Efficient GMM estimation with incomplete data

Chris Muris*

July 1, 2016

Abstract

The standard missing data model classifies data in terms of "binary missingness", that is, as either complete or completely missing. Thus, the model deals with two strata of missingness. However, applied researchers face situations with an arbitrary number of strata of incompleteness. Examples include unbalanced panels, and instrumental variables settings where some observations are missing some instruments, and other observations are missing different instruments. In this paper, I propose a model for settings where observations may be incomplete, with an arbitrary number of strata of incompleteness. I derive a set of moment conditions that generalizes those in Graham

[*] Previous title: "Efficient GMM estimation with general missing data patterns". Department of Economics, Simon Fraser University. Email: cmuris@sfu.ca. I am grateful to Ramon van den Akker, Richard Blundell, Otilia Boldea, Irene Botosaru, Pedro Duarte Bom, Katherine Carman, Miguel Atanasio Carvalho, Toru Kitagawa, Tobias Klein, Andrea Krajina, Jan Magnus, Bertrand Melenberg, Krishna Pendakur, Franco Peracchi, Pedro Raposo, Thomas Vigie, Bas Werker, and Frank Windmeijer for encouraging and insightful discussions. I also thank the seminar participants at Tilburg University, University of Bristol, Institute of Advanced Studies Vienna, Simon Fraser University, Monash University, and Victoria University. I gratefully acknowledge financial support from the Social Sciences and Humanities Research Council through Insight Development Grant 430-2015-00073.

(2011) for the standard missing data setup with two strata. I derive the associated efficiency bound, and propose estimators that attain it. Incompleteness is qualitatively different from binary missingness. In particular, I show that identification can be achieved even if it fails in each stratum of incompleteness.

1 Intro

Incomplete data, where some observations are missing some or all variables, is prevalent in empirical research in economics. For example, Abrevaya and Donald (2015) find that incomplete data occurs in at least 40% of the publications in top economics journals. In 70% of these cases, all incomplete observations are discarded, and the analysis is then carried out with the resulting complete subsample. This strategy fails to exploit all the information in the data, since incomplete observations typically have "some" information about model parameters. This paper shows how to use this information.

I provide a general framework for efficient parameter estimation using incomplete data. To see why a serious treatment of incomplete observations can be useful, consider a linear instrumental variables model with two endogenous variables $X = (X_1, X_2)$ and two instruments W_1 and W_2. The parameter vector β_0 is defined through the moment conditions:

$$E \begin{pmatrix} W_1 \left(y - X\beta_0\right) \\ W_2 \left(y - X\beta_0\right) \end{pmatrix} = 0. \tag{1.1}$$

Now consider a setting where either instrument can be missing. This implies the existence of three strata based on data availability. In the first stratum, both instruments W_1 and W_2 are observed; in stratum 2, only the instrument W_1 is observed; in stratum 3, only W_2 is observed. Although the parameter is not identified in stratum 2, the moment condition $E\left[W_1 \left(y - X\beta_0\right)\right] = 0$ still contains information on β_0. The same is true for stratum 3 through

$E\left[W_2\left(y - X\beta_0\right)\right] = 0$. This paper provides an efficient estimator which uses information from all strata. The approach is general: it allows for arbitary number of strata of incompleteness, and an arbitrary set of nonlinear moment conditions.

Currently available procedures for dealing with incomplete data can be classified into three categories. The first approach is to classify data (or equivalently, moments) in terms of "binary missingness", that is as either complete or completely missing. A second approach is to provide tools that work only in specific applications. The third approach is to impute the missing data.

The approach proposed here is distinct from all of those. I focus on incomplete data that may be partially missing; whereas binary missingness implies the existence of exactly two strata, I allow for an arbitrary finite number of strata based on the availability of each and every moment. My approach accommodates any model that can be expressed in terms of moment conditions. In contrast, model-specific solutions for one type of application may not be useful for another. My approach does not use imputation. Imputation approaches have the obvious drawback that they are inconsistent if the imputation model is misspecified. My approach is consistent, in part, because it does not use an imputation model.

This paper has three methodological contributions. First, I generalize the moment conditions established by Graham (2011) for the binary missingness case to the general incompleteness case. The resulting set of moment conditions consists of one set of Graham's moment conditions for each stratum of incompleteness.

Second, I derive the efficiency bound associated with the complete set of moment conditions, and propose an estimator that attains that bound. I provide conditions under which the estimator is consistent and asymptotically normal. A simulation study shows that the efficiency gain from using incomplete observations can be substantial.

Third, I show that the parameters of interest can be identified by using all the available data, even if identification does not hold in every stratum. As an example, consider a linear IV model with two endogenous variables and two instruments, where the instruments are never observed in the same stratum, but each instrument is available from a different stratum. In this setting, one can still identify the regression parameters.

The results in this paper can also be applied to: (dynamic) panel data models; equation systems where some equations have missing dependent variables for some observations; triangular simultaneous systems with some endogenous explanatory variables missing for some observations; and general nonlinear instrumental variables models. In Section 6.2, I analyze a dynamic panel data model where cross section units may miss observations in any combination of time periods.

The paper is organised as follows. Section 2 provides a literature review. Section 3 describes the model. Section 4 presents the efficiency bound results, and Section 5 presents efficient estimators. Section 6 contains a simulation study.

2 Related literature

The literature on missing data is vast. I discuss the relevant literature in three strands. The first strand considers efficient estimation under the assumption that every observation is either complete or completely missing. The second strand of literature considers estimation with incomplete observations for specific models. The third strand of literature augments incomplete observations using imputation. To the best of my knowledge, my paper is the first that provides a general framework for efficient estimation with incomplete observations without using imputation.

To facilitate this discussion, let p be the number of elements in a moment vector ψ, and let D be a $p \times p$ diagonal matrix with 1 on the main diagonal if a moment is observed, and 0 otherwise. The missing data indicator D thus defines the strata of incompleteness in the data, and the vector $D\psi$ gives the observed elements of ψ. In the linear IV example given above, $p = 2$, and the 2×2 matrix D can take three values corresponding to zeros and ones on the main diagonal. The three values that D can take on correspond to the three strata of data incompleteness. We say a parameter is identified in a stratum D if $D\psi$ contains sufficient information to identify the parameter. In the example above, the only stratum in which the parameter is identified is stratum 1 (for which $D = I_2$).

Strand 1: Binary missingness. There is an extensive literature on missing data models in which each observation contributes either to all, or to none of the sample moments (i.e. the missing data indicator is a binary variable). This literature typically employs the "missing at random" (MAR) assumption. I will call models including a MAR assumption the MAR setup (as in Graham, 2011, p. 438).

The literature on the MAR setup was initiated by Robins et al. (1994), who propose an augmented inverse propensity score weighting (AIPW) procedure. An overview of the AIPW literature in statistics can be found in Tsiatis (2006). Chen et al. (2008) derive the efficiency bound for nonlinear and possibly overidentified models, and propose an efficient estimator for the parameters in the MAR setup that is not based on inverse propensity score weighting (IPW). An important result in this literature is that estimating the propensity score is more efficient than using the true value of the propensity score ("the IPW paradox", see e.g. Hirano et al. 2003; Wooldridge, 2007; Prokhorov and Schmidt, 2009).

Two contributions from this literature that are especially relevant for the discussion in this paper are Graham (2011) and Cattaneo (2010). Graham (2011) shows, in a MAR setup with binary missingness (just 2 strata for D), that the efficiency bound is equivalent to the efficiency bound for the inverse weighted moment conditions of the original (complete data) model plus a set of conditional moment conditions that captures all the information from the MAR assumption. I generalize the moment conditions established by Graham (2011) for the binary missingness case to the general incompleteness case with J strata.

Cattaneo (2010) considers a similar problem to general incompleteness in the context of multi-level program evaluation. Program evaluation models can be thought of as incomplete data, where the incompleteness takes the form of missing dependent variables. With multilevel program evaluation, this incompleteness implies many strata (as many as there are levels of treatment). Cattaneo shows how to optimally combine the estimators from those moment conditions, but his approach requires that the parameter vector is identified in each and every stratum. Consequently, his approach cannot be used for the linear IV example given above. I provide sufficient conditions for an optimal estimator when the parameter vector is identified in just one stratum. Further, I provide special cases where identification is not required in any stratum.

Strand 2: Model-specific solutions. Several papers consider specific GMM settings or specific missing data patterns. Model-specific solutions are also available for the instrumental variable model with incomplete sets of instruments. The problem of partially missing instruments is common; see for example Angrist et al. (2010). Instrumental variables estimation with missing instruments is discussed in Mogstad and Wiswall (2012), who consider a setting with a single instrument that is missing for a subsample of the observations. Abrevaya and Donald (2011) also consider the missing instrument model.

Chen et al. (2010) provide an estimator for the parameters in a static panel data model. Verbeek and Nijman (1992) also considers the static model, and propose to use the different missing data patterns to test for selectivity bias. Hirano et al. (2001) consider a panel data model with three missing data patterns.[1] Abrevaya (2016) shows that the explanatory variables in the static model have information even when the associated dependent variable is missing.

The linear dynamic panel data model with attrition has recently been considered by Pacini and Windmeijer (2015), see also the references therein. Pacini and Windmeijer (2015) show that nonlinear, previously not considered moment conditions are informative when time periods are missing.

My approach accommodates any model that can be expressed in terms of moment conditions, and allows for any structure of incompleteness. In contrast, model-specific solutions restrict the structure of incompleteness, and solutions for one type of application may not be useful for another.

Strand 3: Imputation. There is a substantial literature that considers augmenting incomplete observations by imputing the missing components. A leading example is the linear regression model with missing covariates. Using variables that are always observed, an imputation model can be estimated using the complete observations, and it can then be used to fill in the incomplete observations. Early contributions to the econometric literature on this topic can be found in Dagenais (1973) and Gourieroux and Monfort (1981). To retain consistency, these approaches require a correctly specified imputation model. Such an assumption is not maintained in the model that I consider.[2]

In the context of linear IV example above, imputation would apply to missing instruments. If the imputation were correctly specified, then imputation would not result in bias, and would improve efficiency of the estimator. However, under misspecification, the resulting estimator would typically be biased. My approach does not use imputation, and results in a consistent

[1] An observation is either complete, is subject to attrition, or is part of a refreshment sample.

[2] A more recent contribution by Dardanoni et al. (2011) shows that efficiency gains can be obtained if one is willing to sacrifice consistency. Abrevaya and Donald (2015) propose an GMM estimator that is consistent and efficient under the assumptions required for the consistency of imputation methods.

estimator.

3 Model

The starting point is a GMM framework with complete data. Let $Z = \left(Y_1', X'\right)$ be a random vector, let β be an unknown parameter vector of size $K \times 1$, and let $\psi\left(Z, \beta\right)$ be a $p \times 1$ vector of moment functions, with $p \geq K$.[3] The true value of the parameter, $\beta_0 \in \mathcal{B} \subset \mathbb{R}^K$, is defined by:

Assumption 1. $E\left(\psi\left(Z, \beta\right)\right) = 0 \Leftrightarrow \beta = \beta_0$.

