EFFICIENT ESTIMATION WITH MISSING VALUES IN CROSS SECTION AND PANEL DATA

By

Bhavna Rai

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2021

ABSTRACT

EFFICIENT ESTIMATION WITH MISSING VALUES IN CROSS SECTION AND PANEL DATA

By

Bhavna Rai

Chapter 1: Efficient Estimation with Missing Data and Endogeneity

I study the problem of missing values in both the outcome and the covariates in linear models with endogenous covariates. I propose an estimator that improves efficiency relative to a Two Stage Least Squares (2SLS) based only on the complete cases. My framework also unifies the literature on missing data and combining data sets, and includes the "Two-Sample 2SLS" as a special case. The method is an extension of Abrevaya and Donald (2017), who provide methods of improving efficiency over complete cases estimators in linear models with cross-section data and missing covariates. I also provide guidance on dealing with missing values in the instruments and in commonly used nonlinear functions of the endogenous covariates, likes squares and interactions, without introducing inconsistency in the estimates.

Chapter 2: Imputing Missing Covariate Values in Nonlinear Models

I study the problem of missing covariate values in nonlinear models with continuous or discrete covariates. In order to use the information in the incomplete cases, I propose an inverse probability weighted one-step imputation estimator that provides gains in efficiency relative to the complete cases estimator using a reduced form for the outcome in terms of the always-observed covariates. Unlike the two-step imputation and dummy variable methods commonly used in empirical work, my estimator is consistent for a wide class of nonlinear models. It relies only on the commonly used "missing at random" assumption, and provides a specification test for the resulting restrictions. I show how the results apply to nonlinear models for fractional and nonnegative responses.

Chapter 3: Efficient Estimation of Linear Panel Data Models with Missing Covariates

We study the problem of missing covariates in the context of linear, unobserved effects panel data models. In order to use information on incomplete cases, we propose generalized method of moments (GMM) estimation. By using information on the incomplete cases from all time periods, the proposed estimators provide gains in efficiency relative to the fixed effects (and Mundlak) estimator that use only the complete cases. The method is an extension of Abrevaya and Donald (2017), who consider a linear model with cross-sectional data and incorporate the linear imputation method in the set of moment conditions to obtain gains in efficiency. Our first proposed estimator uses the assumption of strict exogeneity of the covariates as well as the selection, while allowing the selection to be correlated with the observed covariates and unobserved heterogeneity in both the outcome equation and the imputation equation. We also consider the case in which the covariates are only sequentially exogenous and propose an estimator based on the method of forward orthogonal deviations introduced by Arellano and Bover (1995). Our framework suggests a simple test for whether selection is correlated with unobserved shocks, both contemporaneous and those in other time periods.

ACKNOWLEDGEMENTS

My sincere gratitude to my adviser Jeffrey Wooldridge not only for lending his expertise to my dissertation but also for his patience and motivation throughout my Ph.D. I would also like to thank my committee members Peter Schmidt, Todd Elder and Vincenzina Caputo for their insightful comments and discussions.

My deep gratitude to my parents and brother for supporting me both materially and spiritually throughout the program. I am also thankful to my friends and fellow graduate students Katie Bollman, Marissa Eckrote, Pallavi Pal, and Ruonan Xu for their constant support and all the fun times.

The financial support received from the Department of Economics, the Graduate School, and the College of Social Science has been instrumental in completion of this work. Finally, the administrative support received from Lori Jean Nichols and Jay Feight has greatly facilitated navigating the program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

EFFICIENT ESTIMATION WITH MISSING DATA AND ENDOGENEITY

1.1 Introduction

The problem of missing data is highly prevalent in empirical research. While there is a vast literature on methods to deal with missing data, the issue of endogeneity of the covariates with missing values has not been explicitly addressed in the majority of it.[1]

In linear models with endogenous covariates and missing values in either the outcome or the endogenous covariates, a frequently used method is a 2SLS that only uses the “complete cases" - the observations for which all the variables are observed.[2] While consistent under commonly used assumptions, this method can lead to a substantial loss of efficiency due to discarding the information in the incomplete cases. Recent literature has considered the case of missingness only in the endogenous covariates and has suggested some methods that make use of these incomplete cases. The first set of methods is based on “imputation". For instance, McDonough & Millimet (2017) discuss an estimator which replaces the missing covariate values with fitted values from a first stage regression of the endogenous covariate on the instruments. A more efficient estimator is suggested by Abrevaya & Donald (2011), who use the incomplete cases via a reduced form for the outcome in terms of the instruments.

The first contribution of this paper is to extend the framework of Abrevaya & Donald (2011) to allow for missingness in both the outcome and the endogenous covariates. I show that it is possible to obtain strict gains in efficiency for all coefficients relative to the complete cases 2SLS.

My framework also unifies the literature on missing data and that on combining data sets with missing variables. Empirical researchers sometimes have two distinct data sets, one of which contains only the outcome and the instruments, and the other contains only the endogenous

[1] For a comprehensive discussion of methods used to deal with missing data, see Schafer & Graham (2002).
[2] Wooldridge (2010), Section 17.2.1.

covariates and the instruments. A commonly used estimator that combines the two is the "Two-Sample 2SLS" (henceforth TS2SLS).[3] I relax assumptions traditionally used by this estimator and also provide a framework for combining more than two data sets with more general patterns of missing variables.

A second method that makes use of the incomplete cases is the so-called "dummy variable method", which replaces the missing covariate values with zeros and includes an indicator for missingness as an additional covariate in the model. When the covariates are exogenous, Jones (1996) shows that this method produces inconsistent estimates unless some zero restrictions are imposed in the population. I show that this inconsistency carries over to the case of endogenous covariates.

One can also encounter missing values in the instruments, in which case interest lies in continuing to use the observations with missing instruments instead of discarding them. Mogstad & Wiswall (2012) discuss an estimator that imputes missing instrument values. This is a two-step estimator that in the first step replaces the missing instrument values with predicted values from a regression of the instrument on the always-observed exogenous covariates, and in the second step estimates the main model using a 2SLS with both the actual and imputed instrument values. They show that the resulting estimator for the coefficient on the endogenous covariate is numerically equivalent to a complete cases 2SLS. A second contribution of this paper is to propose an imputation estimator for the instruments that can achieve strict gains in efficiency over the complete cases 2SLS for all coefficients.[4] This estimator includes as a special case the estimator suggested by Abrevaya & Donald (2017) in the case where the covariates are exogenous.

Finally, I show how to impute commonly used nonlinear functions of the endogenous covariates like squares and interactions. I show that two-step procedures which in the first step replace the missing values of the nonlinear functions of the covariates with the same nonlinear functions of

[3]TS2SLS was first introduced by Klevmarken (1982), and more recently used by Angrist & Krueger (1995). Inoue & Solon (2010) show that the TS2SLS is more efficient than the related Two-Sample IV estimator. Inoue & Solon (2005) consider GMM estimation with arbitrary heteroskedasticity and stratification. Pacini & Windmeijer (2016) obtain robust standard errors for the traditional TS2SLS with arbitrary heteroskedasticity.

[4]Abrevaya & Donald (2011) also propose an estimator for the case of missing instruments. My estimator is based on different moment conditions and is no less efficient than theirs.

the imputed values generally produce inconsistent estimates. A third contribution of this paper is to propose a consistent imputation estimator in this context that improves upon the efficiency of complete cases 2SLS.

The rest of the paper is organized as follows. Section 1.2 presents the population model of interest and associated assumptions. Section 1.3 describes the missing data scheme and the assumptions on the missingness mechanism for the case of missingness in outcome and endogenous covariates. Section 1.4 describes the resulting moment conditions and the asymptotic distribution of the proposed GMM estimator. Section 1.5 discusses four related estimators: the complete cases 2SLS, the TS2SLS, the imputation estimator, and the dummy variable estimator. Section 1.6 discusses the case of missingness in the instruments. Section 1.7 discusses the case of nonlinearity in the covariates. Section 1.8 presents results from Monte Carlo simulations comparing the proposed estimator with related estimators. Section 1.9 presents an empirical application to the effect of physician's advice on individuals' calorie consumption. Section 1.10 concludes. The Appendices include the proofs and tables.

1.2 The population model and assumptions

Consider the standard linear regression model:

$$y = x_1\beta_1 + x_2\beta_2 + u \equiv x\beta + u, \tag{1.2.1}$$

where $x = (x_1, x_2)$ is the $1 \times (p + k)$ vector of covariates. x_1 is a $1 \times p$ vector of potentially endogenous covariates, while x_2 is a $1 \times k$ vector of exogenous covariates (including the constant). That is,

$$\mathbb{E}(x_2'u) = 0, \tag{1.2.2}$$

and we allow for $\mathbb{E}(x_1'u) \neq 0$. We are interested in estimating $\beta = (\beta_1', \beta_2')'$, where β_1 and β_2 are $p \times 1$ and $k \times 1$ respectively. As is well known, OLS is inconsistent for β under (1.2.2). Suppose we have a set of instruments $z = (z_1, x_2)$, where z_1 is a $1 \times q$ $(q \geq p)$ vector of excluded instruments,

such that

$$\mathbb{E}(z'u) = 0. \tag{1.2.3}$$

The first stage is given by the linear projection

$$x = z_1\Pi_1 + x_2\Pi_2 + r \equiv z\Pi + r, \tag{1.2.4}$$

where Π is the $(q+k)\times(p+k)$ matrix of all the first stage coefficients, and Π_1 and Π_2 are $q\times(p+k)$ and $k \times (p+k)$ matrices of coefficients on z_1 and x_2 respectively. By definition,

$$\mathbb{E}(z'r) = 0, \tag{1.2.5}$$

and by assumption $\Pi \neq 0$.

Then given a random sample and a rank condition, we can use 2SLS to consistently estimate β. Note that the errors u and r are assumed only to satisfy a zero correlation with the instruments in (1.2.3) and (1.2.5), and no other assumptions such as homoskedasticity or zero conditional mean have been imposed on them.

Now, using (1.2.1) and (1.2.4), we get a reduced form for y given by

$$y = z\Pi\beta + v, \quad v \equiv r\beta + u \tag{1.2.6}$$

and using (1.2.3) and (1.2.5), we have

$$\mathbb{E}(z'v) = 0. \tag{1.2.7}$$

Under the missing data scheme described in the next section, equation (1.2.6) allows us to use the incomplete cases for estimating β. When there is no missing data, the information in this equation is redundant given equations (1.2.1)-(1.2.5).

1.3 The missing data scheme

I characterize the potential missingness of the data using selection indicators. For any random draw (x_i, y_i, z_i) from the population, we also draw the selection indicators (s_{1i}, s_{2i}) defined as follows:

$$s_{1i} = \begin{cases} 1 & \text{if } y_i \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

$$s_{2i} = \begin{cases} 1 & \text{if } x_i \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

Two things should be noted. First, I am assuming that z_i is always observed. Since $z_i = (z_{1i}, x_{2i})$, I am allowing for missingness only in the endogenous covariates x_{1i}.[5] Second, I am assuming that either all or none of the elements in x_{1i} are observed. Then our "data" consists of $\{(y_i, x_i, z_i, s_{1i}, s_{2i}) : i = 1, ..., n\}$.

Because identification is properly studied in the population, let s_1 and s_2 denote random variables with distributions of s_{1i} and s_{2i} respectively for all i. In other words, (y, x, z, s_1, s_2) now denotes the population.

This framework allows for several kinds of missing data patterns that arise frequently in practice. Figure 1 shows some of these cases. First, it allows for the case where we have a single sample in which y is missing for certain observations, x is missing for certain other observations, and for the rest of the observations all y, x and z observed (Figure 1.1). Another case is where only x is missing for certain observations (Figure 1.2).[6] In both of these cases, using only the complete cases may lead to a substantial loss of information. A third case is where y and x are missing for disjoint observations such that there are no complete cases (Figure 1.3). Such a sample is typically obtained by combining two samples such that only y and z are observed in one sample and only x and z in the other. The most commonly used estimator in this case is the TS2SLS, which is a special case of the estimator I propose in the next section.

To determine the properties of any estimation procedure using selected samples, we need to know how s_1 and s_2 are related to (y, x, z). I place the following assumptions on the missingness indicators.

[5] I discuss the case of missingness in exogenous covariates in Section 1.6.
[6] This case has briefly been considered in Abrevaya & Donald (2011).

Assumption 1.3.1: $(i)\,\mathbb{E}(s_1 s_2 z'u) = 0 \quad (ii)\,\mathbb{E}(s_1 s_2 z'r) = 0 \quad (iii)\,\mathbb{E}(s_1 z'u) = 0 \quad (iv)\,\mathbb{E}(s_1 z'r) = 0 \quad (v)\,\mathbb{E}(s_2 z'r) = 0$

This assumption essentially implies that the orthogonality assumptions on the errors given in (1.2.3), (1.2.5) and (1.2.7) hold in the selected sub-populations as well. For instance, the first part of this assumption, which is the weakest possible assumption required for the consistency a 2SLS based only on the complete cases, can be written as

$$\mathbb{E}(s_1 s_2 z'u) = \mathbb{E}[\mathbb{E}(s_1 s_2 z'u)|s_1 s_2] = P(s_1 s_2 = 1)\,\mathbb{E}(z'u|s_1 s_2 = 1) = 0, \tag{1.3.1}$$

where the first equality holds by the law of iterated expectations (LIE). If we assume that $P(s_1 s_2 = 1)$ is strictly positive, then we need the population orthogonality condition $\mathbb{E}(z'u) = 0$ to hold in the sub-population where $s_1 = s_2 = 1$ for this assumption to be true. The other parts of this assumption impose similar restrictions on the errors in (1.2.1) and (1.2.4) for different sub-populations.

Sufficient for Assumption 1.3.1 to hold is that $(s_1, s_2) \perp\!\!\!\perp (z, u, r)$, for which a sufficient condition is that $(s_1, s_2) \perp\!\!\!\perp (x, y, z)$. That is, selection is independent of everything else in the model. This is generally known as "missing completely at random" (MCAR) in the missing data literature.[7] For instance, consider the first part of Assumption 1.3.1.

$$\mathbb{E}(s_1 s_2 z'u) = \mathbb{E}(s_1 s_2)\,\mathbb{E}(z'u) = 0 \tag{1.3.2}$$

and similarly for the other parts.

Assumption 1.3.1 also holds if we have correctly specified conditional means and selection is independent of errors in both the model of interest and the first stage conditional on the instruments. That is, strengthening the exogeneity conditions in (1.2.3) and (1.2.5) to $\mathbb{E}(u|z) = 0$ and $\mathbb{E}(r|z) = 0$ respectively and assuming $(s_1, s_2) \perp\!\!\!\perp (u, r)|z$ is sufficient. Again, consider the first part of Assumption 1.3.1.

$$\mathbb{E}(s_1 s_2 z'u) = \mathbb{E}[\mathbb{E}(s_1 s_2 z'u|z, s_1 s_2)] = \mathbb{E}[s_1 s_2 z'\,\mathbb{E}(u|z, s_1 s_2)] = \mathbb{E}[s_1 s_2 z'\,\mathbb{E}(u|z)] = 0 \tag{1.3.3}$$

[7] We do not require s_1 and s_2 to be independent of each other for Assumption 1.3.1 to hold.

where the third equality holds because of the conditional independence and the last one holds because of the zero conditional mean of the errors. An important special case is when selection is a deterministic function of z. But it can also depend on other unobservable random variables under certain conditions. For instance, we can let

$$s_1 s_2 = f(z, w), \tag{1.3.4}$$

where w is an unobserved random variable. Then $\mathbb{E}(s_1 s_2 z' u)$ holds if $\mathbb{E}(u|z) = 0$ and $w \perp\!\!\!\perp (z, u)$, as

$$\mathbb{E}(s_1 s_2 z' u) = \mathbb{E}[\mathbb{E}(s_1 s_2 z' u|z, w)] = \mathbb{E}[s_1 s_2 z' \, \mathbb{E}(u|z, w)] = \mathbb{E}[s_1 s_2 z' \, \mathbb{E}(u|z)] = 0. \tag{1.3.5}$$

What Assumption 1.3.1 rules out is (s_1, s_2) depending on the errors u and r. That is, selection cannot depend on the idiosyncratic errors in either y or x. Whether or not this holds in an empirical application should be carefully considered by the researcher.

1.4 Moment conditions and GMM estimation

Using equations (1.2.1)-(1.2.7) along with Assumption 1.3.1, I define the vector of moment functions as follows.

$$g(\beta, \Pi) = \begin{bmatrix} s_1 s_2 z'(y - x\beta) \\ s_1 s_2 z' \otimes (x - z\Pi)' \\ (1 - s_1) s_2 z' \otimes (x - z\Pi)' \\ s_1 (1 - s_2) z'(y - z\Pi\beta) \end{bmatrix} \equiv \begin{bmatrix} g_1(\beta, \Pi) \\ g_2(\beta, \Pi) \\ g_3(\beta, \Pi) \\ g_4(\beta, \Pi) \end{bmatrix} \tag{1.4.1}$$

where I suppress (y, x, z, s_1, s_2) from $g(.)$ for notational convenience. In the vector $g(.)$, $g_1(.)$ and $g_2(.)$ use the information contained in the complete cases. $g_3(.)$ uses the observations for which x is observed but y is not, while $g_4(.)$ uses the observations for which y is observed but x is not.[8] Then, the following result holds for $g(.)$.

Lemma 1.4.1. *Under Assumption 1.3.1,* $\mathbb{E}[g(\beta, \Pi)] = 0$.

[8]Note that equations (1.2.1)-(1.2.7) and our missing data scheme suggest 5 different moment functions: $g_1(.)$-$g_4(.)$ along with $g_5(.) = s_1 s_2 z'(y - z\Pi\beta)$. However, since $g_5(.)$ is a linear combination of $g_1(.)$ and $g_2(.)$, it is redundant given $g_1(.)$-$g_4(.)$ and hence I exclude it from the set of relevant moment functions.

This gives us a vector of $2(q+k)(1+p+k)$ moment conditions satisfied by the population parameter values (β, Π). We have $(p+k)(1+q+k)$ parameters to estimate, giving us $2(q+k)+(p+k)(q+k-1)$ overidentifying restrictions.

Let $\bar{g}(\beta, \Pi) = n^{-1} \sum_{i=1}^{n} g(y_i, x_i, z_i, s_{1i}, s_{2i}, \beta, \Pi)$, Ω be a square matrix of order $2(q+k)(1+p+k)$ that is nonrandom, symmetric, and positive definite, and $\hat{\Omega}$ be a first step consistent estimate of Ω. Then, the standard two-step GMM estimator minimizes the objective function

$$\bar{g}(\beta, \Pi)' \ \hat{\Omega} \ \bar{g}(\beta, \Pi). \tag{1.4.2}$$

The variance-covariance matrix of the moment functions is given by

$$C \equiv \mathbb{E}[g(\beta, \Pi) \ g(\beta, \Pi)'] = \begin{bmatrix} C_{11} & C_{12} & 0 & 0 \\ C_{12}' & C_{22} & 0 & 0 \\ 0 & 0 & C_{33} & 0 \\ 0 & 0 & 0 & C_{44} \end{bmatrix}$$

where

$$C_{11} = \mathbb{E}(s_1 s_2 u^2 z' z) \quad C_{12} = \mathbb{E}(s_1 s_2 z' u z \otimes r) \quad C_{22} = \mathbb{E}(s_1 s_2 z' \otimes r' z \otimes r)$$

$$C_{33} = \mathbb{E}[(1-s_1) s_2 z' \otimes r' z \otimes r] \quad C_{44} = \mathbb{E}[s_1 (1-s_2) v^2 z' z] \tag{1.4.3}$$

and $g(.)$ is evaluated at the true value of the parameters. The optimal weight matrix is given by the inverse of C. Let $\hat{C}$ be a consistent estimate of C which can be obtained by replacing the expectation by sample average in the definition of C above and replacing u, r and v by consistent estimates obtained using, for instance, GMM estimators that use $g_1(.)$ only, $g_2(.)$ and $g_3(.)$ only, and $g_4(.)$ only respectively. Then, the optimal GMM estimator is defined as the following.

Definition 1.4.1. *Call the estimators of β and Π that minimize (1.4.2) with the optimal weight matrix $\hat{\Omega} = \hat{C}^{-1}$, $\hat{\beta}$ and $\hat{\Pi}$ respectively.*

Further, define the $(k+q)(2+k+p) \times (k+p)(1+k+q)$ matrix of expected derivatives of $g(.)$

w.r.t. $(\beta', vec(\Pi)')'$

$$D \equiv \mathbb{E}[\nabla g(\beta, \Pi)] = \begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \\ 0 & D_{32} \\ D_{41} & D_{42} \end{bmatrix}$$

where

$$D_{11} = -\mathbb{E}(s_1 s_2 z' x) \quad D_{22} = -\mathbb{E}[s_1 s_2 (z'z \otimes c_1, \ldots, z'z \otimes c_{(p+k)})]$$

$$D_{32} = -\mathbb{E}[(1 - s_1) s_2 (z'z \otimes c_1, \ldots, z'z \otimes c_{(p+k)})] \quad D_{41} = -\mathbb{E}[s_1 (1 - s_2) z'z\Pi]$$

$$D_{42} = -\mathbb{E}[s_1 (1 - s_2) \beta' \otimes z'z], \tag{1.4.4}$$

where c_m is a $(p + k) \times 1$ vector with one in the m^{th} row and all other rows being zero, $m = 1, \ldots, (p + k)$. I impose the following rank condition on D for identification of β and Π.

Assumption 1.3.2: $rank(D) = (p + k)(1 + q + k)$

If $P(s_1 s_2 = 1) > 0$, then sufficient for this assumption to hold is that $\mathbb{E}(z'x|s_1 s_2 = 1)$ and $\mathbb{E}(z'z|s_1 s_2 = 1)$ have full column ranks $(p + k)$ and $(p + k)(q + k)$ respectively. In this case, $\mathbb{E}[g_1(\beta)] = 0$ identifies β and $\mathbb{E}[g_2(\Pi)] = 0$ identifies Π. If $P(s_1 s_2 = 1) = 0$, for instance in the TS2SLS case, then sufficient is that $\mathbb{E}(z'z|s_2 = 1)$ and $\mathbb{E}(z'x|s_1 = 1)$ have full column ranks $(p + k)(q + k)$ and $(p + k)$ respectively. In this case, $\mathbb{E}[g_3(\Pi)] = 0$ identifies Π and $\mathbb{E}[g_4(\beta)] = 0$ identifies β since for the purpose of identification, we can treat Π as known.

Then, we have the following result using Hansen (1982).

Theorem 1.4.1 *Under standard regularity conditions and Assumptions 1.3.1 and 1.3.2,*

$$\sqrt{n}[\left((\hat{\beta}', vec(\hat{\Pi})')' - (\beta', vec(\Pi)')'\right] \xrightarrow{d} N(0, (D'C^{-1}D)^{-1})$$

and

$$n\, \bar{g}(\hat{\beta}, \hat{\Pi})'\, \hat{C}^{-1}\, \bar{g}(\hat{\beta}, \hat{\Pi}) \xrightarrow{d} \chi^2_{2(q+k)+(p+k)(q+k-1)}.$$

This statistic can be used for the standard test of overidentifying restrictions. Note that this statistic is just the GMM objective function in (1.4.2) evaluated at the efficient values of the parameters and is distributed as chi-squared with degrees of freedom equal to the number of overidentifying restrictions.

1.5 Comparison with related estimators

1.5.1 Complete cases estimator

The most common practice in the presence of missing data is to just use the complete cases for estimation; that is, only use the observations for which both y and x are observed. In the current framework, the first and the most commonly used estimator that uses only the complete cases is the standard 2SLS. This estimator uses only $g_1(.)$ in estimation as it requires $s_1 = s_2 = 1$, and uses a weight matrix that is optimal when u is homoskedastic.

Definition 1.5.1.1 *Call the estimator of β that minimizes (1.4.2), where $g(.)$ contains only $g_1(.)$ and $\hat{\Omega} = (n^{-1} \sum_{i=1}^{n} s_{1i} s_{2i} z_i' z_i)^{-1}$, the complete cases 2SLS (or $\hat{\beta}_{CC-2SLS}$).*

The weight matrix used by $\hat{\beta}_{CC-2SLS}$ is optimal if $\mathbb{E}(u^2|z, s_1, s_2) = \sigma^2$. When this assumption is violated, a more efficient complete cases estimator can be obtained by using optimal weighting.

Definition 1.5.1.2 *Call the estimator of β that minimizes (1.4.2), where $g(.)$ contains only $g_1(.)$ and $\hat{\Omega} = \hat{C}_{11}^{-1}$, the complete cases GMM (or $\hat{\beta}_{CC-GMM}$).*

This is the optimal GMM estimator based only on the complete cases. Its asymptotic variance is easily obtained using the standard GMM theory.

Lemma 1.5.1.1 *Under Assumption 1.3.1, the complete cases GMM has an asymptotic variance given by*

$$Avar\left(\sqrt{n}(\hat{\beta}_{CC-GMM} - \beta)\right) = (D_{11}' C_{11}^{-1} D_{11})^{-1}.$$

Comparing the asymptotic variances of $\hat{\beta}$ and $\hat{\beta}_{CC-GMM}$, the former is no less efficient than the latter because it uses the information contained in the incomplete cases, while the latter simply

ignores this information. The gain in efficiency follows from the fact that adding valid moment conditions decreases, or at least does not increase the asymptotic variance of a GMM estimator.[9]

Proposition 1.5.1.1 *Under Assumption 1.3.1,*

$$Avar\left(\sqrt{n}(\hat{\beta}_{CC-GMM} - \beta)\right) - Avar\left(\sqrt{n}(\hat{\beta} - \beta)\right) \text{ is positive semi-definite.}$$

Further, I break down the gains in efficiency by β_1 and β_2, the coefficients on the potentially missing endogenous covariates x_1 and the always observed exogenous covariates x_2 respectively. For algebraic convenience, I consider the case where both x_1 and z_1 are scalars.[10]

Proposition 1.5.1.2 *Let $p = q = 1$. Under Assumption 1.3.1,*

$$\textit{(i) } Avar\left(\sqrt{n}(\hat{\beta}_{1-CC-GMM} - \beta_1)\right) - Avar\left(\sqrt{n}(\hat{\beta}_1 - \beta_1)\right) = \begin{bmatrix} A_1' & B_1' \end{bmatrix} E \begin{bmatrix} A_1 \\ B_1 \end{bmatrix} \geq 0$$

$$\textit{(ii) } Avar\left(\sqrt{n}(\hat{\beta}_{2-CC-GMM} - \beta_2)\right) - Avar\left(\sqrt{n}(\hat{\beta}_2 - \beta_2)\right) = \begin{bmatrix} A_2' & B_2' \end{bmatrix} E \begin{bmatrix} A_2 \\ B_2 \end{bmatrix} \geq 0,$$

where $A_j = D_{32}D_{22}^{-1}C_{21}W_j$, $B_j = (D_{41}D_{11}^{-1}C_{11} + D_{42}D_{22}^{-1}C_{21})W_j$, $j = 1, 2$ and W_1, W_2 and E are matrices defined in the appendix, E being a positive definite matrix.

Starting with the first part of Proposition 1.5.1.2, since E is positive definite, the difference is 0 if and only if $A_1 = B_1 = 0$. The corresponding difference for β_2 is 0 if and only if $A_2 = B_2 = 0$. Since neither A_j nor B_j are necessarily 0 under the assumptions made so far, it is possible to obtain strict gains in efficiency for both β_1 and β_2.

Finally, when there is no missingness, the moment conditions in (1.4.1) just give us the standard 2SLS estimator. Because s_1 and s_2 are always 1, $g_3(.)$ and $g_4(.)$ are always zero and we are left with $g_1(.)$ and $g_2(.)$. Since $g_2(.)$ adds equal number of additional parameters as the number of additional moment functions to $g_1(.)$, the GMM estimator of β from $g_1(.)$ will be the same as that from $g_1(.)$ and $g_2(.)$.[11] Thus, estimation is based only on $g_1(.) = z'(y - x\beta)$, which is the usual

[9] Wooldridge (2010), Section 8.6.
[10] The proof for this proposition is an extension of the proof of Proposition 2 in Abrevaya & Donald (2017).
[11] Ahu & Schmidt (1995), Theorem 1.

moment function used by 2SLS along with a weight matrix constructed under homoskedasticity of u.

Proposition 1.5.1.3 *If* $P(s_1 = 1) = P(s_2 = 1) = 1$ *and* $\hat{\Omega}_{11} = (n^{-1}\sum_{i=1}^{n} z_i' z_i)^{-1}$, *where* $\hat{\Omega}_{11}$ *is the upper left* $(q + k) \times (q + k)$ *block of* $\hat{\Omega}$, $\hat{\beta}$ *equals the standard 2SLS estimator.*

1.5.2 Estimators combining different data sets

A special case of missingness occurs when data is combined from more than one data sets, one or more of which do not contain either y or some or all elements of x. For instance, the pattern of missingness in Figure 1.1 can result from combining three data sets, one of which contains all y, x and z, a second is missing y, and a third is missing x. In this case, one can just use the first data set to estimate β, but the second and third can be used to achieve efficiency gains using the framework in Section 1.4. One does have to be careful in making sure that Assumption 1.3.1 holds in order to ensure consistency. For instance, a sufficient condition would be that the different data sets being combined are just random samples of different variables from the same population.

There may also be cases where estimation using a single data set is not possible at all. A prominent example is when one data set contains only y and z, while the second contains only x and z. The most commonly used estimator in this case is the TS2SLS.[12] The TS2SLS is a sequential GMM estimator based only on $g_3(.)$ and $g_4(.)$ since $s_1 s_2 = 0$ in this case.

Definition 1.5.2.1 *Call the estimator of* β *obtained by the following sequential procedure the two-sample two stage least squares (or* $\hat{\beta}_{TS2SLS}$*).*
Step 1: Obtain $\breve{\Pi}$ *by minimizing (1.4.2), where* $g(.)$ *contains only* $g_3(.)$ *and* $\hat{\Omega} = I$.
Step 2: Estimate β *by minimizing (1.4.2), where* $g(.)$ *contains only* $g_4(.)$, $\hat{\Omega} = (n^{-1}\sum_{i=1}^{n} s_{1i} z_i' z_i)^{-1}$, *and* $\Pi = \breve{\Pi}$ *is treated as given.*

There are two differences between $\hat{\beta}$ and $\hat{\beta}_{TS2SLS}$. First is in terms of the assumptions made by the two estimators. The traditional analysis of $\hat{\beta}_{TS2SLS}$ or the related two-sample IV (TSIV)

[12]This estimator is discussed in detail in a GMM context by Inoue & Solon (2010).

estimator either assumes MCAR (Angrist & Krueger, 1995), or imposes restrictions on z and x that essentially follow from assuming MCAR. For instance, Angrist & Krueger (1992), in using the TSIV estimator, assume that $\mathbb{E}(z'x|s_1 = 1) = \mathbb{E}(z'x|s_2 = 1)$. Inoue & Solon (2010) make the same assumption, along with $\mathbb{E}(z'z|s_1 = 1) = \mathbb{E}(z'z|s_2 = 1)$ and that the fourth moments of z conditional on s_1 and s_2 are equal. The framework presented in this paper allows for relaxation of these restrictive assumptions. By allowing s_1 and s_2 to depend on z, I allow for the distribution of z (and x) to be different conditional on s_1 and s_2. However, the coefficient in the linear projection of x on z (that is, Π) remains the same conditional on s_1 and s_2 under Assumption 1.3.1.[13]

The second difference is in terms of the weight matrix used. Note that the weight matrix used in *Step 2* of Definition 1.5.2.1 is the sample counterpart of C_{44}^{-1} (divided by the variance of v, which is just a constant), when v satisfies the following assumption.

$$\mathbb{E}(v^2|z, s_1) = \sigma_v^2. \tag{1.5.1}$$

That is, the variance of v is constant conditional on both the instruments z *and* s_1. If this assumption is not true, then $\hat{\beta}_{TS2SLS}$ uses a sub-optimal weight matrix in *Step 2* and efficiency gains are possible by using the optimal weight matrix.[14] Let

$$\hat{D}_{32} = -\frac{1}{n}\sum_i [s_{2i}(z_i'z_i \otimes c_1, \ldots, z_i'z_i \otimes c_{(p+k)})] \quad \hat{D}_{42} = -\frac{1}{n}\sum_i (s_{1i}\hat{\beta}' \otimes z_i'z_i) \tag{1.5.2}$$

be consistent estimates of D_{32} and D_{42} respectively, and $\hat{C}_{44}$ and $\hat{C}_{33}$ are as defined in Section 1.4, where consistent estimates of β and Π can now be obtained using $\hat{\beta}_{TS2SLS}$.

Definition 1.5.2.2 *Call the estimator of β obtained by replacing*

$$\hat{\Omega} = (\hat{C}_{44} + \hat{D}_{32}(\hat{D}_{42}'\hat{C}_{33}^{-1}\hat{D}_{42})^{-1}\hat{D}_{32}')^{-1}$$

[13]Note that for $\hat{\beta}_{TS2SLS}$ to be consistent, we only need Π to be the same conditional on s_1 and s_2, and not the individual moments involved in the calculation of Π.

[14]Two things should be noted here:

- Because $s_1 s_2 = 0$, $g_3(.) = s_2 z' \otimes (x - z\Pi)'$ and $g_4(.) = s_1 z'(y - z\Pi\beta)$.
- Since $\mathbb{E}[g_2(.)] = 0$ is an exactly identified set of moment conditions, the weight matrix does not matter for estimation in Step 1 of Definition 1.5.2.1.

in step 2 of the procedure in Definition 1.5.2.1, the Optimal TS2SLS estimator (or $\hat{\beta}_{TS2SLS-O}$).

This is the optimal sequential GMM estimator under the assumptions made so far and its asymptotic variance is given in the following result.

Proposition 1.5.2.1 *Under Assumption 1.3.1, $\hat{\beta}_{TS2SLS-O}$ is the optimal sequential GMM estimator of β, and has an asymptotic variance given by*

$$Avar\left(\sqrt{n}(\hat{\beta}_{TS2SLS-O} - \beta)\right) = \{D'_{41}[C_{44} + D_{42}(D_{32}^{-1}C_{33}D_{42}^{'-1})D'_{32}]^{-1}D_{41}\}^{-1}.$$

Since $\hat{\beta}_{TS2SLS}$ uses a sub-optimal weight matrix as opposed to $\hat{\beta}_{TS2SLS-O}$, the latter will be no less efficient than the former.

