Efficient GMM estimation with incomplete data

Chris Muris*

July 1, 2018

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 (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 Introduction

Incomplete data, where some observations are missing some or all variables, is prevalent in empirical research in economics. For example, Abrevaya and Donald (2017) find that

*Previous title: "Efficient GMM estimation with general missing data patterns". Department of Economics, University of Bristol. Email: chris.muris@bristol.ac.uk. I am grateful to Ramon van den Akker, Richard Blundell, Otilia Boldea, Irene Botosaru, Pedro Duarte Bom, Katherine Carman, Matias Cattaneo, Miguel Atanasio Carvalho, Bryan Graham, Hide Ichimura, Toru Kitagawa, Tobias Klein, Andrea Krajina, Jan Magnus, Bertrand Melenberg, David Pacini, Krishna Pendakur, Franco Peracchi, Pedro Raposo, Sami Stouli, 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, Victoria University, and the Bristol Econometrics Study Group. I gratefully acknowledge financial support from the Social Sciences and Humanities Research Council through Insight Development Grant 430-2015-00073.

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:

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

Now consider a setting where either instrument can be unavailable. 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 $E[W_1 (y - X\beta_0)] = 0$ still contains information on β_0. The same is true for stratum 3 through $E[W_2 (y - X\beta_0)] = 0$. 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 incomplete 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 require 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.[1]

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. I also propose and analyze a doubly robust estimator.

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 6.1, 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 an efficient IPW estimator and a locally efficient doubly robust estimator. Section 6 contains a simulation study. Section 7 contains an empirical illustration.

2 Related literature

The literature on missing and incomplete 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

[1] However, I show in Section 5.3 that imputation may be useful if used in the context of doubly robust estimation.

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 incomplete 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 con-

ditions 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.[2] 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. More details on this comparison can be found in Appendix B.

Strand 2: Model-specific solutions. Several papers consider specific GMM settings or specific incomplete data patterns. For example, Abrevaya and Donald (2017) consider the linear regression model. 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 strata of incompleteness.[3] Abrevaya (2016) shows that the explanatory variables in the static model have information even when the associated dependent variable is unavailable.

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 data from some time periods are unavailable.

[2]The relationship between the multivalued treatment effect setting in Cattaneo (2010) and the incomplete data setting here is described in more detail in Appendix B.

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

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 unavailable 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. A more recent contribution by Dardanoni et al. (2011) shows that efficiency gains can be obtained if one is willing to sacrifice consistency.

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 require imputation: I propose an inverse propensity score weighting estimator that is consistent. Nevertheless, I also propose an approach that uses imputation, namely the doubly robust estimator in Section 5.3.

3 Model

This section formalizes the notion of "incomplete data" in this paper, and introduces identification and sampling assumptions that are used throughout the paper.

3.1 Incomplete data

The incomplete data framework starts from moment conditions for complete data. Let $Z = \left(Y_1', X'\right)$ be a random vector of data, 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 $p \geq K$.[4] The true value of the parameter, $\beta_0 \in \mathcal{B} \subset \mathbb{R}^K$, is defined by:

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

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

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

The following three examples illustrate the setup. Additional examples can be found in in Appendix C.

Example 2 (Linear IV)**.** Consider a linear instrumental variables model with a dependent variable y, two endogenous variables $X = (X_1, X_2)$, and two instruments $W = (W_1, W_2)$. Set $Z = (y, X, W)$ and define the moment function

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

so that the parameter vector β_0 is defined through the moment condition

$$E(\psi(Z,\beta_0)) = 0. \tag{3.1}$$

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

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

For d_1, this corresponds to observing all variables:

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

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

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

Similarly, in the stratum with $D = d_3$, only the second instrument W_2 is observed. Finally, the stratum with $D = d_4$ corresponds to the observations for which both instruments are unavailable, or for which the dependent variable or one of the regressors is not observed.

Models with multiple, incompletely observed, instruments are relevant for applied practice. Some examples include Card (1995), Acemoglu et al. (2001), Rodrik et al. (2005), and Angrist et al. (2010). Methodological contributions include Abrevaya and Donald (2011), Mogstad and Wiswall (2012), and Feng (2016).

Example 3 (Rotating dynamic panel)**.** Consider a five-period fixed effects autoregressive distributed lag model with regression equation:

$$Y_{it} = \alpha_i + \rho Y_{i,t-1} + X_{it}\beta_1 + X_{i,t-1}\beta_2 + u_{it},\ t = 1, \cdots, 5. \tag{3.2}$$

Because of the presence of the fixed effects α_i, estimation of the parameters $\theta = (\rho, \beta_1, \beta_2)$ is based on the regression equation in first differences:

$$\Delta Y_{it} = \rho \Delta Y_{i,t-1} + \Delta X_{it}\beta_1 + \Delta X_{i,t-1}\beta_2 + \Delta u_{it},\ t = 2, \cdots, 5. \tag{3.3}$$

In the estimation of empirical growth models, and in the estimation of production functions, it is typically assumed that[5]

$$E\left[\Delta u_{it} \middle| Y_{i,t-3}, Y_{i,t-4}, X_{i,t-3}, X_{i,t-4}\right] = 0.$$

For a hypothetical unit with five time periods, we have:

$$E\begin{bmatrix} Y_{i2}\Delta u_{i5} \\ X_{i2}\Delta u_{i5} \\ Y_{i1}\Delta u_{i5} \\ X_{i1}\Delta u_{i5} \\ Y_{i1}\Delta u_{i4} \\ X_{i1}\Delta u_{i4} \end{bmatrix} = 0.$$

Assume that a rotating panel is available. There are two cohorts, each providing four

[5]Further lags of the dependent and explanatory variables would also qualify as instruments, but are not available for any t because we are only considering five time periods. Closer lags are not valid instruments due to measurement error and endogeneity.

consecutive time periods. The first cohort enters the sample in period 1 and leaves in period 4. The second cohort enters the sample in period 2 and leaves in period 5. In that case, the incomplete data indicators are

$$\widetilde{d}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \widetilde{d}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Example 29 in Appendix C discusses a closely related dynamic panel model with more complex pattern of missingness. Such examples are abundant in empirical work, see for example the applications in Arellano and Bond (1991), Schularick and Steger (2010), Topalova and Khandelwal (2011), Acemoglu et al. (2015), Acemoglu et al. (2018), among many others.

Example 4 (Panel binary choice)**.** Consider a three-period fixed effects logit model for the dependence of a sequence of binary outcomes $Y_i = (Y_{i1}, Y_{i2}, Y_{i3})$ on k-dimensional covariates $X_i = (X_{i1}, X_{i2}, X_{i3})$, with conditional choice probabilities

$$P\left(Y_{it} = 1 | X_i, \alpha_i\right) = \Lambda\left(\alpha_i + X_{it}\beta\right), \, t = 1, 2, 3.$$

With complete data, estimation of the common parameters proceeds by conditional maximum likelihood, based on the conditional probability

$$P\left(Y_i = y | \sum_t y_t = c, X_i\right) = \frac{\exp\left(\sum_t y_t X_{it}\beta\right)}{\sum_{d \in B_c} \exp\left(\sum_t d_t X_{it}\beta\right)}, \tag{3.4}$$

where B_c is the set of all sequences d with $\sum_t d_t = c$, see Chamberlain (1980) and Cameron and Trivedi (2005, p. 2338). Estimation of β based on (3.4) requires that all time periods are available for each individual.

I am not aware of any available estimator for β that allows for data to be incomplete at random.[6] However, the present framework easily accommodates this setting. Con-

[6]For example, Papke and Wooldridge (2008, p. 127) write: "The nonlinear models we apply are difficult to extend to unbalanced panel data – a topic for future research." Their discussion indicates Papke (2005) as an application of the methodology developed here.

sider a combination of two distinct time periods $\{(s,t) : 3 \geq t > s \geq 1\}$. The random variables $(Y_{is}, Y_{it}, X_{is}, X_{it})$ follow a two-period binary choice model, with conditional probability $P(Y_{it} = 1 | Y_{is} + Y_{it} = 1, X_i) = \Lambda((\Delta_{st} X_i)\beta)$, where $\Delta_{st} X_i = X_{it} - X_{is}$, i.e. a cross-sectional logit for a subpopulation of switchers. The score is

$$E[A_{i,st}(\Delta_{st} X_i)(Y_{it} - \Lambda((\Delta_{st} X_i)\beta))] = 0, \tag{3.5}$$

where $A_{i,st} = 1\{Y_{is} + Y_{it} = 1\}$.

The three-period model implies three such two period models, and $3k$ moment conditions:

$$E\begin{bmatrix} A_{i,12}(\Delta_{12} X_i)(Y_{i2} - \Lambda((\Delta_{12} X_i)\beta)) \\ A_{i,13}(\Delta_{13} X_i)(Y_{i3} - \Lambda((\Delta_{13} X_i)\beta)) \\ A_{i,23}(\Delta_{23} X_i)(Y_{i3} - \Lambda((\Delta_{23} X_i)\beta)) \end{bmatrix} = 0. \tag{3.6}$$

For a cross-section unit with complete data the incomplete data indicator is $d_1 = I_{3k}$: all moment functions can be computed. For a cross-section unit that drops out after period 2 (attrition), $d_2 = e_{1,3} \otimes I_k$. For a cross-section unit that enters the sample in period 2, $d_3 = e_{3,3} \otimes I_k$. For a cross-section unit that is not observed in period 2, $d_4 = e_{2,3} \otimes I_k$. A cross-section unit that misses more than one period has $d_5 = O_{3k}$.

Applied research uses panel binary choice models sparingly, but the approach outlined in this example transfer to any panel model with unbalanced data. Unbalanced panels are ubiquitous in applied work across fields, see e.g. Topalova and Khandelwal (2011), de Loecker and Warzynski (2012), Becker and Woessmann (2013), Sturm and De Haan (2015), and Yagan (2015), among many others.

3.2 Identification

The following assumption guarantees identification for the incomplete data setting, given that identification holds for complete data, i.e. Assumption 1 holds.

Assumption 5. *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 5 rules out situations in which a component of ψ is never observed. If Assumption 5 fails, the analysis may proceed after removing the never-observed components from ψ, as long as Assumption 1 holds for the reduced set of moment conditions.

Assumption 5 can hold *even if identification fails in every stratum.* This is an important distinction between the setup here and the multivalued treatment framework in Cattaneo (2010), see Appendix B. The following examples illustrates this for two distinct cases: (i) there exists at least one stratum in which the parameters are identified; (ii) identification fails in every stratum. In case (i), standard results from the MAR setup can be applied to one of those strata, but the resulting procedure will be less efficient than the estimators proposed below. In case (ii), the results proposed below are required for identification.

Example (Linear IV, continued)**.** Recall Example 2. Existing results for missing data can be used to define an estimator based on the subpopulation with both instruments observed (stratum 1, with $d_1 = I_2$). The results in the present work can be applied to obtain more efficient procedures, see the analysis in Section 4.1.

Now consider Example 27 in Appendix C, which differs from Example 2 because no complete observations are available. Instead, for every observation, exactly one instrument is available. This corresponds to strata 2 and 3, with

$$d_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, d_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Note that $d_2 + d_3 = I_2$, so that Assumption 5 is satisfied, even though identification fails in every stratum. The results below can be used to obtain a consistent and efficient estimator that deals with this problematic data setting.

Example (Rotating dynamic panel, continued)**.** Recall Example 3. For the first stratum, only two moment conditions are available to three parameters: stratum identification does not hold. Similarly, stratum identification does not hold for the second stratum. Furthermore, Assumption 5 does not hold for this formulation, since $\widetilde{d}_1 + \widetilde{d}_2 \neq I_6$. In other words, two of the moment functions are not computable for any individual. For this reason, reduce the moment conditions to

$$\psi(Z_i, \theta) = \begin{bmatrix} Y_{i2}\Delta u_{i5} \\ X_{i2}\Delta u_{i5} \\ Y_{i1}\Delta u_{i4} \\ X_{i1}\Delta u_{i4} \end{bmatrix}$$

so that

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

and Assumption 5 is satisfied for the reduced set of moment conditions.

The framework in this paper can now be applied directly to estimate the parameters in the ADL model. Existing results in dynamic panel models suggest that five time periods are required for identification. However, the results below show that identification can be obtained using a rotating panel with four periods per individual. An efficient estimator for the parameters in that model follows immediately. This case is investigated in a simulation study in Section 6.1.

Example (Panel binary choice, continued)**.** Recall Example 4. For this model, the results in this paper are not necessary for identification: the researcher could simply discard strata 2 through 5, and apply results for the standard MAR setup to the balanced subpanel (stratum 1, $d_1 = I_{3k}$). However, the efficiency gains can be substantial when the probability of missingness is large, as will be demonstrated using Monte Carlo simulations in Section 6. Similar efficiency gains may be obtained using the results in Cattaneo (2010, Section 5.5).[7]

3.3 Sampling

The remainder of the paper analyzes the efficient estimation of β_0 under the following restrictions on the sampling design and data availability.

Assumption 6. *(i) Random sampling:* $\{(Z_i, D_i), i = 1, \cdots, n\}$ *is an independent and identically distributed sequence; (ii) the researcher observes* D_i, X_i, *and* $D_i\psi(Z_i, \beta)$ *for all* $\beta \in \mathcal{B}$*; (iii) missing at random:* $Y_1 \perp D | X$*; (iv) overlap: there exists a* $\kappa > 0$ *such that*

$$p_{j,0}(x) = P(D = d_j | X = x) \geq \kappa \tag{3.7}$$

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

[7] Strictly speaking, this would require an extension of the results in Cattaneo (2010) that allows the moment conditions to depend on the stratum. An inspection of his proofs suggests that such an extension is straightforward. See Appendix B for more details on the relationship between Cattaneo's results and those in the present manuscript.

This assumption generalizes the standard assumptions for missing data, in which an observation is either complete or completely missing. In particular, Assumptions 1, 5 and 6 reduce to the standard missing at random (MAR) setup if $J = 1$ and $d_1 = I_p$, see e.g. Graham, 2011. In what follows, I will refer to that case as "missing data" or "the standard MAR setup". One difference with the standard MAR setup is that the conditional independence assumption in part (iii) could be generalized to let the conditioning covariates vary by stratum, using the results in Hristache and Patilea (2016).

The MAR assumption (iii) says that all observable data must be independent of what subset of data is available, conditional on some confounders X. This assumption is best understood in the context of an example. In the linear IV example, MAR requires instrument availability to be independent of the value of the instruments, the covariates, and the error term in the model - once a set of conditioning convariates is taken into account. In the context of the panel binary choice model, it requires that the availability of data for a given cross-section unit in a certain period is independent of the fixed effect of that individual, and that it is also independent of the covariates and error terms in all time periods.[8]

4 Efficiency bound

Assumptions 1 and 6 imply a set of conditional and unconditional moment conditions for each stratum. For each $j \in \{1, \cdots, J\}$, define the stratum indicator $s_j = 1\{D = d_j\}$. The conditional moment restrictions

$$E\left[\frac{s_j}{p_{j,0}(X)} - 1 \middle| X\right] = 0 \text{ for all } j = 1, \cdots, J, \tag{4.1}$$

define the propensity scores (3.7). In the standard MAR setup, Graham (2011) refers to such moment conditions as "auxiliary moments". Furthermore, the unconditional

[8]This assumption is stronger than needed. The crucial assumption on independence is that the moment functions are mean-independent of the incomplete data indicator conditional on the confounders. For example, in the linear IV example, the MCAR assumption can be weakened to: 'in each stratum, the observable instruments should be valid'. However, with some effort, one can construct examples where identification fails under this weaker mean-independence assumption. For this reason, the stronger MAR assumption is maintained in this manuscript.

moment restrictions

$$E\left[\frac{s_j}{p_{j,0}(X)}d_j\psi(Z,\beta_0)\right]=0,\ j=1,\cdots,J, \tag{4.2}$$

hold. These generalize Graham's "identifying moments" to the incomplete data context. The sample analogs of moment conditions (4.1) and (4.2) can be computed with the available data (Assumption 6(ii)).

In what follows, denote by

$$\Gamma_0 \equiv \frac{\partial E[\psi(Z,\beta_0)]}{\partial \beta_0} \tag{4.3}$$

the expected derivative of the moment functions evaluated at the truth, if it exists. Also, denote by

$$\Sigma_0(X) \equiv \mathrm{Var}[\psi(Z,\beta_0)|X] \tag{4.4}$$

the conditional variance of the moment function.

Assumption 7. *(i) The distribution of Z has known, finite support; (ii) $\mathcal{B}$ is open, and there exists a $\beta_0 \in \mathcal{B}$ and $0 < p_{j,0} < 1$, $j = 1,\cdots,J$, such that (4.1) and 4.2 hold; (iii) ψ is continuously differentiable on Θ for all values in the support of Z, and Γ_0 has full rank; (iv) $\Sigma_0(x)$ is invertible for all $x \in supp(X)$.*

These assumptions translate the requirements for Lemma 2 in Chamberlain (1987) and Theorem 1 in Graham (2011) to the incomplete data setting. Below, I follow their results in constructing a semiparametric efficiency bound. Part (i) imposes that the data follow a multinomial distribution. The estimators I propose below do not require this, and still achieve the bound in the upcoming Theorem 8. Remark 10 provides some additional discussion on this restriction. Part (ii) is not restrictive. Part (iii) is a strong assumption on the smoothness of the moment function. In the large sample theory developed in the remainder of this paper, this assumption is relaxed. The proposed estimators allow for nonsmooth moment conditions, and still achieve the efficiency bound. Part (iv) requires sufficient variation in the conditional moments, which is readily checked in a given application.

Theorem 8 (Efficiency bound)**.** *If Assumption 7 holds, then the information bound for*

any regular estimator for β_0 is given by

$$I_0(\beta_0) = \Gamma_0' \left(\sum_j (d_j \Omega_j d_j)^+ \right) \Gamma_0, \tag{4.5}$$

where

$$\Omega_j = E\left[\frac{\Sigma_0(X)}{p_{j,0}(X)} + q_0(X) q_0'(X) \right], \tag{4.6}$$

$$q_0(X) = E[\psi(Z, \beta_0) | X]. \tag{4.7}$$

Proof. See Appendix A.1. □

Section 4.1 provides an interpretation for this bound using the linear IV example. For an interpretation in the general context, recall the information bound for the binary missing data case, see e.g. Graham (2011):

$$I_{\text{MD}} = \Gamma_0' \Omega_1^{-1} \Gamma_0, \tag{4.8}$$

where Ω_1 is a stratum-specific variance as in (4.6), for the complete-data stratum with $p_{j,0} = p_0$ the standard propensity score, and $d_1 = I_p$. First, note that the new bound in 4.5 is therefore a generalization of the bound for the MAR setup with $J = 1$, $d_1 = I_p$.

Second, note that the contribution of stratum j to the information bound is

$$I_j(\beta_0) = \Gamma_0' (d_j \Omega_j d_j)^+ \Gamma_0, \tag{4.9}$$

in the sense that $I_0(\beta_0) = \sum_j I_j(\beta_0)$. Compare (4.8) and (4.9): the new bound in (4.5) thus has the interpretation that it is the sum of the information in J separate binary missing data problems - one for each stratum.

Remark 9. The bound is reminiscent of the bound for the multivalued treatment effects, cf. Cattaneo (2010). Appendix B explores the relationship between the two frameworks in detail, see also "Strand 1" in the literature review. To make a comparison of the bounds, we must consider the case where $d_j = I_p$ for all j. Then the observation in Section 5.5 in Cattaneo (2010) can be applied. In the framework of that section, set their π equal to β_0 in this manuscript, and set take $\beta(\pi) = (\pi, \cdots, \pi)$ so that $\partial\beta(\pi^*) = \iota_J \otimes I_p$. The equivalence of the bounds then follows immediately.