This paper analyzes a framework in which not all elements of the vector $\psi\left(Z, \beta\right)$ are always observable. To model this, let D be a missing data indicator with $J + 1$ outcomes $\{d_1, \cdots, d_{J+1}\}$. Every outcome of D corresponds to a stratum that is defined by data availability. A missing data pattern d_j is an $r \times r$ selection matrix that selects the elements of ψ that are observable for an observation in stratum j. In other words, the researcher observes $D\psi\left(Z, \cdot\right)$. In stratum $J + 1$, none of the components of ψ are observed: $d_{J+1} = O_{p \times p}$.

To fix ideas, consider a linear instrumental variables (IV) model with two regressors $X = (X_1, X_2)$ and two instruments W_1 and W_2. The parameter of interest is defined through the moment conditions

$$E\left(\begin{array}{c} W_1\left(y - X\beta_0\right) \\ W_2\left(y - X\beta_0\right) \end{array}\right) = 0. \tag{3.1}$$

The missing data indicator takes one of $J + 1 = 4$ values

$$D \in \left\{ d_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, d_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, d_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, d_4 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right\}.$$

[3] Wherever possible, I will be using the notation in Graham (2011) to facilitate a comparison with the missing at random setup in that paper.

For d_1, this corresponds to observing all variables:

$$d_1\psi(Z,\beta) = \begin{pmatrix} W_1(y - X\beta) \\ W_2(y - X\beta) \end{pmatrix}$$

for any value of β. In the stratum with $D = d_2$, only the instrument W_1 is available. This corresponds to observing

$$d_2\psi(Z,\beta) = \begin{pmatrix} W_1(y - X\beta_0) \\ 0 \end{pmatrix}.$$

Similarly, in the stratum with $D = d_3$, only the second instrument W_2 is observed. Finally, stratum 4 corresponds to the observations for which both instruments are missing, or for which the dependent or one of the regressors is missing.

Assumption 2. *Consider the following assumptions on the joint distribution of $Z = (Y_1, X)$ and D, and on the sampling process:*

1. *Random sampling: $\{(Z_i, D_i), i = 1, \cdots, n\}$ is an independent and identically distributed sequence;*

2. *Observed data: The researcher observes D_i, X_i, and $D_i\psi(Z_i,\beta)$ for all $\beta \in \mathcal{B}$;*

3. *Missing at random: $Y_1 \perp D|\,X$;*

4. *Overlap: There exists a $\kappa > 0$ such that*

$$p_{j,0}(x) = P(D = d_j|\,X = x) \geq \kappa$$

for all $j = 1, \cdots, J+1$ and for all $x \in supp(X)$.

Assumption 2.3 is standard in the literature on missing data and program evaluation. In particular, Assumptions 1 and 2 reduce to the standard missing at random (MAR) setup if $J = 1$ and $d_1 = I_p$, the case in which an observation is either complete or completely missing (see e.g. Graham, 2011).

Assumption 2.2 formalizes the way incompleteness is modelled in this paper. Note that any combination of elements of ψ may be missing. This flexibility is particularly convenient when modelling panel data with selection and attrition. Then, the framework proposed here can accommodate cross-section units that appear and drop out at any time during the sample time period, possibly multiple times. An example of this flexibility can be found in the dynamic panel data example in Section 6.2.

Assumptions 2.1 and 2.4 are the analogs of standard random sampling and overlap assumptions in the literature on missing data and program evaluation.

Assumption 3. *Every component of ψ is observable in at least one stratum, so that the matrix $\sum_{j=1}^{J} d_j$ has full rank.*

Assumption 3 guarantees identification for the incomplete data setting if identification holds in the complete data setting. It rules out situations in which a component of ψ is never observed. If Assumption 3 fails, the analysis may proceed after removing the never-observed component from ψ, as long as Assumption 1 holds for the modified moment conditions.

Assumption 3 can hold even if identification fails in every stratum. As an example, consider the moment conditions for bivariate mean estimation,

$$E\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right) = 0, \tag{3.2}$$

and assume that only one random variable is available for any observation, i.e. $J = 2$ and

$$D \in \left\{ d_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, d_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, d_3 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right\}.$$

In this case, the parameter $\mu = (\mu_1, \mu_2)$ is identifiable, even though ψ is only ever observed partially.

4 Information

In this section, I first derive a set of moment conditions implied by Assumptions 1, 2, and 3 of Section 3. Second, I derive the efficiency bound associated with those moment conditions. I highlight the implications for the linear instrumental variables model. As in the previous section, my notation here follows that in Graham (2011) to facilitate a comparison with the standard MAR setup.

Existing results. For the complete data problem (Assumption 1), the maximum asymptotic precision with which β_0 can be estimated is given by

$$I_f(\beta_0) = \Gamma_0' \Omega_0^{-1} \Gamma_0, \tag{4.1}$$

where Γ_0 is the expected derivative, and Ω_0 is the variance of ψ at β_0, i.e.

$$\begin{aligned} \Gamma_0 &= E\left[\left. \frac{\partial \psi(Z,\beta)}{\partial \beta} \right|_{\beta=\beta_0} \right], && (4.2) \\ \Omega_0 &= E\left[\psi(Z,\beta_0)\, \psi(Z,\beta_0)' \right]. && (4.3) \end{aligned}$$

The bound in (4.1) was established by Chamberlain (1987).

For the standard MAR setup (Assumptions 1,2,3, with $J = 1$ and $d_1 = I_p$), it is given by

$$I_m(\beta_0) = \Gamma_0' \Lambda_{m,0}^{-1} \Gamma_0, \tag{4.4}$$

where

$$\begin{aligned}
\Lambda_{m,0} &= E\left[\frac{\Sigma_0(X)}{p_{1,0}(X)} + q(X)\,q(X)'\right], && (4.5)\\
\Sigma_0(X) &= V\left[\psi(Z,\beta_0)|\,X\right], && (4.6)\\
q(X) &= E\left[\psi(Z,\beta_0)|\,X\right], && (4.7)
\end{aligned}$$

see for example Robins et al. (1994), Chen et al. (2008) and Graham (2011).

Graham (2011, Theorem 2.1) shows that the expression in (4.4) is equivalent to the information bound for β_0 under the moment conditions

$$\begin{aligned}
E\left[\frac{1\{D=I_p\}}{p_{1,0}(X)}\psi(Z,\beta_0)\right] &= 0, && (4.8)\\
E\left[\left.\frac{1\{D=I_p\}}{p_{1,0}(X)} - 1\right| X\right] &= 0. && (4.9)
\end{aligned}$$

4.1 Moment conditions and efficiency bound

In the incomplete data case, a set of moment conditions similar to 4.8 and 4.9 holds in each stratum. For each $j \in \{1,\cdots,J\}$, define the stratum indicator

$$s_j = 1\{D=d_j\}.$$

For an arbitrary j, the analog of the conditional moment conditions (4.9) is

$$E\left[\left.\frac{s_j}{p_{j,0}(X)} - 1\right| X\right] = 0 \text{ for all } x \in \operatorname{supp}(X). \tag{4.10}$$

This moment condition defines the selection probabilities $p_{j,0}(X)$.

Furthermore, the conditional moment conditions

$$E\left[\left.\left(\frac{s_j}{p_{j,0}(X)} - 1\right) d_j\psi(Z,\beta_0)\right| X\right] = 0 \tag{4.11}$$

hold for each stratum j, and for every value of X.[4] However, since

$$E\left[\left(\frac{s_j}{p_{j,0}(X)}-1\right)\xi(Z)\middle| X\right]=0 \tag{4.12}$$

for any function ξ, those moment conditions cannot be used directly. However, its unconditional implication

$$E\left[\frac{s_j}{p_{j,0}(X)}d_j\psi(Z,\beta_0)\right]=0, \tag{4.13}$$

is informative, and is the analog of equation (4.8).

Given data availability Assumption 2.2, the sample analogs of moment conditions (4.10) and (4.13) are available for each j. Stacking these moment conditions across j obtains

$$\begin{aligned} E\left[\begin{array}{c} \frac{s_1}{p_{1,0}(X)}-1 \\ \vdots \\ \frac{s_J}{p_{J,0}(X)}-1 \end{array}\middle| X\right] &= 0, && (4.14)\\ E\left[\left[\begin{array}{c} \frac{s_1}{p_{1,0}(X)}d_1 \\ \vdots \\ \frac{s_J}{p_{J,0}(X)}d_J \end{array}\right]\psi(Z,\beta_0)\right] &= 0. && (4.15) \end{aligned}$$

The following Theorem presents the efficiency bound under moment condi-

[4] To see this, note that

$$\begin{aligned} E\left[\left(\frac{s_j}{p_{j,0}(X)}-1\right)\psi(Z,\beta_0)\middle| X\right] &= E\left[\frac{s_j}{p_{j,0}(X)}-1\middle| X\right]E\left[\psi(Z,\beta_0)| X\right]\\ &= 0q(X;\beta_0)=0, \end{aligned}$$

where the first equality follows from Assumption 2.3 (MAR), and the second equality follows from (4.10).

tions (4.14) and 4.15. To state the result, define

$$\begin{aligned}
\Lambda_0 &= E\left[R_0^{-1}(X)\otimes\Sigma_0(X)+\iota_J\iota_J^{'}\otimes q(X)\,q(X)^{'}\right], && (4.16)\\
R_0(X) &= \operatorname{diag}\left(p_{1,0}(X),\cdots,p_{J,0}(X)\right), && (4.17)\\
\Delta_2 &= \begin{bmatrix} d_1\\ \vdots\\ d_J\end{bmatrix}. && (4.18)
\end{aligned}$$

Theorem 4. *Assume that (i) the distribution of Z has known, finite support; (ii) there exists a β_0 and*

$$\rho_0=\left(\rho_{11,0},\cdots,\rho_{1L,0},\rho_{21,0},\cdots,\rho_{JL,0}\right),$$

with $\rho_{jl,0}=p_{j,0}(x_l)$ for all j,l, with $\{x_1,\cdots,x_L\}$ the support of X, such that moment conditions (4.14) and 4.15 hold; (iii) Λ_0 is invertible and Γ_0 has full rank; (iv) other regularity conditions hold (see e.g. Chamberlain (1992, Section 2). Then the Fischer information bound for β_0 is given by

$$I(\beta_0) = \Gamma_0^{'}\Delta_2^{'}\Lambda_0^{-1}\Delta_2\Gamma_0. \qquad (4.19)$$

Proof. See Appendix A.1. □

The statement of the result mirrors Graham (2011, Theorem 2.1). In particular, see Graham (2011, p. 442) and Chamberlain (1992, Section 2) for a discussion about condition (i), which assumes that Z has known, finite support. Since any distribution can be approximated arbitrary closely by a multinomial, the bound in (4.19) applies to the general (non-multinomial) problem.

