature, e.g., Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Hahn (1998), Chen, Hong, and Tarozzi (2008), Graham (2011), and Barnwell and Chaudhuri (2018).[4]

[3] The two-phase sampling is a special case where $Z_{(3)}, \ldots, Z_{(R)}$ would also be collected with $Z_{(2)}$ in phase two, and the survey would end there. In turn, the variable probability (VP) sampling studied in Wooldridge (1999), Wooldridge (2007), etc. is a special type of two-phase sampling that would discard all the units for whom only $Z_{(1)}$ was collected. Thus, VP sampling unnecessarily losses information that has already been collected (paid for). The loss is naturally more severe if such a strategy is extended to multiple phases. Hence, we do not consider VP sampling in this article.

[4] It must also be noted here that our presentation in the sequel is incomplete in the following sense. While the idea of planned incompleteness to reduce survey cost and the unintended consequences of unplanned incompleteness is intuitively appealing, our article is silent about the optimality of such survey designs and instead takes a generic

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

Our article proceeds as follows. Section 2 describes our theoretical framework and presents our main theoretical results under MAR. Section 3 takes a closer look at two important special cases of MAR that could be more relevant in practice, and provides auxiliary results along with an analytical demonstration of efficiency gained from the optimal use of all the subsamples for efficient estimation. Section 4 describes efficient estimation and presents the consistency and asymptotic normality of the efficient estimator under high level assumptions. Section 5 demonstrates the efficiency gains (discussed in Section 3) in finite samples using a Monte Carlo experiment. Section 6 concludes. Abridged versions of the proofs of the results presented in our article are collected in an appendix.

There is also a Supplemental Appendix containing three sections: Appendix A, Appendix B, and Appendix C. This is available at Cambridge Journals Online (journals.cambridge.org/ect). Appendix A elaborates on the statements from the main text when they require longer explanations (e.g., Appendices A.1 and A.2 cited above). Appendix B fills in the details in the proofs of the results for Sections 2 and 3. Appendix C expands on the material of Section 4 by: (i) relating our estimation framework to that of Ai and Chen (2012), (ii) filling in the details in the proofs of the results for Section 4, and (iii) describing (with proof) a computationally convenient one-step updating of any $\sqrt{n}$-consistent estimator that leads to asymptotic efficiency. Appendix C also presents Monte Carlo evidence of the good finite-sample properties of the efficient estimator under the setup of Section 5.

Last, we note that recent contributions to data combination in economics include, e.g., Ichimura and Martinez-Sanchis (2005), Ridder and Moffitt (2007), Devereux and Tripathi (2009), Tripathi (2009), Dardanoni, Modica, and Peracchi (2011), Muris (2016), Graham, Pinto, and Egel (2016), Abrevaya and Donald (2017), and the many references therein. It is our focus on: (i) planned incompleteness, (ii) the allowance for a dynamically updating MAR condition, and (iii) the allowance for the parameter of interest to be defined in terms of arbitrary unions of subpopulations, that distinguishes our article from the rest.

2. FRAMEWORK AND THE COMBINATION OF SUBSAMPLES

2.1. Framework

Let $Z := (Z'_{(1)}, \ldots, Z'_{(R)})'$ where $Z_{(r)}$ is a $d_r \times 1$ random vector for $r = 1, \ldots, R$, and $\sum_{r=1}^{R} d_r$ is finite. Following Tsiatis (2006), consider a scalar variable C with support $\mathcal{C} := \{1, \ldots, R\}$ and a transformation $T_C(Z)$ defined as $T_r(Z) := (Z'_{(1)}, \ldots, Z'_{(r)})'$ of dimension $(\sum_{s=1}^{r} d_s) \times 1$ for $r = 1, \ldots, R$. The value of C determines $T_C(Z)$, i.e., how much of Z is observed for a sample unit.

design as given. Indeed, to our knowledge, a general optimality theory is yet to be developed for survey designs with planned incompleteness and this perhaps needs to be addressed on a case-by-case basis [see Reilly (1996)]. While a broader exploration of optimality is the topic of our ongoing research, at this point we only present in Appendix A.2 simple and illustrative examples of the optimality of a planned incomplete survey design in two-phase sampling.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

Let $O := (C, T_C'(Z))'$ denote what is observable for a unit. The observed sample is $\{O_i := (C_i', T_{C_i}'(Z_i))'\}_{i=1}^n$. The r-th subsample is the collection of units with only $T_r(Z)$ observed; it is of size $n_r := \sum_{i=1}^n I(C_i = r)$ for $r = 1, \ldots, R$. Only the R-th subsample is complete, i.e., $T_R(Z) = Z$.

A natural consequence of our description of the multiphase sampling is that it involves selection on observables. Note that, at the end of phase $r = 1, \ldots, R-1$, the researcher is left with the units $\{i = 1, \ldots, n : C_i \geq r\}$ and has observed $T_r(Z_i)$ for each of them. Now, the researcher decides the probability with which each such eligible unit continues to phase $r+1$. This probability can be equal, say $1 - p_r$, for all eligible units, in which case $P(C = r|C \geq r, T_r(Z)) = p_r$; or it can be more involved if it depends on $T_r(Z)$. Regardless, the researcher cannot possibly incorporate in this decision the knowledge of $Z_{(r+1)}, \ldots, Z_{(R)}$ since she does not observe them by the end of phase r. We formalize this statement by maintaining a general selection on observables, i.e., the MAR condition that:

$$P(C = r|C \geq r, T_R(Z)) = P(C = r|C \geq r, T_r(Z)), \text{ equivalently,}$$
$$P(C = r|T_R(Z)) = P(C = r|T_r(Z)) \quad \textbf{(1)}$$

for $r = 1, \ldots, R$. The equivalence in (1) follows from the invertible relation between hazard and probability mass functions [see Appendix A.3]. The second relation in (1) is the MAR condition in the sense of Rubin (1976) [see, e.g., Robins and Rotnitzky (1995) and Tsiatis (2006)] and generalizes to the case of $R > 2$ the MAR assumption found in econometrics where the focus has traditionally been on $R = 2$ [see, e.g., Chen, Hong, and Tamer (2005), Chen et al. (2008), Graham (2011), and Graham, Pinto, and Egel (2012)].

To signify that the incompleteness in the data is by plan/design, we maintain under (1) that:

$$P(C = r|C \geq r, T_r(Z)), \text{ equivalently, } P(C = r|T_r(Z)) \text{ is known for each } r = 1, \ldots, R. \quad \textbf{(2)}$$

The equivalence in (2) follows under (1) [see Appendix A.4]. The condition in (2) is a formality in this context since the researcher actually decides these probabilities as part of the sampling design.

Now, to define the parameter value of interest, consider a generic function $m(Z; \beta) : \text{Support}(Z) \times \mathcal{B} \mapsto \mathbb{R}^d$ of the parameter $\beta \in \mathcal{B} \subset \mathbb{R}^d$. (The case of over-identification is considered in older versions of this article, but it does not provide any new insights.) For a given target population $\lambda \in \Lambda$ where $\Lambda :=$ Power-Set($\mathcal{C}$) excluding the empty set, define the parameter value of interest β_λ^0 as:

$$E[m(Z; \beta)|C \in \lambda] = 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta_\lambda^0. \quad \textbf{(3)}$$

β_λ^0 is defined as a function of λ and may differ across targets $\lambda \in \Lambda$ if C and Z are dependent.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

For a given β, the function $m(Z;\beta)$ can be evaluated from the observed sample only for the n_R units in the complete subsample, i.e., $I(C=R)m(Z;\beta)$. However, point identification of β_λ^0 is still possible by the Horvitz–Thompson reweighting provided that $P(C=R|T_R(Z))>0$ almost surely. This is due to the following relation that holds identically in β [see Appendix A.5 for details]:

$$E\left[\frac{P(C\in\lambda|T_R(Z))}{P(C\in\lambda)}\frac{I(C=R)}{P(C=R|T_R(Z))}m(Z;\beta)\right]=E[m(Z;\beta)|C\in\lambda]. \quad \textbf{(4)}$$

All the terms inside the expectation on the left hand side (LHS) of (4) will be feasible under our assumptions because $P(C\in\lambda)$ will be trivially identified by the observed data under Assumption (A1) below, whereas (1) and (2) already imply that $P(C\in\lambda|T_R(Z))$ and $P(C=R|T_R(Z))$ are known. Hence, the LHS of (4) can serve as the estimating function for β_λ^0. However, such estimation will be based solely on the complete subsample. We will focus on exploring the information contained in the incomplete subsamples and demonstrating how that information can be combined with the information in the complete subsample for the purpose of efficient estimation of β_λ^0 defined in (3).

The discussion of our framework concludes by listing an assumption that we maintain hereafter.

Assumption A

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

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

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

Remark. (A1) is a standard assumption [see, e.g., Tsiatis (2006), Devereux and Tripathi (2009), Tripathi (2011), etc.]. $P(C=R|T_R(Z))>\underline{p}>0$ in (A2) is a strict version of the overlap assumption [see Khan and Tamer (2010)]. The restrictions $P(C=r|T_R(Z))>0$ for $r=1,\ldots,R-1$ are not strictly required but help to avoid more involved proofs peripheral to the main message. However, $P(C=r)>0$ for $r=1,\ldots,R$ is intrinsic to the R-level missing data model. (A3) allows for moment vectors $m(Z;\beta)$ that are not differentiable in β. We do, however, impose differentiability of $E[m(Z;\beta)|C\in\lambda]$ as in, e.g., Chen, Linton, and van Keilegom (2003), Chen et al. (2008), Cattaneo (2010), etc.

The theoretical framework above is closely related to several well-known articles such as Robins and Rotnitzky (1995), Whittemore (1997), Holcroft, Rotnitzky, and Robins (1997), Chen et al. (2005), Chen et al. (2008), Cattaneo (2010), Dardanoni et al. (2011), Lee, Scott, and Wild (2012), and Abrevaya and Donald (2017). In Appendix A.6, we discuss in detail where we actually differ from them. Broadly speaking, the differences are one or more of the following: (i) allowance

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

for a general R, (ii) expansion of the scope to all $(2^R - 1)$ subpopulations (including $\lambda = \mathcal{C}$), (iii) introduction of a dynamically updated sampling design via MAR, and (iv) emphasis on the new insights available only from letting $R > 2$.

2.2. Optimally Combining the Sub-Samples for Efficiency

To state our main result in Proposition 1 let us first, for a given $\lambda \in \Lambda$, define the following $d \times 1$ functions of the observed data O and the $d \times 1$ parameter β as:

$$\varphi_{r,\lambda}(O;\beta) := E\left[\frac{P(C \in \lambda | T_R(Z))}{P(C \in \lambda)} m(T_R(Z);\beta) \middle| T_r(Z)\right] \text{ for } r = 1, \ldots, R, \tag{5}$$

$$\varphi_\lambda(O;\beta) := \frac{I(C = R)}{P(C = R|T_R(Z))}\varphi_{R,\lambda}(O;\beta) + \sum_{r=1}^{R-1}\left[\frac{I(C \geq R-r)}{P(C \geq R-r|T_{R-r}(Z))} - \frac{I(C \geq R-r+1)}{P(C \geq R-r+1|T_{R-r+1}(Z))}\right]\varphi_{R-r,\lambda}(O;\beta). \tag{6}$$

PROPOSITION 1. *Let (1), (2), and (3) and assumption A hold. Let the $d \times d$ matrix $V_\lambda := Var(\varphi_\lambda(O;\beta^0_\lambda))$ be finite and positive definite where $\varphi_\lambda(O;\beta)$ is defined in (6) and β^0_λ is defined in (3). Then, the asymptotic variance lower bound for any regular estimator of β^0_λ is given by $\Omega_\lambda := M_\lambda^{-1} V_\lambda M_\lambda^{-1'}$. A regular estimator whose asymptotic variance equals Ω_λ has the asymptotically linear representation*:[5]

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta^0_\lambda) = -M_\lambda^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_\lambda(O_i;\beta^0_\lambda) + o_p(1).$$

Remarks.

1. Chen et al. (2008)'s results are for $R = 2$ with $\lambda = \{1\}$ and $\lambda = \{1,2\}$. Proposition 1 generalizes Theorem 2 of Chen et al. (2008) to the case of a generic R and a generic target λ. To see this, let $R = 2$. Then, under (1), equations (5) and (6) imply that for $\lambda = \{1,2\}$ and $\{1\}$, respectively:

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

$$\varphi_{\{1\}}(O;\beta) = \frac{P(C=1|T_1(Z))}{P(C=1)}\varphi_{\{1,2\}}(O;\beta),$$

giving exactly the same expressions as in Chen et al. (2008) (p. 830) [see Appendix A.7 for details].

[5] Over-identified models were considered in older versions of the article, but the proof for the corresponding result in those versions had logical gaps since the characterization of the tangent set was too general and did not take account of the over-identifying restrictions. We thank a referee for pointing out this problem with the older versions.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

Interestingly, however, pointing to the crux of the matter related to the planned incompleteness condition (2) is the case where $R = 2$ and $\lambda = \{2\}$. (This case is not considered in Chen et al. (2008).) In this case, our Proposition 1 implies that (following steps as in Appendix A.7):

$$\varphi_{\{2\}}(O;\beta) = \frac{I(C=2)}{P(C=2)} m(T_2(Z);\beta) + \left(\frac{P(C=2|T_1(Z))}{P(C=2)} - \frac{I(C=2)}{P(C=2)} \right) \times E[m(T_2(Z);\beta)|T_1(Z)],$$

i.e., all the subsamples still contribute toward efficient estimation (as evident from the first term inside parentheses on the RHS), a phenomenon that holds for the other targets (λ's) too. On the other hand, the generic result for unplanned incompleteness (i.e., without (2)) under MAR in Proposition 1 of Barnwell and Chaudhuri (2018) would imply that, when $R = 2$ and $\lambda = \{2\}$, then:

$$\varphi_{\{2\}[u]}(O;\beta) = \frac{I(C=2)}{P(C=2)} m(T_2(Z);\beta),$$

rendering the incomplete subsample useless. (The subscript $[u]$ stands for unknown/unplanned.) This comparison makes evident the important benefit of the planned incompleteness approach that makes all the subsamples always usable irrespective of λ. It is also straightforward to see that:

$$Var\left(\varphi_{\{2\}}(O;\beta^0_{\{2\}})\right) = Var\left(\varphi_{\{2\}[u]}(O;\beta^0_{\{2\}})\right) - E\left[\frac{P(C=1|T_1(Z))P(C=2|T_1(Z))}{P^2(C=2)} q(T_1(Z))q'(T_1(Z)) \right].$$

where $q(T_1(Z)) := E[m(T_2(Z);\beta^0_{\{2\}})|T_1(Z)]$. Hence, in this case, the difference between planned and unplanned incompleteness in terms of the efficiency bound boils down to the additional (relevant) information brought by the incomplete subsample, which is reflected by the last term on the RHS.

2. When $R > 2$, Chen et al. (2008)'s selection on observables assumption (Assumption 2) can be generalized as MAR in (1) (or its special cases noted in Section 3). Proposition 1 works under MAR. The result for a generic R under MAR and when the target is $\lambda = \mathcal{C}$ has been known since Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Robins, Rotnitzky, and Zhao (1995), and Holcroft et al. (1997).

On the other hand, the novelty in Proposition 1 is that it allows for any target λ. The key to obtaining this result under a unified framework is how we treat the term $P(C \in \lambda|T_R(Z))$ in (5) (immaterial when $\lambda = \mathcal{C}$ since $P(C \in \mathcal{C}|T_R(Z)) \equiv 1$). Our treatment simplifies to Chen et al. (2008)'s treatment when considering their verify-out-of-sample case, i.e., when $R = 2$ and $\lambda = \{1\}$, from which, however, an extension to the general case in our article may not seem obvious *ex ante*.