Remark 10. The bound in Theorem 8 is for discrete data (Assumption 7(i)). This is an approach that follows Chamberlain (1987), see also Chamberlain (1992a, 1992b) and Graham (2011). An alternative approach would avoid the multinomial assumption.[9] However, the bound in (4.5) can be shown to apply to arbitrary distributions.[10]

4.1 Linear IV case

Consider Example 2 (linear IV) from Section 3, with a set of instruments $W = (W_1, W_2)$ and an error term $u = y - X\beta_0$, $\beta_0 \in \mathbb{R}$ such that the moment conditions are given by

$$E\left[\psi\left(Z, \beta_0\right)\right] = E\begin{bmatrix} W_1 u \\ W_2 u \end{bmatrix} = 0. \tag{4.10}$$

Either instrument can be missing, so $J = 3$ and the incomplete data indicator has support

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

Some additional restrictions will allow us to compare the efficiency bound in (4.5) to several estimators in common use. First, assume that $X = 1$, i.e. incompleteness is completely at random; that the each instrument is missing with probability p, so that $p_{10} = (1-p)^2$ and $p_{20} = p_{30} = p(1-p)$. Second, unbeknownst to the researcher, let $E\left(u^2 | W\right) = \sigma^2$. Then

$$E\left[\psi\left(Z, \beta_0\right)\psi\left(Z, \beta_0\right)'\right] = \sigma^2 E\left[WW'\right] = \sigma^2 \Sigma_Z = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Finally, assume that the instruments are equally correlated with the endogenous variable, $E\left[WX\right] = \sigma_{xw}\iota_2$, with ι the unit vector.

The expression for the bound now simplifies because $q_0\left(X\right) = 0$ and $\Omega_j = \frac{\sigma^2}{1-p_{j0}}\Sigma_Z$, and the expected derivative is $\Gamma_0 = -\sigma_{xw}\iota_2$. The contribution of stratum j to the

[9] See Bickel et al., (1993), Hahn (1998), Chen et al. (2008), Cattaneo (2010). The lack of invertibility apparant from (4.9) creates some technical difficulties in this approach.

[10] See Theorem 2 in Chamberlain (1987) for the unconditional case; Theorem 3 for conditional case. Demonstrating that it can also be done for the mixed conditional/unconditional case is beyond the scope of this paper.

information bound (4.9) is therefore given by

$$I_j(\beta_0) = \frac{\sigma_{xw}^2}{\sigma^2}(1 - p_{j0})\,\iota_2'\left(d_j \Sigma_Z d_j\right)^+ \iota_2.$$

For stratum 2 and 3,

$$I_2(\beta_0) = I_3(\beta_0) = \frac{\sigma_{xw}^2}{\sigma^2}\left(1 - p(1-p)\right). \tag{4.11}$$

For the full data stratum,

$$I_1(\beta_0) = \frac{2\sigma_{xw}^2}{\sigma^2}\frac{(1-p)^2}{1+\rho}.$$

We can now conclude two things. First, the ratio of information in the incomplete strata 2 and 3 relative to stratum 1 is

$$\frac{I_2 + I_3}{I_1} = (1+\rho)\frac{1 - p + p^2}{1 - 2p + p^2}.$$

If $\rho = 0$, the two incomplete strata contain more information than the complete one, demonstrating that the information in the incomplete strata is not negligible. Second, the information bound is

$$I_0(\beta_0) = \sum_j I_j(\beta_0) = \frac{2\sigma_{xw}^2}{\sigma^2}\left(\frac{(1-p)^2}{1+\rho} + (1 - p(1-p))\right). \tag{4.12}$$

We wish to compare this bound for an *optimal* estimator to a few reasonable alternatives. First, the *complete case estimator* (CC) uses only observations with both instruments. This corresponds to the standard MAR setup, and using only stratum 1, so that $I_{\text{CC}}(\beta_0) = I_1(\beta_0)$. Second, the infeasible *full data* (FD) estimator with both instruments always available, with information corresponding to the standard bound for (4.10), $I_{\text{FD}} = I_1(\beta_0)/(1-p)^2$. Third, the *available case* estimator which replaces all instruments by zeros. This amount to estimating each of the moment functions using all the observations for which that moment function is observed, with information is

$$I_{\text{AC}}(\beta_0) = \frac{2\sigma_{xw}^2}{\sigma^2} \times \frac{1-p}{1+\rho}.$$

This corresponds to using the moment conditions $E[DWu] = 0$. To see this, note that

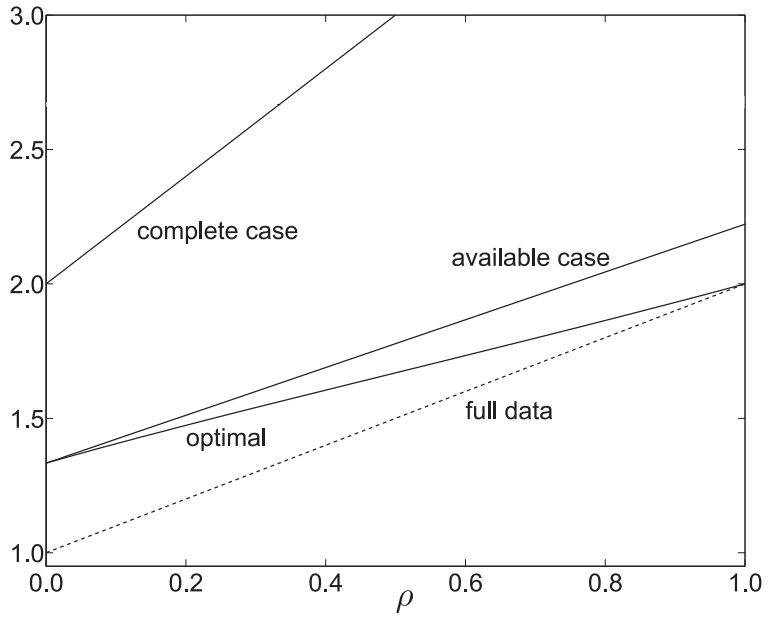

Figure 4.1: Information for different sets of moment conditions, as a function of ρ, for $p_1 = 0.5$.

$\mu_D \equiv E[D] = \begin{bmatrix} 1-p & 0 \\ 0 & 1-p \end{bmatrix}$, so the available case moment conditions have derivative $-\sigma_{xw}\mu_D\iota_2 = -\sigma_{xw}(1-p)\iota_2$ and variance $\sigma^2(1-p)\Sigma_Z$.

In Figure 4.1 we plot the asymptotic variance of the estimators, including an *optimal* one that achieves the efficiency bound in 4.5, as a function of ρ for $p = 0.5$. The key aspect of this analytical comparison is that the two instruments provide similar sources of information. Therefore, as ρ increases, 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.

5 Estimation

Assume that for each stratum $j = 1, \cdots, J$, an estimator $\widehat{p}_j$ for the propensity score $p_{j,0}$ is available. Estimation of β can then be based on a matrix-weighted average of sample analogs of the feasible moment conditions (4.2) with the propensity score estimators $\widehat{p}_j$

plugged in. The matrix weights $A_{j,n}$ are sequences of random $K \times p$ matrices, which lead to the $K-$dimensional sample criterion function:

$$G_n(\beta) = \sum_j A_{j,n} \frac{1}{n} \sum_{i=1}^{n} \frac{s_{ij}}{\hat{p}_j(X_i)} d_j \psi(Z_i, \beta). \tag{5.1}$$

The IPW estimator $\widehat{\beta}_n$ is defined as the value of β that sets that function equal to zero:

$$G_n\left(\widehat{\beta}_n\right) = 0. \tag{5.2}$$

In what follows, we will use $\|A\| = \sqrt{\operatorname{tr}(A'A)}$ to denote the matrix norm for any matrix A. For a function $f : \mathbb{D} \to \mathbb{R}$, denote by $\|f\|_\infty$ its sup-norm $\|f\|_\infty = \sup_{x \in D} |f(x)|$.

5.1 Consistency

To establish consistency of the proposed estimator, we require some conditions on the propensity score estimators, the weight matrices $A_{j,n}$ and their limits, on the function ψ, and on the parameter space $\mathcal{B}$.

Assumption 11. *For each $j = 1, \cdots, J$, the propensity score estimator is consistent:*

$$\|\widehat{p}_j - p_{j,0}\|_\infty = o_p(1).$$

Assumption 11 requires the propensity score estimators to be consistent. Cattaneo (2010, Appendix B) proposes a multinomial logistic series estimator that satisfies Assumption 11 under mild conditions on the regressors. It can be used without modification in the present context.

Assumption 12. *For each j, there exists a $K \times p$ matrix A_j such that (i) $\|A_{j,n} - A_j\| = o_p(1)$; (ii) $A_j d_j = A_j$; and (iii) $rk(A) = K$, where $A = \sum_j A_j$.*

Part (i) is standard. Parts (ii) and (iii) are necessary for identification. They restrict the choice of limiting weights A_j to prevent underidentification. This could happen if A_j assigns zero weight to moment conditions for which the corresponding elements d_j are non-zero. If A_j is chosen as the non-zero rows of d_j, part (iii) reduces to Assumption 5.

Assumption 13. *(i) The class of functions $\{\psi(\cdot, \beta), \beta \in \mathcal{B}\}$ is Glivenko-Cantelli; (ii) $E\left[\sup_{\beta \in \mathcal{B}} \|\psi(Z, \beta)\|\right] < \infty$; (iii) $E[\psi(Z, \beta)]$ is continuous; (iv) $\mathcal{B}$ is compact.*

Part (i) guarantees the uniform convergence of sample averages of the original moment function ψ to its expectation. Together with part (ii) and the assumptions on the propensity scores and their estimators, it implies uniform convergence of the sample criterion function (5.1). Parts (iii) and (iv), combined with the limiting objective function having a unique zero, guarantee that the minimum of the limiting objective function is well-separated (see proof for details). These restrictions are mild. It allows for moment functions that are discontinuous, e.g. a maximum score estimator for the panel data binary choice model with attrition and refreshment.

Theorem 14 (Consistency of IPW estimator)**.** *Under Assumptions 1, 5, 6, 11, 12, and 13,*

$$\widehat{\beta}_n \xrightarrow{p} \beta_0 \text{ as } n \to \infty.$$

5.2 Asymptotic normality

We impose some additional smoothness assumptions on ψ to establish $\sqrt{n}-$asymptotic normality of $\widehat{\beta}_n$ in (5.2).

Assumption 15 (Differentiability)**.** $E\left[\psi\left(Z,\beta\right)\right]$ *is differentiable in* β *at* β_0*, and the derivative* Γ_0 *has full rank.*

Assumption 16. *For some* $\delta > 0$*: (i) The class of functions* $\left\{\psi\left(\cdot,\beta\right), \|\beta-\beta_0\| < \delta\right\}$ *is Donsker; (ii) the second moment is locally uniformly bounded:*

$$E\left[\sup_{\|\beta-\beta_0\|<\delta} \|\psi\left(Z,\beta\right)\|^2\right] < \infty.$$

These assumptions are adapted from Cattaneo (2010, Assumption 6), who uses them in the context of multi-valued treatment effects. They guarantee that the criterion function satisfies stochastic equicontinuity, taking into account the estimated propensity score. These smoothness assumption are mild, requiring differentiability only after smoothing by taking expectations, and requiring it only at the truth. It rules in, among others, a modification of the instrumental variable quantile regression estimator for incomplete data.

Theorem 17 (Limiting distribution of the IPW estimator)**.** *Under the conditions of Theorem 14, Assumption 15, and 16, and*

$$\left\|\widehat{p}_j - p_{j,0}\right\|_\infty = o_p\left(n^{-1/4}\right),$$

then for any β_0 in the interior of $\mathcal{B}$,

$$\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\Gamma_0' A' V_A^{-1} A \Gamma_0\right)^{-1}\right), \tag{5.3}$$

where

$$V_A = \sum_j A_j \Omega_j A_j', \tag{5.4}$$

with Ω_j as in (4.6).

Remark 18 (Efficiency of the IPW estimator). The asymptotic variance is minimized by seting $A_j^* = \Gamma_0' \left(d_j \Omega_j d_j\right)^+$. This resembles the usual optimal choice of weights in moment-based estimation, with the exception of the d_j which guarantee that only observable moment functions are selected for each stratum. Call the resulting estimator $\widehat{\beta}_n^*$. Then

$$\sqrt{n}\left(\widehat{\beta}_n^* - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I_0^{-1}\left(\beta_0\right)\right),$$

i.e. it achieves the semiparametric efficiency bound derived in (4.5).

5.3 Doubly robust estimation

For the doubly robust estimator, the researcher uses possibly misspecified working models for the propensity score and the conditional expectation function.[11] Posit a working model for the propensity scores,

$$p_j\left(X\right) = \zeta_{1j}\left(h_1\left(X\right)\gamma_j\right),\ j = 1, \cdots, J, \tag{5.5}$$

where $h_1\left(X\right)$ is a $K_1 \times 1$ transformation of the confounders X, and the γ_j are the associated regression coefficients.

Posit a working model for the conditional expectation function

$$q_0\left(X\right) = \zeta_{2\beta}\left(h_2\left(X\right)\delta\right),\ \beta \in \mathcal{B}, \tag{5.6}$$

where $h_2\left(X\right)$ is some $K_2 \times 1$ vector of transformations $h_2\left(X\right)$ with regression coefficient

[11] There is a large literature on doubly robust estimation. Some contributions closely related to current setup include Cattaneo (2010), Tan (2010), Graham (2011), Graham et al. (2012), Graham et al. (2016), and Rothe and Firpo (2017).

δ.

Assumption 19 (Correct parametric specification)**.** *(i) For each $j = 1, \cdots, J$, there exists a $\gamma_{j,0} \in \mathbb{R}^{K_1}$ such that $p_{j,0}(X) = \zeta_{1j}(h_1(X)\gamma_{j,0})$ a.s.; (ii) there exists a $\delta_0 \in \mathbb{R}^{K_2}$ such that for all $\beta \in \mathcal{B}$, $q_0(X, \beta) = \zeta_{2\beta}(h_2(X)\delta_0)$ a.s.*

Assumption 19(i) holds if the propensity score working model is correctly specified. This restriction is more stringent than in the usual missing data case: the model must be correctly specified for all strata. Assumption 19(ii) requires the working model for the conditional expectation function to be correctly specified for all β. This is a standard requirement in the analysis of parametric doubly robust estimators.

Assumption 20. *(i) For each $j = 1, \cdots, J$, there exists an estimator $\widehat{\gamma}_{j,n}$ such that $\widehat{\gamma}_{j,n} \xrightarrow{p} \gamma_0$ and $\sqrt{n}(\widehat{\gamma}_{j,n} - \gamma_0) \xrightarrow{d} \mathcal{N}(0, \Omega_{\gamma,j})$ and (ii) there exists an estimator $\widehat{\delta}_n$ such that $\widehat{\delta}_n \xrightarrow{p} \delta_0$ and $\sqrt{n}\left(\widehat{\delta}_n - \delta_0\right) \xrightarrow{d} \mathcal{N}(0, \Omega_\delta)$.*

Assumption 20 requires that estimators are available that are consistent and asymptotically normal at the parametric rate. This is not a restrictive assumption: the parameters in the working models can typically be estimated using maximum likelihood (for $\gamma_{j,0}$) and nonlinear least squares (for δ_0).

Consider the criterion function

$$G_n^{DR}(\beta) = \sum_j A_{j,n} \frac{1}{n} \sum_{i=1}^n \left(\frac{s_{ij} d_j \psi(Z_i, \beta)}{\zeta_{1j}(h_1(X_i)\widehat{\gamma}_{j,n})} - \frac{s_{ij} - \zeta_{1j}(h_1(X_i)\widehat{\gamma}_{j,n})}{\zeta_{1j}(h_1(X_i)\widehat{\gamma}_{j,n})} d_j \zeta_{2\beta}\left(h_2(X_i)\widehat{\delta}_n\right) \right) \tag{5.7}$$

and define the doubly robust estimator $\widetilde{\beta}_n$ to be the solution to

$$G_n^{DR}\left(\widetilde{\beta}_n\right) = 0. \tag{5.8}$$

On top of inverse propensity score weighting, the DR estimator makes a covariate adjustment based on an estimate of the conditional expectation function.

Assumption 21. *(i) The class of functions $\{\zeta_{2\beta}(h_2(\cdot)\delta_0), \beta \in \mathcal{B}\}$ is Glivenko-Cantelli and $E\left[\sup_{\beta \in \mathcal{B}} \|\zeta_{2\beta}(h_2(X)\delta_0)\|\right] < \infty$; (ii) there exists a $\widetilde{\kappa} > 0$ such that $\zeta_{1j}(X\gamma_{j,0}) \geq \widetilde{\kappa}$ for all j.*

These assumptions guarantee that (5.7) converges uniformly to its limit uniformly. Given that the researcher has control of the working models, this is not a restrictive assumption.

Assumption 22. *(i) for each $j = 1, \cdots, J$, the link function $\zeta_{1j}(\cdot)$ has a derivative ζ_{1j}', and there exists an $\epsilon_1 > 0$ such that $\sup_{\|\gamma_j - \gamma_{j0}\| < \epsilon_1} E\left[\left\|\zeta_{1j}'(h_1(X)\gamma_j) h_1(X)\right\|\right] < \infty$; (ii) for each $\beta \in \mathcal{B}$, there exists an $\epsilon_2 > 0$ such that $\sup_{\|\delta - \delta_0\| < \epsilon_2} E[\|\zeta_{2\beta}(h_2(X_i)\delta)\|] < \infty$; (iii) for each $\beta \in \mathcal{B}$, the link function $\zeta_{2\beta}$ has a derivative $\zeta_{2\beta}'$, and there exists an $\epsilon_3 > 0$ such that $\sup_{\|\delta - \delta_0\| < \epsilon_3} E\left[\left\|\zeta_{2\beta}'(h_2(X)\delta) h_2(X)\right\|\right] < \infty$.*

Assumption 22 imposes some conditions on the working models that guarantee that the resulting class of criterion functions is well-behaved. The smoothness assumptions on the working models are stronger than those for the original moment functions, which is reasonable given that the working models are under the control of the researcher.

Theorem 23 (DR consistency)**.** *If Assumptions 1, 5, 6, 12, 13, and 20 hold, and that at least one of Assumption 19(i) or 19(ii) holds. Then $\widetilde{\beta}_n \xrightarrow{p} \beta_0$.*

Theorem 23 provides conditions under which $\widetilde{\beta}_n$ is consistent. In particular, it shows that $\widetilde{\beta}_n$ is indeed doubly robust: only one of the working models needs to be correct for consistency. For asymptotic normality and inference, some additional structure is imposed on the model. The result can be generalized to non-smooth settings by using techniques similar to those in Theorem 17.

Assumption 24 (Additional smoothness)**.** *There exists a $\delta > 0$ such that (i) $\psi(\cdot, \beta)$ is continuously differentiable with respect to β on $\|\beta - \beta_0\| < \delta$, and (ii) $E\left[\sup_{\|\beta - \beta_0\| < \delta} \|\psi(Z, \beta)\|\right] < \infty$.*

Theorem 25. *If the conditions for Theorem 23 are satisfied, and Assumption 24 holds, then*

$$\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\Gamma_0' A' V_A^{-1} A \Gamma_0\right)^{-1}\right),$$

where V_A is as in Theorem 17.

This theorem says that the DR estimator with correctly specified parametric propensity score and conditional expectation function obtains the same limiting distribution as the IPW estimator with nonparametric propensity score. If the limiting weight matrices are chosen as in Remark 18, then the DR estimator is locally efficient.[12]

[12] It attains the efficiency bound (4.5), which does not incorporate the knowledge about the parametric models, if both parametric models are correctly specified.

5.4 Inference

If the working models for the propensity score and the outcome equation are correctly specified, the limiting distributions of the IPW and DR estimators coincide. Recall that the asymptotic variance is given by $\left(\Gamma_0' A' V_A^{-1} A \Gamma_0\right)^{-1}$, where $V_A = \sum_j A_j \Omega_j A_j'$. Consistent standard errors therefore require consistent estimators for Γ_0 and Ω_j, $j = 1, \cdots J$.

