NBER WORKING PAPER SERIES

EFFICIENCY BOUNDS FOR MISSING DATA MODELS WITH SEMIPARAMETRIC RESTRICTIONS

Bryan S. Graham

Working Paper 14376
http://www.nber.org/papers/w14376

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2008

I would like to thank Gary Chamberlain, Jinyong Hahn, Guido Imbens, Michael Jansson and Whitney Newey for comments on earlier draft. Helpful discussions with Oliver Linton, Cristine Pinto, Jim Powell, Geert Ridder as well as participants in the Berkeley Econometrics Reading Group and Seminars are gratefully acknowledged. This revision has benefited from Tom Rothenberg's skepticism, discussions with Michael Jansson, Justin McCrary, Jim Powell, the comments of a co-editor and three especially meticulous/generous anonymous referees. All the usual disclaimers apply. This is a heavily revised version of material which previously circulated under the titles "A note on semiparametric efficiency in moment condition models with missing data", "GMM 'equivalence' for semiparametric missing data models", and "Efficient estimation of missing data models using moment conditions and semiparametric restrictions". The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Efficiency bounds for missing data models with semiparametric restrictions
Bryan S. Graham
NBER Working Paper No. 14376
October 2008, Revised October 2010
JEL No. C1,C14,C21

ABSTRACT

This paper shows that the semiparametric efficiency bound for a parameter identified by an unconditional moment restriction with data missing at random (MAR) coincides with that of a particular augmented moment condition problem. The augmented system consists of the inverse probability weighted (IPW) original moment restriction and an additional conditional moment restriction which exhausts all other implications of the MAR assumption. The paper also investigates the value of additional semiparametric restrictions on the conditional expectation function (CEF) of the original moment function given always-observed covariates. In the program evaluation context, for example, such restrictions are implied by semiparametric models for the potential outcome CEFs given baseline covariates. The efficiency bound associated with this model is shown to also coincide with that of a particular moment condition problem. Some implications of these results for estimation are briefly discussed.

Bryan S. Graham
New York University
19 W 4th Street, 6FL
New York, NY 10012
and NBER
bsg1@nyu.edu

An online appendix is available at:
http://www.nber.org/data-appendix/w14376

1 Introduction

Let $Z = (Y_1', X')'$ be vector of modelling variables, $\{Z_i\}_{i=1}^{\infty}$ be an independent and identically distributed random sequence drawn from the unknown distribution F_0, β a $K \times 1$ unknown parameter vector and $\psi(Z, \beta)$ a known vector-valued function of the same dimension.[2] The only prior restriction on F_0 is that for some $\beta_0 \in \mathcal{B} \subset \mathbb{R}^K$

$$\mathbb{E}\left[\psi\left(Z, \beta_0\right)\right] = 0. \tag{1}$$

Chamberlain (1987) showed that the maximal asymptotic precision with which β_0 can be estimated under (1) (subject to identification and regularity conditions) is given by $\mathcal{I}_{\mathrm{f}}(\beta_0) = \Gamma_0' \Omega_0^{-1} \Gamma_0$, with $\Gamma_0 = \mathbb{E}\left[\partial\psi\left(Z, \beta_0\right) / \partial\beta'\right]$ and $\Omega_0 = \mathbb{V}\left(\psi\left(Z, \beta_0\right)\right)$.[3]

Now consider the case where a random sequence from F_0 is unavailable. Instead only a selected sequence of samples is available. Let D be a binary selection indicator. When $D = 1$ we observe Y_1 and X, when $D = 0$ we observe only X.[4] This paper considers estimation of β_0 under restriction (1) and the following additional assumptions.

Assumption 1.1 (Random Sampling) *$\{Z_i, D_i\}_{i=1}^{\infty}$ is an independent and identically distributed random sequence from F_0.*

Assumption 1.2 (Observed Data) *For each unit we observe D, X and $Y = DY_1$.*

Assumption 1.3 (Conditional Independence) $Y_1 \perp D | X$.

Assumption 1.4 (Overlap) *Let $p_0(x) = \Pr(D = 1 | X = x)$, then $0 < \kappa \leq p_0(x) \leq 1$ for all $x \in \mathcal{X} \subset \mathbb{R}^{\dim(x)}$.*

Restriction (1) and Assumptions 1.1 to 1.4 constitute a semiparametric model for the data. Henceforth I refer to this model as the semiparametric missing data model or the missing at random (MAR) setup. Robins, Rotnitzky and Zhao (1994, Proposition 2.3, p. 850) derived the efficient influence function for this problem and proposed a locally efficient augmented inverse probability weighting (AIPW) estimator (cf., Scharfstein, Rotnitzky and Robins, 1999; Bang and Robins, 2005; Tsiatis, 2006). Cheng (1994), Hahn (1998), Hirano, Imbens and Ridder (2003), Imbens, Newey and Ridder (2005), and Chen, Hong and Tarozzi (2008) develop globally efficient estimators.

The 'MAR setup' has been applied to a number of important econometric and statistical problems, including program evaluation as surveyed by Imbens (2004), non-classical measurement error

[2] Extending what follows to the overidentified case is straightforward.

[3] Throughout upper case letters denote random variables, lower case letters specific realizations of them, and calligraphic letters their support. I use the notation $\mathbb{E}\left[A|c\right] = \mathbb{E}\left[A|C = c\right]$, $\mathbb{V}\left(A|c\right) = Var\left(A|C = c\right)$ and $\mathbb{C}\left(A, B|c\right) = Cov\left(A, B|C = c\right)$.

[4] An earlier version of this paper considered the slightly more general set-up with $\psi(Z, \beta) = \psi_1(Y_1, X, \beta) - \psi_0(Y_0, X, \beta)$ with (X, Y) observed where $Y = DY_1 + (1 - D)Y_0$. Results for this extended model, which contains the standard causal inference model and the two-sample instrumental variables model as special cases (cf., Imbens, 2004; Angrist and Krueger, 1992), follow directly and straightforwardly from those outlined below.

(e.g., Robins, Hsieh and Newey, 1995; Chen, Hong and Tamer, 2005), missing regressors (e.g., Robins, Rotnitzky and Zhao, 1994), attrition in panel data (e.g., Robins, Rotnitzky and Zhao, 1995; Robins and Rotnitzky, 1995; Wooldridge, 2002), and M-estimation under variable probability sampling (e.g., Wooldridge, 1999, 2007). Chen, Hong and Tarozzi (2004), Wooldridge (2007) and Egel, Graham and Pinto (2008) discuss several other applications.

The maximal asymptotic precision with which β_0 can be estimated under the MAR setup has been characterized by Robins, Rotnitzky and Zhao (1994) and is given by

$$\mathcal{I}_{\mathrm{m}}(\beta_0) = \Gamma_0' \Lambda_0^{-1} \Gamma_0, \tag{2}$$

with $\Lambda_0 = \mathbb{E}\left[\Sigma_0(X)/p_0(X) + q(X;\beta_0)\,q(X;\beta_0)'\right]$, where $\Sigma_0(x) = \mathbb{V}(\psi(Z,\beta_0)|\,x)$ and $q(x;\beta) = \mathbb{E}[\psi(Z,\beta)\,|x]$.

The associated efficient influence function, also due to Robins, Rotnitzky and Zhao (1994), is given by

$$\phi(z,\theta_0) = \Gamma_0^{-1} \times \left\{ \frac{d}{p_0(x)}\psi(z,\beta_0) - \frac{q(x;\beta_0)}{p_0(x)}(d - p_0(x)) \right\} \tag{3}$$

for $\theta = (p, q', \beta')'$.

The calculation of (2) is now standard. Knowledge of (2) is useful because it quantifies the cost – in terms of asymptotic precision – of the missing data and because it can be used to verify whether a specific estimator for β_0 is efficient. To simplify what follows I will explicitly assume that $\mathcal{I}_{\mathrm{m}}(\beta_0)$ is well-defined (i.e., that all its component expectations exist and are finite, and that all its component matrices are nonsingular).

This paper shows that the semiparametric efficiency bound for β_0 under the MAR setup, coincides with the bound for a particular augmented moment condition problem. The augmented system consists of the inverse probability of observation weighted (IPW) original moment restriction (1) and an additional conditional moment restriction which exhausts all other implications of the MAR setup. This general equivalence result, while implicit in the form of the efficient influence function (3), is apparently new. It provides fresh intuitions for several 'paradoxes' in the missing data literature, including the well-known results that projection onto, or weighting by the inverse of, a known propensity score results in inefficient estimates (e.g., Hahn, 1998; Hirano, Imbens and Ridder, 2003), that smoothness and exclusion priors on the propensity score do not increase the precision with which β_0 can be estimated (Robins, Hsieh and Newey, 1995; Robins and Rotnitzky, 1995; Hahn, 1998, 2004) and that weighting by a nonparametric estimate of the propensity score results in an efficient estimator (Hirano, Imbens and Ridder, 2003; cf., Hahn, 1998; Wooldridge, 2007; Prokhorov and Schmidt, 2009; Hitomo, Nishiyama and Okui, 2008).

This paper also analyzes the effect of imposing additional semiparametric restrictions on the conditional expectation function (CEF) $q(x;\beta) = \mathbb{E}[\psi(Z,\beta)\,|x]$. If $\psi(Z,\beta) = Y_1 - \beta$, as when the target parameter is $\beta_0 = \mathbb{E}[Y_1]$, then such restrictions may arise from prior information on the form of $\mathbb{E}[Y_1|x]$. Such restrictions may arise in other settings as well. For example, if the goal is to estimate a vector of linear predictor coefficients in the presence of missing regressors, then a semiparametric model for the CEFs of the missing regressors given always-observed variables generates restrictions on the form of $q(x;\beta)$ (cf., Robins, Rotnitzky and Zhao, 1994).[5]

Formally I consider the semiparametric model defined by restriction (1), Assumptions 1.1 to 1.4 and the additional assumption.

Assumption 1.5 (Functional Restriction) *Partition* $X = (X_1', X_2')'$*, then*

$$\mathbb{E}\left[\psi\left(Z,\beta_0\right)|x\right] = q\left(x,\delta_0,h_0\left(x_2\right);\beta_0\right)$$

where $q\left(x,\delta,h\left(x_2\right);\beta\right)$ *is a known* $K\times 1$ *function,* δ *a* $J\times 1$ *finite dimensional unknown parameter, and* $h(\cdot)$ *an unknown function mapping from a subset of* $\mathcal{X}_2 \subset \mathbb{R}^{\dim(X_2)}$ *into* $\mathcal{H}\subset\mathbb{R}^P$*.*

To the best of my knowledge the variance bound for this problem, the MAR setup with 'functional' restrictions, has not been previously calculated. In an innovative paper, Wang, Linton and Härdle (2004) consider a special case of this model where $\psi(Z,\beta) = Y_1 - \beta$. They impose a partial linear structure, as in Engle et al (1986), on $\mathbb{E}\left[Y_1|x\right]$ such that $q\left(x,\delta_0,h_0\left(x_2\right);\beta_0\right) = x_1'\delta_0 + h_0\left(x_2\right) - \beta_0$. In making their variance bound calculation they assume that the conditional distribution of Y_1 given X is normal with a variance that does not depend on X. They do not provide a bound for the general case but conjecture that it is "very complicated" (p. 338). The result given below extends their work to moment condition models, general forms for $q\left(x,\delta,h\left(x_2\right);\beta\right)$ and, importantly, does not require that $\psi(Z,\beta)$ be conditionally normally distributed and/or homoscedastic.

Augmenting the MAR setup with Assumption 1.5 generates a middle ground between the fully parametric likelihood-based approaches to missing data described by Little and Rubin (2002) and those which leave $\mathbb{E}\left[\psi\left(Z,\beta_0\right)|x\right]$ unrestricted (e.g., Cheng, 1994; Hahn, 1998; Hirano, Imbens and Ridder, 2003). Likelihood-based approaches are very sensitive to misspecification (cf., Imbens, 2004), while approaches which utilize only the basic MAR setup require high dimensional smoothing which may deleteriously affect small sample performance (cf., Wang, Linton and Härdle, 2004; Ichimura and Linton, 2005). Assumption 1.5 is generally weaker than a parametric specification for the conditional distribution of $\psi(Z,\beta_0)$ given X, but at the same time reduces the dimension of the nonparametric smoothing problem. Below I show how to efficiently exploit prior information on the form of $\mathbb{E}\left[\psi\left(Z,\beta_0\right)|x\right]$. I also provide conditions under which consistent estimation of β_0 is possible even if the exploited information is incorrect.