3. $\varphi_\lambda(O;\beta)$ in (6) belongs to the class of AIPW (Augmented Inverse Probability Weighted) estimating functions of Robins et al. (1994). The first term

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

$\varphi_{R,\lambda}(O;\beta)$ is the IPW term based on the complete subsample. The rest are the augmentations due to the incomplete subsamples: the r-th term represents the contribution of the $(R-r+1)$-th subsample. Each of these R terms are themselves unbiased estimating function for β_λ^0 but only the first one, i.e., the IPW term, is known without further assumptions [see below (4)]. The augmentation terms reduce the variability of the IPW estimating function and thereby deliver the efficient AIPW estimating function. More precisely:

$$Cov(\text{term}_1, \text{term}_r) = -Var(\text{term}_r) \text{ for } r = 2, \ldots, R$$
$$Cov(\text{term}_s, \text{term}_r) = 0 \text{ for } s \neq r \neq 1,$$
$$\text{and hence } V_\lambda = Var\left(\sum_{r=1}^{R} \text{term}_r\right) = Var(\text{term}_1) - \sum_{r=2}^{R} Var(\text{term}_r).$$

The $(R-r+1)$-th subsample's contribution to the efficiency of estimation for β_λ^0 rises with $Var(\text{term}_r)$ for $r > 1$, countering $Var(\text{term}_1)$ to decrease the variance of the estimating function.[6]

2.3. Contribution of the Observability of the $Z_{(r)}$'s Toward Efficiency

Let us now look into combining the subsamples from an alternative viewpoint that stresses on how the observability of each $Z_{(r)}$ contributes toward efficiency. To this end, rearrange the terms on the RHS of (6) and rewrite $\varphi_\lambda(O;\beta)$ as:

$$\varphi_\lambda(O;\beta) = \varphi_{1,\lambda}(O;\beta) + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r | T_r(Z))} \left[\varphi_{r,\lambda}(O;\beta) - \varphi_{r-1,\lambda}(O;\beta)\right] \quad \textbf{(7)}$$

to slice the contribution of the subsamples differently. (Note that, $I(C \geq R) \equiv I(C = R)$.) Consider the r-th term on the RHS. $\varphi_{r,\lambda}(O;\beta)$ and $\varphi_{r-1,\lambda}(O;\beta)$ differ due to $Z_{(r)}$, which is only observed for all the $(R-r+1)$ subsamples (i.e., for all the units $i = 1, \ldots, n : C_i \geq r$) as is signified by the multiplier $I(C \geq r)$. Thus, the contribution of all the R subsamples toward estimation is represented in this r-th term in an incremental fashion according to their ability in delivering an observable $Z_{(r)}$. This holds for each $r = 1, \ldots, R$, i.e., including the first term on the RHS of (7). Note that the R terms on the RHS of (7) are uncorrelated. Therefore, V_λ is the sum of the variances of the R terms:

$$V_\lambda = Var(\varphi_{1,\lambda}(O;\beta_\lambda^0)) + \sum_{r=2}^{R} E\left[\frac{Var(\varphi_{r,\lambda}(O;\beta_\lambda^0) | T_{r-1}(Z))}{P(C \geq r | T_r(Z))}\right].$$

[6] While $Var(\varphi_{R-r+1,\lambda}(O;\beta)) \geq Var(\varphi_{R-r,\lambda}(O;\beta))$ (in a matrix sense), the order is not always preserved when comparing the relative contribution of $Var(\text{term}_r)$ and $Var(\text{term}_{r+1})$ for $r > 1$ because these latter variances are affected by certain conditional probabilities in a nontrivial way as evident from the expression that $Var(\text{term}_r) = E\left[\frac{P(C=R-r+1|T_{R-r+1}(Z))}{P(C\geq R-r+1|T_{R-r+1}(Z))P(C\geq R-r+2|T_{R-r+2}(Z))}\varphi_{R-r+1,\lambda}(O;\beta_\lambda^0)\varphi'_{R-r+1,\lambda}(O;\beta_\lambda^0)\right]$ for $r = 2, \ldots, R$. This is what complicates a general optimality theory for the survey design, which is the topic of our ongoing work [see note 4].

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

The variance inflation factor $1/P(C \geq r|T_r(Z))$ accounts for the observability of $Z_{(r)}$ by varying inversely with the conditional probability of observing $Z_{(r)}$. Naturally, there is no such variance inflation for the first term on the RHS of (7) since $Z_{(1)}$ is always observed.

Yet another way of looking at these incremental contributions is to design a set of extended moment restrictions whose information content, when combined optimally, equals that in Proposition 1. Accordingly, consider the estimation of β_λ^0 based on the moment restrictions:

$$E\left[\phi_{R,\lambda}(O;\beta)\right] = 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta_\lambda^0, \tag{8}$$

$$E\left[\phi_{R-r}(O)|T_{R-r}(Z)\right] = 0 \text{ almost surely } T_{R-r}(Z) \text{ for } r = 1,\ldots,R-1; \tag{9}$$

$$\text{where } \phi_{R,\lambda}(O;\beta) := \frac{I(C=R)}{P(C=R|T_R(Z))}\varphi_{R,\lambda}(O;\beta) \quad \text{[the IPW term from (6)]},$$

$$\phi_{R-r}(O) := I(C \geq R-r)\left[I(C \geq R-r+1) - P(C \geq R-r+1|C \geq R-r, T_{R-r}(Z))\right]$$

for $r = 1\ldots, R-1$. When considering the expression for $\phi_{R-r}(O)$, note that, by definition, $I(C \geq R-r) = I(C \geq 1) \equiv 1$ when $r = R-1$, and $I(C \geq R-r+1) = I(C \geq R) \equiv I(C = R)$ when $r = 1$.

Under (1) and (2), the moment restriction in (8) already identifies β_λ^0 [see below (4)], and GMM estimation based on it using the complete subsample is the GMM-version of the Horvitz–Thompson method of obtaining IPW estimators [see, e.g., Wooldridge (2007)].

The key to our following discussion, on the other hand, is the moment restrictions in (9). These restrictions do not involve β but bring additional information due to the observability of the $Z_{(r)}$'s in the subsamples. Under the monotonic structure of the observed data, this information is usable due to the MAR condition (1) and, importantly, the planned incompleteness condition (2).

(2) did not play a role in similar discussions in the literature, e.g., Graham (2011), Chaudhuri and Guilkey (2016), etc., since they focused on the full population, i.e., $\lambda = \mathcal{C}$, for which the efficiency bound is the same irrespective of (2). However, we also consider subpopulations; and, hence, (2) will play an important role here without which the contribution of the $Z_{(r)}$'s would be further attenuated.

Additionally, the monotonic structure also plays an important role in our discussion, as is apparent from a comparison with the results in pp. 686–687 of Chaudhuri and Guilkey (2016). The monotonicity is captured by the multiplier $I(C \geq r)$ for the r-th moment function $\phi_r(O)$ for $r = 1,\ldots,R-1$. This multiplier ensures that the corresponding moment restriction reflects the additional information that becomes available due to the observability of $Z_{(r)}$, which is observed if and only if $C \geq r$.

PROPOSITION 2. *Denote* $\phi_r(O)$ *by* ϕ_r *for* $r = 1,\ldots,R-1$, *and define* $\overline{Proj}_{T_r}(Y|\phi_r) := Y - Proj_{T_r}(Y|\phi_r)$ *and* $Proj_{T_r}(Y|\phi_r) :=$

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

$E[Y\phi_r|T_r(Z)]\left(E\left[\phi_r^2|T_r(Z)\right]\right)^{-1}\phi_r$ *for any random variable* Y *whenever the relevant terms in the definition exist. Then, the following results hold.*
(*i*) *If* (*1*) *and Assumptions* (*A1*) *and* (*A2*) *hold, then* $\varphi_\lambda(O;\beta)$ *defined in* (*6*) *satisfies*:

$$\varphi_\lambda(O;\beta) = \overline{Proj}_{T_1}\left(\overline{Proj}_{T_2}\left(\ldots\overline{Proj}_{T_{R-2}}\left(\overline{Proj}_{T_{R-1}}\left(\phi_{R,\lambda}(O;\beta)\middle|\phi_{R-1}\right)\middle|\phi_{R-2}\right)\ldots\middle|\phi_2\right)\middle|\phi_1\right).$$

(*ii*) *Let* (*1*), (*2*) *and Assumption A hold. Then, the asymptotic variance lower bound under* (*8*) *and* (*9*) *for any regular estimator of* β_λ^0 *is* $\mathbf{\Omega}_\lambda$ *as defined in Proposition 1. A regular estimator with asymptotic variance* $\mathbf{\Omega}_\lambda$ *has the same asymptotically linear representation as that in Proposition 1.*

Remarks.

1. The results in (ii) under the moment restrictions (8)–(9) and the conditions (1) and (2) follow directly as a special case of Chamberlain (1992) and Ai and Chen (2012).[7]

2. The result in (i) is essentially a repeated application of equation (15) in Brown and Newey (1998) [see also Theorem 2.1 of Graham (2011)] facilitated by the monotonicity ($T_r(Z)$ nests $T_{r-1}(Z)$) of the conditioning sets in (9). As noted in pp. 686–687 of Chaudhuri and Guilkey (2016), who also refer to an earlier version of our current article, a similar exercise under a nonmonotonic structure would not lead to the efficient influence function, except in very special cases [see their note 5].

3. The broad message of Proposition 2 is that the original problem of optimally combining the subsamples can be boiled down to an equivalent problem of optimally combining a set of carefully chosen moments restrictions, a problem/idea that is perhaps more common in economics [see Appendix A.8 for further discussion and its relation with the calibration literature in survey sampling].

Graham (2011) was first to establish a similar result for the case where $R = 2$ and the target was $\lambda = \mathcal{C}$. Our setup is more involved and thus requires condition (2) and an adequately rich choice for the sequence of functions $(\phi_{R-r}(O))_{r=1}^{R-1}$ in (9) to establish this equivalence result that provides the alternative viewpoint to appreciate the contribution of the subsamples toward efficient estimation.

4. It is important to note that the more involved nature of our setup is not just that we allow for $R > 2$, but also because we allow for the subpopulations to be the target λ. This latter feature helps to highlight a pertinent implication of the planned incompleteness condition in (2). To make this point, take $R = 2$ to match the setup of Hahn (1998), Chen et al. (2008), and, importantly, Graham (2011). Now, note that, under assumption (A2) (that now becomes Graham (2011)'s Assumption 1.4) our augmenting moment restriction (9) becomes $E[I(C = 2) - P(C = 2|T_1(Z))|T_1(Z)] = 0$ almost surely in $T_1(Z)$, i.e., the same as Graham (2011)'s [equation (5)] augmenting (auxiliary) moment restriction.

[7] To match the sequential moment restrictions in Chamberlain (1992) and Ai and Chen (2012), define $T_0(Z)$ as a constant and consider the unconditional expectation in (8) equivalently as the expectation conditional on $T_0(Z)$.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

However, when $\lambda = \{1\}$, Proposition 2 gives the efficiency result only under (2) but not under unplanned incompleteness, and this can be seen simply by comparing Case 1 in Theorems 1 and 2 of Chen et al. (2008) [see Appendix A.9 for details]. This point was moot in Graham (2011) because the efficiency results are identical under planned or unplanned incompleteness when $\lambda = \mathcal{C}$ ($\equiv \{1, 2\}$). Therefore, it is important to recognize that, *in general*, equivalence results such as Proposition 2 hold only under planned incompleteness (along with monotonicity; see Remark 2).

3. A CLOSER LOOK AT TWO SPECIAL CASES OF MAR: CMAR AND INDEP

It is instructive to observe the simplifications in the efficient influence function and, thus, the efficiency bound if instead of the general MAR condition in (1), one maintains the following stronger conditions that rule out dynamically updated survey designs and makes it easier to plan ahead with the survey:

$$\text{CMAR: } P(C = r|T_R(Z)) = P(C = r|T_1(Z)) \text{ for } r = 1, \ldots, R, \tag{10}$$

$$\text{INDEP: } P(C = r|T_R(Z)) = P(C = r) \text{ for } r = 1, \ldots, R. \tag{11}$$

Convenient MAR (CMAR) sampling happens if the sampling design for the later phases is based only on the observed variables from the first phase (baseline). CMAR and MAR are trivially the same in the commonly studied case of $R = 2$, i.e., both generalize Chen et al. (2008). Independent (INDEP) sampling happens if the sampling design is independent of Z. $\lambda = \mathcal{C}$ is the only target of interest under INDEP since β^0_λ does not vary with λ; however, that is not the case under CMAR.

Given the results in Barnwell and Chaudhuri (2018), it is also instructive to compare the results thus obtained with those where one cannot maintain or, as in our article, enforce by design the condition of planned incompleteness in (2). Both issues will be extensively analyzed in this section.

Under CMAR and INDEP, $P(C \in \lambda|T_R(Z))$ becomes $P(C \in \lambda|T_1(Z))$ and $P(C \in \lambda)$, respectively, for all $\lambda \in \Lambda$. Since all subpopulations λ's are the same under INDEP but not under CMAR, the discussion under CMAR is going to be more involved. Hence, in the sequel, we primarily focus on the discussion under CMAR while complementing it with the associated results under INDEP.

3.1. Efficient Influence Functions and Efficiency Bounds

For brevity, we follow the spirit of the equivalent form of $\varphi_\lambda(O; \beta)$ in (7) and define, for all $\lambda \in \Lambda$,

$$\varphi^{\text{CMAR}}_\lambda(O; \beta) := \frac{P(C \in \lambda|T_1(Z))}{P(C \in \lambda)} \left\{ q(T_1(Z); \beta) + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r)} \left[q(T_r(Z); \beta) - q(T_{r-1}(Z); \beta) \right] \right\} \tag{12}$$

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

where:

$$q(T_r(Z);\beta) := E\left[m(T_R(Z);\beta)|T_r(Z)\right] \text{ for } r = 1,\ldots,R. \quad \textbf{(13)}$$

It is straightforward to see using (7) that $\varphi_\lambda(O;\beta)$ in (6) boils down to $\varphi_\lambda^{\text{CMAR}}(O;\beta)$ under CMAR. (While this seems trivial, we will later point out by appealing to Barnwell and Chaudhuri (2018) that such nesting does not hold in general if the planned incompleteness condition in (2) is relaxed.)