An appropriate estimator for Γ_0 depends on the specific application. For differentiable moment conditions, an analog estimator may be based on the expression for the derivative.[13] Cattaneo (2010, Theorem 7) provides a general approach for smooth moment conditions that could be modified for incompletely observed moments. The estimator in Pakes and Pollard (1989, p. 1043) can be used for non-smooth cases. In what follows, it is assumed that any such consistent estimator $\widehat{\Gamma}_n$ is available.

Recall that the inversely weighted moment conditions have the variance that we are after:

$$\Omega_j = E\left[\frac{s_j}{p_j^2(X)}\psi(Z,\beta_0)\,\psi(Z,\beta_0)'\right] \tag{5.9}$$

$$= E\left[\frac{\Sigma_0(X)}{p_{j,0}(X)} + q_0(X)\,q_0'(X)\right]. \tag{5.10}$$

A natural estimator for Ω_j is therefore

$$\widehat{\Omega}_j = \frac{1}{n}\sum_{i=1}^{n} \frac{s_{ij}}{\widehat{p}_j(X_i)} d_j \psi(Z_i,\beta_n)\,\psi(Z_i,\beta_n)'\,d_j,$$

where β_n is a consistent estimator for β_0. The following result clarifies the conditions under which consistent standard errors can be based on $\widehat{\Omega}_j$.

Theorem 26. *If $\beta_n \xrightarrow{p} \beta_0$, $\left\|\widehat{\Gamma}_n - \Gamma_0\right\| = o_p(1)$, Assumptions 11, 12, and 6 hold, and if (i) ψ is continuous at β_0 almost surely; (ii) there exists a $\delta > 0$ such that $E\left[\sup_{\|\beta-\beta_0\|<\delta} \|\psi(Z,\beta)\|^2\right] < \infty$, then*

$$\left(\widehat{\Gamma}_n' A_n' \widehat{V}_{A,n}^{-1} A_n \widehat{\Gamma}_n\right)^{-1} \xrightarrow{p} \left(\Gamma_0' A' V_A^{-1} A \Gamma_0\right)^{-1},$$

[13] As an example, the linear IV example has $\Gamma_0 = -E\begin{bmatrix} W_1 X \\ W_2 X \end{bmatrix}$, which can be estimated consistenty by using the framework outlined in this paper or by using an IPW estimator from the stratum with complete data.

where

$$\widehat{V}_{A,n} = \sum_j A_{j,n} \widehat{\Omega}_j A_{j,n}',$$
$$A_n = \sum_j A_{j,n}.$$

In the next section, we investigate the finite sample performance of this estimator. An alternative estimator for Ω_j would use estimate Σ_0 and q_0 jointly from the J strata using the expression in (5.10).

6 Simulations

This section uses Monte Carlo simulations to investigate the finite sample behavior of the IPW estimator. In the first two subsections, I discuss the dynamic panel example (Example 3) and an extension. The final subsection considers the binary choice panel example (Example 4).

6.1 Dynamic panel data

Recall the dynamic panel model with $T = 5$ periods and two cohorts that each provide four time periods of data (Example 3). Identification fails in each cohort, but can be obtained by combining the information from both cohorts.

For this design, I will now use simulations to establish the efficiency loss from having only four periods of data for each observation. Also, I will look at the accuracy of the proposed methods for inference.[14]

Following Blundell, Bond, Windmeijer (2001), data are generated according to

$$Y_{it} = \alpha_i + \rho Y_{i,t-1} + X_{it}\beta_1 + X_{i,t-1}\beta_2 + \sigma_u u_{it},\ t = 1, \cdots, 5,$$
$$X_{it} = \tau\alpha_i + \gamma_0 + \gamma_1 X_{i,t-1} + \gamma_2 X_{i,t-2} + \sigma_v v_{it},\ t = 1, \cdots, 5,$$

where α_i, u_{it}, and v_{it} are standard normal, and $X_{i,-1} = X_{i0} = Y_{i0} = 0$. Table 1 contains the values for the other parameters. Each cohort contains 100 cross-section units. I

[14] To the best of my knowledge, no competing estimator is available for this case. Later on in this section, I will investigate the relative efficiency of the incomplete data estimator relative to alternatives, using a 6-period version of the DGP.

	ρ	$\{0.7, 0.8, 0.9\}$
	β_1, β_2	$0.5, 0.2$
	σ_u, σ_v	0.1
Cohort 1	τ	0.4
	γ_0	1
	γ_1, γ_2	$0.4, 0.4$
Cohort 2	Δ	$\{0.1, 0.3\}$
	τ	$0.4 + \Delta$
	γ_0	$1 + \Delta$
	γ_1, γ_2	$0.4 + \Delta, 0.4 - \Delta$

Table 1: Parameter values for the dynamic panel simulation study, $T = 5$. Top panel has parameter values that the cohorts have in common. Lower panels have cohort-specific parameters. Curly brackets indicate that the parameter values range over the values inside the curly brackets.

report results (based on 1000 simulations) for the infeasible estimator that uses five periods from each observation ("full", an infeasible estimator), as well as the efficient estimator based on the available data ("incomplete") from Section 5.1 (under MCAR).

Figure 6.1 and Table 2 present the results for the autoregressive parameter ρ. From Figure 6.1, we see that: (i) the efficient estimator using 4 periods of data for each observation is less efficient than the infeasible estimator that uses 5 periods of data; (ii) the variance of the estimators decreases as ρ approaches 1; (iii) both estimators are approximately unbiased.

More detail is provided in Table 2. The top panel corresponds to the boxplot in Figure 6.1, and the bottom panel corresponds to the case where the two cohorts are more similar. The latter design is less favorable for the incomplete data estimator: if the cohorts were identical ($\Delta = 0$) they would provide identical variation. This would lead to underidentification.

From Table 2, we conclude about the performance of the estimators that: (i) the bias of both estimators is negligible for all six combinations of parameter values; (ii) the standard error is 1.5 to 2.5 times as high, and drives a similar discrepancy in the RMSE. With regards to the quality of inference, we see that: (i) the proposed method for computing standard errors seems to work well, given how close columns 6 and 7 are; (ii) coverage probabilities are close to 0.95; (iv) the incomplete data estimator has better coverage than the full data estimator in the $\Delta = 0.1$ case.

Based on the results so far, I conclude that (i) an estimator with decent performance

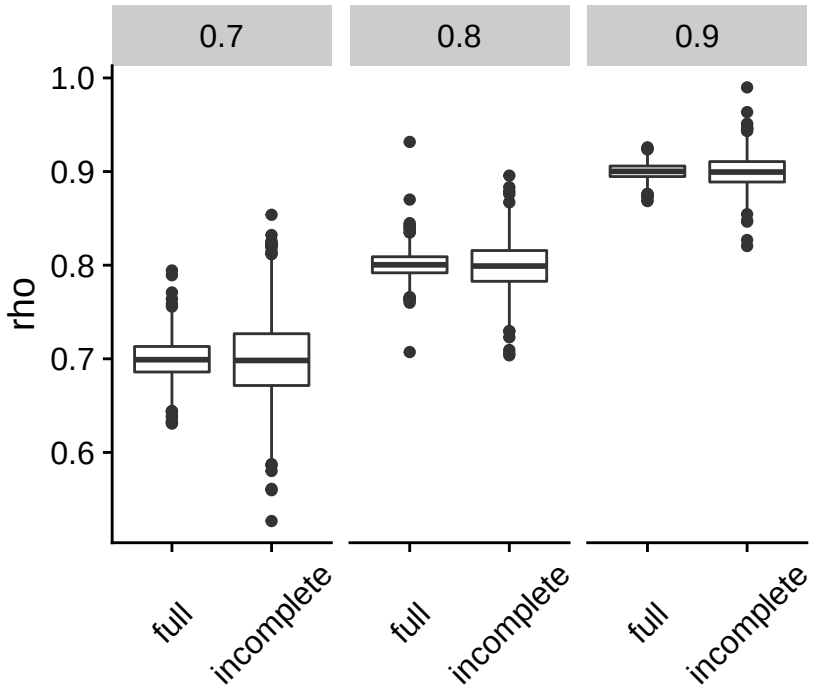

Figure 6.1: Simulation results for dynamic panel model for different values of the autoregressive parameter. The figure compares results from the infeasible "full" estimator to the procedure proposed in this paper ("incomplete").

Design			Performance			Inference		
			RMSE	Bias	SE	$\widehat{SE}$	Length	Coverage
$\Delta = 0.3$	$\rho = 0.7$	Full	0.021	0.000	0.021	0.022	0.086	0.946
		Incomplete	0.043	-0.001	0.043	0.042	0.164	0.942
	$\rho = 0.8$	Full	0.014	0.001	0.014	0.014	0.054	0.945
		Incomplete	0.026	0.000	0.026	0.026	0.104	0.955
	$\rho = 0.9$	Full	0.009	0.000	0.009	0.009	0.034	0.945
		Incomplete	0.018	0.000	0.018	0.018	0.072	0.958
$\Delta = 0.1$	$\rho = 0.7$	Full	0.020	-0.001	0.020	0.019	0.076	0.932
		Incomplete	0.051	0.001	0.051	0.055	0.215	0.951
	$\rho = 0.8$	Full	0.012	0.000	0.012	0.012	0.048	0.954
		Incomplete	0.029	0.000	0.029	0.031	0.120	0.945
	$\rho = 0.9$	Full	0.009	0.000	0.009	0.008	0.032	0.921
		Incomplete	0.019	0.000	0.019	0.020	0.079	0.951

Table 2: Simulation results for ρ in the dynamic panel model for different values of the autoregressive parameter: relative efficiency and inference. The performance measures are based on averages across 1000 simulations. For inference, $\widehat{SE}$ is the average across the standard errors computed for each simulation draw based on the procedure in Section 5.4; "length" is the average length of the confidence interval, and "coverage" is the simulated coverage probability.

can be constructed for a setting in which identification fails in each stratum; (ii) the proposed methods for inference work.

Finally, I establish the efficiency gains of the proposed methods over alternative estimators by considering a modification of the DGP in Section 6.1 for which alternative estimators are available. In particular, I consider a setting with $T = 6$ with 3 cohorts. One cohort provides 6 periods of data for each observation (cohort 0). The others provide five each (cohort 1 is available for periods 1 through 5; cohort 2 is available for periods 2 through 6).

In this case, a few estimators are available. For example, the "complete case" estimator, which uses only the observations from cohort 0. Also, the "available case" estimator, which imputes a "0" whenever a moment function is not available. Other estimators that I include are the "full" and "incomplete" estimator discussed above, and estimators based only on "cohort 1" (and "cohort 2").

Benchmark parameter values are as in 1, with the exception of the parameters governing the DGP of X. The changed values can be found in Table 3.

Table 4 presents the results. The results for the benchmark design are as expected.

	Δ	$\{0.05, 0.15\}$
Cohort 0	τ	0.1
	γ_0	1
	γ_1, γ_2	0.1
Cohort 1	τ	$0.1 + \Delta$
	γ_0	1
	γ_1, γ_2	$0.1 + \Delta, 0.1 + \Delta$
Cohort 2	τ	$0.1 + 2\Delta$
	γ_0	1
	γ_1, γ_2	$0.1 + 2\Delta, 0.1 + 2\Delta$

Table 3: Simulation study parameter values for the benchmark design. Top panel has parameter values that the cohorts have in common. Lower panels have cohort-specific parameters.

From the first panel (design 1), we can see that: (i) the full estimator outperforms the others - this is expected since it uses time periods for cohort 1 and 2 that are not available to the other estimators; (ii) the incomplete estimator outperforms all the other procedures; (iii) the available case estimator is outperformed by the incomplete case estimator - this is expected, because it uses the same data points as the incomplete data estimator, but not in an efficient way; (iv) the available case estimator outperforms the other estimators - which is expected, because they do not use all the available data points.

Moving to panel 2 (design 2), we increase the difference between the DGPs of the three cohorts by increase Δ to 0.15. The ranking of the estimators is unchanged. However, both the incomplete and available case estimators are now relatively more efficient. Moreover, the available case estimator is now as efficient as the incomplete one. Design 2 is such that the performance of the procedure in this paper can be obtained by simpler procedures.

Moving to design 3 (lower persistence), we see that (i) all estimators have an increased RMSE - this is expected from previous simulation studies for dynamic panels; (ii) the relative efficiency of the available case estimator is as in design 1; (iii) the incomplete case estimator regains almost 50% of the efficiency lost by the available case estimator due to the incompleteness of the data.

Finally, with 200 observations (design 4), the incomplete data estimator slightly improves its efficiency relative to the available case estimator, from 95% to 93%. This

Design	ρ	Δ	n	Estimator	RMSE	Relative	Bias	SE
1	0.9	0.05	100	full	0.416	1.000	-0.020	0.416
				incomplete	0.507	1.219	-0.025	0.507
				available case	0.534	1.284	-0.021	0.534
				complete case	0.726		-0.008	0.727
				cohort 1	1.238		-0.115	1.233
				cohort 2	1.432		0.044	1.432
2	0.9	0.15	100	full	0.420	1.000	-0.024	0.419
				incomplete	0.481	1.145	-0.024	0.480
				available case	0.480	1.143	-0.008	0.480
				complete case	0.769		-0.039	0.769
				cohort 1	1.209		0.015	1.209
				cohort 2	1.369		-0.089	1.367
3	0.8	0.05	100	full	0.685	1.000	-0.008	0.685
				incomplete	0.793	1.158	-0.035	0.793
				available case	0.873	1.274	-0.007	0.873
				complete case	1.176		-0.010	1.176
				cohort 1	2.009		0.064	2.009
				cohort 2	2.513		-0.194	2.507
4	0.9	0.05	200	full	0.281	1.000	-0.008	0.281
				incomplete	0.352	1.253	-0.016	0.352
				available case	0.376	1.339	-0.009	0.376
				complete case	0.523		-0.020	0.523
				cohort 1	0.826		-0.038	0.826
				cohort 2	0.900		0.001	0.901

Table 4: Simulation results for the dynamic panel model with $T = 6$ and three cohorts. The designs differ only with respect to ρ, Δ, n. Estimators are described in the main text. "Relative" is the RMSE of an estimator divided by the RMSE of the "full" estimator.

is expected, since there are now more observations available for optimal weight matrix estimation.

6.2 Fixed effects binary choice

This section reports on a simulation study for the three-period fixed effects binary choice logit model in Example 4 (Section 3). We focus on the consistency of the inverse propensity score weighting estimator, and on its efficiency gains.

The results below are based on $S = 10000$ simulations. We simulate data for $n = 1000$ cross-section units. Denote by $\widetilde{D}_{i,t}$ the binary variable that indicates whether

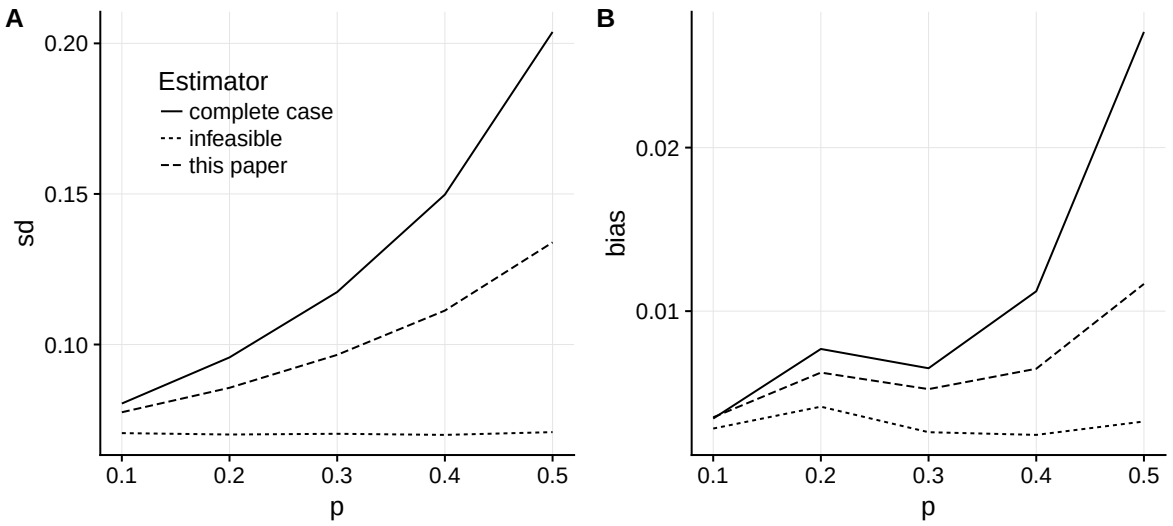

Figure 6.2: Simulation results for a panel data binary choice model, no selection. Panel (a) shows the standard deviation, panel (b) displays the bias. Results are shown as a function of the probability of missingness, for $n = 1000$ cross-section units, based on $S = 10000$ simulations.

measurements are available for unit i in period t. We will assume that the distribution of $\widetilde{D}_{i,t}$ is independent of t and that MCAR holds, i.e. for each t, $(Y_i, X_i) \perp \widetilde{D}_{i,t}$. We are interested in the impact of $p \equiv P\left(\widetilde{D}_{i,t} = 1\right)$ on the relative performance of the infeasible estimator, the complete-case estimator, and the procedure proposed in this paper. To that end, we will vary p while keeping other features of the design constant. In particular, we will generate X_{it} to be i.i.d. standard normal across (i, t), set $\alpha_i = \frac{1}{2}(X_{i1} + X_{i2})$, and set $\beta = 1$.

The results are displayed in Figure 6.2. First, note from the scale of the two figures that the bias is negligible for all estimators, although the complete case estimator is slightly worse than the other two. Second, note that the performance of the estimators are as expected: the infeasible estimator beats the estimator proposed in this paper, which beats the complete case estimator. The amount by which the new proposal beats the complete case estimator depends on the probability of missingness. However, the margin is substantial even for moderate values of p.

We now move to a setup where the incompleteness is at random, as opposed to completely at random. To set up the selection equation, express the binary dependent variable in latent variable form,

$$Y_{it} = 1\left\{\alpha_i + X_{it}\beta + U_{it} \geq 0\right\},$$

which implies the formulation in example 4 under the assumption that the U_{it} are serially independent, standard logistic random variables that are independent of X_i. Assume that missingness depends on

$$Z_{i,t} = 1\left\{U_{it} \geq 0\right\}$$

in a time-invariant fashion, so that the relevant propensity score is

$$P\left(\widetilde{D}_{i,t} = 1 \middle| Z_{i,t} = z\right) = \begin{cases} p_1 & \text{when } z = 1, \\ p_0 & \text{when } z = 0. \end{cases} \tag{6.1}$$

Unless $p_0 = p_1$, estimators that do not correct for selection will be inconsistent.

We adapt the simulation design by replacing the MCAR mechanism above with the MAR mechanism (6.1). We fix $p_1 = 1$, and vary p_0. All other parameters in the simulation design are unchanged.

We consider four estimators. The "complete case" and "infeasible" estimators are as above. The "complete case" estimator does not correct for selection, and the "infeasible" estimator does not need to. The estimator labelled "this paper, MCAR" in Figure 6.3 is the estimator proposed in this paper, but without the correction for selection. The estimator "this paper" is the inverse propensity score weighted estimator proposed in this paper.

The results are displayed in Figure 6.3. The non-weighted estimators are severely biased. That bias deteriorates when p_0 increases, and will disappear when $p_0 = p_1 = 1$, which corresponds to MCAR. The IPW estimator performs well. Its bias is barely above that of the infeasible estimator, and slightly elevated at $p_0 = 0.1$. Its standard deviation is higher than that of the uncorrected estimator, which is severely biased.