Section 2 reports the first result of the paper: an equivalence between the MAR setup and a particular method-of-moments problem. Equivalence, which is suggested by the form of the efficient influence function derived by Robins, Rotnitzky and Zhao (1994), was previously noted for special cases by Newey (1994a) and Hirano, Imbens and Ridder (2003). I discuss the connection between their results and the general result provided below. I also highlight some implications of the equivalence result for understanding various aspects of the MAR setup. Section 3 calculates

[5] The formation of predictive models of this type is the foundation of the imputation approach to missing data described in Little and Rubin (2002).

the variance bound for β_0 when the MAR setup is augmented by Assumption 1.5. I discuss when Assumption 1.5 is likely to be informative and also when consistent estimation is possible even if it is erroneously maintained.

2 Equivalence result

Under the MAR setup the inverse probability weighted (IPW) moment condition

$$\mathbb{E}\left[\frac{D}{p_0(X)}\psi(Z,\beta_0)\right]=0, \tag{4}$$

is valid (e.g., Hirano, Imbens and Ridder, 2003; Wooldridge, 2007). The conditional moment restriction

$$\mathbb{E}\left[\left.\frac{D}{p_0(X)}-1\right|X\right]=0 \quad \forall \quad X\in\mathcal{X}, \tag{5}$$

also holds and nonparametrically identifies $p_0(x)$. While the terminology is inexact, in what follows I call (4) the *identifying moment* and (5) the *auxiliary moment.*

Consider the case where $p_0(x)$ is known such that (5) is truly an auxiliary moment. One efficient way to exploit the information (5) contains is to, following Newey (1994a) and Brown and Newey (1998), reduce the sampling variation in (4) by subtracting from it the fitted value associated with its regression onto the infinite-dimensional vector of unconditional moment functions implied by (5):[6]

$$\begin{aligned} s(Z,\theta_0) &= \frac{D}{p_0(X)}\psi(Z,\beta_0)-\mathbb{E}^*\left[\left.\frac{D}{p_0(X)}\psi(Z,\beta_0)\right|\frac{D}{p_0(X)}-1;X\right] \\ &= \frac{D}{p_0(X)}\psi(Z,\beta_0)-\frac{q(X;\beta_0)}{p_0(X)}(D-p_0(X)). \end{aligned}$$

That this population residual is equal to the efficient score function derived by Robins, Rotnitzky and Zhao (1994) strongly suggests an equivalence between the GMM problem defined by restrictions (4) and (5) and the MAR setup outlined above. One way to formally show this is to verify that the efficiency bounds for β_0 in the two problems coincide.[7] The bound for β_0 under the MAR set-up is given (2) above, while under the moment problem it is established by the following theorem.

Theorem 2.1 (GMM Equivalence) *Suppose that (i) the distribution of Z has a known, finite support, (ii) there is some $\beta_0\in\mathcal{B}\subset\mathbb{R}^K$ and $\rho_0=(\rho_1,\ldots,\rho_L)'$ where $\rho_l=p_0(x_l)\in[\kappa,1]$ for each $l=1,\ldots,L$ and some $0<\kappa<1$ (with $\mathcal{X}=\{x_1,\ldots,x_L\}$ the known support of X) such that restrictions*

[6] The notation $\mathbb{E}^*[Y|X;Z]$ denotes the (mean squared error minimizing) linear predictor of Y given X *within a subpopulation homogenous in* Z:

$$\mathbb{E}^*[Y|X;Z]=X'\pi(Z), \quad \pi(Z)=\mathbb{E}[XX'|Z]^{-1}\times\mathbb{E}[XY|Z].$$

Wooldridge (1999b, Section 4) collects some useful results on conditional linear predictors. See also Newey (1990) and Brown and Newey (1998).

[7] An alternative approach to showing equivalency would involve verifying Newey's (2004) moment spanning condition for efficiency.

(4) and (5) hold, (iii) Λ_0 *and* $\mathcal{I}_{\mathrm{m}}(\beta_0) = \Gamma_0' \Lambda_0^{-1} \Gamma_0$ *are nonsingular and (iv) other regularity conditions hold (cf., Chamberlain (1992b), Section 2), then* $\mathcal{I}_{\mathrm{m}}(\beta_0)$ *is the Fisher information bound for* β_0.

Proof. See the supplemental materials. ■

The proof of Theorem 2.1 involves only some tedious algebra and a straightforward application of Lemma 2 of Chamberlain (1987). Assuming that Z has known, finite support makes the problem fully parametric. The unknown parameters are the probabilities associated with each possible realization of Z, the values of the propensity score at each of the L mass points of the distribution of X, $\rho_0 = (\rho_1, \ldots, \rho_L)'$, and the parameter of interest, β_0.

The multinomial assumption is not apparent in the form of $\mathcal{I}_{\mathrm{m}}(\beta_0)$, which involves only conditional expectations of certain functions of the data. This suggests that the bound holds in general since any F_0 which satisfies (4) and (5) can be arbitrarily well-approximated by a multinomial distribution also satisfying the restrictions. Chamberlain (1992a, Theorem 1) demonstrates that this is indeed the case. Therefore $\mathcal{I}_{\mathrm{m}}(\beta_0)^{-1}$ is the maximal asymptotic precision, in the sense of Hájek's (1972) local minimax approach to efficiency, with which β_0 can be estimated when the only prior restrictions on F_0 are (4) and (5). Since this variance bound coincides with (2) I conclude that (4) and (5) exhaust all of the useful prior restrictions implied by the MAR setup.[8]

The connection between semiparametrically efficient estimation of moment condition models with missing data and augmented systems of moment restrictions has been noted previously for the special case of data missing completely at random (MCAR). In that case Assumptions 1.1 to 1.4 hold with $p_0(X)$ equal to a (perhaps known) constant. Newey (1994a) shows that an efficient estimate of β_0 can be based on the pair of moment restrictions

$$\mathbb{E}\left[D\psi(Z,\beta_0)\right] = 0, \qquad \mathbb{C}\left(D, q(X;\beta_0)\right) = 0,$$

with $q(X;\beta)$ as defined above. Hirano, Imbens and Ridder (2003) discuss a related example with X binary and the data also MCAR. In their example efficient estimation is possible with only a finite number of unconditional moment restrictions. Theorem 2.1 provides a formal generalization of the Newey (1994a) and Hirano, Imbens and Ridder (2003) examples to the missing at random (MAR) case.

The method-of-moments formulation of the MAR setup provides a useful framework for un-

[8] A referee made the insightful observation that the moment condition model (4) and (5) and the MAR setup are equivalent in the stronger sense that they impose identical restrictions on the observed data. This, of course, also implies that they contain identical information on β_0. The complete data vector is given by (D, X, Y_1), with only $(D, X, Y) = (D, X, DY_1)$ observed. Since Y_1 is not observed whenever $D = 0$ we are free specify its conditional distribution given X and $D = 0$ as desired. Choosing $Y_1 | X, D = 0 \overset{D}{\sim} Y_1 | X, D = 1$ ensures conditional independence (Assumption 1.3). Manipulating the identifying moment (4) we then have, writing $\psi(Z,\beta_0) = \psi(X, Y_1, \beta_0)$,

$$\begin{aligned}\mathbb{E}\left[\frac{D}{p_0(X)}\psi(X,Y,\beta_0)\right] &= \mathbb{E}\left[p_0(X)\,\mathbb{E}\left[\frac{D}{p_0(X)}\psi(X,DY_1,\beta_0)\middle| X, D=1\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\psi(X,DY_1,\beta_0)|\, X, D=1\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\psi(X,Y_1,\beta_0)|\, X\right]\right],\end{aligned}$$

which yields (1). Finally, the auxiliary restriction (5) ties down the conditional distribution of D given X and ensures Assumption 1.4 is satisfied. I thank Michael Jansson for several helpful discussions on this point.

derstanding several apparent paradoxes found in the missing data literature. As a simple example consider Hahn's (1998, pp. 324 - 325) result that projection onto a known propensity score may be harmful for estimation of $\beta_0 = \mathbb{E}[Y_1]$. Formally he shows that, for $p_0(x) = Q_0$ constant in x and known, the complete-case estimator, $\widehat{\beta}_{\text{cc}} = \sum_{i=1}^{N} D_i Y_{1i} / \sum_{i=1}^{N} D_i$, while consistent, is inefficient. Observe that for the constant propensity score case $\widehat{\beta}_{\text{cc}}$ is the sample analog of the population solution to (4). It consequently makes no use of any information contained in the auxiliary moment (5). However, that moment will be informative for β_0 if $q(x;\beta_0) = \mathbb{E}[Y_1|\, x] - \beta_0$ varies with x, consistent with Hahn's (1998) finding that the efficiency loss associated with $\widehat{\beta}_{\text{cc}}$ is proportional to $\mathbb{V}(q(X;\beta_0))$. Similar reasoning explains why weighting by the (inverse of) the known propensity score is generally inefficient (cf., Robins, Rotnitzky and Zhao, 1994; Hirano, Imbens and Ridder, 2003; Wooldridge, 2007). The known weights estimator ignores the information contained in (5).

That smoothness and exclusion priors on the propensity score do not lower the variance bound also has a GMM interpretation. Consider the case where the propensity score belongs to a parametric family $p(X;\eta_0)$. If η_0 is known, then an efficient GMM estimator based on (4) and (5) is given by the solution to

$$\frac{1}{N}\sum_{i=1}^{N} s\left(\eta_0, \widehat{q}, \widehat{\beta}\right) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_i}{p(X_i;\eta_0)}\psi(Z_i,\widehat{\beta}) - \frac{\widehat{q}(X_i;\widehat{\beta})}{p(X_i;\eta_0)}(D_i - p(X_i;\eta_0))\right\} = 0,$$

with $\widehat{q}(x;\widehat{\beta})$ a consistent nonparametric estimate of $\mathbb{E}[\psi(Z,\beta_0)|\, x]$. Now consider the effect of replacing η_0 with the consistent estimate $\widehat{\eta}$. From Newey and McFadden (1994, Theorem 6.2), this replacement does not change the first order asymptotic sampling distribution of $\widehat{\beta}$ because $\mathbb{E}[\partial s(\eta_0, q_0, \beta_0)/\partial\eta'] = 0$. Furthermore, if the known propensity score is replaced by a consistent nonparametric estimate, $\widehat{p}(x)$, then the sampling distribution of $\widehat{\beta}$ is also unaffected (Newey 1994b, Proposition 3, p. 1360). Since the M-estimate of β_0 based on its efficient score function has the same asymptotic sampling distribution whether the propensity score is set equal to the truth or instead to a noisy, but consistent, estimate, knowledge of its form cannot increase the precision with which β_0 may be estimated.

Another intuition for redundancy of knowledge of the propensity score can be found by inspecting the information bound for the multinomial problem. Under the conditions of Theorem 2.1 calculations provided in the supplemental materials imply that the GMM estimates of β_0 and ρ_0 (recall that ρ_0 contains the values for the propensity score at each of the mass points of the distribution of X) have an asymptotic sampling distribution of

$$\sqrt{N}\left(\begin{bmatrix}\widehat{\rho}\\ \widehat{\beta}\end{bmatrix} - \begin{bmatrix}\rho_0\\ \beta_0\end{bmatrix}\right) \xrightarrow{D} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\mathcal{I}_{\text{m}}(\rho_0)^{-1} & 0\\ 0 & \mathcal{I}_{\text{m}}(\beta_0)^{-1}\end{bmatrix}\right),$$

with $\mathcal{I}_{\text{m}}(\beta_0)$ as defined in (2) and $\mathcal{I}_{\text{m}}(\rho_0)$ as defined in the supplement. As is well-known, under block diagonality sampling error in $\widehat{\rho}$ does not affect, at least to first order, the asymptotic sampling properties of $\widehat{\beta}$. While block diagonality is formally only a feature of the multinomial problem, the result nonetheless provides another useful intuition for understanding why prior knowledge of the propensity score is not valuable asymptotically.

