Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: www.tandfonline.com/journals/ubes20

Missing Endogenous Variables in Conditional Moment Restriction Models

Antonio Cosma, Andreï Victorovitch Kostyrka & Gautam Tripathi

To cite this article: Antonio Cosma, Andreï Victorovitch Kostyrka & Gautam Tripathi (04 May 2026): Missing Endogenous Variables in Conditional Moment Restriction Models, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2026.2619543

To link to this article: https://doi.org/10.1080/07350015.2026.2619543

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2026, VOL. 0, NO. 0, 1–13
https://doi.org/10.1080/07350015.2026.2619543

Missing Endogenous Variables in Conditional Moment Restriction Models

Antonio Cosma[a], Andreï Victorovitch Kostyrka[b], and Gautam Tripathi[b]

[a]Department of Management, University of Bergamo, Bergamo, Italy; [b]Department of Economics and Management, University of Luxembourg, Esch-sur-Alzette, Luxembourg

ABSTRACT

We estimate finite-dimensional parameters in conditional moment restriction (CMR) models when at least one of the endogenous variables (outcomes and/or explanatory variables) in the model is missing for some individuals in the sample. We demonstrate that efficiency gains in estimation occur if and only if there is at least one endogenous variable—included in or excluded from the CMR model—that is nonmissing (observed for all individuals in the sample), which we show characterizes informative imputation. We propose a semiparametrically efficient estimator which is also "doubly robust." To illustrate the insights our estimator can provide in empirical applications with large sample sizes, we artificially induce missingness in the female labor supply model of Angrist and Evans. Despite medium levels of missingness in female labor income (the outcome) and a sample size exceeding 200,000 observations, the inverse propensity score weighted generalized method of moments (GMM) estimator finds only a statistically insignificant negative effect of having a third child (the endogenous regressor) on labor income. In contrast, our efficient estimator yields point estimates of this effect that are not only comparable to the GMM estimates but are also statistically significant.

ARTICLE HISTORY
Received June 2024
Accepted December 2025

KEYWORDS
Efficient estimation;
Informative imputation;
Missing at random

1. Introduction

Applied researchers frequently estimate models using datasets where certain variables are missing for some individuals in the sample. For example, Abrevaya and Donald (2017) note that almost 40% of the papers that appeared in the American Economic Review, the Journal of Human Resources, the Journal of Labor Economics, and the Quarterly Journal of Economics between 2006 and 2008 dealt with missing data, and in almost 70% of these cases the missing observations were simply dropped. However, dropping each observation with a missing variable and estimating the model only on the subsample where all variables are observed leads to selection bias. The term "selection bias" is a generic description of the problem that arises in identifying features of a "full" population from an "observed" subpopulation without taking into account the relationship between the two (the "full" vs. "observed" terminology, defined in Section 2, is from Robins, Rotnitzky, and Zhao, 1994). If not corrected, selection bias can lead to severely misleading inference. There are two mutually exclusive approaches for dealing with the selection problem: "selection on observables" and "selection on unobservables." In a selection on observables approach—selection on unobservables is briefly discussed in the supplement (see Appendix D.2.1)—a "missing at random" (MAR) assumption is made that, conditional on the nonmissing variables, has the effect of randomly assigning the missingness label to the "potentially missing" variables in the full population. [A random variable is said to be "potentially missing" if the probability that it is not observed for each individual lies in the open interval (0, 1). In contrast, a random variable is "nonmissing" if the probability that it is not observed for each individual is zero.] This random assignment feature of the MAR assumption leads to an "inverse probability weighted" (IPW) scheme that makes the full population and the observed subpopulation statistically indistinguishable, thereby enabling identification of the full population features from the observed subpopulation.

In this paper, we consider the estimation of finite-dimensional parameters in conditional moment restriction (CMR) models when some, or all, of the endogenous variables, that is, those variables that do not appear in the conditioning set, are potentially missing. The missing endogenous variables can either be endogenous outcomes, or endogenous explanatory variables, or both. Endogeneity—pervasive in empirical research and observational studies so much so that it is almost a defining feature of microeconometrics—typically arises in the context of omitted variables, simultaneity or reverse causality, measurement error, and model interpretation. Missingness naturally occurs due to reporting issues or when researchers use multiple data sources to compile their datasets. Thus, endogeneity and missingness are both widespread in empirical applications and ignoring either leads to biased statistical inference. Nevertheless, applied researchers often choose to ignore one or the other to simplify their tasks.

To identify the parameters, we use a selection on observables approach pioneered by Graham (2011) for unconditional moment restriction (UCMR) models, and extended by Hristache and Patilea (2017), henceforth HP, for CMR models, who

CONTACT Gautam Tripathi *gautam.tripathi@uni.lu*. Department of Economics and Management, University of Luxembourg, Esch-sur-Alzette, Luxembourg.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/07350015.2026.2619543.

show that a moment condition model and the MAR assumption in the full population are equivalent to a system of sequential moment restrictions in the observed subpopulation. [Exogenous variables, that is, variables appearing in the conditioning set, can also be missing in empirical applications. However, they have to be handled differently than missing endogenous variables because, unlike the latter, they do not lead to sequential conditional moment restrictions (HP, p. 740). Research on this topic is in progress and will be reported in a subsequent paper.] The framework of HP is very general and accommodates unconditional and conditional moment restrictions (see Remark D.1), infinite-dimensional parameters, missing outcomes, and missing exogenous covariates. Their focus, however, is on establishing their equivalence result. In contrast, our goal is the semiparametrically efficient estimation of the parameters of interest. Crucially, HP do not consider the role that nonmissing endogenous variables—ubiquitous in applied research—play in generating efficiency gains. For this reason, they cannot provide the necessary and sufficient conditions under which imputation is informative, whereas we do (cf. Remark 4.1 for a detailed discussion of this conceptual distinction). The terms "informative" and "uninformative" imputation are defined in the discussion preceding Lemma 4.2. The necessary and sufficient conditions for imputation to be informative that we develop are not merely technical: they offer clear practical guidance for empirical work. As we elaborate below, CMR models with missing endogenous variables arise frequently in applications, yet the conditions under which imputing these variables yields efficiency gains are often misunderstood.

The main contributions of our paper are as follows: (i) To our knowledge, it is the first to show that in CMR models with potentially missing endogenous variables (outcomes and/or covariates), the existence of nonmissing endogenous variables is both necessary and sufficient for achieving efficiency gains in estimation from the observed sample, and this condition is equivalent to imputation being informative (Lemma 4.2). The nonmissing endogenous variables can be endogenous outcomes and/or covariates included in the model, and/or endogenous variables that are excluded from the CMR model but enter the propensity score function (Section 2). The efficiency bound for the model parameters reveals a new CMR for constructing efficient estimators that are also "doubly robust" (Section 4.1). (ii) We propose a smoothed empirical likelihood (SEL) estimator that uses the observed sample and is semiparametrically efficient (Section 4.2). For fast and reliable implementation of the SEL estimator and related inference, we have developed an open-source `R` package called `smoothemplik` (Kostyrka 2025), available on the Comprehensive `R` Archive Network (CRAN) at https://cran.r-project.org/package=smoothemplik. (iii) To illustrate the insights our estimator can provide in empirical applications with large sample sizes, we artificially induce missingness in the female labor supply model of Angrist and Evans (1998). We find that even with medium levels of missingness in female labor income (the outcome variable), having more than 200,000 observations is insufficient for a researcher using IPW generalized method of moments (GMM) to detect a statistically significant negative effect of having a third child (the endogenous explanatory variable) on labor income. In contrast, our efficient estimator yields point estimates of this effect that are not only comparable in sign and magnitude to the GMM estimates but are also statistically significant (Section 5). (iv) A simulation study (supplement Appendix C) reveals that the SEL estimator performs well in medium-sized samples for both point estimation and inference. The efficiency gains achieved are comparable to the maximum gains the simulation design can deliver.

The paper is organized as follows. Section 2 introduces a general CMR model with potentially missing endogenous variables. Section 3 discusses identification, and Section 4 develops efficient estimation and inference. Section 5 presents the empirical illustration, and Section 6 concludes with practical suggestions for researchers dealing with missing endogenous variables in CMR models. Implementation details, the simulation study, additional examples, and all proofs are in Appendix A–D of the online supplement.

2. Model with Missing Endogenous Variables

Let Y_i^*, Z_i, X_i be random (column) vectors for individual $i = 1, \ldots, n$. Vector Y_i^* consists of endogenous variables (outcomes and/or explanatory variables), all of which are simultaneously not observed for some individuals in the sample. In contrast, vector Z_i consists of those endogenous variables (can be endogenous outcomes and/or endogenous covariates) that are observed for each individual in the sample. Similarly, X_i is a vector of exogenous variables which are observed for each individual in the sample. We refer to the coordinates of Y^* as being potentially missing or simply "missing" (for some individuals). Analogously, the coordinates of (Z, X) are referred to as being "nonmissing" (for all individuals).

For each i, we also observe the dummy variable $D_i := 1$ if all coordinates of Y_i^* are observed, and $D_i := 0$ if all coordinates of Y_i^* are missing. We let $Y_i := D_i Y_i^* + (1 - D_i)\mathfrak{m}$ denote the observed version of Y_i^*, where $\mathfrak{m}_{\dim(Y^*)\times 1}$ is a vector of pre-specified numbers for coding missingness, for example, $\mathfrak{m} := (99999, \ldots, 99999)_{\dim(Y^*)\times 1}$. Following Robins, Rotnitzky, and Zhao (1994, p. 848), RRZ hereafter, we refer to (Y_i^*, Z_i, X_i) as the "full data," and (D_i, Y_i, Z_i, X_i) as the "observed data," for individual i. Hence, $(Y_i^*, Z_i, X_i : 1 \le i \le n)$ is the "full sample" and $(D_i, Y_i, Z_i, X_i : 1 \le i \le n)$ the "observed sample." The subsample with no missing observations—obtained from the observed sample by discarding those i for which $D_i = 0$—is called the "validation sample."

A large class of econometric models in applied economics can be written as a system of conditional moment equalities, namely, there exists $\theta^* \in \Theta \subset \mathbb{R}^{\dim(\theta^*)}$ such that

$$\mathbb{E}[g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}, \tag{2.1}$$

where $g := g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*)$ is a vector of residuals (known up to θ^*) from a set of structural equations used by the researcher to model a system of relationships between the missing and the nonmissing variables. [If there is no conditioning, then (2.1) becomes an UCMR model with some variables missing as in Chen, Hong, and Tarozzi (2008) and Graham (2011); cf. Example D.4.] The missing variables Y^* and the nonmissing variables $Z_{\text{in}} \subset Z := (Z_{\text{in}}, Z_{\text{ex}})$ are classified as endogenous (with respect to g) because they appear in g but not in the conditioning set in (2.1). We refer to Z_{in} as the included (in g) endogenous variables.