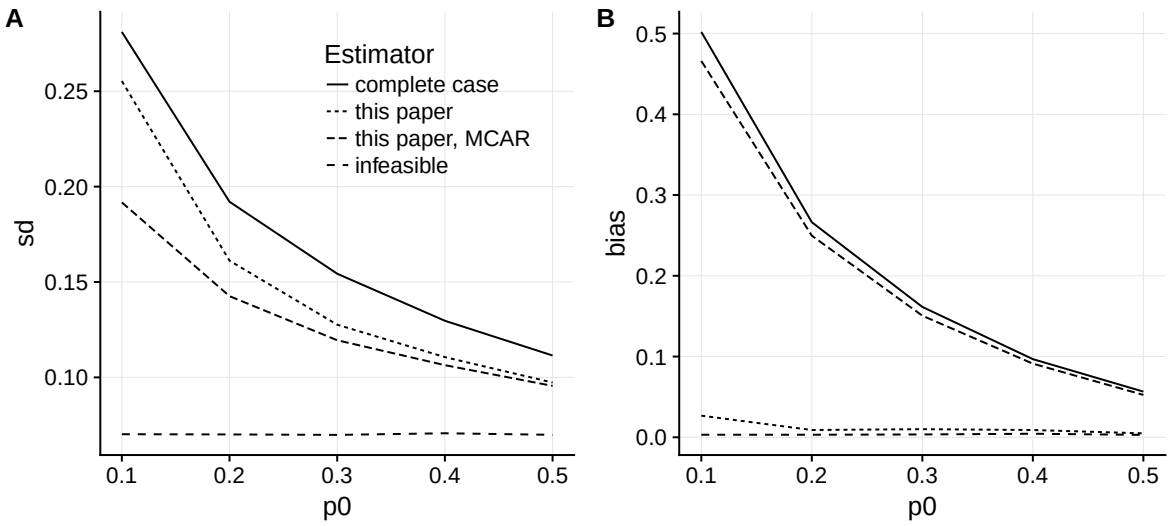

Figure 6.3: Simulation results for a panel data binary choice model with selection. Panel (a) shows the standard deviation, panel (b) displays the bias. Results are shown as a function of the probability of missingness for those with $Z_{it} = 0$. Results are for $n = 1000$ cross-section units, based on $S = 10000$ simulations.

7 Empirical illustration

To demonstrate the potential efficiency gains of the estimators proposed in this paper, this section revisits the analysis in Topalova and Khandelwal (2011, henceforth TK). TK investigate a trade reform in India using firm-level unbalanced panel data. Please see TK for a more detailed data description and substantive background.

TK estimate, among others, parameters in static and dynamic panel models that can be interpreted as the causal effect of output tariffs on total factor productivity. The static panel models considered are (cf. TK's equation (2), with results in Table 4, columns 3 and 4):

$$pr_{ijt} = \alpha_i + \beta trade_{j,t-1} + X_{ijt}\gamma + u_{ijt}, \tag{7.1}$$

where i indicates one of 3108 firms, j indicates a four-digit NIC industry, and t indicates a time period from 1990 to 1996. The dependent variable pr is a productivity measure constructed from production function estimates (see TK, p. 998); $trade$ is output tariff measured at the industry level. In our analysis, X_{ijt} consists only of time dummies.[15]

TK estimate this model using the panel fixed effects estimator. This section instead

[15] I omit the quadratic *age* terms used in TK. This has a minor effect on the results: almost all of the firm age effect is captured by the combination of firm and time fixed effects.

considers a first difference (FD) estimator, because: (i) the estimator for the dynamic model is also an FD estimator; (ii) the FD results are very close to the fixed effects results. The moment conditions for the FD estimator are

$$E\begin{bmatrix} \Delta trade_{j,89}\Delta u_{ij,90} \\ \Delta u_{ij,90} \\ \vdots \\ \Delta trade_{j,95}\Delta u_{ij,96} \\ \Delta u_{ij,96} \end{bmatrix} = 0,$$

where $\Delta u_{ijt} = u_{ijt} - u_{ij,t-1}$, etc. The moment conditions involving $\Delta trade$ follow from the exogeneity of tariff changes to firm level decisions; the other moment conditions define the time dummies.

The panel data set is very unbalanced. Less than half of the firms are observed over the entire period, and there are a total of 46 patterns. Appendix D contains two figures that describe data availability. The incomplete data pattern for a completely observed firm is $D_1 = I_{14}$. A firm that drops out in 1991 has $D_2 = e_{1,2} \otimes I_7$, etc. In what follows, I discard patterns with fewer than 20 firms.

Table 5 contains the results. Additional findings are in Appendix D. The "TK" columns contain the results from TK, the "CM" column contains my replication. The replicated results are slightly different, because: (i) I use optimally weighted FD estimation instead of a fixed effects estimator; (ii) I use bootstrap standard errors rather than robust standard errors; (iii) I did not include *age* and age^2 as control variables.

The complete case estimator uses only firms for which all measurements are available in all time periods. The available case estimator replaces missings by zeros. The "incomplete" estimator is the asymptotically efficient procedure proposed in this paper.

The main finding for the static model is that the standard errors are lowest for "incomplete".[16] From the complementary material (Table 8 in Appendix D) we can see that the relative efficiency varies with the parameter of interest, but that the incomplete case estimator dominates the complete case and available case estimator.

The dynamic model (TK, column 6) adds an autoregressive term on the righthand-

[16]There are at least two reasons why the asymptotically efficient incomplete data estimator could have had a higher standard error: (i) the asymptotic efficiency gains of the optimal procedure are partially offset by the estimation of many additional weight matrix elements for "incomplete"; (ii) the available case estimator uses the incomplete data patterns with < 20 firms, which are discarded in the implementation of "incomplete".

		Complete case		Available case		
		TK (3)	CM	TK (4)	CM	Incomplete
β	Estimate	-0.053	-0.035	-0.059	-0.031	-0.043
	SE	(0.016)	(0.013)	(0.017)	(0.011)	(0.009)
	n	14808	14808	8059	8059	

Table 5: Results for the static panel model.

		TK (6)	CC	Incomplete
β	Estimate	-0.048	-0.041	-0.037
	SE	(0.013)	(0.016)	(0.013)
ρ	Estimate	0.455	0.472	0.228
	SE	(0.068)	(0.057)	(0.032)

Table 6: Results for the dynamic panel model.

side,

$$pr_{ijt} = \alpha_i + \rho pr_{ij,t-1} + \beta trade_{j,t-1} + X_{ijt}\gamma + v_{ijt}. \tag{7.2}$$

TK estimate the parameters in this model using the procedure in Arellano and Bond (1991), which is a GMM estimator based on the moment conditions

$$E\begin{bmatrix} \Delta trade_{j,t-1}\Delta v_{ijt} \\ \Delta v_{ijt} \\ pr_{ijt-s}\Delta v_{ijt} \end{bmatrix} = 0, \text{ for } t = 90, \cdots, 96, \; s = t-2, t-3, \cdots, 90.$$

It is not clear from the paper or code how the incomplete observations are handled, but my replication suggests that the authors are using a complete case approach. Table 6 compares the result in TK with a complete case replication, as well as with "incomplete". First, note that the replication is close but not exact, for reasons mentioned in the discussion of the static case. Second, note that the procedure proposed in this paper provides a substantial improvement in terms of efficient over the complete case estimator for both parameters.

8 Conclusion

In this paper, I propose a moment condition framework for estimation using incomplete data. I derive moment conditions that are valid for the incomplete data, and that generalize those for the standard MAR setup. The efficiency bound associated with those

moment conditions can be obtained using the efficient estimators proposed. Notably, identification can be achieved even if it fails in each stratum of incompleteness. Efficient IPW and DR estimator are proposed, and their large sample properties are established. A simulation study and empirical illustration demonstrate that efficiency gains over standard approaches can be substantial.

References

Abrevaya, J. and S. Donald (2011), "A GMM Approach for Dealing with Missing Data on Regressors and Instruments", Mimeo.

Abrevaya, J. and S. Donald (2017), "A GMM Approach for Dealing with Missing Data on Regressors", The Review of Economics and Statistics 99 (4), 657-662.

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

Acemoglu, D., S. Naidu, P. Restrepo, and J.A. Robinson (2015), "Democracy, Redistribution and Inequality", in A.B. Atkinson and F. Bourguignon, *Handbook of Income Distribution, Volume 2*, pp. 1885-1966. Elsevier.

Acemoglu, D., S. Naidu, P. Restrepo, and J.A. Robinson (2018), "Democracy Does Cause Growth", Journal of Political Economy, forthcoming.

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, 277-297.

Becker S.O. and L. Woessmann (2013), "Not the Opium of the People: Income and Secularization in a Panel of Prussian Counties", American Economic Review 103 (3), 539-544.

Bickel, P.J., C.A.J. Klaasen, Y. Ritov, and J.A. Wellner (1993). Efficient and Adaptive Estimation for Semiparametric Models. Springer: New York.

Blundell, R., S. Bond, and F. Windmeijer (2001), "Estimation in Dynamic Panel Data Models: Improving on the Performance of the Standard GMM Estimator", In B.H. Baltagi, T.B. Fomby, and R. Carter Hill, *Nonstationary Panels, Panel Cointegration, and Dynamic Panels,* pp. 53-91. Emerald Group Publishing Limited.

Botosaru, I. and F. Gutierrez (2018), "Difference-in-Differences When the Treatment Status is Observed in Only One Period", Journal of Applied Econometrics 33 (1), 73-90.

Cameron, A.C. and P.K. Trivedi (2005). Microeconometrics: Methods and Applications. Cambridge: Cambridge University Press.

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling", in L. Christofides, E.K. Grant and R. Swindinsky, *Aspects of Labour Economics: Essays in Honour of John Vanderkamp*, University of Toronto Press.

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

Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data", The Review of Economic Studies 47 (1), 225-238.

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

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

Chamberlain, G. (1992b), "Comment: Sequential Moment Restrictions in Panel Data", Journal of Business and Economic Statistics 10 (1), 20-26.

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.

Chen, X., O. Linton, and I. van Keilegom (2003), "Estimation of Semiparametric Models when the Criterion Function is not Smooth", Econometrica 71 (5), 1591-1608.

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.

de Loecker, J. and F. Warzynski (2012), "Markups and Firm-level Export Status", American Economic Review 102 (6), 2437-2471.

Feng, Q. (2016), "Instrumental Variables Estimation with Missing Instruments", Mimeo.

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.

Graham, B.S., C. Campos de Xavier Pinto, and D. Egel (2012), "Inverse Probability Tilting for Moment Condition Models with Missing Data", Review of Economic Studies 79 (3), 1053-1079.

Graham, B.S., C. Campos de Xavier Pinto, and D. Egel (2016), "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)", Journal of Business and Economic Statistics 34 (2), 288-301.

Heckman, J., H. Ichimura, and P. Todd (1997), "Matching As An Econometric Evaluation Estimator", Review of Economic Studies 64 (4), 605–654.

Hirano, K., G. Imbens, and G. Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", Econometrica 71 (4), 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.

Hristache, M. and V. Patilea (2016), "Semiparametric Efficiency Bounds For Conditional Moment Restriction Models With Different Conditioning Variables", Econometric Theory 32 (4), 917-946.

Hristache, M. and V. Patilea (2017), "Conditional Moment Models with Data Missing at Random", Biometrika 104 (3), 735-742.

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. (1994), "The Asymptotic Variance of Semiparametric Estimators", Econometrica 62 (6), 1349-1382.

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.

Pakes, A. and D. Pollard (1989), "Simulation and the Asymptotics of Optimization Estimators", Econometrica 57 (5), 1027-1057.

Papke, L.E. (2005), "The Effects of Spending on Test Pass Rates: Evidence from Michigan", Journal of Public Economics 89 (5-6), pp. 821-839

Papke, L. E. and J.M. Wooldridge (2008), "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates", Journal of Econometrics, 145 (1-2), 121-133.

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.

Rodrik, D., A. Subramanian and F. Trebbi (2004), "Institutions Rule: the Primacy of Institutions over Geography and Integration in Economic Development." Journal of Economic Growth 9 (2), 131-165.

Rothe, C. and S. Firpo (2017), "Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically", Mimeo.

Schularick, M. and T.M. Steger (2010), "Financial Integration, Investment, and Economic Growth: Evidence from Two Eras of Financial Globalization", The Review of Economics and Statistics 92 (4), 756-768.

Sturm, J., and J. De Haan (2015), "Income Inequality, Capitalism, and Ethno-linguistic Fractionalization", American Economic Review 105 (5), 593-597.

Tan, Z. (2010), "Bounded, Efficient, and Doubly Robust Estimation with Inverse Weighting", Biometrika 97 (3), 661-682.

Topalova, P. and A. Khandelwal (2011), "Trade Liberalization and Firm Productivity: the Case of India", Review of Economics and Statistics 93 (3), 995-1009.

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

van der Vaart, A. and J.A. Wellner (1996). Weak Convergence and Empirical Processes. Springer: New York.

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

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

Yagan, D. (2015), "Capital Tax Reform and the Real Economy: The Effects of the 2003 Dividend Tax Cut", American Economic Review 105 (12), 3531-63.

A Proofs

A.1 Proof of Theorem 8

Proof of Theorem 8. This Theorem generalizes Graham's (2011) Theorem 2.1, who considers the case with $J = 1$ and $d_1 = I$. The proof here mimicks his proof. The main

difference is that the matrix algebra in step 3 of the current proof is more involved than step 3 in Graham's (2011) Theorem 2.1, owing to the presence of multiple strata, multinomial random variables in the propensity score calculations, and generalized inverses to deal with strata that do not individually identify the parameter.

The proof here starts by introducing some additional notation. Step 1 establishes the equivalence between the moment conditions ((4.1)+ 4.2) and a set of unconditional moment restrictions. Step 2 applies Lemma 2 in Chamberlain (1987) to obtain the information bound for those unconditional moment restrictions. Step 3 consists of matrix algebra to obtain the expression for the bound in the main text.

Notation. In the remainder of this proof, "Assumption (iii)" refers to the third assumption in the statement of the result in the main text.

The random object D is an incomplete data indicator that takes one of J values $(d_1, \cdots, d_J)$, and signals which components of ψ are observed. Here, we will work with a modified indicator that omits the zero rows: call the resulting random object $\tilde{D}$, with support $\left\{\tilde{d}_1, \cdots, \tilde{d}_J\right\}$. Each $\tilde{d}_j$ is a rectangular selection matrix of size $r_j \times p$, where r_j is the number of observable components of ψ in stratum j. For example, if $p = 2$, $\tilde{D}$ 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 stratum without any moment function observed does not contribute to the bound.

The set of moment conditions for which we wish to establish the efficiency bound is given by (4.1) and 4.2. We have assumed that Z follows a multinomial distribution ((i) in the statement of the result). 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

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

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

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

Step 1. The moment conditions in (4.1) and 4.2 are equivalent to

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

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

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

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

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

where $V_{12} = E\left[m_1\left(\rho_0\right) m_2\left(\beta_0, \rho_0\right)^{'}\right]$, etcetera. Because of assumption (iv), this matrix

is invertible. 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 well-defined because of assumption (iii). We will denote it as the $(JL + \bar{r}) \times (JL + K)$ matrix

$$M = E\left[\left. \frac{\partial m(\rho, \beta)}{\partial (\rho, \beta)'} \right|_{\beta=\beta_0, \rho=\rho_0} \right] \tag{A.5}$$

$$= \begin{bmatrix} M_{1\rho} & O_{JL\times k} \\ M_{2\rho} & M_{2\beta} \end{bmatrix} \left(\text{dimensions } \begin{bmatrix} JL \times JL & JL \times K \\ \bar{r} \times JL & \bar{r} \times K \end{bmatrix} \right) \tag{A.6}$$

where $M_{2\beta} = E\left[\frac{\partial m_2(\rho,\beta)}{\partial \beta'}\right]$, etc. Note that $M_{2\beta}$ will feature Γ_0 from (4.3).

Because of the assumptions in the statement of the result, all the conditions for Lemma 2 in Chamberlain (1987) are satisfied. To translate my notation to his, set his $\Theta = \mathcal{B} \times (0,1)^{JL}$, which contains the true parameter $\theta_0 = (\beta_0, \rho_0)$; set $\psi(z, \theta) = m(\rho, \beta)$. Chamberlain's condition C1(i) holds because of assumptions (ii) and (iii); C1(ii) holds because of assumption (ii); that C1(iii) holds is implied by assumption (iv); and C1(iv) is assumed in (iii). Finally, we have assumed that Z is multinomial. Then all the conditions for Chamerlain (1987, Lemma 1 and Lemma 2) are satisfied, and it follows that the lower bound on the variance of β_0 is given by

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

The remainder of this proof establishes that the expression (A.7) is equal to that in the main text, (4.5).

The derivations for $I_0(\beta)$ 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

$$I_0 \left(\beta_0 \right) = M_{2\beta}^{'} V^{22} M_{2\beta}. \tag{A.9}$$

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 3c establishes that this is the case. In what follows, we will assume the results of Step 3c, and will proceed to evaluate the expression in (A.9).