When $J=1$, $d_1=I_p$, the expressions above simplify to those for the standard MAR setup. Note that

$$\Lambda_0 = \Lambda_{m,0} = E\left[\frac{\Sigma_0(X)}{p_{10}(X)} + q(X)\, q(X)^{'}\right] \tag{4.20}$$

and $\Delta_2 = I_p$, so that $I(\beta_0) = I_m(\beta_0)$, where $I_m(\beta_0)$ is the bound in (4.4) under the standard MAR setup. Therefore, the bound in Theorem 4 generalizes the bound in the standard MAR setup.

An important special case is obtained by setting $X = 1$, which corresponds to replacing Assumption 2.3 by $Z \perp D$. This special case is typically referred to as MCAR (missing completely at random). Under MCAR, the term in 4.16 simplifies to:

$$\Lambda_{mcar,0}^{-1} = R_0 \otimes \Sigma_0^{-1},$$

so that the bound simplifies to

$$I_{mcar}(\beta_0) = \sum_{j=1}^{J} \frac{1}{p_{j,0}} \Gamma_0^{'} d_j^{'} \Sigma_0^{-1} d_j \Gamma_0. \tag{4.21}$$

To interpret the expression in (4.21), make the additional assumption that $d_j \Gamma_0$ has full rank. In that case, the inverse of the information bound using only the data in stratum j is

$$I_j(\beta_0) = \frac{1}{p_{j,0}} \Gamma_0^{'} d_j^{'} \Sigma_0^{-1} d_j \Gamma_0, \tag{4.22}$$

and the information under MCAR is obtained by summing the information in the strata,

$$I_{mcar}(\beta_0) = \sum_{j=1}^{J} I_j(\beta_0). \tag{4.23}$$

Remark 5. The efficiency result stated above holds for a moment functions ψ, not for the underlying model that generates ψ. For some models, it may be possible to formulate additional moment conditions that are redundant in the full data setting, but are valid and informative under incomplete data. Pacini and Windmeijer (2015) provide an example of this in the linear dynamic panel data setting.

4.2 Example: Linear IV

The moment conditions (4.14) and (4.15) have a natural interpretation in the linear IV model, where they correspond to an augmented set of instruments.

Consider the linear IV model under MCAR ($X = 1$), with a set of instruments (W_1, W_2) and an error term u such that $E[W_1 u] = E[W_2 u] = 0$. Assume three outcomes for the missing data indicator, ignoring the stratum in which no instrument is observed:

$$D \in \left\{ d_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, d_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, d_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}.$$

The set of moment conditions for the observed data implied by Assumptions 1-3 is

$$\begin{aligned} E[s_1 W_1 u] &= 0, \\ E[s_1 W_2 u] &= 0, \\ E[s_2 W_1 u] &= 0, \\ E[s_3 W_1 u] &= 0, \end{aligned}$$

where I have omitted the zero rows in d_2 and d_3, since those are uninformative,[5] and have also omitted the moment conditions that define $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$.

An optimal GMM estimator based on these moment conditions is equivalent to IV estimation using four instruments: $s_1 W_1$, $s_1 W_2$, $s_2 W_1$, and $s_3 W_1$. It attains the efficiency bound in Theorem 4. In particular, it is more efficient

[5] For more information, see the derivation of the efficiency bound in Section A.1.

than using only $s_1 W_1$ and $s_1 W_2$, the approach typically taken in empirical research. The latter approach corresponds to discarding observations from the incomplete strata 2 and 3. See Section 6 for a simulation study that quantifies the efficiency gain.

The set of implied moment conditions depends on the missing data patterns that are observed in the data. If the missing data indicator has outcomes:

$$D \in \left\{ d_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, d_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\},$$

then the set of observable moment conditions shrinks to

$$\begin{aligned} E\left[s_1 W_1 u\right] &= 0, \\ E\left[s_2 W_2 u\right] &= 0. \end{aligned}$$

In this example, the parameter β_0 is not identified in any single stratum, but the framework in this paper shows that it can be identified using the information from all strata simultaneously, since Assumption 3 is satisfied.[6]

5 Estimation

In this section, I propose a GMM estimator to estimate the parameter of interest, β_0, and establish its large sample properties. Notably, the GMM estimator proposed here is efficient, as it attains the bound (4.19) established in Theorem 4.

First, define the plug-in estimator for the selection probabilities $p_{j,o}\left(x_l\right) = P\left(D = d_j \middle| X = x_l\right)$ as

$$\hat{\rho}_j\left(x\right) = \frac{\sum_{i=1}^n 1\left\{D_i = d_j, X_i = x\right\}}{\sum_{i=1}^n 1\left\{X = x\right\}}. \tag{5.1}$$

[6] It is satisfied since $\sum_j d_j = I_2$ has full rank.

Denote the resulting estimator for $p_{j,0}$ by $\hat{p}_j$, and stack the selection probabilities in $p_0 = (p_{1,0}, \cdots, p_{J,0})$, and their plug-in estimators in $\hat{p} = (\hat{p}_1, \cdots, \hat{p}_J)$.

For each stratum j, let $\tilde{d}_j$ be the rectangular selection matrix that selects the non-zero rows of the missing data indicator d_j. Denote by $\bar{m}_n(p, \beta)$ the sample analog of the moment conditions (4.15) for those nonzero rows,

$$\bar{m}_n (p, \beta) = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^{n} \frac{1\{D_i = d_1\}}{\hat{p}_1(X_i)} \tilde{d}_1 \psi (Z_i, \beta) \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} \frac{1\{D_i = d_J\}}{\hat{p}_J(X_i)} \tilde{d}_J \psi (Z_i, \beta) \end{bmatrix}. \tag{5.2}$$

Finally, define the GMM estimator $\hat{\beta}$ as

$$\hat{\beta} = \text{argmin}_{\beta \in \mathcal{B}} m_n (\hat{p}, \beta)' W_n m_n (\hat{p}, \beta). \tag{5.3}$$

for a weight matrix W_n of appropriate dimension.

The following Assumption is useful for establishing the large sample properties of $\hat{\beta}$.

Assumption 6. *(i) The parameter space $\mathcal{B}$ is compact, and β_0 is in the interior of $\mathcal{B}$*; *(ii) the sequence of matrices W_n converges to $I(\beta_0)$; (iii) the moment function ψ is continuously differentiable on $\mathcal{B}$.*

Theorem 7. *Assume that the conditions of Theorem 5 are satisfied, and that Assumptions 1, 2, 3, and 6 hold. Then the limiting distribution of the two-step GMM estimator $\hat{\beta}$ in (5.3) is given by*

$$\sqrt{n} \left(\hat{\beta} - \beta_0 \right) \xrightarrow{p} \mathcal{N} \left(0, I^{-1} (\beta_0) \right). \tag{5.4}$$

Proof. See Appendix A.2. □

Theorem 7 establishes consistency, asymptotic normality, and efficiency of $\hat{\beta}$. The restrictions imposed on Z and ψ can be relaxed. In particular, the multinomial assumption can be dropped in exchange for boundedness of some features of ψ, and the smoothness assumptions on ψ can be weakened.

6 Simulations

This section uses a simulation study to compare the proposed procedure to some standard procedures for dealing with missing data. In the first example, I discuss an instrumental variable model where the instruments are partially observed. The second example is the dynamic panel data estimator proposed by Arellano and Bond (1991). The first example highlights that efficiency is improved when we include observations from subpopulations where the parameter is not identified. The second example highlights the flexibility of the approach in terms of missing data patterns.

6.1 Instrumental variables

The first example is a linear IV model with two instruments. The dependent and explanatory variables are always observed, but we do not always observe both of the instruments. Either instrument can be missing for a subsample.

The dependent variable y is linearly related to an explanatory variable X, $y = X\beta_0 + u$. Two instruments are available, W_1 and W_2, which motivates the following unconditional moment conditions to estimate β_0:

$$E\left(\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} (y - X\beta_0)\right) = 0. \tag{6.1}$$

There are four strata ($J = 3$). For the first stratum, we observe both instruments. For the second stratum, we observe only W_1, and for the third stratum we observe only W_2. Then, the missing data patterns are:

$$\left\{ d_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ d_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},\ d_3 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}.$$

We assume that the instruments are similar in all aspects. First, they are equally likely to be missing, i.e. the selection probabilities are given by: $p_{2,0} = p_{3,0} = (1 - p_{1,0})/2$. Second, they have the same correlation with the explanatory variable: $E(W_1 X) = E(W_2 X) = \lambda$. Third, they have the same first two moments: $E(W_k) = 0$ and $E(W_k^2) = 1$, $k = 1, 2$. The instruments have correlation $\rho = Cov(W_1, W_2)$. Finally, we normalize the variance of the explanatory variable, $Var(x) = 1$. Then we have

$$\text{Var}\begin{pmatrix} X \\ W_1 \\ W_2 \end{pmatrix} = \begin{pmatrix} 1 & \lambda & \lambda \\ \lambda & 1 & \rho \\ \lambda & \rho & 1 \end{pmatrix}.$$

Positive definiteness of $\text{Var}(x, w_1, w_2)$ implies that $\rho > 2\lambda^2 - 1$. We fix $\lambda = \frac{1}{\sqrt{2}}$ so that the lower bound for ρ is 0. The choice of λ does not affect the results.

We consider five estimators: (1) the "full data" estimator is an infeasible estimator that uses both instruments, even when they are not observed; (2) the "complete case" estimator discards observations for which either one of the instruments is not observed; (3) the "available case" estimator replaces missing instruments by zeros. This amount to estimating each of the moment functions using all the observations for which that moment function is observed; (4) the "complete moment" estimator uses only one instrument; (5) the optimal estimator is the estimator proposed in this paper..

In Figure 6.1 we plot the asymptotic variance of our estimators as a function of ρ for $p_1 = 0.5$. The key aspect of this example is that the two instruments provide similar sources of information. Therefore, as ρ increases,

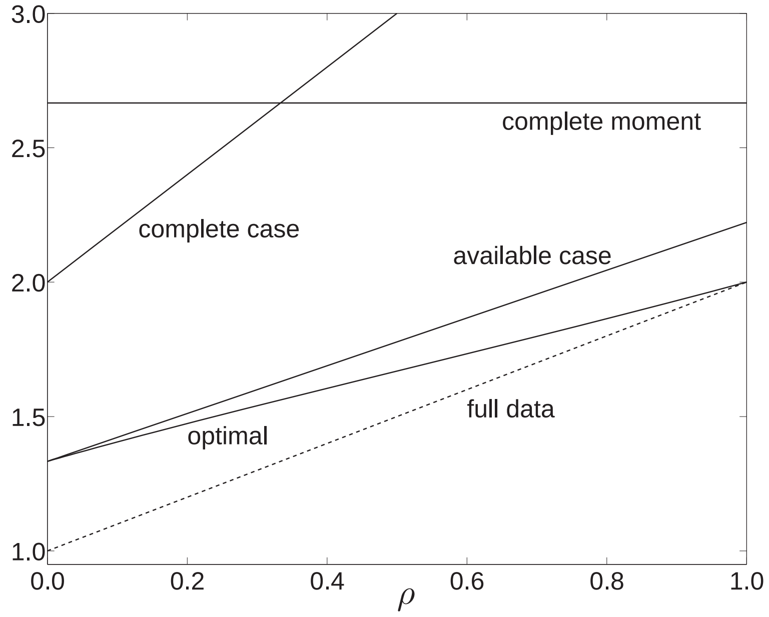

Figure 6.1: Asymptotic variance for various estimators of β as a function of $\rho,\ p_1 = 0.5$.

two effects are expected. First, the total amount of information for β_0 decreases, so we expect the variance of all estimators to increase. Second, the amount of information on the instrument that is missing increases. Since the optimal estimator is constructed such that it efficiently exploits the correlation between the components of the moment conditions, we expect the relative performance of the optimal estimator to increase.