PROPOSITION 3. *Let (2), (3), (10) and Assumption A hold. Let the $d \times d$ matrix $V_\lambda := Var(\varphi_\lambda^{CMAR}(O;\beta_\lambda^0))$ be finite and positive definite where $\varphi_\lambda^{CMAR}(O;\beta)$ is defined in (12) and β_λ^0 is defined in (3). Then, the asymptotic variance lower bound for any regular estimator of β_λ^0 is given by $\Omega_\lambda := M_\lambda^{-1} V_\lambda M_\lambda^{-1'}$. A regular estimator whose asymptotic variance equals Ω_λ has the asymptotically linear representation*:

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -M_\lambda^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_\lambda^{CMAR}(O_i;\beta_\lambda^0) + o_p(1).$$

PROPOSITION 4. *Let (3), (10) and Assumption A hold. Assume $P(C = r|T_1(Z)) = P(C = r|T_1(Z);\gamma^0)$ for some $\gamma^0 \in \Gamma \subset \mathbb{R}^{d_\gamma}$ where $P(C = r|T_1(Z);\gamma)$ is known up to the finite-dimensional unknown γ for $r = 1,\ldots,R$. Let $S_\gamma(C|T_1(Z)) := \sum_{r=1}^{R}\frac{I(C=r)}{P(C=r|T_1(Z))}\frac{\partial}{\partial\gamma}P(C = r|T_1(Z);\gamma^0)$ denote the score function for γ evaluated at $\gamma = \gamma^0$, and assume that $E\left[S_\gamma(C|T_1(Z))S_\gamma(C|T_1(Z))'\right]$ is positive definite. Define*

$$\varphi_{\lambda[pu]}^{CMAR}(O;\beta) := \varphi_\lambda^{CMAR}(O;\beta) + \Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m(T_R(Z);\beta)|T_1(Z)]\middle| S_\gamma(C|T_1(Z))\right),$$

where the subscript $[pu]$ represents that $P(C = r|T_1(Z))$ is partially unknown, i.e., the finite dimensional parameter γ is unknown; $\varphi_\lambda^{CMAR}(O;\beta)$ is as in (12); and for any variables Y and X, let $\Pi(Y|X) := E[YX'](E[XX'])^{-1}X$ denote the population least squares projection when it exists.[8] *Let $V_{\lambda[pu]} := Var(\varphi_{\lambda[pu]}^{CMAR}(O;\beta_\lambda^0))$ be a $d \times d$ finite positive definite matrix. Then, the asymptotic variance lower bound for any regular estimator of β_λ^0 is given by $\Omega_{\lambda[pu]} := M_\lambda^{-1} V_{\lambda[pu]} M_\lambda^{-1'}$. A regular estimator whose asymptotic variance equals $\Omega_{\lambda[pu]}$ has the asymptotically linear representation*:

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

[8] In terms of the notation for the conditional projection $\text{Proj}_{(.|.)}$ in the statement of Proposition 2, $\Pi(Y|X) \equiv \text{Proj}_{T_0(Z)}(Y|X)$ where $T_0(Z)$ is defined as any constant, which makes the conditional projection an unconditional one.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

PROPOSITION 5. *Let (3), (10) and Assumption A hold. Define*

$$\varphi^{CMAR}_{\lambda[u]}(O;\beta) := \frac{I(C\in\lambda)}{P(C\in\lambda)}q(T_1(Z);\beta) + \frac{P(C\in\lambda|T_1(Z))}{P(C\in\lambda)} \sum_{r=2}^{R}\frac{I(C\geq r)}{P(C\geq r)}\left[q(T_r(Z);\beta) - q(T_{r-1}(Z);\beta)\right],$$

where the subscript $[u]$ *represents that* $P(C = r|T_1(Z))$ *is unknown; and* $q(T_r(Z);\beta)$ *is as defined in (13) for* $r = 1,\ldots,R$. *Let* $V_{\lambda[u]} := Var(\varphi^{CMAR}_{\lambda[u]}(O;\beta^0_\lambda))$ *be a* $d\times d$ *finite positive definite matrix. Then, the asymptotic variance lower bound for any regular estimator of* β^0_λ *is given by* $\Omega_{\lambda[u]} := M_\lambda^{-1}V_{\lambda[u]}M_\lambda^{-1'}$. *A regular estimator whose asymptotic variance equals* $\Omega_{\lambda[u]}$ *has the asymptotically linear representation*:

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta^0_\lambda) = -M_\lambda^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi^{CMAR}_{\lambda[u]}(O_i;\beta^0_\lambda) + o_p(1).$$

Remarks.

1. Proposition 3 turns out to be a special case of Proposition 1. Proposition 5 fully relaxes the planned incompleteness condition (2). Proposition 4 is an intermediate result partially relaxing (2).

2. It is straightforward to see (after some algebra) from Propositions 3–5 that:

$$V_{\lambda[u]} = E\left[\frac{P(C\in\lambda|T_1(Z))}{P^2(C\in\lambda)}q(T_1(Z);\beta^0_\lambda)q'(T_1(Z);\beta^0_\lambda) + \frac{P^2(C\in\lambda|T_1(Z))}{P^2(C\in\lambda)}\sum_{r=2}^{R}\frac{Var\left(q(T_r(Z);\beta^0_\lambda)|T_{r-1}(Z)\right)}{P(C\geq r|T_1(Z))}\right]$$

$$V_\lambda = V_{\lambda[u]} - E\left[\frac{P(C\in\lambda|T_1(Z))(1-P(C\in\lambda|T_1(Z)))}{P^2(C\in\lambda)}q(T_1(Z);\beta^0_\lambda)q'(T_1(Z);\beta^0_\lambda)\right]$$

$$V_{\lambda[pu]} = V_\lambda + B\left(E\left[S_\gamma(C|T_1(Z))S_\gamma(C|T_1(Z))'\right]\right)^{-1}B'$$

$$= V_{\lambda[u]} - Var\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}q(T_1(Z);\beta^0_\lambda) - \Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}q(T_1(Z);\beta^0_\lambda)\middle| S_\gamma(C,T_1(Z))\right)\right)$$

where $B := E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}q(T_1(Z);\beta^0_\lambda)S_\gamma(C|T_1(Z))'\right] = E\left[\frac{q(T_1(Z);\beta^0_\lambda)}{P(C\in\lambda)}\sum_{r\in\lambda}\frac{\partial}{\partial\gamma'}P(C=r|T_1(Z);\gamma^0)\right]$ $(=0$ if $\lambda = \mathcal{C})$. Therefore, $V_\lambda = V_{\lambda[pu]} = V_{\lambda[u]}$ if $\lambda = \mathcal{C}$. Otherwise $V_\lambda \leq V_{\lambda[pu]} \leq V_{\lambda[u]}$ in the matrix sense. (Proposition 4 is presented only for the purpose of this remark.) This ordering of the asymptotic variances shows that this well-known result for $R = 2$ and $\lambda = \{1\}$, $\lambda = \{1,2\}$ from Chen et al. (2008) and Hahn (1998) also holds under CMAR for a generic R and a generic target (sub)population λ.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

3. For a generic $R > 2$, CMAR is essentially an extreme dimension reduction assumption that helps to preserve the similarities of the results under planned and unplanned incompleteness. Of course, as we already saw, the forms of the efficient influence functions are different under these two cases; and there are still other important dissimilarities that we will note in the corollaries below.

However, all these dissimilarities are substantively mild compared to what would be the case under the general MAR condition in (1). Relaxing the planned incompleteness condition (2) and imposing restrictions on dimension reduction in MAR in (1), it follows from Proposition 1 in Barnwell and Chaudhuri (2018) that for subpopulations of interest such as $\lambda = \{a, a+1, \ldots, b\}$ where $a \in \{2, \ldots, b\}$ and $b \in \{2, \ldots, R\}$, the units from subsamples $1, \ldots, a-1$ are *not at all* usable for efficient estimation.

By contrast, under planned incompleteness, units from all the subsamples are usable for efficient estimation for any target λ irrespective of whether dimension reduction assumptions such as CMAR are allowed. In this sense, MAR in (1) properly nests all dimension reduction assumptions, with CMAR in (10) being the extreme one, only under planned incompleteness. On the other hand, in contrast to Remark 1 above, our Proposition 5 (under CMAR) is not a special case of Barnwell and Chaudhuri (2018) (under MAR) since neither imposes the planned incompleteness condition (2).

4. Lastly, note that if INDEP in (11) holds, then $P(C \in \lambda | T_R(Z)) = P(C \in \lambda)$ for all λ. In this case, all the subpopulations are the same and hence there is only one population of interest $\lambda = \mathcal{C}$, for which our Proposition 1 (or Proposition 3 or 4 or 5, i.e., irrespective of (2)) implies that:

$$\varphi_{\lambda=\mathcal{C}}(O;\beta) = \varphi^{\text{INDEP}}(O;\beta) := q(T_1(Z);\beta) + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r)} \left[q(T_r(Z);\beta) - q(T_{r-1}(Z);\beta) \right]. \tag{14}$$

3.2. Efficiency Gains From the Existence of Additional Incomplete Sub-Samples

The efficiency gain for a generic target λ from using all the subsamples instead of—(i) only the complete subsample, or (ii) the complete subsample and some but not all incomplete subsamples—was evident from Remark 3 following Proposition 1 (and also from Proposition 2).[9] The underlying premise in that discussion is that all $R-1$ incomplete subsamples exist and, hence, not using any subsample cannot be more beneficial (asymptotically) than using all the subsamples.

The question that we ask in this subsection is different because it changes this premise. More precisely, we ask what is, if any, the benefit *from having an additional incomplete subsample*?

[9] If the $(R-r+1)$-th subsample is not used then Remark 3 following Proposition 1 made it evident that the variance V_λ increases by $Var(\text{term}_r)$ for $r = 2, \ldots, R$. Similar conclusions would follow from Propositions 3 and 5.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

Care is required to avoid trivially positive answers by ensuring that the benefit is not entirely driven by the increase in sample size from the additional incomplete subsample, but rather incorporates the quality of information in this subsample that is actually relevant to the target population of interest (leading to zero benefits in certain cases). Accordingly, for a precise measure of "benefit", define the efficiency loss associated with the j-th element $\beta_{\lambda,j}$ from estimating β_λ based on a collection of subsamples denoted by s instead of another collection of subsamples denoted by s' as:

$$\text{Loss}(\beta_{\lambda,j}; s, s') = \lim_{n\to\infty} \frac{\frac{1}{n_{\{s\}}}\text{Avar}(\widehat{\beta}^{s}_{\lambda,j}) - \frac{1}{n_{\{s'\}}}\text{Avar}(\widehat{\beta}^{s'}_{\lambda,j})}{\frac{1}{n_{\{s'\}}}\text{Avar}(\widehat{\beta}^{s'}_{\lambda,j})} \text{where}$$
$$\lambda, s, s' \in \Lambda \text{and} j = 1, \ldots, d. \quad \textbf{(15)}$$

$n_{\{l\}} := \sum_{r\in l} n_r = \sum_{r\in l}\sum_{i=1}^{n} I(C_i = r)$ is the size of the combined subsamples in l for $l = s, s'$. $\widehat{\beta}^{l}_{\lambda,j}$ is the j-th element of $\widehat{\beta}^{l}_{\lambda}$ for $j = 1, \ldots, d$ and $l = s, s'$.

Crucially, for the question posed here, $\widehat{\beta}^{l}_{\lambda}$ is the *efficient* estimator of β_λ based on the subsamples in l. Hence, $\text{Avar}(\widehat{\beta}^{l}_{\lambda,j})$ is the asymptotic variance ignoring the existence of the subsamples not in l. For example, if $\lambda = \{1\}$ and $s = \{1, R\}$, then we replace $P(C = 1|T_1(Z))$ and $P(C = 1)$ in the result of Proposition 3 or 5 by $P(C = 1|T_1(Z), C \in \{1, R\})$ and $P(C = 1|C \in \{1, R\})$, respectively, as if only two subsamples 1 and R exist (a substitution pattern as in multinomial/conditional logit).

Thus, the estimators not using all the subsamples are not penalized for the suboptimal use of the (available) information since they are actually efficient if the subsamples they use were the only available subsamples. Letting s be included in s', the loss in (15) thus reflects the usable incremental information brought in by the additional subsamples that are included in s' but not in s.

Analytical expressions for this loss under INDEP in (11) and CMAR in (10) are intuitive, and are provided as corollaries to (14) and Proposition 3 in Corollaries 6 and 7. Analogous results without the planned incompleteness condition (2) are provided as corollary to Proposition 5 in Corollary 8.

We take $R = 3$ and always include $\{R\}$ in s, s' for identification [see (4)]. Unless $\lambda = \mathcal{C}$, we include λ in s, s' as a convention. Unless $\lambda = \{3\}$, we do not consider $s = \{3\}$ for brevity (but do so in the Monte Carlo study in Section 5).[10]

For simplicity, let $d = 1$. For $l = s, s'$, let V^{l}_{λ} denote $Var(\varphi_\lambda(O; \beta^0_\lambda))$ when the latter is modified according to the discussion below (15). To avoid clutter, we write: $q(T_r(Z); \beta^0_\lambda)$ in (13) as q_r where λ is omitted from the latter but will be clear from the context; $T_r(Z)$ as T_r; $P(C = r)$ as p_r; $P(C = r|T_1(Z))$ as $p_r(T_1)$; $P(C \in \{r, t\})$ as p_{rt}; and $P(C \in \{r, t\}|T_1(Z))$ as $p_{rt}(T_1)$ for $r, t = 1, 2, 3$.

[10] This is because the relevant comparison in such cases is rooted in the study of the suboptimality of the asymptotic variance of standard IPW estimators that has already been studied extensively in the literature. On the other hand, our focus below is the comparison between two asymptotic variances each of which is optimal under its own assumption on the availability of the subsamples as discussed in and around the definition of the loss in (15).

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

COROLLARY 6. *Let (3), (11) (i.e., INDEP) and Assumption A hold. Under INDEP in (11): (i) β_λ^0 is the same for all $\lambda \in \Lambda$, and (ii) $p_r(T_1) = p_r$ for all $r = 1, \ldots, R$. Thus, there is no distinction between planned versus unplanned incompleteness [see (14)], and hence (2) plays no role. Taking $\lambda = \mathcal{C} := \{1, 2, 3\}$, and assuming that the concerned variances exist, the following hold as $n \to \infty$:*

(*a*) $Loss(\beta_\lambda; s = \{3\}, s' = \{1, 3\}) \times V_\lambda^{\{1,3\}} = \frac{p_1}{p_3} E\left[q_1^2\right]$.

(*b*) $Loss(\beta_\lambda; s = \{3\}, s' = \{2, 3\}) \times V_\lambda^{\{2,3\}} = \frac{p_2}{p_3} E\left[q_2^2\right]$.

(*c*) $Loss(\beta_\lambda; s = \{1, 3\}, s' = \{1, 2, 3\}) \times V_\lambda^{\{1,2,3\}} = \frac{p_2}{p_{13}} E\left[q_1^2\right] + \frac{p_2}{p_3 p_{23}} E\left[Var(q_2|T_1)\right]$.

(*d*) $Loss(\beta_\lambda; s = \{2, 3\}, s' = \{1, 2, 3\}) \times V_\lambda^{\{1,2,3\}} = \frac{p_1}{p_{23}} E\left[q_1^2\right]$.

COROLLARY 7. *Let (2), (3), (10) (i.e., CMAR) and Assumption A hold. Assuming that the concerned variances exist, the following hold as $n \to \infty$:*

(*a*) $Loss(\beta_{\{1\}}; s = \{1, 3\}, s' = \{1, 2, 3\}) \times V_{\{1\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)p_2(T_1)}{p_1}\left\{\frac{q_1^2}{p_{13}(T_1)} + \frac{Var(q_2|T_1)}{p_3(T_1)p_{23}(T_1)}\right\}\middle| C = 1\right]$.

(*b*) $Loss(\beta_{\{2\}}; s = \{2, 3\}, s' = \{1, 2, 3\}) \times V_{\{2\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)p_2(T_1)}{p_2 p_{23}(T_1)} q_1^2 \middle| C = 2\right]$.

(*c1*) $Loss(\beta_{\{3\}}; s = \{3\}, s' = \{1, 3\}) \times V_{\{3\}}^{\{1,3\}} = E\left[\frac{p_{13}}{p_3}\ \frac{p_1(T_1)}{p_{13}(T_1)} q_1^2 \middle| C = 3\right]$.

(*c2*) $Loss(\beta_{\{3\}}; s = \{3\}, s' = \{2, 3\}) \times V_{\{3\}}^{\{2,3\}} = E\left[\frac{p_{23}}{p_3}\ \frac{p_2(T_1)}{p_{23}(T_1)} q_2^2 \middle| C = 3\right]$.