Proposition 1.5.2.2 *Under Assumption 1.3.1,*

$$Avar\left(\sqrt{n}(\hat{\beta}_{TS2SLS-O} - \beta)\right) - Avar(\sqrt{n}(\hat{\beta}_{TS2SLS} - \beta)) \text{ is positive semi-definite.}$$

The proposed estimator $\hat{\beta}$ is then equally efficient as $\hat{\beta}_{TS2SLS-O}$.

Proposition 1.5.2.3 *Under Assumption 1.3.1,*

$$Avar\left(\sqrt{n}(\hat{\beta}_{TS2SLS-O} - \beta)\right) = Avar(\sqrt{n}(\hat{\beta} - \beta)).$$

From Propositions 1.5.2.2 and 1.5.2.3, we can conclude that $\hat{\beta}$ is no less efficient than $\hat{\beta}_{TS2SLS}$.

Proposition 1.5.2.4 *Under Assumption 1.3.1,*

$$Avar\left(\sqrt{n}(\hat{\beta}_{TS2SLS} - \beta)\right) - Avar(\sqrt{n}(\hat{\beta} - \beta)) \text{ is positive semi-definite.}$$

Inoue & Solon (2005) address the issues of optimal weighting using a joint GMM and allowing for conditional heteroskedasticity. Their framework however is more restrictive than necessary. First, they start with zero conditional means of the errors in (2.1) and (2.4), which rules out the important case when (1.2.1) and (1.2.4) are just linear projections and the data is MCAR. Second, they impose restrictions on the second and third moments of x and z, which this framework does not.

Finally, $\hat{\beta}$ is numerically equivalent to $\hat{\beta}_{TS2SLS}$ if β is exactly identified. This is because in case of exact identification, the efficiency due to using the optimal weight matrix is lost as the weight matrix does not matter for estimation.

Proposition 1.5.2.5 *If $p = q$ and Assumption 1.3.1 holds, $\hat{\beta} = \hat{\beta}_{TS2SLS}$. Therefore,*

$$Avar\big(\sqrt{n}(\hat{\beta}_{TS2SLS} - \beta)\big) = Avar\big(\sqrt{n}(\hat{\beta} - \beta)\big).$$

1.5.3 Sequential estimators

Consider the case where y is always observed—that is, $P(s_1 = 1) = 1$—and the only variables that contain missing values are x. Thus, $g_3(.) = 0$ and we are only left with $g_1(.)$, $g_2(.)$ and $g_4(.)$. For this case, McDonough & Millimet (2017) discuss a sequential estimator which is the counterpart of linear imputation in the case where x is exogenous in equation (1.2.1).

Definition 1.5.3.1 *Call the estimator of β obtained by the following procedure the imputation estimator (or $\hat{\beta}_{Imp}$).*
Step 1: Obtain $\hat{\Pi}$ by minimizing (1.4.2), where $g(.)$ contains only $g_2(.)$ and $\hat{\Omega} = I$.
Step 2: Estimate β by minimizing (1.4.2), where $g(.) = g_5(.) = z'\{y - [sx + (1 - s)z\hat{\Pi}]\beta\}$, $\hat{\Omega} = [n^{-1}\sum_{i=1}^{n} g_{5i}(.)g_{5i}(.)']^{-1}$ and $\hat{\Pi}$ is treated as given.

So in the first step, we estimate the first stage coefficients Π. We then replace the missing values of x with $z\hat{\Pi}$ and estimate β in the second step using 2SLS on the full sample and treating the fitted values of x as given. It is straightforward to show that this estimator is no more efficient than $\hat{\beta}$.

Consider the sequential estimator of β that first estimates Π using $g_2(.)$ and then estimates β using $g_1(.)$ and $g_4(.)$, where $g_4(.)$ uses the estimated Π from the first step.

Definition 1.5.3.2 *Call the estimator of β obtained by the following procedure the sequential estimator (or $\hat{\beta}_{Seq}$).*
Step 1: Same as Step 1 in Definition 1.5.3.1.

Step 2: Estimate β by minimizing (1.4.2), where

$$g(\beta, \hat{\Pi}) = \left(g_1(\beta)', \quad g_4(\beta, \hat{\Pi})'\right)', \ \ \hat{\Omega} = [n^{-1} \sum_{i=1}^{n} g_i(.) g_i(.)']^{-1}$$

and $\hat{\Pi}$ is treated as given.

By standard GMM theory, we know that $\hat{\beta}$ is no less efficient than $\hat{\beta}_{Seq}$, since it is a sequential estimator (as opposed to a joint estimator) based on the same moment conditions as $\hat{\beta}$.[15] Moreover, $g_5(.)$, which is the moment condition used in Step 2 of Definition 1.5.3.1 can be obtained by adding $g_1(\beta)$ and $g_4(\beta, \hat{\Pi})$, which are the moment conditions used in step 2 of Definition 1.5.3.2. Since $\hat{\beta}_{Seq}$ uses $g_5(.)$ *and* an additional moment condition, it is no less efficient than $\hat{\beta}_{Imp}$. Thus we can conclude that $\hat{\beta}$ is no less efficient than $\hat{\beta}_{Imp}$ and there is no reason to choose the latter over the former other than computational convenience.

1.5.4 Dummy variable method

A common method used to deal with missingness in x in the case where x is exogenous is the dummy variable method, which entails replacing the missing values of x with zeros and including an indicator for missingness as a covariate. As shown by Abrevaya & Donald (2017), this method is inconsistent unless some zero restrictions are imposed in the population. This method continues to be inconsistent in the current framework where x is endogenous.

Let $P(s_1 = 1) = 1$, that is, y is always observed. Also note that (1.2.4) implies

$$x_1 = z_1 \Pi_{11} + x_2 \Pi_{21} + r_1, \tag{1.5.3}$$

where Π_{11}, Π_{21} and r_1 constitute the first p columns of Π_1, Π_2 and r respectively.[16] Then (1.2.1) and (1.5.2) imply

$$y = [s_2 x_1 + (1 - s_2)(z_1 \Pi_{11} + x_2 \Pi_{21} + r_1)]\beta_1 + x_2 \beta_2 + u. \tag{1.5.4}$$

[15] Prokhorov & Schmidt (2009), Theorem 2.2, part 5.

[16] One can similarly write $x_2 = z_1 \Pi_{12} + x_2 \Pi_{22} + r_2$. However, it is clear that both Π_{12} and r_2 are identically 0 and Π_{22} is a $k \times k$ identity matrix.

Since x_2 contains the constant, write $x_2 = (1, x_{22})$ where x_{22} constitutes the last $(k-1)$ columns of x_2. Correspondingly, write $\Pi_{21} = (\Pi'_{211}, \Pi'_{212})'$, where Π_{211} is the first row of Π_{21} and Π_{212} constitutes the last $(k-1)$ rows of Π_{21}. Plugging this into (1.5.3) and re-arranging gives

$$y = s_2 x_1 \beta_1 + (1 - s_2)\big(z_1 \Pi_{11} + \Pi_{211} + x_{22}\Pi_{212} + r_1\big)\beta_1 + x_2\beta_2 + u. \tag{1.5.5}$$

The dummy variable method omits the covariates $(1 - s_2)z_1$ and $(1 - s_2)x_{22}$ from equation (1.5.4) and estimates using 2SLS the equation

$$y = s_2 x_1 \beta_1 + (1 - s_2)\Pi_{211} + x_2\beta_2 + e \tag{1.5.6}$$

using instruments $(s_2 z_1, 1 - s_2, x_2)$, where $e \equiv (1 - s_2)(z_1\Pi_{11} + x_{22}\Pi_{212})\beta_1 + r_1\beta_1 + u$. However, since each of these instruments is now correlated with the new error e, 2SLS will not yield consistent estimates in general unless we impose some zero restrictions in the population.

Proposition 1.5.4.1 *The 2SLS estimators of β from equation (1.5.5) using instruments $(s_2 z_1, 1 - s_2, x_2)$ are inconsistent unless (i) $\beta_1 = 0$ or (ii) $\Pi_{11} = \Pi_{212} = 0$.*

The first condition implies that x_1 is irrelevant in the model of interest (1.2.1), so the best solution is to drop it. The second implies that neither the excluded instruments z_1 nor the always observed covariates x_{22} help in explaining x_1, in which case any estimation method based on z_1 cannot be used at all.

1.6 Missing instruments

In Sections 1.2-1.5, I discussed the case where y and the endogenous elements of x (that is, x_1) contain missing values, while the instruments z are always observed. In this section, I consider the case where the excluded instruments z_1 contain missing values. This includes as a special case missingness in covariates when all the covariates are exogenous.

Starting with the population model in Section 1.2, I now additionally introduce a linear projection of the excluded instruments z_1 on the always observed exogenous covariates x_2.

$$z_1 = x_2\Gamma + e, \tag{1.6.1}$$

where by definition of a linear projection

$$\mathbb{E}(x_2' e) = 0. \tag{1.6.2}$$

As discussed in Section 1.5.4, (1.2.4) implies that

$$x_1 = z_1\Pi_{11} + x_2\Pi_{21} + r_1. \tag{1.6.3}$$

Plugging (1.6.1) into (1.6.3) gives us a first stage in terms of x_2 only.

$$x_1 = x_2(\Gamma\Pi_{11} + \Pi_{21}) + (e\Pi_{11} + r_1). \tag{1.6.4}$$

Plugging (1.6.4) into (1.2.1) gives us a reduced form for y in terms of x_2 only.

$$y = x_2(\Gamma\Pi_{11}\beta_1 + \Pi_{21}\beta_1 + \beta_2) + (e\Pi_{11}\beta_1 + r_1\beta_1 + u). \tag{1.6.5}$$

Now, for observation i, let

$$s_{3i} = \begin{cases} 1 & \text{if } z_{1i} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

I impose the following assumptions on the missingness mechanism, which can be interpreted in a similar way as Assumption 1.3.1.

Assumption 1.6.1: $(i)\,\mathbb{E}(s_3 z' u) = 0 \quad (ii)\,\mathbb{E}(s_3 z' r) = 0 \quad (iii)\,\mathbb{E}(s_3 x_2' e) = 0.$

This gives us the following moment functions.

$$h(\beta, \Pi, \gamma) = \begin{bmatrix} s_3 z'(y - x\beta) \\ s_3 z' \otimes (x_1 - z_1\Pi_{11} - x_2\Pi_{21})' \\ s_3 x_2' \otimes (z_1 - x_2\Gamma)' \\ (1 - s_3)x_2' \otimes [x_1 - x_2(\Gamma\Pi_{11} + \Pi_{21})]' \\ (1 - s_3)x_2'[y - x_2(\Gamma\Pi_{11}\beta_1 + \Pi_{21}\beta_1 + \beta_2)] \end{bmatrix} = \begin{bmatrix} h_1(\beta, \Pi, \gamma) \\ h_2(\beta, \Pi, \gamma) \\ h_3(\beta, \Pi, \gamma) \\ h_4(\beta, \Pi, \gamma) \\ h_5(\beta, \Pi, \gamma) \end{bmatrix} \tag{1.6.6}$$

This vector of moment functions is basically using the original model of interest and first stage when z_1 is observed $\big(h_1(.)$ and $h_2(.)\big)$. When z_1 is missing, it uses the reduced forms for x_1 and y

in terms of x_2 in order to use the incomplete cases $\big(h_4(.)$ and $h_5(.)\big)$. $h_3(.)$ simply identifies the parameters in the linear projection of z_1 on x_2. Then under Assumption 1.6.1, the following result holds for $h(.)$.

Lemma 1.6.1. *Under Assumption 1.6.1,* $\mathbb{E}[h(\beta, \Pi, \gamma)] = 0$.

This gives us a set of $2k(1+p) + q(1+p+k)$ moment conditions for $(p+k)(1+q+k) + kq$ parameters, giving us $k(1+p) + q - p$ overidentifying restrictions.[17] Then, let $\bar{h}(\beta, \Pi, \Gamma) = n^{-1} \sum_{i=1}^{n} h(y_i, x_i, z_i, s_{3i}, \beta, \Pi, \Gamma)$, Λ be a square matrix of order $2k(1+p) + q(1+p+k)$ that is nonrandom, symmetric, and positive definite, and $\tilde{\Lambda}$ be a first step consistent estimate of Λ. Then, $\big(\tilde{\beta}', vec(\tilde{\Pi})', vec(\tilde{\Gamma})'\big)$ is the standard two-step GMM estimator that minimizes the objective function

$$\bar{h}(\beta, \Pi, \Gamma)' \ \tilde{\Lambda} \ \bar{h}(\beta, \Pi, \Gamma). \tag{1.6.7}$$

Let $\tilde{\beta}_{cc}$ be the complete cases GMM that minimizes (1.6.7) with $h(.) = h_1(.)$ and $\tilde{\Lambda}$ is a consistent estimate of $[\mathbb{E}(h_1(.)h_1(.)')]^{-1}$. Then we know that $\tilde{\beta}$ is no less efficient than $\tilde{\beta}_{cc}$ because the former uses more moment conditions.

Proposition 1.6.1. *Under Assumption 1.6.1,*

$$Avar\big(\sqrt{n}(\tilde{\beta}_{cc} - \beta)\big) - Avar\big(\sqrt{n}(\tilde{\beta} - \beta)\big) \text{ is positive semi-definite.}$$

Similar to Section 1.5, we can break down the efficiency gains by β_1 and β_2, the coefficients on the endogenous and exogenous elements of x respectively, and show that it is possible to obtain strict gains in efficiency for both β_1 and β_2.[18]

This is in contrast with the sequential estimator discussed in Mogstad & Wiswall (2012). They consider the case where $p = q = 1$ and the estimator proceeds in two steps. In the first step, it estimates Γ using $h_3(.)$. It then replaces the missing values of z_1 by the imputed values $x_2\hat{\Gamma}$, where $\hat{\Gamma}$ is the first step estimate of Γ, and then in the second step estimates β by minimizing (1.6.7) where $h(.) = z^{*\prime}(y - x\beta)$ and $z^* = \big(s_3 z_1 + (1 - s_3)x_2\hat{\Gamma}, x_2\big)$.[19] They show that the estimate of β_1 using

[17] Since $\Pi_{21} = 0$ and $\Pi_{22} = I$, the only elements of Π that are being estimated are Π_{11} and Π_{21}.
[18] This proof is analogous to that of Proposition 5.1.2 and is available upon request.
[19] The weight matrix is irrelevant in this case due to exact identification.

this estimator is numerically equivalent to that using complete cases estimator $\tilde{\beta}_{cc}$. Thus, $\tilde{\beta}$ does better than this estimator as it is possible to obtain strict gains in efficiency for both β_1 and β_2.

Abrevaya & Donald (2011) also propose a GMM estimator to deal with missingness in z_1. Their estimator is based on the moment functions

$$h_A(\beta) = z_A'(y - x\beta), \tag{1.6.8}$$

where $z_A = (x_2, (1 - s_3)x_2, s_3 z_1)$. It is clear that the moment functions in (1.6.6) contain (1.6.8) as a linear combination plus some additional moment conditions. Thus, $\tilde{\beta}$ is no less efficient than their estimator.

Now, when x_1 is exogenous in equation (1.2.1) in the sense that

$$\mathbb{E}(x_1'u) = 0, \tag{1.6.9}$$

then $x_1 = z_1$. In this case, $h_2(.) = 0$ and $h_4(.)$ cannot be used anymore.[20] So our vector of moment conditions is

$$\mathbb{E}\begin{bmatrix} s_3 x'(y - x\beta) \\ s_3 x_2' \otimes (x_1 - x_2\Gamma)' \\ (1 - s_3)x_2'[y - x_2(\Gamma\beta_1 + \beta_2)] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{1.6.10}$$

These are the moment conditions used by Abrevaya & Donald (2017) who consider the case of missingness in a single exogenous covariate. Thus, the framework presented here encompasses theirs as a special case when when x_1 is exogenous and $p = 1$.

1.7 Nonlinearity in covariates and instruments

Nonlinear functions of the covariates, like squares and interactions, are frequently used in empirical work. If these covariates are endogenous, one generally uses nonlinear functions of the instruments as well. In general, any sequential procedures that plug in the fitted values of the covariates or the instruments from a first step into nonlinear functions of these variables generally produce inconsistent estimates. For instance, traditional imputation used when the covariates are

[20] $h_2(.) = 0$ because Π_{11} is a $p \times p$ identity matrix since $z_1 = x_1$ and Π_{21} is a matrix of zeros.

exogenous will result in inconsistency if one replaces the missing value of say, the square of a covariate, with square of the imputed value of that covariate. In this section, I provide estimators that are consistent as well as more efficient than these sequential procedures and the complete case methods.

Suppose that the model of interest is now given by

$$y = F_1(x_1, x_2)\beta + u, \tag{1.7.1}$$

where x_1 is a $1 \times p$ vector of potentially endogenous covariates, x_2 is a $1 \times k$ vector of exogenous covariates, $x = (x_1, x_2)$, and $F_1(x_1, x_2)$ is a $1 \times j_1$ vector of potentially nonlinear functions of x_1 and x_2, $j_1 \geq (p + k)$. For instance, suppose $p = k = 1$. Then $F_1(x_1, x_2)$ could equal $(x_1, x_1^2, x_1x_2, x_2)$. We also have a $1 \times q$ vector of instruments z_1 for x_1, $q \geq p$. I assume

$$\mathbb{E}(u|z_1, x_2) = 0, \tag{1.7.2}$$

and allow for $\mathbb{E}(x_1'u) \neq 0$. So I now assume that u has a zero mean conditional on z_1 and x_2.[21] The first stage is given by the linear projection

$$F_1(x_1, x_2) = F_2(z_1, x_2)\Pi + r. \tag{1.7.3}$$

$F_2(z_1, x_2)$ is a $1 \times j_2$ vector of instruments where $F_2(.)$ is chosen by the researcher, and Π is a $j_2 \times j_1$ vector of coefficients. Because $F_1(x_1, x_2)$ contains nonlinear functions of x_1, $F_2(z_1, x_2)$ will most likely also contain nonlinear functions of z_1 and x_2. For instance, as discussed in Wooldridge (2010)[22], if $F_1(x_1, x_2) = (x_1, x_1^2, x_1x_2, x_2)$, one might want to choose $F_2(z_1, x_2) = (z_1, z_1^2, z_1x_2, x_2, x_2^2)$. By definition

$$\mathbb{E}[F_2(z_1, x_2)'r] = 0. \tag{1.7.4}$$

From equations (1.7.1) and (1.7.3), we get a reduced form for y in terms of only z_1 and x_2.

$$y = F_2(z_1, x_2)\Pi\beta + v, \tag{1.7.5}$$

[21] This is a standard assumption made in the literature when the model includes nonlinear functions of covariates and motivates the choice of instruments.

[22] Section 9.5.

where $v \equiv r\beta + u$. Using (1.7.2) and (1.7.4), we have that

$$\mathbb{E}[F_2(z_1, x_2)'v] = 0. \tag{1.7.6}$$

1.7.1 Missingness in outcome and covariates

Starting with the case of missingness in y and x_1, let the scheme of missingness be the same as described in Section 1.3. That is, both y and x_1 contain missing values, while z_1 and x_2 are always observed.

In this case, what seems like the natural extension of the sequential estimator discussed in McDonough & Millimet (2017) will be inconsistent for β because it performs the "forbidden regression" as discussed in Wooldridge (2010).[23] For instance, let $F_1(x_1, x_2) = (x_1, x_1^2, x_1x_2, x_2)$. The sequential estimator would regresses x_1 on $F_2(z_1, x_2)$ and obtain the fitted values (say $\hat{x}_1$) in the first step, replace the missing values of x_1, x_1^2 and x_1x_2 with $\hat{x}_1$, $(\hat{x}_1)^2$ and $\hat{x}_1x_2$ respectively, and then estimate β using 2SLS in the second step treating the fitted values as data. The inconsistency is a result of replacing nonlinear functions of x_1 with the same nonlinear function of fitted values. The correct way to go is to simultaneously estimate the first stage parameters Π and the parameters of interest β.

I first impose the following assumption on the missingness mechanism.

Assumption 1.7.1.1. $(i)\ \mathbb{E}[s_1s_2F_2(z_1, x_2)'u] = 0 \quad (ii)\ \mathbb{E}[s_1s_2F_2(z_1, x_2)'r] = 0$
$(iii)\ \mathbb{E}[s_1F_2(z_1, x_2)'u] = 0 \quad (iv)\ \mathbb{E}[s_1F_2(z_1, x_2)'r] = 0 \quad (v)\ \mathbb{E}[s_2F_2(z_1, x_2)'r] = 0.$

This gives us the following moment conditions.

$$\mathbb{E}[g_{NL}(\beta, \Pi)] = \mathbb{E}\begin{bmatrix} s_1s_2F_2(z_1, x_2)'[y - F_1(x_1, x_2)\beta] \\ s_1s_2F_2(z_1, x_2)' \otimes [F_1(x_1, x_2) - F_2(z_1, x_2)\Pi]' \\ (1 - s_1)s_2F_2(z_1, x_2)' \otimes [F_1(x_1, x_2) - F_2(z_1, x_2)\Pi]' \\ s_1(1 - s_2)F_2(z_1, x_2)'[y - F_2(z_1, x_2)\Pi\beta] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{1.7.7}$$

Compared to Section 1.4, we have simply replaced x with $F_1(x_1, x_2)$ and z with $F_2(z_1, x_2)$. Unlike Section 1.4 though, the traditional imputation is not consistent.

[23] Section 9.5.2.

1.7.2 Missingness in instruments

Next we move to the missing data scenario of Section 1.6. That is, the only variables that contain missing values are the excluded instruments z_1. We re-write equation (1.7.3) as follows by breaking up $F_2(z_1, x_2)$ into elements that do and do not depend on z_1.

$$F_1(x_1, x_2) = F_{21}(z_1, x_2)\Pi_a + F_{22}(x_2)\Pi_b + r, \tag{1.7.8}$$

where $F_2(z_1, x_2)\Pi \equiv F_{21}(z_1, x_2)\Pi_a + F_{22}(x_2)\Pi_b$, $F_{21}(z_1, x_2)$ is a $1 \times j_{21}$ vector that includes all elements of $F_2(z_1, x_2)$ that are functions of z_1, $F_{22}(x_2)$ is a $1 \times j_{22}$ vector that includes all elements of $F_2(z_1, x_2)$ that are functions only of x_2, and $j_2 = j_{21} + j_{22}$. From our example in Section 1.7.1, if $F_2(z_1, x_2) = (z_1, z_1^2, z_1x_2, x_2, x_2^2)$, then $F_{21}(z_1, x_2) = (z_1, z_1^2, z_1x_2)$ and $F_{22}(x_2) = (x_2, x_2^2)$. To handle missingness in z_1, we also need a linear projection of each of the instruments on $F_{22}(x_2)$.[24]

$$F_{21}(z_1, x_2) = F_{22}(x_2)\Gamma + e, \tag{1.7.9}$$

where by definition

$$\mathbb{E}[F_{22}(x_2)'e] = 0. \tag{1.7.10}$$

This gives us the reduced forms of $F_1(x_1, x_2)$ and y in terms of x_2 only. Plugging (1.7.9) into (1.7.8) we get

$$F_1(x_1, x_2) = F_{22}(x_2)(\Gamma\Pi_a + \Pi_b) + e\Pi_a + r. \tag{1.7.11}$$

Similarly, plugging (1.7.11) into (1.7.1) we get

$$y = F_{22}(x_2)(\Gamma\Pi_a + \Pi_b)\beta + (e\Pi_a + r)\beta + u. \tag{1.7.12}$$

Next, I impose the following assumption on the missingness mechanism.

Assumption 1.7.2.1. (i) $\mathbb{E}[s_3F_2(z_1, x_2)'u] = 0$ (ii) $\mathbb{E}[s_3F_2(z_1, x_2)'r] = 0$ (iii) $\mathbb{E}[s_3F_{22}(x_2)'e] = 0$.

[24] Based on the exact functional form of $F_1(.)$, one might want to choose different functions of x_2 in equation (1.7.9) than those in $F_{22}(.)$. This framework can be easily extended to allow for that by replacing $F_{22}(x_2)$ by a different function $F_3(x_2)$ in (1.7.9) and deriving the reduced forms in (1.7.11) and (1.7.12) accordingly. For the ease of exposition, I stick here with the same functions of x_2 in both (1.7.8) and (1.7.9).

This gives us the following moment conditions.

$$\mathbb{E}[h_{NL}(\beta,\Pi,\gamma)] = \mathbb{E}\begin{bmatrix} s_3F_2(z_1,x_2)'[y - F_1(x_1,x_2)\beta] \\ s_3F_2(z_1,x_2)' \otimes [F_1(x_1,x_2) - F_2(z_1,x_2)]' \\ s_3F_{22}(x_2)' \otimes [F_{21}(x_1,x_2) - F_{22}(x_2)\Gamma]' \\ (1-s_3)F_{22}(x_2)' \otimes [F_1(x_1,x_2) - F_{22}(x_2)(\Gamma\Pi_a + \Pi_b)]' \\ (1-s_3)F_{22}(x_2)'\big(y - F_{22}(x_2)(\Gamma\Pi_a + \Pi_b)\beta] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{1.7.13}$$

In the case where x_1 is exogenous (and hence $x_1 = z_1$), this reduces to

$$\mathbb{E}[h_{NL}(\beta,\Pi,\gamma)] = \mathbb{E}\begin{bmatrix} s_3F_2(z_1,x_2)'[y - F_1(x_1,x_2)\beta] \\ s_3F_{22}(x_2)' \otimes [F_{21}(x_1,x_2) - F_{22}(x_2)\Gamma]' \\ (1-s_3)F_{22}(x_2)'[y - F_{22}(x_2)(\Gamma\Pi_a + \Pi_b]\beta] \end{bmatrix} \tag{1.7.14}$$

As discussed in Abrevaya & Donald (2017), when x_1 is exogenous, the second most commonly used method after the complete cases OLS is linear imputation. In the example we have been carrying along where $F_1(x_1,x_2) = (x_1, x_1^2, x_1x_2, x_2)$, it proceeds as follows. In the first step, it regresses x_1 on x_2 and obtains the fitted values (say $\tilde{x}_1$). In the second step, it replaces the missing values of x_1, x_1^2 and x_1x_2 with $\tilde{x}_1$, $\tilde{x}_1^2$, and $\tilde{x}_1x_2$ respectively. Not only does this method not use the optimal instruments for x_1 (as it fails to include the nonlinear functions of x_2 in the imputation equation), it performs a forbidden regression in the second step, and hence results in inconsistent estimates for β.

1.8 Monte Carlo simulations

1.8.1 Missingness in outcome and covariates

The data generating process is as follows.

$$y = 1 + x_1\beta_1 + x_2\beta_2 + u,$$

where x_1 is a scalar and $x_2 = [1 \;\; x_{22} \;\; x_{23}]$ is a 1×3 vector. Moreover,

$$\begin{bmatrix} x_{22} \\ x_{23} \end{bmatrix} \sim N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 & 0.1 \\ 0.1 & 3 \end{bmatrix}\right)$$

$\beta_2 = (\beta_{21}, \beta_{22}, \beta_{23})'$ is fixed at $(1, 1, 1)'$ throughout all designs. The error is $u = \sigma_u u^*$, where u^* is a standard normal, and σ_u will be used to vary the error variance. The vector of instruments $z_1 = (z_{11}, z_{12}, z_{13}, z_{14})$ is 1×4 vector where

$$\begin{bmatrix} z_{11} \\ z_{12} \\ z_{13} \\ z_{14} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.4 & 0.3 \\ & 1 & 0.2 & 0.1 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix} \right)$$

The first stage is given by

$$x_1 = z_1 \Pi_{11} + x_2 \Pi_{21} + r_1,$$

where $\Pi_{11} = (1, 1, 1, 1)'$, $\Pi_{21} = (0.5, 0.5, 0.5)'$ and $r_1 = r_1^* + u^*$, where r_1^* is a standard normal. Thus, u^* is the part of x_1 that is correlated with u and causes x_1 to be endogenous.

The missingness is based on a uniform random variable, making the data MCAR.

$$s^* \sim \mathcal{U}(0, 1), \quad s_1 = 1[s^* < a \text{ or } s^* > b], \quad s_2 = 1[s^* < b].$$

I consider 4 designs.

Design 1: $\beta = 1$, $\sigma_u = 3.5$, $a = 0.5$, $b = 0.75$.

Design 2: $\beta = 1$, $\sigma_u = \sqrt{exp(z_{11}^2)}$, $a = 0.5$, $b = 0.75$.

Design 3: $\beta = 1$, $\sigma_u = \sqrt{exp(z_{11}^2)}$, $a = 0.4$, $b = 0.75$.

Design 4: $\beta = 0.1$, $\sigma_u = \sqrt{exp(z_{11}^2)}$, $a = 0.5$, $b = 0.75$.

The first design is the basic case of homoskedasticity in the model of interest. Design 2 allows for u to be heteroskedastic. Design 3 reduces the percentage of complete cases, and design 4 reduces the magnitude of the coefficient of interest. For all the designs, I do 1000 iterations with $n = 3000$.

I look at five estimators, starting with the most commonly used, which is the complete cases 2SLS. When the data is heteroskedastic, a GMM based on the complete cases will be more efficient than the 2SLS, and that is the second estimator I consider. The third is the imputation estimator

discussed in Section 1.5.3, followed by the dummy variable method and finally the proposed estimator.

The first thing to note is that the proposed estimator works best in terms of efficiency in all cases, with substantial reductions in the standard deviation relative to other estimators. This is true not only for β_{22} and β_{23}, the coefficients on x_2, but also for β_1, the coefficient on the covariate with missing values. The pattern on bias relative to other estimators is less clear, but the proposed estimator still has the smallest root mean squared error out of all the estimators in all cases.

The gains in efficiency of the proposed estimator are more pronounced when we have heteroskedasticity. Relative to the complete cases GMM, the gains increase as the percentage of complete cases decreases, which is to be expected as the proposed estimator now incorporates more additional information into estimation. The gains remain substantial in the case where the coefficient on the covariate with missing values is small.

The complete cases GMM is more efficient than the complete cases 2SLS when there is heteroskedasticity because of the optimal weighting, as expected. Yet it is less efficient than the proposed estimator in all cases, including when the error in the model of interest is homoskedastic. The imputation estimator on the other hand is not guaranteed to bring any efficiency gains relative to the complete cases GMM, and hence has no reason to be preferred over the former. The dummy variable method shows severe bias in all but the last design where the coefficient on the variable with missing value is close to 0, and does not even guarantee gains in efficiency over the complete cases GMM. Thus, this estimator cannot be recommended either.

1.8.2 Missingness in instruments

The data generating process is as follows.

$$y = 1 + x_1\beta_1 + x_2\beta_2 + u.$$

where x_1 is a scalar and $x_2 = [1 \;\; x_{22} \;\; x_{23}]$ is a 1×3 vector. Moreover,

$$\begin{bmatrix} x_{22} \\ x_{23} \end{bmatrix} \sim N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 & 0.2 \\ 0.2 & 1 \end{bmatrix} \right)$$

$\beta_1 = 1$, $\beta_2 = (\beta_{21}, \beta_{22}, \beta_{23})'$ is fixed at $(1, 1, 1)'$ throughout all designs. The error is $u = \sigma_u u^*$, where u^* is a standard normal, and σ_u will be used to vary the error variance. We have a single instrument z_1 such that

$$z_1 = x_2\Gamma + e,$$

where $\Gamma = (1, 0.5, 0.5)'$ and e is standard normal. The first stage is given by

$$x_1 = z_1\Pi_{11} + x_2\Pi_{21} + r_1,$$

where $\Pi_{11} = 1$, $\Pi_{21} = (1, 0.5, 0.5)'$ and $r_1 = r_1^* + u^*$, where r_1^* is a standard normal and u^* is the part of x_1 that is correlated with u.

The missingness is based on a uniform random variable, making the data MCAR.

$$s^* \sim \mathcal{U}(0, 1), \quad s_3 = 1[s^* > a].$$

I consider 3 designs.

Design 5: $\sigma_u = 4$, $a = 0.5$.

Design 6: $\sigma_u = \sqrt{exp(z_{11}^2)}$, $a = 0.5$.

Design 7: $\sigma_u = \sqrt{exp(z_{11}^2)}$, $a = 0.4$.

Design 5 is the case of homoskedasticity in the model of interest, design 6 allows for u to be heteroskedastic, and design 7 increases the percentage of complete cases. For all the designs, I do 1000 iterations with $n = 2000$.

The results are qualitatively similar to those in the previous sub-section. The proposed estimator substantially improves efficiency and has a lower root mean squared error relative to the complete cases 2SLS in all cases including that of homoskedasticity.[25] The gains are more pronounced in the case of heteroskedasticity and increase with a reduction in the percentage of complete cases.

[25]The only exception is β_1 in the case of homoskedasticity, where the two estimators perform equally well.

The imputation estimator for β_1 is numerically equivalent to the complete cases 2SLS, as noted in Mogstad & Wiswall (2012). For β_{22} and β_{23}, this estimator always does no better than the proposed estimator, and sometimes does worse than even the complete cases 2SLS. Since it does not guarantee efficiency gains over the complete cases or the proposed estimator, there is no reason to prefer it over either of the two.

1.9 Empirical application

I estimate the effect of physician's advice to reduce weight on calorie consumption by individuals using the estimator proposed in Section 1.4. As noted by Joshi and Wooldridge (2020), physician's advice is a low cost and precisely targeted intervention that can affect food consumption habits of individuals. The effect of physician's advice on outcomes like smoking, dietary and exercise behavior has been considered by Loureiro & Nayga Jr (2006), Loureiro & Nayga Jr (2007), Secker-Walker et al. (1998), and Ortega-Sanchez et al. (2004), among others.

The data comes from five most recent cycles of National Health and Nutritional Examination Survey (NHANES): 2007-08, 2009-2010, 2011-12, 2013-14, and 2015-16.[26] The NHANES is designed to assess the health and nutritional status of adults and children in the US. It examines a nationally representative sample of about 5000 persons each year and contains demographic, socioeconomic, dietary, and health-related questions.