The excluded (from g) nonmissing endogenous variables Z_{ex} appear neither in g nor in the conditioning set in (2.1), either due to exclusion restrictions imposed by economic theory or because of their auxiliary nature. Nonetheless, Z_{ex} (together with the remaining nonmissing variables) may be present in the probability mechanism generating the missing Y^* (i.e., the propensity score defined in Section 3), and being correlated with g can influence the probability that Y^* is observed (cf. Example D.1). The nonmissing variables $X := (X_{\text{in}}, X_{\text{ex}})$ in the conditioning set in (2.1) are classified as exogenous (with respect to g), where X_{in} are the included instrumental variables (IV) and X_{ex} the excluded IV. Included instruments are exogenous variables in g, whereas excluded instruments are those exogenous variables not in g but, based on theoretical or external considerations, appear in the conditioning set to help identify θ^*. X_{ex} also contains exogenous variables that are in the propensity score but are excluded from g. The conditional distribution of $Y^*, Z \mid X$, and the marginal distribution of X (denoted by P_X), are unknown. The objective is to use the observed sample $(D_i, Y_i, Z_i, X_i : 1 \leq i \leq n)$ to efficiently estimate θ^* in the CMR model (2.1).

Example 2.1 (IV regression with missing outcomes). The canonical example of (2.1) is the linear regression $Y^* = \alpha^* + X_{\text{in}}'\beta^* + Z_{\text{in}}'\gamma^* + U$, where only the scalar outcome Y^* is missing for some observations, Z_{in} is the vector of nonmissing included endogenous regressors, and $\mathbb{E}[U \mid X] \overset{P_X\text{-a.s.}}{=} 0$ signifying that the nonmissing included and excluded regressors $X := (X_{\text{in}}, X_{\text{ex}})$ are exogenous with respect to U. Here, $g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) := U = Y^* - \alpha^* - X_{\text{in}}'\beta^* - Z_{\text{in}}'\gamma^*$ with $\theta^* := (\alpha^*, \beta^*, \gamma^*)$. If all regressors are endogenous, then $X_{\text{in}} = \vec{\varnothing}$, that is, X_{in} is the empty vector, $X := X_{\text{ex}}$, and the definition of θ^* is adjusted by dropping β^*. The case where the outcome variable Y_1^* and some of the included endogenous explanatory variables Y_2^* are missing is handled by letting $g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) := Y_1^* - \alpha^* - X_{\text{in}}'\beta^* - Z_{\text{in}}'\gamma^* - Y_2^{*\prime}\delta^*$ with $Y^* := (Y_1^*, Y_2^*)_{1+\dim(Y_2^*)\times 1}$ and $\theta^* := (\alpha^*, \beta^*, \gamma^*, \delta^*)$. In all cases, if there are nonmissing excluded endogenous variables then $Z_{\text{ex}} \neq \vec{\varnothing}$. A classic empirical application that fits the framework of Example 2.1 is estimating the returns to schooling as in Balestra and Backes-Gellner (2017), where the outcome (earnings) is subject to missingness due to survey nonresponse and a nonmissing included endogenous regressor (educational attainment) is instrumented using an excluded IV (Swiss reforms on compulsory schooling).

Example 2.2 (IV regression with nonmissing outcomes and missing endogenous covariates). If the scalar outcome (denoted by $Z_{1,\text{in}}$) is nonmissing but endogenous explanatory variables Y^* are missing, then let $g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) := Z_{1,\text{in}} - \alpha^* - X_{\text{in}}'\beta^* - Z_{2,\text{in}}'\gamma^* - Y^{*\prime}\delta^*$ with $Z_{\text{in}} := (Z_{1,\text{in}}, Z_{2,\text{in}})_{1+\dim(Z_{2,\text{in}})\times 1}$ and $\theta^* := (\alpha^*, \beta^*, \gamma^*, \delta^*)$. If there are no nonmissing included endogenous covariates, then $Z_{2,\text{in}} := \vec{\varnothing}$ and the definition of θ^* is adjusted by dropping γ^*. If there are nonmissing excluded endogenous variables, then $Z_{\text{ex}} \neq \vec{\varnothing}$. Empirical studies consistent with the setup of Example 2.2 include: Bennedsen, Nielsen, Perez-Gonzalez, and Wolfenzon (2007, Section II.A), who estimate the causal effect of CEO succession on firm profitability, where the outcome is nonmissing but the endogenous dummy regressor (indicating if the incoming CEO is family) is missing due to dataset merging and instrumented by the gender of the outgoing CEO's first child; McDonough and Millimet (2017, Section 4), who instrument missing birth weight (an endogenous regressor) with nutritional program participation in a regression with math test scores as the nonmissing outcome; and Stephens and Unayama (2019, Section III), who estimate a repeated cross-section linear probability model with a nonmissing binary outcome (shared living arrangement) and a missing endogenous regressor (social security benefits), instrumented using cohort-based variation from amendments to the Social Security Act.

3. Identification

To identify, that is, uniquely define, θ^* without modeling the selection equation that generates the missing Y^*, we follow a selection on observables approach and assume that, conditional on all included and excluded nonmissing variables $Z := (Z_{\text{in}}, Z_{\text{ex}})$ and $X := (X_{\text{in}}, X_{\text{ex}})$, the missing observations on Y^* are missing at random, that is,

Assumption 3.1 **(MAR).** For all individuals, $D \perp\!\!\!\perp Y^* \mid Z, X$, where the symbol " $\perp\!\!\!\perp$" denotes stochastic independence.

Let $\pi(Z, X) := \Pr(D = 1 \mid Z, X)$ denote the propensity score function. It is through the propensity score that the excluded nonmissing endogenous variables Z_{ex} enter the missing data mechanism, thereby rendering the imputation "informative," as defined in Section 4. Example D.1 in the supplement illustrates how nonmissing endogenous variables can be excluded from the CMR model and yet still appear in the propensity score. Henceforth, arguments taken by functions are suppressed when there is no danger of confusion, for example, we write $\pi := \pi(Z, X)$ and $g_{\text{obs}} := g(Y, Z_{\text{in}}, X_{\text{in}}, \theta^*)$. The identity $Dg = Dg_{\text{obs}}$ (recall $g := g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*)$) due to the definition of Y is often used.

We can use MAR to evaluate $\mathbb{E}[g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) \mid X]$ when Y^* is missing because $\mathbb{E}[g \mid Z, X] \overset{\text{MAR}}{=} \mathbb{E}[g \mid Z, X, D = 1] \overset{P_{Z,X}\text{-a.s.}}{=} \mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid Z, X]$. Therefore, under MAR,

$$\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \iff \mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}. \tag{3.1}$$

The right-hand-side of (3.1), which does not contain any missing Y^*, employs an IPW moment function to correct the effects of missingness. To emphasize the nonparametric nature of the propensity score function, we assume that

Assumption 3.2. The functional form of $(Z, X) \mapsto \pi(Z, X)$ is fully unknown.

Although π is unknown, it is nonparametrically identified and estimated as the conditional expectation of $D \mid Z, X$ from the observed sample (not just the validation sample) because (D, Z, X) are nonmissing. As shown in Proposition D.1 in the supplement, if the columns of the Jacobian matrix $J_{\dim(g)\times\dim(\theta^*)} := J(X, \theta^*) := \partial_\theta \mathbb{E}[g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) \mid X]$ are linearly independent P_X-a.s., then local identification of θ^* in the CMR (2.1) is equivalent to the local identification of θ^* in the IPW CMR (3.1). Moreover, the same condition leads to the global identification of θ^* whenever g is linear in θ^*. Since

local identification of the parameters of interest in the missing data problem is not lost under MAR, and local identification is necessary for global identification, we maintain that

Assumption 3.3. θ^* is identified.

4. Efficient Estimation and Inference under MAR

Throughout the paper, the observed data $\mathcal{A}_i := (D_i, Y_i, Z_i, X_i)$, $i = 1, \ldots, n$, are assumed to be i.i.d. Unless specified otherwise, limits are taken as the sample size $n \to \infty$.

The equivalence in (3.1) reveals that, under MAR, θ^* can be estimated from the IPW CMR $\mathbb{E}\left[\frac{Dg_{\text{obs}}}{\pi} \mid X\right] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$, which uses only the validation sample. However, in practice, estimating θ^* using the validation sample alone is not advisable due to the efficiency loss from discarding the observations on (Z, X), even though they are not missing. It is, therefore, important to know the efficiency bound for estimating θ^* in (2.1) under MAR (loosely speaking, the efficiency bound is the smallest asymptotic variance of an estimator that optimally utilizes the information from all nonmissing observations). We motivate the efficiency bound for θ^* using HP (Theorem 1), which extends the results in Graham (2011, Theorem 2.1) to CMR models.

Consider the system of $\dim(g) + 1$ equations

$$\mathbb{E}\left[\frac{Dg_{\text{obs}}}{\pi} \mid X\right] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \tag{4.1}$$

$$\mathbb{E}\left[\frac{D}{\pi} - 1 \mid Z, X\right] \overset{P_{Z,X}\text{-a.s.}}{=} 0, \tag{4.2}$$

which do not contain any missing observations (cf. Remark D.3(i)). Eqn. (4.1) identifies θ^* in the validation sample, whereas (4.2) defines π in the observed sample. Remarkably, by Theorem 1 of HP, the CMRs in (4.1)&(4.2) are equivalent to (2.1) and MAR, that is,

$$(4.1) \;\&\; (4.2) \iff (2.1) \;\&\; \text{MAR}. \tag{4.3}$$

The equivalence in (4.3) reveals that, under MAR, the efficiency bound for θ^* in (2.1) is equal to the efficiency bound for estimating θ^* in (4.1)&(4.2), which is a system of sequential CMRs, that is, CMRs with increasing conditioning sets.

To eliminate the effect of estimating π on (4.1), we follow earlier approaches (see Remark D.3(ii)) and transform the sequential system (4.1)&(4.2) into a system of conditional-on-X moment restrictions based on the vector of residuals from projecting Dg_{obs}/π, the moment function in (4.1), coordinatewise onto the tangent space of score functions for π, the "nuisance parameter" in (4.2). These residuals satisfy a conditional-on-X moment restriction, on which estimation of θ^* can be based.

Let $\mu_{\dim(g)\times 1} := \mu(Z, X, \theta^*) := \mathbb{E}[g \mid Z, X] \overset{\text{MAR}}{=} \mathbb{E}[g_{\text{obs}} \mid Z, X, D = 1]$ denote the nonparametric imputation of the moment function g based on (Z, X). [Nonparametric imputation of g is equivalent to nonparametric imputation of Y^* if and only if g is linear in Y^*. If g is nonlinear in Y^*, then for efficient estimation, the moment function g should be nonparametrically imputed rather than the missing variables themselves.] It is shown in Lemma D.2 that

$$\rho_{\dim(g)\times 1} := \rho(\mathcal{A}, \theta^*, \pi, \mu) := \frac{Dg_{\text{obs}}}{\pi} - \mu\left[\frac{D}{\pi} - 1\right] \tag{4.4}$$

is the vector of residuals from projecting Dg_{obs}/π coordinatewise onto the tangent space of score functions for π. Moreover (see Appendix D.3), ρ satisfies the conditional-on-X moment restriction

$$\mathbb{E}[\rho \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}. \tag{4.5}$$

The residual vector ρ is nonparametrically estimable from the observed sample because π is nonparametrically estimable from the observed sample and, under MAR, μ is nonparametrically estimable from the validation sample (see Section 4.2). Therefore, estimation of θ^* can be based on (4.5).