(*c3*) $Loss(\beta_{\{3\}}; s = \{1, 3\}, s' = \{1, 2, 3\}) \times V_{\{3\}}^{\{1,2,3\}} = E\left[\frac{p_2(T_1)p_3(T_1)}{p_3}\left\{\frac{q_1^2}{p_{13}(T_1)} + \frac{Var(q_2|T_1)}{p_3(T_1)p_{23}(T_1)}\right\}\middle| C = 3\right]$.

(*c4*) $Loss(\beta_{\{3\}}; s = \{2, 3\}, s' = \{1, 2, 3\}) \times V_{\{3\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)p_3(T_1)}{p_3 p_{23}(T_1)} q_1^2 \middle| C = 3\right]$.

(*d*) $Loss(\beta_{\{1,3\}}; s = \{1, 3\}, s' = \{1, 2, 3\}) \times V_{\{1,3\}}^{\{1,2,3\}} = E\left[\frac{p_2(T_1)p_{13}(T_1)}{p_{13}}\left\{\frac{q_1^2}{p_{13}(T_1)} + \frac{Var(q_2|T_1)}{p_3(T_1)p_{23}(T_1)}\right\}\middle| C \in \{1, 3\}\right]$.

(*e*) $Loss(\beta_{\{2,3\}}; s = \{2, 3\}, s' = \{1, 2, 3\}) \times V_{\{2,3\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)}{p_{23}(T_1)} q_1^2 \middle| C \in \{2, 3\}\right]$.

(*f1*) $Loss(\beta_{\{1,2,3\}}; s = \{1, 3\}, s' = \{1, 2, 3\}) \times V_{\{1,3\}}^{\{1,2,3\}} = E\left[\frac{p_2(T_1)}{p_{13}(T_1)} q_1^2 + \frac{p_2(T_1)}{p_3(T_1)p_{23}(T_1)} Var(q_2|T_1)\right]$.

(*f2*) $Loss(\beta_{\{1,2,3\}}; s = \{2, 3\}, s' = \{1, 2, 3\}) \times V_{\{2,3\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)}{p_{23}(T_1)} q_1^2\right]$.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

Remarks. Complementing the discussion on efficiency in Section 5 of Wooldridge (2007), let us note here that Corollaries 6 and 7 imply that there may not always be a loss in efficiency in the sense of (15) when one does not use the subsamples that have been assumed in (15) to not exist (i.e., those in s' but not in s; see the discussion below (15)). For example, if $q_2 := E[m(Z;\beta^0_\lambda)|Z_{(1)}, Z_{(2)}] = 0$, then there is never any loss in all the above cases. Similarly, there is no loss in Corollary 6(a),(d) and Corollary 7(b),(c1),(c4),(e),(f2) under a weaker condition that $q_1 := E[m(Z;\beta^0_\lambda)|Z_{(1)}] = 0$.[11]

COROLLARY 8. *Let (3), (10) (i.e., CMAR) and Assumption A hold, but (2), i.e., planned incompleteness, does not hold. Assuming that the concerned variances exist, the following hold as* $n \to \infty$:

(*a*) $Loss(\beta_{\{1\}}; s = \{1,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{1\}} = E\left[\frac{p_1(T_1)p_2(T_1)}{p_1 p_3(T_1) p_{23}(T_1)} Var(q_2|T_1)\middle| C = 1\right]$.

(*b*) $Loss(\beta_{\{2\}}; s = \{2,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{2\}} = E\left[\frac{p_3(T_1)}{p_2 p_{23}(T_1)} Var(q_2|T_1)\middle| C = 2\right]$.

(*c1*) $Loss(\beta_{\{3\}}; s = \{3\}, s' = \{1,3\}) \times V^{\{1,3\}}_{\{3\}} = 0$.

(*c2*) $Loss(\beta_{\{3\}}; s = \{3\}, s' = \{2,3\}) \times V^{\{2,3\}}_{\{3\}} = 0$.

(*c3*) $Loss(\beta_{\{3\}}; s = \{3\}\ or\ s = \{1,3\}\ or\ s = \{2,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{3\}} = E\left[\frac{p_2(T_1)}{p_3 p_{23}(T_1)} Var(q_2|T_1)\middle| C = 3\right]$.

(*d*) $Loss(\beta_{\{1,3\}}; s = \{1,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{1,3\}} = E\left[\frac{p_2(T_1)p_{13}(T_1)}{p_{13}p_3(T_1)p_{23}(T_1)} Var(q_2|T_1)\middle| C \in \{1,3\}\right]$.

(*e*) $Loss(\beta_{\{2,3\}}; s = \{2,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{2,3\}} = 0$.

(*f1*) $Loss(\beta_{\{1,2,3\}}; s = \{1,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{1,3\}} = E\left[\frac{p_2(T_1)}{p_{13}(T_1)} q_1^2 + \frac{p_2(T_1)}{p_3(T_1)p_{23}(T_1)} Var(q_2|T_1)\right]$.

(*f2*) $Loss(\beta_{\{1,2,3\}}; s = \{2,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{2,3\}} = E\left[\frac{p_1(T_1)}{p_{23}(T_1)} q_1^2\right]$.

[11] A similar analysis of the loss in (15) with MAR in (1) under the premise of Section 5.2 is theoretically problematic. To see this, consider comparing the two cases $s = \{1,3\}$ and $s' = \{1,2,3\}$. For MAR in (1) to hold, a similar analysis demands $P(C = 3|Z) = 1 - P(C = 1|Z) = 1 - P(C = 1|Z_{(1)}) = P(C = 3|Z_{(1)})$ in the former case (i.e., if $\{2\}$ does not exist), whereas the latter case can still accommodate for $P(C = 3|Z) = P(C = 3|Z_{(1)}, Z_{(2)}) \neq P(C = 3|Z_{(1)})$ contradicting the requirement in the former. This obstructs our intended analytical comparison in a strict sense. Hence, such comparisons are not done here. Nevertheless, if one still proceeds with a similar, albeit theoretically problematic, comparison under MAR, then the way $P(C \in \lambda|T_R(Z))$ enters (5) and (6) [also see Remark 2 below Proposition 1] suggests that a zero loss under CMAR in Corollary 7 may not imply a zero loss under MAR.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

Remarks. First, it is not surprising that Corollaries 7 and 8 give identical results in (f1) and (f2) since, as evident from Propositions 3 and 5, there is no difference between planned and unplanned incompleteness when $\lambda = \mathcal{C}$. Second, the results in (a), (b), (c3), and (d) imply that leaving out incomplete subsamples now results in zero loss under weaker conditions, i.e., $Var(q_2|T_1) = 0$ as opposed to the dual requirement of $Var(q_2|T_1) = 0$ and $q_1 := E[q_2|T_1] = 0$ under Corollary 7. Third, we note that such differences between planned and unplanned incompleteness can manifest more prominently if the additional subsamples in s' that are not in s are of worse quality (in terms of the observability of the elements of Z) than each subsample in s. This is evident from comparing (c1), (c2), and (e) in Corollaries 7 and 8, respectively. Consequently, a comparison of (c1) or (c2) with (c3) in Corollary 8 shows that an identically zero loss when $R = 2$ need not imply the same when $R = 3$.

4. EFFICIENT ESTIMATION

β_λ^0 in (3) can be estimated efficiently by solving moment restrictions that depend on preliminary estimation of unknown nuisance parameters. In our case, these unknown nuisance parameters enter the moment restrictions in a way that leads to a certain orthogonality condition by virtue of which their estimation does not affect the asymptotic variance of the efficient estimator of β_λ^0; see, e.g., Andrews (1994), Newey and McFadden (1994), Chen et al. (2003), etc. We provide here a brief discussion of efficient estimation of β_λ^0 by highlighting this orthogonality condition. Further details of efficient estimation, including its relation with Ai and Chen (2012), are presented in Appendix C.

To consolidate notation following Chen et al. (2003), and guided by (6), define a $d \times 1$ function:

$$g(O;\beta,h(\beta)) := \frac{I(C=R)}{P(C=R|T_R(Z))}\varphi_{R,\lambda}(O;\beta) + \sum_{r=1}^{R-1}\left[\frac{I(C\geq r)}{P(C\geq r|T_r(Z))} - \frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1}(Z))}\right]h_r(\beta), \tag{16}$$

where $h(\beta) = (h_1'(\beta), \ldots h_{R-1}'(\beta))'$ are the unknown nuisance parameters, and each $h_r(\beta)$ belongs to a class of functions $(Z,\beta) \mapsto \mathbb{R}^d$, call it $\mathcal{H}_r(\beta)$, for $r = 1, \ldots, R-1$. Let $\mathcal{H} := \{\mathcal{H}_1(\beta) \times \cdots \times \mathcal{H}_{R-1}(\beta) : \beta \in \mathcal{B}\}$ be a vector space endowed with a pseudometric $\|.\|_{\mathcal{H}}$, which is the sup-norm metric with respect to the argument β and a pseudometric with respect to the other arguments.

$g(O;\beta,h(\beta)) = \varphi_\lambda(O;\beta)$ defined in (6) if $h_r(\beta) = \varphi_{r,\lambda}(O;\beta)$ for $r = 1, \ldots, R-1$. Denote the true $h_r(\beta)$ as $h_r^0(\beta) := \varphi_{r,\lambda}(O;\beta)$ for $r = 1, \ldots, R-1$. While this suggests restricting $h_r(\beta)$ as $(T_r(Z),\beta) \mapsto \mathbb{R}^d$ for $r = 1, \ldots, R-1$, it turns out that letting $h_r(\beta)$ instead be a function of Z and β does not affect either consistency or asymptotic normality of the GMM estimator defined below.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

In light of this discussion, now define the GMM average moment vector and its expectation as:

$$G_n(\beta, h(\beta)) := \frac{1}{n}\sum_{i=1}^{n} g(O_i; \beta, (h'_{1,i}(\beta), \ldots, h'_{R-1,i}(\beta))') \text{ and } G(\beta, h(\beta)) := E\left[G_n(\beta, h(\beta))\right].$$

Then, given any standard parametric or nonparametric estimator $\widehat{h}(\beta)$ for $h(\beta)$, the GMM estimator:

$$\widehat{\beta}_\lambda \text{ of } \beta^0_\lambda \text{ is defined as a (possibly approximate) solution of } G_n(\beta, \widehat{h}(\beta)) = 0. \tag{17}$$

The key feature of our setup is an orthogonality condition that can be represented as the following identity that for any $\beta \in \mathcal{B}$ and any $h(.) \in \mathcal{H}$ (that need not be $h(\beta)$):

$$G(\beta, h(.)) = E[\varphi_{R,\lambda}(O; \beta)] = E[m(Z; \beta)|C \in \lambda] \tag{18}$$

by (1), (4), and (16). That is, $G(\beta, h(.))$ does not depend on $h(.) \in \mathcal{H}$. Its main implications are:

(F1) $G(\beta^0_\lambda, h(.)) = 0$ for any $h(.) \in \mathcal{H}$ by also using (3). Also, for any $\beta \in \mathcal{B}$ and any $h(.), \bar{h}(.) \in \mathcal{H}$: $G(\beta, h(.)) - G(\beta^0_\lambda, \bar{h}(.)) = 0 \iff E[m(Z; \beta)|C \in \lambda] - E[m(Z; \beta^0_\lambda)|C \in \lambda] = 0 \iff \beta = \beta^0_\lambda$.

(F2) The partial derivative of $G(\beta, h(\beta))$ with respect to β, denote it by $G_\beta(\beta, h(\beta))$, satisfies $G_\beta(\beta, h(\beta)) = M_\lambda(\beta) := \frac{\partial}{\partial \beta'} E\left[m(Z; \beta)|C \in \lambda\right]$, and it exists whenever $M_\lambda(\beta)$ exists.

(F3) $G(\beta, h(.)) - G(\beta, \bar{h}(.)) = 0$ for any $\beta \in \mathcal{B}$ and $h(.), \bar{h}(.) \in \mathcal{H}$. Thus, the pathwise derivative of $G(\beta, h(.))$ with respect to $h(.)$, denote it by $G_h(\beta, h(.))$, exists at all $h(.) \in \mathcal{H}$, in all directions $[\bar{h}(.) - h(.)]$ for $\{h(.) + \tau(\bar{h}(.) - h(.)) : \tau \in [0, 1]\} \subset \mathcal{H}$, and satisfies $G_h(\beta, h(.))[\bar{h}(.) - h(.)] = 0$.

(F1) helps to verify the well-separability (of the true β) assumption for consistent estimation of β^0_λ by $\widehat{\beta}_\lambda$. It is even stronger since it indicates that $\widehat{h}(\beta)$ need not converge in probability to the true $h^0(\beta)$ but can converge to any $h^\dagger(\beta) \in \text{interior}(\mathcal{H})$ without affecting the consistency of $\widehat{\beta}_\lambda$ for β^0_λ [see Proposition 9]. (F2) simplifies the Jacobian formula in the asymptotic variance of $\widehat{\beta}_\lambda$ since it implies that $G_\beta(\beta^0_\lambda, h(\beta^0_\lambda)) = M_\lambda$. The final implication (F3) is stated in a way that helps to verify that the asymptotic variance of $\widehat{\beta}_\lambda$ is unaffected by the estimation of $h(\beta)$ even if $\widehat{h}(\beta)$ converges at a rate slower than $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(n^{-1/4})$; e.g., $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ will suffice. It is worth noting that we do not even require that $h^\dagger(\beta) = h^0(\beta)$, the truth [see Proposition 10(i)]. Of course, semiparametric efficiency for $\widehat{\beta}_\lambda$ requires that $h^\dagger(\beta^0_\lambda) = h^0(\beta^0_\lambda)$, but the rate of convergence of the consistent $\widehat{h}(\beta)$ is still of no consequence as far as the first-order

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

asymptotic properties of $\widehat{\beta}_\lambda$ are concerned [see Proposition 10(ii)]. Naturally, all these nice implications of (18) also provide flexibility in estimating the nuisance parameters – (i) parametrically based on misspecified models, e.g., giving linear projections rather than conditional expectations or (ii) nonparametrically under less than satisfactory conditions that might prevent a faster than $n^{1/4}$-rate convergence of the estimator. Based on this discussion above, Propositions 9 and 10 describe the asymptotic properties of $\widehat{\beta}_\lambda$ under high level conditions that are somewhat weaker than what one typically assumes in similar contexts.

For simplicity, we follow Chen et al. (2003) and write $(\beta, h(\beta))$ as (β, h) unless confusing. Also, we write the Euclidean norm of a matrix/vector A as $\|A\| := \sqrt{\text{trace}(A'A)}$.

PROPOSITION 9. *Let (1), (3), and Assumptions (A1) and (A2) hold. Assume*:

(*B1*) $\|G_n(\widehat{\beta}_\lambda, \widehat{h})\| \leq \inf_{\beta\in\mathcal{B}} \|G_n(\beta, \widehat{h})\| + o_p(1)$ *where* $\mathcal{B}$ *is a compact subset of* $\mathbb{R}^d$;

(*B2*) $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ *for some* $h^\dagger(\beta) \in$ *interior*$(\mathcal{H})$ *for all* β, *and where* $h^\dagger(\beta)$ *is not necessarily equal to* $h^0(\beta)$;

(*B3*) *for all sequences of positive numbers* $\{\delta_n\}$ *with* $\delta_n = o(1)$,

$$\sup_{\beta\in\mathcal{B}, \|h-h^\dagger(\beta)\|_{\mathcal{H}}\leq\delta_n} \frac{\|G_n(\beta, h) - G(\beta, h)\|}{1 + \|G_n(\beta, h)\| + \|G(\beta, h)\|} = o_p(1).$$

Then $\widehat{\beta}_\lambda - \beta^0_\lambda = o_p(1)$.