Finally the combination of redundancy of knowledge of the propensity score, and the structure of the equivalent GMM problem, suggests why the IPW estimator based on a nonparametric estimate of the propensity score is semiparametrically efficient (Hirano, Imbens and Ridder, 2003): when a nonparametric estimate of the propensity score is used the sample analog of both (4) and (5) are satisfied. In contrast the IPW estimator based on a parametric estimate of the propensity score will only satisfy a finite number of the moment conditions implied by (5), hence while it will be more efficient than the estimator which weights by the true propensity score (e.g., Wooldridge, 2007), it will be less efficient than the one proposed by Hirano, Imbens and Ridder (2003).

3 Semiparametric functional restrictions

Consider the MAR setup augmented by Assumption 1.5. To the best of my knowledge, the maximal asymptotic precision with which β_0 can be estimated in this model has not been previously characterized. In order to calculate the bound for this problem I first consider the conditional moment problem defined by (4) and (5) and

$$\mathbb{E}\left[\rho\left(Z,\delta_0,h_0\left(X_2\right);\beta_0\right)|X\right]=0, \tag{6}$$

with $\rho\left(Z,\delta_0,h_0\left(X_2\right);\beta_0\right)=\psi\left(Z,\beta_0\right)-q\left(x,\delta_0,h_0\left(x_2\right);\beta_0\right)$. I apply Chamberlain's (1992a) approach to this problem to calculate a variance bound for β_0. I then show that this bound coincides with the semiparametric efficiency bound for the problem defined by restriction (1) and Assumptions 1.1 to 1.5 using the methods of Bickel, Klaassen, Ritov and Wellner (1993). The value of first considering the conditional moment problem is that it provides a conjecture for the form of the efficient influence function, therefore sidestepping the need to directly calculate what is evidently a complicated projection.

To present these results I begin by letting $q_0\left(X\right)=q\left(X,\delta_0,h_0\left(X_2\right);\beta_0\right)$, $\rho\left(Z;\beta_0\right)=\psi\left(Z,\beta_0\right)-q_0\left(X\right)$,

$$\underset{P\times P}{\Upsilon_0^h}\left(X_2\right)=\mathbb{E}\left[D\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)'\Sigma_0\left(X\right)^{-1}\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)\middle|X_2\right]$$

$$\underset{K\times J}{\Upsilon_0^{h\delta}}\left(X_2\right)=\mathbb{E}\left[D\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)'\Sigma_0\left(X\right)^{-1}\left(\frac{\partial q_0\left(X\right)}{\partial \delta'}\right)\middle|X_2\right]$$

$$\underset{K\times J}{G_0}\left(X\right)=\frac{\partial q_0\left(X\right)}{\partial \delta'}-\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)\Upsilon_0^h\left(X_2\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right),\quad \underset{K\times P}{H_0}\left(X_2\right)=\mathbb{E}\left[\frac{\partial q_0\left(X\right)}{\partial h'}\middle|X_2\right]$$

$$\underset{J\times J}{\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}}\left(\delta_0\right)=\mathbb{E}\left[DG_0\left(X\right)'\Sigma_0\left(X\right)^{-1}G_0\left(X\right)\right],$$

and

$$\underset{K\times K}{\Xi_0} = \mathbb{E}\left[H_0(X_2)\Upsilon_0^h(X_2)^{-1}H_0(X_2)'\right] + \mathbb{E}\left[G_0(X)\right]\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\delta_0)^{-1}\mathbb{E}\left[G_0(X)\right]' + \mathbb{E}\left[q_0(X)\,q_0(X)'\right].$$

The variance bound for β_0 in the conditional moment problem defined by (4), (5) and (6) is established by the following Theorem.

Theorem 3.1 (Efficiency with Functional Restrictions, Part 1) *Suppose that (i) the distribution of Z has a known, finite support, (ii) there is some $\beta_0 \in \mathcal{B} \subset \mathbb{R}^K$, $\rho_0 = (\rho_1, \ldots, \rho_L)'$ where $\rho_l = p_0(x_l) \in [\kappa, 1]$ for each $l = 1, \ldots, L$ and some $0 < \kappa < 1$ (with $\mathcal{X} = \{x_1, \ldots, x_L\}$ the known support of X), $\delta_0 \in \mathcal{D} \subset \mathbb{R}^J$ and $h_0(x_{2,m}) = \lambda_{0,m} \in \mathcal{L} \subset \mathbb{R}^P$ for each $m = 1, \ldots, M$ (with $\mathcal{X}_2 = \{x_{2,1}, \ldots, x_{2,M}\}$ the known support of X_2) such that restrictions (4), (5) and (6) hold, (iii) Ξ_0 and $\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\beta_0) = \Gamma_0'\Xi_0^{-1}\Gamma_0$ are nonsingular and (iv) other regularity conditions hold (cf., Chamberlain 1992b, Section 2), then $\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\beta_0)$ is the Fisher information bound for β_0.*

Proof. See the supplemental materials. ■

Note that if $X_1 = \varnothing$ and $X_2 = X$, such that $\mathbb{E}\left[\psi(Z, \beta_0)|\, x\right]$ is unrestricted, then $\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\beta_0)$ simplifies to $\mathcal{I}_{\mathrm{m}}(\beta_0)$ above. Therefore, Theorem 2.1 may be viewed as a special case of Theorem 3.1. As with Theorem 2.1, the validity of the bound for the non-multinomial case follows from Theorem 1 of Chamberlain (1992a).

The form of Ξ_0 suggests a candidate efficient influence function of

$$\phi_\beta^{\mathrm{f}}(Z, \eta_0, \beta_0) = \Gamma_0^{-1}\Big\{DH_0(X_2)\Upsilon_0^h(X_2)^{-1}\left(\frac{\partial q_0(X)}{\partial h'}\right)'\Sigma_0(X)^{-1}\rho(Z;\beta_0) + D\mathbb{E}\left[G_0(X)\right]\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\delta_0)^{-1}G_0(X)'\Sigma_0(X)^{-1}\rho(Z;\beta_0) + q(X;\beta_0)\Big\}. \tag{7}$$

where $\eta = \left(h, \delta, H, \Upsilon^h, \Upsilon^{h\delta}, \Sigma, \overline{G}\right)$, with $\overline{G} = \mathbb{E}\left[G(X)\right]$. Note that each of the three components of (7) are mutually uncorrelated. The next Theorem verifies that (7) is the efficient influence function under the MAR setup with Assumption 1.5 also imposed.

Theorem 3.2 (Efficiency with Functional Restrictions, Part 2) *The semiparametric efficiency bound for β_0 in the problem defined by restriction (1) and Assumptions 1.1 to 1.5 is equal to $\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\beta_0)$ with an efficient influence function of $\phi_\beta^{\mathrm{f}}(Z, \eta_0, \beta_0)$.*

Proof. See the supplemental materials. ■

Theorem 3.1 implies that Assumption 6 can be exploited to more efficiently estimate β_0. However its use also carries risk, if false, yet nevertheless erroneously maintained by the data analyst, an inconsistent estimate of β_0 may result. This tension, between efficiency and robustness, is formalized by the next two Propositions which together provide guidance as to whether prior information of the type given by Assumption 1.5 should be utilized in practice.

The first Proposition characterizes the magnitude of the efficiency gain associated with correctly exploiting Assumption 1.5. Define:

$$\underset{K\times 1}{\xi_1}\left(Z,\eta_0,\beta_0\right) = D\left\{\frac{I_K}{p_0\left(X\right)} - H_0\left(X_2\right)\Upsilon_0^h\left(X_2\right)^{-1}\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)'\Sigma_0\left(X\right)^{-1}\right\}\rho\left(Z;\beta_0\right)$$
$$\underset{J\times 1}{\xi_2}\left(Z,\eta_0,\beta_0\right) = DG_0\left(X\right)'\Sigma_0\left(X\right)^{-1}\rho\left(Z;\beta_0\right).$$

Proposition 3.1 *Under (1) and Assumptions 1.1 to 1.5*

$$\mathcal{I}_{\mathrm{m}}\left(\beta_0\right)^{-1} - \mathcal{I}_{\mathrm{m}}^{\mathrm{f}}\left(\beta_0\right)^{-1} = \Gamma_0^{-1}\left(\mathbb{V}\left(\xi_1\right) - \mathbb{C}\left(\xi_1,\xi_2'\right)\mathbb{V}\left(\xi_2\right)^{-1}\mathbb{C}\left(\xi_1,\xi_2'\right)'\right)\Gamma_0^{-1\prime} \geq 0. \tag{8}$$

Proof. See the supplemental materials. ■

Equation (8) has an intuitive interpretation. The first term in parentheses

$$\mathbb{V}\left(\xi_1\right) = \mathbb{E}\left[\frac{\Sigma_0\left(X\right)}{p_0\left(X\right)} - H_0\left(X_2\right)\Upsilon_0^h\left(X_2\right)^{-1}H_0\left(X_2\right)'\right],$$

equals the asymptotic variance reduction that would be available by additionally imposing Assumption 6 if δ_0 were *known*.

The additional (asymptotic) sampling uncertainty induced by having to estimate δ_0 is captured by the second term

$$\mathbb{C}\left(\xi_1,\xi_2\right)\mathbb{V}\left(\xi_2\right)^{-1}\mathbb{C}\left(\xi_1,\xi_2\right) = \mathbb{E}\left[G_0\left(X\right)\right]\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}\left(\delta_0\right)^{-1}\mathbb{E}\left[G_0\left(X\right)\right]',$$

where $\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}\left(\delta_0\right)$ is the information bound for δ_0 in the semiparametric regression problem (cf., Chamberlain, 1992a):

$$D\psi\left(Z,\beta_0\right) = Dq\left(X,\delta_0,h_0\left(X_2\right);\beta_0\right) + DV, \quad \mathbb{E}\left[V\mid X,D=1\right] = \mathbb{E}\left[V\mid X\right] = 0.$$

The more precisely determined δ_0, the greater the efficiency gain from imposing Assumption 1.5. The size of $\mathbb{E}\left[G_0\left(X\right)\right]$ also governs the magnitude of the efficiency gain. Conditional on X_2, $\left(\frac{\partial q_0(X)}{\partial h'}\right)\Upsilon_0^h\left(X_2\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right)$ is a weighted linear predictor of $\frac{\partial q_0(X)}{\partial \delta'}$ given $\frac{\partial q_0(X)}{\partial h'}$ *in the* $D=1$ *subpopulation.* That is[9]

$$\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)\Upsilon_0^h\left(X_2\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right) = \mathbb{E}_{\Sigma_0(X)}^*\left[\frac{\partial q_0\left(X\right)}{\partial \delta'}\middle|\frac{\partial q_0\left(X\right)}{\partial h'};X_2,D=1\right],$$

and hence $G_0\left(X\right)$ is equal to the difference between $\frac{\partial q_0(X)}{\partial \delta'}$ and its predicted value based on a

[9] The notation $\mathbb{E}_{\omega(X)}^*\left[Y\mid X;Z,D=1\right]$ denotes the weighted conditional linear predictor

$$\mathbb{E}_{\omega(X)}^*\left[Y\mid X;Z,D=1\right] = X\mathbb{E}\left[DX\omega\left(X\right)^{-1}X'\middle|Z\right]^{-1}\times\mathbb{E}\left[DX\omega\left(X\right)^{-1}Y\middle|Z\right].$$

This is the population analog of the fitted value from a generalized least squares regression in a subpopulation homogenous in Z and with $D=1$.

weighted least squares regression in the $D = 1$ subpopulation. The average of these differences, $\mathbb{E}\left[G_0\left(X\right)\right]$, is taken across the *entire* population; it will be large in absolute value when the distribution of X_1 conditional on X_2 differs in the $D = 1$ versus $D = 0$ subpopulations. This will occur whenever X_1 is highly predictive for missingness (conditional on X_2). In such situations the efficiency costs of sampling uncertainty in $\widehat{\delta}$ are greater (relative to the known δ_0 case) because estimation of β_0 requires greater levels of extrapolation.