The optimal estimator is efficient among the feasible estimators. Except for $\rho = 0$, it outperforms the available case estimator. As ρ increases, the relative performance of the optimal estimator with respect to the available estimator increases: the available case estimator uses all the available data but does not efficiently use the correlation between the instruments. As ρ approaches 1, the variance of the optimal missing data estimator approach that of the infeasible estimator. The complete case and complete moment estimators are always outperformed by the available case estimator.

	Missing components			
	None	$y_{i,1}$	$y_{i,4}$	$(y_{i,1}, y_{i,4})$
$y_{i,1}\Delta\epsilon_{i,3}$	X	·	X	·
$y_{i,1}\Delta\epsilon_{i,4}$	X	·	·	·
$y_{i,1}\Delta\epsilon_{i,5}$	X	·	·	·
$y_{i,2}\Delta\epsilon_{i,4}$	X	X	·	·
$y_{i,2}\Delta\epsilon_{i,5}$	X	X	·	·
$y_{i,3}\Delta\epsilon_{i,5}$	X	X	·	·

Table 1: Missing data patterns for dynamic panel data estimation using the estimator in [**?**], $T = 5$.

6.2 Dynamic panel data

The goal of this section is to demonstrate the performance of our method in a more complicated model and to provide an example where the variance matrix is not known. In particular, we look at a dynamic panel data model, and use continuous updating GMM to estimate the parameters of the model.

The parameter of interest ρ describes the relationship between current and lagged values of a random variable $y_{i,t}$:

$$y_{i,t} = \alpha_i + \rho y_{i,t-1} + u_{i,t},\ 2 \leq t \leq T. \tag{6.2}$$

We normalize $\mathrm{E}(\alpha_i) = 0$, $\mathrm{Var}(\alpha_i) = \sigma_a^2$, and $\mathrm{E}(u_{it}) = 0$, $Var(u_{it}) = 1$. Furthermore, we assume no autocorrelation in the error terms: $E(u_{i,t}u_{i,s}) = 0$ whenever $s \neq t$. Arellano and Bond (1991) propose an estimator that is widely used: the optimal GMM estimator based on the $(T-2)(T-1)/2$ moment conditions $E(y_{i,t-s}\Delta u_{i,t}) = 0,\ t \geq 3, s \geq 2$.

For any observation i, if $y_{i,t}$ is not observed, then several components of the moment function are not observed. For an example with $T = 5$, see Table 1. For the purposes of this simulation, we consider the case $T = 9$, which corresponds to the example in Blundell and Bond (1998). This gives 28 moment conditions for 1 parameter. If any of the $y_{i,t}$ are missing, the moment function is incompletely observed: if $y_{i,1}$ is not observed, 7 components of the moment function are not observed; if $y_{i,4}$ is not observed, 12 components of the moment function are not observed.

This section reports the results from a Monte Carlo analysis to compare the relative performance of the estimator introduced in this paper to the full data, complete case, and available case estimators. We use a continuous updating version of the Arellano-Bond estimator to estimate ρ. When estimating the variance matrix, we assume that $\Omega_j = \Omega$ for each j.

We consider different values for the variance of the individual effect $\sigma_\alpha^2 \in \{0.1, 1\}$ and the parameter of interest $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. We set the number of observations to $n = 10000$ and perform $s = 1000$ simulations per parameter combination. There are 10 missing data patterns. Patterns $j = 1, \ldots, 9$ have $y_{i,j}$ missing and the other variables observed. Pattern 10 corresponds to the subsample with all variables observed. The severity of the missing data problem is determined by a parameter $p = P(R = S_j)$. In the population, a proportion $P(R = S_{10}) = 1 - 9p$ of the observations are complete. I consider $p \in \{0.02, 0.06\}$ so that 82% (respectively 46%) of the observations are complete.

Table 2 reports the variance of the complete case, available case, and optimal estimator divided by the variance of the full-data estimator. The complete case estimator is always outperformed by the available-case estimator, except for $(\sigma_\alpha^2, \rho, p) = (1, 0.8, 0.02)$. The optimal estimator always outperforms the other two estimators, which provides evidence in favor of the missing data GMM estimator proposed in this paper. The optimal estimator seems to gain relative efficiency as p increases. For some parameter configurations, the efficiency gain can be substantial.

σ_α^2	ρ	p	cc	ac	opt
0.1	0.1	0.02	1.19	1.12	1.08
		0.06	2.29	1.46	1.41
	0.2	0.02	1.29	1.23	1.18
		0.06	2.37	1.34	1.27
	0.5	0.02	1.82	1.77	1.69
		0.06	3.35	2.50	2.25
	0.8	0.02	8.61	8.11	7.74
		0.06	15.95	11.76	10.45
1	0.1	0.02	1.71	1.47	1.46
		0.06	3.04	1.89	1.84
	0.2	0.02	1.91	1.70	1.68
		0.06	3.75	2.35	2.21
	0.5	0.02	5.10	4.75	4.59
		0.06	8.61	5.85	5.33
	0.8	0.02	2.04	2.20	1.92
		0.06	3.47	3.30	2.62

Table 2: Simulation results for a continuous updating version of the Arellano-Bond estimator in a dynamic panel data context with $T = 9$ time periods. Reported results are variances relative to the full data estimator, for the (i) complete case estimator (cc); (ii) available case estimator (ac); and (iii) optimal (opt) estimator. Results of a Monte Carlo study with $n = 1000$ observations, and $s = 1000$ replications. of a continuous updating version of the Arellano-Bond estimator. Details of the simulation design are in Section 6.2.

References

Abrevaya, J. and D. Abrevaya (2011), "A GMM approach for dealing with missing data on regressors and instruments", Mimeo.

Abrevaya, J. and D. Abrevaya (2015), "A GMM approach for dealing with missing data on regressors", Working paper.

Abrevaya, J. (2016), "Missing dependent variables in Fixed-Effects Models", Working paper.

Angrist, J., V. Lavy, and A. Schlosser (2010): "Multiple Experiments for the Causal Link Between the Quantity and Quality of Children," Journal of Labor Economics 28 (4), 773–824.

Arellano, M. and S. Bond (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," The Review of Economic Studies, 58, 277.

Cattaneo, M.D. (2010), "Efficient Semiparametric Estimation of Multi-valued Treatment Effects Under Ignorability", Journal of Econometrics 155, 138–154.

Chamberlain, G. (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", Journal of Econometrics 34, 305–334.

Chamberlain, G. (1992), "Efficiency Bounds for Semiparametric Regression", Econometrica 60 (3), 567–596.

Chen, B., G. Yi, and R. Cook (2010), "Weighted Generalized Estimating Functions for Longitudinal Response and Covariate Data That Are Missing at Random," Journal of the American Statistical Association, 105, 336–353.

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

Dagenais, M.G. (1973), "The Use of Incomplete Observations in Multiple Regression Analysis: A Generalized Least Squares Approach." Journal of Econometrics 1 (4), 317–328.

Dardanoni, V., S. Modica and F. Peracchi (2011), "Regression with Imputed Covariates: A Generalized Missing-Indicator Approach", Journal of Econometrics 162 (2), 362–368.

Gourieroux, C. and A. Monfort (1981), "On the Problem of Missing Data in Linear Models", The Review of Economic Studies 48 (4), 579–586.

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

Heckman, J., H. Ichimura, and P. Todd (1997), "Matching As An Econometric Evaluation Estimator", Review of Economic Studies 64, p. 605–654.

Hirano, K., G. Imbens, and G. Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", Econometrica 71, 1161–1189.

Hirano, K., G. Imbens, G. Ridder, and D.B. Rubin (2001), "Combining Panel Data Sets with Attrition and Refreshment Samples", Econometrica 69 (6), 1645–1659.

Meyer, C.D. (1973), "Generalized Inverses and Ranks of Block Matrices", SIAM Journal on Applied Mathematics 25 (4), p. 597–602.

Mogstad, M., and M. Wiswall (2012). "Instrumental Variables Estimation with Partially Missing Instruments", Economics Letters 114 (2), 186–189.

Newey, W.K., and D.L. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in R.F. Engle and D.L. McFadden, *Handbook of Econometrics, Volume 4*, pp. 2111-2245. Amsterdam: Elsevier.

Pacini, D. and F. Windmeijer (2016), "Moment conditions for AR(1) Panel Data Models with Missing Outcomes", Bristol Economics Discussion Papers 15/660, University of Bristol, UK.

Prokhorov, A. and P. Schmidt (2009), "GMM Redundancy Results for General Missing Data Problems", Journal of Econometrics 151, p. 47–55.

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

Tsiatis, A.A. (2010), "Semiparametric Theory and Missing Data", New York: Springer.

Verbeek, M. and T. Nijman (1992), "Testing for Selectivity Bias in Panel Data Models", International Economic Review 33(3), 681–703.

Wooldridge, J. (2007), "(2007): "Inverse ProbabilityWeighted Estimation for General Missing Data Problems," Journal of Econometrics, 141, 1281–1301."

A Proofs

A.1 Proof of Theorem 4

For $J = 1$ and $d_1 = I$, the Theorem reduces to Graham's (2011) Theorem 2.1. For that reason, the derivation here is similar to that in Graham (2011, Supplementary Material).

The random object D is a missing data indicator that takes one of J values $(d_1, \cdots, d_J)$, and signals which of the components of ψ are observed. Here, we will work with a modification of the missing data indicator. We omit the zero rows, and call the resulting object $\tilde{D}$, with outcomes $\tilde{d}_j$, rectangular selection matrices of size $r_j \times r$, where r_j is the number of observable components of ψ in stratum j.

For example, if $r = 2$, $\tilde{D}$ may take values

$$\begin{aligned} \tilde{d}_1 &= I_2, \\ \tilde{d}_2 &= \begin{bmatrix} 1 & 0 \end{bmatrix}, \\ \tilde{d}_3 &= \begin{bmatrix} 0 & 1 \end{bmatrix}, \end{aligned}$$

so that $\tilde{d}_1$ signals that both components are observed, $\tilde{d}_2$ corresponds to observing only the first component, and $\tilde{d}_3$ corresponds to observing only the second component. Note that d_{J+1} disappears from the analysis.