PROPOSITION 10. *Let (1), (3) and Assumptions A hold. Let* $\beta^0_\lambda \in$ *interior*$(\mathcal{B})$ *and* $h^\dagger(\beta) \in$ *interior*$(\mathcal{H})$ *for all* β, *and where* $h^\dagger(\beta)$ *is not necessarily equal to* $h^0(\beta)$. *For a small* $\delta > 0$ *define the neighborhoods* $\mathcal{B}_\delta := \{\beta \in \mathcal{B} : \|\beta - \beta^0_\lambda\| \leq \delta\}$ *and* $\mathcal{H}_\delta := \{h \in \mathcal{H} : \|h - h^\dagger(\beta)\|_{\mathcal{H}} \leq \delta\}$. *Let* $\widehat{\beta}_\lambda - \beta^0_\lambda = o_p(1)$ *and* $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$. *Assume*:

(*C1*) $\|G_n(\widehat{\beta}_\lambda, \widehat{h})\| \leq \inf_{\beta\in\mathcal{B}_\delta} \|G_n(\beta, \widehat{h})\| + o_p(n^{-1/2})$;

(*C2*) $G_\beta(\beta, h^\dagger)$ *exists for* $\beta \in \mathcal{B}_\delta$ *and is continuous at* $\beta = \beta^0_\lambda$; (*note that* $G_\beta(\beta^0_\lambda, h^\dagger)$ *is nonsingular by* (*A3*) *and* (*F2*));

(*C3*) *for all sequences of positive numbers* $\{\delta_n\}$ *with* $\delta_n = o(1)$,

$$\sup_{\beta\in\mathcal{B}_{\delta_n}, h\in\mathcal{H}_{\delta_n}} \frac{\|G_n(\beta, h) - G(\beta, h) - G_n(\beta^0_\lambda, h^\dagger)\|}{n^{-1/2} + \|G_n(\beta, h)\| + \|G(\beta, h)\|} = o_p(1);$$

(*C4*) $\sqrt{n}G_n(\beta^0_\lambda, h^\dagger) \xrightarrow{d} N(0, \Sigma)$ *where* $\Sigma := E\left[g(O; (\beta^0_\lambda, h^\dagger))\, g(O; (\beta^0_\lambda, h^\dagger))'\right]$ *is finite.*

Then, the following results hold:

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

(*i*) $\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -M_\lambda^{-1}\sqrt{n}G_n(\beta_\lambda^0, h^\dagger) + o_p(1) \xrightarrow{d} N\left(0, M_\lambda^{-1}\Sigma M_\lambda^{-1'}\right)$;

(*ii*) *if, additionally,* $h^\dagger(\beta_\lambda^0) = h^0(\beta_\lambda^0)$ *then* $\Sigma = V_\lambda$ *as in Proposition 1 and*:

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -M_\lambda^{-1}\sqrt{n}G_n(\beta_\lambda^0, h^0) + o_p(1) \xrightarrow{d} N\left(0, \Omega_\lambda = M_\lambda^{-1}V_\lambda M_\lambda^{-1'}\right),$$

i.e., by Proposition 1, the estimator $\widehat{\beta}_\lambda$ *becomes semiparametrically efficient.*

Example

We conclude this section with a concrete example of the estimator $\widehat{\beta}_\lambda$ in (17). This example also informs our simulation study in Section 5. Consider a moment vector of the form $m(Z;\beta) = X(y - X'\beta)$. Let $R = 3$. For $i = 1, \ldots, n$, let $T_{ji} := T_j(Z_i)$ for $j = 1, 2, 3$, $a_{3i} := I(C_i = 3)/P(C = 3|T_{3i})$, $a_{2i} := I(C_i \geq 2)/P(C \geq 2|T_{2i}) - a_{3i}$, $a_{1i} := 1 - a_{2i} - a_{3i}$, $q := P(C \in \lambda|T_3(Z))$ and $q_i := P(C \in \lambda|T_{3i})$. Hence, ignoring the denominator $P(C \in \lambda)(\neq 0)$, (17) implies that $\widehat{\beta}_\lambda$ solves:

$$0 = \sum_{i=1}^n \left(a_{3i}q_i X_i(y_i - X_i'\beta) + a_{2i}\widehat{E}[qX(y - X'\beta)|T_{2i}] + a_{1i}\widehat{E}[qX(y - X'\beta)|T_{1i}]\right)$$

$$\Rightarrow \widehat{\beta}_\lambda = \left(\sum_{i=1}^n \left\{a_{3i}q_i X_i X_i' + a_{2i}\widehat{E}\left[qXX'|T_{2i}\right] + a_{1i}\widehat{E}\left[qXX'|T_{1i}\right]\right\}\right)^{-1}$$
$$\times \sum_{i=1}^n \left\{a_{3i}q_i X_i y_i + a_{2i}\widehat{E}\left[qXy|T_{2i}\right] + a_{1i}\widehat{E}\left[qXy|T_{1i}\right]\right\},$$

where $\widehat{E}$ denotes the estimated conditional expectations giving the estimated nuisance parameters as $\widehat{h}_2(\beta) := \widehat{E}[qX(y - X'\beta)|T_2]$ and $\widehat{h}_1(\beta) := \widehat{E}[qX(y - X'\beta)|T_1]$. Appendix C.4 provides a one-step updating efficient estimator that is useful when, unlike here, a closed form for $\widehat{\beta}_\lambda$ is unavailable.

5. SIMULATION STUDY

Now we numerically study the benefit, if any, of using all the subsamples for efficient estimation of β_λ by estimating (15) in a standard linear regression setup using a small scale Monte Carlo experiment.

The following observation motivates our experiment. In their editorial introduction, McKenzie and Rosenzweig (2012) note how the different measurements of the same variables (e.g., consumption) can dramatically alter the conclusion of analyses using survey data. But, the "good" measures can be substantially more expensive.[12] Hence, under budget constraint, one could obtain the good but expensive measures for only subsets of units, and the other measures for larger subsets or everyone.

[12] For example, collecting consumption data through maintaining a personal diary can be 6 to 10 times more expensive than a 7-day recall [see the last 3 rows of column 6 in Table 10 of Beegle et al. (2012)]. On the other hand, the 7-day recall with a short list of aggregated consumption items can understate food consumption by 30% as compared

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

Accordingly, consider a linear regression of a random variable y on a constant and another random variable X. Let X_c and X_e be mismeasured X, possibly dependent also on y. Let $Z_{(1)} = (y, X_c')'$, $Z_{(2)} = X_e$, and $Z_{(3)} = X$, a data structure that can be justified if y and X_c ("c" for cheap) are cheap to observe, while X_e ("e" for expensive) is more expensive but still cheaper to observe than X (e.g., where X is true consumption). Now, defining $\vec{X} := (1, X)'$ and $\beta_\lambda := (\beta_{\lambda,1}, \beta_{\lambda,2})'$, our experiment involves efficient estimation of β_λ by taking the moment vector in (3) as $m(Z; \beta) = \vec{X}(y - \vec{X}'\beta)$.

The generic expression of the efficient estimator for any target λ relevant for this section is given at the end of Section 4. However, when using the collection of subsamples in s, i.e., the nested collection, these estimators need to be adapted to the premise of the discussion in Section 3.2 [see below (15)]. Let us give an example to make this adaptation clear. Consider $\lambda = \{1\}$, $s' = \{1, 2, 3\}$, and $s = \{1, 3\}$. Then, the efficient estimators using s' and s are as in (19) and (20), respectively.

$$\widehat{\beta}^{s'=\{1,2,3\}}_{\lambda=\{1\}} = \left(\sum_{i=1}^{n} q_i \left\{a_{3i}\vec{X}_i\vec{X}_i' + a_{2i}\widehat{E}\left[\vec{X}\vec{X}'|T_2(Z_i)\right] + (1 - a_{2i} - a_{3i})\widehat{E}\left[\vec{X}\vec{X}'|T_1(Z_i)\right]\right\}\right)^{-1}$$
$$\times \sum_{i=1}^{n} q_i \left\{a_{3i}\vec{X}_i + a_{2i}\widehat{E}\left[\vec{X}|T_2(Z_i)\right] + (1 - a_{2i} - a_{3i})\widehat{E}\left[\vec{X}|T_1(Z_i)\right]\right\} y_i, \quad \textbf{(19)}$$

where $\widehat{E}[.|.]$ denotes the estimated conditional expectation.[13] Under MAR, $a_{3i} := I(C_i = 3)/P(C = 3|T_2(Z_i))$, $a_{2i} := [1 - I(C_i = 1)]/[1 - P(C = 1|T_1(Z_i))] - a_{3i}$ and $q_i := P(C = \lambda|T_3(Z_i) = P(C = 1|T_1(Z_i))$ for $i = 1, \ldots, n$. The only difference under CMAR is that $a_{3i} := I(C_i = 3)/P(C = 3|T_1(Z_i))$ for $i = 1, \ldots, n$. Under INDEP, $a_{3i} := I(C_i = 3)/P(C = 3)$, $a_{2i} := I(C_i \in \{2, 3\})/P(C \in \{2, 3\}) - a_{3i}$, whereas $q_i := P(C = 1)$ (a constant, which cancels out making λ moot) for $i = 1, \ldots, n$. ((19) is simplified from the generic expression in Illustration 1 in Appendix C.5 that provides further details.)

$$\widehat{\beta}^{s=\{1,3\}}_{\lambda=\{1\}} = \left(\sum_{i=1}^{n} q_i \left\{a_{3i}\vec{X}_i\vec{X}_i' + (1 - a_{3i})\widehat{E}\left[\vec{X}\vec{X}'|T_1(Z_i)\right]\right\}\right)^{-1}$$
$$\sum_{i=1}^{n} q_i \left\{a_{3i}\vec{X}_i + (1 - a_{3i})\widehat{E}\left[\vec{X}|T_1(Z_i)\right]\right\} y_i. \quad \textbf{(20)}$$

to personal diaries [compare rows 4 and 8 of column 2 in Table 2 of Beegle et al. (2012)]. These figures are based on eight different measures of consumption, with accuracy varying inversely with cost, collected for nonmonotonic subsamples of 4,029 sample units in Tanzania. For simplicity, and similarity with Section 3.2, we will consider (i) only three different measures and (ii) monotonic subsamples. (ii) helps us to focus on efficient estimation and avoid ad hoc comparisons.

[13] In our simulation experiment below, these $\widehat{E}[.|T_r(Z)]$'s are series estimators for which we always use cubic polynomials of the elements of $(1, T_r(Z)')'$ *irrespective* of the sample size n. Hence, the resulting estimators for the parameters of interest β_λ's can alternatively be considered parametric in the sense of Ackerberg, Chen, and Hahn (2012).

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

$a_{3i} := I(C_i = 3)/P(C = 3|C \in \{1,3\}, T_1(Z_i))$ and $q_i := P(C = 1|C \in \{1,3\}, T_1(Z_i))$ for $i = 1,\ldots,n$ under MAR and CMAR. While this conditioning does not affect the terms with $q_i a_{3i}$, this does affect the terms with $q_i(1 - a_{3i})$. Under INDEP, $a_{3i} := I(C_i = 3)/P(C = 3|C \in \{1,3\})$ for $i = 1,\ldots,n$ (q_i is moot as before). Conditioning on the event $C \in \{1,3\}$ adapts (20) to the premise of Section 3.2.

5.1. Simulation Design

We draw n i.i.d. copies of the concerned variables $Z = (y, X, X_c, X_e)'$ defined as follows.

$$y_i = \alpha + \delta X_i + \epsilon_i, \quad X_{ci} = X_i + I(y_i > 0)\sqrt{2}\epsilon_{ci}, \quad X_{ei} = X_i + I(y_i > 0)\epsilon_{ei},$$

where $\epsilon_i, \epsilon_{ci}, \epsilon_{ei}, X_i$ are mutually independent and i.i.d. $N(0,1)$ for all $i = 1,\ldots,n$. We take $\alpha = \delta = 1$. While $E[X] = E[X_e] = E[X_c]$, X_e is relatively less variable than X_c as a measure of X because the former has smaller variance when $y > 0$ (precisely, $Var(X_c) - Var(X_e) = P(y > 0) = \Phi(1/\sqrt{2})$). The specification is arbitrary except only to motivate the data structure from a multiphase sampling design—$Z_{(1)} = (y, X_c)'$, $Z_{(2)} = X_e$, and $Z_{(3)} = X$—because the relative accuracy of X_e helps to envision X_e as more expensive to observe than X_c but less expensive to observe than X.

In the context of dependent sampling, without regard to the optimality of the sampling design, we generate the variable $C_i \in \mathcal{C} := \{1,2,3\}$ for $i = 1,\ldots,n$ as i.i.d. copies of C such that:

$$\begin{aligned} P(C = 1|Z_i) &= F_{t_1}(\gamma_c(X_{ci} + y_i - 1)), \quad P(C = 2|C_i \geq 1, Z_i) \\ &= 1 - F_{t_1}(\gamma_e X_{ei} + \gamma_c(X_{ci} + y_i - 2)), \end{aligned}$$

and $P(C = 3|Z_i) = 1 - P(C = 1|Z_i) - P(C = 2|Z_i)$ where $F_{t_1}(a)$ is the cumulative distribution function of a t_1-distributed random variable evaluated at $a \in \mathbb{R}$. (The fat tail of the t_1 distribution helps to partly offset problems with limited overlap [see Assumption (A2) and Chaudhuri and Hill (2016)].) We design the selection mechanisms MAR in (1) and CMAR in (10) by taking $\gamma_c = \gamma_e = 0.25$ and $\gamma_c = 0.25, \gamma_e = 0$, respectively. Although $\gamma_c = \gamma_e = 0$ gives INDEP in (11), this hinders comparability with MAR and CMAR since this results in $P(C = 2) = P(C = 3) = 0.25$ while MAR and CMAR give $P(C = 2) \approx n_2/n \approx 0.31$ and $P(C = 3) \approx n_3/n \approx 0.19$ (obtained as average over 10,000 Monte Carlo trials). Hence, we directly design INDEP as $(n_1, n_2, n_3) \sim \text{Trinomial}(n, 0.5, 0.31)$.

The subsamples are made incomplete by deleting X_i if $C_i \neq 3$ and X_{ei} if $C_i = 1$ for $i = 1,\ldots,n$. We take $n = 600$, 1,200, 1,800.

The true value of the parameters of interest β_1 (Intercept) and β_2 (Slope) is $(1,1)'$, i.e., $(\alpha, \delta)'$, under INDEP. The same holds under CMAR and MAR when $\lambda = \{1,2,3\}$. However, it is difficult to analytically obtain the true values under

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

TABLE 1. Obtained as averages over 10,000 Monte Carlo trials of ordinary least squares estimates of Intercept and Slope from the regression of y on X using the correct (infeasible) subsample ($s = \lambda$) when $n = 1$ million

Target	CMAR sampling					MAR sampling				
λ	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,3\}$	$\{2,3\}$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,3\}$	$\{2,3\}$
Intercept	1.1375	0.7602	1.0087	1.1006	0.8652	1.1375	0.7624	0.9991	1.0985	0.8652
slope	0.9630	0.9318	0.9562	0.9675	0.9685	0.9630	0.9239	0.9473	0.9628	0.9685

CMAR and MAR when $\lambda \neq \{1,2,3\}$. Since the study of bias is not our focus, we take the values listed in Table 1 as (roughly) the truth for the other β_λ's.