An example clarifies the discussion given above. Assume that $\psi\left(Z, \beta_0\right) = Y_1 - \beta_0$ with

$$q\left(X, \delta_0, h_0\left(X_2\right); \beta_0\right) = X_1'\delta_0 + h_0\left(X_2\right) - \beta_0.$$

This is the model considered by Wang, Linton and Härdle (2004). In addition to being of importance in its own right, it provides insight into the program evaluation problem (where the means of two missing outcomes, as opposed to just one, need to be estimated). Wang, Linton and Härdle's (2004) prior restriction includes the condition that $\mathbb{V}\left(Y_1 \middle| X\right) = \sigma_1^2$ is constant in X. For clarity of exposition I also assume homoscedasticity holds, but that this fact is not known by the econometrician. Let $e_0\left(X_2\right) = \mathbb{E}\left[p\left(X\right) \middle| X_2\right] = \Pr\left(D = 1 \middle| X_2\right)$; specializing the general results given above to this model and evaluating (8) gives

$$\begin{aligned} &\mathcal{I}_{\mathrm{m}}\left(\beta_0\right)^{-1} - \mathcal{I}_{\mathrm{m}}^{\mathrm{f}}\left(\beta_0\right)^{-1} \\ &\quad = \sigma_1^2\left\{\mathbb{E}\left[\mathbb{E}\left[\left.\frac{1}{p\left(X\right)}\right| X_2\right] - \frac{1}{e_0\left(X_2\right)}\right]\right. \\ &\quad\quad \left. - \frac{\left(\mathbb{E}\left[\mathbb{E}\left[X_1 | X_2\right] - \mathbb{E}\left[X_1 | X_2, D=1\right]\right]\right)'\left(\mathbb{E}\left[\mathbb{E}\left[X_1 | X_2\right] - \mathbb{E}\left[X_1 | X_2, D=1\right]\right]\right)}{\mathbb{E}\left[e_0\left(X_2\right)\mathbb{V}\left(X_1 | X_2, D=1\right)\right]}\right\} \geq 0, \end{aligned}$$

which shows that the efficiency gain associated with correctly exploiting Assumption 1.5 reflects three forces. First, substantial convexity in $p\left(X\right)^{-1}$, which will occur when overlap is limited, increases the efficiency gain.[10] This gain reflects the semiparametric restriction allowing for extrapolation in the presence of conditional covariate imbalance. The next two effects reflect the fact that the first source of efficiency gain is partially nullified by having to estimate δ_0. If X_1 varies strongly given X_2 in the $D = 1$ subpopulation then the information for δ_0 is large which, in turn, increases the precision with which β_0 may be estimated. On the other hand if there are large (average) differences in the conditional mean of X_1 given X_2 across the $D = 1$ and $D = 0$ subpopulations, then estimating β_0 requires greater extrapolation which – when δ_0 is unknown – decreases the precision with which it may be estimated.

Proposition 3.1 provides insight into when correctly imposing Assumption 1.5 is likely to be informative. A related question concerns the consequences of misspecifying the form of $q\left(X, \delta, h\left(X_2\right); \beta\right)$. Under such misspecification the conditional moment restriction (6) will be invalid. Nevertheless the efficient score function may continue to have an expectation of zero at $\beta = \beta_0$. This suggest that an M-estimator based on an estimate of the efficient score function may be consistent even if As-

[10] When some subpopulations have low propensity scores $\mathbb{E}\left[1/p\left(X\right) \middle| X_2\right] - 1/\mathbb{E}\left[p\left(X\right) \middle| X_2\right]$ will tend to be large (Jensen's Inequality).

sumption 1.5 does not hold. The following proposition provides one set of conditions under which such a robustness property holds.

Proposition 3.2 (Double Robustness) *Let* $q_*(X) = q(X, \delta_*, h_*(X_2); \beta_0)$ *with* δ_* *and* $h_*(X_2)$ *arbitrary,* $\rho_*(Z;\beta_0) = \psi(Z,\beta_0) - q_*(X)$*, and redefine* $\Sigma_0(X) = \mathbb{V}(\rho_*(Z;\beta_0)|X)$*,* $H_0(X_2) = \mathbb{E}\left[\left.\frac{\partial q_*(X)}{\partial h'}\right| X_2\right]$ *and* $\Upsilon_0^h(X_2)$*,* $\Upsilon_0^{h\delta}(X_2)$*, and* $\overline{G}_0$ *similarly. Under restriction (1) and Assumptions 1.1 to 1.4* $\phi_\beta^{\mathrm{f}}(Z,\eta,\beta_0)$ *is mean zero if either (i)* $\beta = \beta_0$*,* $\eta = \eta_0$ *and Assumption 1.5 holds or (ii)* $\beta = \beta_0$*,* $\eta = \eta_* = \left(h_*, \delta_*, H_0, \Upsilon_0^h, \Upsilon_0^{h\delta}, \Sigma_0, \overline{G}_0\right)$ *and (a)* $p_0(x) = e_0(x_2)$ *for all* $x \in \mathcal{X}$*, (b)* $\Sigma_0(x) = \Theta_0(x_2)$ *for all* $x \in \mathcal{X}$*, and (c) at least one element of* $h_*(x_2)$ *enters linearly in each row of* $q_*(X)$*.*

Proof. See the supplemental materials. ■

Note that there is a tension between the robustness property of Proposition 3.2 and the efficiency gain associated with Assumption 1.5. Mean-zeroness of $\phi_\beta^{\mathrm{f}}(Z,\eta,\beta_0)$ under misspecification requires that those variables entering $q(X,\delta,h(X_2);\beta_0)$ parametrically do not affect either the probability of missingness or the conditional variance of the moment function (1). Under such conditions an estimator based on $\phi_\beta^{\mathrm{f}}(Z,\eta,\beta_0)$ will perform no better, at least asymptotically, than one based on the efficient score function derived by Robins, Rotnitzky and Zhao (1994). In particular we have:

Corollary 3.1 *Under the conditions of part (ii) of Proposition 3.2*

$$\mathcal{I}_{\mathrm{m}}(\beta_0)^{-1} - \mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\beta_0)^{-1} = 0.$$

Proof. See the supplemental materials. ■

Collectively Propositions 3.1 and 3.2 suggest that estimation while maintaining Assumption 1.5 will be most valuable when the econometrician is highly confident in the imposed semiparametric restriction.

References

Angrist, Joshua D. and Alan B. Krueger. (1992). "The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples," *Journal of the American Statistical Association* 87 (418): 328 - 336.

Bang, Heejung and James M. Robins. (2005). "Doubly robust estimation in missing data and causal inference models," *Biometrics* 61 (4): 962 - 972.

Bickel, Peter J., Chris A.J. Klaassen, Ya'acov Ritov and Jon A. Wellner. (1993). *Efficient and adaptive estimation for semiparametric models.* New York: Springer-Verlag, Inc.

Brown, Bryan W. and Whitney K. Newey. (1998). "Efficient semiparametric estimation of expectations," *Econometrica* 66 (2): 453 - 464.

Chamberlain, Gary. (1987). "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics* 34 (1): 305 - 334.

Chamberlain, Gary. (1992a). "Efficiency bounds for semiparametric regression," *Econometrica* 60 (3): 567 - 596.

Chamberlain, Gary. (1992b). "Comment: sequential moment restrictions in panel data," *Journal of Business and Economic Statistics* 10 (1): 20 - 26.

Chen, Xiaohong, Han Hong, Elie T. Tamer. (2005). "Measurement error models with auxiliary data," *Review of Economic Studies* 72 (2): 343 - 366.

Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2004). "Semiparametric efficiency in GMM models of nonclassical measurement errors, missing data and treatment effects, *Mimeo.*

Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2008). "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics* 36 (2): 808 - 843.

Cheng, Philip E. (1994). "Nonparametric estimation of mean functionals with data missing at random," *Journal of the American Statistical Association* 89 (425): 81 - 87.

Egel, Daniel, Bryan S. Graham, and Cristine Pinto. (2008). "Inverse probability tilting and missing data problems," *NBER Working Paper No.* 13981.

Engle, Robert F., C. W. J. Granger, John Rice and Andrew Weiss. (1986). "Semiparametric estimates of the relation between weather and electricity sales," *Journal of the American Statistical Association* 81 (394): 310 - 320.

Hahn, Jinyong. (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica* 66 (2): 315 - 331.

Hahn, Jinyong. (2004). "Functional restriction and efficiency in causal inference," *Review of Economics and Statistics* 86 (1): 73 - 76.

Hájek, Jaroslav. (1972). "Local asymptotic minimax and admissibility in estimation," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* 1: 175 - 194 (L. M. Le Cam, J. Neyman & E. L. Scott, Eds.). Berkeley: University of California Press.

Hirano, Keisuke, Guido W. Imbens and Geert Ridder. (2003). "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica* 71 (4): 1161 - 1189.

Hitomo, Kohtaro, Yoshihiko Nishiyama and Ryo Okui. (2008). "A puzzling phenomenon in semiparametric estimation problems with infinite-dimensional nuisance parameters," *Econometric Theory* 24 (6): 1717 - 1728.

Ichimura, Hidehiko and Oliver Linton. (2005). "Asymptotic expansions for some semiparametric program evaluation estimators," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*: 149 -170 (D.W.K Andrews & J.H. Stock, Eds). Cambridge: Cambridge University Press.

Imbens, Guido W. (2004). "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics* 86 (1): 4 - 29.

Imbens, Guido W., Whitney K. Newey and Geert Ridder (2005). "Mean-square-error calculations for average treatment effects," *IEPR Working Paper 05.34*.

Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data.* Hoboken, N.J.: John Wiley & Sons, Inc.

Newey, Whitney K. (1990). "Semiparametric efficiency bounds," *Journal of Applied Econometrics* 5 (2): 99 - 135.

Newey, Whitney K. (1994a). "Series estimation of regression functionals," *Econometric Theory* 10 (1): 1 - 28.

Newey, Whitney K. (1994b). "The asymptotic variance of semiparametric estimators," *Econometrica* 62 (6): 1349 - 1382.

Newey, Whitney K. (2004). "Efficient semiparametric estimation via moment restrictions," *Econometrica* 72 (6): 1877 - 1897.

Newey, Whitney K. and Daniel McFadden. (1994). "Large sample estimation and hypothesis testing," *Handbook of Econometrics 4*: 2111 - 2245 (R.F. Engle & D.L. McFadden). Amsterdam: North Holland.

Prokhorov, Artem and Peter J, Schmidt. (2009). "GMM redundancy results for general missing data problems," *Journal of Econometrics* 151 (1): 47 - 55.

Robins, James M., Fushing Hsieh and Whitney Newey. (1995). "Semiparametric efficient estimation of a conditional density function with missing or mismeasured covariates," *Journal of the Royal Statistical Society* B 57 (2): 409 - 424.

Robins, James M. and Andrea Rotnitzky. (1995). "Semiparametric efficiency in multivariate regression models," *Journal of the American Statistical Association* 90 (429): 122 - 129.

Robins, James M. and Andrea Rotnitzky. (2001). "Comments," *Statistica Sinica* 11 (4): 920 - 936.

Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. (1994). "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association* 89 (427): 846 - 866.

Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. (1995). “Analysis of semiparametric regression-models for repeated outcomes in the presence of missing data,” *Journal of the American Statistical Association* 90 (429): 106 - 121.

Scharfstein, Daniel O., Andrea Rotnitzky and James M. Robins. (1999). “Rejoinder,” *Journal of the American Statistical Association* 94 (448): 1135- 1146.

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

Wang, Qihua, Oliver Linton and Wolfgang Härdle. (2004). “Semiparametric regression analysis with missing response at random,” *Journal of the American Statistical Association* 99 (466): 334 - 345.

Wooldridge, Jeffrey M. (1999a). “Asymptotic properties of weighted M-estimators for variable probability samples,” *Econometrica* 67 (6): 1385 - 1406.

Wooldridge, Jeffrey M. (1999b). “Distribution-free estimation of some nonlinear panel data models,” *Journal of Econometrics* 90 (1): 77 - 97.

Wooldridge, Jeffrey M. (2002). “Inverse probability weighted M-estimators for sample selection, attrition and stratification,” *Portuguese Economic Journal* 1 (2): 117 - 139.

Wooldridge, Jeffrey M. (2007). “Inverse probability weighted estimation for general missing data problems,” *Journal of Econometrics* 141 (2): 1281 - 1301.