The set of moment conditions for which we wish to establish the efficiency bound is given by (4.14) and 4.15. We have assumed that Z follows a multinomial distribution. It will be useful to introduce some more notation related to the multinomial nature of (Z, X). Let L be the number of support points of X, so that X takes values in $\{x_1, \cdots, x_L\}$. The $L \times 1$ vector B converts X into L binary variables

$$B = (1\{X = x_1\}, \cdots, 1\{X = x_L\}).$$

Denote the probability that a unit with $X = x_l$ selects into missing data pattern j by

$$\rho_{jl,0} = P(D = d_j | X = x_l),$$

and stack the selection probabilities for pattern j into

$$\rho_{j,0} = (\rho_{j1,0}, \cdots, \rho_{jL,0}).$$

Then we can write

$$p_{j,0}(X) = B^{'}\rho_{j,0}.$$

Step 1: Equivalence to unconditional moments.

The moment conditions in (4.14) and 4.15 are equivalent to

$$E\left[m_1(\rho_0)\right] = E\begin{bmatrix} m_{11}(\rho_{1,0}) \\ \vdots \\ m_{1J}(\rho_{J,0}) \end{bmatrix} = E\left[\begin{bmatrix} \frac{s_1}{B^{'}\rho_{1,0}} - 1 \\ \vdots \\ \frac{s_J}{B^{'}\rho_{J,0}} - 1 \end{bmatrix} \otimes B\right] = 0 \tag{A.1}$$

$$E\left[m_2(\rho_0, \beta_0)\right] = E\begin{bmatrix} m_{21}(\rho_{1,0}, \beta_0) \\ \vdots \\ m_{2J}(\rho_{J,0}, \beta_0) \end{bmatrix} = E\begin{bmatrix} \frac{s_1}{B^{'}\rho_{1,0}}\tilde{d}_1\psi(Z, \beta_0) \\ \vdots \\ \frac{s_J}{B^{'}\rho_{J,0}}\tilde{d}_J\psi(Z, \beta_0) \end{bmatrix} = 0 \tag{A.2}$$

The dimension of m_1 is $JL \times 1$, and the dimension of m_2 is $\bar{r} \times 1$, where $\bar{r} = \sum_j r_j$ is the total number of components selected by the d_j's. The equivalence is then a straightforward extension of Graham (2011, Supplementary material, Section A, Step 1).

Step 2: Applying Lemma 2 of Chamberlain (1987).

Define $m = (m_1, m_2)$, and define the variance of the moment conditions as the $(JL + \bar{r}) \times (JL + \bar{r})$ matrix

$$V = E\left[m(\rho, \beta)\, m(\rho, \beta)^{'}\right] \tag{A.3}$$

$$= \begin{bmatrix} V_{11} & V_{12} \\ V_{12}^{'} & V_{22} \end{bmatrix} \left(\text{dimensions } \begin{bmatrix} JL \times JL & \bar{r} \times \bar{r} \\ \bar{r} \times \bar{r} & JL \times \bar{r} \end{bmatrix}\right) \tag{A.4}$$

where $V_{12} = E\left[m_1(\rho_0)\, m_2(\beta_0, \rho_0)'\right]$, etcetera. Define its inverse as

$$V^{-1} = \begin{bmatrix} V^{11} & V^{12} \\ \ddots & V^{22} \end{bmatrix},$$

with components of equal dimensions as the corresponding components of V.

The expected derivative of the moment conditions is the $(JL + \bar{r}) \times (JL + K)$ matrix

$$\begin{aligned} M &= E\left[\left. \frac{\partial m(\beta, \rho)}{\partial (\beta, \rho)'} \right|_{\beta=\beta_0, \rho=\rho_0} \right] & \text{(A.5)} \\ &= \begin{bmatrix} M_{1\rho} & O_{JL\times k} \\ M_{2\rho} & M_{2\beta} \end{bmatrix} \left(\text{dimensions } \begin{bmatrix} JL \times JL & JL \times K \\ \bar{r} \times JL & \bar{r} \times K \end{bmatrix} \right) & \text{(A.6)} \end{aligned}$$

where $M_{2\beta} = E\left[\frac{\partial m_2(\beta,\rho)}{\partial \beta'}\right]$, etc. Chamberlain showed that the lower bound on the variance of β is given by

$$I(\beta_0) = \left\{ \left(M' V^{-1} M \right)^{-1} \right\}_{22}, \tag{A.7}$$

which is the objective of the present derivation.

Remark 8. Chamberlain's result requires that M has full rank, and that V is invertible. Step 5 establishes that this is the case.

The derivations for $I(\beta)$ simplify if $M'V^{-1}M$ is block-diagonal. Note that

$$\begin{bmatrix} M_{1\rho}' & M_{2\rho}' \\ 0 & M_{2\beta}' \end{bmatrix} \begin{bmatrix} V^{11} & V^{12} \\ \ddots & V^{22} \end{bmatrix} \begin{bmatrix} M_{1\rho} & 0 \\ M_{2\rho} & M_{2\beta} \end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\ A_2' & M_{2\beta}' V^{22} M_{2\beta} \end{bmatrix},$$

where

$$\begin{aligned} A_1 &= M_{1\rho}^{'} V^{11} M_{1\rho} + M_{2\rho}^{'} V^{12'} M_{1\rho} + M_{1\rho}^{'} V^{12} M_{2\rho} + M_{2\rho}^{'} V^{22} M_{2\rho}, \\ A_2 &= M_{1\rho}^{'} V^{12} M_{2\beta} + M_{2\rho}^{'} V^{22} M_{2\beta}. \end{aligned} \tag{A.8}$$

If $A_2 = 0$, then $M^{'} V^{-1} M$ is block-diagonal, so that

$$I^{-1}(\beta) = M_{2\beta}^{'} V^{22} M_{2\beta}. \tag{A.9}$$

It can be seen from the expression (A.8) that this happens when $M_{1\rho}^{'} V^{12} M_{2\beta} = -M_{2\rho}^{'} V^{22} M_{2\beta}$. Prokhorov and Schmidt (2009) show that this happens when $V_{12}^{'} V_{11}^{-1} M_{1\rho} = M_{2\rho}$. Step 5 establishes that this is the case. In what follows, we will assume the results of Step 5, and will proceed to evaluate the expression in (A.9).

Step 3: Computing the bound's ingredients

In this step, I obtain expressions for (dimensions): (i) $M_{2\beta}$ $(\bar{r} \times K)$; (ii) V_{22} $(\bar{r} \times \bar{r})$; (iii) V_{12} $(JL \times \bar{r})$; (iv) V_{11} $(JL \times JL)$, and its inverse V_{11}^{-1}; (v) $V^{22} = \left(V_{22} - V_{12}^{'} V_{11}^{-1} V_{12}\right)^{-1}$ $(\bar{r} \times \bar{r})$. I aim for expressions of the form

$$\sum_{l=1}^{L} A_l \otimes e_l e_l^{'},$$

where e_l is the unit vector of appropriate dimension, with zeros everywhere and a 1 in position l. This will facilitate the matrix products and inverses in steps (iv) and (v).

(i) Derivative $M_{2\beta}$. The moment function associated with the j-th stratum is an $r_j \times 1$ vector

$$m_{2j}(\rho_j, \beta) = \frac{s_j}{p_j(X)} \tilde{d}_j \psi(Z, \beta).$$

The derivative is the $r_j \times K$ matrix

$$\frac{\partial m_{2j}}{\partial \beta'} = \tilde{d}_j \frac{s_j}{p_j(X)} \frac{\partial \psi(Z,\beta)}{\partial \beta'}$$

so that (evaluating all derivatives at the true value of the parameters)

$$\begin{aligned} M_{2j,\beta} &= \tilde{d}_j E\left[\frac{s_j}{p_{j,0}(X)} \frac{\partial \psi(Z,\beta)}{\partial \beta'}\right] \\ &= \tilde{d}_j E\left[E\left[\left.\frac{s_j}{p_{j,0}(X)} \frac{\partial \psi(Z,\beta)}{\partial \beta'}\right| X\right]\right] \\ &= \tilde{d}_j E\left[\frac{1}{p_{j,0}(X)} E\left[s_j | X\right] E\left[\left.\frac{\partial \psi}{\partial \beta'}\right| X\right]\right] \qquad \text{(A.10)} \\ &= \tilde{d}_j E\left[E\left[\left.\frac{\partial \psi}{\partial \beta'}\right| X\right]\right] \\ &= \tilde{d}_j \Gamma_0 \end{aligned}$$

where (A.10) is implied by Assumption 2.3.

To obtain the expected derivative $M_{2\beta}$, we introduce some more notation. The matrix

$$\tilde{\Delta}_1 = \begin{bmatrix} \tilde{d}_1 & 0 & 0 \\ 0 & \ddots & \\ 0 & & \tilde{d}_J \end{bmatrix}$$

is $\bar{r} \times Jr$, and the matrix

$$\tilde{\Delta}_2 = \begin{bmatrix} \tilde{d}_1 \\ \vdots \\ \tilde{d}_J \end{bmatrix}$$

is $\bar{r} \times r$. Note that $\tilde{\Delta}_2 = \tilde{\Delta}_1 (\iota_J \otimes I_r)$, where ι_J is the $J \times 1$ vector of ones.

Also note that $\tilde{\Delta}_1\tilde{\Delta}_1^{'} = I_{\bar{r}}$, and $\tilde{\Delta}_1^{'}\tilde{\Delta}_1 = \Delta$, where

$$\Delta = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & \ddots & \\ 0 & & d_J \end{bmatrix}$$

is a square, block-diagonal matrix with square selection matrices d_j as blocks. To see this, let $r = 3$, and

$$\tilde{d}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

then $\tilde{d}_2\tilde{d}_2^{'} = I_2$, and

$$\tilde{d}_2^{'}\tilde{d}_2 \quad = \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = d_2.$$

The expected derivative of m_2 with respect to β at β_0 is the $\bar{r} \times K$ matrix

$$M_{2\beta} = \begin{bmatrix} \tilde{d}_1 \\ \vdots \\ \tilde{d}_J \end{bmatrix} \Gamma_0 = \tilde{\Delta}_2\Gamma_0.$$

(ii) Variance V_{22}. For the lower right block of V, we have the $\bar{r} \times \bar{r}$ matrix

$$V_{22} = E\left[m_2(\beta,\rho)\, m_2(\beta,\rho)^{'}\right].$$

This matrix is blockdiagonal with $r_j \times r_j$ blocks $E\left[m_{2j}m_{2j}^{'}\right]$. It is block-diagonal because $E[s_j s_k] = 0 \Leftrightarrow j \neq k$, which implies $E\left[m_{2j}m_{2k}^{'}\right] = 0$ whenever $j \neq k$.

The diagonal blocks are as in Graham (2011, Theorem 2.1), i.e.

$$\begin{aligned} E\left[m_{2j}m_{2j}^{'}\right] &= E\left[\frac{s_j}{p_{j,0}^2(X)}\tilde{d}_j\psi(Z,\beta)\psi(Z,\beta)^{'}\tilde{d}_j^{'}\right] \\ &= \tilde{d}_jE\left[E\left[\left.\frac{s_j}{p_{j,0}^2(X)}\right|X\right]E\left[\psi(Z,\beta)\psi(Z,\beta)^{'}\Big|X\right]\right]\tilde{d}_j^{'} \\ &= \tilde{d}_jE\left[\frac{1}{p_{j,0}(X)}E\left[\psi(Z,\beta)\psi(Z,\beta)^{'}\Big|X\right]\right]\tilde{d}_j^{'} \\ &= \tilde{d}_jE\left[\frac{1}{p_{j,0}(X)}E\left[\Sigma(X)+q(X)q(X)^{'}\right]\right]\tilde{d}_j^{'} \end{aligned}$$

where $q(X) = E(\psi(Z,\beta_0)|X)$ and $\Sigma(X) = V(\psi(Z,\beta_0)|X)$.