5.2. Simulation Results

Tables 2 and 3 list the estimated loss (in percent) defined in (15) for various s with respect to $s' = \{1,2,3\}$ under INDEP, and CMAR and MAR, respectively. Although, this particular demonstration may not be strictly correct theoretically under MAR [see note 11], we nevertheless report the MAR results to get a sense of the concerned loss. Incidentally, here the losses under MAR turn out to be quite close to those under CMAR, and hence are not given special attention.

If all the subsamples contained the same variables, then these losses should more or less reflect the smaller than n size of the collection of subsamples in s. For example, the first row of Table 2 would be $100 \times (1/n_3 - 1/n)/(1/n) \approx 100 \times (1/P(C=3) - 1) \approx 426$, and similarly the second and third rows would be approximately 45 and 100, respectively. The actual loss will invariably be much smaller in the first and third rows because the units in the additional subsamples in $s' = \{1,2,3\}$ that are not in $s = \{3\}$ and $s = \{2,3\}$, i.e., the subsamples $\{1,2\}$ and $\{1\}$, respectively, are uniformly worse in terms of their information content than those in s. This is however not true for the second row since the extra subsample in s' is $\{2\}$, and a unit in it is more informative than a unit in the subsample $\{1\}$ but less so than a unit in the other subsample $\{3\}$ in s. Thus, it is not clear a priori in this case, i.e., $s = \{1,3\}$, if the actual loss will also be much smaller. All these intuitions are reflected in the tables, not only for INDEP (Table 2) but also for CMAR and MAR (Table 3).

TABLE 2. Estimated Loss$(\beta_{\lambda,j}; s, s' = \{1,2,3\})$ (in percent) defined in (15) for $j = 1$ (Intercept) and $j = 2$ (Slope). Results are based on the analytically estimated Avar averaged over 10,000 Monte Carlo trials

Target popln.	Used sample	INDEP sampling					
		Intercept			Slope		
λ	s	$n = 600$	$n = 1,200$	$n = 1,800$	$n = 600$	$n = 1,200$	$n = 1,800$
$\{1,2,3\}$	$\{3\}$	155	159	157	104	107	104
$\{1,2,3\}$	$\{1,3\}$	30	32	33	21	24	23
$\{1,2,3\}$	$\{2,3\}$	33	34	33	23	23	21

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

TABLE 3. Estimated Loss($\beta_{\lambda,j}; s, s' = \{1,2,3\}$) (in percent) defined in (15) for $j = 1$ (Intercept) and $j = 2$ (Slope). Results are based on the analytically estimated Avar averaged over 10,000 Monte Carlo trials

Target popln. λ	Used sample s	CMAR sampling						MAR sampling					
		Intercept			Slope			Intercept			Slope		
		n			n			n			n		
		600	1,200	1,800	600	1,200	1,800	600	1,200	1,800	600	1,200	1,800
{1,2,3}	{3}	156	160	159	123	128	125	165	168	167	134	139	136
{1,2,3}	{1,3}	37	38	39	43	41	39	40	42	42	47	46	44
{1,2,3}	{2,3}	47	44	43	47	45	44	49	45	43	54	48	48
{1}	{3}	126	129	127	103	110	107	134	135	133	105	111	109
{1}	{1,3}	24	26	26	17	18	17	29	31	30	18	20	19
{2}	{3}	168	174	173	120	136	135	165	174	172	139	152	151
{2}	{2,3}	24	25	25	1	5	5	22	23	21	2	4	4
{3}	{3}	151	156	155	102	107	104	148	155	153	95	101	99
{3}	{1,3}	35	37	37	32	32	30	33	36	36	25	26	25
{3}	{2,3}	42	41	41	35	33	32	41	41	40	40	36	34
{1,3}	{3}	134	137	136	105	110	108	140	142	140	104	110	109
{1,3}	{1,3}	28	29	30	21	21	20	30	32	32	20	22	21
{2,3}	{3}	170	176	175	123	134	132	176	184	184	142	151	149
{2,3}	{2,3}	35	35	35	14	16	16	36	36	36	16	17	17

There are cases like $\lambda = \{2\}, s = \{2,3\}$ under CMAR and MAR sampling where the loss for the Slope estimator is minimal and close to zero, and this is in spite of the fact that the estimator based on $s = \{2,3\}$ uses roughly half the number of observations used by the estimator based on $s' = \{1,2,3\}$. The loss can, however, be quite substantial in many cases and would be even larger if we had not restricted the definition of loss in (15) from penalizing the suboptimal use of information by the subsamples in s. Taken together, these simulation results show the obvious benefit of using all the subsamples for estimation. Furthermore, comparing the first three (i.e., the only comparable) rows of Tables 2 and 3, it is evident that such benefits could be more under dependent sampling.

Appendix C.6 reports simulation evidence of reasonably good finite-sample properties of the efficient estimator used here under all the cases considered.[14] This lends credibility to the above simulation results on efficiency loss. In turn, the simulation results help to appreciate our detailed analytical exposition of efficiency gain/loss (from Section 3.2) by quantifying them numerically.

[14] These finite-sample properties and the magnitude of the losses remain stable when the same experiment is conducted with sample sizes such as $n = 1{,}000$, 2,000, 5,000 (in older versions). For smaller sample sizes such as $n = 400$, the results for certain subpopulations, however, fluctuate without tail trimming. (The complete subsample is quite small, i.e., $n_3 \approx 76$, when $n = 400$.) While the use of adaptive negligible trimming to address the associated problem of limited overlap in related scenarios is a topic of our ongoing research [see Chaudhuri and Hill (2016) for estimation of effects of binary treatments], to our knowledge, a rigorous trimming-strategy with proper bias-correction is yet to be developed for more involved cases such as that in our present article. Hence, no trimming is done in our experiment.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

6. CONCLUSION

Our article provided a thorough discussion of efficient estimation based on planned incomplete data obtained by sampling designs that generate monotonic incompleteness in the sample. Similar cost-effective sampling designs are widely used in biostatistics, epidemiology, psychology and statistics. Recently, such designs have also found use in the laboratory and field experiments in economics.

Generalizing the related literature, our theoretical framework allowed for a general $R \geq 2$ level of incompleteness in the data and a dynamically updating sampling design to provide more flexibility in designing and implementing the survey. However, similar to the related literature, our article is actually silent about the optimality of the sampling design itself. A general optimality theory for planned incomplete designs turns out to be a difficult problem, and is a topic of our ongoing work.

Our results on efficient estimation are important because planned incompleteness, by definition, entails loss in information that makes it imperative that estimators based on such data use all available information in the data optimally. By obtaining the efficiency bound for estimation of a variety of target parameters, we established what "optimal" means in such contexts. Subsequently, we provided a thorough demonstration of how the sample units, that differ in their information content due to the planned incompleteness, contribute toward this optimality, and contrasted the characteristics and implications of this contribution from that found in the related literature.

The estimator presented in our article makes the optimal use of available information in a direct way because we designed it based on the efficient influence function that we had obtained for the concerned target parameter. This estimator turned out to be a simple two-step estimator based on preliminary estimation of nuisance parameters, and its consistency and asymptotic normality were shown to hold under weaker than standard conditions.

We hope that our presentation of the analytical and simulation evidence of the benefits from this optimal use of information, and the simplicity of the underlying efficient estimation would encourage further research to facilitate the adoption of planned incomplete surveys in the face of budget constraints.

REFERENCES

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

Ackerberg, D., X. Chen, & J. Hahn (2012) A practical asymptotic variance estimator for two-step semiparametric estimators. *The Review of Economics and Statistics* 94, 481–498.

Ai, C. & X. Chen (2012) The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics* 170, 442–457.

Andrews, D.W.K. (1994) Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* 62, 43–72.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

Ashraf, N., J. Berry, & J.M. Shapiro (2010) Can higher prices stimulate product use? Evidence from a field experiment in zambia. *American Economic Review* 100, 2383–2413.

Ashraf, N., E. Field, & J. Lee (2014) Household bargaining and excess fertility: An experimental study in zambia. *American Economic Review* 104, 2210–2237.

Barnwell, J.L. & S. Chaudhuri (2018) Efficient Estimation in Sub and Full Populations with Monotonically Missing at Random Data. Technical report, McGill University.

Beaman, L., D. Karlan, B. Thusbaert, & C. Udry (2015) *Self-Selection into Credit Markets*: *Evidence from Agriculture in Mali*. Mimeo.

Beegle, K., J.D. Weerdt, J. Friedman, & J. Gibson (2012) Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics* 98, 3–18.

Brown, B. & W. Newey (1998) Efficient semiparametric estimation of expectations. *Econometrica* 66, 453–464.

Carroll, R., D. Ruppert, & L. Stefanski (1995) *Measurement Error in Nonlinear Models*. Chapman and Hall.

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

Chamberlain, G. (1992). Comment: Sequential moment restrictions in panel data. *Journal of Business and Economic Statistics* 10, 20–26.

Chatterjee, N. & Y. Li (2010) Inference in semiparametric regression models under partial questionnaire design and nonmonotone missing data. *Journal of the American Statistical Association* 105, 787–797.

Chaudhuri, S. (2014) *A Note on Efficiency Gains from Multiple Incomplete Subsamples.* Mimeo.

Chaudhuri, S. & D.K. Guilkey (2016) GMM with multiple missing variables. *Journal of Applied Econometrics* 31, 678–706.

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

Chen, X., H. Hong, & E. Tamer (2005) Measurement error models with auxiliary data. *Review of Economic Studies* 72, 343–366.

Chen, X., H. Hong, & A. Tarozzi (2008) Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics* 36, 808–843.

Chen, X., O. Linton, & I. van Keilegom (2003) Estimation of semiparametric models when the criteria function is not smooth. *Econometrica* 71, 1591–1608.

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

Devereux, P.J. & G. Tripathi (2009) Optimally combining censored and uncensored datasets. *Journal of Econometrics* 151, 17–32.

Graham, B.S. (2011) Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica* 79, 437–452.

Graham, B.S., C. Pinto, & D. Egel (2012) Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies* 79, 1053–1079.

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

Graham, J.W., S.M. Hofer, & D.P. MacKinnon (1996) Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research* 31, 197–218.

Graham, J.W., B.J. Taylor, A.E. Olchowski, & P.E. Cumsille (2006) Planned missing data designs in psychological research. *Psychological Methods* 11, 323–342.

Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66, 315–331.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

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

Holt, C.A. & S.K. Laury (2002) Risk aversion and incentive effects. *The American Economic Review* 92, 1644–1655.

Ichimura, I. & E. Martinez-Sanchis (2005) Identification and Estimation of GMM Models by Combining Two Data Sets. Working paper.

Khan, S. & E. Tamer (2010) Irregular identification, support conditions, and inverse weight estimation. *Econometrica* 78, 2021–2042.

Lee, A.J., A.J. Scott, & C.J. Wild (2012) Efficient estimation in multiphase case–control studies. *Biometrika* 97, 361–374.

Little, R. & D. Rubin (2002) *Statistical Analysis with Missing Data*. Wiley.

McKenzie, D. & M. Rosenzweig (2012) Preface for symposium on measurement and survey design. *Journal of Development Economics* 98, 1–2.

Muris, C. (2016) Efficient GMM Estimation with a General Missing Data Pattern. Technical report, Simon Frasier University.

Newey, W.K. & D.L. McFadden (1994) Large sample estimation and hypothesis testing. In R.F. Engle & D. McFadden (eds.), *Handbook of Econometrics*, vol. IV, chapter 36. pp. 2212–2245. Elsevier Science Publisher.

Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–1057.

Raghunathan, T.E. & J.E. Grizzle (1995) A split questionnaire survey design. *Journal of the American Statistical Association* 90, 54–63.

Reilly, M. (1996) Optimal sampling strategies for two-stage studies. *American Journal of Epidemiology* 143, 92–100.

Ridder, G. & R. Moffitt (2007) The econometrics of data combination. In J.J. Heckman & E.E. Leamer (eds.), *Handbook of Econometrics*, vol. 6B, chapter 75. pp. 5470–5547. Elsevier Science Publisher.

Robins, J. & A. Rotnitzky (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of American Statistical Association* 90, 122–129.

Robins, M., A. Rotnitzky, & L. Zhao (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association* 427, 846–866.

Robins, M., A. Rotnitzky, & L. Zhao (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of American Statistical Association* 429, 106–121.

Rotnitzky, A. & J. Robins (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.

Shoemaker, D.M. (1973) *Principles and Procedures of Multiple Matrix Sampling*. Ballinger.

Thornton, R.L. (2008) The demand for, and impact of, learning HIV status. *American Economic Review* 98, 1829–1863.

Tripathi, G. (2009) Optimally combining censored and uncensored datasets. *Journal of Econometrics* 151, 17–32.

Tripathi, G. (2011) Moment-based inference with stratified data. *Econometric Theory* 27, 47–73.

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

Wacholder, S., R.J. Carroll, D. Pee, & M.H. Gail (1994) The partial questionnaire design for case-control studies. *Statistics in Medicine* 13, 623–634.

Whittemore, A.S. (1997) Multistage sampling designs and estimating equations. *Journal of Royal Statistical Society, Series B* 59, 589–602.

Wooldridge, J. (1999) Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica* 69, 1385–1406.

Wooldridge, J. (2007) Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 141(2), 1281–1301.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

APPENDIX: Proofs

The proofs are presented here in an abridged form, but all the details can be found in Appendix B for the proofs of the results for Sections 2 and 3, and in Appendix C for those for Section 4.

The proofs of Propositions 1, 3, 4, and 5 involve obtaining the semiparametric efficiency bound and the efficient influence function, under different assumptions, following Chen et al. (2008). They follow in two steps. Step 1 characterizes the tangent set for all regular parametric submodels satisfying the semiparametric assumptions on the observed data. Step 2 obtains the efficient influence function and, thereby, the asymptotic variance lower bound as the expectation of its outer product. f and F denote the density and distribution functions, with the concerned random variables specified inside parentheses. $L_0^2(F)$ denotes the space of mean-zero, square integrable functions with respect to F.

Proof of Proposition 1.

Step 1. Consider a regular parametric submodel indexed by a parameter θ for the distribution of the observed data $O = (C', T_C'(Z))'$. The log of the distribution can be expressed in terms of the full data $(C, Z')'$ as:

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

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

$$S_\theta(O) = s_\theta(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})$$

where $s_\theta(Z_{(1)}) := \frac{\partial}{\partial \theta} \log f_\theta(Z_{(1)})$ and $s_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}) := \frac{\partial}{\partial \theta} \log f_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})$. (We will omit the subscript θ from the quantities evaluated at $\theta = \theta_0$.) The tangent set is the mean square closure of all d dimensional linear combinations of $S_\theta(O)$ for all such smooth parametric submodels, and it takes the form:

$$\mathcal{T} := a_1(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) a_r(Z_{(1)}, \ldots, Z_{(r)}), \tag{A.1}$$

where $a_1(Z_{(1)}) \in L_0^2(F(Z_{(1)}))$ and $a_r(Z_{(1)}, \ldots, Z_{(r)}) \in L_0^2(F(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}))$.