In fact, ρ being free from the influence of estimating π suggests that (4.5) can also deliver an efficient estimator of θ^*. Indeed, as shown in (D.4), (D.5), (D.6) in the supplement, the Jacobian $\partial_{\theta^*}\mathbb{E}[\rho \mid X] \overset{P_X\text{-a.s.}}{=} J$, whereas $\partial_\pi \mathbb{E}[\rho \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ and $\partial_\mu \mathbb{E}[\rho \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times\dim(g)}$. Hence, by Ai and Chen (2003, Theorems 4.1 and 6.1), the efficiency bound for estimating θ^* in (4.5) is given by $(\mathbb{E}J'\Omega_\rho^{-1}J)^{-1}$, where $\Omega_\rho := \mathbb{E}[\rho\rho' \mid X] \overset{(4.5)}{=} \text{var}[\rho \mid X]$. Furthermore, as confirmed by Lemma 4.1, $(\mathbb{E}J'\Omega_\rho^{-1}J)^{-1}$ is also the semiparametric efficiency bound for estimating θ^* in (2.1). Therefore, efficient estimation of θ^* can be based on (4.5).

Lemma 4.1. Let Assumptions 3.1, 3.2, 3.3 hold. Then, under the regularity conditions specified in Assumption D.1 in the supplement, the semiparametric efficiency bound for estimating θ^* in (2.1) is given by $\text{l.b.}(\theta^*) := (\mathbb{E}J'\Omega_\rho^{-1}J)^{-1}$. The efficiency bound does not decrease if the propensity score function is parametrically specified up to a finite-dimensional parameter, or even if it is fully known.

The abbreviation "l.b." stands for "lower bound" because the semiparametric efficiency bound is the greatest lower bound for the asymptotic variance of any $n^{1/2}$-consistent regular estimator. If there is no missingness, that is, $Y^* \overset{\text{w.p.1}}{=} Y$, then $\rho = g$ and the bound in Lemma 4.1 becomes $(\mathbb{E}J'\Omega_g^{-1}J)^{-1}$ with $\Omega_g := \mathbb{E}[gg' \mid X]$, which is the well-known efficiency bound for estimating θ^* in the CMR model $\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$.

The efficiency bound in Lemma 4.1 can be obtained by applying Ai and Chen (2012, Theorem 2.1) to (4.1)&(4.2). For completeness, Appendix D.3 contains an alternative derivation. Since $\theta \mapsto g(Y^*, Z, X, \theta)$ is not required to be differentiable, the bound is valid for non-smooth moment functions, for example, quantile regression. The bound remains unchanged whether π is fully unknown, fully known, or known up to a finite-dimensional parameter, due to the propensity score function being ancillary to θ^* (Hahn, 1998, p. 319). This is expected, as π does not enter the moment condition (2.1) through which θ^* is defined. As noted in Chen, Hong, and Tarozzi (2008, p. 822) and Graham (2011, p. 439), ancillarity of π implies that to obtain an asymptotically efficient estimator of θ^*, the propensity score should be nonparametrically estimated, even if it is parametrically specified or fully known.

To measure the efficiency gain when all data in the observed sample—and not just those in the validation sample—are used to estimate θ^*, the efficiency bound in Lemma 4.1 is compared with $\text{l.b.}_{\text{VS}}(\theta^*)$, the efficiency bound for θ^* based on the IPW

CMR $\mathbb{E}\left[\frac{Dg_{\text{obs}}}{\pi} \mid X\right] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$. It is shown in Appendix D.3 that

$$\text{l.b.}(\theta^*) \leq_L \text{l.b.}_{\text{VS}}(\theta^*), \tag{4.6}$$

where the inequality $M_1 \leq_L M_2$ for symmetric matrices M_1, M_2 means that $M_1 - M_2$ is negative semidefinite (Löwner order).

The next result characterizes the necessary and sufficient conditions under which a semiparametrically efficient estimator of θ^*, based on the moment function ρ in (4.4) and utilizing all data in the observed sample, beats any estimator relying on the IPW moment function Dg_{obs}/π and the validation sample. Lemma 4.2, proved in Appendix D.3, is the key result of our paper and is central to understanding the conceptual difference between our work and the existing literature (cf. Remark 4.1). To facilitate the interpretation of Lemma 4.2, we introduce the following terminology: We define the nonparametric imputation of g to be "uninformative" if it is zero w.p.1, that is, $\mu \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1}$, in which case $\rho = Dg_{\text{obs}}/\pi$ and estimation using the observed sample yields no efficiency gains. In contrast, we say that the nonparametric imputation of g is "informative" if it is nonzero with positive probability ("w.p.p."), that is, $\mu \overset{\text{w.p.p.}}{\neq} 0_{\dim(g)\times 1}$, thereby leading to maximal efficiency gains in estimation. As shown in Lemma 4.2, for efficiency gains to be realized when θ^* is estimated using the observed sample, it is necessary and sufficient that the nonparametric imputation of g be informative, which occurs if and only if there are nonmissing endogenous variables included in or excluded from the CMR (2.1). Henceforth, keep in mind that $Z = \vec{\varnothing} \iff Z_{\text{in}} = \vec{\varnothing}$ & $Z_{\text{ex}} = \vec{\varnothing}$, whereas $Z \neq \vec{\varnothing} \iff Z_{\text{in}} \neq \vec{\varnothing}$ or $Z_{\text{ex}} \neq \vec{\varnothing}$.

Lemma 4.2. Inequality (4.6) is sharp, meaning $\text{l.b.}(\theta^*) = \text{l.b.}_{\text{VS}}(\theta^*)$ holds if and only if $Z = \vec{\varnothing} \iff \mu \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1}$. Consequently, efficiency gains in estimation, measured by the coordinatewise ratio $\frac{\text{l.b.}_{\text{VS}}(\theta^*)}{\text{l.b.}(\theta^*)} \overset{(4.6)}{>} 1$, occur if and only if $Z \neq \vec{\varnothing} \iff \mu \overset{\text{w.p.p.}}{\neq} 0_{\dim(g)\times 1}$.

Lemma 4.2 establishes that estimation of θ^* using the validation subsample alone is asymptotically efficient if and only if there are no nonmissing endogenous variables included in or excluded from (2.1), that is, $Z = \vec{\varnothing}$; and this is equivalent to the nonparametric imputation of g being uninformative, that is, $\mu \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ as imputation is based solely on the nonmissing exogenous variables (included and excluded). Uninformative imputation does not yield efficiency gains in estimation. For efficiency gains to be realized when θ^* is estimated using the observed sample, it is necessary and sufficient that $Z \neq \vec{\varnothing} \iff \mu \overset{\text{w.p.p.}}{\neq} 0_{\dim(g)\times 1}$. In other words, efficiency gains in estimation from the observed sample occur if and only if there are nonmissing endogenous variables included in or excluded from the "structural" moment function g in the CMR (2.1); and this happens if and only if the nonparametric imputation of g is informative, that is, $\mu \overset{\text{w.p.p.}}{\neq} 0_{\dim(g)\times 1}$ as it is based on the nonmissing endogenous variables (whether included or excluded) in addition to the nonmissing exogenous variables (whether included or excluded). It is informative imputation that leads to maximal efficiency gain in estimation from the observed sample. To attain maximal efficiency gains in estimation, imputation must be nonparametric, meaning that μ should be estimated nonparametrically from the validation sample. If a parametric model for μ is used for imputation, then efficient estimation is possible only if it is correctly specified (see Section 4.1).

Remark 4.1. The CMR model (2.1) differs conceptually from the models of HP, HP21 (Hristache and Patilea, 2021), and indeed the rest of the literature, in that these works never consider the possibility that nonmissing endogenous variables can generate efficiency gains in estimation. To see this, note that HP include in their propensity score auxiliary variables T that are excluded from their partially linear regression model with no endogenous regressors (HP, p. 736), without ever specifying whether T itself is exogenous or endogenous. When discussing the consequence of missing outcomes, HP (p. 740) state that "...there is no information loss if the observations for which the outcome is missing are deleted from the sample," which suggests that, in models with missing outcomes, there is no scope for efficiency gains from using the observed sample. But Lemma 4.2 shows that this claim does not hold: Even in HP's model (where $Z_{\text{in}} = \vec{\varnothing}$), efficiency gains arise when $Z_{\text{ex}} = T$, that is, when the auxiliary variables in their propensity score are endogenous. Unlike us, HP and HP21 do not consider endogenous covariates, and therefore cannot provide the necessary and sufficient conditions under which imputation is informative, that is, when the observed sample yields maximal efficiency gains. For example, when is imputing missing outcomes informative in linear regression models, the workhorse of the missing data literature?

As noted earlier, Lemma 4.2 offers valuable insights into when to impute missing endogenous variables, which may be particularly appealing to applied researchers. It says that efficiency gains in estimating θ^* using all nonmissing observations in the observed sample—rather than just those in the validation sample—arise if and only if the nonparametric imputation of g is informative, for which it is necessary and sufficient that $Z \neq \vec{\varnothing} \iff Z_{\text{in}} \neq \vec{\varnothing}$ or $Z_{\text{ex}} \neq \vec{\varnothing}$. The following example demonstrates that imputing missing outcomes in linear regression models with no nonmissing endogenous regressors is uninformative and, hence, does not lead to efficiency gains in estimation.