To rewrite this using the discrete support of $X \in \{x_1, \cdots, x_L\}$, let

$$\begin{aligned} \tau_l &= P(X = x_l), \\ q_l &= q(x_l), \\ \Sigma_l &= \Sigma(x_l). \end{aligned}$$

Then the j-th block can be written as

$$E\left[m_{2j}m_{2j}^{'}\right] = \tilde{d}_j\left[\sum_{l=1}^{L}\frac{\tau_l}{\rho_{jl,0}}\left(\Sigma_l + q_lq_l^{'}\right)\right]\tilde{d}_j^{'}. \tag{A.11}$$

To construct V_{22}, denote the $J \times J$ matrix of selection probabilities given $X = x_l$ by

$$R_{l,0}^{-1} = \begin{bmatrix} \frac{1}{\rho_{1l,0}} & 0 & 0 \\ 0 & \ddots & \\ 0 & & \frac{1}{\rho_{Jl,0}} \end{bmatrix}.$$

Conditional on $X = x_l$, D follows a multinomial distribution with probabilities $R_{l,0}\iota_J$. Then

$$\begin{aligned} V_{22} &= \begin{bmatrix} \tilde{d}_1 & 0 & 0 \\ 0 & \ddots & \\ 0 & & \tilde{d}_J \end{bmatrix} \left[\sum_{l=1}^{L} \tau_l \begin{bmatrix} \frac{1}{\rho_{1l,0}} & 0 & 0 \\ 0 & \ddots & \\ 0 & & \frac{1}{\rho_{Jl,0}} \end{bmatrix} \otimes \left(\Sigma_l + q_l q_l^{'} \right) \right] \begin{bmatrix} \tilde{d}_1 & 0 & 0 \\ 0 & \ddots & \\ 0 & & \tilde{d}_J \end{bmatrix}^{'} \\ &= \tilde{\Delta}_1 \left(\sum_{l=1}^{L} \tau_l R_{l,0}^{-1} \otimes \left(\Sigma_l + q_l q_l^{'} \right) \right) \tilde{\Delta}_1^{'}. \qquad \text{(A.12)} \end{aligned}$$

(iii) Variance V_{12}. The covariance $V_{12} = E\left[m_1 m_2^{'}\right]$ consists of J^2 blocks. Off-diagonal blocks are of dimension $L \times r_k$ block

$$\begin{aligned} V_{12,jk} &= E\left[m_{1j} m_{2k}^{'}\right] \\ &= E\left[B\left(\frac{s_j}{p_{j,0}(X)} - 1\right) \frac{s_k}{p_{k,0}(X)} \psi(Z,\beta)^{'} \tilde{d}_k^{'}\right] \\ &= E\left[B\left(\frac{s_j s_k}{p_{j,0}(X)\, p_{k,0}(X)} - \frac{s_k}{p_{k,0}(X)}\right) \psi(Z,\beta)^{'}\right] \tilde{d}_k^{'} \\ &= E\left[BE\left[\left(-\frac{s_k}{p_{k,0}(X)}\right)\middle| X\right] E\left[\psi(Z,\beta)^{'}\middle| X\right]\right] \tilde{d}_k^{'} \\ &= -E\left[Bq(X)^{'}\right] \tilde{d}_k^{'}. \qquad \text{(A.13)} \end{aligned}$$

In $\sum_l$-format, it obtains that

$$\begin{aligned} V_{12,jk} &= -\begin{bmatrix} \tau_1 q_1^{'} \\ \vdots \\ \tau_L q_L^{'} \end{bmatrix} \tilde{d}_k^{'} \\ &= -\left(\sum_{l=1}^{L} \tau_l \otimes e_l q_l^{'} \right) \tilde{d}_k^{'}. \qquad \text{(A.14)} \end{aligned}$$

Next, consider the diagonal blocks (dimensions $L \times r_j$)

$$\begin{aligned}
V_{12,jj} &= E\left[m_{1j} m_{2j}^{'}\right] \\
&= E\left[B\left(\frac{s_j}{p_{j,0}(X)} - 1\right)\frac{s_j}{p_{j,0}(X)}\psi(Z,\beta_0)^{'}\tilde{d}_j^{'}\right] \\
&= E\left[B\left(\frac{s_j}{p_{j,0}^2(X)} - \frac{s_j}{p_{j,0}(X)}\right)\psi(Z,\beta_0)^{'}\right]\tilde{d}_j^{'} \\
&= E\left[BE\left[\left(\frac{s_j}{p_{j,0}^2(X)} - \frac{s_j}{p_{j,0}(X)}\right)\middle| X\right] E\left[\psi(Z,\beta_0)^{'}\middle| X\right]\right]\tilde{d}_j^{'} \\
&= E\left[B\left(\frac{1}{p_{j,0}(X)} - 1\right) q(X)^{'}\right]\tilde{d}_j^{'}. \qquad \text{(A.15)}
\end{aligned}$$

Remember that B is an $L \times 1$ vector and that $q(X)$ is an $r \times 1$ vector. Continue the derivation above by expressing (A.15) as a Kronecker product:

$$\begin{aligned}
E\left[B\left(\frac{1}{p_{j,0}(X)} - 1\right) q(X)^{'}\right]\tilde{d}_j^{'} &= \begin{bmatrix} \tau_1\left(\frac{1}{\rho_{j1,0}} - 1\right) q_1^{'} \\ \vdots \\ \tau_L\left(\frac{1}{\rho_{jL,0}} - 1\right) q_L^{'} \end{bmatrix}\tilde{d}_j^{'} \\
&= \left(\sum_{l=1}^{L}\tau_l\left(\frac{1}{\rho_{jl,0}} - 1\right)\otimes e_l q_l^{'}\right)\tilde{d}_j^{'} \\
&= \left(\sum_{l=1}^{L}\frac{\tau_l}{\rho_{jl,0}}\otimes e_l q_l^{'}\right)\tilde{d}_j^{'} - \left(\sum_{l=1}^{L}\tau_l\otimes e_l q_l^{'}\right)\tilde{d}_j^{'} \qquad \text{(A.16)}
\end{aligned}$$

Now, arrange the blocks into V_{12}. The structure is

$$\begin{aligned}
V_{12} &= \begin{bmatrix} \left(\sum_{l=1}^{L}\frac{\tau_l}{\rho_{1l}}\otimes e_l q_l^{'}\right)\tilde{d}_1^{'} & 0 & 0 \\ 0 & \ddots & \\ 0 & & \left(\sum_{l=1}^{L}\frac{\tau_l}{\rho_{Jl}}\otimes e_l q_l^{'}\right)\tilde{d}_J^{'} \end{bmatrix} - \begin{bmatrix} \left(\sum_{l=1}^{L}\tau_l\otimes e_l q_l^{'}\right)\tilde{d}_1^{'} & \cdots & \left(\sum_{l=1}^{L}\tau_l\otimes e_l q_l^{'}\right)\tilde{d}_J^{'} \\ \vdots & & \vdots \\ \left(\sum_{l=1}^{L}\tau_l\otimes e_l q_l^{'}\right)\tilde{d}_1^{'} & \cdots & \left(\sum_{l=1}^{L}\tau_l\otimes e_l q_l^{'}\right)\tilde{d}_J^{'} \end{bmatrix} \\
&= \begin{bmatrix} \left(\sum_{l=1}^{L}\frac{\tau_l}{\rho_{1l}}\otimes e_l q_l^{'}\right) & 0 & 0 \\ 0 & \ddots & \\ 0 & & \left(\sum_{l=1}^{L}\frac{\tau_l}{\rho_{Jl}}\otimes e_l q_l^{'}\right) \end{bmatrix}\tilde{\Delta}_1^{'} - \begin{bmatrix} \left(\sum_{l=1}^{L}\tau_l\otimes e_l q_l^{'}\right) & \cdots & \left(\sum_{l=1}^{L}\tau_l\otimes e_l q_l^{'}\right) \\ \vdots & & \vdots \\ \left(\sum_{l=1}^{L}\tau_l\otimes e_l q_l^{'}\right) & \cdots & \left(\sum_{l=1}^{L}\tau_l\otimes e_l q_l^{'}\right) \end{bmatrix}\tilde{\Delta}_1^{'},
\end{aligned}$$

which reveals that

$$\begin{aligned} V_{12} &= \left(\sum_{l=1}^{L}\tau_l R_{l,0}^{-1}\otimes e_l q_l^{'}\right)\tilde{\Delta}_1^{'} - \left(\sum_{l=1}^{L}\tau_l \iota_J \iota_J^{'}\otimes e_l q_l^{'}\right)\tilde{\Delta}_1^{'} \\ &= \left(\sum_{l=1}^{L}\tau_l \left(R_{l,0}^{-1} - \iota_J \iota_J^{'}\right)\otimes e_l q_l^{'}\right)\tilde{\Delta}_1^{'}. \qquad \text{(A.17)} \end{aligned}$$

(iv) Variance V_{11}. Denote by $F_j = \frac{s_j}{p_j(X)} - 1$, let $F = (F_1, \cdots, F_J)$, so that $m_1(\rho_0) = F \otimes B$. We are after the inverse of

$$\begin{aligned} V_{11} &= E\left(m_1(\rho_0)\, m_1(\rho_0)^{'}\right), \\ &= E\left[(F\otimes B)(F\otimes B)^{'}\right], \\ &= E\left[(F\otimes B)\left(F^{'}\otimes B^{'}\right)\right], \\ &= E\left[\left(FF^{'}\right)\otimes\left(BB^{'}\right)\right], \\ &= E\left[E\left[\left(FF^{'}\right)\otimes\left(BB^{'}\right)\Big| X\right]\right], \\ &= \sum_{l=1}^{L}\tau_l E\left[FF^{'}\Big| X = x_l\right]\otimes e_l e_l^{'}, \qquad \text{(A.18)} \end{aligned}$$

where, for the third equality, we use that $(A\otimes B)^{'} = A^{'}\otimes B^{'}$, and the fourth equality uses $(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$.