Efficiency bounds for missing data models with semiparametric restrictions, supplemental material: proofs

This appendix contains proofs of the results contained in the main paper. All notation is as defined in the main text unless explicitly noted otherwise. Equation numbering continues in sequence with that established in the main text. To simplify notation let β denote the true parameter value β_0 unless explicitly stated otherwise (similarly the '0' subscript is removed from other objects, such as the propensity score, when doing so does not cause confusion).

A Proof of Theorem 2.1

The proof closely follows that of Theorem 1 in Chamberlain (1992) and consists of three steps.

Step 1: Demonstration of equivalence with an unconditional GMM problem The first step is to show that restrictions (4) and (5) are, in the multinomial case, equivalent to a finite set of unconditional moment restrictions. Under the multinomial assumption we have $X \in \{x_1, \ldots, x_L\}$ for some L. Let the $L \times 1$ vector B have a 1 in the l^{th} row if $X = x_l$ and zeros elsewhere and $\tau_l = \Pr(X = x_l)$ (observe that $\sum_{l=1}^{L} \tau_l = 1$). Denote the value of the selection probability at $X = x_l$ by ρ_l and define $\rho = \{\rho_1, \ldots, \rho_L\}'$; this vector gives the values of $p(\cdot)$ at each of the mass points of X. Using this notation we can write $p(X) = B'\rho$.

Under the multinomial assumption restrictions (4) and (5) are equivalent to the $L + K \times 1$ vector of unconditional moment restrictions

$$\mathbb{E}\left[m(Z, \beta, \rho)\right] = \mathbb{E}\begin{bmatrix} m_1(Z, \rho) \\ m_2(Z, \beta, \rho) \end{bmatrix} = \mathbb{E}\begin{bmatrix} B\left(\frac{D}{B'\rho} - 1\right) \\ \frac{D}{B'\rho}\psi(Z, \beta) \end{bmatrix} = 0.$$

To verify that this is the case note that by iterated expectations

$$\mathbb{E}\left[m_1(Z, \rho)\right] = \begin{pmatrix} \tau_1 \mathbb{E}\left[\left(\frac{D}{p(X)} - 1\right)\Big| X = x_1\right] \\ \vdots \\ \tau_L \mathbb{E}\left[\left(\frac{D}{p(X)} - 1\right)\Big| X = x_L\right] \end{pmatrix},$$

and hence $\mathbb{E}\left[m_1(Z, \rho)\right] = 0$ if and only if $\mathbb{E}\left[\frac{D}{p(X)} - 1\Big| X\right] = 0$ for all $X \in \{x_1, \ldots, x_L\}$. We also have

$$\mathbb{E}\left[m_2(Z, \beta, \rho)\right] = \mathbb{E}\left[\frac{D}{p(X)}\psi(Z, \beta)\right] = 0,$$

so $\mathbb{E}\left[m(Z, \beta, \rho)\right] = 0$ if and only if (4) and (5) are satisfied as claimed.

Step 2: Application of Lemma 2 of Chamberlain (1987) Chamberlain (1987, Lemma 2) shows that for Z a multinomial random variable the variance bound for β under the sole restriction that $\mathbb{E}\left[m(Z, \beta, \rho)\right] = 0$ is

$$\left\{\left(M'V^{-1}M\right)^{-1}\right\}_{22}$$

where $\left\{\left(M'V^{-1}M\right)^{-1}\right\}_{22}$ is the lower-right $K \times K$ block of $\left(M'V^{-1}M\right)^{-1}$ with

$$V \stackrel{def}{\equiv} \mathbb{E}\left[m(Z, \beta, \rho)\, m(Z, \beta, \rho)'\right], \quad M \stackrel{def}{\equiv} \mathbb{E}\left[\frac{\partial m(Z, \beta, \rho)}{\partial \rho'}, \frac{\partial m(Z, \beta, \rho)}{\partial \beta'}\right].$$

The application of Chamberlain's result requires that M has full column rank and that V is non-singular. The calculations made in Step 3 below demonstrate that these conditions are implied by the assumption that Γ has full column rank, $p(X)$ is bounded away from zero and non-singularity of Λ.

Step 3: Calculation of the bound The final step is to solve for an explicit expression for $\left\{\left(M'V^{-1}M\right)^{-1}\right\}_{22}$. This requires some simple, albeit tedious, algebra. Partitioning V_0

$$\underset{L+K\times L+K}{V} = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}' & V_{22} \end{pmatrix},$$

we have the lower right-hand block, letting $\psi = \psi(Z,\beta)$ and $q(X) = \mathbb{E}[\psi|X]$, given by

$$\begin{aligned} \underset{K\times K}{V_{22}} &= \mathbb{E}\left[m_2(Z,\beta,\rho)\, m_2(Z,\beta,\rho)'\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}\left[\psi\psi'|X\right]}{p(X)}\right] \\ &= \mathbb{E}\left[\frac{\mathbb{V}(\psi|X)}{p(X)} + \frac{1-p(X)}{p(X)} q(X)\, q(X)' + q(X)\, q(X)'\right] \\ &= \sum_{l=1}^{L} \tau_l \left[\frac{\Sigma_l}{\rho_l} + \frac{1-\rho_l}{\rho_l} q_l q_l' + q_l q_l'\right], \end{aligned} \tag{9}$$

where $q_l = \mathbb{E}[\psi(Z,\beta)|x_l]$ and $\Sigma_l = \mathbb{V}(\psi|x_l)$.

The upper right-hand block is similarly derived as

$$\begin{aligned} \underset{L\times K}{V_{12}} &= \mathbb{E}\left[m_1(Z,\beta)\, m_2(Z,\beta,\rho)'\right] \\ &= \mathbb{E}\left[B\left(\frac{D}{B'\rho} - 1\right)\left\{\frac{D\psi(Z,\beta)}{B'\rho}\right\}'\right] \\ &= \mathbb{E}\left[B\left(\frac{1-p(X)}{p(X)} q(X)'\right)\right] \\ &= \begin{pmatrix} \tau_1 \frac{1-\rho_1}{\rho_1} q_1 & \cdots & \tau_L \frac{1-\rho_L}{\rho_L} q_L \end{pmatrix}'. \end{aligned} \tag{10}$$

Finally the upper left-hand block is given by

$$\begin{aligned} \underset{L\times L}{V_{11}} &= \mathbb{E}\left[B\left(\frac{D}{B'\rho} - 1\right)\left(\frac{D}{B'\rho} - 1\right)B'\right] \\ &= \mathbb{E}\left[BB'\left(\frac{1-p(X)}{p(X)}\right)\right] \\ &= diag\left\{\begin{matrix} \tau_1 \frac{1-\rho_1}{\rho_1} & \cdots & \tau_L \frac{1-\rho_L}{\rho_L} \end{matrix}\right\}. \end{aligned} \tag{11}$$

Now partition M

$$\underset{L+K\times L+K}{M} = \begin{pmatrix} M_{1\rho} & 0 \\ M_{2\rho} & M_{2\beta} \end{pmatrix},$$

where, from similar calculations to those made above, we have

$$\underset{L\times L}{M_{1\rho}} = -diag\left\{\begin{matrix} \frac{\tau_1}{\rho_1} & \cdots & \frac{\tau_L}{\rho_L} \end{matrix}\right\}, \quad \underset{K\times L}{M_{2\rho}} = -\begin{pmatrix} \tau_1 \frac{q_1}{\rho_1} & \cdots & \tau_L \frac{q_L}{\rho_L} \end{pmatrix}, \quad \underset{K\times K}{M_{2\beta}} = \Gamma. \tag{12}$$

Applying standard results on partitioned inverses then yields

$$M^{-1} = \begin{pmatrix} M_{1\rho}^{-1} & 0 \\ -M_{2\beta}^{-1} M_{2\rho} M_{1\rho}^{-1} & M_{2\beta}^{-1} \end{pmatrix}.$$

Note that the existence of $M_{1\rho}^{-1}$ and $M_{2\beta}^{-1}$ follows from the assumptions that $p(X)$ is bounded away from zero and the assumption that Γ has full column rank.

Redundancy of knowledge of the propensity score suggests that $M^{-1}VM^{-1\prime}$ will be block diagonal. A sufficient condition for this is that (cf., Prokhorov and Schmidt, 2009)

$$V_{12}' = M_{2\rho} M_{1\rho}^{-1} V_{11}. \tag{13}$$

To verify that this condition holds use (11) and (12) to show that

$$M_{2\rho} M_{1\rho}^{-1} V_{11} = \begin{pmatrix} \tau_1 \frac{1-\rho_1}{\rho_1} q_1 & \cdots & \tau_L \frac{1-\rho_L}{\rho_L} q_L \end{pmatrix},$$

which equals V_{12}' as required. Exploiting the resulting simplifications yields

$$M^{-1} V M^{-1\prime} = \begin{pmatrix} M_{1\rho}^{-1} V_{11} M_{1\rho}^{-1} & 0 \\ 0 & M_{2\beta}^{-1} \left(V_{22} - V_{12}' V_{11}^{-1} V_{12} \right) M_{2\beta}^{-1\prime} \end{pmatrix},$$

and hence

$$\left(M^{-1} V M^{-1\prime} \right)_{22} = M_{2\beta}^{-1} \left(V_{22} - V_{12}' V_{11}^{-1} V_{12} \right) M_{2\beta}^{-1\prime}.$$

By $M_{2\rho} M_{1\rho}^{-1} = (q_1, \ldots, q_L)$ and (13) we have $V_{12}' V_{11}^{-1} V_{12}$ equal to

$$\begin{aligned} V_{12}' V_{11}^{-1} V_{12} &= M_{2\rho} M_{1\rho}^{-1} V_{11} M_{1\rho}^{-1\prime} M_{2\rho}' \\ &= \sum_{l=1}^{L} \tau_l \frac{1-\rho_l}{\rho_l} q_l q_l' \\ &= \mathbb{E}\left[\frac{1-p(X)}{p(X)} q(X) q(X)' \right], \end{aligned}$$

and hence, using (9),

$$V_{22} - V_{12}' V_{11}^{-1} V_{12} = \mathbb{E}\left[\frac{\mathbb{V}(\psi|X)}{p(X)} + q(X) q(X)' \right] = \Lambda.$$

Using this result and taking the partitioned determinant gives

$$\det(V) = \det(V_{11}) \det\left(V_{22} - V_{12}' V_{11}^{-1} V_{12} \right) = \mathbb{E}\left[\frac{1-p(X)}{p(X)} \right] \det\{\Lambda\},$$

and hence V is non-singular under overlap (Assumption 1.4) and non-singularity of Λ.

Since $M_{2\beta} = \Gamma$ we have $\mathcal{I}_{\mathrm{m}}(\beta_0) = \Gamma' \Lambda^{-1} \Gamma$ as claimed. For completeness the upper left-hand portion of the full variance covariance matrix is given by

$$M_{11}^{-1} V_{11} M_{11}^{-1\prime} = \mathcal{I}_{\mathrm{m}}^{-1}(\rho_0) = diag\left\{ \frac{p(x_1)(1-p(x_1))}{f(x_1)}, \cdots, \frac{p(x_L)(1-p(x_L))}{f(x_L)} \right\}$$

where $f(x) = \sum_{l=1}^{L} \tau_l \times \mathbf{1}(x = x_l)$.

B Proof of Theorem 3.1

The first two steps of the proof of Theorem 3.1 are analogous to those of Theorem 2.1 and therefore omitted. The actual calculation of the bound, while conceptually straightforward, is considerably more tedious. Details of this step are provided here.

Assume that the marginal distributions of X_1 and X_2 have I and M points of support with probabilities $\pi_1, \ldots, \pi_I$ and $\varsigma_1, \ldots, \varsigma_M$. Let $L = I \times M$ and τ_{im} denote the joint probability $\Pr\left(X_1 = x_{1,i}, X_2 = x_{2,m}\right)$. Let $\lambda = (\lambda_1, \ldots, \lambda_M)'$ be the values of $h(\cdot)$ at each of the mass points of X_2 (for simplicity I assume that $\dim(h(x_2)) = P = 1$ in the calculations below, but the results generalize). Let C be a $M \times 1$ vector with a 1 in the m^{th} row if $X_2 = x_{2,m}$ and zeros elsewhere. Finally it is convenient to use the shorthand $\Psi = q(X) q(X)'$. In what follows I use both the single and double subscript notation to denote a point on the support of X as is convenient. We can map between the two notations by observing that $x_{im} = x_l$ for $l = (i-1)M + m$.