Step 3a. 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

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

The derivative is the $r_j \times K$ matrix

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

so that $M_{2j,\beta} = \tilde{d}_j \Gamma_0$: we can interchange expectation and derivative because of the multinomial assumption (everything is bounded). 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 $\bar{r} \times Jp$, and the matrix

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

is $\bar{r} \times p$. Note that $\tilde{\Delta}_2 = \tilde{\Delta}_1 (\iota_J \otimes I_p)$, 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 $p = 3$, and

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

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

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

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.

$$\begin{aligned}
E\left[m_{2j} m_{2j}'\right] &= E\left[\frac{s_j}{p_{j,0}^2(X)} \tilde{d}_j \psi(Z,\beta)\,\psi(Z,\beta)'\,\tilde{d}_j'\right] \\
&= \tilde{d}_j E\left[E\left[\frac{s_j}{p_{j,0}^2(X)}\middle| X\right] E\left[\psi(Z,\beta)\,\psi(Z,\beta)'\middle| X\right]\right]\tilde{d}_j' \\
&= \tilde{d}_j E\left[\frac{1}{p_{j,0}(X)} E\left[\psi(Z,\beta)\,\psi(Z,\beta)'\middle| X\right]\right]\tilde{d}_j' \\
&= \tilde{d}_j E\left[\frac{1}{p_{j,0}(X)} E\left[\Sigma_0(X) + q_0(X)\, q_0(X)'\right]\right]\tilde{d}_j',
\end{aligned}$$

where $q_0(X)$ and $\Sigma_0(X)$ are defined in the statement of the result. To rewrite this using the discrete support of $X \in \{x_1, \cdots, x_L\}$, let

$$\begin{aligned}
\tau_l &= P(X = x_l), \\
q_l &= q_0(x_l), \\
\Sigma_l &= \Sigma_0(x_l).
\end{aligned}$$

Then the j-th block can be written as

$$E\left[m_{2j} m_{2j}'\right] = \tilde{d}_j \left[\sum_{l=1}^{L} \frac{\tau_l}{\rho_{jl,0}} \left(\Sigma_l + q_l q_l'\right)\right] \tilde{d}_j'. \tag{A.10}$$

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

$$
\begin{aligned}
V_{22} &= \begin{bmatrix} \tilde{d}_1 & 0 & 0 \\ 0 & \ddots & \\ 0 & & \tilde{d}_J \end{bmatrix} \left[\sum_{l=1}^{L} \tau_l \begin{bmatrix} \frac{1}{\rho_{1l,0}} & 0 & 0 \\ 0 & \ddots & \\ 0 & & \frac{1}{\rho_{Jl,0}} \end{bmatrix} \otimes \left(\Sigma_l + q_l q_l^{'} \right) \right] \begin{bmatrix} \tilde{d}_1 & 0 & 0 \\ 0 & \ddots & \\ 0 & & \tilde{d}_J \end{bmatrix}^{'} \\
&= \tilde{\Delta}_1 \left(\sum_{l=1}^{L} \tau_l R_{l,0}^{-1} \otimes \left(\Sigma_l + q_l q_l^{'} \right) \right) \tilde{\Delta}_1^{'}. \qquad \text{(A.11)}
\end{aligned}
$$

(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

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

In $\sum_l$-format, it obtains that

$$
\begin{aligned}
V_{12,jk} &= -\begin{bmatrix} \tau_1 q_1^{'} \\ \vdots \\ \tau_L q_L^{'} \end{bmatrix} \tilde{d}_k \\
&= -\left(\sum_{l=1}^{L} \tau_l \otimes e_l q_l^{'} \right) \tilde{d}_k. \qquad \text{(A.13)}
\end{aligned}
$$

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

$$\begin{aligned} V_{12,jj} &= E\left[m_{1j} m_{2j}^{'}\right] \\ &= E\left[B\left(\frac{s_j}{p_{j,0}(X)} - 1\right)\frac{s_j}{p_{j,0}(X)}\psi(Z,\beta_0)^{'}\tilde{d}_j^{'}\right] \\ &= E\left[B\left(\frac{s_j}{p_{j,0}^2(X)} - \frac{s_j}{p_{j,0}(X)}\right)\psi(Z,\beta_0)^{'}\right]\tilde{d}_j^{'} \\ &= E\left[BE\left[\left(\frac{s_j}{p_{j,0}^2(X)} - \frac{s_j}{p_{j,0}(X)}\right)\middle| X\right]E\left[\psi(Z,\beta_0)^{'}\middle| X\right]\right]\tilde{d}_j^{'} \\ &= E\left[B\left(\frac{1}{p_{j,0}(X)} - 1\right)q(X)^{'}\right]\tilde{d}_j^{'}. \end{aligned} \tag{A.14}$$

Remember that B is an $L \times 1$ vector and that $q(X)$ is an $p \times 1$ vector. Continue the derivation above by expressing (A.14) as a Kronecker product:

$$\begin{aligned} E\left[B\left(\frac{1}{p_{j,0}(X)} - 1\right)q(X)^{'}\right]\tilde{d}_j^{'} &= \begin{bmatrix} \tau_1\left(\frac{1}{\rho_{j1,0}} - 1\right)q_1^{'} \\ \vdots \\ \tau_L\left(\frac{1}{\rho_{jL,0}} - 1\right)q_L^{'} \end{bmatrix}\tilde{d}_j^{'} \\ &= \left(\sum_{l=1}^{L}\tau_l\left(\frac{1}{\rho_{jl,0}} - 1\right)\otimes e_l q_l^{'}\right)\tilde{d}_j^{'} \\ &= \left(\sum_{l=1}^{L}\frac{\tau_l}{\rho_{jl,0}}\otimes e_l q_l^{'}\right)\tilde{d}_j^{'} - \left(\sum_{l=1}^{L}\tau_l \otimes e_l q_l^{'}\right)\tilde{d}_j^{'} \end{aligned} \tag{A.15}$$

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

$$\begin{aligned} V_{12} &= \left(\sum_{l=1}^{L} \tau_l R_{l,0}^{-1} \otimes e_l q_l^{'}\right) \tilde{\Delta}_1^{'} - \left(\sum_{l=1}^{L} \tau_l \iota_J \iota_J^{'} \otimes e_l q_l^{'}\right) \tilde{\Delta}_1^{'} \\ &= \left(\sum_{l=1}^{L} \tau_l \left(R_{l,0}^{-1} - \iota_J \iota_J^{'}\right) \otimes e_l q_l^{'}\right) \tilde{\Delta}_1^{'}. \end{aligned} \tag{A.16}$$

(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

$$\begin{aligned} V_{11} &= E\left(m_1(\rho_0)\, m_1(\rho_0)^{'}\right), \\ &= E\left[(F \otimes B)(F \otimes B)^{'}\right], \\ &= E\left[(F \otimes B)\left(F^{'} \otimes B^{'}\right)\right], \\ &= E\left[\left(FF^{'}\right) \otimes \left(BB^{'}\right)\right], \\ &= E\left[E\left[\left(FF^{'}\right) \otimes \left(BB^{'}\right) \middle| X\right]\right], \\ &= \sum_{l=1}^{L} \tau_l E\left[FF^{'} \middle| X = x_l\right] \otimes e_l e_l^{'}, \end{aligned} \tag{A.17}$$

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 4.1 implies that

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

so that (A.17) becomes

$$V_{11} = \sum_{l=1}^{L} \tau_l V(F|X = x_l) \otimes e_l e_l^{'}. \tag{A.18}$$

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:

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

so that

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

where

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

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,

$$V\left(D|X = x_l\right) = R_{l,0} - R_{l,0}\iota_J \iota_J^{'} R_{l,0}, \tag{A.19}$$

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.18), note that for any $l \neq k$, the product $\left(e_l e_l^{'}\right)\left(e_k e_k^{'}\right) = 0$. Therefore:

$$\begin{aligned} V_{11}^{-1} &= \left(\sum_{l=1}^{L} \tau_l V\left(F|X = x_l\right) \otimes e_l e_l^{'} \right)^{-1} \\ &= \sum_{l=1}^{L} \frac{1}{\tau_l} \left(R_{l,0}^{-1} - \iota_J \iota_J^{'} \right)^{-1} \otimes e_l e_l^{'}. \end{aligned} \tag{A.20}$$

(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.11, A.16, and (A.20). Remember that $(A \otimes B)^{'} = \left(A^{'} \otimes B^{'}\right)$ and

$(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$, so that

$$\begin{aligned} V_{12}^{'} V_{11}^{-1} &= \tilde{\Delta}_1 \sum_{l_1=1}^{L} \sum_{l_2=1}^{L} \left(\tau_{l_1} \left(R_{l_1,0}^{-1} - \iota_J \iota_J^{'} \right) \otimes e_{l_1} q_{l_1}^{'} \right)^{'} \left(\frac{1}{\tau_{l_2}} \left(R_{l_2,0}^{-1} - \iota_J \iota_J^{'} \right)^{-1} \otimes e_{l_2} e_{l_2}^{'} \right) \\ &= \tilde{\Delta}_1 \sum_{l_1=1}^{L} \sum_{l_2=1}^{L} \frac{\tau_{l_1}}{\tau_{l_2}} \left(\left(R_{l_1,0}^{-1} - \iota_J \iota_J^{'} \right)^{'} \otimes q_{l_1} e_{l_1}^{'} \right) \left(\left(R_{l_2,0}^{-1} - \iota_J \iota_J^{'} \right)^{-1} \otimes e_{l_2} e_{l_2}^{'} \right) \\ &= \tilde{\Delta}_1 \sum_{l_1=1}^{L} \sum_{l_2=1}^{L} \frac{\tau_{l_1}}{\tau_{l_2}} \left(\left(R_{l_1,0}^{-1} - \iota_J \iota_J^{'} \right) \left(R_{l_2,0}^{-1} - \iota_J \iota_J^{'} \right)^{-1} \otimes q_{l_1} e_{l_1}^{'} e_{l_2} e_{l_2}^{'} \right) \end{aligned} \tag{A.21}$$

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.21) reduces to

$$V_{12}^{'} V_{11}^{-1} = \tilde{\Delta}_1 \sum_{l=1}^{L} \left(I_J \otimes q_l e_l^{'} \right). \tag{A.22}$$

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.22). Finally,

$$\begin{aligned} V_{22} - V_{12}^{'} V_{11}^{-1} V_{12} &= \tilde{\Delta}_1 \left(\sum_{l=1}^{L} \tau_l R_{l,0}^{-1} \otimes \left(\Sigma_l + q_l q_l^{'} \right) \right) \tilde{\Delta}_1^{'} - \tilde{\Delta}_1 \sum_{l=1}^{L} \tau_l \left(\left(R_{l,0}^{-1} - \iota_J \iota_J^{'} \right) \otimes q_l q_l^{'} \right) \tilde{\Delta}_1^{'} \\ &= \tilde{\Delta}_1 \left(\sum_{l=1}^{L} \tau_l \left(R_{l,0}^{-1} \otimes \left(\Sigma_l + q_l q_l^{'} \right) - \left(R_{l,0}^{-1} - \iota_J \iota_J^{'} \right) \otimes q_l q_l^{'} \right) \right) \tilde{\Delta}_1^{'} \\ &= \tilde{\Delta}_1 \left(\sum_{l=1}^{L} \tau_l \left(R_{l,0}^{-1} \otimes \Sigma_l + \iota_J \iota_J^{'} \otimes q_l q_l^{'} \right) \right) \tilde{\Delta}_1^{'} \\ &\equiv \tilde{\Delta}_1 \Lambda_0 \tilde{\Delta}_1^{'}. \end{aligned} \tag{A.23}$$

where

$$\begin{aligned} \Lambda_0 &= \sum_{l=1}^{L} \tau_l \left(R_{l,0}^{-1} \otimes \Sigma_l + \iota_J \iota_J^{'} \otimes q_l q_l^{'} \right) \\ &= E \left(R_0^{-1}(X) \otimes \Sigma_0(X) + \iota_J \iota_J^{'} \otimes q_0(X) q_0(X)^{'} \right) \\ R_0(X) &= \operatorname{diag}\left(p_{1,0}(X), \cdots, p_{J,0}(X) \right). \end{aligned}$$

In terms of the notation of the result,

$$V_{22} - V_{12}^{'} V_{11}^{-1} V_{12} = \begin{bmatrix} \widetilde{d}_1 \Omega_1 \widetilde{d}_1^{'} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \widetilde{d}_J \Omega_J \widetilde{d}_J^{'} \end{bmatrix}, \tag{A.24}$$

Assuming invertibility of Λ_0 (see Step 3c),

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

so that it can be seen from (A.24) that

$$V^{22} = \begin{bmatrix} \left(\widetilde{d}_1 \Omega_1 \widetilde{d}_1^{'} \right)^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \left(\widetilde{d}_J \Omega_J \widetilde{d}_J^{'} \right)^{-1} \end{bmatrix}. \tag{A.25}$$

with Σ_j defined in (4.6).

Step 3b. Finally, we assemble the pieces of step 3a into the desired expression of the bound $M_{2\beta}^{'} V^{22} M_{2\beta}$. Straightforward matrix multiplication for V^{22} in (A.25) and $\tilde{\Delta}_2 \Gamma_0$,

$$\begin{aligned} I(\beta_0) &= M_{2\beta}^{'} V^{22} M_{2\beta} \\ &= \Gamma_0^{'} \tilde{\Delta}_2^{'} \begin{bmatrix} \left(\widetilde{d}_1 \Omega_1 \widetilde{d}_1^{'}\right)^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \left(\widetilde{d}_J \Omega_J \widetilde{d}_J^{'}\right)^{-1} \end{bmatrix} \tilde{\Delta}_2 \Gamma_0 \\ &= \Gamma_0^{'} \left(\sum_j \widetilde{d}_j^{'} \left(\widetilde{d}_j \Omega_j \widetilde{d}_j^{'}\right)^{-1} \widetilde{d}_j \right) \Gamma_0 \\ &= \Gamma_0^{'} \left(\sum_j \left(d_j \Omega_j d_j\right)^{+} \right) \Gamma_0, \end{aligned}$$

where the last step follows because the construction of d_j and $\widetilde{d}_j$ implies

$$\left(d_j \Omega_j d_j\right)^{+} = \tilde{d}_j^{'} \left(\tilde{d}_j \Omega_j \tilde{d}_j^{'}\right)^{-1} \tilde{d}_j.$$

Step 3c. The validity of the bound calculations in Steps 3a,b require that M has full rank, that V is invertible, and that $M' V^{-1} M$ is blockdiagonal. In this final section, we show that these requirements are satisfied.

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.28) 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.23)). Then, since V_{11} is invertible (see (A.20)) and the second term is invertible when Λ_0 is, V is invertible under Assumptions 1-3 and the conditions of Theorem 8.

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

$$V_{12}^{'} V_{11}^{-1} M_{1\rho} = M_{2\rho}. \tag{A.26}$$

From (A.22), we know that

$$V_{12}^{'} V_{11}^{-1} = \tilde{\Delta}_1 \sum_{l=1}^{L} \left(I_J \otimes q_l e_l^{'} \right). \tag{A.27}$$

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.22).

First, note that

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

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

$$M_{1\rho} = -\sum_{l=1}^{L} \tau_l R_{l,0}^{-1} \otimes e_l e_l^{'}. \tag{A.28}$$

Using a similar expression as for the inverse of V_{11} in (A.20), 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.21),

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

□

A.2 Proof of Theorem 14

Proof of Theorem 14. This proof follows the structure in Cattaneo (2010, Theorem 2) and verifies the conditions of Pakes and Pollard (1989, henceforth PP89), Corollary 3.2. To link their notation to the present case, set $\theta = \beta$ and $G(\theta) = \sum_j A_j E\left[\frac{1\{D=d_j\}}{p_{j,0}(X)} d_j \psi(Z, \beta)\right]$. The definition of the IPW estimator in equation (5.2) implies their condition (i). We check conditions (ii, identification) and (iii, uniform convergence) below.

Identification. Informally, Assumption 1 assumes that the original moment conditions have a unique zero; Assumption 5 guarantees that identification is not lost because of the incomplete nature of the data; Assumptions 12(ii) and (iii) guarantees that the weight matrices are chosen in such a way that identification is not lost.

We now show formally that well-separatedness (condition (ii) in PP89's Corollary 3.2)) holds. A sufficient condition is that $G(\theta)$ has a unique zero at θ_0, is continuous, and that the parameter space is compact. The latter two requirements are immediately implied by our Assumption 13(iii) and (iv). It remains to show that

$$G(\theta) = 0 \Leftrightarrow \theta = \theta_0.$$

To see that this is the case, note that

$$\begin{aligned} G(\theta) &= \sum_j A_j d_j E_{Z|X}\left[\frac{E[1\{D=d_j\}|X]}{p_{j,0}(X)} E[\psi(Z,\theta)|X]\right] \\ &= \left(\sum_j A_j\right) E_{Z|X}[E[\psi(Z,\theta)|X]] \\ &= AE[\psi(Z,\theta)] \end{aligned} \tag{A.29}$$

where the first step follows from MAR (Assumption 6(iii)) and a law of iterated expectations; the second step follows from the definition of $p_{j,0}(X)$, Assumption 12(ii) ($A_j d_j = A_j$), and the fact that the expectation does not depend on j; the final step follows from the definition of A and the law of iterated expectations.

Assumption 1 states that $AE[\psi(Z,\theta_0)] = 0$. Take any $\theta \neq \theta_0$. Then

$$E[\psi(Z,\theta)]' A'AE[\psi(Z,\theta)] > 0$$

because $E[\psi(Z,\theta)] \neq 0$ because of Assumption 1, and because $A'A$ has full rank (Assumption 12(iii)). Therefore, $AE[\psi(Z,\theta)] \neq 0$, which concludes our demonstration of identification.

Uniform convergence. Condition (iii) in PP89's Corollary 3.2 is implied by

$$\sup_\theta \|G_n(\theta) - G(\theta)\| = o_p(1), \tag{A.30}$$

where the supremum here and in the remainder of this proof is understood to be over the entire parameter space. By the triangle inequality,

$$\sup_\theta \|G_n(\theta) - G_0(\theta)\| \leq \sum_{j=1}^{J} (R_{1j} + R_{2j} + R_{3j})$$

where, letting $s_{ij} = 1\{D_i = d_j\}$,

$$
\begin{aligned}
R_{1j} &= \sup_\theta \left\| A_{j,n} \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{1}{\hat{p}_j(X_i)} - \frac{1}{p_{j,0}(X_i)} \right) s_{ij} d_j \psi(Z_i, \theta) \right) \right\|, \\
R_{2j} &= \sup_\theta \left\| (A_{j,n} - A_j) \left(\frac{1}{n} \sum_{i=1}^n \frac{s_{ij}}{p_{j,0}(X_i)} d_j \psi(Z_i, \theta) \right) \right\|, \\
R_{3j} &= \sup_\theta \left\| A_j \left(\frac{1}{n} \sum_{i=1}^n \frac{s_{ij}}{p_{j,0}(X_i)} d_j \psi(Z_i, \theta) - E\left[\frac{s_j}{p_{j,0}(X)} d_j \psi(Z, \theta) \right] \right) \right\|.
\end{aligned}
$$

If we can demonstrate that each of these terms is $o_p(1)$ for all j, we have demonstrated (A.30). First,

$$
\begin{aligned}
R_{1j} &\leq \|A_{j,n}\| \sup_\theta \left\| \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{\hat{p}_j(X_i)} - \frac{1}{p_{j,0}(X_i)} \right) s_{ij} d_j \psi(Z_i, \theta) \right\| \\
&\leq \|A_{j,n}\| \frac{1}{n} \sum_{i=1}^n \sup_\theta \left| \frac{1}{\hat{p}_j(X_i)} - \frac{1}{p_{j,0}(X_i)} \right| \|s_{ij} d_j\| \|\psi(Z_i, \theta)\| \\
&\leq \sqrt{p} \|A_{j,n}\| \frac{1}{n} \sum_{i=1}^n \left| \frac{1}{\hat{p}_j(X_i)} - \frac{1}{p_{j,0}(X_i)} \right| \sup_\theta \|\psi(Z_i, \theta)\| \qquad \text{(A.31)} \\
&\leq \sqrt{p} \|A_{j,n}\| \frac{1}{n} \sum_{i=1}^n \left| \frac{p_{j,0}(X_i) - \hat{p}_j(X_i)}{\hat{p}_j(X_i) p_{j,0}(X_i)} \right| \sup_\theta \|\psi(Z_i, \theta)\| \\
&\leq \frac{\sqrt{p}}{\kappa^2} \|A_{j,n}\| \frac{1}{n} \sum_{i=1}^n |p_{j,0}(X_i) - \hat{p}_j(X_i)| \sup_\theta \|\psi(Z_i, \theta)\| \\
&\leq \frac{\sqrt{p}}{\kappa^2} \|A_{j,n}\| \|p_{j,0} - \hat{p}_j\|_\infty \frac{1}{n} \sum_{i=1}^n \sup_\theta \|\psi(Z_i, \theta)\| = o_p(1),
\end{aligned}
$$

where the first inequality follows from the submultiplicative property of the norm; the second follows from the submultiplicative property and the fact that the sup of sums is smaller than the sum of sups; the third follows because $s_{ij} \in \{0, 1\}$ and $\|d_j\| \leq \|I\| = p$, and that the propensity scores do not depend on θ; the fourth rewrites the term with the propensity score term; the fifth uses the fact that the propensity scores are bounded below by $\kappa > 0$; the sixth replaces the differences in propensity scores at a point by the supremum over all points. The conclusion then follows from the fact that $A_{j,n}$ converges so that it is $O_p(1)$, that the propensity score estimator is uniformly consistent

by Assumption 11, and the fact that

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta}\left\|\psi\left(Z_i,\theta\right)\right\| \xrightarrow{p} E\left[\sup_{\theta}\left\|\psi\left(Z_i,\theta\right)\right\|\right] < \infty, \tag{A.32}$$

from a standard LLN and Assumption 13(ii).

Second,

$$\begin{aligned}
R_{2j} &\leq \left\|A_{j,n}-A_j\right\| \sup_{\theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{s_{ij}}{p_{j,0}\left(X_i\right)}d_j\psi\left(Z_i,\theta\right)\right\| \\
&\leq \left\|A_{j,n}-A_j\right\| \frac{1}{n}\sum_{i=1}^{n}\frac{s_{ij}}{p_{j,0}\left(X_i\right)}\left\|d_j\right\| \sup_{\theta}\left\|\psi\left(Z_i,\theta\right)\right\| \\
&\leq \frac{\sqrt{p}}{\kappa}\left\|A_{j,n}-A_j\right\| \frac{1}{n}\sum_{i=1}^{n}\sup_{\theta}\left\|\psi\left(Z_i,\theta\right)\right\| \\
&= \frac{\sqrt{p}}{\kappa}o_p\left(1\right)O_p\left(1\right) = o_p\left(1\right),
\end{aligned}$$

where the first two inequalities are as in the previous derivation; the third inequality uses that $s_{ij} \in \{0,1\}$ and $\left\|d_j\right\| \leq \left\|I_p\right\| = \sqrt{p}$. For the first equality, the $o_p\left(1\right)$ restates Assumption 12(i), whereas the $O_p\left(1\right)$ term follows from a law of large numbers and Assumption 13(ii), as in (A.32).