Example 4.1 (When should missing outcomes be imputed?). Consider the linear regression model $Y^* = \alpha_0^* + X_{\text{in}}'\beta_0^* + U$ with $\mathbb{E}[U \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0$, where the outcomes may be missing, there are no nonmissing endogenous variables included and excluded ($Z_{\text{in}} = \vec{\varnothing}$ and $Z_{\text{ex}} = \vec{\varnothing}$), and no excluded instruments ($X_{\text{ex}} = \vec{\varnothing}$). Here, $g := U = Y^* - \alpha_0^* - X_{\text{in}}'\beta_0^*$; hence, $\mu := \mu(X_{\text{in}}, \alpha_0^*, \beta_0^*) = \mathbb{E}[U \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0$ implying that nonparametric imputation of g in this model is uninformative. Consequently, the validation sample alone can be used to construct a semiparametrically efficient estimator of (α_0^*, β_0^*). Indeed, Lemma 4.1 shows that the efficiency bound for estimating (α_0^*, β_0^*) is given by $(\mathbb{E}\pi(X_{\text{in}})J'\Omega_g^{-1}J)^{-1}$, where the propensity score $\pi(X_{\text{in}}) := \mathbb{E}[D \mid X_{\text{in}}]$ depends only on

X_{in}, $J = -[\,1\ X'_{\text{in}}\,]$, and $\Omega_g = \mathbb{E}[gg' \mid X_{\text{in}}]$. By Lemma 4.1, this coincides with the efficiency bound using the validation sample alone; moreover, based on the moment condition $\mathbb{E}[D(Y - \alpha_0^* - X'_{\text{in}}\beta_0^*) \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0$, which holds only in the validation sample, the estimator proposed later in (4.12) attains the bound. In this model, imputing the missing Y^* using X_{in} (as no other nonmissing endogenous/exogenous variables are present) and employing the imputed values to estimate (α_0^*, β_0^*) does not lead to any efficiency gains. This is easily seen for the least-squares (LS) estimator, which is not semiparametrically efficient but serves to illustrate the point. Since $\mathbb{E}[Y^* \mid X_{\text{in}}, D = 1] \overset{\text{MAR}}{=} \mathbb{E}[Y^* \mid X_{\text{in}}] = \alpha_0^* + X'_{\text{in}}\beta_0^*$, the missing Y^* can be replaced by their imputed value $\hat{Y} := \hat{\alpha}_{0\text{VS}} + X'_{\text{in}}\hat{\beta}_{0\text{VS}}$, where $(\hat{\alpha}_{0\text{VS}}, \hat{\beta}_{0\text{VS}}) := \text{argmin}_{\alpha,\beta} \sum_{i=1}^n D_i(Y_i - \alpha - X'_{\text{in},i}\beta)^2$ is the estimator of (α_0^*, β_0^*) from the validation sample alone. It is shown in the supplement (Appendix D.3.1) that the imputation $\hat{Y}$ does not provide any information about missing outcomes beyond what is available from the regression model itself so that $(\hat{\alpha}_{0\text{VS}}, \hat{\beta}_{0\text{VS}}) = (\hat{\alpha}_{0\text{LS}}, \hat{\beta}_{0\text{LS}})$, the LS estimator obtained from the observed sample by replacing the missing outcomes with their imputed values. Therefore, imputing the missing outcomes in linear regression models that have no nonmissing endogenous regressors does not lead to efficiency gains.

But imputing the missing outcome when nonmissing endogenous regressors are present is informative and, hence, does lead to efficiency gains.

Example 4.2 (Example 4.1 contd.). We now allow for nonmissing included endogenous regressors ($Z_{\text{in}} \neq \vec{\varnothing}$) and excluded instruments ($X_{\text{ex}} \neq \vec{\varnothing}$); nonmissing excluded endogenous variables may or may not be present, that is, Z_{ex} may or may not be empty. The model now is $Y^* = \alpha^* + X'_{\text{in}}\beta^* + Z'_{\text{in}}\gamma^* + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] \overset{P_X\text{-a.s.}}{=} 0$. Here, $g := \varepsilon$; hence, $\mu := \mu(Z, X, \alpha^*, \beta^*, \gamma^*) = \mathbb{E}[\varepsilon \mid Z, X] \overset{\text{w.p.p.}}{\neq} 0$ implying that nonparametric imputation of g in this model is informative. Consequently, no estimator of $(\alpha^*, \beta^*, \gamma^*)$ using the validation sample alone is semiparametrically efficient. Indeed, the efficiency bound (cf. Example 4.3)—which is attained by the estimator in (4.12) with ρ defined in (4.4)—is strictly smaller than the efficiency bound for estimating $(\alpha^*, \beta^*, \gamma^*)$ from the validation sample alone because $\mu \overset{\text{w.p.p.}}{\neq} 0$. In this model, imputing the missing Y^* using (Z, X) and employing the imputed values to estimate $(\alpha^*, \beta^*, \gamma^*)$ does lead to efficiency gains. This is easily seen for the two-stage least-squares (2SLS) estimator, which is not semiparametrically efficient but illustrates the point nicely. Since $\mathbb{E}[Y^* \mid Z, X, D = 1] \overset{\text{MAR}}{=} \mathbb{E}[Y^* \mid Z, X] = \alpha^* + X'_{\text{in}}\beta^* + Z'_{\text{in}}\gamma^* + \mu$, the missing Y^* are imputed by $\hat{Y}_{\text{imp}} := \hat{\alpha}_{\text{VS}} + X'_{\text{in}}\hat{\beta}_{\text{VS}} + Z'_{\text{in}}\hat{\gamma}_{\text{VS}} + \hat{\mu}(Z, X)$, where $(\hat{\alpha}_{\text{VS}}, \hat{\beta}_{\text{VS}}, \hat{\gamma}_{\text{VS}}) := \text{argmin}_{\alpha,\beta,\gamma} \sum_{i=1}^n \frac{D_i}{\hat{\pi}_i}(Y_i - \alpha - X'_{\text{in},i}\beta - \hat{Z}'_{\text{in},i}\gamma)^2$ is the 2SLS estimator in the validation sample, $\hat{Z}_{\text{in}}$ is the predicted Z_{in} from the reduced form equations for Z_{in} in the observed sample, $\hat{\mu}(Z, X) := \hat{\mu}(Z, X, \hat{\alpha}_{\text{VS}}, \hat{\beta}_{\text{VS}}, \hat{\gamma}_{\text{VS}})$ is obtained from the validation sample by nonparametrically regressing $\hat{\varepsilon}_{\text{VS}} := \hat{\alpha}_{\text{VS}} - X'_{\text{in}}\hat{\beta}_{\text{VS}} - Z'_{\text{in}}\hat{\gamma}_{\text{VS}}$ on (Z, X), and $\hat{\pi}_i := \hat{\pi}(Z_i, X_i)$ is the estimated propensity score (Section 4.2). It is shown in the (supplement Appendix D.3.1) that $\hat{Y}_{\text{imp}}$ provides information about missing outcomes that is not available from the regression model itself so that $(\hat{\alpha}_{\text{VS}}, \hat{\beta}_{\text{VS}}, \hat{\gamma}_{\text{VS}}) \neq (\hat{\alpha}_{2\text{SLS}}, \hat{\beta}_{2\text{SLS}}, \hat{\gamma}_{2\text{SLS}})$, the 2SLS estimator in the observed sample. Therefore, imputing the missing outcome when nonmissing endogenous regressors are present leads to efficiency gains.

Example 4.3 (Example 2.1 contd.). In the IV regression model with missing outcomes, $g = Y^* - \alpha^* - X'_{\text{in}}\beta^* - Z'_{\text{in}}\gamma^*$. By Lemma 4.1, the efficiency bound for θ^* is $(\mathbb{E}J'J/\Omega_\rho)^{-1}$, where $J = -[\,1\ X'_{\text{in}}\ \mathbb{E}[Z'_{\text{in}}|X]\,]$, $\Omega_\rho \overset{(D.11)}{=} \mathbb{E}[\pi^{-1}\text{var}(Y^* \mid Z, X) \mid X] + \mathbb{E}[\mu^2 \mid X]$, and $\mu = \mathbb{E}[Y^* \mid Z, X] - \alpha^* - X'_{\text{in}}\beta^* - Z'_{\text{in}}\gamma^* \overset{\text{w.p.p.}}{\neq} 0$ because Z is endogenous with respect to g. Hence, imputation is informative and efficiency gains in estimation exist.

Example 4.4 (Example 2.2 contd.). In the IV regression model where the outcome $Z_{1,\text{in}}$ is nonmissing but endogenous explanatory variables Y^* are missing, $g := Z_{1,\text{in}} - \alpha^* - X'_{\text{in}}\beta^* - Z'_{2,\text{in}}\gamma^* - Y^{*\prime}\delta^*$. By Lemma 4.1, the efficiency bound for θ^* is $(\mathbb{E}J'J/\Omega_\rho)^{-1}$, where $J = -[1\ X'_{\text{in}}\mathbb{E}[Z'_{2,\text{in}} \mid X]\ \mathbb{E}[Y^{*\prime} \mid X]]$, $\Omega_\rho \overset{(D.11)}{=} \delta^{*\prime}\mathbb{E}[\pi^{-1}\text{var}(Y^* \mid Z, X) \mid X]\delta^* + \mathbb{E}[\mu^2 \mid X]$, and $\mu = Z_{1,\text{in}} - \alpha^* - X'_{\text{in}}\beta^* - Z'_{2,\text{in}}\gamma^* - \mathbb{E}[Y^{*\prime} \mid Z, X]\delta^* \overset{\text{w.p.p.}}{\neq} 0$ because Z is endogenous with respect to g. Hence, imputation is informative and efficiency gains in estimation exist.

Example 4.5 (Missing completely at random (MCAR)). Y^* is said to be MCAR if $D \perp\!\!\!\perp (Y^*, Z, X)$. MCAR, which implies MAR, is too strong to be of much empirical interest. Nonetheless, it is worth noting that the results under MCAR can be obtained as a special case of the results under MAR. To see this, note that MCAR is equivalent to MAR plus the condition that $D \perp\!\!\!\perp (Z, X)$. But $D \perp\!\!\!\perp (Z, X)$ if and only if the propensity score function $(Z, X) \mapsto \pi(Z, X)$ is constant, that is, there exists $\pi_{\text{MCAR}} \in (0, 1)$ such that $\pi(Z, X) \overset{P_{Z,X}\text{-a.s.}}{=} \pi_{\text{MCAR}}$. Therefore, results under MCAR follow from those under MAR by simply replacing $\pi(Z, X)$ in Lemma 4.1 and Lemma 4.2 by π_{MCAR}. Efficiency gains exist under MCAR in the presence of nonmissing endogenous variables because the nonparametric imputation of g is then informative, that is, μ is a nonzero function of (Z, X).

4.1. Double Robustness of Estimators Based on (4.5)

Before focusing on efficient estimation, we highlight the "double robustness" property of estimators of θ^* based on (4.5), which refers to θ^* being consistently estimable when either the selection model for D, or the model for imputing g, is correctly specified.