For the multinomial case the conditional moment problem defined by (4), (5) and (6) is equivalent to the unconditional problem

$$\mathbb{E}\left[m\left(Z,\theta\right)\right]=\mathbb{E}\left[\begin{array}{c} m_1\left(Z,\rho\right) \\ m_2\left(Z,\rho,\lambda,\delta,\beta\right) \\ m_3\left(Z,\rho,\beta\right)\end{array}\right]=0,$$

with $\theta=\left(\rho',\lambda',\delta',\beta'\right)'$ and

$$\underset{L\times 1}{m_1}\left(Z,\rho\right)=B\left(\frac{D}{B'\rho}-1\right),\quad \underset{LK\times 1}{m_2}\left(Z,\rho,\lambda,\delta,\beta\right)=\left(B\otimes I_K\right)\left(\frac{D}{B'\rho}\left(\psi\left(Z,\beta\right)-q(X,\delta,C'\lambda;\beta)\right)\right),$$
$$\underset{K\times 1}{m_3}\left(Z,\rho,\beta\right)=\frac{D}{B'\rho}\psi\left(Z,\beta\right).$$

Partition $V=\mathbb{E}\left[m\left(Z,\theta\right)m\left(Z,\theta\right)'\right]$ as

$$\underset{L+KL+K\times L+KL+K}{V}=\left(\begin{array}{ccc} V_{11} & & \\ V_{21} & V_{22} & \\ V_{31} & V_{32} & V_{34}\end{array}\right),$$

where, using calculations similar to those given in the proof of Theorem 2.1, we have

$$\underset{L\times L}{V_{11}}=diag\left\{\tau_1\frac{1-\rho_1}{\rho_1},\ldots,\tau_L\frac{1-\rho_L}{\rho_L}\right\},\quad \underset{L\times KL}{V_{12}}=\left(\underline{0},\ldots,\underline{0}\right),\quad \underset{KL\times KL}{V_{22}}=diag\left\{\tau_1\frac{\Sigma_1}{\rho_1},\ldots,\tau_L\frac{\Sigma_L}{\rho_L}\right\}$$
$$\underset{K\times L}{V_{31}}=\left(\tau_1\frac{1-\rho_1}{\rho_1}q_1,\ldots,\tau_L\frac{1-\rho_L}{\rho_L}q_L\right),\quad \underset{K\times KL}{V_{32}}=\left(\tau_1\frac{\Sigma_1}{\rho_1},\ldots,\tau_L\frac{\Sigma_L}{\rho_L}\right),\quad \underset{K\times K}{V_{33}}=\sum\nolimits_{l=1}^{L}\tau_l\left[\frac{\Sigma_l}{\rho_l}+\frac{1-\rho_l}{\rho_l}q_lq_l'+q_lq_l'\right].$$

We can partition the Jacobian matrix

$$\underset{L+KL+K\times L+M+J+K}{M}=\left(\begin{array}{cccc} M_{1\rho} & 0 & 0 & 0 \\ 0 & M_{2\lambda} & M_{2\delta} & 0 \\ M_{3\rho} & 0 & 0 & M_{3\beta}\end{array}\right),$$

where

$$\underset{L\times L}{M_{1\rho}}=-diag\left\{\frac{\tau_1}{\rho_1},\ldots,\frac{\tau_L}{\rho_L}\right\},\quad \underset{KL\times M}{M_{2\lambda}}=-\left(H_1',\ldots,H_I'\right)',\quad \underset{KL\times J}{M_{2\delta}}=-\left(\begin{array}{c}\tau_1\nabla_\delta q_1 \\ \vdots \\ \tau_L\nabla_\delta q_L\end{array}\right)$$
$$M_{3\rho}=-\left(\begin{array}{ccc}\tau_1\frac{q_1}{\rho_1} & \cdots & \tau_L\frac{q_L}{\rho_L}\end{array}\right),\quad M_{3\beta}=\Gamma.$$

where $H_i=diag\left\{\tau_{i1}\nabla_h q_{i1},\ldots,\tau_{iM}\nabla_h q_{iM}\right\}$ for $i=1,\ldots,I$ with $q_{im}=q(x_{im},\delta,h\left(x_{2,m}\right);\beta)$.

The variance bound for β is given by the lower right-hand $K\times K$ block of $\left(M'V^{-1}M\right)^{-1}$. We begin by calculating V^{-1}. Partition V

$$V=\left(\begin{array}{cc} B_{11} & B_{12} \\ B_{12}' & B_{22}\end{array}\right),$$

with

$$\underset{L+KL\times L+KL}{B_{11}}=diag\left\{\begin{array}{cc} V_{11} & V_{22}\end{array}\right\},\quad \underset{L+KL\times K}{B_{12}}=\left(\begin{array}{cc} V_{31} & V_{32}\end{array}\right)',\quad B_{22}=V_{33}.$$

Now partition V^{-1} as

$$V_0^{-1}=\left(\begin{array}{cc} C_{11} & C_{12} \\ C_{12}' & C_{22}\end{array}\right), \tag{14}$$

where the partitioned inverse formula gives

$$\underset{L+KL\times L+KL}{C_{11}}=diag\left\{\begin{array}{cc} V_{11}^{-1} & V_{22}^{-1}\end{array}\right\}+D'\mathbb{E}\left[\Psi\right]^{-1}D,\quad \underset{K\times L+KL}{C_{12}'}=-\mathbb{E}\left[\Psi\right]^{-1}D,\quad \underset{K\times K}{C_{22}}=\mathbb{E}\left[\Psi\right]^{-1},$$

with $D = \begin{pmatrix} A' & (\iota_L \otimes I_K)' \end{pmatrix} = B_{12}' B_{11}^{-1}$ and $A = \begin{pmatrix} q_1 & \cdots & q_L \end{pmatrix}'$ a $L \times K$ matrix.

Expression (14) follows since

$$\begin{aligned} C_{22} &= \left(B_{22} - B_{12}' B_{11}^{-1} B_{12} \right)^{-1} \\ &= \sum_{l=1}^{L} \tau_l \left[\frac{\Sigma_l}{\rho_l} + \frac{1-\rho_l}{\rho_l} q_l q_l' + q_l q_l' \right] - \sum_{l=1}^{L} \tau_l \left[\frac{1-\rho_l}{\rho_l} q_l q_l' + \frac{\Sigma_l}{\rho_l} \right] \\ &= \left\{ \sum_{l=1}^{L} \tau_l q_l q_l' \right\}^{-1} = \mathbb{E}\left[\Psi\right]^{-1}. \end{aligned}$$

We also have $C_{12}' = -C_{22} B_{12}' B_{11}^{-1} = -\mathbb{E}\left[\Psi\right]^{-1} D$ and

$$C_{11} = B_{11}^{-1} + B_{11}^{-1} B_{12} C_{22} B_{12}' B_{11}^{-1} = diag\left\{ \begin{matrix} V_{11}^{-1} & V_{22}^{-1} \end{matrix} \right\} + D' \mathbb{E}\left[\Psi\right]^{-1} D.$$

We now evaluate $\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\theta) = M' V^{-1} M$ to

$$\left(\begin{matrix} M_{1\rho}' V_{11}^{-1} M_{1\rho} \\ 0 \\ 0 \\ 0 \end{matrix} \right.$$

$$\begin{matrix} 0 & 0 \\ M_{2\lambda}' \left[V_{22}^{-1} + (\iota_L \otimes I_K) \mathbb{E}\left[\Psi\right]^{-1} (\iota_L \otimes I_K)' \right] M_{2\lambda} & M_{2\lambda}' \left[V_{22}^{-1} + (\iota_L \otimes I_K) \mathbb{E}\left[\Psi\right]^{-1} (\iota_L \otimes I_K)' \right] M_{2\delta} \\ M_{2\delta}' \left[V_{22}^{-1} + (\iota_L \otimes I_K) \mathbb{E}\left[\Psi\right]^{-1} (\iota_L \otimes I_K)' \right] M_{2\lambda} & M_{2\delta}' \left[V_{22}^{-1} + (\iota_L \otimes I_K) \mathbb{E}\left[\Psi\right]^{-1} (\iota_L \otimes I_K)' \right] M_{2\delta} \\ -M_{3\beta}' \mathbb{E}\left[\Psi\right]^{-1} (\iota_L \otimes I_K)' M_{2\lambda} & -M_{3\beta}' \mathbb{E}\left[\Psi\right]^{-1} (\iota_L \otimes I_K)' M_{2\delta} \end{matrix}$$

$$\left. \begin{matrix} 0 \\ -M_{2\lambda}' (\iota_L \otimes I_K) \mathbb{E}\left[\Psi\right]^{-1} M_{3\beta} \\ -M_{2\delta}' (\iota_L \otimes I_K) \mathbb{E}\left[\Psi\right]^{-1} M_{3\beta} \\ M_{3\beta}' \mathbb{E}\left[\Psi\right]^{-1} M_{3\beta} \end{matrix} \right),$$

where I have made use of the equality $M_{1\rho}' A = M_{3\rho}'$.

Observe that, as in the standard semiparametric missing data model, $\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\theta)$ satisfies Stein's condition for redundancy of knowledge of the propensity score for β. However the structure of the bound does indicate that knowledge of the finite dimensional parameters and nonparametric portions of the CEF of $\psi(Z, \beta)$ given X does increase the precision with which β can be estimated.

The variance bound for β_0 is given by the lower right-hand $K \times K$ block of the inverse of this matrix. Because of block diagonality we only need to consider the lower right-hand block. Partition this block as

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{12}' & B_{22} \end{pmatrix},$$

where B_{11}, B_{12} and B_{22} are redefined to equal

$$B_{11} = \begin{pmatrix} M_{2\lambda}' V_{22}^{-1} M_{2\lambda} & M_{2\lambda}' V_{22}^{-1} M_{2\delta} \\ M_{2\delta}' V_{22}^{-1} M_{2\lambda} & M_{2\delta}' V_{22}^{-1} M_{2\delta} \end{pmatrix} + \begin{pmatrix} M_{2\lambda}' (\iota_L \otimes I_K) \\ M_{2\delta}' (\iota_L \otimes I_K) \end{pmatrix} \mathbb{E}\left[\Psi\right]^{-1} \begin{pmatrix} (\iota_L \otimes I_K)' M_{2\lambda} & (\iota_L \otimes I_K)' M_{2\delta} \end{pmatrix}$$

$$B_{12} = \begin{pmatrix} M_{2\lambda}' (\iota_L \otimes I_K) \\ M_{2\delta}' (\iota_L \otimes I_K) \end{pmatrix} \mathbb{E}\left[\Psi\right]^{-1} M_{3\beta}, \qquad B_{33} = M_{3\beta}' \mathbb{E}\left[\Psi\right]^{-1} M_{3\beta}.$$

The information bound is therefore given by

$$\begin{aligned}
\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\theta) &= B_{22} - B_{12}' B_{11}^{-1} B_{12} \\
&= M_{3\beta}' \mathbb{E}[\Psi]^{-1} M_{3\beta} - M_{3\beta}' \mathbb{E}[\Psi_0]^{-1} \begin{pmatrix} (\iota_L \otimes I_K)' M_{2\lambda} & (\iota_L \otimes I_K)' M_{2\delta} \end{pmatrix} \\
&\times \left\{ \begin{pmatrix} M_{2\lambda}' V_{22}^{-1} M_{2\lambda} & M_{2\lambda}' V_{22}^{-1} M_{2\delta} \\ M_{2\delta}' V_{22}^{-1} M_{2\lambda} & M_{2\delta}' V_{22}^{-1} M_{2\delta} \end{pmatrix} \right. \\
&+ \begin{pmatrix} M_{2\lambda}' (\iota_L \otimes I_K) \\ M_{2\delta}' (\iota_L \otimes I_K) \end{pmatrix} \mathbb{E}[\Psi]^{-1} \\
&\times \left. \begin{pmatrix} (\iota_L \otimes I_K)' M_{2\lambda} & (\iota_L \otimes I_K)' M_{2\delta} \end{pmatrix} \right\}^{-1} \begin{pmatrix} M_{2\lambda}' (\iota_L \otimes I_K) \\ M_{2\delta}' (\iota_L \otimes I_K) \end{pmatrix} \mathbb{E}[\Psi]^{-1} M_{3\beta} \\
&= M_{3\beta}' \left[\mathbb{E}[\Psi] + \begin{pmatrix} (\iota_L \otimes I_K)' M_{2\lambda} & (\iota_L \otimes I_K)' M_{2\delta} \end{pmatrix} \right. \\
&\times \left. \begin{pmatrix} M_{2\lambda}' V_{22}^{-1} M_{2\lambda} & M_{2\lambda}' V_{22}^{-1} M_{2\delta} \\ M_{2\delta}' V_{22}^{-1} M_{2\lambda} & M_{2\delta}' V_{22}^{-1} M_{2\delta} \end{pmatrix}^{-1} \begin{pmatrix} M_{2\lambda}' (\iota_L \otimes I_K) \\ M_{2\delta}' (\iota_L \otimes I_K) \end{pmatrix} \right]^{-1} M_{3\beta},
\end{aligned}$$

where I have used the identity $A^{-1} - A^{-1}U(B^{-1} + U'A^{-1}U)^{-1}U'A^{-1} = \left(A + UBU'\right)^{-1}$.