Finally,

$$\begin{aligned}
R_{3j} &\leq \left\|A_j\right\| \sup_{\theta}\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\frac{s_{ij}}{p_{j,0}\left(X_i\right)}d_j\psi\left(Z_i,\theta\right) - E\left[\frac{s_j}{p_{j,0}\left(X\right)}d_j\psi\left(Z,\theta\right)\right]\right)\right\| \\
&= \left\|A_j\right\| o_p\left(1\right) = o_p\left(1\right),
\end{aligned}$$

because $\left\|A_j\right\|$ is bounded because it is a limit, and the $o_p\left(1\right)$ term follows because

$$\left\{\frac{1\left\{\cdot = d_j\right\}}{p_{j,0}\left(\cdot\right)}d_j\psi\left(\cdot,\theta\right),\, \theta \in \mathcal{B}\right\} \tag{A.33}$$

is Glivenko-Cantelli, which follows from Assumption 13(i); that $p_{j,0}$ is bounded from below by κ (Assumption 6(iv)); and that Assumption 13(iii) implies a bounded envelope of the class in (A.33). □

A.3 Proof of Theorem 17

Proof of Theorem 17. This proof proceeds by verifying the conditions in PP89, Theorem 3.3, with $\theta = \beta$,

$$G(\theta) = E\left[\sum_j A_j \frac{1\{D = d_j\}}{p_{j,0}(X)} d_j \psi(Z, \beta)\right],$$

and

$$G_n(\theta) = \sum_j A_{j,n}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1\{D_i = d_j\}}{\hat{p}_j(X_i)} d_j \psi(Z_i, \beta)\right).$$

Condition (i) is satisfied by the definition of the IPW estimator in (5.2). To see that their condition (ii) is satisfied, remember from the proof of consistency (Theorem 14, page 56, equation (A.29)) that

$$G(\theta) = AE[\psi(Z, \theta)]$$

so that $G(\theta)$ is differentiable at θ_0 with derivative matrix $A\Gamma$. That $A\Gamma$ has full rank follows from the Assumption that A has full rank (Assumption 12(iii)) and that Γ has full rank (Assumption 15).

Instead of Pakes and Pollard's condition (iii), I will verify the slightly stronger condition that for all $\delta_n = o(1)$,

$$\sup_{\|\theta - \theta_0\| < \delta_n} \|G_n(\theta) - G(\theta) - G_n(\theta_0)\| = o_p\left(n^{-1/2}\right). \tag{A.34}$$

With the upcoming decomposition in mind, I define

$$H_{jn}(\theta, p) = \frac{1}{n}\sum_{i=1}^{n} \frac{1\{D_i = d_j\}}{p(X_i)} \psi(Z_i, \theta),$$
$$H_j(\theta, p) = E\left[\frac{1\{D = d_j\}}{p(X)} \psi(Z, \theta)\right],$$

so that

$$G_n(\theta) - G(\theta) - G_n(\theta_0) = \sum_j \left[A_{j,n} d_j H_{jn}(\theta, \widehat{p}_j) - A_j d_j H_j(\theta, p_{j,0}) - A_{j,n} d_j H_{jn}(\theta_0, \widehat{p}_j)\right].$$

Following Cattaneo (2010, proof of Theorem 4), the condition (A.34) is established by

verifying each of terms in the following decomposition is $o_p(1)$:

$$\sum_{j=1}^{J}\left(R_{1j}+R_{2j}+R_{3j}+R_{4j}\right),$$

where

$$\begin{aligned}
R_{1j} &\equiv \sup_{\|\theta-\theta_0\|<\delta_n}\sqrt{n}\left\|H_{jn}\left(\theta,p_{j,0}\right)-H_j\left(\theta,p_{j,0}\right)-H_{jn}\left(\theta_0,p_{j,0}\right)\right\|,\\
R_{2j} &\equiv \sup_{\|\theta-\theta_0\|<\delta_n}\sqrt{n}\left\|H_{jn}\left(\theta,\widehat{p}_j\right)-H_{jn}\left(\theta,p_{j,0}\right)-\Delta_{jn}\left(\theta,\widehat{p}_j-p_{j,0}\right)\right\|,\\
R_{3j} &\equiv \sup_{\|\theta-\theta_0\|<\delta_n}\sqrt{n}\left\|H_{jn}\left(\theta_0,p_{j,0}\right)-H_{jn}\left(\theta_0,\widehat{p}_j\right)-\left(-\Delta_{jn}\left(\theta_0,\widehat{p}_j-p_{j,0}\right)\right)\right\|,\\
R_{4j} &\equiv \sup_{\|\theta-\theta_0\|<\delta_n}\sqrt{n}\left\|\Delta_{jn}\left(\theta,\widehat{p}_j-p_{j,0}\right)-\Delta_{jn}\left(\theta_0,\widehat{p}_j-p_{j,0}\right)\right\|,
\end{aligned}$$

where

$$\Delta_{jn}\left(\theta,\widehat{p}_j-p_{j,0}\right)=-\frac{1}{n}\sum_{i=1}^{n}\frac{1\left\{D_i=d_j\right\}}{p_{j,0}^2\left(X_i\right)}\psi\left(Z_i,\theta\right)\left[\widehat{p}_j\left(X_i\right)-p_{j,0}\left(X_i\right)\right]$$

is the differential (with respect to the propensity score) of H_{jn} at $(\theta,p_{j,0})$. By the triangle inequality, demonstrating that each one of those terms is $o_p(1)$ implies condition (A.34).[17]

To see that R_{1j} is $o_p(1)$, it is sufficient to show that the class of functions

$$\left\{\frac{1\left\{\cdot=d_j\right\}}{p_{j,0}\left(\cdot\right)}\psi\left(\cdot,\theta\right):\theta\in\mathcal{B},\|\theta-\theta_0\|<\delta_n\right\}$$

is Donsker with a finite envelope. But this follows immediately from a preservation theorem for Donsker classes, e.g. van der Vaart and Wellner (1996, Theorem 2.10.6), given the local Donskerity of the original class of moment functions (Assumption 16(i)), which is assumed to have a finite envelope (Assumption 16(iii)), and by the assumption that the propensity score $p_{j,0}$ is bounded away from 0 uniformly (Assumption 6(iv)).

To see why $R_{2j}=o_p(1)$, note that the moment function, at every value θ, depends smoothly on the infinite-dimensional parameter $p_j(\cdot)$, and that the denominator is bounded away from 0. As a result, "linearization" is satisfied, see e.g. Newey (1994,

[17] The presence of the weight matrices does not create additional difficulties beyond notational ones, and the analysis with them included is therefore omitted.

Assumption 5.1). In other words, there exists a function $b(Z_i)$ such that

$$\frac{1\{D_i = d_j\}}{p_j(X_i)}\psi(Z_i,\theta) - \frac{1\{D_i = d_j\}}{p_{j,0}(X_i)}\psi(Z_i,\theta) - \frac{1\{D_i = d_j\}}{p_{j,0}^2(X_i)}\psi(Z_i,\theta)[p_j(X_i) - p_{j,0}(X_i)] \leq b(Z_i)\|p_j - p_0\|^2.$$

This upper bound does not depend on β. Because we assumed in the statement of the theorem that $\|\widehat{p}_j - p_{j,0}\|_\infty = o_p\left(n^{-1/4}\right)$, we can conclude that $R_{2j} = o_p(1)$. The same applies to R_{3j}.

Finally, for R_{4j},

$$\begin{aligned}
R_{4j} &= \sup_{\|\theta-\theta_0\|<\delta_n} \sqrt{n}\left\|\frac{1}{n}\sum_{i=1}^n \frac{1\{D_i = d_j\}}{p_{j,0}^2(X_i)}\left(\psi(Z_i,\theta) - \psi(Z_i,\theta_0)\right)\left(\widehat{p}_j(X_i) - p_{j,0}(X_i)\right)\right\| \\
&\leq \sup_{\|\theta-\theta_0\|<\delta_n} \sqrt{n}\frac{1}{n}\sum_{i=1}^n \frac{1\{D_i = d_j\}}{p_{j,0}^2(X_i)}\|\psi(Z_i,\theta) - \psi(Z_i,\theta_0)\|\,|\widehat{p}_j(X_i) - p_{j,0}(X_i)| \\
&\leq \|\widehat{p}_j - p_{j,0}\|_\infty \times \sup_{\|\theta-\theta_0\|<\delta_n} \sqrt{n}\frac{1}{n}\sum_{i=1}^n \frac{1\{D_i = d_j\}}{p_{j,0}^2(X_i)}\|\psi(Z_i,\theta) - \psi(Z_i,\theta_0)\|. \quad \text{(A.35)} \\
&= o_p(1).
\end{aligned}$$

The first inequality follows from the triangle inequality and the submultiplicativity of the norm, and the second from the definition of the sup-norm. The final equality follows from the uniform consistency of $\widehat{p}_j$ (Assumption 11) and the demonstration in the following paragraph.

For n large enough, consider the class of functions

$$\mathcal{Q} = \left\{\frac{1\{\cdot = d_j\}}{p_{j,0}^2(\cdot)}\|\psi(\cdot,\theta) - \psi(\cdot,\theta_0)\| : \theta \in \mathcal{B}, \|\theta - \theta_0\| < \delta_n\right\}.$$

This class is Donsker, because of a preservation theorem for Donsker classes, e.g. van der Vaart and Wellner (1996, Theorem 2.10.6). The conditions for preservation hold because of local Donskerity of the original class of moment functions (Assumption 16(i)), the assumption that it has a finite envelope (which follows from the triangle inequality and then applying Assumption 16(ii)), and the assumption that the propensity score is bounded away from 0 uniformly (Assumption 6(iv)). Letting

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^n \frac{1\{D_i = d_j\}}{p_{j,0}^2(X_i)}\|\psi(Z_i,\theta) - \psi(Z_i,\theta_0)\|$$

is the sample average across a function from $\mathcal{Q}$. By construction $Q_n(\theta_0) = 0$, so that the second term in (A.35) is a stochastic equicontinuity condition on $\mathcal{Q}$.

For condition (iv): that a central limit theorem applies to the sample criterion function follows from Lemma 5.1 in Newey (1994), henceforth N94. To see that the required Assumptions for N94's Lemma 5.1 hold, note that we have already used "linearization" (N94's Assumption 5.1) in verifying that $R_{2j} = o_p(1)$ and $R_{3j} = o_p(1)$. N94's Assumption 5.2 (stochastic equicontinuity of the derivative) was establish in the verification that $R_{4j} = o_p(1)$. Assumption 5.3 is verified in existing work on binary missing data, see the references below. All the conditions for Lemma 5.1 in Newey (1994) are therefore satisfied, and all that is required is to obtain an expression for the asymptotic variance.

To see that the asymptotic variance is V_A as in (5.4), note that the asymptotic variance for the sample objective function is

$$\begin{aligned}\operatorname{Var}\left[\sum_j A_j \frac{1\{D_i = d_j\}}{\widehat{p}_j(X)} d_j \psi(Z, \beta_0)\right] &= \sum_j A_j d_j \left(\operatorname{Var}\left[\frac{1\{D = d_j\}}{\widehat{p}_j(X)}\psi(Z, \beta_0)\right]\right) d_j A_j' \\ &= \sum_j A_j d_j \Omega_j d_j A_j',\end{aligned}$$

where the first equality follows from independence across strata and the second equality defines Ω_j.

The stratum-specific variance Ω_j is identical to that of the binary missing data case and can be imported from Chen et al. (2008), Cattaneo (2010), or Graham (2011). Because of linearization (see the analysis of R_{2j} and R_{3j} above), the asymptotic variance of the IPW moment conditions can be expressed as

$$\Omega_j = \operatorname{Var}\left[\frac{1\{D = d_j\}}{p_{j,0}(X)}\psi(Z, \beta_0) - \frac{1\{D = d_j\} - p_{j,0}(X)}{p_{j,0}(X)} q_0(X)\right], \tag{A.36}$$

where the first term corresponds to the case of a known propensity score, and the second term corrects for the fact that it is estimated. It is straightforward to show (see e.g. Cattaneo (2010)) that

$$\Omega_j = E\left[\frac{\Sigma_0(X)}{p_{j,0}(X)} + q_0(X) q_0'(X)\right], \tag{A.37}$$

so that the asymptotic variance of the limiting objective function is

$$V_A = \sum_j A_j d_j E\left[\frac{\Sigma_0(X)}{p_{j,0}(X)} + q_0(X)\, q_0'(X)\right] d_j A_j'.$$

Because condition (v) was assumed in the statement of the theorem, this concludes the proof. □

A.4 Proof of Theorem 23

Proof of Theorem 23. This proof follows closely that of Theorem 14, where we verified the conditions in PP89, Corollary 3.2. For the present case, set their θ to our β, and $G_n(\theta) = G_n^{DR}(\beta)$. Finally, set their limiting objective function $G(\theta)$ to

$$\begin{aligned} G^{DR}(\beta) &= E\left[\sum_j A_j \left(\frac{s_j}{\zeta_{1j}(h_1(X)\gamma_{j,0})} d_j \psi(Z,\beta) - \frac{s_j - \zeta_{1j}(h_1(X)\gamma_{j,0})}{\zeta_{1j}(h_1(X)\gamma_{j,0})} d_j \zeta_{2\beta}(h_2(X)\delta_0)\right)\right] \\ &= E\left[\sum_j A_j d_j g_j(Z,\beta)\right] \\ g_j(Z,\beta) &= \frac{s_j}{\zeta_{1j}(h_1(X)\gamma_{j,0})}\psi(Z,\beta) - \frac{s_j - \zeta_{1j}(h_1(X)\gamma_{j,0})}{\zeta_{1j}(h_1(X)\gamma_{j,0})}\zeta_{2\beta}(h_2(X)\delta_0) \end{aligned}$$

Identification. Their condition (i) is satisfied by the construction of $\widetilde{\beta}_n$ in (5.8). Because of compactness of $\mathcal{B}$ (Assumption 13(iv) and continuity of $E[\psi(Z,\beta)]$ in β at β_0 (Assumption 13(iii)), well-separatedness (their condition (ii)) holds if $G^{DR}(\beta) = 0 \Leftrightarrow \beta = \beta_0$. To see that this is the case, first note that

$$\begin{aligned} E[g_j(Z,\beta)] &= E\left[\frac{s_j}{\zeta_{1j}(h_1(X)\gamma_{j,0})}\psi(Z,\beta) - \frac{s_j - \zeta_{1j}(h_1(X)\gamma_{j,0})}{\zeta_{1j}(h_1(X)\gamma_{j,0})}\zeta_{2\beta}(h_2(X)\delta_0)\right] \\ &= E\left[\frac{p_{j,0}(X)}{\zeta_{1j}(h_1(X)\gamma_{j,0})} q(X,\beta) - \frac{p_{j,0}(X) - \zeta_{1j}(h_1(X)\gamma_{j,0})}{\zeta_{1j}(h_1(X)\gamma_{j,0})}\zeta_{2\beta}(h_2(X)\delta_0)\right]. \end{aligned} \tag{A.38}$$

If the propensity score is correctly specified, then $p_{j,0}(X) = \zeta_{1j}(h_1(X)\gamma_{j,0})$ so that (A.38) simplifies to

$$E[g_j(Z,\beta)] = E[1q(X,\beta) + 0\zeta_{2\beta}(h_2(X)\delta_0)] = E[\psi(Z,\beta)].$$

Alternatively, if the conditional expectation function is correctly specified, then $q(X,\beta) = \zeta_{2\beta}(h_2(X)\delta_0)$ so that

$$\begin{aligned} E[g_j(Z,\beta)] &= E\left[\frac{p_{j,0}(X)(q(X,\beta) - \zeta_{2\beta}(h_2(X)\delta_0))}{\zeta_{1j}(h_1(X)\gamma_{j,0})} + \zeta_{2\beta}(h_2(X)\delta_0)\right] \\ &= E[\zeta_{2\beta}(h_2(X)\delta_0)] = E[\psi(Z,\beta)]. \end{aligned}$$

Therefore, if either of the working models is correctly specified, $G^{DR}(\beta)$ has a unique root at zero. This follows from the restrictions imposed on the matrices A_j, d_j, as in the proof of Theorem 14. For more details, see the proof around equation (A.29) on page 56.

Uniform convergence. The final condition (iii) in PP89's Corollary 3.2 to check is that

$$\sup_{\beta\in\mathcal{B}} \left\| G_n^{DR}(\beta) - G^{DR}(\beta) \right\| = o_p(1).$$

By the triangle inequality,

$$\sup_{\beta\in\mathcal{B}} \left\| G_n^{DR}(\beta) - G^{DR}(\beta) \right\| \leq \sup_{\beta\in\mathcal{B}} \sum_j \left\| A_{jn} d_j G_{1n}(\beta) - A_j d_j G_1(\beta) \right\|$$

with

$$\begin{aligned} G_{1n}(\beta) &= \frac{1}{n}\sum_{i=1}^n \left(\frac{s_{ij}}{\zeta_{1j}(h_1(X_i)\widehat{\gamma}_{j,n})}\psi(Z_i,\beta) - \frac{s_{ij} - \zeta_{1j}(h_1(X_i)\widehat{\gamma}_{j,n})}{\zeta_{1j}(h_1(X_i)\widehat{\gamma}_{j,n})}\zeta_{2\beta}\left(h_2(X_i)\widehat{\delta}_n\right)\right) \\ &= \frac{1}{n}\sum_{i=1}^n \left(\frac{s_{ij}\left(\psi(Z_i,\beta) - \zeta_{2\beta}\left(h_2(X_i)\widehat{\delta}_n\right)\right)}{\zeta_{1j}(h_1(X_i)\widehat{\gamma}_{j,n})} + \zeta_{2\beta}\left(h_2(X_i)\widehat{\delta}_n\right)\right) \\ G_1(\beta) &= E\left[\frac{s_j(\psi(Z,\beta) - \zeta_{2\beta}(h_2(X_i)\delta_0))}{\zeta_{1j}(h_1(X_i)\gamma_{j,0})} + \zeta_{2\beta}(h_2(X_i)\delta_0)\right]. \end{aligned}$$

Using the triangle inequality, and the fact that A_{jn}, A_j, d_j can be ignored in $o_p(1)$ calculations (see proof of Theorem 14), implies that it is sufficient to establish that the

following two terms are $o_p(1)$:

$$R_{1j} = \sup_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\frac{s_{ij}}{\zeta_{1j}\left(h_1\left(X_i\right)\gamma_{j,0}\right)}\psi\left(Z_i,\beta\right) - \frac{s_{ij}-\zeta_{1j}\left(h_1\left(X_i\right)\gamma_{j,0}\right)}{\zeta_{1j}\left(h_1\left(X_i\right)\gamma_{j,0}\right)}\zeta_{2\beta}\left(h_2\left(X_i\right)\delta_0\right)\right) \right.$$
$$\left. -E\left[\frac{s_j}{\zeta_{1j}\left(h_1\left(X\right)\gamma_{j,0}\right)}\psi\left(Z,\beta\right) - \frac{s_j-\zeta_{1j}\left(h_1\left(X\right)\gamma_{j,0}\right)}{\zeta_{1j}\left(h_1\left(X\right)\gamma_{j,0}\right)}\zeta_{2\beta}\left(h_2\left(X\right)\delta_0\right)\right] \right\|,$$

$$R_{2j} = \sup_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\frac{s_{ij}}{\zeta_{1j}\left(h_1\left(X_i\right)\widehat{\gamma}_{j,n}\right)}\psi\left(Z_i,\beta\right) - \frac{s_{ij}-\zeta_{1j}\left(h_1\left(X_i\right)\widehat{\gamma}_{j,n}\right)}{\zeta_{1j}\left(h_1\left(X_i\right)\widehat{\gamma}_{j,n}\right)}\zeta_{2\beta}\left(h_2\left(X_i\right)\widehat{\delta}_n\right)\right) \right. \tag{A.39}$$
$$\left. -\frac{1}{n}\sum_{i=1}^{n} \left(\frac{s_{ij}}{\zeta_{1j}\left(h_1\left(X_i\right)\gamma_{j,0}\right)}\psi\left(Z_i,\beta\right) - \frac{s_{ij}-\zeta_{1j}\left(h_1\left(X_i\right)\gamma_{j,0}\right)}{\zeta_{1j}\left(h_1\left(X_i\right)\gamma_{j,0}\right)}\zeta_{2\beta}\left(h_2\left(X_i\right)\delta_0\right)\right) \right\|.$$

R1j. That $R_{1j} = o_p(1)$ follows because

$$\mathcal{F} = \left\{ \frac{s_j\psi\left(\cdot,\beta\right)}{\zeta_{1j}\left(h_1\left(\cdot\right)\gamma_{j,0}\right)} - \frac{s_j-\zeta_{1j}\left(h_1\left(\cdot\right)\gamma_{j,0}\right)}{\zeta_{1j}\left(h_1\left(\cdot\right)\gamma_{j,0}\right)}\zeta_{2\beta}\left(h_2\left(\cdot\right)\delta_0\right), \beta\in\mathcal{B}\right\}$$

is Glivenko-Cantelli (GC). To see that this is the case, first note that

$$\mathcal{F}_1 = \left\{ \frac{s_j\psi\left(\cdot,\beta\right)}{\zeta_{1j}\left(h_1\left(\cdot\right)\gamma_{j,0}\right)}, \beta\in\mathcal{B}\right\}$$

is GC because of Assumption 13(i) (that the class of ψ is GC) and that $\zeta_{1j}\left(h_1\left(\cdot\right)\gamma_{j,0}\right)$ is bounded away from 0 (Assumption 21(ii)). Second, note that

$$\mathcal{F}_2 = \left\{ \frac{s_j-\zeta_{1j}\left(h_1\left(\cdot\right)\gamma_{j,0}\right)}{\zeta_{1j}\left(h_1\left(\cdot\right)\gamma_{j,0}\right)}\zeta_{2\beta}\left(h_2\left(\cdot\right)\delta_0\right), \beta\in\mathcal{B}\right\}$$

because the $\zeta_{2\beta}\left(h_2\left(\cdot\right)\delta_0\right)$ form a GC class (Assumption 21(i)), and that $\zeta_{1j}\left(h_1\left(\cdot\right)\gamma_{j,0}\right)$ is bounded away from 0 (Assumption 21(ii)).