The dependent variable (y) is the log of calorie intake of individuals. The endogenous covariate (x_1) is a binary variable which equals one if the physician advised the individual to lose weight. The excluded instruments (z_1) are binary variables indicating whether the individual has health insurance and a regular source of care. Other explanatory variables (x_2) include demographic variables like age, gender, race, education, and income of the individual as well as health-related variables such as the individual's body mass index (BMI), and indicators for whether they have high blood pressure, high cholesterol, Arthritis, a heart condition and Diabetes. Also included are year fixed effects and all variables have been demeaned.

[26] I would like to thank Riju Joshi for providing me with neatly compiled and cleaned data.

I restrict the sample to overweight individuals, that is, those with BMI greater than or equal to 25. I also exclude from the sample women who are pregnant, and individuals for whom the covariates x_2 or the excluded instruments z_1 are missing. The final sample consists of 11,512 observations with y missing for 952 observations and x_1 missing for 2173 observations.

Table B8 reports the results for two estimators: the complete cases GMM and the estimator proposed in Section 1.4 which uses the incomplete cases. The former results in the coefficient of interest being insignificant, which continues to hold true with the reduced standard error resulting from the proposed estimator. The standard errors for all other coefficients are smaller as well using the proposed estimator, while the coefficients for most variables remain similar to those obtained using the complete cases GMM.

1.10 Conclusion

I have offered some simple GMM estimators that improve efficiency over the currently used methods in the presence of missing data in linear regression models with endogenous covariates. I consider the cases of missingness in the outcomes and the endogenous covariates as well as that of missingness in the instruments. The latter includes the missingness in exogenous covariates as a special case. I also consider models that are nonlinear in the covariates and need a more careful treatment to ensure consistency. Thus, my framework can be used to deal with missingness in a wide variety of models frequently used in empirical work. In ongoing work, I am extending these methods to the case of panel data and models nonlinear in the parameters.

CHAPTER 2

IMPUTING MISSING COVARIATE VALUES IN NONLINEAR MODELS

2.1 Introduction

Nonlinear models are widely considered better suited to explain limited dependent variables than linear models. With missing covariate values - a ubiquitous problem in empirical research - nonlinear models become even more important because unlike the case where all variables are observed, estimates from linear models are now not necessarily consistent for parameters in the best linear approximations to nonlinear models.[1] Yet not much of the vast literature on missing data has explicitly addressed the unique issues that arise when dealing with missingness in nonlinear models.

Economists deal with missing covariate values predominantly in three ways. The most common thing to do is to just use the “complete cases" - the observations for which all the covariates are observed. While easy to use, this method can lead to substantial loss of efficiency because of discarding the incomplete cases. This has inspired methods that make use of these incomplete cases. The first commonly used method in this regard is the dummy variable method (DVM), which replaces the missing values with 0 and includes an indicator for missingness as an additional covariate. The second commonly used method is two-step regression imputation. In the first step, it regresses the covariate with missing values (CMV) on the always-observed covariates using the complete cases and uses the estimated coefficients to predict missing values of the CMV. In the second step, it estimates the model of interest using all observations with this “composite" CMV, which consists of both observed and predicted values (Dagenais, 1973). Table D1 summarizes the usage of these methods in 5 highly ranked economics journals in the last 3 years. Out of 846 papers, about 26% reported having missing data. Out of these, about 62%, 19% and 14% used the complete cases estimator, the DVM and the two-step regression imputation respectively.[2]

[1] I discuss this issue in detail in Section 2.6.4. Also see Wooldridge (2002).

[2] Of all the other methods used, no single category stood out. About 18% of the papers use other methods, most

The choice of method comes down to consistency and relative efficiency. The complete cases estimator generally requires the least number of assumptions in both linear and nonlinear models to be consistent. For instance, when the econometric model is correctly specified, say a model of a mean or a distribution conditional on the covariates, it only requires that the missingness depends only on the covariates (Wooldridge, 2002). However, as mentioned above, it can be inefficient relative to the other two estimators that use the incomplete cases. The DVM on the other hand is generally inconsistent even in linear models (Jones, 1996) and as I show in this paper, in nonlinear models as well, unless some very strong zero assumptions are imposed. Even with these assumptions, it does not guarantee efficiency improvements over the complete cases estimator (Abrevaya & Donald, 2017). Yet this method is still widely used as is evident from Table D1, perhaps because of its ease of use.

Two-step regression imputation also imposes additional assumptions on the model relative to the complete cases estimator, but these assumptions are much more plausible than those imposed by DVM. Practically, the most important one is ruling out the dependence of missingness on the CMV itself. Under this assumption, it is generally consistent in linear models.

However, in this paper I show that even under this assumption, this method is generally inconsistent in nonlinear models. Most notable are models based on conditional means, including commonly used models like probit, tobit, and Poisson regression. The reason for inconsistency is that this method simply plugs the imputed values in the same objective function that one would minimize if there were no missing values. However, in nonlinear models, this objective function does not necessarily capture the correct relationship between the observed variables in observations with missing values. The core issue is that conditional expectation does not pass through nonlinear functions, unlike linear ones. For instance, in binary choice models, simply plugging imputed values in the standard probit response probability and maximizing the resulting log likelihood will generally result in inconsistency in estimators of both the structural parameters and other quantities

of which are ad-hoc. This includes methods like replacing missing values with observations from the previous or following time period in case of panel data (5%), replacing missing values with 0 (4%), and dropping or combining variables with missingness (2%). Some papers also used hot deck (3%) and context specific imputation methods (2%). There were 2 instances each of multiple imputation and weighting.

of interest, such as average partial effects. To my knowledge, this issue has not been addressed in the literature and on the contrary, it has been claimed that this method *is* consistent in binary choice models (DeCanio & Watkins, 1998).

The key contribution of this paper is to propose a one-step imputation estimator which relies on the same assumptions as two-step imputation, but is consistent in nonlinear models. It simultaneously estimates the model of interest and the imputation model using the complete cases and a "reduced form" using all observations. The reduced form is a version of the main model in which we have "integrated out" the CMV using the imputation model, and hence it is able to make use of the incomplete cases. The key is that it correctly captures the relationship between the observed variables when the CMV is missing.

The estimator provides potentially strict efficiency gains over the complete cases estimator for all coefficients, and using a generalized method of moments (GMM) framework provides the overidentification test as a test for underlying restrictions. The method is an extension of Abrevaya & Donald (2017), who proposed a one-step imputation estimator for linear models. I provide a unified treatment of linear and nonlinear models using an M-estimation framework. Special cases include linear and nonlinear least squares, conditional maximum likelihood, and quasi maximum likelihood methods.

A second contribution is that I allow for nonlinearity in the imputation model itself. As mentioned above, the presence of missing data heightens the concerns about using linear models for limited dependent variables. Therefore, when imputing say a binary CMV, a probit may be more appropriate than a linear probability model. To my knowledge, regression imputation literature has solely focused on linear imputation models, though some of these nonlinear models have been discussed in the context of multiple imputation which is a Bayesian method of imputing (Rubin, 1987, Van Buuren, 2007).

The rest of this paper is organized as follows. Section 2.2 lays out the population minimization problems obtained from the underlying model of interest and imputation model. Section 2.3 describes the selection problem and estimation of selection probabilities. Section 2.4 derives the

proposed estimator, its asymptotic distribution and a simple estimator of the asymptotic variance. Section 2.5 discusses two practically important examples: nonlinear models for fractional responses and nonnegative responses, including count responses. Within each model, I consider a continuous and a binary CMV. Section 2.6 compares the proposed estimator to three other estimators: complete cases, two-step imputation and DVM. Section 2.7 provides simulation results showing the relative performance of these estimators. Section 2.8 provides an empirical application to the estimation of association between grade variance and educational attainment as considered in Sandsor (2020). Section 9 concludes. Proofs, tables and figures are given in appendices.

2.2 The population optimization problems

We start with the population optimization problem which defines the parameters of interest. Let y be a $1 \times J$ random vector taking values in $\mathbb{Y} \subset \mathbb{R}^J$ and x be a $1 \times (K+1)$ random vector taking values in $\mathbb{X} \subset \mathbb{R}^{K+1}$. We are interested in explaining y in terms of x. Some aspect of the joint distribution of (y, x) depends on a $L_1 \times 1$ parameter vector, α, contained in a parameter space $\mathbb{A} \subset \mathbb{R}^{L_1}$. Let $f_1(y, x_1, x_2, \alpha)$ denote an objective function.

Assumption 2.2.1. α_0 is the *unique* solution to the population minimization problem

$$\min_{\alpha \in \mathbb{A}} \mathbb{E}[f_1(y, x_1, x_2, \alpha)]. \tag{2.2.1}$$

Often, α_0 indexes some correctly specified feature of the distribution of y conditional on x, such as a conditional mean or a conditional median. But we will derive consistency and asymptotic normality results for a general class of problems in which the underlying population model can be misspecified in some way.

Next, let $x = (x_1, x_2)$, where x_1 is a scalar,[3] and x_2 is a $1 \times K$ random vector taking values in $\mathbb{X}_1 \subset \mathbb{R}$ and $\mathbb{X}_2 \subset \mathbb{R}^K$ respectively, and $\mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2$. As discussed in Section 2.3, we will allow x_1 to contain missing values and assume that (y, x_2) are always observed. Thus, we are interested in imputing x_1 using x_2. Let some aspect of the joint distribution of (x_1, x_2) depends on a $L_2 \times 1$

[3]The discussion for a random vector x_1, all elements of which are missing and observed at the same time, is essentially the same.

parameter vector β, contained in a parameter space $\mathbb{B} \subset \mathbb{R}^{L_2}$. Let $f_2(x_1, x_2, \beta)$ denote an objective function, and consider the population optimization problem which characterizes the imputation parameters.

Assumption 2.2.2. β_0 is the *unique* solution to the population minimization problem

$$\min_{\beta \in \mathbb{B}} \mathbb{E}[f_2(x_1, x_2, \beta)]. \tag{2.2.2}$$

Similar to the model of interest, the underlying population model here can be misspecified in some way.

The case that has been well studied in the classical imputation literature is where the underlying models for both $f_1(y, x, \alpha)$ and $f_2(x_1, x_2, \beta)$ are linear. The framework presented here allows for both the underlying models to be nonlinear as long as they are estimable using M-estimators, which includes maximum likelihood, quasi-maximum likelihood, nonlinear least squares, and many other procedures. For instance, if both y and x_1 are binary, we can let both $f_1(y, x, \alpha)$ and $f_2(x_1, x_2, \beta)$ be negative of probit log-likelihoods, instead of basing them on linear models. Alternatively, y could be a nonnegative count variable and x_1 could be continuous, in which case we can let $f_1(y, x, \alpha)$ be the negative of Poisson log-likelihood and let $f_2(x_1, x_2, \beta)$ come from a linear model. We consider these examples in detail in Section 2.5.

Next, we define a reduced form M-estimation problem which is based only on the always-observed variables (y, x_2). This reduced form is what allows us to use the incomplete cases, and hence is the key to the efficiency gains of the proposed estimator.

Let $\gamma = q(\alpha, \beta)$ be a (potentially nonlinear) $L_3 \times 1$ function of the parameters of interest α and the imputation parameters β, where γ is contained in a parameter space $\Gamma \subset \mathbb{R}^{L_3}$ and $L_3 \leq L_1 + L_2$. We assume that we can obtain a "reduced form" objective function $f_3(y, x_2, \gamma)$ in terms of the always-observed variables y and x_2 as well as γ such that $\gamma_0 = q(\alpha_0, \beta_0)$ uniquely minimizes this function.

Assumption 2.2.3. γ_0 is the *unique* solution to the population minimization problem

$$\min_{\gamma \in \Gamma} \mathbb{E}[f_3(y, x_2, \gamma)]. \tag{2.2.3}$$

The reduced form model underlying $f_3(y, x_2, \gamma)$ is derived by "integrating out" x_1 from the model of interest using the imputation model. When the model of interest is a linear projection or a model of conditional mean linear in the parameters, the reduced form can be derived using iterated projections or iterated expectations properties without having to do explicit integration. This is the case considered in Abrevaya & Donald (2017). In commonly used models nonlinear in the parameters like probit and Poisson regression, "substituting" for x_1 using the imputation model eliminates the need for explicit integration. We consider these examples in Section 2.5.

The dimension of γ warrants some discussion. It is possible that $L_3 < L_1 + L_2$, that is, the reduced form only identifies certain functions of α_0 and β_0, and not each element of α_0 and β_0 separately. Some examples are the case of linear projections considered in Abrevaya & Donald (2017) and the case of probit with continuous x_1 considered in Section 5.1.1. It is however, also possible that $L_3 = L_1 + L_2$, in which case $\gamma_0 = (\alpha_0', \beta_0')'$, for instance in the case of probit with binary x_1 considered in Section 2.5.1.2.[4]

Assumptions (2.2.1)-(2.2.3) imply that (α_0, β_0) is the unique solution to the following equations, provided that we can interchange the expectation and the derivative.

$$\mathbb{E}[g^*(y, x_1, x_2, \alpha, \beta)] = \mathbb{E}\begin{bmatrix} g_1^*(y, x_1, x_2, \alpha) \\ g_2^*(x_1, x_2, \beta) \\ g_3^*(y, x_2, \alpha, \beta) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{2.2.4}$$

where $g_1^*(y, x_1, x_2, \alpha) \equiv \nabla_\alpha f_1(y, x_1, x_2, \alpha)'$ is the $L_1 \times 1$ score of $f_1(y, x_1, x_2, \alpha)$, $g_2^*(x_1, x_2, \beta) \equiv \nabla_\beta f_2(x_1, x_2, \beta)'$ is the $L_2 \times 1$ score of $f_2(x_1, x_2, \beta)$, and $g_3^*(y, x_2, \alpha, \beta) \equiv \nabla_\gamma f_3(y, x_2, \gamma)'$ is the $L_3 \times 1$ score of $f_3(y, x_2, \gamma)$. (2.2.4) gives us a set of moment conditions, a transformation of which will be the basis of the proposed estimator as discussed in Section 2.4.

[4]As a note on notation, I express functions as explicitly depending on γ only when it is necessary to take into account the nature of γ. For instance, when $L_3 < L_1 + L_2$, the score of $f_3(y, x_2, \gamma)$ should only contain partial derivatives with respect to γ and not with respect to individual elements of α and β to prevent redundancy in the resulting moment conditions. But for the most part, when looking at the derivatives of $f_1(.)$, $f_2(.)$, and $f_3(.)$, we only need to acknowledge the fact that they are functions of (α, β).

2.3 Non random sampling and inverse probability weighting

I characterize nonrandom sampling through a selection indicator. For any random draw (y_i, x_{1i}, x_{2i}) from the population, we also draw s_i, a binary indicator equal to unity if x_{1i} is observed, and zero otherwise. We assume that y_i and x_{2i} are always observed. A generic element from the population is now denoted (y, x_1, x_2, s). Then the following assumption characterizes the nature of selection.

Assumption 2.3.1 (i) x_1 is observed whenever $s = 1$, (y, x_2) is always observed. (ii) There is a random vector z such that $P(s = 1|y, x, z) = P(s = 1|z) \equiv p(z)$. (iii) For all $z \in \mathbb{Z} \subset \mathbb{R}^M$, $p(z) > 0$. (iv) z is always observed.

Part (i) simply defines data observability. Parts (ii) and (ii) are the key assumptions. They state that selection is based on observable variables. This is the same as the "missing at random" assumption used in statistics literature (Rubin, 1976). Part (ii) states that s is independent of (y, x) conditional on z. Because the only variable assumed to contain missing values is x_1, we can, at a minimum, allow z to contain (y, x_2). Although apart from this, z can also contain some "outside" variables that are good predictors of selection and are always observed. Then Assumption 2.3.1 is more general than allowing s to depend only on the covariates x_2, which is the case considered in Abrevaya & Donald (2017) in the context of linear models.

Moreover, the framework presented here can also be used when y contains missing values. We simply redefine s to equal 1 when both y and x_1 are observed, and rule out z containing y in addition to z containing x_1. Then the proposed estimator discussed in the next section will impute using the observations for which *only* x_1 is missing, and discard the y-missing observations.

For selection as described in Assumption 2.3.1, note that the first and second moment functions in (2.2.4) can only use the $s = 1$ observations since they depend on x_1, and the third moment function is able to use the $s = 0$ observations. We will weight each of the moment functions by the inverse of appropriate probabilities in order to account for this selection. To this end, we specify a model for the selection probability. We assume that a conditional density determining selection is correctly specified, and that the standard regularity conditions required for maximum likelihood

estimation (MLE) of the selection model are satisfied. Let $D(.|.)$ denote conditional distribution.

Assumption 2.3.2 (i) $G(z, \delta)$ is a parametric model for $p(z)$, where $\delta \in \Delta \subset \mathbb{R}^P$ and $G(z, \delta) > 0$, all $z \in \mathbb{Z} \subset \mathbb{R}^M$, $\delta \in \Delta$. (ii) There exists $\delta_0 \in \Delta$ such that $p(z) = G(z, \delta_0)$. (iii) The estimator $\hat{\delta}$ solves the binary response problem

$$\max_{\delta \in \Delta} \sum_{i=1}^{N} \{s_i log[G(z_i, \delta)] + (1 - s_i) log[1 - G(z_i, \delta)]\}. \tag{2.3.1}$$

Given $\hat{\delta}$, we can form $G(z_i, \hat{\delta})$ for all i. This leads us to the problem of estimation.

2.4 Moment conditions and GMM

The proposed estimator is a GMM estimator based on the following transformation of the moment functions in (2.2.4).

$$g_i(\alpha, \beta; \delta) = \begin{bmatrix} g_{1i}(\alpha, \beta; \delta) \\ g_{2i}(\alpha, \beta; \delta) \\ g_{3i}(\alpha, \beta; \delta) \end{bmatrix} \equiv \begin{bmatrix} [s_i / G(z_i, \delta)] g_1^*(y_i, x_{1i}, x_{2i}, \alpha) \\ [s_i / G(z_i, \delta)] g_2^*(x_{1i}, x_{2i}, \beta) \\ g_3^*(y_i, x_{2i}, \alpha, \beta) \end{bmatrix}. \tag{2.4.1}$$

Because both $g_1^*(y_i, x_{1i}, x_{2i}, \alpha)$ and $g_2^*(x_{1i}, x_{2i}, \beta)$ are functions of x_{1i}, they can only use the complete cases - the observations for which $s_i = 1$. We thus multiply these by s_i and weight by the inverse of selection probability in the usual inverse probability weighting (IPW) fashion (Wooldridge, 2002, 2007). Since $g_3^*(y_i, x_{2i}, \alpha, \beta)$ is a function only of the always-observed variables y_i and x_{2i}, it can use all the observations including the incomplete cases and hence we do not need to weight it.

For a generic element from the population (y, x_1, x_2, z, s), denote this vector of moment functions by $g(\alpha, \beta; \delta)$ and its individual elements by $g_j(\alpha, \beta; \delta)$, $j = 1, 2, 3$. This is a set of overidentified moment functions. $g_1(.)$ exactly identifies the parameters of interest α_0 and $g_2(.)$ exactly identifies the imputation parameters β_0. The overidentification (and hence the efficiency gains) in the system come from $g_3(.)$. The number of overidentifying restrictions is L_3, the dimension of the reduced form parameters γ_0. Given the first step estimate $\hat{\delta}$, we can write the sample analogue

of moment conditions based on (2.4.1) as

$$\bar{g}_j(\alpha, \beta; \hat{\delta}) = N^{-1} \sum_{i=1}^{N} g_{ji}(\alpha, \beta; \hat{\delta}), \quad j = 1, 2, 3, \tag{2.4.2}$$

and $\bar{g}(\alpha, \beta; \hat{\delta}) = [\bar{g}_1(\alpha, \beta; \hat{\delta})', \ \bar{g}_2(\alpha, \beta; \hat{\delta})', \ \bar{g}_3(\alpha, \beta; \hat{\delta})']'$. A GMM estimator based on (2.4.1) minimizes the following objective function with respect to (α, β).

$$\hat{Q}(\alpha, \beta; \hat{\delta}) = \bar{g}(\alpha, \beta; \hat{\delta})' \ \hat{W} \ \bar{g}(\alpha, \beta; \hat{\delta}), \tag{2.4.3}$$

where $\hat{W}$ is an estimated weight matrix such that $\hat{W} \xrightarrow{p} W$.

We first discuss identification of (α_0, β_0). The limit function for $\hat{Q}(\alpha, \beta; \hat{\delta})$ is $Q(\alpha, \beta; \delta_0) = \mathbb{E}[g(\alpha, \beta; \delta_0)]' \ W \ \mathbb{E}[g(\alpha, \beta; \delta_0)]$.

Lemma 2.4.1. *(Identification) Assume that W is a symmetric positive definite matrix. Then under Assumptions 2.2.1-2.2.3, 2.3.1, and 2.3.2, $Q(\alpha, \beta; \delta_0)$ has a unique minimum at (α_0, β_0).*

For a nonsingular W, the GMM identification condition reduces to $\mathbb{E}[g(\alpha, \beta; \delta_0)] \neq 0$ if $(\alpha, \beta) \neq (\alpha_0, \beta_0)$. Sufficient is to show that a corresponding condition holds for each element of $g(\alpha, \beta; \delta_0)$. For instance, $\mathbb{E}[g_1(\alpha, \beta; \delta_0)] \neq 0$ if $\alpha \neq \alpha_0$ follows from identification of α_0 in the population (Assumption 2.2.1) and the assumptions on selection (Assumptions 2.3.1 and 2.3.2). A formal proof of Lemma 2.4.1, along with all other proofs in the rest of the paper are given in the appendix.

The GMM estimator based on the general weight matrix $\hat{W}$ is defined as the following.

Definition 2.4.1: *Call the estimator of (α, β) that minimizes (2.4.3), $(\hat{\alpha}, \hat{\beta})$.*

Consistency of $(\hat{\alpha}, \hat{\beta})$ follows from Lemma 2.4.1 and standard regularity conditions given in the following theorem.

Theorem 2.4.1 *(Consistency) Assume that*

1. *$\{(y_i, x_i, z_i, s_i) : i = 1, \ldots, N\}$ are random draws from the population satisfying Assumptions 2.3.1 and 2.3.2.*

2. *The assumptions in Lemma 2.4.1 hold.*

3. $\mathbb{A}$, $\mathbb{B}$, Γ, Δ, $\mathbb{A}\times\Delta$, $\mathbb{B}\times\Delta$, *and* $\mathbb{A}\times\mathbb{B}\times\Gamma$ *are compact subsets of* $\mathbb{R}^{L_1}$, $\mathbb{R}^{L_2}$, $\mathbb{R}^{L_3}$, $\mathbb{R}^{P}$, $\mathbb{R}^{L_1+P}$, $\mathbb{R}^{L_2+P}$, *and* $\mathbb{R}^{L_1+L_2+L_3}$ *respectively.*

4. $f_1(y,x,\alpha)$, $f_2(x,\beta)$ *and* $f_3(y,x_2,\gamma)$ *are twice differentiably continuous on* $\mathbb{A}$, $\mathbb{B}$, Γ *respectively for each* (y,x), x *and* (y,x_2) *in* $\mathbb{Y}\times\mathbb{X}$, $\mathbb{X}$ *and* $\mathbb{Y}\times\mathbb{X}_2$ *respectively.*

5. $G(z,\delta)$ *is continuous in* Δ *for each* $z\in\mathbb{Z}$, *twice continuously differentiable on* $int(\Delta)$, *and* $\delta_0\in int(\Delta)$. *For some* $a>0$, $G(z,\delta)\geq a$ *for all* $z\in\mathbb{Z}$, $\delta\in\Delta$.

6. *For all* $(\alpha,\beta,\gamma)\in\mathbb{A}\times\mathbb{B}\times\Gamma$, $|g^*(y,x,\alpha,\beta,\gamma)|\leq b(y,x)$, *where*

$$b(y,x)\equiv[b_1(y,x)',b_2(x)',b_3(y,x_2)']'$$

and $b(.)$ *is a function such that* $\mathbb{E}[b(y,x)]<\infty$.

Then $(\hat{\alpha},\hat{\beta})\xrightarrow{p}(\alpha_0,\beta_0)$ *as* $N\to\infty$.

The consistency of $(\hat{\alpha},\hat{\beta})$ follows from standard arguments involving consistency of two-step M-estimators. First, analogous to the discussion in Wooldridge (2002), Lemma 2.4 of Newey & McFadden (1994) applies to show that $g_1(\alpha,\beta;\delta)$, $g_2(\alpha,\beta;\delta)$ and $g_3(\alpha,\beta;\delta)$ satisfy the uniform weak law of large numbers over $\mathbb{A}\times\Delta$, $\mathbb{B}\times\Delta$, Γ respectively under Assumptions 1, 3, 4, 5 and 6 of Theorem 2.4.1. Then the averages in (2.4.2) can be shown to converge to

$$\mathbb{E}[g_j(\alpha,\beta;\delta_0)],\quad j=1,2,3, \tag{2.4.4}$$

uniformly over $\mathbb{A}$, $\mathbb{B}$, and Γ respectively. Along with the identification from Lemma 2.4.1, this can be shown to imply consistency of $(\hat{\alpha},\hat{\beta})$ for (α_0,β_0).

Now, assuming that $\mathbb{E}[g(\alpha,\beta;\delta_0)]$ is differentiable at (α_0,β_0), its derivative is defined as the following.

$$D_0\equiv\mathbb{E}[\nabla_{(\alpha',\beta')'}g(\alpha,\beta;\delta_0)|_{(\alpha,\beta)=(\alpha_0,\beta_0)}]=\mathbb{E}[\nabla_{(\alpha',\beta')'}g^*(\alpha,\beta)|_{(\alpha,\beta)=(\alpha_0,\beta_0)}]=\begin{bmatrix}D_{11}^0 & 0\\ 0 & D_{22}^0\\ D_{31}^0 & D_{32}^0\end{bmatrix}, \tag{2.4.5}$$

where $D^0_{j1} = \partial g^*_j(\alpha,\beta)/\partial\alpha|_{(\alpha,\beta)=(\alpha_0,\beta_0)}$ and $D^0_{j2} = \partial g^*_j(\alpha,\beta)/\partial\beta|_{(\alpha,\beta)=(\alpha_0,\beta_0)}$, $j = 1,2,3$ and the first equality follows by the standard IPW argument given Assumptions 2.3.1 and 2.3.2. Then the following result gives the asymptotic distribution of $(\hat{\alpha},\hat{\beta})$.

Theorem 2.4.2 *(Asymptotic normality) Assume that*

1. *The assumptions in Theorem 2.4.1 hold.*
2. $(\alpha_0,\beta_0) \in int(\mathbb{A}\times\mathbb{B})$.
3. $g(\alpha,\beta;\delta)$ *is twice continuously differentiable on* $int(\mathbb{A}\times\mathbb{B}\times\Delta)$.
4. D_0 *is of full rank* $L_1 + L_2$.
5. $\mathbb{E}[\sup_{(\alpha,\beta;\delta)\in\mathbb{A}\times\mathbb{B}\times\Delta}|\nabla_{(\alpha,\beta,\delta)}g(\alpha,\beta,\delta)|] < \infty$.

Then,

$$\sqrt{N}[(\hat{\alpha}',\hat{\beta}')' - (\alpha_0',\beta_0')'] \xrightarrow{d} Normal[0,(D_0'WD_0)^{-1}D_0'WF_0WD_0(D_0'WD_0)^{-1}], \quad (2.4.6)$$

where $F_0 = \mathbb{E}(g_ig_i') - \{\mathbb{E}(g_id_i')[\mathbb{E}(d_id_i')]^{-1}\mathbb{E}(d_ig_i')\}\circ R$, $g_i \equiv g_i(\alpha_0,\beta_0;\delta_0)$, $d_i \equiv s_i(\nabla_\delta G_i'/G_i) - (1-s_i)[\nabla_\delta G_i'/(1-G_i)]$ *is the* $P\times 1$ *score of the binary response log-likelihood,* R *is a square matrix of order* $L_1+L_2+L_3$ *with all elements being unity except the lower right* $L_3\times L_3$ *block which is a 0 matrix,*[5] $G_i \equiv G(z_i,\delta_0)$, $H_0 \equiv \mathbb{E}[\nabla_\delta g(\alpha_0,\beta_0;\delta_0)]$ *and* $\psi(s_i,z_i) = -[\mathbb{E}(d_id_i')]^{-1}d_i$.

Standard GMM theory dictates that the optimal weight matrix to be used in (2.4.3) is $\hat{W} = \hat{F}^{-1}$, where $\hat{F}$ is a consistent estimate of F_0 which can be obtained as

$$\hat{F} = \left(N^{-1}\sum_{i=1}^N \hat{g}_i\hat{g}_i'\right) - \left[\left(N^{-1}\sum_{i=1}^N \hat{g}_i\hat{d}_i'\right)\left(N^{-1}\sum_{i=1}^N \hat{d}_i\hat{d}_i'\right)^{-1}\left(N^{-1}\sum_{i=1}^N \hat{d}_i\hat{g}_i'\right)\right]\circ R, \quad (2.4.7)$$

where

$$\hat{g}_i \equiv g_i(\hat{\alpha},\hat{\beta};\hat{\delta}), \quad \hat{d}_i \equiv s_i\left[\frac{\nabla_\delta G(z_i,\hat{\delta})'}{G(z_i,\hat{\delta})}\right] - (1-s_i)\left[\frac{\nabla_\delta G(z_i,\hat{\delta})'}{1-G(z_i,\hat{\delta})}\right]. \quad (2.4.8)$$

Then, the proposed estimator is the optimal GMM estimator based on (2.4.1), as defined below.

[5] $\circ$ *denotes a Hadamard product.*

Definition 2.4.2: *Call the estimator of* (α, β) *that minimizes (2.4.3) with* $\hat{W} = \hat{F}^{-1}$*, the weighted joint GMM estimator or* $(\hat{\alpha}_{WJ}, \hat{\beta}_{WJ})$.

Because $(\hat{\alpha}_{WJ}, \hat{\beta}_{WJ})$ uses the optimal weight matrix, the asymptotic variance in (2.4.6) reduces to $(D_0' F_0^{-1} D_0)^{-1}$. A consistent estimator can be obtained using $\hat{F}$ and a consistent estimator of D_0 defined as

$$\hat{D} = N^{-1} \sum_{i=1}^{N} [\nabla_{(\alpha', \beta')'} g_i(\hat{\alpha}, \hat{\beta}; \hat{\delta})]. \tag{2.4.9}$$

Then the following result follows from Theorem 2.4.2.

Theorem 2.4.3 *(Asymptotic Normality of the optimal GMM) Let all assumptions of Theorem 2.4.2 hold. Then,*

$$\sqrt{N}[(\hat{\alpha}_{WJ}', \hat{\beta}_{WJ}')' - (\alpha_0', \beta_0')'] \xrightarrow{d} Normal[0, (D_0' F_0^{-1} D_0)^{-1}], \tag{2.4.10}$$

and a consistent estimator of $Avar\{\sqrt{N}[(\hat{\alpha}_{WJ}', \hat{\beta}_{WJ}')' - (\alpha_0', \beta_0')']\}$ *is given by*

$$(\hat{D}' \hat{F}^{-1} \hat{D})^{-1}, \tag{2.4.11}$$

where $\hat{F}$ *is given in (2.4.7) and* $\hat{D}$ *is given in (2.4.9).*

Further, we can use the standard test of overidentifying restrictions based on the objective function evaluated at the parameter estimates proposed by Hansen (1982). The original result was obtained for a standard GMM. It is straightforward to extend the proof to the case where the moment functions depend on an estimate of δ from a first step.

Proposition 2.4.1: *Let all assumptions of Theorem 2.4.2 hold. Then under the null hypothesis that* $\mathbb{E}[g(\alpha_0, \beta_0; \delta_0)] = 0$,

$$N \; \bar{g}(\hat{\alpha}_{WJ}, \hat{\beta}_{WJ}; \hat{\delta})' \;\; \hat{F}^{-1} \;\; \bar{g}(\hat{\alpha}_{WJ}, \hat{\beta}_{WJ}; \hat{\delta}) \xrightarrow{p} \chi^2_{L_3}. \tag{2.4.12}$$

2.5 Examples

The proposed estimator can be applied to many cases relevant for empirical research. I provide two important examples: a binary or fractional y and a nonnegative y, both of which are estimated using quasi-MLE.

2.5.1 Models for binary and fractional responses

Binary response models are one of the most commonly used nonlinear models in empirical research. Suppose that y is a variable taking values in the unit interval, $[0, 1]$. This includes the case where y is binary but also allows y to be a continuous proportion. Further, y can have both discrete and continuous characteristics (for instance, y can be a proportion that takes on zero or one with positive probability). We start by assuming that the mean of y conditional on x has a probit form.

$$\mathbb{E}(y|x_1, x_2) = \Phi(\alpha_{10}x_1 + x_2\alpha_{20}) \equiv \Phi(x\alpha_0), \tag{2.5.1}$$

where x_1 is a scalar and x_2 is a $1 \times k$ vector. If x_1 was always observed, we would simply estimate α_0 using quasi-MLE with a Bernoulli log likelihood, which identifies the parameters in a correctly specified conditional mean by the virtue of being in the linear exponential family (Gourieroux et al., 1984). But because x_1 is sometimes missing, now we additionally specify a model to impute x_1 using x_2 and use it to obtain the reduced form conditional mean of y given x_2.

I consider two cases: where x_1 is continuous, and where it is binary.