Using the partitioned inverse formula and multiplying out the expression in $[\cdot]$ above then gives

$$\begin{aligned}
\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\theta) = M_{3\beta}' \times [\mathbb{E}[\Psi] + (\iota_L \otimes I_K)' &\Big[M_{2\lambda} \left(M_{2\lambda}' V_{22}^{-1} M_{2\lambda} \right) M_{2\lambda}' \\
&+ \left(M_{2\delta} - M_{2\lambda} \left(M_{2\lambda}' V_{22}^{-1} M_{2\lambda} \right)^{-1} M_{2\lambda}' V_{22}^{-1} M_{2\delta} \right) \\
&\times \left(M_{2\delta}' V_{22}^{-1} M_{2\delta} - M_{2\delta}' V_{22}^{-1} M_{2\lambda} \left(M_{2\lambda}' V_{22}^{-1} M_{2\lambda} \right)^{-1} M_{2\lambda}' V_{22}^{-1} M_{2\delta} \right)^{-1} \\
&\times \left(M_{2\delta} - M_{2\lambda} \left(M_{2\lambda}' V_{22}^{-1} M_{2\lambda} \right)^{-1} M_{2\lambda}' V_{22}^{-1} M_{2\delta} \right)' \Big] (\iota_L \otimes I_K) \Big]^{-1} \times M_{3\beta}.
\end{aligned}$$

We can now use the explicit expressions for V_0 and M_0 give above to generate an interpretable bound. The required calculations are tedious but straightforward (details are available from the author upon request), they give an information bound of $\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\theta)$ as defined in the main text of the paper.

C Proof of Theorem 3.2

In calculating the efficiency bound for the semiparametric missing data model defined by restriction (1) and Assumptions 1.1 to 1.5 above, I follow the general approach outlined by Bickel, Klaassen, Ritov and Wellner (1993) and, especially, Newey (1990, Section 3). First, I characterize the nuisance tangent space. Second, I demonstrate pathwise differentiability of the parameter of interest, β. The efficient influence function for this model equals the projection of the pathwise derivative onto the tangent space. In the present example the direct calculation of this projection appears to be particularly difficult. However inspection of the variance bound associated with the conditional moment problem defined by restrictions (4), (5) and (6) provides a conjecture for the form of the efficient influence function. The third and final step of the proof therefore involves demonstrating that (i) this conjectured influence function lies in the model tangent space and (ii) that it is indeed the required projection (i.e., that it satisfies equation (9) in Newey (1990, p. 106)). The result then follows from an application of Theorem 3.1 in Newey (1990).

Step 1: Characterization of the nuisance tangent space Recalling that $Y = DY_1$, the joint density function for (Y, X, D), making use of Assumption 1.3, is given by

$$f(y, x, d) = f(y_1 | x)^d \, p(x)^d \, [1 - p(x)]^{1-d} f(x).$$

Assumption 1.5 also requires that $f(y_1 | x)$ satisfy the restriction

$$\int \rho(z, \delta_0, h_0(x_2); \beta_0) \, f(y_1 | x) \, \mathrm{d}y_1 = 0,$$

where $\psi(z,\beta) = \psi(x, y_1, \beta)$ and

$$\rho(z, \delta, h(x_2); \beta) = \psi(x, y_1, \beta) - q(x, \delta, h(x_2); \beta).$$

Consider a regular parametric submodel with $f(y, x, d; \eta) = f(y, x, d)$ at $\eta = \eta_0$. The submodel joint density is given by

$$f(y, x, d; \eta) = f(y_1 | x; \eta)^d p(x; \eta)^d [1 - p(x; \eta)]^{1-d} f(x; \eta) \tag{15}$$

and satisfies the restriction

$$\int \rho(z, \delta(\eta), h(x_2; \eta); \beta_0) f(y_1 | x; \eta) \mathrm{d}y_1 = 0. \tag{16}$$

The submodel score vector equals

$$s_\eta(y, x, d; \eta) = d s_\eta(y_1 | x; \eta) + \frac{d - p(x; \eta)}{p(x; \eta)[1 - p(x; \eta)]} \nabla_\eta p(x; \eta) + t_\eta(x; \eta), \tag{17}$$

where

$$s_\eta(y, x, d; \eta) = \nabla_\eta \log f(y, x, d; \eta), \quad s_\eta(y_1 | x; \eta) = \nabla_\eta \log f(y_1 | x; \eta), \quad t_\eta(x; \eta) = \nabla_\eta \log f(x; \eta).$$

By the usual mean zero property of (conditional) scores we have

$$\mathbb{E}[s_\eta(Y_1 | X) | X] = \mathbb{E}[t_\eta(X)] = 0, \tag{18}$$

where suppression of η in a function means that it is evaluated at its population value (e.g., $t_\eta(x) = t_\eta(x; \eta_0)$).

Condition (16) imposes additional restrictions on $s_\eta(Y_1 | X)$ beyond conditional mean zeroness. To see the structure of these restrictions differentiate (16) with respect to η through the integral and evaluate the result at $\eta = \eta_0$:

$$\frac{\partial q_0(X)}{\partial \delta'} \frac{\partial \delta(\eta_0)}{\partial \eta'} + \frac{\partial q_0(X)}{\partial h'} \frac{\partial h(X_2; \eta_0)}{\partial \eta'} = \mathbb{E}\left[\rho(Z, \delta_0, h_0(X_2); \beta_0) s_\eta(Y_1 | X)' \middle| X\right].$$

The conditional covariance between $\rho(Z, \delta_0, h_0(X_2); \beta_0)$ and $s_\eta(Y_1 | X)$ has a particular structure induced by the semiparametric restrictions on the form of $\mathbb{E}[\psi(Z, \beta) | x]$.

From (17), (18) and the above equality the tangent set is evidently

$$\mathcal{T} = \{d s(y_1 | x) + a(x)[d - p(x)] + t(x)\}, \tag{19}$$

where $a(x)$ is unrestricted and $t(x)$ and $s(y_1 | x)$ satisfy

$$\begin{aligned} \mathbb{E}[t(X)] &= 0 \\ \mathbb{E}[s(Y_1 | X) | X] &= 0 \\ \mathbb{E}\left[\rho(Z, \delta_0, h_0(X_2); \beta_0) s(Y_1 | X)' \middle| X\right] &= \left(\frac{\partial q_0(X)}{\partial \delta'}\right) c + \left(\frac{\partial q_0(X)}{\partial h'}\right) k(X_2), \end{aligned}$$

with c a constant matrix and $k(x_2)$ an unrestricted matrix-valued function of x_2.

Step 2: Demonstration of pathwise differentiability Under the parametric submodel $\beta(\eta)$ is identified by the unconditional moment restriction

$$\mathbb{E}_\eta[\psi(Z; \beta(\eta))] = 0.$$

Differentiating under the integral and evaluating at $\eta = \eta_0$ gives

$$\frac{\partial \beta(\eta_0)}{\partial \eta'} = -\Gamma_0^{-1} \mathbb{E}\left[\psi(Z; \beta_0) \frac{\partial \log f(Y_1, X; \eta_0)}{\partial \eta'}'\right].$$

To demonstrate pathwise differentiability of β we require $F(Y, X, D)$ such that

$$\frac{\partial \beta(\eta_0)}{\partial \eta'} = \mathbb{E}\left[F(Y, X, D) s_\eta(Y, X, D)'\right].$$

It is easy to verify that the function

$$F(Y,X,D) = -\Gamma_0^{-1}\left\{\frac{D}{p_0(X)}\rho(Z,\delta_0,h_0(X_2);\beta_0)\right\} + q(X,\delta_0,h_0(X_2);\beta_0),$$

satisfies this condition (cf., Hahn 1998).

Step 3: Verification that conjectured efficient influence function equals the required projection Inspection of the variance bounds associated with the conditional moment problem suggests the candidate efficient influence given by (7) in the main text. I first verify that $\phi_\beta^{\mathrm{f}}(Z,\eta_0,\beta_0)$ lies in the model tangent space. The last term in (7) plays the role of $t(x)$. A zero plays the role of $a(x)[d-p(x)]$. Finally the first two terms in (7) play the role of $ds(y_1|x)$. To see this note that in addition to being both conditionally mean zero we have

$$\begin{aligned}
&\mathbb{E}\left[\rho(Z;\beta_0)\left\{H_0(X_2)\Upsilon_0^h(X_2)^{-1}\left(\frac{\partial q_0(X)}{\partial h'}\right)'\Sigma_0(X)^{-1}\rho(Z;\beta_0)\right.\right.\\
&\left.\left.+\mathbb{E}[G_0(X)]\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\delta_0)^{-1}G_0(X)'\Sigma(X)^{-1}\rho(Z;\beta_0)\right\}'\right|X\Bigg]\\
&=\left(\frac{\partial q_0(X)}{\partial\delta'}\right)c+\left(\frac{\partial q_0(X)}{\partial h'}\right)k(X_2)
\end{aligned}$$

with

$$\begin{aligned}
c &= \mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\delta_0)^{-1}\mathbb{E}[G_0(X)]'\\
k(X_2) &= \Upsilon_0^h(X_2)^{-1}\left\{H_0(X_2)'-\Upsilon_0^{h\delta}(X_2)c\right\}.
\end{aligned}$$

The candidate efficient influence function therefore belongs to the model tangent space as required.