The conditional moment restriction 4.10 implies that

$$E[F|X] = 0,$$

so that (A.18) becomes

$$V_{11} = \sum_{l=1}^{L}\tau_l V(F|X = x_l)\otimes e_l e_l^{'}. \qquad \text{(A.19)}$$

The conditional variance of F, $V(F|\, X = x_l)$ is obtained from standard results on the multinomial. To see this, note that the conditional distribution of F is a linear transformation of a multinomial:

$$(F|X = x_l) \overset{d}{=} \left(R_{l,0}^{-1} \begin{bmatrix} s_1 \\ \vdots \\ s_J \end{bmatrix} - \iota_J | X = x_l \right),$$

so that

$$V\left(F|X = x_l\right) = R_{l,0}^{-1} V \left(\begin{bmatrix} s_1 \\ \vdots \\ s_J \end{bmatrix} \middle| X = x_l \right) R_{l,0}^{-1},$$

where

$$\begin{bmatrix} s_1 \\ \vdots \\ s_J \end{bmatrix} \middle| X = x_l \sim MN\left(\rho_{1l}, \cdots, \rho_{JL}\right).$$

Since $R_{l,0}\iota = (\rho_{1l}, \cdots, \rho_{Jl})$ is the $J \times 1$ column vector of multinomial probabilities, and since $R_{l,0}$ is symmetric, we can use the well-known expression for the variance of a multionomial,

$$V\left(D|X = x_l\right) = R_{l,0} - R_{l,0}\iota_J\iota_J^{'}R_{l,0}, \tag{A.20}$$

and it follows that

$$V\left(F|\, X = x_l\right) = R_{l,0}^{-1} - \iota_J\iota_J^{'}.$$

To compute the inverse of V_{11} in (A.19), note that for any $l \neq k$, the product $\left(e_l e_l^{'}\right)\left(e_k e_k^{'}\right) = 0$. Therefore:

$$\begin{aligned} V_{11}^{-1} &= \left(\sum_{l=1}^{L} \tau_l V\left(F|X = x_l\right) \otimes e_l e_l^{'} \right)^{-1} \\ &= \sum_{l=1}^{L} \frac{1}{\tau_l} \left(R_{l,0}^{-1} - \iota_J\iota_J^{'} \right)^{-1} \otimes e_l e_l^{'}. \end{aligned} \tag{A.21}$$

(v) Variance V^{22}. We have obtained the ingredients of $V^{22} = \left(V_{22} - V_{12}^{'} V_{11}^{-1} V_{12}\right)^{-1}$ in equations A.12, A.17, and (A.21). Remember that $(A \otimes B)^{'} = \left(A^{'} \otimes B^{'}\right)$ and $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$, so that

$$\begin{aligned} V_{12}^{'} V_{11}^{-1} &= \tilde{\Delta}_1 \sum_{l_1=1}^{L} \sum_{l_2=1}^{L} \left(\tau_{l_1} \left(R_{l_1,0}^{-1} - \iota_J \iota_J^{'} \right) \otimes e_{l_1} q_{l_1}^{'} \right)^{'} \left(\frac{1}{\tau_{l_2}} \left(R_{l_2,0}^{-1} - \iota_J \iota_J^{'} \right)^{-1} \otimes e_{l_2} e_{l_2}^{'} \right) \\ &= \tilde{\Delta}_1 \sum_{l_1=1}^{L} \sum_{l_2=1}^{L} \frac{\tau_{l_1}}{\tau_{l_2}} \left(\left(R_{l_1,0}^{-1} - \iota_J \iota_J^{'} \right)^{'} \otimes q_{l_1} e_{l_1}^{'} \right) \left(\left(R_{l_2,0}^{-1} - \iota_J \iota_J^{'} \right)^{-1} \otimes e_{l_2} e_{l_2}^{'} \right) \\ &= \tilde{\Delta}_1 \sum_{l_1=1}^{L} \sum_{l_2=1}^{L} \frac{\tau_{l_1}}{\tau_{l_2}} \left(\left(R_{l_1,0}^{-1} - \iota_J \iota_J^{'} \right) \left(R_{l_2,0}^{-1} - \iota_J \iota_J^{'} \right)^{-1} \otimes q_{l_1} e_{l_1}^{'} e_{l_2} e_{l_2}^{'} \right) \qquad \text{(A.22)} \end{aligned}$$

where I have also used symmetry of $\left(R_{l,0}^{-1} - \iota_J \iota_J^{'}\right)$. Because

$$e_{l_1}^{'} e_{l_2} = \begin{cases} 0 & \text{if } l_1 \neq l_2, \\ 1 & \text{if } l_1 = l_2, \end{cases}$$

the expression in (A.22) reduces to

$$V_{12}^{'} V_{11}^{-1} = \tilde{\Delta}_1 \sum_{l=1}^{L} \left(I_J \otimes q_l e_l^{'} \right). \qquad \text{(A.23)}$$

Using similar tools, I obtain

$$\begin{aligned}
V_{12}^{'}V_{11}^{-1}V_{12} &= \tilde{\Delta}_1 \sum_{l=1}^{L}\left(I \otimes q_l e_l^{'}\right)\left(\sum_{l=1}^{L}\tau_l\left(R_{l,0}^{-1} - \iota_J\iota_J^{'}\right)\otimes e_l q_l^{'}\right)\tilde{\Delta}_1^{'} \\
&= \tilde{\Delta}_1 \sum_{l_1=1}^{L}\sum_{l_2=1}^{L}\left(I \otimes q_{l_1} e_{l_1}^{'}\right)\left(\tau_{l_2}\left(R_{l_2,0}^{-1} - \iota_J\iota_J^{'}\right)\otimes e_{l_2} q_{l_2}^{'}\right)\tilde{\Delta}_1^{'} \\
&= \tilde{\Delta}_1 \sum_{l_1=1}^{L}\sum_{l_2=1}^{L}\left(\tau_{l_2}\left(R_{l_2,0}^{-1} - \iota_J\iota_J^{'}\right)\otimes q_{l_1} e_{l_1}^{'} e_{l_2} q_{l_2}^{'}\right)\tilde{\Delta}_1^{'} \\
&= \tilde{\Delta}_1 \sum_{l=1}^{L}\tau_l\left(\left(R_{l,0}^{-1} - \iota_J\iota_J^{'}\right)\otimes q_l q_l^{'}\right)\tilde{\Delta}_1^{'}
\end{aligned}$$

using steps similar to those used to arrive at (A.23).

Finally,

$$\begin{aligned}
V_{22} - V_{12}^{'}V_{11}^{-1}V_{12} &= \tilde{\Delta}_1\left(\sum_{l=1}^{L}\tau_l R_{l,0}^{-1}\otimes\left(\Sigma_l + q_l q_l^{'}\right)\right)\tilde{\Delta}_1^{'} - \tilde{\Delta}_1\sum_{l=1}^{L}\tau_l\left(\left(R_{l,0}^{-1} - \iota_J\iota_J^{'}\right)\otimes q_l q_l^{'}\right)\tilde{\Delta}_1^{'} \\
&= \tilde{\Delta}_1\left(\sum_{l=1}^{L}\tau_l\left(R_{l,0}^{-1}\otimes\left(\Sigma_l + q_l q_l^{'}\right) - \left(R_{l,0}^{-1} - \iota_J\iota_J^{'}\right)\otimes q_l q_l^{'}\right)\right)\tilde{\Delta}_1^{'} \\
&= \tilde{\Delta}_1\left(\sum_{l=1}^{L}\tau_l\left(R_{l,0}^{-1}\otimes\Sigma_l + \iota_J\iota_J^{'}\otimes q_l q_l^{'}\right)\right)\tilde{\Delta}_1^{'} \\
&\equiv \tilde{\Delta}_1\Lambda_0\tilde{\Delta}_1^{'}. \qquad (A.24)
\end{aligned}$$

where

$$\begin{aligned}
\Lambda_0 &= \sum_{l=1}^{L}\tau_l\left(R_{l,0}^{-1}\otimes\Sigma_l + \iota_J\iota_J^{'}\otimes q_l q_l^{'}\right) \\
&= E\left(R_0^{-1}(X)\otimes V\left(\psi\left(Z,\beta_0\right)\right| X\right) + \iota_J\iota_J^{'}\otimes q(X)\, q(X)^{'}\Big) \\
R_0(X) &= \operatorname{diag}\left(p_{1,0}(X), \cdots, p_{J,0}(X)\right).
\end{aligned}$$

Step 4: Computing the bound

Assuming invertibility of Λ_0 (see Step 5),

$$\begin{aligned} V^{22} &= \left(V_{22} - V_{12}^{'} V_{11}^{-1} V_{12}\right)^{-1} \\ &= \left(\tilde{\Delta}_1 \Lambda_0 \tilde{\Delta}_1^{'}\right)^{-1} \\ &= \tilde{\Delta}_1 \Lambda_0^{-1} \tilde{\Delta}_1^{'}. \end{aligned}$$

To see this, remember that $\tilde{\Delta}_1 \tilde{\Delta}_1^{'} = I_{\bar{r}}$, and that

$$\tilde{\Delta}_1^{'} \tilde{\Delta}_1 = \Delta = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & \ddots & \\ 0 & & d_J \end{bmatrix}.$$

It follows that

$$\begin{aligned} \tilde{\Delta}_1 \Lambda_0^{-1} \tilde{\Delta}_1^{'} \tilde{\Delta}_1 \Lambda_0 \tilde{\Delta}_1^{'} &= \tilde{\Delta}_1 \tilde{\Delta}_1^{'} \tilde{\Delta}_1 \Lambda_0^{-1} \Lambda_0 \tilde{\Delta}_1^{'} \\ &= \tilde{\Delta}_1 \tilde{\Delta}_1^{'} \tilde{\Delta}_1 \tilde{\Delta}_1^{'} \\ &= I_{\bar{r}} I_{\bar{r}}. \end{aligned}$$

Remember that

$$\begin{aligned} M_{2\beta} &= \tilde{\Delta}_2 \Gamma_0, \\ \tilde{\Delta}_2 &= \tilde{\Delta}_1 \left(\iota_J \otimes I_r\right), \end{aligned}$$

which leads to our final expression for the bound:

$$\begin{aligned} I^{-1}(\beta_0) &= M_{2\beta}^{'} V^{22} M_{2\beta} \\ &= \left[\Gamma_0^{'} (\iota_J \otimes I_r)^{'} \tilde{\Delta}_1^{'}\right] \tilde{\Delta}_1 \Lambda_0^{-1} \tilde{\Delta}_1^{'} \left[\tilde{\Delta}_1 (\iota_J \otimes I_r) \Gamma_0\right] \\ &= \Gamma_0^{'} (\iota_J \otimes I_r)^{'} \Delta \Lambda_0^{-1} \Delta (\iota_J \otimes I_r) \Gamma_0 \\ &= \Gamma_0^{'} \Delta_2^{'} \Lambda_0^{-1} \Delta_2 \Gamma_0. \end{aligned}$$

Step 5: Intermediate calculations

The validity of the bound calculations in Steps 3 and 4 require that M has full rank, that V is invertible, and that $M'V^{-1}M$ is blockdiagonal.

Rank of M. It follows from Theorem 4.2 in Meyer (1973) that M has full rank if $M_{1\rho}$ and $M_{2\beta}$ have full rank. If Γ_0 has full rank (K), then the $\bar{r} \times K$ matrix

$$M_{2\beta} = \tilde{\Delta}_2 \Gamma_0$$

because Assumption 3 guarantees that $\tilde{\Delta}_2$ has full rank. To see that $M_{1\rho}$ has full rank, see equation (A.27) below, from which it is clear that $M_{1\rho}$ is invertible because the probabilities are bounded away from 0.