Since $Dg = Dg_{\text{obs}}$, we can write $\rho \overset{(4.4)}{=} \frac{Dg_{\text{obs}}}{\pi} - \mu\left[\frac{D}{\pi} - 1\right] = g + \left[\frac{D}{\pi} - 1\right][g - \mu]$. As π and μ are unknown functions, they have to be estimated nonparametrically. However, to

avoid employing nonparametric methods, applied researchers often use "working" approximations of π and μ, denoted by $\pi_{\text{work}} := \pi_{\text{work}}(Z,X)$ and $\mu_{\text{work}} := \mu_{\text{work}}(Z,X)$, that are easier to estimate than π and μ (see Remark D.5). If $\pi_{\text{work}} \neq \pi$ and $\mu_{\text{work}} \neq \mu$, then (see Remark D.6)

$$\begin{aligned}\rho(\pi_{\text{work}}, \mu_{\text{work}}) &:= \frac{Dg_{\text{obs}}}{\pi_{\text{work}}} - \mu_{\text{work}}\left[\frac{D}{\pi_{\text{work}}} - 1\right] \\ &= g + \left[\frac{D}{\pi_{\text{work}}} - 1\right][g - \mu_{\text{work}}] \end{aligned} \tag{4.7}$$

can be regarded as a noisy version of ρ with the additive term $\left[\frac{D}{\pi_{\text{work}}} - 1\right][g - \mu_{\text{work}}]$ capturing the error from simultaneously misspecifying π (the true propensity score function) and μ (the nonparametric imputation of g). As shown in Appendix D.3, under MAR,

$$\begin{aligned}\pi_{\text{work}} \overset{P_{Z,X}\text{-a.s.}}{=} \pi \quad \text{or} \quad \mu_{\text{work}} \overset{P_{Z,X}\text{-a.s.}}{=} \mu \implies \\ \mathbb{E}[\rho(\pi_{\text{work}}, \mu_{\text{work}}) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}. \end{aligned} \tag{4.8}$$

By (4.8), consistent estimation of θ^* can be based either on the CMR $\mathbb{E}[\rho(\pi, \mu_{\text{work}}) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ (when only the working model for D is correctly specified, that is, $\pi_{\text{work}} = \pi$ but $\mu_{\text{work}} \neq \mu$), or on the CMR $\mathbb{E}[\rho(\pi_{\text{work}}, \mu) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ (when only the working model for imputing g is correctly specified, that is, $\mu_{\text{work}} = \mu$ but $\pi_{\text{work}} \neq \pi$). However, neither leads to an efficient estimator of θ^* because $\rho(\pi, \mu_{\text{work}}) \neq \rho$ and $\rho(\pi_{\text{work}}, \mu) \neq \rho$. Since efficient estimation is possible only when $\pi_{\text{work}} = \pi$ and $\mu_{\text{work}} = \mu$, in which case $\rho(\pi_{\text{work}}, \mu_{\text{work}}) = \rho$, Section 4.2 focuses on constructing an efficient estimator.

As shown in Remark D.7, under MAR, $\text{var}\rho(\pi, \mu) \leq_L \text{var}\rho(\pi, \mu_{\text{work}})$, that is, $\text{var}\rho(\pi, \cdot)$ is minimized when $\mu_{\text{work}} = \mu$. Therefore, if the working model for D is correctly specified, then, in the spirit of RRZ (Sections 2.6 and 2.7), ρ is the "least noisy" version of $\text{var}\rho(\pi, \cdot)$, on which we can base efficient estimation of θ^*.

4.2. Efficient Estimation by Empirical Likelihood Smoothing

If π and μ are fully known, then the smoothed empirical likelihood (SEL) estimator of θ^* (Kitamura, Tripathi, and Ahn, 2004, henceforth, KTA) based on (4.5) is asymptotically efficient, that is, its asymptotic variance equals the semiparametric efficiency bound in Lemma 4.1, because $J \overset{(D.4)}{=} \partial_\theta \mathbb{E}[\rho(\mathcal{A}, \theta^*, \pi, \mu) \mid X]$ P_X-a.s. In fact, the asymptotic variance of the SEL estimator does not change if π and μ are replaced by their nonparametric estimators (cf. Lemma D.3). Therefore, we estimate θ^* using the SEL approach, which maximizes the empirical likelihood of the data in the observed sample subject to (4.5). Smoothing the empirical likelihood is required because (4.5) is a conditional restriction and the coordinates of X are assumed to be continuously distributed (cf. the discussion in Appendix A.1 in the supplement following Assumption A.1). Since $\mu = \mathbb{E}[g_{\text{obs}} \mid Z, X, D = 1] = \dfrac{\mathbb{E}[Dg_{\text{obs}} \mid Z, X]}{\pi(Z,X)} = \dfrac{\mathbb{E}[Dg_{\text{obs}} \mid Z, X]\text{pdf}_{Z,X}(Z,X)}{\mathbb{E}[D \mid Z, X]\text{pdf}_{Z,X}(Z,X)}$, we estimate π and μ with the kernel estimators

$$\begin{aligned}\hat{\pi}_c(z,x) &:= \frac{(nc_n^{\dim(Z)+\dim(X)})^{-1}\sum_{k=1}^n D_k H_c(Z_k - z, X_k - x)}{\hat{f}_{Z,X}(z,x)} \\ \hat{\mu}_d(z,x,\theta) &:= \frac{(nd_n^{\dim(Z)+\dim(X)})^{-1}\sum_{k=1}^n D_k g(Y_k, Z_{\text{in},k}, X_{\text{in},k}, \theta) H_d(Z_k - z, X_k - x)}{\hat{f}^{\text{VS}}_{Z,X}(z,x)}, \end{aligned}$$

where $H_c(\cdot) := H(\cdot/c_n)$ and $H_d(\cdot) := H(\cdot/d_n)$ is a kernel of sufficiently high order to deal with the estimation bias in $(\hat{\pi}_c, \hat{\mu}_d)$, the subscripts $c := (c_n)$ and $d := (d_n)$ denote the bandwidths used to estimate π and μ, and

$$\begin{aligned}\hat{f}_{Z,X}(z,x) &:= \frac{\sum_{k=1}^n H_c(Z_k - z, X_k - x)}{nc_n^{\dim(Z)+\dim(X)}} \\ \hat{f}^{\text{VS}}_{Z,X}(z,x) &:= \frac{\sum_{k=1}^n D_k H_d(Z_k - z, X_k - x)}{nd_n^{\dim(Z)+\dim(X)}} \end{aligned}$$

estimate the joint density of (Z,X) in the observed and validation samples. Note that $\hat{\pi}_c$ is based on the observed sample, whereas $\hat{\mu}_d$ is based only on the validation sample.

For the remainder of the paper, let $\rho_j(\theta) := \rho(\mathcal{A}_j, \theta, \pi(Z_j, X_j), \mu(Z_j, X_j, \theta))$ and $\hat{\rho}_j(\theta) := \rho(\mathcal{A}_j, \theta, \hat{\pi}_c(Z_j, X_j), \hat{\mu}_d(Z_j, X_j, \theta))$. To motivate the SEL approach, for $i, j = 1, \ldots, n$, let p_{ij} denote the conditional probability $\Pr(\mathcal{A} = \mathcal{A}_j \mid X = X_i)$ arising from a discrete distribution with support $(X_1, \ldots, X_n) \times (\mathcal{A}_1, \ldots, \mathcal{A}_n)$. Using these p_{ij} and kernel weights

$$w_{ij} := \frac{K_b(X_i - X_j)}{\sum_{k=1}^n K_b(X_i - X_k)} = \frac{(nb_n^{\dim(X)})^{-1} K_b(X_i - X_j)}{\hat{f}_X(X_i)},$$

where $K_b(\cdot) := K(\cdot/b_n)$ is a 2^{nd} order kernel, $b := (b_n)$ the bandwidth, and $\hat{f}_X(X_i) := \sum_{j=1}^n K_b(X_i - X_j)/nb_n^{\dim(X)}$, construct the smoothed loglikelihood $\sum_{i=1}^n \sum_{j=1}^n w_{ij} \log p_{ij}$. Then, given θ, solve the following optimization problem that finds the optimal discrete distribution to enforce the sample analog of (4.5), namely,

$$\begin{aligned}&\max_{\substack{p_{ij}\in(0,1)\\ i,j=1,\ldots,n}} \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log p_{ij} \\ &\text{s.t.} \quad \begin{cases} \sum_{j=1}^n p_{1j} = 1, \ldots, \sum_{j=1}^n p_{nj} = 1 \\ \sum_{j=1}^n \hat{\rho}_j(\theta) p_{1j} = 0_{\dim(g)\times 1}, \ \ldots, \ \sum_{j=1}^n \hat{\rho}_j(\theta) p_{nj} = 0_{\dim(g)\times 1}. \end{cases} \end{aligned} \tag{4.9}$$

The solution to (4.9), denoted by $(\hat{p}_{ij}(\theta))_{i,j=1,\ldots,n}$, is given by (see Appendix D.3)

$$\hat{p}_{ij}(\theta) \overset{(D.43)}{=} \frac{w_{ij}}{1 + \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)}, \quad i, j = 1, \ldots, n, \tag{4.10}$$

where $\hat{\lambda}_i(\theta)$, the Lagrange multipliers imposing the moment condition in (4.9), satisfy

$$\sum_{j=1}^{n} \frac{w_{ij}\hat{\rho}_j(\theta)}{1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)} = 0_{\dim(g)\times 1}, \quad i=1,\ldots,n. \tag{4.11}$$

As $\hat{p}_{ij}(\theta) > 0$ for small enough b_n (cf. Footnote 26 in Appendix D.3), the smoothed empirical loglikelihood of θ is the value function of (4.9) given by

$$\widehat{\mathrm{SEL}}(\theta) := \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log\hat{p}_{ij}(\theta) \overset{(D.10)}{=} \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(w_{ij}) - \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)),$$

where the $\hat{\lambda}_i(\theta)$ satisfy (4.11) (the "hat" in $(\widehat{\mathrm{SEL}}, \hat{\lambda}_i)$ emphasizes that it is based on $(\hat{\pi}, \hat{\mu})$). The estimator of θ^* is defined to be the maximizer of a trimmed version of $\widehat{\mathrm{SEL}}(\cdot)$, that is,

$$\hat{\theta} := \operatorname*{argmax}_{\theta\in\Theta} \widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta), \tag{4.12}$$

where—ignoring the term $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(w_{ij})$ as it does not depend on θ—the trimmed SEL objective function is given by

$$\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) := -\sum_{i=1}^{n} \hat{\mathbb{T}}_{1i} \sum_{j=1}^{n} \hat{\mathbb{T}}_{2j} w_{ij}\log(1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)), \tag{4.13}$$

and (abusing notation) the $\hat{\lambda}_i(\theta)$ in (4.13) satisfy the trimmed version of (4.11), namely,

$$\sum_{j=1}^{n} \frac{\hat{\mathbb{T}}_{2j} w_{ij}\hat{\rho}_j(\theta)}{1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)} = 0_{\dim(g)\times 1}, \quad i=1,\ldots,n. \tag{4.14}$$

The variables $\hat{\mathbb{T}}_{1i} := 1(\hat{f}_X(X_i) \geq b_n^{\tau_b})$ and $\hat{\mathbb{T}}_{2j} := 1(\hat{f}_{Z,X}(Z_j, X_j) \geq c_n^{\tau_c}, \hat{f}_{Z,X}^{\mathrm{VS}}(Z_j, X_j) \geq d_n^{\tau_d})$, where $\tau_b, \tau_c, \tau_d \in (0,1)$, are trimming indicators included to control the instability of the local empirical likelihood $\theta \mapsto \sum_{j=1}^{n} \hat{\mathbb{T}}_{2j} w_{ij}\log(1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta))$ caused by the denominators of $w_{ij}, \hat{\pi}_c, \hat{\mu}_d$ becoming too small in the tails. Since $(\hat{\mathbb{T}}_{1i}, \hat{\mathbb{T}}_{2j}) \xrightarrow{\mathrm{P}} (1)$, this trimming scheme ensures that $\hat{\theta}$ is efficient by guaranteeing that, asymptotically, no data is lost.