I next show that $\phi_\beta^{\mathrm{f}}(Z,\eta_0,\beta_0)$ is indeed the required projection by verifying that it satisfies

$$\mathbb{E}\left[\left\{F(Y,X,D)-\phi_\beta^{\mathrm{f}}(Z,\eta_0,\beta_0)\right\}\mathsf{t}'\right]=0,\quad \text{for all } \mathsf{t}\in\mathcal{T}$$

(cf., equation (9) in Newey (1990, p. 106)). We have

$$\begin{aligned}
F(Y,X,D)-\phi_\beta^{\mathrm{f}}(Z,\eta_0,\beta_0) = \Gamma_0^{-1}D&\left\{\frac{1}{p_0(X)}-H_0(X_2)\Upsilon_0^h(X_2)^{-1}\left(\frac{\partial q_0(X)}{\partial h'}\right)'\Sigma(X)^{-1}\right.\\
&\left.-\mathbb{E}[G_0(X)]\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\delta_0)^{-1}G_0(X)'\Sigma(X)^{-1}\right\}\rho(Z;\beta_0).
\end{aligned}$$

By the conditional independence of Y_1 and D given X (Assumption 1.3) and conditional mean zeroness of $\rho(Z;\beta_0)$ it is easy to show that $F(Y,X,D)-\phi_\beta^{\mathrm{f}}(Z,\eta_0,\beta_0)$ is orthogonal to any functions of the form $a(x)[d-p(x)]$ and $t(x)$. All that remains is to show orthogonality with $ds(y_1|x)$. We have

$$\begin{aligned}
&\mathbb{E}\left[\left\{F(Y,X,D)-\phi_\beta^{\mathrm{f}}(Z,\eta_0,\beta_0)\right\}Ds(Y_1|X)'\right]\\
&=\mathbb{E}\left[\Gamma_0^{-1}\left\{I_K-H_0(X_2)\Upsilon_0^h(X_2)^{-1}\left(\frac{\partial q_0(X)}{\partial h'}\right)'\left(\frac{\Sigma(X)}{p(X)}\right)^{-1}\right.\right.\\
&\left.-\mathbb{E}[G_0(X)]\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}(\delta_0)^{-1}G_0(X)'\left(\frac{\Sigma(X)}{p(X)}\right)^{-1}\right\}\\
&\left.\times\left\{\left(\frac{\partial q_0(X)}{\partial\delta'}\right)c+\left(\frac{\partial q_0(X)}{\partial h'}\right)k(X_2)\right\}\right],
\end{aligned}$$

where I have made use of the special structure of the conditional covariance $\mathbb{E}\left[\rho(Z;\beta_0)s_\eta(Y_1|X)'\middle|X\right]$. Multiplying out terms yields

$$
\begin{aligned}
&\mathbb{E}\left[\left\{F(Y,X,D)-\phi_{\beta}^{\mathrm{f}}\left(Z,\eta_0,\beta_0\right)\right\}Ds\left(Y_1|\,X\right)\right]\\
&=\Gamma_0^{-1}\mathbb{E}\left[\left\{\frac{\partial q_0\left(X\right)}{\partial\delta'}c+H_0\left(X_2\right)k\left(X_2\right)\right.\right.\\
&-H_0\left(X_2\right)\Upsilon_0^h\left(X_2\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right)c-H_0\left(X_2\right)k\left(X_2\right)\\
&-\mathbb{E}\left[G_0\left(X\right)\right]\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}\left(\delta_0\right)^{-1}\Upsilon_0^{\delta}\left(X_2\right)c\\
&+\mathbb{E}\left[G_0\left(X\right)\right]\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}\left(\delta_0\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right)'\Upsilon_0^h\left(X_2\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right)c\\
&-\mathbb{E}\left[G_0\left(X\right)\right]\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}\left(\delta_0\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right)'k\left(X_2\right)\\
&\left.+\mathbb{E}\left[G_0\left(X\right)\right]\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}\left(\delta_0\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right)'k\left(X_2\right)\right]\\
&=\Gamma_0^{-1}\left\{\mathbb{E}\left[G_0\left(X\right)\right]c-\mathbb{E}\left[G_0\left(X\right)\right]c\right\}=0,
\end{aligned}
$$

where the first equality follows from iterated expectations and the second from the definitions of $G_0\left(X\right)$ and $\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}\left(\delta_0\right)$ in the main text.

The result then follows from an application of Theorem 3.1 in Newey (1990).

D Proof of Proposition 3.1

The difference in the variance bounds is given by

$$
\mathcal{I}_{\mathrm{m}}\left(\beta_0\right)^{-1}-\mathcal{I}_{\mathrm{m}}^{\mathrm{f}}\left(\beta_0\right)^{-1}=\Gamma_0^{-1}\left(\Lambda_0-\Xi_0\right)\Gamma_0^{-1\prime},
$$

with Λ_0 and Ξ_0 as defined in the main text.

First observe that $\mathbb{E}\left[G_0\left(X\right)\right]$ has the covariance representation

$$
\mathbb{E}\left[G_0\left(X\right)\right]=\mathbb{E}\left[\frac{\partial q_0\left(X\right)}{\partial\delta'}-\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)\Upsilon_0^h\left(X_2\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right)\right]=\mathbb{C}\left(\xi_1,\xi_2'\right),
$$

with ξ_1 and ξ_2 as defined in the main text. This follows since

$$
\begin{aligned}
&\mathbb{E}\left[\frac{D}{p_0\left(X\right)}\rho\left(Z;\beta_0\right)\rho\left(Z;\beta_0\right)'\Sigma_0\left(X\right)^{-1}\left\{\frac{\partial q_0\left(X\right)}{\partial\delta'}-\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)\Upsilon_0^h\left(X_2\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right)\right\}\right]\\
&\qquad=\mathbb{E}\left[\frac{\partial q_0\left(X\right)}{\partial\delta'}-\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)\Upsilon_0^h\left(X_2\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right)\right],
\end{aligned}
$$

and also

$$
\begin{aligned}
&\mathbb{E}\left[DH_0\left(X_2\right)\Upsilon_0^h\left(X_2\right)^{-1}\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)'\Sigma_0\left(X\right)^{-1}\rho\left(Z;\beta_0\right)\rho\left(Z;\beta_0\right)'\Sigma_0\left(X\right)^{-1}\right.\\
&\qquad\left.\times\left\{\frac{\partial q_0\left(X\right)}{\partial\delta'}-\left(\frac{\partial q_0\left(X\right)}{\partial h'}\right)\Upsilon_0^h\left(X_2\right)^{-1}\Upsilon_0^{h\delta}\left(X_2\right)\right\}\right]=0.
\end{aligned}
$$

Similar calculations yield the variance representations

$$
\mathbb{V}\left(\xi_1\right)=\mathbb{E}\left[\frac{\Sigma_0\left(X\right)}{p_0\left(X\right)}-H_0\left(X_2\right)\Upsilon_0^h\left(X_2\right)^{-1}H_0\left(X_2\right)'\right],\quad\mathbb{V}\left(\xi_2\right)=\mathbb{E}\left[DG_0\left(X\right)'\Sigma_0\left(X\right)^{-1}G_0\left(X\right)\right],
$$

with the result directly following.

E Proof of Proposition 3.2

Part (i) follows from Theorem 3.2. For part (ii) condition (a) implies the equality.

$$\mathbb{E}\left[DH_0(X_2)\Upsilon_0^h(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)'\Sigma_0(X)^{-1}\rho_*(Z;\beta_0)\middle| X_2\right]$$
$$= H_0(X_2)\mathbb{E}\left[\left(\frac{\partial q_*(X)}{\partial h'}\right)'\Sigma_0(X)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\middle| X_2\right]^{-1}\mathbb{E}\left[\left(\frac{\partial q_*(X)}{\partial h'}\right)'\Sigma_0(X)^{-1}\rho_*(Z;\beta_0)\middle| X_2\right].$$

Condition (b) implies that $\Sigma_0(X) = \Phi_0(X_2)$. Let $L(X_2)L(X_2)' = \Phi_0(X_2)$ be the Cholesky decomposition of $\Phi_0(X_2)$. This implies that the term to the right of the last equality equals

$$H_0(X_2)\mathbb{E}\left[\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}'\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}\middle| X_2\right]^{-1}\mathbb{E}\left[\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}'L(X_2)^{-1}\rho_*(Z;\beta_0)\middle| X_2\right].$$

Since all expectations in the above expression condition on X_2, $L(X_2)$ may be treated as non-stochastic so that

$$L(X_2)^{-1}H_0(X_2) = \mathbb{E}\left[L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\middle| X_2\right].$$

Recall that a linear predictor passes through the mean of the outcome variable at the means of the predictor variables (when a constant is included). Condition (c) implies that each row of $\partial q_*(X)/\partial h'$ includes such a constant and hence that

$$\begin{aligned}L(X_2)^{-1}\mathbb{E}\left[\rho_*(Z;\beta_0)|X_2\right] = L(X_2)^{-1}H_0(X_2)&\\ \times\mathbb{E}\left[\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}'\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}\middle| X_2\right]^{-1}&\\ \times\mathbb{E}\left[\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}'\left\{L(X_2)^{-1}\rho(Z;\beta_0)\right\}\middle| X_2\right],&\end{aligned}$$

and therefore that

$$\mathbb{E}\left[DH_0(X_2)\Upsilon_0^h(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)'\Sigma_0(X)^{-1}\rho_*(Z;\beta_0)\middle| X_2\right] = \mathbb{E}\left[\rho_*(Z;\beta_0)\right] = -\mathbb{E}\left[q_*(X)\right].$$

This implies that the first part of $\phi_\beta^{\mathrm{f}}(Z,\eta_*,\beta_0)$ has mean $-\mathbb{E}\left[q_*(X)\right]$.

Using conditions (a), (b), (c), and arguments analogous to those given immediately above we have

$$\begin{aligned}L(X_2)^{-1}G_0(X) = L(X_2)^{-1}\frac{\partial q_0(X)}{\partial\delta'}&\\ - L(X_2)^{-1}\left(\frac{\partial q_0(X)}{\partial h'}\right)\mathbb{E}\left[\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}'\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}\middle| X_2\right]^{-1}&\\ \times\mathbb{E}\left[\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}'\left\{L(X_2)^{-1}\frac{\partial q_0(X)}{\partial\delta'}\right\}\middle| X_2\right],&\end{aligned}$$

so that $\mathbb{E}\left[L(X_2)^{-1}G_0(X)\middle| X_2\right] = L(X_2)^{-1}\mathbb{E}\left[G_0(X)|X_2\right] = 0$. The law of iterated expectations then gives $\mathbb{E}\left[G_0(X)\right] = 0$. This implies that the second part of $\phi_\beta^{\mathrm{f}}(Z,\eta_*,\beta_0)$ is mean zero. The third part of $\phi_\beta^{\mathrm{f}}(Z,\eta_*,\beta_0)$ has mean $\mathbb{E}\left[q_*(X)\right]$. The result follows as claimed.

F Proof of Corollary 3.1

From the proof to Proposition 3.2 we have $\mathbb{E}[G_0(X)] = 0$. So the result follows if

$$\mathbb{E}\left[\frac{\Sigma_0(X)}{p_0(X)} - H_0(X_2)\Upsilon_0^h(X_2)^{-1}H_0(X_2)'\right] = 0.$$

Under conditions (a) and (b) of part (ii) of Proposition 3.2 we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\Sigma_0(X)}{p_0(X)} - H_0(X_2)\Upsilon_0^h(X_2)^{-1}H_0(X_2)'\right] \\
&= \mathbb{E}\left[\frac{\Phi_0(X_2)}{e_0(X_2)} - \frac{1}{e_0(X_2)}H_0(X_2)\mathbb{E}\left[\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}'\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}\middle| X_2\right]^{-1}H_0(X_2)'\right] \\
&= \mathbb{E}\left[\frac{\Phi_0(X_2)}{e_0(X_2)} - \frac{L(X_2)}{e_0(X_2)}\mathbb{E}\left[L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\middle| X_2\right]\right. \\
&\left.\times\mathbb{E}\left[\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}'\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}\middle| X_2\right]^{-1}\mathbb{E}\left[L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\middle| X_2\right]'L(X_2)'\right],
\end{aligned}$$

where $L(X_2)L(X_2)' = \Phi_0(X_2)$ as above. Observe that

$$\begin{aligned}
&\mathbb{E}\left[L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\middle| X_2\right] \\
&\times\mathbb{E}\left[\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}'\left\{L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\right\}\middle| X_2\right]^{-1}\mathbb{E}\left[L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\middle| X_2\right]',
\end{aligned}$$

is equal to the multivariate conditional linear predictor of the $K \times K$ identity matrix given $L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)$ evaluated at $\mathbb{E}\left[L(X_2)^{-1}\left(\frac{\partial q_*(X)}{\partial h'}\right)\middle| X_2\right]$; therefore this object equals I_K and we have

$$\mathbb{E}\left[\frac{\Sigma_0(X)}{p_0(X)} - H_0(X_2)\Upsilon_0^h(X_2)^{-1}H_0(X_2)'\right] = \mathbb{E}\left[\frac{\Phi_0(X_2)}{e_0(X_2)} - \frac{L(X_2)L(X_2)'}{e_0(X_2)}\right] = 0,$$

as required.

References

Bickel, Peter J., Chris A.J. Klaassen, Ya'acov Ritov and Jon A. Wellner. (1993). *Efficient and adaptive estimation for semiparametric models.* New York: Springer-Verlag, Inc.

Chamberlain, Gary. (1987). "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics* 34 (1): 305 - 334.

Chamberlain, Gary. (1992). "Comment: sequential moment restrictions in panel data," *Journal of Business and Economic Statistics* 10 (1): 20 - 26.

Hahn, Jinyong. (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica* 66 (2): 315 - 331.

Newey, Whitney K. (1990). "Semiparametric efficiency bounds," *Journal of Applied Econometrics* 5 (2): 99 - 135.