Invertibility of V. To see that V is invertible, use the formula for the determinant of a block matrix,

$$\det(V) = \det(V_{11}) \det\left(V_{22} - V_{12}^{'} V_{11}^{-1} V_{12}\right)$$

and the fact that $V_{22} - V_{12}^{'} V_{11}^{-1} V_{12} = \tilde{\Delta}_1 \Lambda_0 \tilde{\Delta}_1^{'}$ (see (A.24)). Then, since V_{11} is invertible (see (A.21)) and the second term is invertible when Λ_0 is, V is invertible under Assumptions 1-3 and the conditions of Theorem 4.

Blockdiagonality of $M^{'}V^{-1}M$. Using Prokhorov and Schmidt (2009, Theorem 2.2, statement 9), blockdiagonality of $M^{'}V^{-1}M$ occurs when

$$V_{12}^{'} V_{11}^{-1} M_{1\rho} = M_{2\rho}. \tag{A.25}$$

From (A.23), we know that

$$V_{12}^{'}V_{11}^{-1} \quad = \quad \tilde{\Delta}_1 \sum_{l=1}^{L} \left(I_J \otimes q_l e_l^{'} \right). \tag{A.26}$$

Because $M_{1\rho}$ is invertible (this follows from the following derivations), we can proceed by showing that $M_{2\rho}M_{1\rho}^{-1}$ is also equal to the expression in (A.23).

First, note that

$$\frac{\partial m_{1j}}{\partial \rho_{j,0}} = -\frac{s_j}{(B^{'}\rho_{j,0})} BB^{'}$$

so that

$$E\left[\frac{\partial m_{1j}}{\partial \rho_{j,0}}\right] = -\sum_{l=1}^{L} \tau_l \frac{1}{\rho_{jl,0}} e_l e_l^{'}$$

and

$$M_{1\rho} = -\sum_{l=1}^{L} \tau_l R_{l,0}^{-1} \otimes e_l e_l^{'}. \tag{A.27}$$

Using a similar expression as for the inverse of V_{11} in (A.21), it obtains that

$$M_{1\rho}^{-1} = -\sum_{l=1}^{L} \frac{1}{\tau_l} R_{l,0} \otimes e_l e_l^{'}.$$

Second, we need an expression for $M_{2\rho}$. The ingredient is

$$\frac{\partial m_{2j}}{\partial \rho_{j,0}} = -\frac{s_j}{(B^{'}\rho_{j,0})} \tilde{d}_j \psi\left(Z, \beta\right) B^{'}$$

and its expectation is

$$E\left[\frac{\partial m_{2j}}{\partial \rho_{j,0}}\right] = -\tilde{d}_j \sum_{l=1}^{L} \tau_l \frac{1}{\rho_{jl,0}} q_l e_l^{'}.$$

Stacking these across strata j yields

$$M_{2\rho} = -\tilde{\Delta}_1 \sum_{l=1}^{L} \tau_l R_{l,0}^{-1} \otimes q_l e_l^{'}.$$

Finally, using matrix algebra tools similar to the ones used to simplify (A.22),

$$M_{1\rho}^{-1} M_{2\rho} = \sum_{l=1}^{L} I_J \otimes q_l e_l^{'}.$$

A.2 Proof of Theorem 7

Proof. This Theorem establishes the consistency and asymptotic normality of the plug-in estimator $\hat{\beta}$. The starting point for the asymptotic analysis is the recognition that the plugin estimator $\left(\hat{p}, \hat{\beta}\right)$ is equivalent to the optimal GMM estimator

By stacking the moment conditions for the (exactly identified) selection probabilities and the parameter of interest, β_0 , the plugin estimator is equivalent to joint estimation

It consists of two steps. First, I show that the optimal GMM estimator that jointly estimates the selection probabilities and the parameter β_0 is consistent, asymptotically normal, and that it attains the efficiency bound. In step 2, I show that the plugin estimator is asymptotically equivalent to that joint estimator.

Step 1. I first repeat here the unconditional moment conditions from Step 1 of Section A.1, and some required notation. We have assumed that $Z = (Y_1, X)$ follows a multinomial distribution. Let L be the number of support points of X, so that X takes values in $\{x_1, \cdots, x_L\}$. The $L \times 1$ vector B converts X into L binary variables

$$B = \left(1\left\{X = x_1\right\}, \cdots, 1\left\{X = x_L\right\}\right).$$

Denote the probability that a unit with $X = x_l$ selects into missing data pattern j by

$$\rho_{jl,0} = P\left(D = d_j \middle| X = x_l\right),$$

and stack the selection probabilities for pattern j into

$$\rho_{j,0} = \left(\rho_{j1,0}, \cdots, \rho_{jL,0}\right).$$

Then we can write

$$p_{j,0}\left(X\right) = B^{'}\rho_{j,0}.$$

Consider then the moment conditions:

$$E\left[m_1\left(\rho_0\right)\right] = E\begin{bmatrix} m_{11}\left(\rho_{1,0}\right) \\ \vdots \\ m_{1J}\left(\rho_{J,0}\right)\end{bmatrix} = E\begin{bmatrix}\begin{bmatrix} \frac{s_1}{B^{'}\rho_{1,0}} - 1 \\ \vdots \\ \frac{s_J}{B^{'}\rho_{J,0}} - 1\end{bmatrix} \otimes B\end{bmatrix} = \tag{A.28}$$

$$E\left[m_2\left(\rho_0, \beta_0\right)\right] = E\begin{bmatrix} m_{21}\left(\rho_{1,0}, \beta_0\right) \\ \vdots \\ m_{2J}\left(\rho_{J,0}, \beta_0\right)\end{bmatrix} = E\begin{bmatrix} \frac{s_1}{B^{'}\rho_{1,0}}\tilde{d}_1\psi\left(Z, \beta_0\right) \\ \vdots \\ \frac{s_J}{B^{'}\rho_{J,0}}\tilde{d}_J\psi\left(Z, \beta_0\right)\end{bmatrix} = \tag{A.29}$$

The dimension of m_1 is $JL \times 1$, and the dimension of m_2 is $\bar{r} \times 1$, where $\bar{r} = \sum_j r_j$ is the total number of components selected by the d_j's. Also note that the dimension of ρ_0 is equal to $JL \times 1$ (and that it has the same dimension as the the vector m_1), and that the dimension of β_0 is $K \leq r$.

Define $\left(\tilde{\rho}, \tilde{\beta}\right)$ as the optimal GMM estimator based on (A.28) and (A.29). The associated sequence of (optimal) weight matrices converges to V^{-1}. The consistency and asymptotic normality of $\left(\tilde{\rho}, \tilde{\beta}\right)$ is obtained by verifying the conditions of Theorems 2.6 and 3.4 in Newey and McFadden (1994). The numbering of their assumptions is in bold, and unbolded assumptions refer to the current paper.

For consistency, I verify the conditions of Theorem 2.6 of Newey and

McFadden. Assumption 2.1 guarantees that the data are **i.i.d**. **2.6(i):** Assumption 8.2 guarantees convergence of the weight matrix. The positive definiteness of its limit V^{-1} was established in Step 5 of the proof in Section A.1. The identification condition follows from this invertibility, in conjunction with Assumptions 1 and 3, and the fact that $p_j(x) \in [\kappa, 1]$, which is the overlap condition in Assumption 2.4. **2.6(ii)** corresponds to Assumption 8.1. **2.6(iii)** follows from Assumption 8.3. **2.6(iv)** is satisfied under the multinomial assumption in the conditions for Theorem 5.

For asymptotic normality, I verify the conditions of Theorem 3.4 of Newey and McFadden. Condition **3.4(i)** corresponds to Assumption 8.1. Condition **3.4(ii)** follows from Assumption 8.3, and the fact that $\frac{s_j}{p_j(X)}$ is differentiable in ρ_{jl}, and bounded (Assumption 2.4). **3.4(iii)** Assumption 1 implies that the moment conditions are mean zero: see also the construction of the IPW moment conditions in Section 4.1. The fact that the variance is bounded, and that condition **3.4(iv)** holds follows from the multinomial assumption for Theorem 5. Finally, the invertibility of **3.4(v)** is implied by the condition that Λ_0 in the conditions for Theorem 5.

Since the conditions for Newey and McFadden's Theorem 3.4 is satisfied, and because the sequence of weighting matrices is assumed to converge to V^{-1}, the limiting distribution of the optimal GMM estimator is

$$\sqrt{n}\begin{bmatrix}\tilde{\rho}-\rho_0\\ \tilde{\beta}-\beta_0\end{bmatrix} \overset{a}{\sim} \mathcal{N}\left(0, \left(M'V^{-1}M\right)^{-1}\right), \tag{A.30}$$

where the matrices M and V were defined in equations (A.3) and (A.5). The lower right block of $\left(M'V^{-1}M\right)^{-1}$ is $I^{-1}(\beta_0)$ was analyzed in the proof of Theorem 4. It follows that $\tilde{\beta}$ attains the efficiency bound from Theorem 4.

Step 2. The joint estimator $\left(\tilde{\rho}, \tilde{\beta}\right)$ above is the optimal estimator based on the possibly overidentifying moment conditions

$$E\begin{pmatrix} m_1(p_0) \\ m_2(p_0,\beta_0) \end{pmatrix} = 0. \tag{A.31}$$

An alternative estimator that is also efficient and therefore asymptotically equivalent to $\left(\tilde{\rho},\tilde{\beta}\right)$ is the estimator based on the just-identified, optimally weighted moment conditions A.31, namely the $JL+K$ moments:

$$\left(M^{'}V^{-1}M\right)^{1/2} E\begin{pmatrix} m_1(p_0) \\ m_2(p_0,\beta_0) \end{pmatrix} = 0. \tag{A.32}$$

Because $M^{'}V^{-1}M$ is blockdiagonal (see Step 5 in Section A.1), this system can be expressed as

$$E\begin{pmatrix} \tilde{m}_1(p_0) \\ \tilde{m}_2(p_0,\beta_0) \end{pmatrix} = E\begin{pmatrix} A_1 m_1(p_0) \\ A_2 m_2(p_0,\beta_0) \end{pmatrix} = 0.$$

The blockdiagonality preserves the structure of the problem, which is that the first set of moment conditions depends on p, and the second set of moment conditions depends on p and β. The dimension of the first block is equal to the dimension of p_0, and the dimension of the second block is equal to β_0. For that reason, by statement 6 of Prokhorov and Schmidt (2009, Theorem 2.2), the plug-in estimator in the text (Prokhorov and Schmidt's "two-step" estimator) is equal to the optimal estimator based on (A.32) (Prokhorov and Schmidt's "one-step" estimator). That estimator was asymptotically equivalent to the estimator analyzed in Step 1, and it follows that the plug-in estimator has the asymptotic distribution in (A.30). Therefore, the conclusion at the end of Step 1 applies to the plug-in estimator in the main text.

□