R2j. That $R_{2j} = o_p(1)$ follows because of an application of the mean-value theorem. It states that for $D_n\left(\widetilde{\gamma}_{j,n},\widetilde{\delta}_n\right)$ - the derivative of (A.39), evaluated at some intermediate

value $\left(\widetilde{\gamma}_{j,n}, \widetilde{\delta}_n\right)$,

$$\begin{aligned} R_{2j} &= \sup_{\beta \in \mathcal{B}} \left\| D_n\left(\widetilde{\gamma}_{j,n}, \widetilde{\delta}_n\right) \begin{bmatrix} \widehat{\gamma}_{j,n} - \gamma_0 \\ \widehat{\delta}_n - \delta_0 \end{bmatrix} \right\| \\ &\leq \left\| \begin{bmatrix} \widehat{\gamma}_{j,n} - \gamma_0 \\ \widehat{\delta}_n - \delta_0 \end{bmatrix} \right\| \sup_{\beta \in \mathcal{B}} \left\| D_n\left(\widetilde{\gamma}_{j,n}, \widetilde{\delta}_n\right) \right\| \end{aligned}$$

which is $o_p(1)$ if $\sup_{\beta \in \mathcal{B}} \left\| D_n\left(\widetilde{\gamma}_{j,n}, \widetilde{\delta}_n\right) \right\| = O_p(1)$. The derivative with respect to γ is,

$$D_{n,\gamma}(\gamma, \delta) = -\frac{1}{n} \sum_i \left[\frac{s_{ij} \left(\psi\left(Z_i, \beta\right) - \zeta_{2\beta}\left(h_2\left(X_i\right) \delta\right)\right)}{\zeta_{1j}^2\left(h_1\left(X_i\right) \widehat{\gamma}_{j,n}\right)} \zeta_{1j}^{'}\left(h_1\left(X_i\right) \gamma\right) h_1(X) \right],$$

where $\zeta_{1j}^{'}$ is the derivative of ζ_{1j} with respect to its index argument. For the first term, note that

$$\begin{aligned} E\left[\sup_{\beta \in \mathcal{B},} \left\| \frac{s_{ij} \left(\psi\left(Z_i, \beta\right) - \zeta_{2\beta}\left(h_2\left(X_i\right) \delta\right)\right)}{\zeta_{1j}^2\left(h_1\left(X_i\right) \gamma\right)} \right\| \right] &\leq \frac{1}{\kappa^2} E\left[\sup_{\beta \in \mathcal{B}} \left\| \psi\left(Z_i, \beta\right) - \zeta_{2\beta}\left(h_2\left(X_i\right) \delta\right) \right\| \right] \\ &\leq \frac{1}{\kappa^2} \left(E\left[\sup_{\beta \in \mathcal{B},} \left\| \psi\left(Z_i, \beta\right) \right\| \right] + E\left[\left\| \zeta_{2\beta}\left(h_2\left(X_i\right) \delta\right) \right\| \right] \right) \end{aligned}$$

where the first term is bounded because of Assumption 13(i) and the second term is bounded because of Assumption 22(ii). For the second term,

$$E\left[\zeta_{1j}^{'}\left(h_1\left(X_i\right) \widetilde{\gamma}_n\right) h_1(X) \right] < \infty$$

because of Assumption 22(i).

The derivative with respect to δ is

$$D_{n,\gamma}(\gamma, \delta) = -\frac{1}{n} \sum_i \left[\frac{s_{ij} - \zeta_{1j}\left(h_1\left(X_i\right) \gamma\right)}{\zeta_{1j}\left(h_1\left(X_i\right) \gamma\right)} \zeta_{2\beta}^{'}\left(h_2\left(X_i\right) \delta\right) h_2\left(X_i\right) \right]$$

so that the desired boundedness follows from $\kappa \leq \zeta_{1j}\left(h_1\left(X_i\right) \gamma\right) \leq 1 - (J-1)\kappa$ and Assumption 22(iii). □

A.5 Proof of Theorem 25

Proof of Theorem 25. All the conditions for Newey and McFadden (1994, Theorem 6.1) are satisfied. All that remains is to derive the expression for the asymptotic variance under Assumption 19. This is the key part of our result since it implies local efficiency of the DR estimator.

We first derive the asymptotic variance of the moment function for the case that γ_0 and δ_0 are known. Then, we show that the adjustment term from plugging in parametric estimators for the working model parameters is zero.

First, the known-(γ_0, δ_0) variance of the moment conditions h_j,

$$\begin{aligned} h_j(Z,\beta) &= \frac{s_j \psi(Z,\beta)}{\zeta_{1j}(h_1(X)\gamma_{j,0})} - \frac{s_j - \zeta_{1j}(h_1(X)\gamma_{j,0})}{\zeta_{1j}(h_1(X)\gamma_{j,0})}\zeta_{2\beta}(h_2(X)\delta_0) \\ &= h_{1j} - h_{2j}. \end{aligned}$$

Because $E[h_{1j}] = E[h_{2j}] = 0$, and because the cross-term will be shown to be symmetric,

$$\mathrm{Var}[h_j(Z,\beta)] = E\left[h_{1j}(Z,\beta)\, h_{1j}'(Z,\beta)\right] + E\left[h_{2j}(Z,\beta)\, h_{2j}'(Z,\beta)\right] - 2E\left[h_{1j}(Z,\beta)\, h_{2j}'(Z,\beta)\right].$$

Dealing with the terms in order of appearance, and using that $\zeta_{1j}(h_1(X)\gamma_{j,0}) = p_{j,0}(X)$ and $\zeta_{2\beta}(h_2(X)\delta_0) = q(X,\beta)$, we obtain

$$\begin{aligned} E\left[h_{1j}(Z,\beta)\, h_{1j}'(Z,\beta)\right] &= E\left[E\left[\frac{s_j \psi^2(Z,\beta)}{p_{j,0}^2(X)}\middle| X\right]\right] \\ &= E\left[\frac{\Sigma_0(X) + q(X,\beta)\, q(X,\beta)'}{p_{j,0}(X)}\right], \end{aligned}$$

and

$$\begin{aligned} E\left[h_{2j}(Z,\beta)\, h_{2j}'(Z,\beta)\right] &= E\left[\frac{(s_j - p_{j,0}(X))^2}{p_{j,0}^2(X)} q(X,\beta)\, q(X,\beta)'\right] \\ &= E\left[\frac{1 - p_{j,0}(X)}{p_{j,0}(X)} q(X,\beta)\, q(X,\beta)'\right], \end{aligned}$$

and

$$E\left[h_{1j}\left(Z,\beta\right)h_{2j}^{'}\left(Z,\beta\right)\right]=E\left[\frac{s_{j}\left(1-p_{j,0}\left(X\right)\right)}{p_{j,0}^{2}\left(X\right)}\psi\left(Z,\beta\right)q^{'}\left(X,\beta\right)\right]$$
$$=E\left[\frac{1-p_{j,0}\left(X\right)}{p_{j,0}\left(X\right)}q\left(X,\beta\right)q^{'}\left(X,\beta\right)\right]$$
$$=E\left[h_{2}\left(Z,\beta\right)h_{1}^{'}\left(Z,\beta\right)\right].$$

Then the variance is

$$\mathrm{Var}\left[h_{j}\left(Z,\beta\right)\right]=E\left[\frac{\Sigma_{0}\left(X\right)+q\left(X,\beta\right)q\left(X,\beta\right)^{'}}{p_{j,0}\left(X\right)}\right]+E\left[\frac{1-p_{j,0}\left(X\right)}{p_{j,0}\left(X\right)}q\left(X,\beta\right)q\left(X,\beta\right)^{'}\right]$$
$$-2E\left[\frac{1-p_{j,0}\left(X\right)}{p_{j,0}\left(X\right)}q\left(X,\beta\right)q^{'}\left(X,\beta\right)\right] \tag{A.40}$$
$$=E\left[\frac{\Sigma_{0}\left(X\right)}{p_{j,0}\left(X\right)}+q\left(X,\beta\right)q\left(X,\beta\right)^{'}\right]. \tag{A.41}$$

Next, we account for the adjustment term owing to the fact that the parameters in the working models are plugged in. For the propensity score, the adjustment term is

$$E\left[\left.\frac{\partial h_{j}\left(Z,\beta\right)}{\partial\gamma_{j}}\right|_{\gamma_{j}=\gamma_{j,0}}\right]=E\left[\left.\frac{\partial\left(\frac{s_{j}\left[\psi\left(Z,\beta\right)-q\left(X,\beta\right)\right]}{\zeta_{1j}\left(h_{1}\left(X\right)\gamma\right)}+\zeta_{2\beta}\left(h_{2}\left(X\right)\delta_{0}\right)\right)}{\partial\gamma_{j}}\right|_{\gamma_{j}=\gamma_{j,0}}\right]$$
$$=E\left[\left.-\frac{s_{j}\left[\psi\left(Z,\beta\right)-q\left(X,\beta\right)\right]}{\zeta_{1j}^{2}\left(h_{1}\left(X\right)\gamma\right)}D_{1j}\left(h_{1}\left(X\right)\gamma\right)h_{1}\left(X\right)\right|_{\gamma_{j}=\gamma_{j,0}}\right]$$
$$=0$$

because $E\left[\psi\left(Z,\beta\right)-q\left(X,\beta\right)\right|X]=0$. Similarly,

$$E\left[\left.\frac{\partial h_{j}\left(Z,\beta\right)}{\partial\delta}\right|_{\delta=\delta_{0}}\right]=E\left[\left.\frac{\partial\left(\frac{s_{j}\psi\left(Z,\beta\right)}{\zeta_{1j}\left(h_{1}\left(X\right)\gamma_{j,0}\right)}-\frac{s_{j}-\zeta_{1j}\left(h_{1}\left(X\right)\gamma_{j,0}\right)}{\zeta_{1j}\left(h_{1}\left(X\right)\gamma_{j,0}\right)}\zeta_{2\beta}\left(h_{2}\left(X\right)\delta\right)\right)}{\partial\delta}\right|_{\delta=\delta_{0}}\right]$$
$$=E\left[\left.-\frac{s_{j}-\zeta_{1j}\left(h_{1}\left(X\right)\gamma_{j,0}\right)}{\zeta_{1j}\left(h_{1}\left(X\right)\gamma_{j,0}\right)}\times\frac{\partial\left(\zeta_{2\beta}\left(h_{2}\left(X\right)\delta\right)\right)}{\partial\delta}\right|_{\delta=\delta_{0}}\right]$$
$$=0$$

because $E\left[s_{j}-\zeta_{1j}\left(h_{1}\left(X\right)\gamma_{j,0}\right)\right|X]=0$. Because both adjustment terms are zero,

the asymptotic variance of the estimator follows from the known-$(\gamma_{j,0}, \delta_0)$ variance in (A.41). □

A.6 Proof of Theorem 26

Proof of Theorem 26. We show that $\widehat{\Omega}_j$ is consistent for $d_j \Omega_j d_j$. The result then follows from the restrictions on the weight matrices (Assumption 12, in particular $A_j d_j = A_j$) and standard results for consistent variance matrix estimation for GMM, see e.g. Theorem 4.2 in Newey and McFadden (1994).

To see that $\widehat{\Omega}_j$ is consistent under the assumptions in the statement of the theorem, first note that it is sufficient to show that

$$\left\| \widehat{\Omega}_j^0 - \Omega_j \right\| = o_p(1),$$

where

$$\widehat{\Omega}_j^0 = \frac{1}{n} \sum_{i=1}^{n} \frac{s_{ij}}{\widehat{p}_j(X_i)} \psi(Z_i, \beta_n) \psi(Z_i, \beta_n)',$$

i.e. we can drop the selection matrices d_j. We will investigate the two terms on the righthandside of

$$\left\| \widehat{\Omega}_j^0 - \Omega_j \right\| \leq \left\| \widehat{\Omega}_j^0 - \Omega_j^{00} \right\| + \left\| \widehat{\Omega}_j^{00} - \Omega_j \right\|, \tag{A.42}$$

where

$$\widehat{\Omega}_j^{00} = \frac{1}{n} \sum_{i=1}^{n} \frac{s_{ij}}{p_{j,0}(X_i)} \psi(Z_i, \beta_n) \psi(Z_i, \beta_n)'.$$

That the second term in (A.42) is $o_p(1)$ follows from Newey and McFadden (1994, Lemma 4.3): we have assumed the required conditions for the application of that Lemma in the statement of the result. That the first term is zero follows from manipulations that we have done previously for the proof of consistency of the IPW estimator, in particular equation (A.31) (page 57). It requires Assumption (ii) in the statement of the theorem, and uses the fact that the propensity scores are bounded below by a constant $\kappa > 0$. □

B Multivalued treatment effects

The goal of this Appendix is to clarify the difference between existing results on multivalued treatment effects (MVT) in Cattaneo (2010) and the results on incomplete data

(ID) in the present manuscript.

The MVT and ID settings both distinguish strata in the population of interest. In the MVT setting, a stratum consists of all the units assigned to one of the J treatment levels. For example, in a setting with two different types of treatment, A and B, one stratum corresponds to control units, another to the units that were exposed to treatment A, and a third to those that were exposed to treatment B. In contrast, in the ID setting, two units are in the same stratum if and only if each unit has measurements on the same set of variables. For example, in a panel data setting, one stratum may correspond to the cross-section units with a complete time series; another stratum to the units that drop out at some point; and another stratum consists of a refreshment sample.

In the standard case with two strata ($J = 2$), MVT reduces to the standard treatment effects model under unconfoundedness, whereas ID reduces to the missing data case under MAR. In that special case, results are easily ported between the treatment effect and missing data setting.

However, with multiple strata ($J > 2$), this is no longer the case. There are two key differences between the MVT and ID setting. The first difference concerns the moment functions and parameters of interest. In MVT, (a) all moment functions are observable for each stratum, and (b) the true value of the parameter varies across strata. For the ID case, (a) a different subset of moment functions is observable in each stratum, and (b) the true parameter value does not change between strata. I elaborate on this point in Section B.1, and show that the ID setting nests the MVT setting.

The second difference is that MVT requires identification of the parameter in each stratum. ID only requires identification from the union of all strata. I elaborate on this point in Section B.2, and provide empirically relevant settings where the MVT results are not applicable, but the ID results are.

B.1 Strata and moment conditions

The MVT and ID framework were constructed for different purposes. In an ID setting, the target parameter is the same as that in a setting with complete data, call it θ_0. The absence of complete information is an obstacle to learning about θ_0. It makes this task more complicated than it would be in the presence of complete information. However, the task is unchanged.

Consequently, the population parameter in the ID setting is common to all strata.

However, the available information on θ_0 varies across strata. The starting point for the ID analysis is a moment condition that holds in the population,

$$E\left[\psi\left(Z,\theta_0\right)\right]=0. \tag{B.1}$$

Because of incomplete information, only some of the elements of ψ are observed in each stratum. The observable elements are indicated by a missing data matrix d_j. In conjunction with the MAR assumption, the implied observable moment conditions for stratum $j=1,\cdots,J$, are

$$E\left[\frac{d_j}{p_j\left(X\right)}\psi\left(Z,\theta_0\right)\right]=0. \tag{B.2}$$

In the ID setting, the true parameter value does not depend on j. The implied moment conditions B.2 do, and so do the elements that are observable. Interest is in the population parameter θ_0, and the different strata provide complementary information about it.

In the MVT setting, strata are defined through treatment assignment. Each stratum (treatment level) j is associated with a potential outcome, $Z(j)$. The j-th stratum provides information about the features of the j-th potential outcome. The parameter θ_{j0} quantifies the feature of interest in stratum j, e.g. the mean or quantile of a response for that level of treatment. The goal of an MVT analysis is to investigate how this parameter varies across strata.

For this reason, the value of the parameter changes between strata:

$$E\left[\phi\left(Z\left(j\right),\theta_{j0}\right)\right]=0.$$

Interest centers on e.g. $(\theta_{j0}-\theta_{10},\, j=2,\cdots,J)$, some feature of the outcome in each stratum relative to a baseline ("dose-response").

The unconfoundedness assumption then implies that for each $j=1,\cdots,J$, the following moment conditions hold:

$$E\left[\frac{1}{p_j\left(X\right)}\phi\left(Z,\theta_{j0}\right)\right]=0, \tag{B.3}$$

where Z is the observed outcome.

Despite the very different objectives of the ID and MVT frameworks, the implied moment conditions (B.2 for ID; B.3 for MVT) look very similar. There are two differ-

ences. First, the ID framework allows for different elements of the moment function to be observable in different strata, as can be seen from the presence of d_j in B.2. Second, the MVT framework allows for the parameter to be strata-dependent, as can be seen from the presence of θ_{j0} in B.3.