2.5.1.1 Continuous covariate with missing values

We assume that the imputation model is linear.

$$x_1 = x_2\theta_0 + r, \tag{2.5.2}$$

$$r|x_2 \sim Normal[0, \sigma_0^2 exp(2x_{21}\lambda_0)], \tag{2.5.3}$$

where $x_{21} \subset x_2$. That is, x_1 is assumed to be normally distributed conditional on x_2. To make the model more flexible, we allow the error to be heteroskedastic with variance dependent on x_{21}. Typically, x_{21} will include all elements of x_2 except the constant, so that the case where r is homoskedastic with variance σ_0^2 is obtained as a special case by setting $\lambda_0 = 0$. The conditional pdf of x_1 is given by

$$f(x_1|x_2, \beta_0) = \frac{1}{\sqrt{2\pi\sigma_0^2 exp(2x_{21}\lambda_0)}} exp\left[-\frac{(x_1 - x_2\theta_0)^2}{2\sigma_0^2 exp(2x_{21}\lambda_0)} \right]. \tag{2.5.4}$$

In order to find $\mathbb{E}(y|x_2)$, we integrate out x_1 from $\mathbb{E}(y|x_1, x_2)$ given in (2.5.1) using the density given in (2.5.4).

$$\begin{aligned}\mathbb{E}(y|x_2) &= \int_{-\infty}^{\infty} \Phi(x\alpha_0)\sigma_0^{-1} exp(-x_{21}\lambda_0)\phi\left(\frac{x_1 - x_2\theta_0}{\sigma_0 exp(x_{21}\lambda_0)}\right)dx_1 \\ &= \Phi\left(\frac{x_2(\alpha_{10}\theta_0 + \alpha_{20})}{\sqrt{1 + \alpha_{10}^2\sigma_0^2 exp(2x_{21}\lambda_0)}}\right). \end{aligned} \tag{2.5.5}$$

We can derive $\mathbb{E}(y|x_2)$ without carrying out the explicit integration as well. Define a binary variable as following.

$$w^* = \alpha_{10}x_1 + x_2\alpha_{20} + u \equiv x\alpha_0 + u, \tag{2.5.6}$$

$$u|x_1, x_2 \sim Normal(0, 1), \tag{2.5.7}$$

$$w = 1[w^* > 0]. \tag{2.5.8}$$

Next, note that

$$\mathbb{E}(w|x_1, x_2) = \mathbb{E}(y|x_1, x_2) = \Phi(\alpha_{10}x_1 + x_2\alpha_{20}), \tag{2.5.9}$$

and so, by iterated expectations,

$$\mathbb{E}(w|x_2) = \mathbb{E}(y|x_2). \tag{2.5.10}$$

(2.5.10) is what allows us to obtain $\mathbb{E}(y|x_2)$. Substituting (2.5.2) into (2.5.6) gives

$$w^* = x_2(\alpha_{10}\theta_0 + \alpha_{20}) + v, \tag{2.5.11}$$

where $v \equiv u + \alpha_{10}r$ and $v|x_2 \sim Normal[0, 1 + \alpha_{10}^2\sigma_0^2 exp(2x_{21}\lambda_0)]$ under the assumptions made so far. Therefore,

$$w = 1[x_2(\alpha_{10}\theta_0 + \alpha_{20}) + v > 0], \tag{2.5.12}$$

which implies

$$\mathbb{E}(w|x_2) = P(w = 1|x_2) = P[v > -x_2(\alpha_{10}\theta_0 + \alpha_{20})|x_2], \tag{2.5.13}$$

which gives the same expression as (2.5.5). Now we can use quasi-MLE with a Bernoulli log likelihood for both the model of interest (2.5.1) and the reduced form (2.5.5), and full MLE for the

imputation model using (2.5.4). The objective functions in (2.2.1)-(2.2.3) are given by

$$
\begin{aligned}
f_1(y, x, \alpha) &= -log\{\Phi(x\alpha)^y[1-\Phi(x\alpha)]^{(1-y)}\} \\
f_2(x_1, x_2, \beta) &= -log\left\{\frac{1}{\sqrt{2\pi\sigma^2 exp(2x_{21}\lambda)}} exp\left[-\frac{(x_1 - x_2\theta)^2}{2\sigma^2 exp(2x_{21}\lambda)}\right]\right\} \\
f_3(y, x_2, \gamma) &= -log\{\Phi[h_1(x_2, \gamma)]^y\{1-\Phi[h_1(x_2, \gamma)]\}^{(1-y)}\}, \qquad (2.5.14)
\end{aligned}
$$

where $h_1(x_2, \gamma) \equiv [x_2(\alpha_1\theta + \alpha_2)]/\sqrt{1 + \alpha_1^2\sigma^2 exp(2x_{21}\lambda)}$ and in the general notation of Section 2.2, $\beta = (\theta, \sigma^2, \lambda)$ and $\gamma = [(\alpha_1\theta + \alpha_2), \alpha_1^2\sigma^2, \lambda]$.

The issue of defining γ warrants some discussion. It can be shown that first, the partial derivatives of $f_3(y, x_2, \gamma)$ with respect to (α_1, θ) are linear combinations of those with respect to $(\alpha_2, \sigma^2, \lambda)$. Since we use the the weighted versions of these partial derivatives as moment functions, we should use only those taken with respect to $(\alpha_2, \sigma^2, \lambda)$ to prevent redundancy in the resulting moment conditions. Second, the partial derivatives with respect to $(\alpha_2, \sigma^2, \lambda)$ are just scaled versions of those with respect to γ as defined above, which makes this definition of γ preferable both intuitively and for algebraic simplicity.

The objective functions in (2.5.14) result in the following score functions.

$$g_1^*(y,x,\alpha) = x' \frac{[y-\Phi(x\alpha)]\phi(x\alpha)}{\Phi(x\alpha)[1-\Phi(x\alpha)]}$$

$$g_2^*(x_1,x_2,\beta) = \begin{bmatrix} \dfrac{x_2'(x_1-x_2\theta)}{\sigma^2 exp(2x_{21}\lambda)} \\ \dfrac{(x_1-x_2\theta)^2}{exp(2x_{21}\lambda)\sigma^4} - \dfrac{1}{\sigma^2} \\ x_{21}'\left[\dfrac{(x_1-x_2\theta)^2}{\sigma^2 exp(2x_{21}\lambda)} - 1\right] \end{bmatrix}$$

$$g_3^*(y,x_2,\gamma) = \begin{bmatrix} \dfrac{x_2'}{\sqrt{1+\alpha_1^2\sigma^2 exp(2x_{21}\lambda)}} \\ \dfrac{exp(2x_{21}\lambda)x_2(\theta\alpha_1+\alpha_2)}{[1+\alpha_1^2\sigma^2 exp(2x_{21}\lambda)]^{3/2}} \\ \dfrac{exp(2x_{21}\lambda)x_2(\theta\alpha_1+\alpha_2)x_{21}'}{[1+\alpha_1^2\sigma^2 exp(2x_{21}\lambda)]^{3/2}} \end{bmatrix} \phi[h_1(x_2,\gamma)]\left[\frac{y-\Phi[h_1(x_2,\gamma)]}{\Phi[h_1(x_2,\gamma)]\{1-\Phi[h_1(x_2,\gamma)]\}}\right]. \tag{2.5.15}$$

In the case where $\lambda_0 = 0$ and hence r is homoskedastic, the third elements of $g_2^*(.)$ and $g_3^*(.)$, which are the partial derivatives with respect to λ of $f_2(.)$ and $f_3(.)$ respectively go away. Moreover, the second element of $g_3^*(.)$ in that case is just a linear function of the first element of $g_3^*(.)$ and hence should be removed to prevent redundancy.

Given these score functions and $\hat{\delta}$ obtained in Section 2.3, it is straightforward to form the moment functions in (2.4.1) and estimate (α_0, β_0) by minimizing (2.4.3).

2.5.1.2 Binary covariate with missing values

We now consider the case where x_1 is binary. Equations (2.5.2) and (2.5.3) are replaced by

$$x_1^* = x_2\theta_0 + r, \tag{2.5.16}$$

$$r|x_2 \sim Normal[0, exp(2x_{21}\lambda_0)], \tag{2.5.17}$$

$$x_1 = 1[x_1^* > 0], \tag{2.5.18}$$

where $x_{21} \subset x_2$. Just as in Section 2.5.1.1, x_{21} typically includes all elements of x_2 except the constant, so that we can get a standard probit with unit variance as a special case by setting $\lambda_0 = 0$. Now, (2.5.16)-(2.5.18) imply that

$$P(x_1 = 1|x_2) = \Phi[exp(-x_{21}\lambda_0)x_2\theta_0] \equiv \Phi[h_2(x_2, \beta_0)], \tag{2.5.19}$$

where in the general notation of Section 2.2, $\beta = (\theta, \lambda)$. Using (2.5.1) and iterated expectations,

$$\begin{aligned}\mathbb{E}(y|x_2) &= \mathbb{E}[\mathbb{E}(y|x_1, x_2)|x_2] = \mathbb{E}(y|x_1 = 1, x_2)P(x_1 = 1|x_2) + \mathbb{E}(y|x_1 = 0, x_2)P(x_1 = 0|x_2) \\ &= \Phi(\alpha_{10} + x_2\alpha_{20})\Phi[exp(-x_{21}\lambda_0)x_2\theta_0] + \Phi(x_2\alpha_{20})\{1 - \Phi[exp(-x_{21}\lambda_0)x_2\theta_0]\} \\ &\equiv h_3(x_2, \gamma_0),\end{aligned} \tag{2.5.20}$$

where in the general notation of Section 2.2, $\gamma = (\alpha, \beta)$. Analogous to the previous section, we use quasi-MLE with a Bernoulli log likelihood for the model of interest (2.5.1) and the reduced form (2.5.20), and full MLE for the imputation model using (2.5.19). The objective functions are given by

$$\begin{aligned}f_1(y, x, \alpha) &= -log\{\Phi(x\alpha)^y[1 - \Phi(x\alpha)]^{(1-y)}\} \\ f_2(x_1, x_2, \beta) &= -log(\Phi[h_2(x_2, \beta)]^{x_1}\{1 - \Phi[h_2(x_2, \beta)]\}^{(1-x_1)}) \\ f_3(y, x_2, \gamma) &= -log\{h_3(x_2, \gamma)^y[1 - h_3(x_2, \gamma)]^{(1-y)}\}.\end{aligned} \tag{2.5.21}$$

This results in the following score functions.

$$g_1^*(y, x, \alpha) = x' \frac{[y - \Phi(x\alpha)]\phi(x\alpha)}{\Phi(x\alpha)[1 - \Phi(x\alpha)]} \tag{2.5.22}$$

$$g_2^*(x_1, x_2, \beta) = \begin{bmatrix} exp(-x_{21}\lambda)x_2' \\ h_2(x_2, \beta)x_{21}' \end{bmatrix} \phi[h_2(x_2, \beta)] \frac{\{x_1 - \Phi[h_2(x_2, \beta)]\}\phi[h_2(x_2, \beta)]}{\Phi[h_2(x_2, \beta)]\{1 - \Phi[h_2(x_2, \beta)]\}} \tag{2.5.23}$$

$$g_3^*(y, x_2, \gamma) = \begin{bmatrix} \phi(\alpha_1 + x_2\alpha_2)\Phi[h_2(x_2, \beta)] \\ x_2'\{\phi(\alpha_1 + x_2\alpha_2)\Phi[h_2(x_2, \beta)] + \phi(x_2\alpha_2)\{1 - \Phi[h_2(x_2, \beta)]\}\} \\ x_2' exp(-x_{21}\lambda)\phi[h_2(x_2, \beta)][\Phi(\alpha_1 + x_2\alpha_2) - \Phi(x_2\alpha_2)] \\ x_{21}' h_2(x_2, \beta)\phi[h_2(x_2, \beta)][\Phi(x_2\alpha_2) - \Phi(\alpha_1 + x_2\alpha_2)] \end{bmatrix} h_4(y, x_2, \gamma), \tag{2.5.24}$$

where $h_4(y, x_2, \gamma) \equiv \dfrac{y - h_3(x_2, \gamma)}{h_3(x_2, \gamma)[1 - h_3(x_2, \gamma)]}$.

2.5.1.3 Average partial effects

In a probit, usually the average partial effects (APEs) are the quantities of interest rather than the coefficients themselves. It is important to note that the APEs of interest are still derived from the model of interest in (2.5.1), just as in the case where there is no missing data. The partial effect (PE) of the j^{th} element of x, $x_{(j)}$ on $\mathbb{E}(y|x)$ is given by[6]

$$PE_j(x) = \frac{\partial \mathbb{E}(y|x)}{\partial x_{(j)}} = \alpha_{(j)0}\phi(x\alpha_0) = \alpha_{(j)0}\phi(\alpha_{10}x_1 + x_2\alpha_{20}). \tag{2.5.25}$$

The average partial effect of $x_{(j)}$, APE_j, is the expected value of $PE_j(x)$ with respect to x.

$$APE_j(x) = \mathbb{E}_x\left[\frac{\partial \mathbb{E}(y|x)}{\partial x_{(j)}}\right] = \alpha_{(j)0}\,\mathbb{E}[\phi(x\alpha_0)]. \tag{2.5.26}$$

In the absence of missing data, this can be consistently estimated using

$$\tilde{\alpha}_{(j)}\left[N^{-1}\sum_{i=1}^{N}\phi(x_i\tilde{\alpha})\right], \tag{2.5.27}$$

where $\tilde{\alpha}$ is any consistent estimate of α_0. That is, one simply computes the partial effect for each unit in the sample and then averages over the entire sample.

However, when we have missing data on x_1, this quantity is not estimable as we cannot calculate the partial effect for individuals with missing x_1. A quantity that *is* feasible to compute is the average of partial effects over the complete cases only. This is given by

$$\widehat{APE}_j^c(x) = \hat{\alpha}_{WJ(j)}\left[N_c^{-1}\sum_{i=1}^{N}s_i\phi(x_i\hat{\alpha}_{WJ})\right],$$

where $N_c = \sum_{i=1}^{N} s_i$ is the number of complete cases in the sample. That is, we average the individual partial effects over the complete cases only. This estimator however, is not consistent for $APE_j(x)$ unless $s \perp\!\!\!\perp x$. If s depends on say x_2, then $\widehat{APE}_j^c(x)$ will be inconsistent for $APE_j(x)$.

[6] If $x_{(j)}$ is discrete, the derivative is replaced with a difference.

The current framework, however, makes it possible to recover $APE_j(x)$ using IPW.

$$\mathbb{E}\{[s/p(z)]\phi(x\alpha)\} = \mathbb{E}\{\mathbb{E}([s/p(z)]\phi(x\alpha)|y,x,z)\} = \mathbb{E}\{[\mathbb{E}(s|y,x,z)/p(z)]\phi(x\alpha)\} = \mathbb{E}[\phi(x\alpha)], \tag{2.5.28}$$

where the last equality follows from Assumption 2.3.1. Therefore, a consistent estimator of $APE_j(x)$ is

$$\widehat{APE}_j(x) = \hat{\alpha}_{WJ(j)} N^{-1} \sum_{i=1}^{N} \frac{s_i}{G(z_i,\hat{\delta})} \phi(x_i \hat{\alpha}_{WJ}). \tag{2.5.29}$$

2.5.2 Exponential models

Next we consider exponential models for nonnegative responses y, including but not restricted to count variables. We focus on a continuous x_1.[7] The model of interest is characterized by the conditional mean

$$\mathbb{E}(y|x) = exp(\alpha_{10}x_1 + x_2\alpha_{20}) \equiv exp(x\alpha_0), \tag{2.5.30}$$

where in the absence of missing data, α_0 can be estimated using a Poisson quasi log likelihood. We consider the same linear imputation model as in Section 2.5.1.1.

$$x_1 = x_2\theta_0 + r, \tag{2.5.31}$$

$$r|x_2 \sim Normal[0, \sigma_0^2 exp(2x_{21}\lambda_0)]. \tag{2.5.32}$$

The reduced form conditional mean can be obtained using (2.5.30)-(2.5.32) and an iterated expectations argument.

$$\begin{aligned}\mathbb{E}(y|x_2) &= \mathbb{E}[exp(\alpha_{10}x_1 + x_2\alpha_{20})|x_2] = exp(x_2\alpha_{20})\,\mathbb{E}[exp(\alpha_{10}x_1)|x_2] \\ &= exp[x_2(\theta_0\alpha_{10} + \alpha_{20})]\,\mathbb{E}[exp(r\alpha_{10})|x_2],\end{aligned} \tag{2.5.33}$$

where the third equality follows from substituting for x_1 using (2.5.31). Moreover, (2.5.32) implies that $exp(r\alpha_{10})$ conditional on x_2 follows a lognormal distribution with

$$\mathbb{E}[exp(r\alpha_{10})|x_2] = exp[\alpha_{10}^2\sigma_0^2 exp(2x_{21}\lambda_0)/2]. \tag{2.5.34}$$

[7]The discussion for a binary x_1 follows easily given the discussion in Section 2.5.1.2.

Plugging into (2.5.33), we get

$$\mathbb{E}(y|x_2) = exp[x_2(\theta_0\alpha_{10} + \alpha_{20}) + \alpha_{10}^2\sigma_0^2 exp(2x_{21}\lambda_0)/2]. \tag{2.5.35}$$

Thus, we have $\beta = (\theta, \sigma^2, \lambda), \gamma = (\theta\alpha_1+\alpha_2, \sigma^2, \lambda), h_5(x_2, \gamma) \equiv x_2(\theta\alpha_1+\alpha_2)+\alpha_1^2\sigma^2 exp(2x_{21}\lambda)/2$ and the objective functions are given by

$$\begin{aligned}
f_1(y, x, \alpha) &= exp(x\alpha) - yx\alpha \\
f_2(x_1, x_2, \beta) &= -log\left\{\frac{1}{\sqrt{2\pi\sigma^2 exp(2x_{21}\lambda)}} exp\left[-\frac{(x_1 - x_2\theta)^2}{2\sigma^2 exp(2x_{21}\lambda)}\right]\right\} \\
f_3(y, x_2, \gamma) &= exp[h_5(x_2, \gamma)] - y[h_5(x_2, \gamma)].
\end{aligned} \tag{2.5.36}$$

This results in the following score functions.

$$\begin{aligned}
g_1^*(y, x, \alpha) &= x'[y - exp(x\alpha)] \\
g_2^*(x_1, x_2, \beta) &= \begin{bmatrix} \frac{x_2'(x_1 - x_2\theta)}{\sigma^2 exp(2x_{21}\lambda)} \\ \frac{(x_1 - x_2\theta)^2}{exp(2x_{21}\lambda)\sigma^4} - \frac{1}{\sigma^2} \\ x_{21}'\left[\frac{(x_1 - x_2\theta)^2}{\sigma^2 exp(2x_{21}\lambda)} - 1\right] \end{bmatrix} \\
g_3^*(y, x_2, \gamma) &= \begin{bmatrix} x_2' \\ exp(2x_{21}\lambda) \\ exp(2x_{21}\lambda)x_{21}' \end{bmatrix} \{y - exp[h_5(x_2, \gamma)]\}.
\end{aligned} \tag{2.5.37}$$

Similar to Section 2.5.1.1, when $\lambda_0 = 0$, the third element of $g_2^*(.)$ and the second and third elements of $g_3^*(.)$ become redundant.

2.6 Comparison with related estimators

2.6.1 Complete cases

The most common practice when dealing with missing covariate values is to just use the complete cases for estimation; that is, use only the observations for which x_1 is observed. The inverse

probability weighted complete cases estimator has been discussed in detail by Wooldridge (2002). In this section, I show that the weighted joint GMM does no worse than the weighted complete cases estimator in terms of asymptotic variance, and can potentially provide strict efficiency gains.

Definition 2.6.1.1. *Call the estimator of α_0 that minimizes (2.4.3), where $g(.)$ contains only $g_1(.)$ and $\hat{W} = I$, the weighted complete cases estimator (or $\hat{\alpha}_{Wcc}$).*

Define the upper-left $P_1 \times P_1$ block of F_0 as

$$F_{11}^0 \equiv \mathbb{E}(g_{1i}g_{1i}') - \mathbb{E}(g_{1i}d_i')[\mathbb{E}(d_i d_i')]^{-1}\,\mathbb{E}(d_i g_{1i}'), \tag{2.6.1}$$

where $g_i = [g_{1i}', g_{2i}', g_{3i}']'$. Then the asymptotic variance of the weighted complete cases estimator as derived in Wooldridge (2002) is given in the following lemma, where we have used the fact that D_{11}^0 is symmetric.

Lemma 2.6.1.1 *Under the assumptions of Theorems 4.1 and 4.2,*

$$Avar[\sqrt{N}(\hat{\alpha}_{Wcc} - \alpha_0)] = [D_{11}^0(F_{11}^0)^{-1}D_{11}^0]^{-1}.$$

Then we know that $\hat{\alpha}_{WJ}$ is no less efficient than $\hat{\alpha}_{Wcc}$, since standard GMM theory dictates that a GMM estimator that uses more valid moment conditions is no less efficient.

Proposition 2.6.1.1. *Under the assumptions of Theorem 2.4.1 and 2.4.2,*

$$Avar[\sqrt{N}(\hat{\alpha}_{Wcc} - \alpha_0)] - Avar[\sqrt{N}(\hat{\alpha}_{WJ} - \alpha_0)] \text{ is positive semidefinite.}$$

We can further disaggregate the efficiency gains by α_{10} and α_{20}. In linear models, the "plug-in" imputation estimators, as discussed in the next section, are generally equivalent to the complete cases estimators for α_{10} and may provide some efficiency gains for α_{20}.[8] Abrevaya & Donald (2017) were the first to propose an estimator that provides potential gains for α_{10} as well in the linear case. I extend their result to the case discussed in Section 2.5.1.1 with the simplifying assumption that $\lambda_0 = 0$, and show that efficiency gains are possible for both α_{10} and α_{20}.

[8]For instance, Abrevaya & Donald (2011) show that in the case where both the main model and the imputation model are linear, the plug-in estimator that estimates the main model using ordinary least squares (OLS) or feasible generalized least squares with missing values being replaced by predicted values using a first step OLS is numerically equivalent to the complete cases estimator for α_{10}.

Proposition 2.6.1.2. *Consider the case in Section 2.5.1.1 with $\lambda_0 = 0$. Under the assumptions of Theorems 2.4.1 and 2.4.2,*

1. $Avar[\sqrt{N}(\hat{\alpha}_{1Wcc} - \alpha_{10})] - Avar[\sqrt{N}(\hat{\alpha}_{1WJ} - \alpha_{10})] = L_1' K L_1 \geq 0$

2. $Avar[\sqrt{N}(\hat{\alpha}_{2Wcc} - \alpha_{20})] - Avar[\sqrt{N}(\hat{\alpha}_{2WJ} - \alpha_{20})] = L_2' K L_2 \geq 0$,

where L_1, L_2 and K are matrices defined in the appendix. I show that K is a positive definite matrix and neither L_1 nor L_2 are necessarily zero under the assumptions made so far, and hence it is possible to obtain strict efficiency gains for both α_{10} and α_{20}.

2.6.2 Sequential procedures

Traditionally, imputation is done in two steps using a "plug-in" method (Dagenais, 1973). In the first step, the missing values of x_1 are replaced with predicted values from a regression of x_1 on x_2 and in the second step, the main model is estimated using the observed values as well as the predicted values. Methods like mean imputation,[9] where the missing values are replaced by the sample mean of x_1, can be considered a special case of this method where the first step regression only includes the constant as a covariate.

Definition 2.6.2.1: *Call the estimator of α_0 obtained using the following procedure the plug-in estimator (or $\hat{\alpha}_P$).*
Step 1: Obtain $\hat{\beta}_{Wcc}$ by minimizing (2.4.3) where $g(.)$ contains only $g_2(\beta)$ and $\hat{W} = I$.
Step 2: Estimate α_0 by minimizing (2.4.3) where $g(.)$ contains only $g_1(\tilde{x}_1, x_2, \alpha)$ and $\tilde{x}_{1i} = s_i x_{1i} + (1 - s_i) h(x_{2i}, \hat{\beta}_{Wcc})$ and $h(.)$ is the function defining predicted values.

In the first step, β_0 is consistently estimated using only the complete cases and the missing values of x_1 are replaced with predicted values based on the imputation model. The function $h(.)$ depends on what the imputation model is. For instance, in the linear case, $h(x_{2i}, \hat{\beta}_{Wcc}) = x_{2i}\hat{\beta}_{Wcc}$. We denote this new variable by $\tilde{x}_1$. In the second step, α_0 is estimated by solving the sample counterpart of (2.2.1) with x_1 being replaced by $\tilde{x}_1$.

[9](Little & Rubin, 2002)

While this procedure can be consistent when the model of interest is linear, contrary to prior claims in the literature (DeCanio & Watkins, 1998), it is generally inconsistent when the model of interest is nonlinear in the parameters.[10] This is because under the assumptions made so far, α_0 is generally not a solution to

$$\min_{\alpha \in \mathbb{A}} \mathbb{E}[f_1(y, x_1^*, x_2, \alpha)], \tag{2.6.2}$$

where $x_1^* = sx_1 + (1-s)h(x_2, \beta_0)$.

To see why this procedure is inconsistent, consider the model in Section 2.5.1.1. Suppose y is binary, that is, $y = w$ (and $y^* \equiv w^*$). For simplicity, assume that $\lambda_0 = 0$ and $z = x_2$, that is, the imputation error is homoskedastic and selection is independent of (y, x_1) conditional on x_2. Since $\mathbb{E}(x_1|x_2) = x_2\theta_0$ and θ_0 is consistently estimated by Ordinary Least Squares (OLS) of x_1 on x_2 using the complete cases only (call this estimator $\hat{\theta}_{cc}$), it is tempting to replace the missing values of x_1 by $x_2\hat{\theta}_{cc}$ and estimate α_0 from the probit of y on $\tilde{x}_1 \equiv sx_1 + (1-s)x_2\hat{\theta}_{cc}$ and x_2. Standard two-step M-estimation theory[11] states that for this procedure to be consistent, we require that α_0 uniquely solves

$$\min_{\alpha \in \mathbb{A}} -\mathbb{E}\{y \, log \, \Phi(\alpha_1 x_1^* + x_2\alpha_2) + (1-y) log[1 - \Phi(\alpha_1 x_1^* + x_2\alpha_2)]\}, \tag{2.6.3}$$

where $x_1^* \equiv sx_1 + (1-s)x_2\theta_0$. However, α_0 does not minimize (2.6.3) in general since for that to be true, we would need

$$P(y = 1|sx_1, x_2, s) = \Phi(\alpha_{10} x_1^* + x_2\alpha_{20}). \tag{2.6.4}$$

However, (2.5.2) and (2.5.6) imply

$$\begin{aligned} y^* &= \alpha_{10}[sx_1 + (1-s)x_2\theta_0] + x_2\alpha_{20} + u + (1-s)r\alpha_{10} \\ &\equiv \alpha_{10} x_1^* + x_2\alpha_{20} + u + (1-s)r\alpha_{10}, \end{aligned} \tag{2.6.5}$$

and

$$\mathbb{E}\{1[\alpha_{10} x_1^* + x_2\alpha_{20} + u + (1-s)r\alpha_{10}]|sx_1, x_2, s\} \neq \Phi(\alpha_{10} x_1^* + x_2\alpha_{20}). \tag{2.6.6}$$

[10] This procedure also requires extra caution when the model of interest is nonlinear in the variables, as discussed in Rai (2020).

[11] Wooldridge (2010) Section 17.4.

The core issue is that expectation does not pass through nonlinear operators, in this case the indicator function $1[.]$. In fact, in this example,

$$\begin{aligned}\mathbb{E}(y=1|sx_1,x_2,s) &= P(y=1|sx_1,x_2,s)\\ &= P\{[u+(1-s)r\alpha_{10}] > -(\alpha_{10}x_1^* + x_2\alpha_{20})|sx_1,x_2,s\}\\ &= \Phi\left[\frac{\alpha_{10}x_1^* + x_2\alpha_{20}}{\sqrt{1+(1-s)\alpha_{10}^2\sigma_0^2}}\right],\end{aligned} \tag{2.6.7}$$

since $u+(1-s)r\alpha_{10}|sx_1,x_2,s \sim Normal[0,1+(1-s)\alpha_{10}^2\sigma_0^2]$ under Assumption 2.3.1, which makes the main estimation problem a heteroskedastic probit. The correct log likelihood function is therefore based on (2.6.7), and α_0 is not a solution to (2.6.3).

Proposition 2.6.2.1: *Consider the case in Section 2.5.1.1. Let Assumptions 2.2.1-2.2.3, 2.3.1, 2.3.2 and the assumptions in Theorems 2.4.1 and 2.4.2 hold. Additionally assume that $z = x_2$ and $\lambda_0 = 0$. Then $\hat{\alpha}_P$ is inconsistent for α_{10} unless $\alpha_{10} = 0$.*

However, $\alpha_{10} = 0$ implies that x_1 is irrelevant in the model of interest, in which case the best solution is to just drop it from the model.

As a second example, consider the exponential model from Section 2.5.2 and again for simplicity, assume that $\lambda_0 = 0$ and $z = x_2$. The plug-in method would entail estimating α_0 using Poisson quasi-MLE with the conditional mean function $exp(\alpha_1\tilde{x}_1 + x_2\alpha_2)$. For this estimator to be consistent, we would require that α_0 uniquely solves

$$\min_{\alpha\in\mathbb{A}} \; -\mathbb{E}[y(\alpha_1x_1^* + x_2\alpha_2) - exp(\alpha_1x_1^* + x_2\alpha_2)], \tag{2.6.8}$$

which would be true if

$$\mathbb{E}(y|sx_1,x_2,s) = exp(\alpha_{10}x_1^* + x_2\alpha_{20}). \tag{2.6.9}$$

However, under Assumption 2.3.1, equations (2.5.30) and (2.5.35) imply that

$$\begin{aligned}\mathbb{E}(y|sx_1,x_2,s) &= exp\{\alpha_{10}[sx_1 + (1-s)x_2\theta_0] + x_2\alpha_{20} + (1-s)\alpha_{10}^2\sigma_0^2/2\}\\ &\equiv exp[\alpha_{10}x_1^* + x_2\alpha_{20} + (1-s)\alpha_{10}^2\sigma_0^2/2].\end{aligned} \tag{2.6.10}$$

Since the log likelihood in (2.6.8) is based on an incorrect specification of the conditional mean of y, α_0 will generally not solve (2.6.8).

Proposition 2.6.2.2: *Consider the case in Section 2.5.2. Let Assumptions 2.2.1-2.2.3, 2.3.1, 2.3.2 and the assumptions in Theorems 2.4.1 and 2.4.2 hold. Additionally assume that $z = x_2$ and $\lambda_0 = 0$. Then $\hat{\alpha}_P$ is inconsistent unless $\alpha_{10} = 0$.*

A sequential procedure that *would* be consistent is plugging $\hat{\beta}_{Wcc}$ in $g_3(.)$, and estimating α_0 using $g_1(\alpha)$ and $g_3(\alpha, \hat{\beta}_{Wcc})$ in a joint GMM procedure.

Definition 2.6.2.2: *Call the estimator of α_0 obtained using the following procedure the sequential estimator (or $\hat{\alpha}_{Seq}$).*

Step 1: Obtain $\hat{\beta}_{Wcc}$ by minimizing (2.4.3) where $g(.)$ contains only $g_2(\beta)$ and $\hat{W} = I$.

Step 2: Estimate α_0 by minimizing (2.4.3) where $g(.)$ contains only $g_1(\alpha)$ and $g_3(\alpha, \hat{\beta}_{Wcc})$, and $\hat{W} = \hat{F}^{-1}$, where $\hat{F}^{-1}$ can be obtained using equation (2.4.7) and imposing $\hat{g}_i = [g_{1i}(\tilde{\alpha})' \; g_{2i}(\tilde{\alpha}, \hat{\beta}_{Wcc})']'$, $\tilde{\alpha}$ being a first step consistent estimate of α_0.

Even though $\hat{\alpha}_{Seq}$ is consistent, it is going to be less efficient than $\hat{\alpha}_{WJ}$ because the former does not utilize the correlation between the moment functions $g_1(.)$ and $g_2(.)$. From a GMM perspective, it is well known that a sequential procedure using the same moment conditions is no more efficient than its joint counterpart.

Proposition 2.6.2.3. *Under Assumptions 2.2.1-2.2.3, 2.3.1, 2.3.2, and the assumptions made in Theorems 2.4.1 and 2.4.2,*

$$Avar[\sqrt{N}(\hat{\alpha}_{Seq} - \alpha_0)] - Avar[\sqrt{N}(\hat{\alpha}_{WJ} - \alpha_0)] \text{ is positive semi-definite.}$$

Thus, there is no reason to prefer $\hat{\alpha}_{Seq}$ over $\hat{\alpha}_{WJ}$ other than computational convenience.

2.6.3 Dummy variable method

The dummy variable estimator ($\hat{\alpha}_D$) replaces the missing values of x_1 with zeros and uses an indicator for missingness as an additional covariate. Jones (1996) and Rai (2020) show that the

resulting estimator is generally inconsistent for α_0 in linear models with exogenous and endogenous x_1 respectively. This inconsistency continues to hold in nonlinear models.

Consider again the example in Section 2.5.1.1 with $\lambda_0 = 0$ and $z = x_2$. The DVM would entail doing a probit of y on $(sx_1, 1-s, x_2)$. Analogous to the discussion in Section 2.6.2, this estimator would be consistent if

$$P(y = 1|sx_1, x_2, s) = \Phi[\alpha_{10} sx_1 + (1-s)\theta_{10}\alpha_{10} + x_2\alpha_{20}], \tag{2.6.11}$$

which is not true in general. Too see this, let $x_2 = (1, x_{22})$ and $\theta_0 = (\theta_{10}, \theta'_{20})'$ and note that we can rewrite equation (2.6.7) as

$$P(y = 1|sx_1, x_2, s) = \Phi\left\{\frac{\alpha_{10} sx_1 + (1-s)\theta_{10}\alpha_{10} + (1-s)x_{22}\theta_{20}\alpha_{10} + x_2\alpha_{20}}{\sqrt{1 + (1-s)\alpha_{10}^2\sigma_0^2}}\right\}. \tag{2.6.12}$$

As can be seen from this equation, $\hat{\alpha}_D$ is inconsistent for two reasons. The first issue, which is unique to this method, is that it omits the covariates $(1-s)x_{22}$, leading to endogeneity unless $\alpha_{10} = 0$ and/or $\theta_{20} = 0$. The second issue, which is common with the plug-in method, is that it ignores the scale factor in the denominator which remains unless $\alpha_{10} = 0$.

Proposition 2.6.3.1: *Consider the case in Section 2.5.1.1. Let Assumptions 2.2.1-2.2.3, 2.3.1, 2.3.2 and the assumptions in Theorems 2.4.1 and 2.4.2 hold. Additionally assume that $z = x_2$ and $\lambda_0 = 0$. Then $\hat{\alpha}_D$ is inconsistent unless (i) $\alpha_{10} = 0$ or (ii) $\theta_{20} = \sigma_0^2 = 0$.*

Similar to Section 2.6.2, if $\alpha_{10} = 0$, the best solution is to drop x_1. The second condition requires that both the imputation coefficients and the imputation error variance are zero at the same time, which is not possible.