Step 2. Differentiating the moment conditions in (3) with respect to θ under the integral, and noting that $P(C \in \lambda|Z)$ (which is known) does not depend on θ but $P(C \in \lambda)$ (which is unknown) does, we obtain by using (1), (3) and Assumption (A3) that:

$$\frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} = -M_\lambda^{-1} E\left[m(Z; \beta_\lambda^0) \left\{ s(Z_{(1)})' + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})' \right\} \middle| C \in \lambda \right].$$

Pathwise differentiability follows if we can find a $\psi(O) \in \mathcal{T}$ such that:

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

$$E[\psi(O)S(O)'] = \frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'}. \tag{A.2}$$

Let us conjecture that $\psi(O) = -M_\lambda^{-1}\varphi_\lambda(O;\beta_\lambda^0)$, and then verify (A.2) by equivalently showing that:

$$E[\varphi_\lambda(O;\beta_\lambda^0)S(O)'] = E\left[m(Z;\beta_\lambda^0)\left\{s(Z_{(1)})' + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})'\right\}\middle| C \in \lambda\right].$$

Consider the left hand side (LHS) and, in accordance with the partition of $\varphi_\lambda(O)$ (we work with the alternative specification in (7) for convenience), write it as $\sum_{q=1}^{R} B_q$ where, for $q = 2,\ldots,R$:

$$B_q := E\left[\frac{I(C \geq q)}{P(C \geq q|T_q(Z)}\left[\varphi_{q,\lambda}(O;\beta_\lambda^0) - \varphi_{q-1,\lambda}(O;\beta_\lambda^0)\right]S(O)'\right]$$

while $B_1 := E\left[\varphi_{1,\lambda}(O;\beta_\lambda^0)S(O)'\right]$.

To avoid notational clutter, in the rest of Step 2 we write $m(Z;\beta_\lambda^0)$ as m; $T_q(Z)$ as T_q; $\varphi_{q,\lambda}(O;\beta_\lambda^0)$ as $\varphi_{q,\lambda}$ for $q = 1,\ldots,R$; and also write $s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})$ as $s(Z_{(r)}|T_{r-1})$ for $r = 2,\ldots,R$.

First, we note that:

$$\begin{aligned} B_1 &= E\left[E\left[\frac{P(C \in \lambda|T_R)}{P(C \in \lambda)}m\middle| T_1\right]s(Z_{(1)})'\right] \\ &\quad + \sum_{r=2}^{R} E\left[E\left[\frac{P(C \in \lambda|T_R)}{P(C \in \lambda)}m\middle| T_1\right]I(C \geq r)s(Z_{(r)}|T_{r-1})'\right] \\ &= E[ms(Z_{(1)})'|C \in \lambda] + 0 \end{aligned}$$

by using MAR in (1) and the fact that $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ for $r = 2,\ldots,R$.

Now consider B_q. Since $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$, by using MAR in (1) we obtain that:

$$\begin{aligned} B_q &= \sum_{r=1}^{q-1} E\left[\frac{I(C \geq q)}{P(C \geq q|T_q)}(\varphi_{q,\lambda} - \varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right] \\ &\quad + \sum_{r=q}^{R} E\left[\frac{I(C \geq r)}{P(C \geq q|T_q)}(\varphi_{q,\lambda} - \varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right] \\ &= 0 + E[ms(Z_{(q)}|T_{q-1})'|C \in \lambda] \end{aligned}$$

for $q = 2,\ldots,R$, and combining this with the B_1 obtained above verifies (A.2).

That $\psi(O) \in \mathcal{T}$ follows from matching terms as follows. (i) $-M_\lambda^{-1}\varphi_{1,\lambda}$ is a function of only T_1, and $E[\varphi_{1,\lambda}] = 0$ and, hence, satisfies the properties of $a_1(T_1)$ in (A.1). (ii) Using the definition of φ_r, the conditional on T_{r-1} expectation of the r-th term ($r = 2,\ldots,R$, without the multiplier $I(C \geq r)$) on the RHS of $\psi(O)$ is 0 by using (1). Therefore, this

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

term, which is a function of only T_r, is also in $L_0^2(F(Z_{(r)}|T_{r-1}))$, and hence satisfies the properties of $a_r(T_r)$ in (A.1).

Therefore, we have now verified that the projection of the influence function $-M_\lambda^{-1}m(Z;\beta_\lambda^0)$ on to the tangent set $\mathcal{T}$ is $\psi(O) := -M_\lambda^{-1}\varphi_\lambda(O;\beta_\lambda^0)$. Hence, $\psi(O)$ is the efficient influence function and, therefore, the asymptotic variance lower bound is $E[\psi(O)\psi(O)'] = M_\lambda^{-1}V_\lambda M_\lambda^{-1'} =: \Omega_\lambda$. ■

Proof of Proposition 2.

(i) Let us start with $r = 1$, i.e., the residual from the projection, $\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)\middle|\phi_{R-1}\right)$, inside the innermost parenthesis on the RHS. Then we apply induction arguments. For brevity, write $\varphi_{R,\lambda}(O;\beta)$ as $\varphi_{R,\lambda}$ and $T_r(Z)$ as T_r.

First, note that direct computation and MAR in (1) give:

$$\text{Proj}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)\middle|\phi_{R-1}\right) = \left[\frac{I(C=R)}{P(C=R|T_R)} - \frac{I(C\geq R-1)}{P(C\geq R-1|T_{R-1})}\right]E[\varphi_{R,\lambda}|T_{r-1}],$$

which implies that:

$$\begin{aligned}\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)\middle|\phi_{R-1}\right) &= \frac{I(C=R)}{P(C=R|T_R)}\underbrace{\left(\varphi_{R,\lambda} - E[\varphi_{R,\lambda}|T_{R-1}]\right)} \\ &\quad + \frac{I(C\geq R-1)}{P(C\geq R-1|T_{R-1})}E[\varphi_{R,\lambda}|T_{R-1}].\end{aligned}$$

Consider the under-braced part in the RHS of the expression for $\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)\middle|\phi_{R-1}\right)$. Using $T_{R-1}\setminus T_{R-2} = Z_{(R-1)}$ and (1), note that $E\left[\left(\varphi_{R,\lambda} - E[\varphi_{R,,\lambda}|T_{R-1}]\right)\phi_{R-2}|T_{R-2}\right]$ is a $d\times 1$ vector of zeros, and hence has no contribution in the successive projections. (Terms with no contribution in the successive projections are marked by under-braces in this proof.) On the other hand,

$$E\left[\frac{I(C\geq R-1)}{P(C\geq R-1|T_{R-1})}E[\varphi_{R,\lambda}|T_{R-1}]\phi_{R-2}\middle|T_{R-2}\right] = \frac{P(C=R-2|T_{R-2})}{P(C\geq R-2|T_{R-2})}E[\varphi_{R,\lambda}|T_{R-2}].$$

For our proof by induction, first assume that the following holds for a general $r\in\{2,\ldots,R-2\}$:

$$\begin{aligned}&\overline{\text{Proj}}_{T_{R-r}}\left(\ldots\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)\middle|\phi_{R-1}\right)\ldots\middle|\phi_{R-r}\right) \\ &= \sum_{s=0}^{r-1}\frac{I(C\geq R-s)}{P(C\geq R-s|T_{R-s})}\underbrace{\left(E[\varphi_{R,\lambda}|T_{R-s}] - E[\varphi_{R,\lambda}|T_{R-s-1}]\right)} \\ &+ \frac{I(C\geq R-r)}{P(C\geq R-r|T_{R-r})}E[\varphi_{R,\lambda}|T_{R-r}].\end{aligned}$$

Now, once again using MAR in (1), we obtain that:

$$\begin{aligned}&\left(E[\overline{\text{Proj}}_{T_{R-r}}\left(\ldots\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)\middle|\phi_{R-1}\right)\ldots\middle|\phi_{R-r}\right)\phi_{R-r-1}|T_{R-r-1}]\right)\left(E[\phi_{R-r-1}^2|T_{R-r-1}]\right)^{-1} \\ &= \left(\frac{P(C=R-r-1|T_{R-r-1})}{P(C\geq R-r-1|T_{R-r-1})}E[\varphi_{R,\lambda}|T_{R-r-1}]\right)\left(\frac{P(C\geq R-r-1|T_{R-r-1})}{P(C\geq R-r|T_{R-r})P(C=R-r-1|T_{R-r-1})}\right).\end{aligned}$$

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

Hence, the proof follows by induction since the form is also valid for $r+1$, i.e.,

$$\begin{aligned}
&\overline{\text{Proj}}_{T_{R-r-1}}\left(\ldots\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)\middle|\phi_{R-1}\right)\ldots\middle|\phi_{R-r-1}\right)\\
&=\sum_{s=0}^{r}\frac{I(C\geq R-s)}{P(C\geq R-s|T_{R-s})}\left(E[\varphi_{R,\lambda}|T_{R-s}]-E[\varphi_{R,\lambda}|T_{R-s-1}]\right)\\
&\quad+\frac{I(C\geq R-r-1)}{P(C\geq R-r-1|T_{R-r-1})}E[\varphi_{R,\lambda}|T_{R-r-1}].
\end{aligned}$$

(ii) The proof follows in the same way as that of Theorem 1 in Chamberlain (1992) or, more generally, as that of Theorem 1 of Ai and Chen (2012). ■

Proof of Proposition 3. This proof follows in the same way as that of Proposition 1. The efficient influence function in this case turns out to be exactly the same as in Proposition 1 if CMAR is imposed on the latter. ■

We present the proofs of Propositions 4 and 5 in reverse order because the proof for the latter makes a reference to that for the former. Certain details of lesser importance are omitted below because they were already made explicit in the proof of Proposition 1.

To avoid notational clutter in these two proofs, when convenient, we write $m(Z;\beta_\lambda^0)$ as m; $T_r(Z)$ as T_r for $r=1,\ldots,R$; and also write $s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})$ as $s(Z_{(r)}|T_{r-1})$ for $r=2,\ldots,R$.

Proof of Proposition 5.

Step 1. Consider a regular parametric submodel indexed by θ for the joint distribution of the observed data $O=(C,T_C'(Z))'$. Because of CMAR in (10), the log of the distribution can be expressed in terms of the full data $(C,Z')'$ as:

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

(We use the same notation as before.) Then, the score function with respect to θ is:

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

where $\dot{P}_\theta(C=r|Z_{(1)}):=\frac{\partial}{\partial\theta}P_\theta(C=r|Z_{(1)})$. The tangent space is characterized by functions of the form:

$$\mathcal{T}:=a_1(Z_{(1)})+\sum_{r=2}^{R}I(C\geq r)a_r(Z_{(1)},\ldots,Z_{(r)})+\sum_{r=1}^{R}I(C=r)\frac{b_r(Z_{(1)})}{bb_r(Z_{(1)})}, \tag{A.3}$$

where $a_1(Z_{(1)})\in L_0^2(F(Z_{(1)}))$; $a_r(Z_{(1)},\ldots,Z_{(r)})\in L_0^2(F(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)}))$ for $r=2,\ldots,R$; $\sum_{r=1}^{R}b_r(Z_{(1)}))=0$, $\sum_{r=1}^{R}bb_r(Z_{(1)})=1$, and $\sum_{r=1}^{R}I(C=r)\frac{b_r(Z_{(1)})}{bb_r(Z_{(1)})}\in L_0^2(F(C|Z_{(1)}))$.

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

For a given $\lambda \in \Lambda$, the following relation obtained by two different factorization of the joint distribution of $(I(C \in \lambda), T_1(Z) \equiv Z_{(1)})$ helps us to switch between different factorizations:

$$\begin{aligned} & s(T_1) + I(C \in \lambda)\frac{\dot{P}(C \in \lambda|T_1)}{P(C \in \lambda|T_1)} + I(C \notin \lambda)\frac{\dot{P}(C \notin \lambda|T_1)}{P(C \notin \lambda|T_1)} \\ & = I(C \in \lambda)\left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(T_1|C \in \lambda)\right] + I(C \notin \lambda)\left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(T_1|C \notin \lambda)\right]. \end{aligned} \quad \textbf{(A.4)}$$

Step 2. Differentiating (3) with respect to θ under the integral:

$$\frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} = -M_\lambda^{-1} E\left[m \left\{ s(T_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})' \right\} \middle| C \in \lambda \right].$$

Then, as in the proof of Proposition 1, here we will need to correspondingly verify that:

$$E[\varphi_{\lambda[u]}^{\mathrm{CMAR}}(O;\beta_\lambda^0)S(O)'] = E\left[m \left\{ s(T_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})' \right\} \middle| C \in \lambda \right]. \quad \textbf{(A.5)}$$

We do this term by term for $\varphi_{\lambda[u]}^{\mathrm{CMAR}}(O;\beta_\lambda^0)$ and show equality of the terms on the LHS and RHS.

Consider the first term of $\varphi_{\lambda[u]}^{\mathrm{CMAR}}(O;\beta_\lambda^0)$. Since $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ for $r = 2, \ldots, R$, we can use (10) to take conditional expectations and, then using (A.4) to replace $s(T_1)$, write:

$$\begin{aligned} E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m|T_1]S(O)'\right] &= E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m|T_1]\left\{\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(T_1|C \in \lambda) - \frac{\dot{P}(C \in \lambda|T_1)}{P(C \in \lambda|T_1)}\right\}'\right] \\ &\quad + E\left[\frac{1}{P(C \in \lambda)} E[m|T_1]\dot{P}(C \in \lambda|T_1)'\right] \\ &= E[m|C \in \lambda]\frac{\dot{P}(C \in \lambda)'}{P(C \in \lambda)} + E\left[E[m|T_1]s(T_1|C \in \lambda)'|C \in \lambda\right] + 0 \\ &= 0 + E[ms(T_1|C \in \lambda)'|C \in \lambda] + 0 \end{aligned} \quad \textbf{(A.6)}$$

where the first zero in last line follows from (3). The second term follows by using (10) and noting that $E\left[E[m|T_1]s(T_1|C \in \lambda)'|C \in \lambda\right] = E\left[E[ms(T_1|C \in \lambda)'|T_1, C \in \lambda]|C \in \lambda\right] = E\left[ms(T_1|C \in \lambda)'|C \in \lambda\right]$.

Now consider the r-th term of $\varphi_{\lambda[u]}^{\mathrm{CMAR}}(O;\beta_\lambda^0)$ for $r = 2, \ldots, R$. Taking expectation conditional on T_{r-1}, and using (10) and that $s(Z_{(s)}|T_{s-1}) \in L_0^2(F(Z_{(s)}|T_{s-1}))$ for $s = r, \ldots, R$, we obtain that:

$$\begin{aligned} & E\left[\frac{P(C \in \lambda|T_1)}{P(C \in \lambda)}(E[m|T_r] - E[m|T_{r-1}])S(O)'\right] \\ & = E\left[\frac{P(C \in \lambda|Z_1)}{P(C \in \lambda)}(E[m|T_r] - E[m|T_{r-1}])\sum_{s=r}^{R} s(Z_{(s)}|T_{s-1})\right] \\ & = E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m|T_r]s(Z_{(r)}|T_{r-1})'\right] \\ & = E\left[ms(Z_{(r)}|T_{r-1})'|C \in \lambda\right]. \end{aligned} \quad \textbf{(A.7)}$$

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

Therefore, (A.6) and (A.7) verify (A.5). That $\varphi^{\text{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ belongs to $\mathcal{T}$ in (A.3) can be shown as follows. (i) Match the term $a(T_r)$ in $\mathcal{T}$ with the r-th term of $\varphi^{\text{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ for $r > 1$. (ii) Distribute the first term $s(Z_{(1)})$ in $\mathcal{T}$ according to the relation (A.4) and match the term $I(C \in \lambda)s(Z_{(1)}|C \in \lambda)$ with the first term of $\varphi^{\text{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ while keeping in mind that, by definition, $s(Z_{(1)}|C \in \lambda) \in L^2_0(F(Z_{(1)}|C \in \lambda))$. It is straightforward to verify that all the corresponding conditional expectations, as required by the definition in (A.3) and also (A.4), are zeros. The remaining terms in $\mathcal{T}$ (including the one due to distributing the terms in (ii)) are represented in $\varphi^{\text{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ by zeros. ■

Proof of Proposition 4. The references in the steps of this proof are mainly to that of Proposition 3 (i.e., effectively to that of Proposition 1) and to that of Proposition 5.