Remark 4.2.

(i) (SEL standard errors). The Hessian of the SEL objective function yields the SEL standard errors $\mathrm{se}_{\mathrm{SEL}}(\hat{\theta}^{(k)}) := \sqrt{[(-\nabla^2_{\theta\theta}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\hat{\theta}))^{-1}]_{kk}}$.

(ii) (Consistency and asymptotic normality). Under conditions that control the estimation error when (π, μ) is nonparametrically estimated by $(\hat{\pi}_c, \hat{\mu}_d)$, consistency and asymptotic normality of $\hat{\theta}$ can be shown by replicating KTA (*Theorems 3.1 and 3.2*). However, a rigorous proof of the consistency and asymptotic normality of $\hat{\theta}$ will only add length to our paper, without substantially increasing its contribution. Therefore, to minimize technical details, in *Appendix* D.3.3 we assume that $\hat{\theta}$ is consistent for θ^* and justify why, asymptotically, $n^{1/2}(\hat{\theta} - \theta^*)$ is distributed as a Gaussian random vector with mean zero and variance-covariance matrix equal to the efficiency bound in *Lemma 4.1*.

(iii) (Why SEL?). CMR models with unknown functions can also be efficiently estimated using the sieve minimum distance (SMD) approach of Ai and Chen (2003), which is suited for cases where these functions are not identifiable as conditional expectations or densities. Since the unknown functions (π, μ) in (4.5) are conditional expectations easily handled by kernel estimators, we employ the SEL approach instead. The two methods are closely related: Footnote 4 of Ai and Chen notes that kernel estimators can be used within SMD, and *Appendix* D.3.3 shows that the continuous-updating SMD objective function (Ai and Chen, *Eqn. 23*) is a large-sample quadratic approximation of the SEL objective function. Therefore, as elaborated in *Appendix* A.3, both the SEL and SMD approaches can be effectively employed for efficient estimation in our setting, though other methods can also be used. Although both yield asymptotically efficient estimators, SEL inference based on the likelihood-ratio (LR) statistic tends to be more accurate in small samples than SMD's Wald-based inference (see *Section 4.3* and *Appendix* A.3).

(iv) (Easing the computational burden of SEL). The SEL method can be computationally demanding due to its nonparametric smoothing. To facilitate its use, we developed the `R` package `smoothemplik` available on CRAN. All empirical and simulation results in this paper were obtained using `smoothemplik`. See *Appendix B.3* for details.

4.3. Inference

The SEL estimator $\hat{\theta}$ is a nonparametric maximum likelihood estimator of θ^* under the constraint $\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0$, enabling likelihood-ratio (LR)-based inference via the statistic $\mathrm{LR}(\theta) := 2[\widehat{\mathrm{SEL}}_{\mathbb{T}}(\hat{\theta}) - \widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta)]$. For hypotheses $\mathrm{H}_0 : R(\theta^*) = 0_{\dim(R)\times 1}$ with R a smooth vector function, rejection occurs for large $\mathrm{LR}(\hat{\theta}_R)$ with $\hat{\theta}_R := \operatorname*{argmax}_{\theta\in\Theta: R(\theta)=0_{\dim(R)\times 1}} \widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta)$. Under H_0, $\mathrm{LR}(\hat{\theta}_R) \overset{\mathrm{d}}{\approx} \chi^2_{\dim(R)}$, so the statistic is asymptotically pivotal. Unlike the Wald statistic, the LR statistic is internally studentized and does not require variance estimation. Inverting $\mathrm{LR}(\theta)$ yields asymptotically valid, transformation-invariant confidence regions that respect parameter bounds. In small samples, LR confidence regions better capture the shape of the sampling distribution—for example, they may be asymmetric or unbounded—whereas Wald regions (obtained by inverting the Wald statistic), being always symmetric and bounded, can yield unreliable inference. This difference is illustrated in Appendix C.3, where we describe the main findings of our simulation study. As noted in Appendix C.3.1 and C.3.2, for both simulation designs (58% missingness and 36% missingness), LR confidence intervals can be unbounded in one direction for small sample sizes (Figures C.5 and C.7). In contrast, for the same level of missingness and the same sample size, the Wald confidence intervals are always symmetric and bounded by construction, which may not accurately reflect the uncertainty in the sampling distribution.

5. Empirical Illustration

Angrist and Evans (1998), hereafter AE, employ U.S. census data to establish a causal link between family size and female labor supply. Their dataset can be downloaded from https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive, and the variables names in this section are what AE use in their code. AE find that having more than two children has a strong negative economic impact on the labor supply of working mothers. For example, labor market outcomes such as their labor income, their work status, the number of weeks worked, and the number of hours worked per week, are all negatively affected by the birth of a 3rd child. The number of children is potentially endogenous because fertility is likely to be simultaneously determined with the labor market outcomes. Therefore, to identify the causal effect of the number of children on a labor market outcome, AE use as excluded instruments $X_{\text{ex}} := (boys2, girls2)_{2\times 1}$, two dummy variables indicating whether both kids are boys or girls. These instruments are, at least intuitively, both valid and relevant: the former because the sex of a child is typically not influenced by the parents, and the latter because some parents prefer mixed-sex siblings, so having same-sex children increases the likelihood that the parents will conceive another child. [AE maintain that sex is randomly assigned (cf. their p. 451), and the validity of their instruments is not rejected in most of their specifications. We work with their model specification with labor income as the outcome variable where the instrument validity is not rejected (pvalue > 0.50).]

To study the robustness of AE's finding to missing outcomes—because non-response to income-related questions is common in surveys—we carry out the following counterfactual exercise using the AE dataset (the "full" sample in the terminology of Section 2): We artificially induce missingness in the outcome variable (labor income) ranging from 1% to 46%. The resulting observed sample with the missing outcomes is then used to estimate a model of female labor earnings using the semiparametrically efficient SEL estimator (4.12). The validation sample is used to obtain the IPW-SEL and IPW-GMM estimators. By comparing the performance of these estimators for different levels of missingness, we can demonstrate the extent—had AE encountered missing outcomes in their data—to which the aforementioned finding in their paper is robust in the presence of missing observations. As discussed in Section 5.3, the results from this counterfactual exercise illustrate that our methodology can be used to address the issue of missing endogenous variables in applied work in a fruitful manner.

5.1. Labor Earnings for Working Mothers

AE model the labor earnings for working mothers as $Y^* = \alpha^* + X_{\text{in}}'\beta^* + \gamma^* Z_{\text{in}} + U$, where the outcome $Y^* := incomem$ (annual labor income of mother in thousands of 1995 dollars) is potentially missing, the vector of exogenous explanatory variables $X_{\text{in}} := (agem1, agefstm, boy1st)_{3\times 1}$ contains the current age of mother, the age of mother at first birth, and the sex of the first child, and $Z_{\text{in}} := morekids$ is an endogenous dummy variable indicating that a mother has three or more kids. As in AE, we treat (*agem1*, *agefstm*) as continuously distributed and assume that $\mathbb{E}[U \mid X] \overset{P_X\text{-a.s.}}{=} 0$ with $X := (agem1, agefstm, boy1st, boys2, girls2)_{5\times 1}$. The AE moment function $g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) := Y^* - \alpha^* - X_{\text{in}}'\beta^* - \gamma^* Z_{\text{in}}$ with $\mathbb{E}[g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0$ and $\theta^* := (\alpha^*, \beta^*, \gamma^*)_{5\times 1}$. There are no nonmissing excluded endogenous variables, that is, $Z_{\text{ex}} = \vec{\varnothing}$, so that $Z = Z_{\text{in}}$ throughout the empirical illustration.

AE consider six specifications, each corresponding to a different outcome variable. For each outcome, they report 12 estimators, namely, (LS, just-identified IV, over-identified 2SLS) $\times$ (1980 or 1990 data) $\times$ (all women or married women). For simplicity, we restrict our analysis to the sub-sample of white married females in 1980 (approx. 82% of all surveyed females are white in the 1980 Public Use Micro Sample), which yields a sample of size $n = 227{,}146$ and explains the minor discrepancy between the GMM estimates in Table 1 and the 2SLS estimates of AE (Table 7). The GMM estimates in Table 1 correspond to the 2SLS estimates in AE (Table 7, p. 465, row "Labor income," column 6), although, for convenience, we divide the labor income by 1000. Table 1 verifies that SEL estimates replicate the 2SLS estimates in AE when there are no missing outcomes. In Table 1, the model $\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0$ is estimated using SEL, and the over-identified UCMR model $\mathbb{E}[\tilde{X}g] = 0_{6\times 1}$, where $\tilde{X} := (1, agem1, agefstm, boy1st, boys2, girls2)_{6\times 1}$, is estimated using iterated GMM to remove the dependence of the 2-step optimal GMM estimator on the initial estimator/weight matrix. As noted in Table 1, the hypothesis $\mathbb{E}[\tilde{X}g] = 0_{6\times 1}$ is not rejected by the J-test (pvalue $= 0.613$).

The GMM and SEL point estimates in Table 1 are similar, with smaller standard errors for the SEL estimates as expected. Estimates for (*agem1*, *agefstm*, *boy1st*) are virtually identical. Estimates for *morekids* may appear a bit different numerically ($\hat{\gamma}_{\text{GMM}} = -1.499, \hat{\gamma} = -2.046$), but they are statistically indistinguishable at levels of significance $\leq 1\%$. This is evident from a Hausman test of the null hypothesis that $\hat{\gamma}_{\text{GMM}}$ and $\hat{\gamma}$ estimate the same parameter (cf. Remark B.1). Hence, as their point estimates are similar, it is their standard errors that determine which estimator delivers a statistically significant estimate of the effect of having a 3rd child on labor income. The hypothesis that *morekids* is irrelevant for explaining labor income—testing $\gamma^* = 0$ against the alternative that $\gamma^* \neq 0$—is rejected at all reasonable significance levels by the SEL estimate (pvalue $= 0.00008$), but not by the GMM estimate (pvalue $= 0.009$). Thus, even with >200,000 observations, the GMM point estimate—although economically relevant—is not statistically different enough from zero to convincingly reject the irrelevance of *morekids* at the 1% level of significance. In contrast, the SEL point estimate remains both economically relevant and statistically significant at all reasonable significance levels. Therefore, as this problem with the GMM estimate only gets exacerbated when there is missingness in labor income, we carry out the counterfactual exercise outlined earlier.

5.2. Missingness in Labor Income

As described in Appendix B.1, we artificially induce missingness in labor income and create datasets—drawn randomly from the original AE dataset—containing missing outcomes. To ensure

Table 1. Estimated female labor earnings model with no missingness in labor income.