A second example is the exponential model discussed in Section 2.5.2. Consider again the case where $z = x_2$ and $\lambda_0 = 0$. The DVM would entail using $(sx_1, 1-s, x_2)$ as covariates for a Poisson quasi-MLE, which would be consistent if

$$\mathbb{E}(y|sx_1, x_2, s) = exp[\alpha_{10} sx_1 + (1-s)(\theta_{10}\alpha_{10} + \alpha_{10}^2\sigma_0^2/2) + x_2\alpha_{20}]. \tag{2.6.13}$$

However, we can re-write (2.6.10) as

$$\mathbb{E}(y|sx_1, x_2, s) = exp[\alpha_{10} sx_1 + (1-s)(\theta_{10}\alpha_{10} + \alpha_{10}^2\sigma_0^2/2) + (1-s)x_{22}\theta_{20}\alpha_{10} + x_2\alpha_{20}]. \tag{2.6.14}$$

Similar to the probit case, the DVM omits the covariates $(1-s)x_{22}$ from the above conditional mean function.

Proposition 2.6.3.2: *Consider the case in Section 2.5.2. Let Assumptions 2.2.1-2.2.3, 2.3.1, 2.3.2 and the assumptions in Theorems 2.4.1 and 2.4.2 hold. Additionally assume that $z = x_2$ and $\lambda_0 = 0$. Then $\hat{\alpha}_D$ is inconsistent unless (i) $\alpha_{10} = 0$ or (ii) $\theta_{20} = 0$.*

That is, $\hat{\alpha}_D$ is inconsistent unless x_1 is irrelevant in the model of interest or x_{22} does not help in predicting x_1.

2.6.4 Unweighted estimators

The key to efficiency gains of $\hat{\alpha}_{WJ}$ over $\hat{\alpha}_{Wcc}$ is that the former uses the information in the incomplete cases. Weighting the moment functions in (2.4.1) allows for more flexibility in terms of what variables selection can depend on and estimation of interesting parameters in the presence of misspecification, but that core reason for efficiency gains is independent of weighting. In other words, the joint GMM based on the unweighted version of the moment functions in (2.4.1) will still be more efficient than the unweighted complete cases estimator. These two unweighted estimators are defined below.

Definition 6.4.1: *Call the estimator of α_0 that minimizes (2.4.3) where $g(.) = s \cdot g_1^*(y, x, \alpha)$ and $\hat{W} = I$, the unweighted complete cases estimator, or $\hat{\alpha}_{Ucc}$.*

The unweighted joint estimator is based on the following vector of moment conditions.

$$g_i(\alpha, \beta) = \begin{bmatrix} g_{1i}(\alpha, \beta) \\ g_{2i}(\alpha, \beta) \\ g_{3i}(\alpha, \beta) \end{bmatrix} \equiv \begin{bmatrix} s_i g_{1i}^*(y_i, x_i, \alpha) \\ s_i g_{2i}^*(x_{1i}, x_{2i}, \beta) \\ g_{3i}^*(y_i, x_{2i}, \alpha, \beta) \end{bmatrix}. \tag{2.6.15}$$

For a generic element from the population (y, x_1, x_2, s), denote this vector of moment functions by $g(\alpha, \beta)$. Then the variance-covariance matrix of $g(\alpha, \beta)$ evaluated at the true parameter values is given by

$$C_0 = \mathbb{E}[g(\alpha_0, \beta_0)\ g(\alpha_0, \beta_0)'], \tag{2.6.16}$$

and the optimal GMM estimator based on (2.6.15) is defined as follows.

Definition 2.6.4.2. *Call the estimator of* (α_0, β_0) *that solves*

$$\min_{(\alpha,\beta)\in\mathbb{A}\times\mathbb{B}} \bar{g}(\alpha,\beta)' \ \hat{C}^{-1} \ \bar{g}(\alpha,\beta),$$

the unweighted joint estimator, or $(\hat{\alpha}_{UJ}, \hat{\beta}_{UJ})$*, where* $\bar{g}(\alpha,\beta) = N^{-1}\sum_{i=1}^{N} g_i(\alpha,\beta)$ *and* $\hat{C} \xrightarrow{p} C_0$*.*

I provide the asymptotic distribution of this estimator in Appendix E. The key point to note is that just like $\hat{\alpha}_{WJ}$ is no less efficient than $\hat{\alpha}_{Wcc}$, $\hat{\alpha}_{UJ}$ is no less efficient than $\hat{\alpha}_{Ucc}$.

Proposition 2.6.4.1. *Under the assumptions of Theorems E.1 and E.2,*

$$Avar[\sqrt{N}(\hat{\alpha}_{Ucc} - \alpha_0)] - Avar[\sqrt{N}(\hat{\alpha}_{UJ} - \alpha_0)] \text{ is positive semidefinite.}$$

The proof of this proposition is very similar to that of Proposition 2.6.1.1, and hence is omitted.

The natural question that arises then is whether one should weight when using the joint estimator, and whether $\hat{\alpha}_{WJ}$ is preferred over $\hat{\alpha}_{Ucc}$, which is the most commonly used estimator out of all four.[12] The issue of whether to weight has previously been considered in Wooldridge (2002), but the use of an imputation model here brings in some new issues. In looking at these two alternatives to $\hat{\alpha}_{WJ}$, there are two issues to address: consistency and asymptotic efficiency.

Start with $\hat{\alpha}_{UJ}$. From the point of view of consistency, $\hat{\alpha}_{WJ}$ is always preferred over $\hat{\alpha}_{UJ}$ as the former is always consistent when the latter is, but the converse is not true. This is because while both estimators rule out z containing x_1 to be consistent,[13] $\hat{\alpha}_{WJ}$ allows z to contain y as well as some outside predictors of selection, while $\hat{\alpha}_{UJ}$ does not. A related issue is that of correct specification of the models underlying $f_1(y,x,\alpha)$, $f_2(x_1,x_2,\beta)$, and $f_3(y,x_2,\gamma)$ in (2.2.1)-(2.2.3), by which I mean that $(\alpha_0,\beta_0,\gamma_0)$ characterize a correctly specified feature of $D(y|x)$, $D(x_1|x_2)$ and $D(y|x_2)$ respectively.[14] For instance, this can be a model of a conditional mean, conditional median, conditional distribution, and so on. When $z = x_2$, $\hat{\alpha}_{WJ}$ is always consistent for α_0 and β_0

[12]That is, out of $\hat{\alpha}_{Ucc}$, $\hat{\alpha}_{Wcc}$, $\hat{\alpha}_{UJ}$ and $\hat{\alpha}_{WJ}$.

[13]$\hat{\alpha}_{UJ}$ rules out z containing x_1 because it uses the imputation equation in estimation in addition to the main equation. Since unweighted estimators can only allow selection to depend on covariates in order to maintain consistency, x_1 being the outcome variable in the imputation model means that we cannot allow s to depend on x_1, conditional on x_2. This is the cost of getting more efficiency using the imputation model. $\hat{\alpha}_{WJ}$ rules out this dependence because the weights cannot be estimated using a variable that contains missing values. Therefore, irrespective of whether one uses the imputation model, weighted estimation cannot allow z to contain x_1.

[14]I make this notion precise in Assumption B.1.

that solve (2.2.1) and (2.2.2) irrespective of whether the underlying models are correctly specified, but $\hat{\alpha}_{UJ}$ is consistent for α_0 and β_0 only if they characterize some correctly specified feature of the respective distributions.

For instance, consider the linear case discussed in Abrevaya & Donald (2017) where the 3 M-estimation problems are given by

$$\min_{\alpha \in \mathbb{A}} \mathbb{E}[s \cdot (y - \alpha_1 x_1 - x_2 \alpha_2)^2] \tag{2.6.17}$$

$$\min_{\beta \in \mathbb{B}} \mathbb{E}[s \cdot (x_1 - x_2 \beta)^2] \tag{2.6.18}$$

$$\min_{\gamma \in \Gamma} \mathbb{E}[(y - x_2 \gamma)^2] \tag{2.6.19}$$

where $\gamma \equiv \alpha_1 \beta + \alpha_2$. Consider first the problem in (2.6.17). Suppose that y is binary with a nonlinear conditional mean $\mathbb{E}(y|x) = \Phi(x\kappa_0)$, and the linear projection of y on x is $x\alpha_0$. When x_1 is always observed, the usual motivation for using a linear model here is that it gives consistent estimates of the linear projection parameters α_0, and linear projection is the best linear approximation to the true conditional mean $\Phi(x\kappa_0)$. That is, the solution to

$$\min_{\alpha \in \mathbb{A}} \mathbb{E}[(y - \alpha_1 x_1 - x_2 \alpha_2)^2] \tag{2.6.20}$$

is α_0.

However, this result does not always carry over to the case with missing data. Suppose s depends on x_2. Then the solution to (2.6.17) will generally neither be κ_0 and more importantly nor be α_0 (Wooldridge, 2002). So by estimating a linear model using only the complete cases, we are not getting consistent estimates of anything interesting in the population.[15]

In general, if we want the solution to

$$\min_{\alpha \in \mathbb{A}} \mathbb{E}[s \cdot f_1(y, x_1, x_2, \alpha)] \tag{2.6.21}$$

to be the conditional mean parameters, we want to make sure that the we have correctly specified the conditional mean. In the above example, one way to do that here is to use a better model of $\mathbb{E}(y|x)$,

[15] An exception is the case where s is independent of both y and x, also known as "missing completely at random". In this case the solution to (6.17) is still α_0. However, this case rarely holds in practice.

that is, a probit instead of a linear probability model. This highlights the importance of nonlinear models with missing data, even if one is generally satisfied with using a linear approximation when x_1 was always observed.

The weighted estimator on the other hand recovers the linear projection parameters even when using only the complete cases. In other words, the solution to

$$\min_{\alpha \in \mathbb{A}} \mathbb{E}\{[s/p(z)](y - \alpha_1 x_1 - x_2 \alpha_2)^2\} \tag{2.6.22}$$

is α_0.

A similar discussion holds for the imputation problem in (2.6.18). If x_1 is binary, then we should either weight the imputation model in order to consistently estimate the linear projection parameters or impute using a probit if not using weights.

The second consideration is that of asymptotic efficiency. When $z = x_2$ and the models underlying $f_1(y, x, \alpha)$, $f_2(x_1, x_2, \beta)$, and $f_3(y, x_2, \gamma)$ are correctly specified, both estimators are consistent. A theoretical comparison of the asymptotic variances of the two estimators in this case will likely depend on whether a generalized conditional information matrix equality (GCIME), discussed in Wooldridge (2002), holds for each of the three models underlying (2.2.1)-(2.2.3). For instance, the GCIME always holds for conditional MLE under correct specification of the conditional density and for quasi-MLE in the linear exponential family under the so-called generalized linear models assumption. Wooldridge (2002) shows that in this case, $\hat{\alpha}_{Ucc}$ is more efficient than $\hat{\alpha}_{Wcc}$ when GCIME holds. So it is reasonable to expect that $\hat{\alpha}_{UJ}$ will be more efficient than $\hat{\alpha}_{WJ}$ as well. I do not undertake a theoretical comparison here but provide some simulation evidence in the next section in support of this speculated efficiency ranking.

The other unweighted alternative to $\hat{\alpha}_{WJ}$ is $\hat{\alpha}_{Ucc}$, and it is not clear from the perspective of consistency whether it is preferred to $\hat{\alpha}_{WJ}$ (or the weighted complete cases estimator $\hat{\alpha}_{Wcc}$). Suppose that α_0 characterizes a correctly specified feature of $D(y|x)$ in (2.2.1). Then if selection is exogenous and depends on x_1 after conditioning on x_2, that is, $z = (x_1, x_2)$, then $\hat{\alpha}_{Ucc}$ is consistent for α_0. However, both the weighted estimators $\hat{\alpha}_{WJ}$ and $\hat{\alpha}_{Wcc}$ are inconsistent. This is because the estimation of weights cannot depend on x_1 which is missing for some observations. Therefore,

the weights will generally not be consistently estimated. However, if selection depends on y after conditioning on x, that is, z contains y, then $\hat{\alpha}_{Ucc}$ is inconsistent while both $\hat{\alpha}_{WJ}$ and $\hat{\alpha}_{Wcc}$ are consistent.

The other consideration is that of correct specification of the model underlying $f_1(y, x, \alpha)$ in (2.2.1). Suppose that $z = x_2$. Then $\hat{\alpha}_{WJ}$ will be consistent for α_0, the solution to (2.2.1), whether or not there is any model misspecification. But under misspecification, $\hat{\alpha}_{Ucc}$ will generally not be consistent for α_0.

For instance, let us go back to the linear model given by (2.6.17)-(2.6.19). Suppose $\mathbb{E}(u|x_1, x_2) = 0$. That is, α_0 in (6.17) are actually the coefficients in the conditional mean of y given (x_1, x_2). If $z = (x_1, x_2)$, then $\hat{\alpha}_{Ucc}$ is consistent for α_0, but $\hat{\alpha}_{WJ}$ and $\hat{\alpha}_{Wcc}$ are inconsistent since the weights can only be based on x_2. If $z = (y, x_2)$, then $\hat{\alpha}_{WJ}$ and $\hat{\alpha}_{Wcc}$ with weights based on (y, x_2) are consistent but $\hat{\alpha}_{Ucc}$ is inconsistent.

On the other hand, suppose $z = x_2$. Then $\hat{\alpha}_{WJ}$ will be consistent for α_0, the linear projection parameters, whether or not they are the conditional mean parameters as well. However, $\hat{\alpha}_{Ucc}$ will be inconsistent for α_0 if they are only the linear projection parameters, and not the conditional mean parameters.

As far as asymptotic efficiency goes, when $z = x_2$ and the model underlying $f_1(y, x, \alpha)$ is correctly specified, both $\hat{\alpha}_{Ucc}$ and $\hat{\alpha}_{WJ}$ are consistent, and we can again expect the efficiency comparison to depend on the GCIME. Again, I do not provide a theoretical comparison but the next section gives some simulation evidence that when the GCIME holds, $\hat{\alpha}_{WJ}$ is still more efficient than $\hat{\alpha}_{Ucc}$ despite of the former being a weighted estimator.

In conclusion, one can choose whether or not to weight when using the joint GMM based on the nature of selection and model specification, but in either case, the joint estimator is no less (and generally more) efficient than its complete cases counterpart.

2.7 Empirical application

I apply the proposed estimation method to the setting of Sandsor (2020), who studies the association between individuals' grade variance and educational attainment. One measure of individuals' cognitive skills is their grades received in school, which are generally summarized using the grade point average (GPA), the mean of the grades. The author looks at the importance of grade variance on educational attainment for a given level of GPA. That is, is it better to specialize in some subjects or to be a "jack-of-all-subjects". She finds that grade variance is negatively associated with educational attainment, that is, students who are jack-of-all-subjects have higher educational attainment.

The data comes from the National Longitudinal Survey of Youth, 1979 (NLSY79). The NLSY79 is a nationally representative sample of 12,686 young men and women between the ages of 14 and 22. Following the author, I only use the sub-sample of 6111 respondents representing the non-institutionalized civilian segment of the population. The data includes high school transcripts, educational attainment, socio-economic characteristics and other measures of cognitive and non-cognitive skills. GPA is measured as the mean of all grades received in upper secondary education (grades 9 to 12). The measure of grade variance is the standard deviation of an individual's grades (GSD). The outcome of interest I consider is whether the individual has a four year college degree at age 30. Again, following the author, I restrict the sample to individuals with at least 10 valid grades and with non-missing data on all variables other than family income in 1979, which is the covariate with missing values I focus on. This leaves me with a sample of 3942 individuals out of which family income is missing for 723 (about 18%) individuals.

I model the relationship between GSD and attainment of a four year college degree as a probit. Since family income is a continuous variable, we are in the general framework of Section 2.5.1.1. The model of interest is given by:

$$y_i = 1[\alpha_{10} linc_i + \alpha_{210} GSD_i + x_{22i}\alpha_{220} + u_i > 0], \tag{2.7.1}$$

$$u_i | x_i \sim Normal(0, 1), \tag{2.7.2}$$

where y_i is a binary variable equal to 1 if individual i has a college degree by the age of 30 and 0 otherwise. $linc_i$ is the log of family income of individual i in 1979, and GSD_i is the grade standard deviation of individual i, the covariate of interest. x_{22i} is the vector of other covariates which includes individual's GPA, gender, race, ethnicity, area of residence, and parental education. It also includes measures of cognitive and noncognitive abilities which are based on the Armed Services Vocational Aptitude Battery (ASVAB) test and a combination of Rotter Locus of Control Scale and Rosenberg Self-Esteem Scale respectively. In our general notation from Section 2.5.1.1, $x_{1i} = linc_i$, $x_{2i} = (GSD_i, x_{22i})$, and $x_i = (x_{1i}, x_{2i})$.

Note that Assumption 2.3.1 in this context states that conditional on x_{2i}, the missingness of $linc_i$ is independent of $linc_i$ itself. This assumption is the basis of many standard procedures used to impute income. For instance, the method of hot decking used by the Current Population Survey is based on this assumption, so is multiple imputation used by the National Health Interview Survey. The standard two-step regression imputation is also based on this assumption.

The imputation model is given by

$$linc_i = x_{2i}\theta_0 + r_i, \quad r_i|x_{2i} \sim Normal(0, \sigma^2). \tag{2.7.3}$$

Table D2 presents the results. Columns 1 and 2 give the coefficient estimates and standard errors from the complete cases probit and the joint GMM respectively. Columns 3 gives the percentage reduction in standard errors of the joint GMM. The standard errors fall for all coefficients, and quite substantially so for many coefficients. While there is not much gain for the coefficient on log of family income, there is about a 10% reduction in the standard error for GSD_i, the variable of interest. The reduction for coefficients on other variables range from about $7\% - 12\%$. The last row of the table gives the Hansen's J-statistic discussed in Proposition 2.4.1. The null hypothesis of correct specification is not rejected at any reasonable significance level, giving us some confidence in the assumptions underlying the joint GMM.

Columns 4 and 5 give the estimates and standard errors for the plug-in method and DVM respectively. In this particular case, both estimators give quite similar results as the joint GMM estimator, which is not surprising given that the coefficient of log(income) is fairly small in

magnitude. As the simulations suggest, the plug-in estimator performs similarly to CC and the joint GMM in terms of both bias and efficiency when α_{10} is small in magnitude. The DVM also has small biases and a smaller standard deviation than the joint GMM for such values, although that efficiency gain does not seem to be present in this application. Moreover, the joint GMM here still has the additional advantage of providing an overidentification test for the assumptions underlying the imputation procedure.

2.8 Conclusion

I have provided a new method of consistently imputing missing covariate values in nonlinear models. The estimator uses the standard assumptions used in the imputation literature, but unlike other imputation estimators based on classical principles, it is consistent in nonlinear models for both the structural parameters and other quantities of interest like average partial effects. I have provided two practically important examples: fractional and nonnegative responses with binary or continuous CMV. The proposed estimator provides substantial efficiency gains over the complete cases estimator, and as a byproduct of using GMM, the overidentification test provides a way to test the extra restrictions imposed by the imputation estimator compared to the complete cases estimator. I have also provided a comprehensive framework for imputing using a variety of nonlinear models for cases where a linear model might be unrealistic.

I have provided the weighted and unweighted versions of the estimator, both of which provide efficiency gains over their complete cases counterparts. This allows the empirical researcher to choose the version best suited for their particular model and the nature of missingness in their specific data.

CHAPTER 3

EFFICIENT ESTIMATION OF LINEAR PANEL DATA MODELS WITH MISSING COVARIATES*

3.1 Introduction

The problem of missingness is ubiquitous in empirical research. In this paper, we provide some methods to deal with missing covariate values in linear panel data models with unobserved heterogeneity.

Economists use a variety of methods to deal with missing covariate values in panel data. One common method is to just use the “complete cases" - the observations for which all covariates are observed [for instance Cabral et al. (2018), David & Venkateswaran (2019)]. While easy to use, methods based only on complete cases can lead to substantial loss of efficiency when missingness is large because of discarding the potentially useful information in the incomplete cases. This has inspired methods that make use of these incomplete cases. One method used in this regard is the “last observation carried forward" (LOCF), which replaces the missing observations in a given time period with observations from the previous time period [for instance, Doraszelski et al. (2018), Giroud & Rauh (2019)].[1] Another method is the dummy variable method (DVM), which replaces the missing values with zeros and includes an indicator for missingness as an additional covariate in the model [for instance, Antecol et al. (2018)]. A third method we consider is regression imputation. This is a two-step method which in the first step, regresses the covariate with missing values (CMV) on the always-observed covariates using complete cases and uses the estimated coefficients to predict missing values of the CMV. In the second step, it estimates the model of interest using all observations with this “composite" CMV, which consists of both observed and predicted values.[2]

*This chapter is co-authored with Professor Jeffrey Wooldridge.

[1]Sometimes the missing observations are also replaced with observations from the following time period.

[2]Moffitt et al. (2020) use this method for imputing a variable which is used to define a covariate in the model of interest.

In this paper, we consider the issue of proper imputation specifically when using the fixed effects estimator, which is perhaps the most frequently used method to estimate linear panel data models with unobserved heterogeneity. We propose a new method of imputing when using fixed effects that improves upon the performance of the estimators mentioned above. The choice of method comes down to consistency and relative efficiency. The complete cases fixed effects estimator [as described in Wooldridge (2019)] generally requires the least number of assumptions to be consistent. However, as mentioned above, it can be inefficient relative to the estimators that make use of the incomplete cases. LOCF has been shown to be generally biased and inconsistent even under the strongest assumptions on missingness (Lane, 2008). We show that DVM is also generally inconsistent unless some very strong zero restrictions are imposed in the model, including the assumption that the CMV does not contain individual specific unobserved heterogeneity - an assumption generally unlikely to hold in practice. Regression imputation is consistent under less restrictive assumptions than the DVM, but still requires that the CMV does not contain unobserved heterogeneity.

The key contribution of this paper therefore is to propose a new imputation estimator which is consistent under assumptions that are much less restrictive than those required by the estimators above. We do not impose the zero restrictions required by the DVM, allow for unobserved heterogeneity in the CMV, and allow for missingness to depend on the always-observed covariates. We propose imputation methods for the cases of both strict as well as sequential exogeneity of the covariates, the latter allowing for things like lagged dependent variables and feedback effects.

A second contribution we make is proposing a novel variable addition test (VAT) for exogeneity of missingness. The VATs proposed so far in this context have only been able to test for missingness in *other* time periods being uncorrelated with unobservables in a given time period (Wooldridge, 2010). We propose a test for missingness in the *same* time period being uncorrelated with the unobservables in a given time period, which is the kind of exogeneity one is most likely to be concerned about in practice.

The rest of the paper proceeds as follows. Section 2 presents the population model of interest

and the associated assumptions of strict exogeneity of the covariates. Section 3 describes the missing data scheme and the assumptions on the missingness mechanism. Section 4 presents the proposed estimator and its asymptotic distribution. Section 5 compares the proposed estimator to some commonly used alternatives. Section 6 proposes an imputation estimator under sequential exogeneity of the covariates and the novel VAT for the exogeneity of missingness. Section 7 concludes. Proofs and extensions to the cases of missing vectors and time-varying unobserved heterogeneity are given in the appendix.

3.2 Population model

We consider a standard linear model with additive heterogeneity. Assume that an underlying population consists of a large number of units for whom data on T time periods are potentially available. We assume random sampling from this population, and let i denote a random draw. Along with the outcome y_{it} and covariates $x_{it} = [x_{1it} \;\; x_{2it}]$, we also draw scalars c_i and d_i, which are the unobserved heterogeneities in y_{it} and x_{1it} respectively.

The linear model with additive heterogeneity is

$$y_{it} = \beta_1 x_{1it} + x_{2it}\beta_2 + c_i + u_{it} \equiv x_{it}\beta + c_i + u_{it}, \quad t = 1, \ldots, T, \tag{3.2.1}$$

where x_{1it} is a scalar, x_{2it} is a $1 \times k$ vector which includes the constant term[3], and $\beta = [\beta_1 \;\; \beta_2']'$. We are interested in estimators of β that allow for correlation between c_i and the history of the covariates, $\{x_{it} : t = 1, \ldots, T\}$.

We first define the histories of all variables. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iT})$, $\mathbf{x}_{1i} = (x_{1i1}, \ldots, x_{1iT})$, $\mathbf{x}_{2i} = (x_{2i1}, \ldots, x_{2iT})$, and $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})$. We place the following assumption on the idiosyncratic error u_{it} in equation (3.2.1).

Assumption 3.2.1. $\mathbb{E}(\mathbf{x}_t' u_{it}) = 0, \quad t = 1, \ldots, T.$

This is a kind of strict exogeneity assumption of the covariates with respect to the idiosyncratic error. It implies that x_{is} is uncorrelated with u_{it}, $s = 1, \ldots, T$. In other words, the idiosyncratic error at time t is uncorrelated with the covariates in *all* time periods. Note that this assumption

[3]where x_{2it} can include a full set of time dummies, or other aggregate time variables.

does not restrict the relationship between $\mathbf{x}_i$ and the unobserved heterogeneity c_i, which can be arbitrarily correlated.

The model which underlies the gains in efficiency in this paper is the following linear imputation model with unobserved heterogeneity, which explains x_{1it} in terms of x_{2it}.

$$x_{1it} = x_{2it}\pi + d_i + r_{it}. \tag{3.2.2}$$

We impose an assumption analogous to Assumption 3.2.1 on the idiosyncratic error r_{it}.

Assumption 3.2.2: $\mathbb{E}(\mathbf{x}'_{2i} r_{it}) = 0, \quad t = 1, \ldots, T.$

Again, this assumption implies that x_{2is} is orthogonal to the idiosyncratic error r_{it} in every time period $s = 1, \ldots, T$. Moreover, it does not restrict the relation between $\mathbf{x}_{2i}$ and the unobserved heterogeneity d_i.

Using the imputation model which explains x_{1it} in terms of x_{2it}, we are able to obtain a "reduced form" for y_{it} in terms of only x_{2it}. Plugging (3.2.2) in (3.2.1) gives

$$y_{it} = \beta_1(x_{2it}\pi + d_i + r_{it}) + x_{2it}\beta_2 + c_i + u_{it} \equiv x_{2it}\gamma + h_i + v_{it}, \tag{3.2.3}$$

where $\gamma \equiv \beta_1\pi + \beta_2$, $h_i \equiv \beta_1 d_i + c_i$, and $v_{it} \equiv \beta_1 r_{it} + u_{it}$. As we will discuss in Section 3, we allow x_{1it} to contain missing values while assuming that x_{2it} is always observed. Equation (3.2.3) allows us to utilize the observations for which x_{1it} is not observed but y_{it} and x_{2it} are.

Note that Assumptions 3.2.1 and 3.2.2 imply that

$$\mathbb{E}(\mathbf{x}'_{2i} v_{it}) = \mathbb{E}[\mathbf{x}'_{2i}(\beta_1 r_{it} + u_{it})] = 0. \tag{3.2.4}$$

That is, x_{2is} is orthogonal to the idiosyncratic error v_{it} in equation (3.2.3) for all $s = 1, \ldots, T$.

3.3 The missing data mechanism

To allow for unbalanced panels, we introduce a series of selection indicators for each i, $\mathbf{s}_i = \{s_{i1}, \ldots, s_{iT}\}$, where $s_{it} = 1$ if x_{1it} is observed; otherwise $s_{it} = 0$. In this paper, we only allow x_{1it} to contain missing values. Hence, s_{it} indicates whether we have a "complete case" for unit i in period t.

Our main estimation method is based on the well-known fixed effects estimator. Define $T_i = \sum_{q=1}^{T} s_{iq}$ as the total number of time periods for which x_{1it} is observed for individual i. Unlike T, T_i is random, since s_{it} is random for every $t = 1, \ldots, T$. We impose the following assumption on T_i.

Assumption 3.3.1. $P(T_i = 0) = 0$.

This assumption simply says that for every individual i in the population, the probability that their x_{1it} is not observed in *any* time period $t = 1 \ldots, T$ is zero.

Further, define the time-demeaned covariates as $\ddot{x}_{it} = x_{it} - T_i^{-1} \sum_{q=1}^{T} s_{iq} x_{iq}$, where the time demeaning here has been done using the complete cases only. We can write $\ddot{x}_{it} = [\ddot{x}_{1it} \quad \ddot{x}_{2it}]$, where $\ddot{x}_{1it} = x_{1it} - T_i^{-1} \sum_{q=1}^{T} s_{iq} x_{1iq}$, and $\ddot{x}_{2it} = x_{2it} - T_i^{-1} \sum_{q=1}^{T} s_{iq} x_{2iq}$. Moreover, $\dot{x}_{2it} = x_{2it} - (T - T_i)^{-1} \sum_{q=1}^{T} (1 - s_{iq}) x_{iq}$ are the time demeaned covariates where the time demeaning has been done using the incomplete cases only. Under Assumption 3.3.1, $\ddot{x}_{it}$ and $\dot{x}_{2it}$ are well defined.[4]

For consistent estimation in the selected samples using fixed effects, we impose the following assumptions on the population distribution.

Assumption 3.3.2. For every $t = 1, \ldots, T$, (i) $\mathbb{E}(s_{it} \ddot{x}'_{it} u_{it}) = 0$ (ii) $\mathbb{E}(s_{it} \ddot{x}'_{2it} r_{it}) = 0$ (iii) $\mathbb{E}[(1 - s_{it}) \dot{x}'_{2it} v_{it}] = 0$.

One case where this assumption would hold is when $\mathbf{s}_i \perp\!\!\!\perp (\mathbf{x}_i, \mathbf{u}_i, \mathbf{r}_i, c_i, d_i)$. That is, selection is independent of everything else in the model, a case we will call "missing completely at random" (MCAR). For instance, data will be MCAR when we have a randomly rotating panel. Then, part (i) of Assumption 3.3.2 becomes

$$\begin{aligned}
\mathbb{E}(s_{it} \ddot{x}'_{it} u_{it}) &= \mathbb{E}(s_{it} x'_{it} u_{it}) - \mathbb{E}(s_{it} T_i^{-1} \sum_{q=1}^{T} s_{iq} x'_{iq} u_{it}) \\
&= \mathbb{E}(s_{it}) \, \mathbb{E}(x'_{it} u_{it}) - \sum_{q=1}^{T} \mathbb{E}(s_{it} T_i^{-1} s_{iq}) \, \mathbb{E}(x'_{iq} u_{it}) \\
&= 0. \qquad (3.3.1)
\end{aligned}$$

The third equality follows from Assumption 3.2.1 under which $\mathbb{E}(\mathbf{x}'_i u_{it}) = 0$. Similarly, part (ii) of

[4]So are all other time demeaned variables defined in Section 3.

Assumption 3.3.2 becomes

$$\begin{aligned}\mathbb{E}(s_{it}\ddot{x}'_{2it}r_{it}) &= \mathbb{E}(s_{it}x'_{2it}r_{it}) - \mathbb{E}(s_{it}T_i^{-1}\sum_{q=1}^{T} s_{iq}x'_{2iq}r_{it}) \\ &= \mathbb{E}(s_{it})\,\mathbb{E}(x'_{2it}r_{it}) - \sum_{q=1}^{T}\mathbb{E}(s_{it}T_i^{-1}s_{iq})\,\mathbb{E}(x'_{2iq}r_{it}) \\ &= 0. \end{aligned} \tag{3.3.2}$$

The third equality follows from Assumption 3.2.2 under which $\mathbb{E}(\mathbf{x}'_{2i}r_{it}) = 0$. As we will see in Section 4, time demeaning using complete cases gets rid of the unobserved heterogeneities c_i and d_i in equations (3.2.1) and (3.2.2) respectively. Therefore, Assumption 3.3.2 does not put any restrictions on the unobserved heterogeneities, and we do not need selection to be independent of the unobserved heterogeneities for this assumption to hold. So along with Assumptions 3.2.1 and 3.2.2, MCAR is sufficient for Assumption 3.3.2 to hold, but we can get by with the weaker assumption $\mathbf{s}_i \perp\!\!\!\perp (\mathbf{x}_i, \mathbf{u}_i, \mathbf{r}_i)$.[5]

We can also allow selection to be a function of the always-observed covariates x_{2it} or unobserved random variables outside the model, but we have to strengthen the exogeneity Assumptions 3.2.1 and 3.2.2 to the following zero conditional mean assumptions.

Assumption 3.2.1' $\mathbb{E}(u_{it}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, c_i, \mathbf{s}_i) = 0, \quad t = 1, \ldots, T.$

Assumption 3.2.1' is a version of strict exogeneity of selection (along with strict exogeneity of the covariates) conditional on c_i. It implies that observing x_{1it} in any time period t cannot be systematically related to the idiosyncratic errors $\mathbf{u}_i$. As a practical matter, Assumption 3.2.1 allows selection s_{it} at time period t to be arbitrarily correlated with $(\mathbf{x}_{1i}, \mathbf{x}_{2i}, c_i)$, that is, with the covariates in any time period and the unobserved heterogeneity in y_{it}.

We also need to strengthen Assumption 3.2.2 to the following zero conditional mean assumption.

Assumption 3.2.2': $\mathbb{E}(r_{it}|\mathbf{x}_{2i}, d_i, \mathbf{s}_i) = 0, \quad t = 1, \ldots, T.$

Assumption 3.2.2' implies that observing x_{1it} in any time period t cannot be systematically related to $\mathbf{r}_i$, where $\mathbf{r}_i = (r_{i1}, \ldots, r_{iT})$. But it can be arbitrarily correlated with $(\mathbf{x}_{2i}, d_i)$, that is,

[5]It is however hard to think of situations where selection is independent of the covariates and the idiosyncratic errors but not the unobserved heterogeneities.

with the always-observed covariates and the unobserved heterogeneity in x_{1it}.

Together, Assumptions 3.2.1' and 3.2.2' allow s_{it} to be arbitrarily correlated with the always-observed covariates $\mathbf{x}_{2i}$, as well as with the unobserved heterogeneity in both y_{it} and x_{1it}, that is, c_i and d_i. But it rules out s_{it} being a function of the idiosyncratic errors $\mathbf{u}_i$ and $\mathbf{r}_i$.