As before, we obtain the score function for a parametric submodel indexed by θ as:

$$S_\theta(O) = s_\theta(T_1) + \sum_{r=2}^{R} I(C \geq r)s_\theta(Z_{(r)}|T_{r-1}) + \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|T_1)} \left(\frac{\partial P(C=r|T_1;\gamma^0)}{\partial \gamma'} \frac{\partial \gamma^0}{\partial \theta'} \right)'.$$

Recall that $S_\gamma(C|T_1) := \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|T_1)} \frac{\partial}{\partial \gamma} P(C=r|T_1;\gamma^0)$. Let b denote constant matrices of dimension same as that of $\frac{\partial \gamma^0}{\partial \theta'}$. Then, the tangent set is characterized by the set of functions:

$$\mathcal{T} := a_1(T_1) + b' S_\gamma(C|T_1) + \sum_{r=2}^{R} I(C \geq r)a_r(T_r),$$

where $a_1(T_1) \in L^2_0(F(T_1))$, $S_\gamma(C|T_1) \in L^2_0(F(C|T_1))$ and $a_r(T_r) \in L^2_0(F(Z_{(r)}|T_{r-1}))$.

Recognizing that $P(C=r|T_1) = P(C=r|T_1;\gamma^0)$ is known up to the finite (d_γ) dimensional parameter γ, alters the relationship in (A.4) as follows:

$$s(T_1) + \frac{\partial \gamma^{0\prime}}{\partial \theta} \left[I(C \in \lambda) \frac{\frac{\partial}{\partial \gamma} P(C \in \lambda|T_1;\gamma^0)}{P(C \in \lambda|T_1)} + I(C \notin \lambda) \frac{\frac{\partial}{\partial \gamma} P(C \notin \lambda|T_1;\gamma^0)}{P(C \notin \lambda|T_1)} \right]$$
$$= I(C \in \lambda) \left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(T_1|C \in \lambda) \right] + I(C \notin \lambda) \left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(T_1|C \notin \lambda) \right].$$

As before, differentiating (3) (equivalently, (4)) with respect to θ, and using the above relationship gives:

$$\frac{\partial \beta^0_\lambda(\theta_0)}{\partial \theta'} = -M^{-1}_\lambda E \left[\frac{P(C \in \lambda|T_1)}{P(C \in \lambda)} m \left\{ s(T_1)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})' \right\} \right]$$
$$-M^{-1}_\lambda E \left[E[m|T_1] \frac{\frac{\partial}{\partial \gamma'} P(C \in \lambda|T_1;\gamma^0)}{P(C \in \lambda)} \frac{\partial \gamma^0}{\partial \theta'} \right].$$

Therefore, utilizing the expression of the efficient influence function in Proposition 3 and its relation to that in Proposition 4, the verification of pathwise differentiability reduces to verifying that:

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

$$E\left[\Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1)]\middle|S_\gamma(C|T_1)\right)S(O)'\right]=E\left[E[m|T_1]\frac{\frac{\partial}{\partial\gamma'}P(C\in\lambda|T_1;\gamma^0)}{P(C\in\lambda)}\frac{\partial\gamma^0}{\partial\theta'}\right].$$

Now, using that $E[S_\gamma(C|T_1)|T_1]=0$, $s(Z_{(r)}|T_{r-1})\in L_0^2(F(Z_{(r)}|T_{r-1})$ and CMAR in (1), it can be shown that $E\left[S_\gamma(C|T_1)\left\{s(T_1)'+\sum_{r=2}^{R}I(C\geq r)s(Z_{(r)}|T_{r-1})'\right\}\right]=0$. Subsequently, using the expressions for $S(O)$ and $S_\gamma(C|T_1)$, the verification of the above equality follows. (Certain steps that were omitted here for brevity may be less obvious. These are presented in detail in Appendix B.) ■

Proofs of Corollary 6, 7, 8. Straightforward but tedious manipulations of the results of Propositions 3 and 5 give Corollaries 7 and 8, respectively [see Chaudhuri (2014) for the proof of the latter]. Corollary 6 follows by imposing INDEP on the result of either Proposition 3 or Proposition 5. ■

Proof of Proposition 9. (F1) already implies the standard well-separability of β^0 by virtue of (3). Hence, for all $\delta>0$ there exists $\epsilon(\delta)>0$ such that $P(\|\widehat{\beta}_\lambda-\beta_\lambda^0\|>\delta)\leq P(\|G(\widehat{\beta}_\lambda,h^\dagger)\|\geq\epsilon(\delta))$.

Therefore, to establish that $\widehat{\beta}_\lambda\xrightarrow{P}\beta_\lambda^0$, it is sufficient to show that $\|G(\widehat{\beta}_\lambda,h^\dagger)\|=o_p(1)$. Assumption (B2) implies that $P(\widehat{h}(\beta)\in\mathcal{H})\to1$ uniformly in $\beta\in\mathcal{B}$ as $n\to\infty$. Hence, we work conditional on the event $\{\widehat{h}(\widehat{\beta}_\lambda)\in\mathcal{H}\}$, and, instead, show that $\|G(\widehat{\beta}_\lambda,h^\dagger)\|=o_p(1)$ conditional on $\{\widehat{h}(\widehat{\beta}_\lambda)\in\mathcal{H}\}$, since this would imply that $\|G(\widehat{\beta}_\lambda,h^\dagger)\|=o_p(1)$ unconditionally. To this end, first note that:

$$\begin{aligned}\|G(\widehat{\beta}_\lambda,h^\dagger)\|&\leq\|G(\widehat{\beta}_\lambda,h^\dagger)-G(\widehat{\beta}_\lambda,\widehat{h})\|+\|G(\widehat{\beta}_\lambda,\widehat{h})-G_n(\widehat{\beta}_\lambda,\widehat{h})\|+\|G_n(\widehat{\beta}_\lambda,\widehat{h})\|\\&=\|G(\widehat{\beta}_\lambda,\widehat{h})-G_n(\widehat{\beta}_\lambda,\widehat{h})\|+\|G_n(\widehat{\beta}_\lambda,\widehat{h})\|.\end{aligned}\tag{A.8}$$

The inequality holds by the triangle inequality (kept implicit hereafter). The equality holds by (F3).

Using (B3) and then (F3), we obtain that:

$$\begin{aligned}\|G(\widehat{\beta}_\lambda,\widehat{h})-G_n(\widehat{\beta}_\lambda,\widehat{h})\|&\leq o_p(1)\{1+\|G_n(\widehat{\beta}_\lambda,\widehat{h})\|+\|G(\widehat{\beta}_\lambda,\widehat{h})\|\}\\&\leq o_p(1)\{1+\|G_n(\widehat{\beta}_\lambda,\widehat{h})\|+\|G(\widehat{\beta}_\lambda,h^\dagger)\|\}.\end{aligned}$$

Using this along with (A.8) and (B1) gives:

$$\|G(\widehat{\beta}_\lambda,h^\dagger)\|\times(1-o_p(1))\leq o_p(1)+\|G_n(\widehat{\beta}_\lambda,\widehat{h})\|\times(1+o_p(1))\leq\inf_{\beta\in\mathcal{B}}\|G_n(\beta,\widehat{h})\|(1+o_p(1)).\tag{A.9}$$

Now, as before, and by using (18), i.e., $G(\beta_\lambda^0,h)=0$ for all $h\in\mathcal{H}$, note that:

$$\|G_n(\beta,\widehat{h})\|\leq\|G_n(\beta,\widehat{h})-G(\beta,\widehat{h})\|+\|G(\beta,\widehat{h})-G(\beta,h^\dagger)\|+\|G(\beta,h^\dagger)-G(\beta_\lambda^0,h^\dagger)\|\tag{A.10}$$

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

where $\|G(\beta,\widehat{h}) - G(\beta,h^\dagger)\| = 0$ by (F3). Since $G(\beta_\lambda^0,h^\dagger) = 0$ and $\|G(\beta,\widehat{h}) - G(\beta,h^\dagger)\| = 0$, we have:

$$\|G_n(\beta,\widehat{h}) - G(\beta,\widehat{h})\| \leq o_p(1)\{1 + \|G_n(\beta,\widehat{h})\| + \|G(\beta,h^\dagger)\| + o_p(1)\} = o_p(1) + \|G_n(\beta,\widehat{h})\| \times o_p(1)$$

by using (B3). Therefore, (A.10) gives that:

$$\|G_n(\beta,\widehat{h})\| \leq o_p(1) + \|G_n(\beta,\widehat{h})\| \times o_p(1) + 0 + \|G(\beta,h^\dagger) - G(\beta_\lambda^0,h^\dagger)\| \times o_p(1).$$

Hence, $\|G_n(\beta,\widehat{h})\| \times (1 - o_p(1)) \leq o_p(1) + \|G(\beta,h^\dagger) - G(\beta_\lambda^0,h^\dagger)\| \times o_p(1)$ where all the $o_p(1)$ terms are uniform with respect to $\beta \in \mathcal{B}$. This implies that:

$$\inf_{\beta\in\mathcal{B}} \|G_n(\beta,\widehat{h})\| \leq \sup_{\beta\in\mathcal{B}} o_p(1) + \inf_{\beta\in\mathcal{B}} \|G(\beta,h^\dagger) - G(\beta_\lambda^0,h^\dagger)\| \times \sup_{\beta\in\mathcal{B}} o_p(1) = o_p(1)$$

since $\inf_{\beta\in\mathcal{B}} \|G(\beta,h^\dagger) - G(\beta_\lambda^0,h^\dagger)\| = 0$. Hence, if follows from (A.9) that $\|G(\widehat{\beta}_\lambda,h^\dagger)\| = o_p(1)$. ■

Proof of Proposition 10.

(i) The proof consists of two parts. First, one shows the $\sqrt{n}$-consistency of $\widehat{\beta}_\lambda$, and then, using it, one shows the asymptotic normality of $\widehat{\beta}_\lambda$. Under the conditions of the proposition, $\sqrt{n}$-consistency of $\widehat{\beta}_\lambda$ follows if one shows that $\|G(\widehat{\beta},h^\dagger)\| = O_p(n^{-1/2})$. This part of the proof is very similar to that in Chen et al. (2003) with minor adjustments made to it as we did in the proof of Proposition 9. Hence, we omit the first part and instead refer the reader to the same proof (under over-identification) for Proposition 14 in Appendix C where all the details are presented.

Subsequently, by taking the $\sqrt{n}$-consistency of $\widehat{\beta}_\lambda$ as given, we now proceed to show the asymptotic normality of $\widehat{\beta}_\lambda$. Define the linearization of $G_n(\beta_\lambda^0,h^\dagger)$ as $L_n(\beta) = G_n(\beta_\lambda^0,h^\dagger) + M_\lambda(\beta - \beta_\lambda^0)$. Note that the differences from the linearization in Chen et al. (2003) arise due to (F2) and (F3). This gives:

$$\begin{aligned}
&\|G_n(\widehat{\beta}_\lambda,\widehat{h}) - L_n(\widehat{\beta}_\lambda)\| \\
&= \|G_n(\widehat{\beta}_\lambda,\widehat{h}) - G_n(\beta_\lambda^0,h^\dagger) - M_\lambda(\widehat{\beta}_\lambda - \beta_\lambda^0)\| \\
&\leq \|G_n(\widehat{\beta}_\lambda,\widehat{h}) - G_n(\beta_\lambda^0,h^\dagger) - G(\widehat{\beta}_\lambda,\widehat{h})\| + \|G(\widehat{\beta}_\lambda,\widehat{h}) - G(\widehat{\beta}_\lambda,h^\dagger)\| + \|G(\widehat{\beta}_\lambda,h^\dagger) \\
&\quad - M_\lambda(\widehat{\beta}_\lambda - \beta_\lambda^0)\| \\
&\leq \|G_n(\widehat{\beta}_\lambda,\widehat{h}) - G_n(\beta_\lambda^0,h^\dagger) - G(\widehat{\beta}_\lambda,\widehat{h})\| + \|G(\widehat{\beta}_\lambda,h^\dagger) - M_\lambda(\widehat{\beta}_\lambda - \beta^0)\| \text{ [by (F3)]} \\
&\leq o_p(1) \times \{1 + \|G_n(\widehat{\beta}_\lambda,\widehat{h})\| + \|G(\widehat{\beta}_\lambda,\widehat{h})\|\} + \|G(\widehat{\beta}_\lambda,h^\dagger) - G(\beta_\lambda^0,h^\dagger) - M_\lambda(\widehat{\beta}_\lambda - \beta^0)\|,
\end{aligned}$$

where the term inside braces follows from (C3). Note that, the inclusion of $G(\beta_\lambda^0,h^\dagger)$ in the last term is innocuous since $G(\beta_\lambda^0,h^\dagger) = 0$. Now, by the definition of M_λ, Assumptions (C2), (A3), and (F2), it follows that $\|G(\widehat{\beta}_\lambda,h^\dagger) - G(\beta_\lambda^0,h^\dagger) - M_\lambda(\widehat{\beta}_\lambda - \beta^0)\| = o_p(\|\widehat{\beta}_\lambda - \beta_\lambda^0\|)$, which is $o_p(n^{-1/2})$ since $\widehat{\beta}_\lambda - \beta_\lambda^0 = O_p(n^{-1/2})$. On the other hand, one can also show that $\|G_n(\widehat{\beta}_\lambda,\widehat{h})\| \leq \inf_{\beta\in\mathcal{B}_\delta} \|G_n(\beta,\widehat{h})\| + o_p(n^{-1/2}) = O_p(n^{-1/2})$, for which all the details are given in the proof for Proposition 14 in Appendix C. Finally, since $\|G(\widehat{\beta}_\lambda,\widehat{h})\| \leq \|G(\widehat{\beta}_\lambda,\widehat{h}) - G(\widehat{\beta}_\lambda,h^\dagger)\| + \|G(\widehat{\beta}_\lambda,h^\dagger)\| = O_p(n^{-1/2})$ because the first term is 0 by (F3) and the second term is $O_p(n^{-1/2})$ from the first part of the proof, we obtain that

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press

$\|G_n(\widehat{\beta}_\lambda, \widehat{h}) - L_n(\widehat{\beta}_\lambda)\| \leq o_p(n^{-1/2})$. Similarly, for $\bar{\beta} := \arg\min_\beta \|L_n(\beta)\|$, that, by construction, satisfies $\sqrt{n}(\bar{\beta} - \beta_\lambda^0) = -M_\lambda^{-1}\sqrt{n}G_n(\beta_\lambda^0, h^\dagger)$, we can show that $\|G_n(\bar{\beta}, \widehat{h}) - L_n(\bar{\beta})\| \leq o_p(n^{-1/2})$. Now that the proximity of $G_n(\beta, \widehat{h})$ and $L_n(\beta)$ has been established at $\widehat{\beta}_\lambda$ and $\bar{\beta}$, respectively, the rest of the proof is to show that $\sqrt{n}(\bar{\beta} - \widehat{\beta}_\lambda) = o_p(1)$. As was the case in Chen et al. (2003), this does not involve anything particularly related to the key feature of our setup, and hence follows exactly in the same way as in the proof of Theorem 3.3 and Lemma 3.5 in Pakes and Pollard (1989).

(ii) This follows by noting that $g(O; \beta, h^0(O; \beta)) = \varphi_\lambda(O; \beta)$ defined in (6). ■

https://doi.org/10.1017/S0266466619000239 Published online by Cambridge University Press