As it turns out, the MVT framework is nested by ID. Starting from the MVT setting, define

$$\theta_0 \equiv (\theta_{10}, \cdots, \theta_{J0}),$$

and stack the implied moment conditions associated with treatment levels $1, 2, \cdots, J$ into a vector ψ.

$$\Phi(Z, \theta_0) = \begin{bmatrix} \frac{1}{p_1(X)}\phi(Z, \theta_{10}) \\ \frac{1}{p_2(X)}\phi(Z, \theta_{20}) \\ \vdots \\ \frac{1}{p_J(X)}\phi(Z, \theta_{J0}) \end{bmatrix}.$$

Now set $d_j = e_j \otimes I_q$, where e_j is the unit vector and I_p is the identity matrix of size p (the number of elements in the moment function $\tilde{\psi}_j$). Then d_j indicates the set of moment conditions that are observable in each stratum. In other words, stratum 1 provides information on the first block of Φ, stratum 2 provides information on the second block of Ψ, and so on. We have therefore reformulated an arbitrary MVT model as an ID model, by setting $\psi = \Phi$ and $d_j = (e_j \otimes I_q)$, $j = 1, \cdots, J$.

B.2 Identification

Assumption 1 in MVT is "identification in every stratum": for each $j = 1, \cdots, J$, there exists a θ_{j0} such that

$$E[\psi_j(Z, \theta)] = 0 \Leftrightarrow \theta = \theta_{j0}.$$

In contrast, identification in the ID setting only requires "global identification", see Assumption 5. Global identification amounts to

$$E\begin{bmatrix} \psi_1(Z, \theta) \\ \psi_2(Z, \theta) \\ \vdots \\ \psi_J(Z, \theta) \end{bmatrix} = 0 \Leftrightarrow \theta = \theta_0,$$

i.e. the information from all strata should jointly be sufficient for identification of θ_0.

This difference in identification assumptions can be important. It can show up in one of two ways: (1) the parameter is identified in one or more strata, but not in all; (2) the parameter is not identified in any stratum. To see the difference between cases 1 and 2, consider the linear instrumental variables examples (Examples 2 and 27 on page 7). In Example 2, the regression coefficient is identified in stratum 1, so case (1) applies. In Example 27, case (2) applies because the regression coefficient cannot be recovered using any one stratum.

In case (1), the full results in the manuscript are not necessary for identification. For example, one could use the following simple procedure: discard all observations not in stratum 1, and use only the observations in stratum 1 to estimate the regression coefficient. Identification is retained, but at the cost of a loss of efficiency. Similarly, in the case of unbalanced panel data, a complete-data analysis can be applied to a balanced subpanel. See Examples 4 and 29. The efficiency loss from using such simple procedures can be substantial, as demonstrated in the Monte Carlo results in Section 6.

Case (2) is more problematic for existing estimators. In case (2), the MVT results are no longer informative, because the identification assumption is violated and the results from the present manuscript are required for identification and estimation. Appendix C contains two cases in addition to the linear IV example 27. In Example 28, a case of difference-in-differences where the period-0 treatment status is not observed; and in Example 3 an example of dynamic panels with a rotating panel data structure, where any one cohort is insufficient to identify the parameters of interest.

C More examples

This Appendix contains some additional examples to which the general results in this paper apply.

Example 27 (Linear IV without stratum identification)**.** Starting from Example 2, suppose that exactly one of the instruments is missing. In that case, there are three strata, i.e. $J = 2$. In stratum 1, only the instrument W_1 is observed. In stratum 2, only the instrument W_2 is observed, i.e.:

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

The regression coefficient is identified in stratum j if the observable moment function, $d_j\psi$, contains sufficient information to identify it. In Example 2, the only stratum in which the parameter is identified is stratum 1. In the current example, the regression coefficient is not identified in any of the strata, since in both strata only one instrument is observed.

Example 28 (Botosaru and Gutierrez, 2018)**.** Consider a difference-in-differences setting with two time periods. A cross-section is available from each time period. The random variables in the standard model are $(T, D, Y_0(0), Y_0(1), Y_1(0), Y_1(1))$, where $T \in \{0, 1\}$ indicates the time period, $D \in \{0, 1\}$ indicates treatment status, and $Y_t(D)$ is the potential outcome under treatment status D in period t. The parameter of interest is the average treatment effect for the treated,

$$ATT = E\left[Y_1(1) - Y_1(0) \middle| D = 1\right].$$

Denote by $Y_t = DY_t(1) + (1 - D)Y_t(0)$ the observed outcome in period t. The ATT can be recovered from the distributions (D, Y_0) and (D, Y_1), both of which are observable in the standard difference-in-differences setup.

Botosaru and Gutierrez (2018) consider the case where D is not observable for the individuals with $T = 0$: the pre-intervention treatment status is not recorded. They propose an estimator that relies on the availability of a proxy Z_t and some additional exclusion and stationarity conditions - see their paper for additional details. Under an additional assumption (not made in the original paper) that the propensity score is

logistic, i.e.

$$P\left(D=1|\,Z_t\right)=\Lambda\left(Z_t\gamma\right)\equiv\frac{\exp\left(Z_t\gamma\right)}{1+\exp\left(Z_t\gamma\right)}.$$

and that the reduced-form conditional outcome equations are as in their setting,

$$\begin{aligned}E\left(Y_1|\,D,Z_1\right)&=\delta_{10}+\delta_{11}D+\delta_2 Z_1,\\ E\left(Y_0|\,Z_0\right)&=\delta_{00}+\delta_{01}\Lambda\left(Z_0\gamma\right)+\delta_2 Z_0,\end{aligned}$$

Botosaru and Gutierrez (2018) show that the model parameters satisfy the moment conditions

$$\begin{aligned}0&=E\begin{bmatrix}Z_1'\left(D-\Lambda\left(Z_1\gamma\right)\right)\\ \left(1,D,Z_1\right)'\left(Y_1-\delta_{10}-\delta_{11}D-\delta_2 Z_1\right)\\ \left(1,Z_0\right)'\left(Y_0-\delta_{00}-\delta_{01}\Lambda\left(Z_0\gamma\right)-\delta_2 Z_0\right)\end{bmatrix}\\ &\equiv E\begin{bmatrix}\psi_1\left(Z_1,D;\gamma\right)\\ \psi_2\left(Y_1,Z_1,D;\delta_{10},\delta_{11},\delta_2\right)\\ \psi_3\left(Y_0,Z_0;\gamma,\delta_{00},\delta_{01},\delta_2\right)\end{bmatrix}.\end{aligned}\tag{C.1}$$

The first moment condition pins down the propensity score parameter γ. The second pins down parameters in the period-1 outcome equation. The third uses the propensity score to pin down the parameters in the period-0 outcome equation. Identification of the model parameters leads to identifcation of the $ATT=\delta_{11}-\delta_{01}$.

In this example, there are two strata. The first stratum (d_1 in the display below) consists of the period-0 observations. For these observations, we observe (Y_0,Z_0). The second stratum consists of the period-1 outcomes (d_2 in the display below). From inspection of the moment functions in (C.1), we conclude that

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

Using any single stratum, we cannot identify the ATT. However, global identification holds, so that the identification and estimation results in the present paper apply.

The results in the present manuscript apply immediately to the following generalizations: (1) D may not be available in either period 1, but identification can be

	Unavailable components				
	None	$Y_{i,1}$	$Y_{i,4}$	$(Y_{i,1}, Y_{i,4})$	$Y_{i,2}$
$Y_{i,1}\Delta u_{i,3}$	X	·	X	·	·
$Y_{i,1}\Delta u_{i,4}$	X	·	·	·	X
$Y_{i,1}\Delta u_{i,5}$	X	·	·	·	X
$Y_{i,2}\Delta u_{i,4}$	X	X	·	·	·
$Y_{i,2}\Delta u_{i,5}$	X	X	·	·	·
$Y_{i,3}\Delta u_{i,5}$	X	X	·	·	X

Table 7: Strata for dynamic panels.

obtained by obtaining (D, Z) in a third stratum (see note 12 in Botosaru and Gutierrez, 2018); (2) The moment functions in (C.1) can be modified to allow for more flexible, per-treatment-status outcome equations in period 1, which would lead to four moment conditions with three strata; (3) they allow for a straightforward extension to the case of non-binary treatment D, any number of time periods, etc.

Example 29 (Dynamic panel data)**.** The goal of this example is to illustrate that the framework in this paper remains useful under complex missing data patterns. Consider the panel data AR(1) model

$$Y_{i,t} = \alpha_i + \rho Y_{i,t-1} + u_{i,t},\ t = 2, \cdots, T, \tag{C.2}$$

where interest lies in the autoregressive parameter ρ. Covariates can be added at the cost of additional notation.

Assume that $E(\alpha_i) = 0,\ Var(\alpha_i) = \sigma_a^2$,$E(u_{it}) = 0,\ Var(u_{it}) = 1$, and that there is no autocorrelation in the error terms: $E(u_{i,t}u_{i,s}) = 0$ for all $s \neq t$. For this case, 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 $E(Y_{i,t-s}\Delta u_{i,t}) = 0,\ t \geq 3, s \geq 2$.

if for any observation i, the dependent variable $Y_{i,t}$ is not observed in a given time period, then several components of the moment function are not observed. Table 7 illustrates the relationship between incompleteness of an observation and the incompletely observed moment function for the case of $T = 5$, with six moment conditions.7.

If $Y_{i,1}$ is missing, observation i still contributes to three sample moments. If $Y_{i,4}$ is missing, only one component of the moment function can be evaluated. More generally, the estimator proposed in this paper efficiently accommodates static and dynamic panel

Parameter	Available	Complete	Incomplete
β	0.0106	0.0128	0.0084
90-89	0.0025	0.0030	0.0026
91-90	0.0044	0.0031	0.0031
92-91	0.0033	0.0024	0.0024
93-92	0.0057	0.0067	0.0049
94-93	0.0039	0.0043	0.0033
95-94	0.0038	0.0042	0.0029
96-95	0.0042	0.0049	0.0037

Table 8: Standard deviations of all parameters, static model, three methods.

data models with unbalanced panels with different starting points, different endpoints, and any combination of gaps.

Existing approaches that use all available measurements can still be inefficient. One such approach is the available case estimator, which replaces missing moments by zeros before applying the full data estimation procedure. The available-case estimator is consistent if there is no selection: it corresponds to the procedure suggested in Arellano and Bond (1991, p. 281). The working paper version of this paper contains a simulation study highlighting the efficiency gains of using the efficient estimators proposed in this paper.

D Empirical illustration material

This Section contains supporting material for the empirical illustration. The first figure displays a histogram of the number of time periods available per firm. The second figure shows data availability for the first few firms. The third figure presents box plots based on the bootstrap for different estimators in the static model.

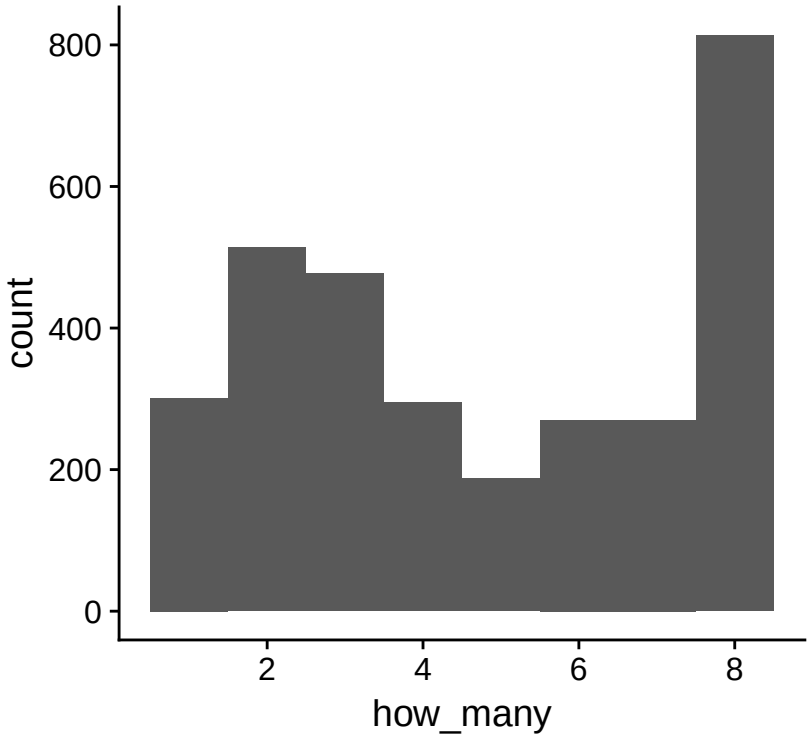

Figure D.1: Histogram for the number of time periods per firm.

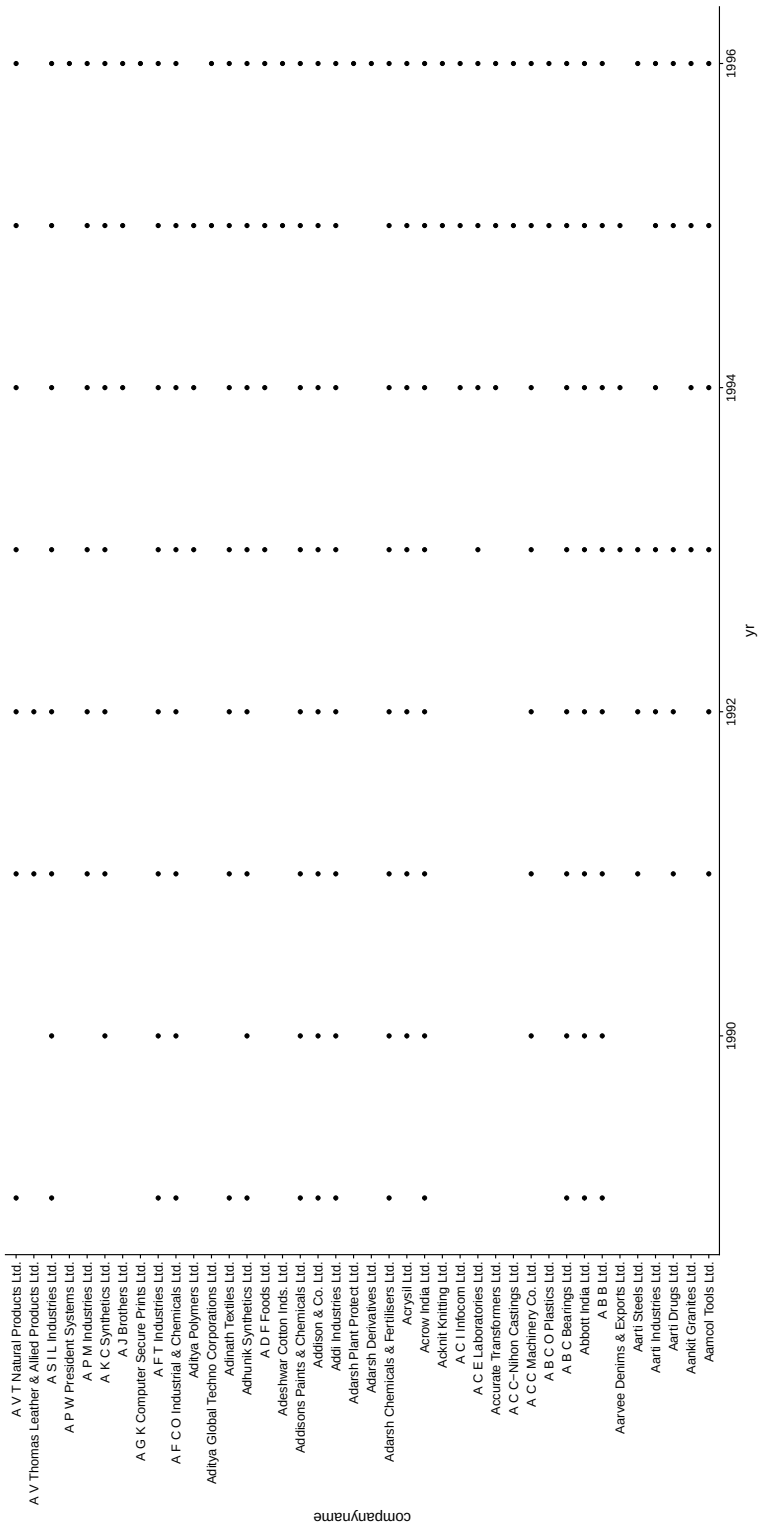

Figure D.2: Data patterns for a few firms.

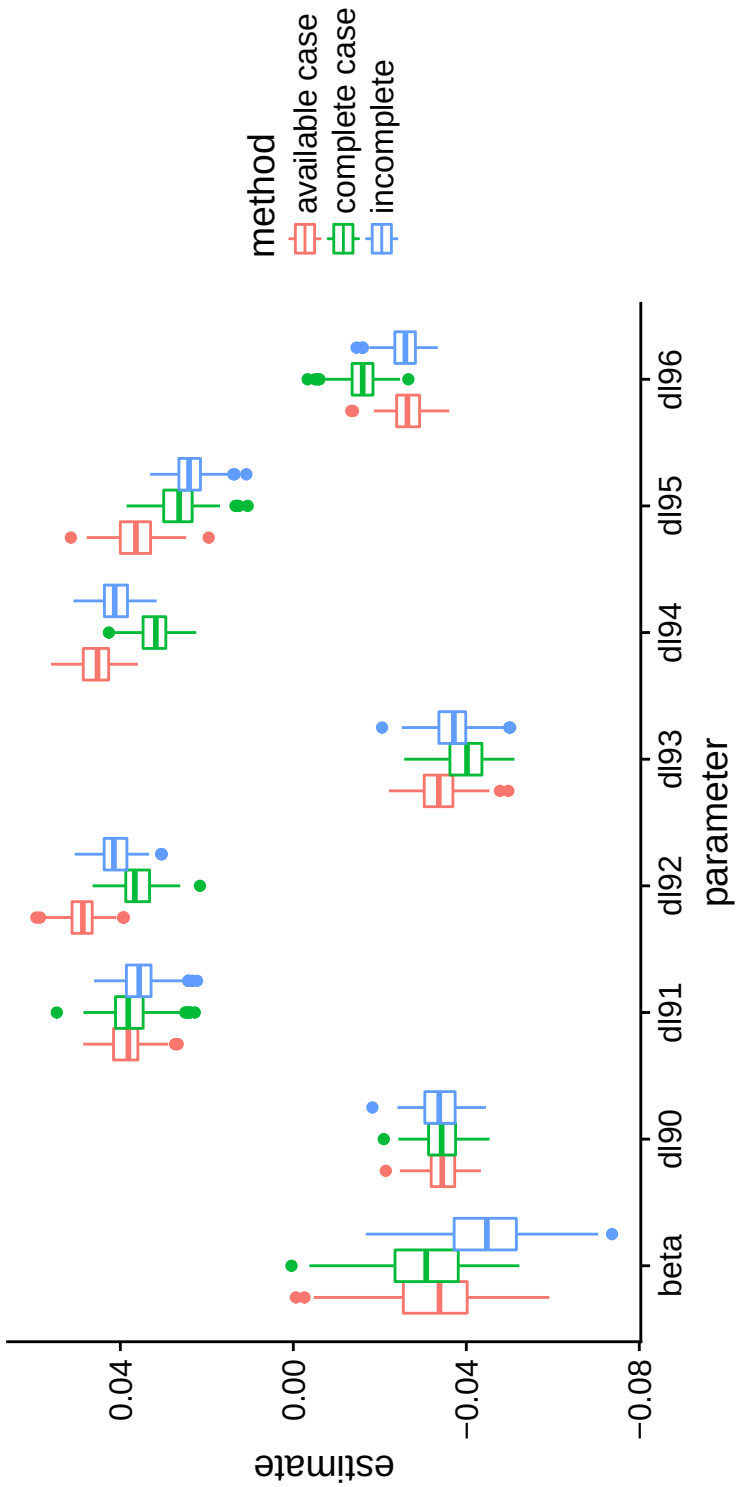

Figure D.3: Boxplots for estimates for different parameters and methods, based on bootstrap and the static model. "beta" is the regression coefficient, and the other parameters are difference in time dummies, i.e. "dl90" is the time dummy in 1990 minus the time dummy in 1989, etc.