To see that Assumption 3.3.2 holds under Assumptions 3.2.1' and 3.2.2', consider part (i) of Assumption 3.3.2.

$$\mathbb{E}(s_{it}\ddot{x}'_{it}u_{it}) = \mathbb{E}[\mathbb{E}(s_{it}\ddot{x}'_{it}u_{it}|\mathbf{x}_i, \mathbf{s}_i)] = \mathbb{E}[s_{it}\ddot{x}'_{it}\,\mathbb{E}(u_{it}|\mathbf{x}_i, \mathbf{s}_i)] = 0. \tag{3.3.3}$$

The first equality follows from the Law of Iterated Expectations (LIE), and the third follows from the fact that under Assumption 3.2.1', $\mathbb{E}(u_{it}|\mathbf{x}_i, \mathbf{s}_i) = 0$ using the LIE. Similarly, part (ii) of Assumption 3.3.2 becomes

$$\mathbb{E}(s_{it}\ddot{x}'_{2it}r_{it}) = \mathbb{E}[\mathbb{E}(s_{it}\ddot{x}'_{2it}r_{it}|\mathbf{x}_{2i}, \mathbf{s}_i)] = \mathbb{E}[s_{it}\ddot{x}'_{2it}\,\mathbb{E}(r_{it}|\mathbf{x}_{2i}, \mathbf{s}_i)] = 0, \tag{3.3.4}$$

where the third equality follows from the fact that under Assumption 3.2.2', $\mathbb{E}(r_{it}|\mathbf{x}_{2i}, \mathbf{s}_i) = 0$ using the LIE.

3.4 Moment conditions and GMM

It is well known that the fixed effects (within) estimator that uses only the complete cases is generally consistent under Assumption 3.2.1'. One way to characterize this estimator is to multiply equation (3.2.1) through by the selection indicator to get

$$s_{it}y_{it} = \beta_1 s_{it}x_{1it} + s_{it}x_{2it}\beta_2 + s_{it}c_i + s_{it}u_{it}, \quad t = 1, \ldots, T. \tag{3.4.1}$$

Averaging this equation across t for each i gives

$$\bar{y}_i = \beta_1 \bar{x}_{1i} + \bar{x}_{2i}\beta_2 + c_i + \bar{u}_i, \quad t = 1, \ldots, T, \tag{3.4.2}$$

where $\bar{y}_i = T_i^{-1} \sum_{q=1}^{T} s_{iq}y_{iq}$ is the average of the selected observations. The other averages in (3.4.2) are defined similarly. If we now multiply (3.4.2) by s_{it} and subtract from (3.4.1), we remove c_i.

$$s_{it}(y_{it} - \bar{y}_i) = \beta_1 s_{it}(x_{1it} - \bar{x}_{1i}) + s_{it}(x_{2it} - \bar{x}_{2i})\beta_2 + s_{it}(u_{it} - \bar{u}_i), \quad t = 1, \ldots, T. \tag{3.4.3}$$

Equivalently,

$$s_{it}\ddot{y}_{it} = \beta_1 s_{it}\ddot{x}_{1it} + s_{it}\ddot{x}_{2it}\beta_2 + s_{it}\ddot{u}_{it}, \quad t = 1, \ldots, T, \tag{3.4.4}$$

where $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$, $\ddot{u}_{it} \equiv u_{it} - \bar{u}_i$, and $\ddot{x}_{1it}$ and $\ddot{x}_{2it}$ are as defined in Section 3. These are the time-demeaned variables, where the demeaning has been done using the complete cases. Then pooled OLS on (3.4.4) gives consistent estimates of β under part (i) of Assumption 3.3.2. Estimating β using pooled OLS is equivalent to GMM estimation using the following moment conditions.

$$\mathbb{E}[f_{1i}(\beta, \pi)] \equiv \mathbb{E}\left[\sum_{t=1}^{T} s_{it}\ddot{x}'_{it}(\ddot{y}_{it} - \ddot{x}_{1it}\beta_1 - \ddot{x}_{2it}\beta_2)\right] = 0. \tag{3.4.5}$$

These moment conditions give the fixed effects estimator based only on complete cases. Even though this estimator is consistent, it leaves room for gains in efficiency as it ignores the information contained in those observations for which x_{1it} is missing but y_{it} and x_{2it} are observed. In order to utilize those observations, we augment the above moment conditions with those from the imputation model and the reduced form for y_{it}.

We can time demean the imputation model (3.2.2) in a similar fashion as (3.2.1), that is, using the complete cases. This gives

$$s_{it}\ddot{x}_{1it} = s_{it}\ddot{x}_{2it}\pi + s_{it}\ddot{r}_{it}, \quad t = 1, \ldots, T, \tag{3.4.6}$$

where $\ddot{r}_{it} \equiv r_{it} - \bar{r}_i$ and $\bar{r}_i = T_i^{-1}\sum_{q=1}^{T} s_{iq}r_{iq}$. Again, the unobserved heterogeneity d_i is eliminated by the time demeaning. Estimating π using pooled OLS in this equation is equivalent to GMM estimation using the moment functions

$$f_{2i}(\beta, \pi) = \sum_{t=1}^{T} s_{it}\ddot{x}'_{2it}(\ddot{x}_{1it} - \ddot{x}_{2it}\pi). \tag{3.4.7}$$

For the reduced form, we use the incomplete cases to time demean the data. Define

$$\dot{y}_{it} \equiv y_{it} - (T - T_i)^{-1}\sum_{q=1}^{T}(1 - s_{iq})y_{iq}$$

$$\dot{x}_{2it} \equiv x_{2it} - (T - T_i)^{-1}\sum_{q=1}^{T}(1 - s_{iq})x_{2iq}.$$

Then estimating γ using pooled OLS on the equation

$$(1 - s_{it})\dot{y}_{it} = (1 - s_{it})\dot{x}_{2it}\gamma + (1 - s_{it})\dot{v}_{it}, \quad t = 1, \ldots, T \tag{3.4.8}$$

is equivalent to GMM estimation using the following moment functions for the reduced form.

$$f_{3i}(\beta, \pi) = \sum_{t=1}^{T} (1 - s_{it})\dot{x}'_{2it}[\dot{y}_{it} - \dot{x}_{2it}(\beta_1\pi + \beta_2)]. \tag{3.4.9}$$

The full vector of moment functions is given by:

$$f_i(\beta, \pi) = \begin{bmatrix} \sum_{t=1}^{T} s_{it}\ddot{x}'_{it}(\ddot{y}_{it} - \ddot{x}_{1it}\beta_1 - \ddot{x}_{2it}\beta_2) \\ \sum_{t=1}^{T} s_{it}\ddot{x}'_{2it}(\ddot{x}_{1it} - \ddot{x}_{2it}\pi) \\ \sum_{t=1}^{T} (1 - s_{it})\dot{x}'_{2it}\big(\dot{y}_{it} - \dot{x}_{2it}(\beta_1\pi + \beta_2)\big) \end{bmatrix} \equiv \begin{bmatrix} f_{1i}(\beta, \pi) \\ f_{2i}(\beta, \pi) \\ f_{3i}(\beta, \pi) \end{bmatrix}. \tag{3.4.10}$$

Lemma 3.4.1: *Under Assumptions 3.2.1', 3.2.2', 3.3.1 and 3.3.2,* $\mathbb{E}[f_i(\beta, \pi)] = 0$.

This is a set of $3k + 1$ moment conditions with $2k + 1$ parameters, giving us k over-identifying restrictions. It is the availability of these over-identifying restrictions that leads to gains in efficiency in this model. As the following result shows, using either $f_{1i}(.)$ and $f_{2i}(.)$ or $f_{1i}(.)$ and $f_{3i}(.)$ leads to an estimator of β that is identical to the estimator that uses only $f_{1i}(.)$ and hence utilizes only the complete cases.

Lemma 3.4.2: *Under Assumptions 3.2.1', 3.2.2', 3.3.1 and 3.3.2, GMM estimators of* β *based on moment functions* $[f_{1i}(.)'f_{2i}(.)']'$ *or moment functions* $[f_{1i}(.)'f_{3i}(.)']'$ *are identical to that based only on* $f_{1i}(.)$.

Lemma 3.4.2 follows directly from the result in Ahu & Schmidt (1995)[6] that adding equal number of additional parameters and extra moment conditions does not change the GMM estimate of the original parameter. Both $f_{2i}(.)$ and $f_{3i}(.)$ are a set of k moment functions which add k extra parameters π.

To define the GMM estimator based on the entire vector $f_i(.)$, let $\bar{f}(\beta, \pi) = N^{-1}\sum_{i=1}^{N} f_i(\beta, \pi)$, Ω be a square matrix of order $3k + 1$ that is nonrandom, symmetric, and positive definite, and $\hat{\Omega}$ be a first step consistent estimate of Ω. Then the standard two-step GMM minimization problem is

[6]p3. Thoerem 1

given by:

$$\min_{\beta,\pi} \bar{f}(\beta,\pi)' \;\; \hat{\Omega} \;\; \bar{f}(\beta,\pi). \tag{3.4.11}$$

The variance-covariance matrix of the moment functions is given by:

$$C \equiv \mathbb{E}[f_i(\beta,\pi) \;\; f_i(\beta,\pi)'] = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix},$$

where

$$C_{11} = \mathbb{E}\big(\sum_{t=1}^{T} s_{it}\ddot{x}'_{it}u_{it}\sum_{r=1}^{T} s_{ir}\ddot{x}_{ir}u_{ir}\big) \quad C_{12} = \big(\sum_{t=1}^{T} s_{it}\ddot{x}'_{it}u_{it}\sum_{r=1}^{T} s_{ir}\ddot{x}_{2ir}r_{ir}\big)$$

$$C_{13} = \big(\sum_{t=1}^{T} s_{it}\ddot{x}'_{it}r_{it}\sum_{r=1}^{T}(1-s_{ir})\dot{x}_{2ir}v_{ir}\big) \quad C_{22} = \mathbb{E}\big(\sum_{t=1}^{T} s_{it}\ddot{x}'_{2it}r_{it}\sum_{r=1}^{T} s_{ir}\ddot{x}_{2ir}r_{ir}\big)$$

$$C_{23} = \mathbb{E}\big(\sum_{t=1}^{T} s_{it}\ddot{x}'_{2it}r_{it}\sum_{r=1}^{T}(1-s_{ir})\dot{x}_{2ir}v_{ir}\big) \quad C_{33} = \mathbb{E}\big(\sum_{t=1}^{T}(1-s_{it})\dot{x}'_{2it}v_{it}\sum_{r=1}^{T}(1-s_{ir})\dot{x}_{2ir}v_{ir}\big),$$

and $f_i(.)$ is evaluated at the true value of the parameters. The optimal weight matrix is given by the inverse of C_i. Let $\hat{C}$ be a consistent estimate of C.[7] Then the *joint GMM* is defined as follows.

Definition 3.4.1. *Call the estimator of* $[\beta' \;\; \pi']'$ *that solves (3.4.11), where* $\hat{\Omega} = \hat{C}^{-1}$ *the joint GMM estimator (or* $[\hat{\beta}'_{JointFE} \;\; \hat{\pi}'_{JointFE}]'$*).*

Further, define the gradient as follows:

$$D \equiv \mathbb{E}[\nabla f_i(\beta,\Pi)] = \begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \\ D_{31} & D_{32} \end{bmatrix},$$

where

$$D_{11} = -\mathbb{E}\big(\sum_{t=1}^{T} s_{it}\ddot{x}'_{it}\ddot{x}_{it}\big) \quad D_{22} = -\mathbb{E}\big(\sum_{t=1}^{T} s_{it}\ddot{x}'_{2it}\ddot{x}_{2it}\big)$$

$$D_{31} = -\mathbb{E}\big(\sum_{t=1}^{T}(1-s_{it})\dot{x}'_{2it}\dot{x}_{2it}\pi \;\; \sum_{t}(1-s_{it})\dot{x}'_{2it}\dot{x}_{it}\big) \quad D_{32} = -\mathbb{E}\big(\sum_{t=1}^{T}(1-s_{it})\dot{x}'_{2it}\dot{x}_{2it}\beta_1\big).$$

We impose the following rank condition on D for identification of β and π.

[7] which can be obtained by replacing the expectations with sample averages and substituting the estimated errors.

Assumption 3.4.1: $rank(D_{11}) = k + 1$ *and* $rank(D_{22}) = k$.

Under this assumption, $f_{1i}(\beta)$ identifies β and $f_{2i}(\pi)$ identifies π. Then we have the following result using Hansen (1982).

Theorem 3.4.1 *Under standard regularity conditions and Assumptions 3.2.1', 3.2.2', 3.3.1, 3.3.2, and 3.4.1, the estimators* $[\hat{\beta}'_{JointFE}\ \hat{\pi}'_{JointFE}]'$ *are consistent and asymptotically normal, with asymptotic variance given by* $(D'C^{-1}D)^{-1}$*, and*

$$N\ \bar{f}(\hat{\beta}, \hat{\Pi})'\ \hat{C}^{-1}\ \bar{f}(\hat{\beta}, \hat{\Pi}) \xrightarrow{d} \chi^2_k.$$

This statistic can be used for the standard test of over-identifying restrictions. Note that this statistic is just the GMM objective function in (3.4.11) evaluated at the efficient values of the parameters, and is distributed as chi-squared with degrees of freedom equal to the number of over-identifying restrictions.

3.5 Comparison to related estimators

3.5.1 Complete cases estimator

The most common practice in the presence of missing data is to just use the complete cases for estimation; that is, only use the observations for which x_1 is observed. One estimator that uses only complete cases is a GMM estimator based only on $h_{1i}(.)$ which is defined as follows.

Definition 3.5.1.1 *Call the estimator of* β *that solves (3.4.11), where* $f_i(.)$ *contains only* $f_{1i}(.)$ *and* $\hat{\Omega} = I$*, the complete cases estimator (or* $\hat{\beta}_{CC}$*).*

Since $f_{1i}(.)$ is an exactly identified set of moment functions, the weight matrix is irrelevant for this estimation procedure. The asymptotic variance of this estimator is given in the following result.

Lemma 3.5.1.1 *Under Assumptions 3.2.1', 3.3.1, 3.3.2 and 3.4.1, the complete cases estimator* $\hat{\beta}_{CC}$ *has an asymptotic variance given by*

$$Avar[\sqrt{N}(\hat{\beta}_{CC} - \beta)] = (D'_{11}C_{11}^{-1}D_{11})^{-1}.$$

This estimator simply ignores the information in the observations with missing x_1. $\hat{\beta}_{JointFE}$ allows for utilization of this information, leading to potential efficiency gains. The gain in efficiency just follows from the fact that adding valid moment conditions [in this case, $f_{2i}(.)$ and $f_{3i}(.)$] decreases, or at least does not increase, the asymptotic variance of a GMM estimator.

Proposition 3.5.1.1 *Under Assumptions 3.2.1', 3.2.2', 3.3.1, 3.3.2, and 3.4.1,*

$$Avar[\sqrt{N}(\hat{\beta}_{CC} - \beta)] - Avar[\sqrt{N}(\hat{\beta}_{JointFE} - \beta)] \text{ is positive semi-definite.}$$

3.5.2 Dummy variable method

For cross section data, the dummy variable method refers to setting the missing values of the covariate to zero and using an indicator for whether the covariate is missing as an additional covariate. Jones (1996) showed that this generally leads to biased and inconsistent estimates for the case of cross section data.

For panel data, one way the dummy variable method could proceed is the following. Note that using (3.2.1) and (3.2.2), we can write

$$y_{it} = \beta_1[s_{it}x_{1it} + (1 - s_{it})(x_{2it}\pi + d_i + r_{it})] + x_{2it}\beta_2 + c_i + u_{it}. \tag{3.5.1}$$

Now, separating the intercept in the imputation model (3.2.2), we get

$$x_{1it} = \pi_1 + x_{22it}\pi_2 + d_i + r_{it}, \tag{3.5.2}$$

where $x_{2it} = [1 \quad x_{22it}]$. Substituting (3.5.2) in (3.5.1) and rearranging gives

$$y_{it} = \beta_1 s_{it}x_{1it} + \beta_1\pi_1(1 - s_{it}) + x_{2it}\beta_2 + e_{it}, \tag{3.5.3}$$

where $e_{it} \equiv \beta_1(1 - s_{it})(x_{22it}\pi_2 + d_i + r_{it}) + c_i + u_{it}$.

The dummy variable method omits the term $(1 - s_{it})x_{22it}\pi_2\beta_1$ from the model and includes it in the error term. This omitted variable bias is the source of inconsistency of this method, and hence even when the data is missing completely at random, neither POLS nor fixed effects consistently estimates the parameters in the model under the assumptions made so far.

As is expected, POLS on (3.5.3) is additionally inconsistent because e_{it} contains c_i and d_i which are correlated with x_{it}. But even fixed effects estimation of (3.5.3) is additionally inconsistent as it does not get rid of the term $(1 - s_{it})d_i$ in the error, which is correlated with x_{it}. The fixed effects estimator where we time demean using all observations proceeds as follows.

Averaging (3.5.3) across t for each i and then subtracting the averaged equation from (3.5.3) gives

$$\grave{y}_{it} = \beta_1 \grave{s}_{it}\grave{x}_{1it} + \beta_1 \pi_1 (1 - \grave{s}_{it}) + \grave{x}_{2it}\beta_2 + \grave{e}_{it}, \tag{3.5.4}$$

where $\grave{y}_{it} = y_{it} - T^{-1}\sum_{q=1}^{T} y_{iq}$, $\grave{s}_{it}\grave{x}_{1it} = s_{it}x_{1it} - T^{-1}\sum_{q=1}^{T} s_{iq}x_{1iq}$ and so on. Estimating this equation using POLS gives the dummy variable estimator $\hat{\beta}_D$. This estimator is inconsistent unless we impose the restrictions that certain objects are zero in the model.

Proposition 3.5.2.1. *Under Assumptions 3.2.1', 3.2.2', and 3.4.1, $\hat{\beta}_D$ is inconsistent unless (i) $\beta_1 = 0$ or (ii) $\pi_2 = 0$ and $d_i = 0 \; \forall \; i$.*

The first condition is setting $\beta_1 = 0$, which clearly gets rid of both sources of inconsistency in this model. If $\beta_1 = 0$, $\grave{e}_{it} = \grave{u}_{it}$, which is clearly uncorrelated with the regressors in (3.5.3) under Assumption 3.2.1. Intuitively, this condition implies that x_{1it} is irrelevant in model of interest (3.2.1). In this case, the best solution is to drop it and use all observations to estimate β_2 in (3.2.1) using a standard fixed effects estimator that is used when there is no missingness. The second condition implies that first, there is no unobserved heterogeneity in the variable with missing values x_{1it}. As mentioned above, this condition is required because the fixed effects transformation does not get rid of d_i in (3.5.3) because it is now multiplied by $(1 - s_{it})$. But even if $d_i = 0 \; \forall \; i$, this estimator is inconsistent because of omitting the term $(1 - s_{it})x_{22it}\pi_2\beta_1$. Therefore, we need an additional condition that $\pi_2 = 0$, which intuitively means that x_{2it} does not help in predicting x_{1it}.

3.5.3 Regression imputation

Regression imputation is a two-step method which proceeds as following. In the first step, estimate π in (3.2.2) using POLS and complete cases only (call it $\tilde{\pi}$). In the second step, plug $\tilde{\pi}$ in

the equation

$$y_{it} = \omega_1 x^*_{1it} + x_{2it}\omega_2 + error_{it}, \tag{3.5.5}$$

where $x^*_{1it} \equiv s_{it}x_{1it} + (1 - s_{it})x_{2it}\pi$. This is the "composite" x_1 which contains the true values of x_1 when it is observed (i.e. when $s_{it} = 1$) and the predicted values from the imputation equation (3.2.2) when it is missing (i.e. when $s_{it} = 0$). Then estimate ω_1 and ω_2) using fixed effects.

To establish the performance of this estimator, recall that we can write using (3.2.1) and (3.2.2)

$$y_{it} = \beta_1[s_{it}x_{1it} + (1 - s_{it})(x_{2it}\pi + d_i + r_{it})] + x_{2it}\beta_2 + c_i + u_{it}. \tag{3.5.6}$$

This boils down to the model of interest (3.2.1) when $s_{it} = 1$ and to the reduced form (3.2.3) when $s_{it} = 0$. Re-arrange this and write as

$$\begin{aligned} y_{it} &= \beta_1[s_{it}x_{1it} + (1 - s_{it})x_{2it}\pi] + x_{2it}\beta_2 + [(1 - s_{it})d_i\beta_1 + c_i] + [(1 - s_{it})r_{it}\beta_1 + u_{it}] \\ &\equiv \beta_1 x^*_{1it} + x_{2it}\beta_2 + [(1 - s_{it})d_i\beta_1 + c_i] + [(1 - s_{it})r_{it}\beta_1 + u_{it}]. \end{aligned} \tag{3.5.7}$$

Comparing (3.5.7) with (3.5.5), we note that the error in (3.5.5) contains both of the last two terms in (3.5.7), that is, the term that occurs due to the idiosyncratic errors in the model of interest and the imputation model as well as the term that occurs due to the unobserved heterogeneities in the two models. The issue with plugging $\tilde{\pi}$ in (3.5.7) and then estimating using fixed effects is twofold. First, estimating (3.2.2) using POLS and not fixed effects will lead to an inconsistent estimator of π due to the presence of d_i in (3.2.2). Second, and more importantly, even if one gets a consistent estimate of π using fixed effects on (3.2.2) and plugs it in (3.5.7), a standard fixed effects on this equation does not produce consistent estimates of β_1 and β_2 because the unobserved heterogeneity term $[(1 - s_{it})d_i\beta_1 + c_i]$ is not time constant anymore and hence cannot be eliminated by the standard fixed effects transformation. This method is therefore generally going to be inconsistent due to the presence of d_i in the imputation model.

A sequential estimator that *is* consistent is the following. First estimate π using $f_{2i}(.)$, plug the estimated π into $f_{3i}(.)$, and then estimate β using $f_{1i}(.)$ and $f_{3i}(.)$ together.

Definition 3.5.3.1. *Call the following two-step estimator the sequential GMM (or* $[\hat{\beta}'_{Seq}\ \hat{\pi}'_{Seq}]'$*).*

Step 1: Obtain $\hat{\pi}_{Seq}$ by solving (3.4.11), where $f_i(.)$ contains only $f_{2i}(.)$ and $\hat{\Omega} = I$.

Step 2: Obtain $\hat{\beta}_{Seq}$ by solving (3.4.11), where

$$f_i(\beta, \hat{\pi}_{Seq}) = \begin{bmatrix} \sum_{t=1}^{T} s_{it} \ddot{x}'_{it}(\ddot{y}_{it} - \ddot{x}_{1it}\beta_1 - \ddot{x}_{2it}\beta_2) \\ \sum_{t=1}^{T} (1 - s_{it}) \dot{x}'_{2it}\left(\dot{y}_{it} - \dot{x}_{2it}(\beta_1 \hat{\pi}_{Seq} + \beta_2)\right) \end{bmatrix} \equiv \begin{bmatrix} f_{1i}(\beta, \pi) \\ f_{3i}(\beta, \hat{\pi}_{Seq}) \end{bmatrix}$$

and

$$\hat{\Omega} = \left[N^{-1} \sum_{i=1}^{N} f_i(\beta, \hat{\pi}_{Seq})\, f_i(\beta, \hat{\pi}_{Seq})' \right]^{-1}.$$

As is well known, sequential GMM estimators are generally less, or at least no more, efficient than joint GMM estimators that use the same moment conditions. Therefore, $\hat{\beta}_{Seq}$ is generally less efficient than $\hat{\beta}_{JointFE}$[8] and there would be no reason to choose it other than computational convenience.

3.5.4 Mundlak device

In the case of balanced panels, it is well known that the Mundlak device which adds time averages of the covariates as additional explanatory variables in equation (3.2.1) and estimates the model using POLS is numerically equivalent to the fixed effects estimator (Mundlak, 1978). Wooldridge (2019) shows that this numerical equivalence carries over to the case of unbalanced panels as well. In equation (3.2.1), if we include time averages of x_{it} computed using only the complete cases as additional covariates and estimate the model using POLS on complete cases only, then this estimator is numerically equivalent to the complete cases fixed effects estimator $\hat{\beta}_{CC}$.

This suggests an alternative to the joint fixed effects GMM estimator introduced in Section 4. Instead of time demeaning each of the equations (3.2.1)-(3.2.3), we can use the Mundlak device for each of them. Consider first equation (3.2.1) and write

$$c_i = \psi_1 + \xi_{11}\bar{x}_{1i} + \bar{x}_{2i}\xi_{12} + a_{1i} \equiv \psi_1 + \bar{x}_i\xi_1 + a_{1i}. \tag{3.5.8}$$

[8]Prokhorov and Schmidt (2009), Theorem 2.2, part 5.

This is a model that explains the unobserved heterogeneity c_i in terms of the time averages of covariates in equation (3.2.1), where the averaging has been done using the complete cases only. We impose the following zero conditional mean assumption on the error a_{1i}.

Assumption 3.5.4.1. $\mathbb{E}(a_{1i}|\mathbf{x}_i, \mathbf{s}_i) = 0$.

This implies first that $\mathbb{E}(c_i|\bar{x}_i) = \psi_1 + \bar{x}_i\xi_1$. Second, it implies that selection in all time periods is uncorrelated with the error a_{1i}. Plugging (3.5.8) into (3.2.1), we get

$$y_{it} = x_{it}\beta + \psi_1 + \bar{x}_i\xi_1 + a_{1i} + u_{it}. \tag{3.5.9}$$

Let $\acute{x}_{it} = [1 \;\; x_{it} \;\; \bar{x}_i]$. Estimating this model using POLS with the $s_{it} = 1$ observations is equivalent to doing GMM with the following moment functions

$$g_{1i}(\beta, \psi_1, \xi_1) = s_{it}\acute{x}'_{it}(y_{it} - x_{it}\beta - \psi_1 - \bar{x}_i\xi_1). \tag{3.5.10}$$

Similarly, for the unobserved heterogeneity in the imputation model in equation (3.2.2), we can write

$$d_i = \psi_2 + \bar{x}_{2i}\xi_2 + a_{2i}. \tag{3.5.11}$$

Analogous to Assumption 3.5.4.1, we place the following assumption on the error term a_{2i}, which implies that $\mathbb{E}(d_i|\bar{x}_{2i}) = \psi_2 + \bar{x}_{2i}\xi_2$ and that selection in all time periods is uncorrelated with a_{2i}.

Assumption 3.5.4.2. $\mathbb{E}(a_{2i}|\mathbf{x}_{2i}, \mathbf{s}_i) = 0$.

Plugging (3.5.11) into equation (3.2.2), we get

$$x_{1it} = x_{2it}\pi + \psi_2 + \bar{x}_{2i}\xi_2 + a_{2i} + r_{it}. \tag{3.5.12}$$

Let $\acute{x}_{2it} = [1 \;\; x_{2it} \;\; \bar{x}_{2i}]$. Estimating this model using POLS with the $s_{it} = 1$ observations is equivalent to doing GMM with the following moment functions.

$$g_{2i}(\pi, \psi_2, \xi_2) = s_{it}\acute{x}'_{2it}(x_{1it} - x_{2it}\pi - \psi_2 - \bar{x}_{2i}\xi_2). \tag{3.5.13}$$

For the reduced form in equation (3.2.3), we first plug in for the unobserved heterogeneity h_i using (3.5.8) and (3.5.11). Recall that $h_i \equiv \beta_1 d_i + c_i$. We first obtain c_i as a function of $\bar{x}_{2i}$. To do this,

we substitute for $\bar{x}_{1i}$ in (3.5.8) using equation (3.2.2). Averaging (3.2.2) over all time periods for which $s_{it} = 1$, we get

$$\bar{x}_{1i} = \bar{x}_{2i}\pi + d_i + \bar{r}_i. \tag{3.5.14}$$

Plugging in for d_i from (3.5.11) in this equation, we have

$$\bar{x}_{1i} = \bar{x}_{2i}(\pi + \xi_2) + \psi_2 + a_{2i} + \bar{r}_i. \tag{3.5.15}$$

Plugging this into equation (3.5.8),

$$c_i = \psi_1 + \xi_{11}[\bar{x}_{2i}(\pi + \xi_2) + \psi_2 + a_{2i} + \bar{r}_i] + \bar{x}_{2i}\xi_{12} + a_{1i}. \tag{3.5.16}$$

Thus, using equations (3.5.11) and (3.5.16), we can write h_i as

$$h_i \equiv \beta_1 d_i + c_i = \beta_1(\psi_2 + \bar{x}_{2i}\xi_2 + a_{2i}) + \psi_1 + \xi_{11}[\bar{x}_{2i}(\pi + \xi_2) + \psi_2 + a_{2i} + \bar{r}_i] + \bar{x}_{2i}\xi_{12} + a_{1i}. \tag{3.5.17}$$

Plugging this into equation (3.2.3) and re-arranging, we get

$$y_{it} = x_{2it}\gamma + \psi + \bar{x}_{2i}\delta + error_{it}. \tag{3.5.18}$$

where $\psi \equiv \psi_1 + \xi_{11}\psi_2 + \beta_1\psi_2$, $\delta \equiv \xi_{11}(\pi + \xi_2) + \xi_{12} + \beta_1\xi_2$, and $error_{it} \equiv \xi_{11}(a_{2i} + \bar{r}_i) + a_{1i} + \beta_1 a_{2i} + v_{it}$. Estimating this model using POLS with the $s_{it} = 0$ observations is equivalent to doing GMM with the following moment functions.

$$g_{3i}(\beta, \pi, \psi_1, \psi_2, \xi_1, \xi_2) = (1 - s_{it})\acute{x}'_{2it}(y_{it} - x_{2it}\gamma - \psi - \bar{x}_{2i}\delta). \tag{3.5.19}$$

So the final set of moment functions is given by

$$g_i(\beta, \pi, \psi_1, \psi_2, \xi_1, \xi_2) = \begin{bmatrix} \sum_{t=1}^{T} s_{it}\acute{x}'_{it}(y_{it} - x_{it}\beta - \psi_1 - \bar{x}_i\xi_1) \\ \sum_{t=1}^{T} s_{it}\acute{x}'_{2it}(x_{1it} - x_{2it}\pi - \psi_2 - \bar{x}_{2i}\xi_2) \\ \sum_{t=1}^{T}(1 - s_{it})\acute{x}'_{2it}(y_{it} - x_{2it}\gamma - \psi - \bar{x}_{2i}\delta) \end{bmatrix} \equiv \begin{bmatrix} g_{1i}(\beta, \psi_1, \xi_1) \\ g_{2i}(\pi, \psi_2, \xi_2) \\ g_{3i}(\beta, \pi, \psi_1, \psi_2, \xi_1, \xi_2) \end{bmatrix}. \tag{3.5.20}$$

Lemma 3.5.4.1. *Under Assumptions 3.5.4.1 and 3.5.4.2,* $\mathbb{E}[g_i(\beta, \pi, \psi_1, \psi_2, \xi_1, \xi_2)] = 0$.

The rest of the GMM estimation proceeds as usual using the moment conditions in (3.5.21). Define the variance-covariance matrix of the moment functions in (3.5.20) as

$$\Lambda \equiv \mathbb{E}[g_i(\beta, \pi, \psi_1, \psi_2, \xi_1, \xi_2)g_i(\beta, \pi, \psi_1, \psi_2, \xi_1, \xi_2)']. \tag{3.5.21}$$

and let $\hat{\Lambda}$ be a consistent estimate of Λ. Then we define the optimal GMM estimator based on moment conditions (3.5.20) as follows.

Definition 3.5.4.1. *Call the estimator of* $[\beta' \ \pi' \ \psi_1' \ \psi_2' \ \xi_1' \ \xi_2']'$ *that solves*

$$\min_{\beta,\pi} \bar{g}(\beta,\pi,\psi_1,\psi_2,\xi_1,\xi_2)' \ \hat{\Omega} \ \bar{g}(\beta,\pi,\psi_1,\psi_2,\xi_1,\xi_2), \tag{3.5.22}$$

where $\bar{g}(\beta,\pi,\psi_1,\psi_2,\xi_1,\xi_2) = \sum_{i=1}^{N} g_i(\beta,\pi,\psi_1,\psi_2,\xi_1,\xi_2)$ *and* $\hat{\Omega} = \hat{\Lambda}^{-1}$, *the joint Mundlak estimator. Denote the estimator of* β *from this vector as* $\hat{\beta}_{JointMundlak}$.

3.6 Estimation under sequential exogeneity

As is well known, the strict exogeneity Assumption 3.2.1 rules out lagged dependent variables and feedback from past shocks to current covariates in the model of interest (3.2.1).[9] For instance, if x_{it} contains a policy variable, then Assumption 3.2.1 imposes that there is no feedback where policy is more likely to occur based on past shocks. Or if (3.2.1) is a wage equation and one of the covariates is union status, then it rules out a negative wage shock today leading to someone deciding to join the union next year. Assumption 3.2.2' imposes these restrictions on the imputation model (3.2.2).

In order to allow for such effects, we relax Assumption 3.2.1 and 3.2.2 to sequential exogeneity Assumptions 3.6.1 and 3.6.2.

Assumption 3.6.1. $\mathbb{E}(\mathbf{x}_t^{t\prime} u_{it}) = 0, \quad t = 1,\ldots,T,$

where $\mathbf{x}_i^t = (x_{it}, x_{i,t-1}, \ldots, x_{i1})$. This assumes correct distributed lag dynamics but is silent on feedback as it allows for u_{it} to be arbitrarily correlated with $x_{i,t+s}$ for $s \in \{1,\ldots,T-t\}$. For the imputation model (3.2.2), we relax Assumption 3.2.2 to the following.

Assumption 3.6.2: $\mathbb{E}(\mathbf{x}_{2i}^{t\prime} r_{it}) = 0,$

where $\mathbf{x}_{2i}^t = (x_{2it}, x_{2i,t-1}, \ldots, x_{2i1})$.

Under these assumptions, we can use an alternative transformation called "forward orthogonalization" suggested by Arellano & Bover (1995). It demeans data using average over future time

[9] Wooldridge (2010), Chapter 10

periods instead of average over all time periods. It thus preserves sequential exogeneity while still using as much data as possible.