Variable	$\hat{\theta}_{\text{GMM}}$	$\text{se}_{\text{GMM}}(\hat{\theta}_{\text{GMM}})$	p value$(\hat{\theta}_{\text{GMM}})$	$\hat{\theta}$	$\text{se}_{\text{SEL}}(\hat{\theta})$	p value$(\hat{\theta})$
const.	−1.067	0.282	.00016	−0.899	0.258	.0005
agem1	{-1}0.459	0.018	3.3×10^{-143}	{-1}0.484	0.017	5.3×10^{-185}
agefstm	−0.313	0.026	2.0×10^{-34}	−0.347	0.023	1.8×10^{-50}
boy1st	{-1}0.040	0.041	.325	{-1}0.040	0.040	.31
morekids	−1.499	0.574	.009	−2.046	0.520	.00008
n					227,146	
Weak instruments *F*					651.7 (pvalue $= 7.8 \times 10^{-284}$)	
Endogeneity of *morekids F*					6.91 (p value = .009)	
Over-identification *J*					0.256 (p value = .613)	

NOTE: $\hat{\theta}$ is the SEL estimator of θ^* in the model $\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0$ with bandwidth $b_n = 1.2$ (cf. Appendix B.4). $\hat{\theta}_{\text{GMM}}$ is the iterated GMM estimator of θ^* in the over-identified model $\mathbb{E}[\tilde{X}g] = 0_{6\times 1}$, and se_{GMM} are the GMM standard errors.

Table 2. Fraction of 1000 experiments (as a function of missingness in labor income) for which the *t*-test fails to reject the hypothesis that *morekids* is irrelevant.

Missingness in	size = 10%			size = 5%			size = 1%		
labor income	$\hat{\gamma}_{\text{GMM,IPW}}$	$\hat{\gamma}_{\text{SEL,IPW}}$	$\hat{\gamma}$	$\hat{\gamma}_{\text{GMM,IPW}}$	$\hat{\gamma}_{\text{SEL,IPW}}$	$\hat{\gamma}$	$\hat{\gamma}_{\text{GMM,IPW}}$	$\hat{\gamma}_{\text{SEL,IPW}}$	$\hat{\gamma}$
1%	0.00	0.00	0.00	0.00	0.00	0.00	0.43	0.00	0.00
7%	0.00	0.00	0.00	0.01	0.00	0.00	0.61	0.00	0.00
14%	0.03	0.00	0.00	0.12	0.00	0.00	0.67	0.01	0.01
20%	0.07	0.00	0.00	0.23	0.01	0.00	0.71	0.10	0.06
27%	0.18	0.01	0.00	0.34	0.05	0.02	0.74	0.27	0.20
33%	0.29	0.11	0.04	0.46	0.22	0.09	0.81	0.48	0.38
39%	0.43	0.35	0.13	0.61	0.49	0.24	0.88	0.76	0.57
46%	0.65	0.72	0.28	0.77	0.81	0.43	0.93	0.93	0.76

NOTE: The *t*-test of $\gamma^* = 0$ vs. $\gamma^* \neq 0$ compares $|\frac{\hat{\gamma}_{\text{GMM,IPW}}}{\text{se}_{\text{GMM}}(\hat{\gamma}_{\text{GMM,IPW}})}|$, $|\frac{\hat{\gamma}_{\text{SEL,IPW}}}{\text{se}_{\text{SEL}}(\hat{\gamma}_{\text{SEL,IPW}})}|$, and $|\frac{\hat{\gamma}}{\text{se}_{\text{SEL}}(\hat{\gamma})}|$ with the two-sided critical values from the normal distribution.

that our analysis is not influenced by a specific level of missingness in the labor income or a specific draw from the AE dataset, we consider 22 levels of missingness ranging from 1% to 46% and for each level of missingness we independently draw 1000 datasets with missing outcomes, and for each dataset we compute the following estimators:

- We estimate θ^* in the over-identified model $\mathbb{E}[\tilde{X}\frac{Dg_{\text{obs}}}{\pi}] = 0_{6\times 1}$ using iterated GMM, and undersmoothed bandwidth $\hat{c}_n^{\text{CV}}/3$ to estimate π, where $\hat{c}_n^{\text{CV}}$ is the cross-validated bandwidth for estimating π. We call this estimator $\hat{\theta}_{\text{GMM,IPW}}$ and compute $\text{se}_{\text{GMM}}(\hat{\theta}_{\text{GMM,IPW}})$, the GMM standard error. The 1/3 factor and choice of smoothing bandwidths b_n, c_n, d_n are discussed in Appendix B.4.
- We estimate θ^* in the model $\mathbb{E}\left[\frac{Dg_{\text{obs}}}{\pi} \mid X\right] \overset{P_X\text{-a.s.}}{=} 0$ using SEL and bandwidths $b_n = 1.2$ and $\hat{c}_n^{\text{CV}}/3$. We call this estimator $\hat{\theta}_{\text{SEL,IPW}}$ and compute $\text{se}_{\text{SEL}}(\hat{\theta}_{\text{SEL,IPW}})$.
- We estimate θ^* in (4.5) using the SEL estimator $\hat{\theta}$ defined in (4.12) with bandwidths $b_n = 1.2$, $\hat{c}_n^{\text{CV}}/3$, and $\hat{d}_n^{\text{CV}}/3$. We compute $\text{se}_{\text{SEL}}(\hat{\theta})$, the SEL standard error of $\hat{\theta}$.

For each level of missingness ranging from 1% to 46%, we therefore have 1000 i.i.d. copies of $(\hat{\theta}_{\text{GMM,IPW}}, \text{se}_{\text{GMM}}(\hat{\theta}_{\text{GMM,IPW}}))$, $(\hat{\theta}_{\text{SEL,IPW}}, \text{se}_{\text{SEL}}(\hat{\theta}_{\text{SEL,IPW}}))$, and $(\hat{\theta}, \text{se}_{\text{SEL}}(\hat{\theta}))$. It is useful to interpret them as 1000 independent researchers each possessing these three estimators and their standard errors for each of the 22 levels of missingness. The discussion of the results in Figure 2 and Table 2 is based on this interpretation. Additional results for higher levels of missingness are in the supplement (see Appendix B.2).

5.3. Results and Discussion

Based on the 1000 experiments, Figure 1 plots the median standard errors of the estimated slope coefficients $(\beta^*, \gamma^*)_{4\times 1}$ as a function of the missingness. Since the size of the validation sample decreases as missingness increases, the standard errors of $(\hat{\beta}_{\text{SEL,IPW}}, \hat{\gamma}_{\text{SEL,IPW}})$ and $(\hat{\beta}_{\text{GMM,IPW}}, \hat{\gamma}_{\text{GMM,IPW}})$ are strictly increasing in the level of missingness. Moreover, the standard errors of $(\hat{\beta}_{\text{SEL,IPW}}, \hat{\gamma}_{\text{SEL,IPW}})$ are systematically smaller than the standard errors of $(\hat{\beta}_{\text{GMM,IPW}}, \hat{\gamma}_{\text{GMM,IPW}})$ because $\hat{\theta}_{\text{SEL,IPW}}$ is more efficient than the GMM estimator (SEL estimates a CMR model whereas GMM estimates an UCMR model). In turn, the standard errors of $(\hat{\beta}, \hat{\gamma})$ are smaller than the standard errors of $(\hat{\beta}_{\text{SEL,IPW}}, \hat{\gamma}_{\text{SEL,IPW}})$ because we have already shown that the SEL estimator $\hat{\theta}$ using all nonmissing observations—and not just those in the validation sample—is semiparametrically efficient. For missingness rates under 25%, the two are almost identical because efficiency gains at low missingness rates are offset by the noise from estimating the nonparametric imputation μ. Efficiency gains emerge as missingness increases, and we find that for missingness = (30%, 37%, 42%, 46%) the standard error of $\hat{\gamma}_{\text{SEL,IPW}}$ is approximately (1%, 5%, 11%, 28%) larger than that of $\hat{\gamma}$ (these numbers, difficult to read from Figure 1, are from the data used to create the figure). The one exception where the efficient estimator does not dominate its validation sample counterpart is the effect of *boy1st*, the gender of the first-born. This can be explained by the fact that *boy1st* is an uninformative predictor: Its *t*-statistic is ≈ 1 even in this large a sample (Table 1), and the induced missingness does not depend on *boy1st*.

To visually compare the relative performance of $\hat{\gamma}_{\text{GMM,IPW}}$, $\hat{\gamma}_{\text{SEL,IPW}}$, and $\hat{\gamma}$ as a function of the missingness in labor income, Figure 2 displays some features of the sampling distributions of

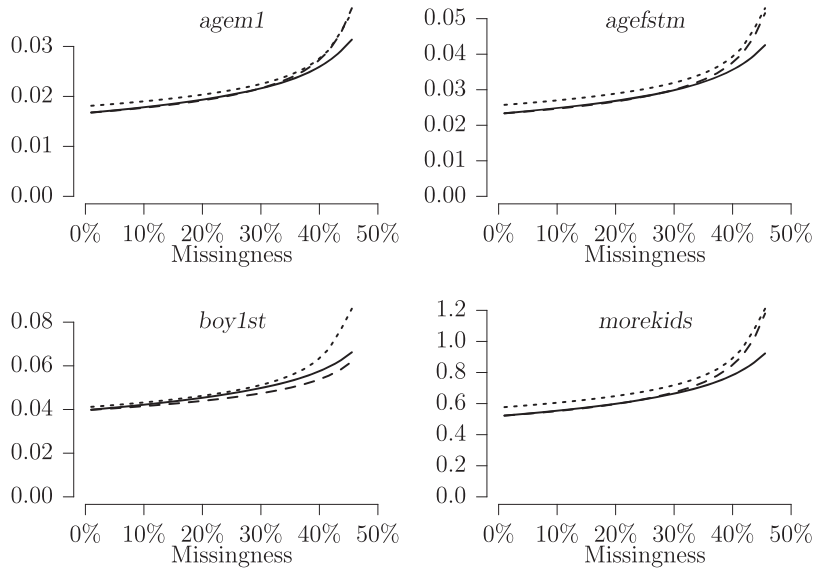

Figure 1. Based on 1000 experiments, the standard errors of the estimated slope coefficients $(\beta^*, \gamma^*)_{4\times1}$ as a function of missingness in labor income.For each level of missingness: Dotted line is median of $\text{se}_{\text{GMM}}(\hat{\theta}^{(1)}_{\text{GMM,IPW}}), \ldots, \text{se}_{\text{GMM}}(\hat{\theta}^{(1000)}_{\text{GMM,IPW}})$. Dashed line is median of $\text{se}_{\text{SEL}}(\hat{\theta}^{(1)}_{\text{SEL,IPW}}), \ldots, \text{se}_{\text{SEL}}(\hat{\theta}^{(1000)}_{\text{SEL,IPW}})$. Solid line is median of $\text{se}_{\text{SEL}}(\hat{\theta}^{(1)}), \ldots, \text{se}_{\text{SEL}}(\hat{\theta}^{(1000)})$.