We begin with the model of interest (3.2.1). At time $t \leq T-1$, consider the equations for $t+1, \ldots, T$.

$$y_{i,t+1} = \beta_1 x_{1i,t+1} + x_{2i,t+1}\beta_2 + c_i + u_{i,t+1}$$

$$\vdots$$

$$y_{iT} = \beta_1 x_{1iT} + x_{2iT}\beta_2 + c_i + u_{iT}.$$

In order to time demean (3.2.1), we can naturally use only those future time periods for which x_1 is observed. Define

$$T_i(t) = \sum_{q=t+1}^{T} s_{iq} \tag{3.6.1}$$

as the number of time periods for which x_1 is observed after time t for unit i. Multiply each equation for $t+1 \leq q \leq T$ by s_{iq} and sum

$$\sum_{q=t+1}^{T} s_{iq} y_{iq} = \beta_1 \left(\sum_{q=t+1}^{T} s_{iq} x_{1iq} \right) + \left(\sum_{q=t+1}^{T} s_{iq} x_{2iq} \right) \beta_2 + T_i(t) c_i + \left(\sum_{q=t+1}^{T} s_{iq} u_{iq} \right). \tag{3.6.2}$$

Multiplying through by $T_i(t)^{-1}$ gives

$$\bar{y}_i(t) = \beta_1 \bar{x}_{1i}(t) + \bar{x}_{2i}(t)\beta_2 + c_i + \bar{u}_i(t), \tag{3.6.3}$$

where $\bar{y}_i(t) = T_i(t)^{-1} \sum_{q=t+1}^{T} s_{iq} y_{iq}$ is the average of the observed y_{iq} after time t and $\bar{x}_{1i}(t)$, $\bar{x}_{2i}(t)$ and $\bar{u}_i(t)$ are defined similarly.

Subtracting this equation from (3.2.1), which is the equation at time t gives

$$y_{it} - \bar{y}_i(t) = \beta_1 [x_{1it} - \bar{x}_{1i}(t)] + [x_{2it} - \bar{x}_{2i}(t)]\beta_2 + [u_{it} - \bar{u}_i(t)] \tag{3.6.4}$$

or

$$\tilde{y}_i(t) = \beta_1 \tilde{x}_{1i}(t) + \tilde{x}_{2i}(t)\beta_2 + \tilde{u}_i(t). \tag{3.6.5}$$

Subtracting the forward averages thus eliminates c_i just as with the usual within transformation. Now we use x_{1ip} and x_{2ip}, $p \leq t$ as instrumental variables in this equation, and use only those time

periods for which $s_{it} = 1$, i.e. the complete cases. This gives the following moment functions.

$$m_{1i}(\beta) = \begin{bmatrix} s_{ip} x_{1ip} s_{it} [\tilde{y}_i(t) - \beta_1 \tilde{x}_{1i}(t) - \tilde{x}_{2i}(t)\beta_2] \\ x'_{2ip} s_{it} [\tilde{y}_i(t) - \beta_1 \tilde{x}_{1i}(t) - \tilde{x}_{2i}(t)\beta_2] \end{bmatrix} \quad p \leq t, \quad t = 1, \ldots, T-1. \tag{3.6.6}$$

We require an additional selection indicator for the first set of moment conditions here as in addition to x_{1it}, these moment conditions also require x_{1ip} to be observed for it to be used as an instrumental variable.

Since the moment conditions in (3.6.6) utilize only the complete cases, they leave room for gains in efficiency by utilizing the incomplete cases. We can again implement forward orthogonalization with time demeaning using complete cases to estimate π in (3.2.2). Similar to (3.6.4), we can write

$$x_{1it} - \bar{x}_{1i}(t) = [x_{2it} - \bar{x}_{2i}(t)]\pi + [r_{it} - \bar{r}_i(t)], \tag{3.6.7}$$

where $\bar{r}_i(t) = T_i(t)^{-1} \sum_{q=t+1}^{T} s_{iq} r_{iq}$. Multiplying through by $T_i(t)^{-1}$, we get

$$\tilde{x}_{1i}(t) = \tilde{x}_{2i}(t)\pi + \tilde{r}_i(t). \tag{3.6.8}$$

Using x_{2ip}, $p \leq t$ as instrumental variables and using only the complete cases, we get the moment functions

$$m_{2i}(\pi) = x'_{2ip} s_{it} [\tilde{x}_{1i}(t) - \tilde{x}_{2i}(t)\pi] \quad p \leq t, \quad t = 1, \ldots, T-1. \tag{3.6.9}$$

Similar to Section 4, the moment conditions that allow gains in efficiency come from the reduced form (3.2.3). Here we do the forward orthogonalization using incomplete cases. Let

$$\breve{y}_i(t) = y_{it} - \big(T - t - T_i(t)\big)^{-1} \sum_{q=t+1}^{T} (1 - s_{iq}) y_{iq}$$

$$\breve{x}_{2i}(t) = x_{2it} - \big(T - t - T_i(t)\big)^{-1} \sum_{q=t+1}^{T} (1 - s_{iq}) x_{2iq}.$$

We can then write

$$\breve{y}_i(t) = \breve{x}_{2i}(t)\gamma + \breve{v}_{it}. \tag{3.6.10}$$

We estimate $\gamma \equiv (\beta_1 \pi + \beta_2)$ using incomplete cases as well. This gives moment functions

$$m_{3i}(\beta, \pi) = x'_{2ip}(1 - s_{it})[\breve{y}_i(t) - \breve{x}_{2i}(t)(\beta_1 \pi + \beta_2)] \quad p \leq t, \quad t = 1, \ldots, T-1. \tag{3.6.11}$$

The full set of moment functions is given by

$$m_i(\beta,\pi) = \begin{bmatrix} m_{1i}(\beta) \\ m_{2i}(\pi) \\ m_{3i}(\beta,\pi) \end{bmatrix} = \begin{bmatrix} s_{ip}x_{1ip}s_{it}[\tilde{y}_i(t) - \beta_1\tilde{x}_{1i}(t) - \tilde{x}_{2i}(t)\beta_2] \\ x'_{2ip}s_{it}[\tilde{y}_i(t) - \beta_1\tilde{x}_{1i}(t) - \tilde{x}_{2i}(t)\beta_2] \\ x'_{2ip}s_{it}[\tilde{x}_{1i}(t) - \tilde{x}_{2i}(t)\pi] \\ x'_{2ip}(1-s_{it})[\breve{y}_i(t) - \breve{x}_{2i}(t)(\beta_1\pi + \beta_2)] \end{bmatrix} \quad p \leq t, \quad t = 1,\ldots,T-1. \tag{3.6.12}$$

The moment functions $m_i(\beta,\pi)$ have a zero mean if Assumptions 3.6.1 and 3.6.2 hold and $\mathbf{s}_i \perp\!\!\!\perp (\mathbf{x}_i, \mathbf{u}_i, \mathbf{r}_i)$.[10] However, if we want to allow the selection to be more general (for instance, depend on $\mathbf{x}_{2i}$ or other unobserved variables), we need to strengthen Assumptions 3.6.1 and 3.6.2 to the following zero conditional mean assumptions.

Assumption 3.6.1': $\mathbb{E}(u_{it}|\mathbf{x}_i^t, \mathbf{s}_i, c_i) = 0, \quad t = 1,\ldots,T.$

Assumption 3.6.2': $\mathbb{E}(r_{it}|\mathbf{x}_{2i}^t, \mathbf{s}_i, d_i) = 0, \quad t = 1,\ldots.T.$

Note that although Assumptions 3.6.1' and 3.6.2' allow the covariates to be sequentially exogenous in both the model of interest (3.2.1) and the imputation model (3.2.2), selection is assumed to be strictly exogenous in both models. This is because in the moment functions $m_{1i}(\beta)$ and $m_{2i}(\pi)$, $\bar{y}_i(t)$, $\bar{x}_{1i}(t)$ and $\bar{x}_{2i}(t)$ depend non-linearly on all selection indicators from $t+1$ to T and we use instruments with $p \leq t$. Therefore, we need selection to be strictly, and not just sequentially, exogenous for these moment functions to have a zero mean. Moreover, Assumption 3.6.1' allows selection to be arbitrarily correlated with $\mathbf{x}_i$ and c_i. Assumption 3.6.2' allows selection to be arbitrarily correlated with $\mathbf{x}_{2i}$ and d_i, but it rules out selection depending on x_1 once we condition on x_2. Thus together, Assumptions 3.6.1' and 3.6.2' allow selection to depend on $\mathbf{x}_{2i}$, c_i and d_i, but not $\mathbf{r}_i$ or $\mathbf{u}_i$.

We summarize the conditions under which the moment functions in (3.6.12) have an expected value of zero in the following lemma.

Lemma 3.6.1: $\mathbb{E}[m_i(\beta,\pi)] = 0$ *if either of the following conditions hold.*

(i) $\mathbf{s}_i \perp\!\!\!\perp (\mathbf{x}_i, \mathbf{u}_i, \mathbf{r}_i)$ *and Assumptions 3.6.1 and 3.6.2 hold.*

[10] Recall that this is weaker than MCAR as it allows $\mathbf{s}_i$ to depend on c_i and d_i.

(ii) s_{it} is a function of $\mathbf{x}_{2i}$ or some other random variable w_{it} and Assumptions 3.6.1' and 3.6.2' hold..

Then, $\mathbb{E}[m_i(\beta, \pi)] = 0$ gives us a set of $(3k + 1)T(T - 1)/2$ moment conditions with $2k + 1$ parameters and hence number of over-identifying restrictions depends on T. We can use the regular two-step GMM estimator using these moment conditions.

One way to test for exogeneity of $\mathbf{s}_i$ with respect to $\{u_{it} : t = 1, \ldots, T\}$ is to include selection indicators from other time periods as covariates in equation (3.2.1) and check for their significance at time t. For instance, one might be concerned that a shock today causes people to drop out from the sample in the next time period. Then one can add $s_{i,t+1}$ as a covariate at time t (so that the last time period is lost), estimate the model using the moment conditions in (3.6.6), and compute the robust t-statistic on $s_{i,t+1}$. Another option is to use $s_{i,t-1}$ as a covariate, but that does not work in the case of attrition when it is an absorbing state because if $s_{it} = 1$ for i, then so is $s_{i,t-1}$.

Note that this test can be used even if one is only using the complete cases[11], as it does not even require us to write down the imputation equation (3.2.2). But when using the GMM based on full set of moment conditions in (3.6.12), one can also test for the exogeneity of $\mathbf{s}_i$ with respect to $\{r_{it} : t = 1, \ldots, T\}$ by including $s_{i,t+1}$ as a covariate in the imputation equation (3.2.2) at time t, estimating the model using moment conditions in (3.6.9), and computing the robust t-statistic on $s_{i,t+1}$.

However, what we are most likely to be concerned about in an application is the contemporaneous selection problem, that is, s_{it} being correlated with u_{it}. But one cannot test for s_{it} by including it as a covariate in either (3.2.1) or (3.2.2). This is because both of these models are estimated using complete cases and hence s_{it} will always equal 1 for the observations used in moment conditions in (3.6.6) and (3.6.9). The reduced form in (3.2.3), however, provides a way to test for s_{it} as it can be used for all observations i irrespective of whether $s_{it} = 0$ or $s_{it} = 1$.

Since y_{it} and x_{2it} are observed for all observations, instead of (3.6.10), we can use the following

[11] that is, only the moment conditions in (3.6.6)

moment conditions.

$$\mathbb{E}[x'_{2ip}(\check{y}_{it} - \check{x}_{2it}(\beta_1\pi + \beta_2))] = 0 \quad p \leq t, \quad t = 1, \ldots, T-1. \tag{3.6.13}$$

We have simply removed the $(1 - s_{it})$ from (3.6.10), which means that instead of restricting these moment conditions to the incomplete cases, we are using all observations. Then we can test for the exogeneity of s_{it} with respect to $\{v_{it} : t = 1, \ldots, T\}$ by including s_{it} as a covariate in the reduced form (3.2.3) at time t, estimating the model using the moment conditions in (3.6.12), and computing the robust t-statistic on s_{it}. The null hypothesis here is that s_{it} is uncorrelated with v_{it}. Since $v_{it} = u_{it} + \beta_1 r_{it}$, if we reject the null, then we can conclude that s_{it} is correlated with either u_{it} or r_{it} or both. Since we require both of these correlations to be zero in order for the moment conditions in (3.6.12) to be valid, a rejection would bring the validity of this method into question irrespective of which idiosyncratic error s_{it} is correlated with.

Finally, we can also use this test for s_{it} in the framework of Section 4 where we are assuming strict exogeneity of the covariates with respect to the idiosyncratic errors. In that case, we simply include s_{it} as a covariate in the reduced form (3.2.3) at time t and estimate the model using the following moment conditions

$$\mathbb{E}[\sum_{t=1}^{T} \dot{x}'_{2it}(\dot{y}_{it} - \dot{x}_{2it}(\beta_1\pi + \beta_2))] = 0. \tag{3.6.14}$$

instead of those in (3.6.13), and computing the robust t-statistic on s_{it}. The moment conditions in (3.6.14) are essentially the same as in (3.4.9) except that we have removed the selection indicator $(1 - s_{it})$ just like in the case of sequential exogeneity. Note that all the tests discussed here require that $T \geq 3$.

3.7 Conclusion

We have provided new methods of consistently imputing missing covariate values in linear panel data models with unobserved heterogeneity when using fixed effects. We provide imputation estimators under both strict and sequential exogeneity of the covariates. We relax some substantial assumptions made by currently used imputation estimators, most notably allowing the covariate

with missing values to contain individual specific unobserved heterogeneity. We provide two tests for the assumptions underlying our imputation procedure. The first is a GMM overidentification test which tests the validity of the moment conditions, the second is a novel variable addition test for the missingness in a given time period being uncorrelated with the unobservables, both in the same time period as well as in other time periods.

APPENDICES

APPENDIX A

PROOFS FOR CHAPTER 1

Proof of Proposition 1.5.1.2

We know that

$$Avar(\sqrt{n}[(\hat{\beta}'\ vec(\hat{\Pi})')' - (\beta'\ vec(\Pi)')']) = (D'C^{-1}D)^{-1}$$

Now,

$$\begin{aligned} D'C^{-1}D &= \begin{bmatrix} D_{11}' & 0 \\ 0 & D_{22}' \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} \\ C_{12}' & C_{22} \end{bmatrix}^{-1} \begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \end{bmatrix} + \begin{bmatrix} 0 & D_{41}' \\ D_{32}' & D_{42}' \end{bmatrix} \begin{bmatrix} C_{33} & 0 \\ 0 & C_{44} \end{bmatrix}^{-1} \begin{bmatrix} 0 & D_{32} \\ D_{41} & D_{42} \end{bmatrix} \\ &\equiv G + HIH' \end{aligned}$$

where

$$G = \begin{bmatrix} D_{11}' & 0 \\ 0 & D_{22}' \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} \\ C_{12}' & C_{22} \end{bmatrix}^{-1} \begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \end{bmatrix}, \quad H = \begin{bmatrix} 0 & D_{32} \\ D_{41} & D_{42} \end{bmatrix}, \quad I = \begin{bmatrix} C_{33} & 0 \\ 0 & C_{44} \end{bmatrix}^{-1}$$

Using the matrix inversion lemma,

$$(D'C^{-1}D)^{-1} = (G + HIH')^{-1} = G^{-1} - G^{-1}H(I + H'G^{-1}H)^{-1}H'G^{-1}$$

and thus

$$G^{-1} - (D'C^{-1}D)^{-1} = G^{-1}H(I^{-1} + H'G^{-1}H)^{-1}H'G^{-1} \tag{A.1}$$

Let $E \equiv (I + H'G^{-1}H)^{-1}$. Now,

$$G^{-1} = \begin{bmatrix} D_{11}^{-1}C_{11}D_{11}^{-1\prime} & D_{11}^{-1}C_{12}D_{22}^{-1\prime} \\ D_{22}^{-1}C_{21}D_{11}^{-1\prime} & D_{22}^{-1}C_{22}D_{22}^{-1\prime} \end{bmatrix}$$

and the asymptotic variance of the complete cases GMM is given by the upper left $(k+1) \times (k+1)$ block of G^{-1}. Therefore, the difference between the asymptotic variances of the complete cases estimator and the proposed estimator is given by the upper left $(k+1) \times (k+1)$ block of the

expression on the right hand side of (A.1). For this we need the first $(k+1)$ columns of $H'G^{-1}$, which are given by

$$\begin{bmatrix} D_{32}D_{22}^{-1}C_{21}D_{11}^{-1'} \\ D_{41}D_{11}^{-1}C_{11}D_{11}^{-1'} + D_{42}D_{22}^{-1}C_{21}D_{11}^{-1'} \end{bmatrix} \tag{A.2}$$

For the difference corresponding to β_1, we need the first column of this matrix. To find that, consider

$$\begin{aligned} D_{11}^{-1} &= [\mathbb{E}(s_1 s_2 x' z)]^{-1} \\ &= \begin{bmatrix} J^{-1} & -J^{-1}K_1K_2^{-1} \\ -K_2^{-1}K_4J^{-1} & (K_2 - K_4K_3^{-1}K_1)^{-1} \end{bmatrix} \end{aligned}$$

where $J \equiv \left(\mathbb{E}(s_1 s_2 x_1' z_1) - \mathbb{E}(s_1 s_2 x_1 x_2)[\mathbb{E}(s_1 s_2 x_2' x_2)]^{-1}\,\mathbb{E}(s_1 s_2 x_2' z_1)\right)$, $K_1 \equiv \mathbb{E}(s_1 s_2 x_1 x_2)$, $K_2 \equiv \mathbb{E}(s_1 s_2 x_2' x_2)$, $K_3 \equiv \mathbb{E}(s_1 s_2 x_1 z_1)$, and $K_4 \equiv \mathbb{E}(s_1 s_2 x_2' z_1)$. The first column and the last k columns of this matrix are given by W_1 and W_2 respectively, where

$$W_1 = \begin{bmatrix} 1 \\ -K_2^{-1}K_4 \end{bmatrix} J^{-1} \tag{A.0.1}$$

$$W_2 = \begin{bmatrix} -J^{-1}K_1K_2^{-1} \\ (K_2 - K_4K_3^{-1}K_1)^{-1} \end{bmatrix} \tag{A.0.2}$$

Now, the first column of the matrix in (A.2) is given by

$$\begin{bmatrix} D_{32}D_{22}^{-1}C_{21}W_1 \\ (D_{41}D_{11}^{-1}C_{11} + D_{42}D_{22}^{-1}C_{21})W_1 \end{bmatrix} \equiv \begin{bmatrix} A_1 \\ B_1 \end{bmatrix}$$

Similarly, the last k columns of the matrix in (A.2) are given by

$$\begin{bmatrix} D_{32}D_{22}^{-1}C_{21}W_2 \\ (D_{41}D_{11}^{-1}C_{11} + D_{42}D_{22}^{-1}C_{21})W_2 \end{bmatrix} \equiv \begin{bmatrix} A_2 \\ B_2 \end{bmatrix}$$

Thus, the difference corresponding to β_j, $j = 1, 2$ is

$$\begin{bmatrix} A_j' & B_j' \end{bmatrix} E \begin{bmatrix} A_j \\ B_j \end{bmatrix}$$

as stated in the proposition. ■

Proof of proposition 1.5.2.1

When we have two distinct samples containing (y, z) and (x, z), and hence the estimation is based only on $g_3(.)$ and $g_4(.)$. Thus

$$h(\beta, \Pi) = \left[g_3(\Pi) g_4(\beta, \Pi) \right] \quad C = \begin{bmatrix} C_{33} & 0 \\ 0 & C_{44} \end{bmatrix} \quad D = \begin{bmatrix} 0 & D_{32} \\ D_{41} & D_{42} \end{bmatrix}.$$

The first step solves

$$\frac{1}{n} \sum_{i=1}^{n} g_3(x_i, z_i, s_{2i}, \breve{\Pi}) = 0.$$

By standard GMM theory,

$$\sqrt{n}(\breve{\Pi} - \Pi) \xrightarrow{d} N(0, V_2) \quad \text{where} \quad V_2 = D_{32}^{-1} C_{33} D_{32}^{\prime -1}.$$

The second step solves

$$\min_{\beta} \quad \bar{h}_4(\beta, \breve{\Pi})' \ \breve{\Omega}_1 \ \bar{h}_4(\beta, \breve{\Pi}),$$

where $\bar{h}_4(\beta, \Pi) = \frac{1}{n} \sum_{i=1}^{n} g_4(y_i, z_i, s_{1i}, \beta, \Pi)$. The first order condition is given by

$$\hat{D}_{41} \breve{\Omega}_1 \bar{h}_4(\breve{\beta}, \breve{\Pi}) = 0 \tag{A.3}$$

where $\hat{D}_{41} = \dfrac{\partial \bar{h}_4(\breve{\beta}, \breve{\Pi})}{\partial \beta}$, $\breve{\Omega}_1 \xrightarrow{p} \Omega_1$, and Ω_1 is a general weight matrix. A Taylor expansion of $\bar{h}_4(\breve{\beta}, \breve{\Pi})$ around β gives

$$\bar{h}_4(\breve{\beta}, \breve{\Pi}) = \bar{h}_4(\beta, \breve{\Pi}) + \bar{D}_{41}(\breve{\beta} - \beta),$$

where $\bar{D}_{41} = \dfrac{\partial \bar{h}_4(\bar{\beta}, \breve{\Pi})}{\partial \beta}$ and $\bar{\beta} \in [\beta, \breve{\beta}]$. Substituting in (A.3)

$$\hat{D}_{41} \breve{\Omega}_1 \bar{h}_4(\beta, \breve{\Pi}) + \hat{D}_{41} \breve{\Omega}_1 \bar{D}_{41}(\breve{\beta} - \beta) = 0.$$

Thus,

$$\sqrt{n}(\breve{\beta} - \beta) = -(\hat{D}_{41} \breve{\Omega}_1 \bar{D}_{41})^{-1} \hat{D}_{41} \breve{\Omega}_1 \sqrt{n} \ \bar{h}_4(\beta, \breve{\Pi}).$$

Now, a Taylor expansion of $\bar{h}_4(\beta, \breve{\Pi})$ around Π gives

$$\bar{h}_4(\beta, \breve{\Pi}) = \bar{h}_4(\beta, \Pi) + \bar{D}_{42}(\breve{\Pi} - \Pi),$$

where $\bar{D}_{42} = \dfrac{\partial \bar{h}_4(\beta, \bar{\Pi})}{\partial\, vec\Pi}$ and $\bar{\Pi} \in [\Pi, \breve{\Pi}]$. Thus,

$$\sqrt{n}(\breve{\beta} - \beta) = -(\hat{D}_{41}\breve{\Omega}_1\bar{D}_{41})^{-1}\hat{D}_{41}\breve{\Omega}_1[\sqrt{n}\ \bar{h}_4(\beta, \Pi) + \bar{D}_{42}\sqrt{n}\ (\breve{\Pi} - \Pi)].$$

Now, let $Z \equiv [\sqrt{n}\ \bar{h}_4(\beta, \Pi) + \bar{D}_{42}\sqrt{n}\ (\breve{\Pi} - \Pi)]$. Since

$$\sqrt{n}\ \bar{h}_4(\beta, \Pi) \xrightarrow{d} N(0, C_{44}) \quad \text{and} \quad \sqrt{n}\ (\breve{\Pi} - \Pi) \xrightarrow{d} N(0, V_2),$$

therefore

$$Z \xrightarrow{d} N(0, \Sigma) \quad \text{where} \quad \Sigma = C_{44} + D_{42}V_2D_{42}'.$$

Moreover,

$$\hat{D}_{41} \xrightarrow{p} D_{41} \quad \bar{D}_{41} \xrightarrow{p} D_{41} \quad \bar{D}_{32} \xrightarrow{p} D_{42} \quad \breve{\Omega}_1 \xrightarrow{p} \Omega_1.$$

Let $\breve{\beta} \equiv \hat{\beta}_{TS2SLS-O}$. Then,

$$\begin{aligned}\sqrt{n}(\breve{\beta} - \beta) &= -[(\hat{D}_{41}\breve{\Omega}_1\bar{D}_{41})^{-1}\hat{D}_{41}\breve{\Omega}_1 - (D_{41}\Omega_1D_{41})^{-1}D_{41}\Omega_1]Z - (D_{41}\Omega_1D_{41})^{-1}D_{41}\Omega_1Z \\ &= o_p(1) - (D_{41}\Omega_1D_{41})^{-1}D_{41}\Omega_1Z\end{aligned}$$

where $[(\hat{D}_{41}\breve{\Omega}_1\bar{D}_{41})^{-1}\hat{D}_{41}\breve{\Omega}_1 - (D_{41}\Omega_1D_{41})^{-1}D_{41}\Omega_1]$ is $o_p(1)$ because of the Slutsky's theorem, Z is $O_p(1)$, and $o_p(1).O_p(1) = o_p(1)$. Then, by the asymptotic equivalence lemma,

$$\sqrt{n}(\breve{\beta} - \beta) \xrightarrow{d} N(0, V_1)$$

where

$$V_1 = (D_{41}'\Omega_1D_{41})^{-1}D_{41}'\Omega_1\ \Sigma\ \Omega_1D_{41}(D_{41}'\Omega_1D_{41})^{-1}.$$

By standard GMM theory, the optimal weight matrix for this step is Σ^{-1}. Using this matrix gives

$$V_1^* = (D_{41}'\Sigma^{-1}D_{41})^{-1}.$$

■

Proof of proposition 1.5.2.3

The asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta)$ is given by the upper left $(p + k) \times (p + k)$ block of $(D'C^{-1}D)^{-1}$. Now,

$$(D'C^{-1}D)^{-1} = \left[\begin{bmatrix} 0 & D'_{41} \\ D'_{32} & D'_{42} \end{bmatrix}\begin{bmatrix} C_{33}^{-1} & 0 \\ 0 & C_{44}^{-1} \end{bmatrix}\begin{bmatrix} 0 & D_{32} \\ D_{41} & D_{42} \end{bmatrix}\right]^{-1}$$

$$= \begin{bmatrix} D'_{41}C_{44}^{-1}D_{41} & D'_{41}C_{44}^{-1}D_{42} \\ D'_{42}C_{44}^{-1}D_{41} & D'_{32}C_{33}^{-1}D_{32} + D'_{42}C_{44}^{-1}D_{42} \end{bmatrix}^{-1}$$

Using the formula for the inversion of a block matrix, the upper left $(p + k) \times (p + k)$ block of this inverse is

$$(D'_{41}C_{44}^{-1}D_{41} - D'_{41}C_{44}^{-1}D_{42}(D'_{32}C_{33}^{-1}D_{32} + D'_{42}C_{44}^{-1}D_{42})^{-1}D'_{42}C_{44}^{-1}D_{41})^{-1} \tag{A.4}$$

On the other hand, we know $Avar(\sqrt{n}(\hat{\beta}_{TS2SLS-O} - \beta))$

$$= (D'_{41}\Sigma^{-1}D_{41})^{-1} \tag{A.5}$$

$$= (D'_{41}(C_{44} + D_{42}(D'_{32}C_{33}^{-1}D_{32})^{-1}D'_{42})^{-1}D_{41})^{-1}$$

$$= (D'_{41}(C_{44}^{-1} - C_{44}^{-1}D_{42}(D'_{32}C_{33}^{-1}D_{32} + D'_{42}C_{44}^{-1}D_{42})^{-1}D'_{42}C_{44}^{-1})D_{41})^{-1}$$

$$= (D'_{41}C_{44}^{-1}D_{41} - D'_{41}C_{44}^{-1}D_{42}(D'_{32}C_{33}^{-1}D_{32} + D'_{42}C_{44}^{-1}D_{42})^{-1}D'_{42}C_{44}^{-1}D_{41})^{-1} \tag{A.0.3}$$

where the third equality uses the matrix inversion lemma. The result follows from the fact that (A.4) = (A.5). ■

Proof of proposition 1.5.2.5

With exact identification, $\hat{\beta}$ simply solves

$$\frac{1}{n}\sum_{i=1}^{n} h(y_i, x_i, z_i, s_{1i}, s_{2i}, \hat{\Pi}, \hat{\beta}) = 0 \quad \text{where} \quad h(.) = \begin{bmatrix} g_3(.) \\ g_4(.) \end{bmatrix}.$$

This is the same as first solving

$$\frac{1}{n}\sum_{i=1}^{n} g_3(x_i, z_i, s_{2i}, \breve{\Pi}) = 0$$

for $\breve{\Pi}$, and then solving

$$\frac{1}{n}\sum_{i=1}^{n} g_4(y_i, z_i, s_{1i}, \breve{\Pi}, \hat{\beta}_{TS2SLS}) = 0$$

for $\hat{\beta}_{TS2SLS}$. ■

APPENDIX B

TABLES FOR CHAPTER 1

Table B.1: Monte Carlo simulations, Design 1

	β_1			β_{22}			β_{23}		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Complete cases 2SLS	0.008	0.035	0.036	-0.000	0.066	0.066	-0.023	0.055	0.059
Complete cases GMM	0.008	0.035	0.036	-0.001	0.066	0.066	-0.023	0.055	0.060
Imputation	0.009	0.029	0.031	-0.013	0.056	0.057	-0.011	0.047	0.048
Dummy variable method	0.008	0.035	0.036	0.154	0.065	0.167	0.153	0.054	0.162
Proposed GMM	0.008	0.027	0.028	-0.004	0.051	0.051	-0.011	0.043	0.044

Table B.2: Monte Carlo simulations, Design 2

	β_1			β_{22}			β_{23}		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Complete cases 2SLS	0.013	0.069	0.070	-0.008	0.074	0.074	-0.024	0.065	0.069
Complete cases GMM	0.004	0.052	0.052	-0.005	0.067	0.068	-0.018	0.059	0.062
Imputation	0.008	0.060	0.061	-0.019	0.066	0.069	-0.010	0.058	0.059
Dummy variable method	0.013	0.069	0.070	0.146	0.069	0.161	0.151	0.060	0.163
Proposed GMM	0.010	0.041	0.042	-0.012	0.055	0.056	-0.011	0.049	0.050

Table B.3: Monte Carlo simulations, Design 3

	β_1			β_{22}			β_{23}		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Complete cases 2SLS	0.009	0.076	0.077	0.003	0.083	0.083	-0.018	0.074	0.076
Complete cases GMM	0.002	0.056	0.056	0.004	0.074	0.074	-0.014	0.067	0.068
Imputation	0.006	0.065	0.065	-0.015	0.074	0.075	-0.005	0.064	0.064
Dummy variable method	0.009	0.076	0.077	0.176	0.078	0.193	0.181	0.066	0.193
Proposed GMM	0.010	0.044	0.045	-0.009	0.059	0.060	-0.009	0.053	0.054

Table B.4: Monte Carlo simulations, Design 4

	β_1			β_{22}			β_{23}		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Complete cases 2SLS	0.013	0.069	0.070	-0.008	0.074	0.074	-0.024	0.065	0.069
Complete cases GMM	0.004	0.052	0.052	-0.005	0.067	0.068	-0.018	0.059	0.062
Imputation	0.008	0.060	0.060	-0.015	0.064	0.066	-0.013	0.057	0.058
Dummy variable method	0.013	0.070	0.070	0.001	0.060	0.060	0.003	0.052	0.052
Proposed GMM	0.010	0.040	0.041	-0.011	0.053	0.054	-0.011	0.047	0.049

Table B.5: Monte Carlo simulations, Design 5

	β_1			β_{22}			β_{23}		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Complete cases 2SLS	-0.001	0.118	0.118	-0.006	0.145	0.146	0.023	0.166	0.168
Imputation	-0.001	0.118	0.118	-0.004	0.136	0.136	0.019	0.148	0.149
Proposed GMM	0.000	0.119	0.119	-0.006	0.136	0.136	0.018	0.149	0.150

Table B.6: Monte Carlo simulations, Design 6

	β_1			β_{22}			β_{23}		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Complete cases 2SLS	0.013	0.180	0.180	-0.018	0.171	0.172	0.035	0.192	0.195
Imputation	0.013	0.180	0.180	-0.023	0.174	0.175	0.020	0.185	0.186
Proposed GMM	0.005	0.155	0.155	-0.012	0.150	0.150	0.030	0.165	0.167

Table B.7: Monte Carlo simulations, Design 7

	β_1			β_{22}			β_{23}		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Complete cases 2SLS	0.003	0.198	0.198	-0.014	0.186	0.187	0.039	0.206	0.210
Imputation	0.003	0.198	0.198	-0.014	0.192	0.192	0.030	0.199	0.201
Proposed GMM	0.005	0.162	0.162	-0.012	0.155	0.156	0.031	0.172	0.175

Table B.8: Effect of physician's advice on calorie consumption: complete cases versus the proposed estimator

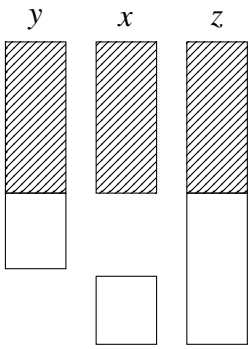
Estimator	Complete cases GMM	Proposed GMM
Physician advised to lose weight	0.126	0.119
	(0.099)	(0.091)
Age	-0.004	-0.004
	(0.00040)	(0.00036)
Female	-0.294	-0.300
	(0.011)	(0.010)
Black	-0.054	-0.056
	(0.013)	(0.011)
Other race	-0.040	-0.142
	(0.013)	(0.011)
9 to 12 years of schooling	0.083	0.085
	(0.024)	(0.021)
High school grad or equivalent	0.074	0.074
	(0.022)	(0.020)
Some college or AA	0.049	0.063
	(0.021)	(0.019)
College or above	0.053	0.060
	(0.023)	(0.021)
Married	-0.015	-0.019
	(0.010)	(0.009)
Has high BP	-0.002	-0.007
	(0.015)	(0.014)
Has high cholesterol	0.005	-0.002
	(0.019)	(0.016)
Has Arthritis	-0.0005	0.006
	(0.013)	(0.012)
Has heart condition	-0.074	-0.073
	(0.025)	(0.023)
Has Diabetes	-0.079	-0.085
	(0.020)	(0.019)
BMI	0.0007	0.0003
	(0.003)	(0.002)
Monthly income < $2100	-0.019	-0.033
	(0.016)	(0.014)
Monthly income between $2100 and $5400	-0.003	-0.013
	(0.014)	(0.012)
Monthly income between $5400 and $8400	-0.017	-0.027
	(0.015)	(0.013)
Is employed	0.086	0.081
	(0.011)	(0.010)

APPENDIX C

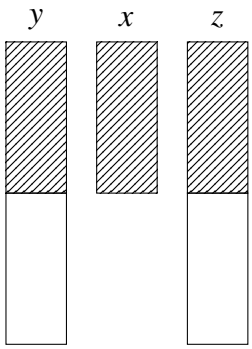
FIGURES FOR CHAPTER 1

Figure C.1: Some admissible patterns of missingness (shaded areas represent complete cases)

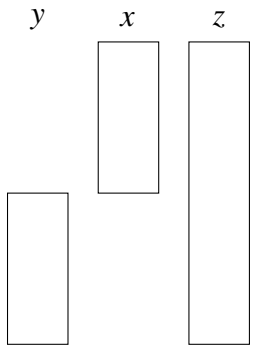
1.1: Partial overlap



1.2: Univariate missing data



1.3: The TS2SLS case



APPENDIX D

PROOFS FOR CHAPTER 2

Proof of Lemma 2.4.1

Since W is nonsingular by assumption, it suffices to show that $\mathbb{E}[g(\alpha, \beta; \delta_0)] \neq 0$ for $(\alpha, \beta) \neq (\alpha_0, \beta_0)$.[1] We show this element-by-element of $g(\alpha, \beta; \delta_0)$.