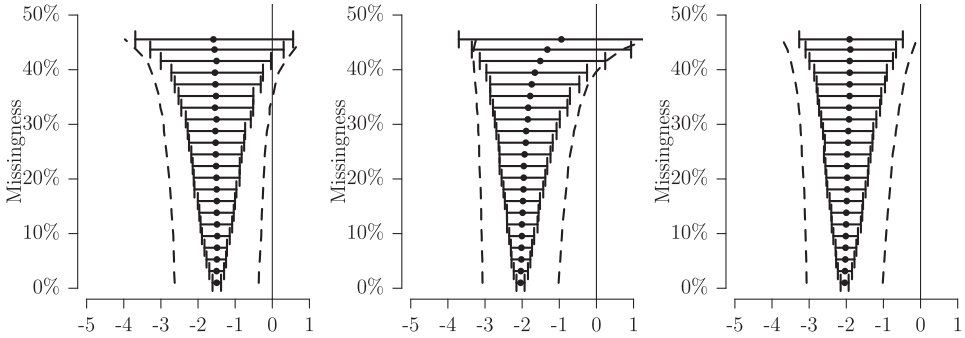

Figure 2. Based on 1000 experiments, three quantiles of $\gamma \in \{\hat{\gamma}_{\text{GMM,IPW}}, \hat{\gamma}_{\text{SEL,IPW}}, \hat{\gamma}\}$, and the medians of $\gamma \pm 1.96\text{se}(\gamma)$, as a function of missingness in labor income.For each level of missingness: Dots are medians of the γ. Solid whiskers are .025 and .975 quantiles of the γ. Left dashed line is median of $\gamma - 1.96\text{se}(\gamma)$. Right dashed line is median of $\gamma + 1.96\text{se}(\gamma)$.

these estimators across the 1000 researchers. For each estimator, the horizontal solid bars (the dot and the whiskers) indicate the median and the 0.025^{th} and 0.975^{th} quantiles of the point estimates across the researchers for different levels of missingness. The dashed lines are the medians, across all researchers, of the left- and right-endpoints of the 95% confidence intervals (CIs) for the true effect γ^*. As the missingness increases, the distribution of $\hat{\gamma}_{\text{GMM,IPW}}$ becomes more dispersed (the length of the whiskers around the median increases), and the left- and right-endpoints of the 95% CIs move steadily away from the point estimates. Compared to the GMM estimator, $\hat{\gamma}_{\text{SEL,IPW}}$ has smaller variance at 1%–20% levels of missingness; however, at high missingness levels, its dispersion increases, and a drift appears. In marked contrast, the semiparametrically efficient estimator $\hat{\gamma}$ yields 95% CIs of shorter length (the gap between the dashed lines) for high levels of missingness, lower variance of the estimator itself (the whisker width), and the median right-endpoint of the 95% CIs (the right dashed line) is always less than zero, indicating that the semiparametrically efficient estimator rejects the irrelevance of *morekids* at 5% significance in at least half the experiments for each level of missingness.

Table 2 complements Figure 2 by providing some additional information. It reports the fraction of 1000 experiments, expressed as a function of the missingness in labor income, for which the *t*-test fails to reject the hypothesis that *morekids* is irrelevant at the 10%, 5%, and 1% significance levels. The results in Table 2 can be interpreted as meaning that an overwhelming majority of the 1000 independent researchers using the GMM estimator would likely conclude that—irrespective of the extent of missingness in labor income—there is no statistically significant negative relationship between having more than two kids and labor income at the 1% significance level. Starting with just 20% missingness, we also see that almost 7% (resp. 23%) of the researchers using the GMM estimator would fail to reject the irrelevance of *morekids* at the 10% (resp. 5%) significance level. In contrast to the GMM estimator, with 20% missingness, all 1000 independent researchers using the SEL estimator $\hat{\gamma}$ with its smaller standard errors reject the irrelevance of *morekids* at the 10% and 5% significance levels, and only 6% of the researchers fail to reject that *morekids* is irrelevant at the 1% significance level. As the missingness in labor income increases, so does the failure to reject the irrelevance of *morekids*.

With 46% missingness, almost (65%, 77%, 93%) of independent researchers using $\hat{\gamma}_{\text{GMM,IPW}}$ would find *morekids* irrelevant at the (10%, 5%, 1%) significance levels. In contrast, for the same missingness, only (28%, 43%, 76%) of independent researchers using the efficient estimator $\hat{\gamma}$ are likely to conclude irrelevance of *morekids* at the (10%, 5%, 1%) levels of significance.

In summary, the reduction in the earnings of working mothers due to having a 3rd child, that is, the negative causal effect of the endogenous explanatory variable *morekids* on labor income, can have manifold economic and social implications. However, our analysis of the AE dataset shows that if there is even medium missingness in the labor income, then having more than 200,000 observations may not be enough for the inverse propensity score weighted GMM estimator to deliver statistically significant estimates of this effect. In contrast, for the same levels of missingness, the semiparametrically efficient SEL estimator utilizes information from the nonmissing endogenous variable *morekids* to produce statistically significant point estimates of its effect on labor income that are comparable in sign and magnitude to the GMM estimates. The choice of the smoothing bandwidths (b_n, c_n, d_n) seems to have little impact on the reliability of SEL-based inference (cf. Appendix B.4). Hence, despite the lack of theory regarding how to choose the smoothing bandwidths for the SEL approach, practitioners can use reasonably small bandwidths to smooth the empirical likelihood and nonparametrically estimate π and μ. Replication codes are available at https://github.com/Fifis/msely.

6. Conclusion

We believe the literature has overlooked the possibility that nonmissing endogenous variables—whether included in or excluded from CMR models—can lead to informative imputation and deliver efficiency gains in estimation when some endogenous variables (outcomes and/or covariates) are missing. Our findings are therefore highly relevant for applied researchers confronting missing endogenous variables in CMR models, and we conclude by offering the following practical recommendations:

(i) When specifying CMR models, researchers must distinguish not only between the endogenous and exogenous variables, but also whether the endogenous and exogenous variables are included in or excluded from the CMR model.
(ii) Auxiliary variables that are in the propensity score but are excluded from the CMR model must also be classified as endogenous or exogenous.
(iii) Efficiency gains in estimation from the observed sample occur if and only if there exist nonmissing endogenous variables (outcomes and/or covariates) that are included in or excluded from the CMR model.
(iv) Imputation should be based on all nonmissing variables (endogenous or exogenous) that are included in or excluded from the CMR model; that is, in our notation, imputation should be based on $(X_{\text{in}}, X_{\text{ex}}, Z_{\text{in}}, Z_{\text{ex}})$. However, nonparametric imputation is informative, that is, yields maximal efficiency gains in estimation, if and only if there exist nonmissing endogenous variables that are included in or excluded from the CMR model. That is, if and only if $Z_{\text{in}} \neq \vec{\varnothing}$ or $Z_{\text{ex}} \neq \vec{\varnothing}$. In particular, imputing missing outcomes in linear regression models is informative if and only if there exist nonmissing endogenous covariates that are included in or excluded from the regression.
(v) To achieve maximal efficiency gains in estimation from the observed sample, imputation must be nonparametric. If a parametric model is used for imputation, then efficient estimation is possible only if the imputation model is correctly specified.

Acknowledgments

We thank the editor Ivan Canay, an associate editor, and two anonymous referees for comments that greatly improved the paper. We are also grateful to Xiaohong Chen, Paul Devereux, Patrick Gagliardini, Jinyong Hahn, Valentin Patilea, and seminar participants at the University of Luxembourg, EcoSta 2023 (Waseda University), the 2023 Econometric Society Summer Meeting (Barcelona), the 2024 Asia Meeting (Ho Chi Minh City), and the 2025 World Congress (Seoul) for helpful suggestions. Andreï V. Kostyrka acknowledges financial support from FNR-Luxembourg through a PRIDE grant for the Migration and Labor (MINLAB) doctoral training unit. Simulation experiments were conducted using the University of Luxembourg HPC facilities.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Antonio Cosma http://orcid.org/0000-0002-2455-9760
Andreï Victorovitch Kostyrka http://orcid.org/0009-0001-8524-1182
Gautam Tripathi http://orcid.org/0000-0002-6757-0652

References

Abrevaya, J., and Donald, S. G. (2017), "A GMM Approach for Dealing with Missing Data on Regressors," *The Review of Economics and Statistics*, 99, 657–662. DOI: 10.1162/rest$_{a0}$0645. [1]

Ai, C., and Chen, X. (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843. DOI: 10.1111/1468-0262.00470. [4,8]

——— (2012), "The Semiparametric Efficiency Bound for Models of Sequential Moment Restrictions Containing Unknown Functions," *Journal of Econometrics*, 170, 442–457. DOI: 10.1016/j.jeconom.2012.05.015. [4]

Angrist, J. D., and Evans, W. N. (1998), "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 88, 450–477. [2,9]

Balestra, S., and Backes-Gellner, U. (2017), "Heterogeneous Returns to Education over the Wage Distribution: Who Profits the Most?" *Labour Economics*, 44, 89–105. DOI: 10.1016/j.labeco.2017.01.001. [3]

Bennedsen, M., Nielsen, K. M., Perez-Gonzalez, F., and Wolfenzon, D. (2007), "Inside the Family Firm: The Role of Families in Succession Decisions and Performance," *The Quarterly Journal of Economics*, 122, 647–691. DOI: 10.1162/qjec.122.2.647. [3]

Chen, X., Hong, H., and Tarozzi, A. (2008), "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics*, 36, 343–366. [2,4]

Graham, B. S. (2011), "Efficiency Bounds for Missing Data Models with Semiparametric Restrictions," *Econometrica*, 79, 437–452. [1,2,4]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. DOI: 10.2307/2998560. [4]

Hristache, M., and Patilea, V. (2017), "Conditional Moment Models with Data Missing at Random," *Biometrika*, 104, 735–742. DOI: 10.1093/biomet/asx025. [1]

——— (2021), "Equivalent Models for Observables under the Assumption of Missing at Random," *Econometrics and Statistics*, 20, 153–165. DOI: 10.1016/j.ecosta.2020.03.002. [5]

Kitamura, Y., Tripathi, G., and Ahn, H. (2004), "Empirical Likelihood Based Inference in Conditional Moment Restriction Models," *Econometrica*, 72, 1667–1714. DOI: 10.1111/j.1468-0262.2004.00550.x. [7]

Kostyrka, A. V. (2025), "smoothemplik: Smoothed empirical likelihood for efficient estimation and specification testing," `R` package version 0.0.17. Available at https://cran.r-project.org/package=smoothemplik. [2]

McDonough, I. K., and Millimet, D. L. (2017), "Missing Data, Imputation, and Endogeneity," *Journal of Econometrics*, 199, 141–155. DOI: 10.1016/j.jeconom.2017.05.006. [3]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. DOI: 10.1080/01621459.1994.10476818. [1,2]

Stephens, M., Jr., and Unayama, T. (2019), "Estimating the Impacts of Program Benefits: Using Instrumental Variables with Underreported and Imputed Data," *The Review of Economics and Statistics*, 101, 468–475. DOI: 10.1162/rest$_{a0}$0769. [3]