Starting with the weighted moment functions from the model of interest, given Assumptions 2.3.1 and 2.3.2 and the standard IPW argument, we know that

$$\begin{aligned}\mathbb{E}\{[s/G(z, \delta_0)]g_1^*(y, x, \alpha_0)\} = \mathbb{E}\{[s/p(z)]g_1^*(y, x, \alpha_0)\} &= \mathbb{E}\{\mathbb{E}([s/p(z)]g_1^*(y, x, \alpha_0)|y, x, z)\} \\ &= \mathbb{E}\{[\mathbb{E}(s|y, x, z)/p(z)]g_1^*(y, x, \alpha_0)\} \\ &= \mathbb{E}[g_1^*(y, x, \alpha_0)].\end{aligned}$$

Now, Assumption 2.2.1 implies that $\mathbb{E}[g_1^*(y, x, \alpha)] \neq 0$ for any $\alpha \neq \alpha_0$. It follows that for any $\alpha \neq \alpha_0$,

$$\mathbb{E}\{[s/G(z, \delta_0)]g_1^*(y, x, \alpha)\} \neq 0.$$

Moving on to the imputation model, first note that by iterated expectations,

$$\mathbb{E}(s|x_1, x_2, z) = \mathbb{E}[\mathbb{E}(s|y, x_1, x_2, z)|x_1, x_2, z] = \mathbb{E}[\mathbb{E}(s|z)|x_1, x_2, z] = \mathbb{E}(s|z) \equiv p(z),$$

where the second equality follows from Assumption 2.3.1. Now consider the weighted moment functions from the imputation model.

$$\begin{aligned}\mathbb{E}\{[s/G(z, \delta_0)]g_2^*(x_1, x_2, \beta_0)\} &= \mathbb{E}\{[s/p(z)]g_2^*(x_1, x_2, \beta_0)\} \\ &= \mathbb{E}\{\mathbb{E}([s/p(z)]g_2^*(x_1, x_2, \beta_0)|x_1, x_2, z)\} \\ &= \mathbb{E}\{[\mathbb{E}(s|x_1, x_2, z)/p(z)]g_2^*(x_1, x_2, \beta_0)\} \\ &= \mathbb{E}[g_2^*(x_1, x_2, \beta_0)] = 0\end{aligned}$$

[1] Note that even though $g_3(\alpha, \beta; \delta)$ sometimes only identifies functions of (α_0, β_0) and not each element of (α_0, β_0) separately, the entire vector $g(\alpha, \beta; \delta)$ still identifies (α_0, β_0) separately because $g_1(\alpha; \delta)$ identifies α_0 and $g_2(\beta; \delta)$ identifies β_0.

and the same argument as above applies for identification of β_0 using Assumption 2.2.2.

For the reduced form moment functions, identification of γ_0 simply follows from Assumption 2.2.3.

Proof of Theorem 2.4.1

Identification of (α_0, β_0) follows from Lemma 2.4.1 and $\hat{\delta} \xrightarrow{p} \delta_0$ follows from Assumption 2.3.2 and standard MLE theory. To complete the proof, we simply show that the objective function satisfies the weak uniform law of large numbers. By 5 and 6,

$$|g_1(y, x, z, s, \alpha; \delta_0)| \leq a^{-1} b_1(y, x), \text{ all } (z, s),$$
$$|g_2(x, z, s, \beta; \delta_0)| \leq a^{-1} b_2(x), \text{ all } (z, s),$$
$$|g_3(y, x_2, \gamma)| \leq b_3(y, x_2).$$

and by 6, $\mathbb{E}[b(y, x)] < \infty$, where $g_1(y, x, z, s, \alpha; \delta_0)$, $g_2(x, z, s, \beta; \delta_0)$, and $g_3(y, x_2, \gamma)$ are as defined in (2.4.1). It follows from Lemma 2.4 in Newey and McFadden (1994) that

$$\sup_{(\alpha,\beta,\gamma)\in\mathbb{A}\times\mathbb{B}\times\Gamma} \left\| N^{-1} \sum_{i=1}^{N} g(y_i, x_i, z_i, s_i, \alpha, \beta, \gamma; \hat{\delta}) - \mathbb{E}[g(y, x, z, s, \alpha, \beta, \gamma; \delta_0)] \right\| \xrightarrow{p} 0.$$

The rest of the proof is standard, see Wooldridge (2010, Section 12.4.1).

Proof of Theorem 2.4.2

For notational convenience, let $\tau \equiv (\alpha', \beta')'$. First we will show that

$$\sqrt{N} \nabla_\tau \hat{Q}(\tau_0; \hat{\delta}) \xrightarrow{d} Normal(0, D_0' W F_0 W D_0).$$

Since $\hat{Q}(\tau; \hat{\delta}) = \bar{g}(\tau; \hat{\delta})' \ \hat{W} \ \bar{g}(\tau; \hat{\delta})$,

$$\implies \nabla_\tau \hat{Q}(\tau; \hat{\delta}) = [\nabla_\tau \bar{g}(\tau; \hat{\delta})]' \ \hat{W} \ \bar{g}(\tau; \hat{\delta})$$
$$\implies \sqrt{N} \nabla_\tau \hat{Q}(\tau_0; \hat{\delta}) = [\nabla_\tau \bar{g}(\tau_0; \hat{\delta})]' \ \hat{W} \ \sqrt{N} \bar{g}(\tau_0; \hat{\delta}).$$

Carrying out an element-by-element mean value expansion of $\sqrt{N}\nabla_\tau \hat{Q}(\tau_0;\hat{\delta})$ around δ_0 gives,

$$\begin{aligned}
\sqrt{N}\nabla_\tau \hat{Q}(\tau_0;\hat{\delta}) &= [D_0 + o_p(1)]' \ \hat{W} \ [\sqrt{N}\bar{g}(\tau_0;\delta_0) + \nabla_\delta \bar{g}(\tau_0;\bar{\delta})\sqrt{N}(\hat{\delta}-\delta_0)] \qquad \text{(D.1)} \\
&= [D_0 + o_p(1)]' \ \hat{W} \ \{\sqrt{N}\bar{g}(\tau_0;\delta_0) + [H_0 + o_p(1)]\sqrt{N}(\hat{\delta}-\delta_0)\} \\
&= [D_0 + o_p(1)]' \ \hat{W} \ \{\sqrt{N}\bar{g}(\tau_0;\delta_0) + [H_0 + o_p(1)][N^{-\frac{1}{2}}\sum_{i=1}^{N}\psi(s_i,z_i) + o_p(1)]\} \\
&= D_0' \ W \ \{N^{-\frac{1}{2}}\sum_{i=1}^{N}[g_i + H_0\psi(s_i,z_i)]\} + o_p(1), \qquad \text{(D.0.1)}
\end{aligned}$$

where $\bar{\delta}$ lies between $\hat{\delta}$ and δ_0 (thus $\bar{\delta} \xrightarrow{p} \delta_0$), $H_0 \equiv \mathbb{E}[\nabla_\delta g(\tau_0,\delta_0)]$ and $\psi(s_i,z_i) = -[\mathbb{E}(d_i d_i')]^{-1}d_i$ is the influence function for $\sqrt{N}(\hat{\delta}-\delta_0)$. Moreover, by central limit theorem,

$$N^{-\frac{1}{2}}\sum_{i=1}^{N}[g_i + H_0\psi(s_i,z_i)]\} \xrightarrow{d} N(0,F_0),$$

where

$$F_0 \equiv \mathbb{E}[g_i g_i' + H_0\psi(s_i,z_i)g_i' + g_i\psi(s_i,z_i)'H_0' + H_0\psi(s_i,z_i)\psi(s_i,z_i)'H_0'].$$

Now, note that by definition,

$$H_0 = \mathbb{E}\begin{bmatrix} -(s_i/G_i)g_{1i}^*(\nabla_\delta G_i/G_i) \\ -(s_i/G_i)g_{2i}^*(\nabla_\delta G_i/G_i) \\ 0 \end{bmatrix} = -\mathbb{E}(\tilde{g}_i d_i'),$$

where $\tilde{g}_i \equiv (g_{1i}', g_{2i}', 0)'$ and the third element is a $1 \times L_3$ zero vector. This is because

$$\begin{aligned}
\mathbb{E}(\tilde{g}_i d_i') &= \mathbb{E}\begin{bmatrix} (s_i/G_i)g_{1i}^*\{s_i(\nabla_\delta G_i/G_i) - (1-s_i)[\nabla_\delta G_i/(1-G_i)]\} \\ (s_i/G_i)g_{2i}^*\{s_i(\nabla_\delta G_i/G_i) - (1-s_i)[\nabla_\delta G_i/(1-G_i)]\} \\ 0 \end{bmatrix} \\
&= \mathbb{E}\begin{bmatrix} (s_i/G_i)g_{1i}^*(\nabla_\delta G_i/G_i) \\ (s_i/G_i)g_{2i}^*(\nabla_\delta G_i/G_i) \\ 0 \end{bmatrix},
\end{aligned}$$

since $s_i^2 = s_i$ and $(1-s_i)^2 = (1-s_i)$. This implies

$$\mathbb{E}[H_0\psi(s_i,z_i)g_i'] = -\mathbb{E}(\tilde{g}_i d_i')[\mathbb{E}(d_i d_i')]^{-1}\mathbb{E}(d_i g_i'),$$

and

$$\mathbb{E}[H_0\psi(s_i,z_i)\psi(s_i,z_i)'H_0'] = \mathbb{E}(\tilde{g}_i d_i')[\mathbb{E}(d_i d_i')]^{-1}\,\mathbb{E}(d_i\tilde{g}_i').$$

Therefore,

$$F_0 = \mathbb{E}(g_i g_i') - \{\mathbb{E}(g_i d_i')[\mathbb{E}(d_i d_i')]^{-1}\,\mathbb{E}(d_i g_i') \circ R\},$$

where R is a square matrix of order $L_1 + L_2 + L_3$ with all elements being unity except the lower right $L_3 \times L_3$ block which is a 0 matrix, and $\circ$ denotes Hadamard product. Then using (D.1) and the asymptotic equivalence lemma,

$$\sqrt{N}\nabla_\tau \hat{Q}(\tau_0;\hat{\delta}) \xrightarrow{d} Normal(0, D_0' W F_0 W D_0). \tag{D.2}$$

Next, an element-by-element mean value expansion of $\nabla_\tau \hat{Q}(\hat{\tau};\hat{\delta})$ around τ_0 gives,

$$\nabla_\tau \hat{Q}(\hat{\tau};\hat{\delta}) = \nabla_\tau \hat{Q}(\tau_0;\hat{\delta}) + [D_0' W D_0 + o_p(1)](\hat{\tau} - \tau_0)$$

$$\implies \sqrt{N}(\hat{\tau} - \tau_0) = -(D_0' W D_0)^{-1}\sqrt{N}\nabla_\tau \hat{Q}(\tau_0;\hat{\delta}) + o_p(1). \tag{D.3}$$

Combining (D.2) and (D.3) and using the asymptotic equivalence lemma gives

$$\sqrt{N}(\hat{\tau} - \tau_0) \xrightarrow{d} Normal[0, (D_0' W D_0)^{-1} D_0' W F_0 W D_0 (D_0' W D_0)^{-1}], \tag{D.0.2}$$

which is the desired result.

Proof of Proposition 2.4.1

For notational convenience, let $\hat{\tau}_{WJ} \equiv (\hat{\alpha}_{WJ}', \hat{\beta}_{WJ}')'$. We want to show that under the null hypothesis , $N\,\hat{Q}(\hat{\tau}_{WJ};\hat{\delta}) \xrightarrow{d} \chi^2_{L_3}$, where $\hat{W} = \hat{F}^{-1}$.

First note that a mean value expansion around δ_0 yields

$$\sqrt{N}\bar{g}(\tau_0;\hat{\delta}) = \sqrt{N}\bar{g}(\tau_0;\delta_0) + \nabla_\delta \bar{g}(\tau_0;\bar{\delta})\sqrt{N}(\hat{\delta} - \delta_0) \xrightarrow{d} N(0, F_0).$$

by equation (A.9). This implies

$$-F_0^{-\frac{1}{2}}\sqrt{N}\bar{g}(\tau_0;\hat{\delta}) = U_N \xrightarrow{d} U \sim Normal(0, I). \tag{D.4}$$

Moreover, the first order conditions for the objective function in (4.3) imply that

$$\begin{aligned}
&\sqrt{N}\nabla_\tau \hat{Q}(\hat{\tau};\hat{\delta}) = [\nabla_\tau \bar{g}(\hat{\tau};\hat{\delta})]' \;\; \hat{F}^{-1} \;\; \sqrt{N}\bar{g}(\hat{\tau};\hat{\delta}) = 0 \qquad \text{(D.5)}\\
&\implies D_0' F_0^{-1}\sqrt{N}\bar{g}(\hat{\tau};\hat{\delta}) + o_p(1) = 0\\
&\implies D_0' F_0^{-1}[-F_0^{\frac{1}{2}} U_N + D_0\sqrt{N}(\hat{\tau}-\tau_0)] + o_p(1) = 0\\
&\implies \sqrt{N}(\hat{\tau}-\tau_0) = (D_0' F_0^{-1} D_0)^{-1} D_0' F_0^{-\frac{1}{2}} U_N + o_p(1). \qquad \text{(D.0.3)}
\end{aligned}$$

Now, a mean value expansion of the sample moments around τ_0 gives

$$\sqrt{N}\bar{g}(\hat{\tau};\hat{\delta}) = \sqrt{N}\bar{g}(\tau_0;\hat{\delta}) + \nabla_\tau \bar{g}(\bar{\tau};\hat{\delta})\sqrt{N}(\hat{\tau}-\tau_0), \qquad \text{(D.6)}$$

where $\bar{\tau}$ lies between $\hat{\tau}$ and τ_0. Substituting (D.4) and (D.5) into (D.6), we get

$$\begin{aligned}
\sqrt{N}\bar{g}(\hat{\tau};\hat{\delta}) &= -F_0^{-\frac{1}{2}} U_N + D_0(D_0' F_0^{-1} D_0)^{-1} D_0' F_0^{-\frac{1}{2}} U_N + o_p(1)\\
&= -F_0^{-\frac{1}{2}} R_0 U_N + o_p(1),
\end{aligned}$$

where $R_0 = I - F_0^{-\frac{1}{2}} D_0(D_0' F_0^{-1} D_0)^{-1} D_0' F_0^{-\frac{1}{2}}$ is idempotent of rank L_3. Then,

$$N\,\hat{Q}(\hat{\tau};\hat{\delta}) = U_N' R_0 U_N + o_p(1) \xrightarrow{d} \chi^2_{L_3}.$$

Proof of Proposition 2.6.1.1

I drop the 0 subscripts/superscripts for notational convenience, but all expressions in this proof are evaluated at the true values of the parameters, that is, at $(\alpha_0, \beta_0, \gamma_0)$.

First note that the GMM estimator of α_0 that minimizes (2.4.3) with

$$g(\alpha,\beta;\hat{\delta}) = [g_1(\alpha;\hat{\delta})', g_2(\beta;\hat{\delta})']'$$

and $\hat{W} = I$ is numerically equivalent to $\hat{\alpha}_{Wcc}$, which is based only on $g_1(\alpha;\hat{\delta})$. This is because $g_2(\beta;\hat{\delta})$ simply adds equal number of parameters to be estimated and moment conditions to the system.[2] To characterize the asymptotic variance of this estimator, first define the following

[2] Ahu & Schmidt (1995)

quantities.

$$D_1 \equiv \begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \end{bmatrix} \quad D_2 \equiv \begin{bmatrix} D_{31} & D_{32} \end{bmatrix} \quad F_1 \equiv \begin{bmatrix} F_{11} & F_{12} \\ F_{12}' & F_{22} \end{bmatrix} \quad F_2 \equiv \begin{bmatrix} F_{13} \\ F_{23} \end{bmatrix} \quad F_3 \equiv F_{33}, \tag{D.0.4}$$

with $F_{jn} = \mathbb{E}(g_j g_n') - \mathbb{E}(g_j d')\left[\mathbb{E}(dd')\right]^{-1}\mathbb{E}(dg_n')$, $j, n = 1, 2, 3$ except F_{33} which equals $\mathbb{E}(g_3 g_3')$. Then the asymptotic variance of this estimator is given by $(D_1' F_1^{-1} D_1)^{-1}$, and the required difference in the proposition is given by the upper-left $L_1 \times L_1$ block of $(D_1' F_1^{-1} D_1)^{-1} - (D' F^{-1} D)^{-1}$. We will now characterize this difference.

First note that

$$F^{-1} = \begin{bmatrix} F_1 & F_2 \\ F_2' & F_3 \end{bmatrix}^{-1} = \begin{bmatrix} F_1^{-1}(I + F_2 H F_2' F_1^{-1}) & -F_1^{-1} F_2 H \\ -H F_2' F_1^{-1} & H \end{bmatrix}, D = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}, \tag{D.0.5}$$

where $H \equiv (F_3 - F_2' F^{-1} F_2)^{-1}$. Therefore,

$$D' F^{-1} D = \begin{bmatrix} D_1' & D_2' \end{bmatrix} \begin{bmatrix} F_1^{-1}(I + F_2 H F_2' F_1^{-1}) & -F_1^{-1} F_2 H \\ -H F_2' F_1^{-1} & H \end{bmatrix} \begin{bmatrix} D_1 \\ D_2 \end{bmatrix} = D_1 F_1^{-1} D_1 + J' H J, \tag{D.0.6}$$

where $J \equiv F_2' F_1^{-1} D_1 - D_2$. Therefore, using the Sherman Morrison formula,

$$\begin{aligned} (D' F^{-1} D)^{-1} &= (D_1' F_1^{-1} D_1 + J' H J)^{-1} \\ &= (D_1' F_1^{-1} D_1)^{-1} - (D_1' F_1^{-1} D_1)^{-1} J' [H^{-1} + J (D_1' F_1^{-1} D_1)^{-1} J']^{-1} J (D_1' F_1^{-1} D_1)^{-1}, \end{aligned} \tag{D.0.7}$$

which implies that

$$\begin{aligned} (D_1' F_1^{-1} D_1)^{-1} - (D' F^{-1} D)^{-1} &= (D_1' F_1^{-1} D_1)^{-1} J' [H^{-1} + J (D_1' F_1^{-1} D_1)^{-1} J']^{-1} J (D_1' F_1^{-1} D_1)^{-1} \\ &\equiv (D_1' F_1^{-1} D_1)^{-1} J' K J (D_1' F_1^{-1} D_1)^{-1}, \end{aligned} \tag{D.0.8}$$

where $K \equiv [H^{-1} + J (D_1' F_1^{-1} D_1)^{-1} J']^{-1}$ is a positive definite matrix. The matrix in (A.32) is clearly positive semidefinite, which proves the proposition.

For use in the next proof, we want to characterize the difference corresponding specifically to α_0, which is given by the upper-left $L_1 \times L_1$ block of the matrix in (A.32). For this difference, we focus on the first L_1 columns of $J(D_1' F_1^{-1} D_1)^{-1}$. Note that

$$\begin{aligned} J(D_1' F_1^{-1} D_1)^{-1} &= (F_2' F_1^{-1} D_1 - D_2)(D_1' F_1^{-1} D_1)^{-1} = (F_2' F_1^{-1} D_1 - D_2) D_1^{-1} F_1 D_1^{-1} \\ &= (F_2' - D_2 D_1^{-1} F_1) D_1^{-1}, \end{aligned} \tag{D.0.9}$$

where we have used the fact that D_1 is symmetric. Substituting the definitions of F_1, F_2, D_1, and D_2, we get $J(D_1' F_1^{-1} D_1)^{-1}$ equals

$$\left[(F_{13}' - D_{31} D_{11}^{-1} F_{11} - D_{32} D_{22}^{-1} F_{12}') D_{11}^{-1} \quad (F_{23}' - D_{31} D_{11}^{-1} F_{12} - D_{32} D_{22}^{-1} F_{22}) D_{22}^{-1} \right]. \tag{D.0.10}$$

The first L_1 columns of this matrix are given by the left block, which is

$$L \equiv F_{13}' D_{11}^{-1} - D_{31} D_{11}^{-1} F_{11} D_{11}^{-1} - D_{32} D_{22}^{-1} F_{12}' D_{11}^{-1}. \tag{D.0.11}$$

Let $L = [L_1 \;\; L_2]$, where L_1 is the first column of L and L_2 is the matrix of last $L_1 - 1$ columns of L. Then the difference in asymptotic variances corresponding to α_1 and α_2 is $L_1' K L_1$ and $L_2' K L_2$ respectively.

Proof of Proposition 6.1.2

We want to show that neither L_1 nor L_2 derived in the proof of Proposition 6.1.1 is zero in general. For notational simplicity, I drop the 0 sub/superscripts in this proof, but all expressions are evaluated at the true parameter values.

By standard second order conditions for a probit and a normal MLE,

$$\begin{aligned} D_{11} &= -\mathbb{E}(x' x e_1) \quad D_{22} = \mathbb{E}\begin{bmatrix} \sigma^{-2} x_2' x_2 & 0 \\ 0 & \sigma^{-4}/2 \end{bmatrix} \\ D_{31} &= -\mathbb{E}(x_2' x_2 e_2) h_\alpha \quad D_{32} = -\mathbb{E}(x_2' x_2 e_2) h_\beta \end{aligned} \tag{D.0.12}$$

where

$$
\begin{aligned}
h_\alpha &= [h_{\alpha_1} \;\; h_{\alpha_2}] \\
&\equiv \begin{bmatrix} \dfrac{\theta - (\theta\alpha_1 + \alpha_2)(1+\alpha_1^2\sigma^2)^{-1}\alpha_1\sigma^2}{\sqrt{1+\alpha_1^2\sigma^2}} & \dfrac{1}{\sqrt{1+\alpha_1^2\sigma^2}} I_k \end{bmatrix} \\
h_\beta &= [h_\theta \;\; h_{\sigma^2}] \\
&\equiv \begin{bmatrix} \dfrac{\alpha_1}{\sqrt{1+\alpha_1^2\sigma^2}} I_k & -\dfrac{(\theta\alpha_1+\alpha_2)\alpha_1^2}{2(1+\alpha_1^2\sigma^2)^{3/2}} \end{bmatrix},
\end{aligned} \tag{D.0.13}
$$

$e_1 \equiv [\phi(x\alpha)]^2/\{\Phi(x\alpha)[1-\Phi(x\alpha)]\}$, $e_2 \equiv [\phi(x_2\gamma)]^2/\{\Phi(x_2\gamma)[1-\Phi(x_2\gamma)]\}$. Then we can write

$$
D_{11} = -\mathbb{E}\begin{bmatrix} x_1'x_1e_1 & x_1'x_2e_1 \\ x_2'x_1e_1 & x_2'x_2e_1 \end{bmatrix} \tag{D.0.14}
$$

Let $\mathbb{E}(x_2'x_2e_1) \equiv \Gamma_1$, $\mathbb{E}(x_2're_1) \equiv \Gamma_2$, and $\mathbb{E}(r^2e_1) = \sigma^2_{r^2e_1}$. Then using $x_1 = x_2\theta + r$, we can write

$$
D_{11} = -\begin{bmatrix} \theta'\Gamma_1\theta + 2\Gamma_2'\theta + \sigma^2_{r^2e_1} & \theta'\Gamma_1 + \Gamma_2' \\ \Gamma_1\theta + \Gamma_2 & \Gamma_1 \end{bmatrix} \tag{D.0.15}
$$

Let $\Gamma_3 \equiv (\sigma^2_{r^2e_1} - \Gamma_2'\Gamma_1^{-1}\Gamma_2)$. Using the partitioned inverse formula, we can write

$$
D_{11}^{-1} = \begin{bmatrix} \Gamma_3^{-1} & -\Gamma_3^{-1}(\theta' + \Gamma_2'\Gamma_1^{-1}) \\ -\Gamma_3^{-1}(\theta + \Gamma_1^{-1}\Gamma_2) & \Gamma_1^{-1} + (\theta + \Gamma_1^{-1}\Gamma_2)\Gamma_3^{-1}(\theta' + \Gamma_2'\Gamma_1^{-1}) \end{bmatrix} \tag{D.0.16}
$$

To calculate the first term in (A.35), we begin by deriving F_{13}.

$$
F_{13} = \mathbb{E}(g_1g_3') - \mathbb{E}(g_1d')[\mathbb{E}(dd')]^{-1}\mathbb{E}(dg_3'). \tag{D.0.17}
$$

Let $u_1 \equiv [y-\Phi(x\alpha)]\phi(x\alpha)/\Phi(x\alpha)[1-\Phi(x\alpha)]$ be the generalized residual for the model of interest, $v_1 \equiv [y-\Phi(x_2\gamma)]\phi(x_2\gamma)/\Phi(x_2\gamma)[1-\Phi(x_2\gamma)]$ be the generalized residual for the reduced form, $\Omega_{u_1v_1} \equiv \mathbb{E}\{[s/p(z)]x_2'x_2u_1v_1\}$, $\Omega_{ru_1v_1} \equiv \mathbb{E}\{[s/p(z)]x_2ru_1v_1\}$, $\Omega_{v_1d} \equiv \mathbb{E}(x_2'v_1d')$, $\Omega_{u_1d} \equiv \mathbb{E}\{[s/p(z)]x_2'u_1d'\}$, and $\Omega_{ru_1d} \equiv \mathbb{E}\{[s/p(z)^2]ru_1d'\}$. Then

$$
F_{13} = \begin{bmatrix} \theta'\Omega_{u_1v_1} + \Omega_{ru_1v_1} - (\theta'\Omega_{u_1d} + \Omega_{ru_1d})[\mathbb{E}(dd')]^{-1}\Omega_{v_1d}' \\ \Omega_{u_1v_1} - \Omega_{u_1d}[\mathbb{E}(dd')]^{-1}\Omega_{v_1d}' \end{bmatrix} \tag{D.0.18}
$$

Using the definitions of F_{13} and D_{11}^{-1}, we get that the first column of $F_{13}'D_{11}^{-1}$ is

$$Q_{11} \equiv \{\Omega'_{ru_1v_1} - \Omega_{v_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{ru_1d}\}\Gamma_3^{-1} - \{\Omega'_{u_1v_1} - \Omega_{v_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1d}\}\Gamma_3^{-1}\Gamma_1^{-1}\Gamma_2 \tag{D.0.19}$$

and the last k columns of $F_{13}'D_{11}^{-1}$ are

$$\begin{aligned} Q_{12} \equiv & -\{\Omega'_{ru_1v_1} - \Omega_{v_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{ru_1d}\}\Gamma_3^{-1}(\theta' + \Gamma_2'\Gamma_1^{-1}) \\ & + \{\Omega'_{u_1v_1} - \Omega_{v_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1d}\}\Gamma_1^{-1}[I_k + \Gamma_2\Gamma_3^{-1}(\theta' + \Gamma_2'\Gamma_1^{-1})] \end{aligned} \tag{D.0.20}$$

Next we derive the second term in (A.35).

$$\begin{aligned} h_\alpha D_{11}^{-1} &= [h_{\alpha_1} \;\; h_{\alpha_2}]D_{11}^{-1} \\ &= [\,\Gamma_3^{-1}[h_{\alpha_1} - h_{\alpha_2}(\theta + \Gamma_1^{-1}\Gamma_2)] \; -h_{\alpha_1}\Gamma_3^{-1}(\theta' + \Gamma_2'\Gamma_1^{-1}) + h_{\alpha_2}[\Gamma_1^{-1} + (\theta + \Gamma_1^{-1}\Gamma_2)\Gamma_3^{-1}(\theta' + \Gamma_2'\Gamma_1^{-1})]\,] \end{aligned} \tag{D.0.21}$$

Let $h \equiv h_{\alpha_1} - h_{\alpha_2}\theta$ and $\Gamma_4 \equiv \Gamma_3^{-1}(h - h_{\alpha_2}\Gamma_1^{-1}\Gamma_2)$. Then

$$h_\alpha D_{11}^{-1} = \begin{bmatrix} \Gamma_4 & -\Gamma_4(\theta' + \Gamma_2'\Gamma_1^{-1}) + h_{\alpha_2}\Gamma_1^{-1} \end{bmatrix} \tag{D.0.22}$$

Now consider F_{11}. Let $\Omega_{u_1^2} \equiv \mathbb{E}\{[s/p(z)^2]x_2'x_2u_1^2\}$, $\Omega_{ru_1^2} \equiv \mathbb{E}\{[s/p(z)^2]x_2'ru_1^2\}$, and $\Omega_{r^2u_1^2} \equiv \mathbb{E}\{[s/p(z)^2]r^2u_1^2\}$. Then,

$$F_{11} = [F_{111} \quad F_{112}] \tag{D.0.23}$$

where

$$F_{111} \equiv \begin{bmatrix} \theta'\Omega_{u_1^2}\theta + 2\theta'\Omega_{ru_1^2} + \Omega_{r^2u_1^2} - (\theta'\Omega_{u_1d} + \Omega_{ru_1d})[\mathbb{E}(dd')]^{-1}(\Omega'_{u_1d}\theta + \Omega'_{ru_1d}) \\ \Omega'_{u_1^2}\theta + \Omega_{ru_1^2} - \Omega_{u_1d}[\mathbb{E}(dd')]^{-1}(\Omega'_{u_1d}\theta + \Omega'_{ru_1d}) \end{bmatrix} \tag{D.0.24}$$

$$F_{112} \equiv \begin{bmatrix} \theta'\Omega_{u_1^2} + \Omega'_{ru_1^2} - (\theta'\Omega_{u_1d} + \Omega_{ru_1d})[\mathbb{E}(dd')]^{-1}\Omega'_{u_1d} \\ \Omega_{u_1^2} - \Omega_{u_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1d} \end{bmatrix} \tag{D.0.25}$$

Using the definitions of $h_\alpha D_{11}^{-1}$, and F_{11}, we find that the first column of $h_\alpha D_{11}^{-1} F_{11} D_{11}^{-1}$ is given by

$$\begin{aligned} Q_{21}^* \equiv \Gamma_4\Gamma_3^{-1}(&\{\Omega_{r^2u_1^2} - \Omega_{ru_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{ru_1d}\} \\ &- \{\Omega'_{ru_1^2} - \Omega_{ru_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1d}\}\Gamma_1^{-1}\Gamma_2 - \Omega'_{ru_1^2}\theta) \\ &- (\Gamma_4\Gamma_2' + h_{\alpha_2})\Gamma_1^{-1}\Gamma_3^{-1}[\{\Omega_{ru_1^2} - \Omega_{u_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{ru_1d}\} \\ &- \{\Omega_{u_1^2} - \Omega_{u_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1d}\}\Gamma_1^{-1}\Gamma_2 - (\theta + \Gamma_1^{-1}\Gamma_2)] \end{aligned} \tag{D.0.26}$$

and the last k columns of $h_\alpha D_{11}^{-1} F_{11} D_{11}^{-1}$ are given by

$$\begin{aligned} Q_{22}^* \equiv &-\Gamma_4\Gamma_3^{-1}[\{\Omega_{r^2u_1^2} - \Omega_{ru_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{ru_1d}\} - \Omega'_{ru_1^2}(\theta + \Gamma_1^{-1}\Gamma_2)](\theta' + \Gamma_2'\Gamma_1^{-1}) \\ &+ (\Gamma_4\Gamma_2' + h_{\alpha_2})\Gamma_1^{-1}\{\Omega_{ru_1^2} - \Omega_{u_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{ru_1d}\}\Gamma_3^{-1}(\theta' + \Gamma_2'\Gamma_1^{-1}) + \Gamma_4\Omega'_{ru_1^2}\Gamma_1^{-1} \\ &- [\Gamma_4\Omega_{ru_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1d} + (\Gamma_4\Gamma_2' + h_{\alpha_2})\Gamma_1^{-1}\{\Omega_{u_1^2} - \Omega_{u_1d}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1d}\}] \\ &\Gamma_1^{-1}[I_k + \Gamma_2\Gamma_3^{-1}(\theta' + \Gamma_2'\Gamma_1^{-1})] \end{aligned} \tag{D.0.27}$$

Let $Q_{21} \equiv -\mathbb{E}(x_2'x_2e_2)Q_{21}^*$ and $Q_{22} \equiv -\mathbb{E}(x_2'x_2e_2)Q_{22}^*$. Then,

$$D_{31}D_{11}^{-1}F_{11}D_{11}^{-1} = [Q_{21} \quad Q_{22}] \tag{D.0.28}$$

Clearly, neither Q_{21} nor Q_{22} is zero.

Next we want to find $D_{32}D_{22}^{-1}F_{12}'D_{11}^{-1}$. First note that

$$D_{22}^{-1} = \begin{bmatrix} \sigma^2\,\mathbb{E}(x_2'x_2)^{-1} & 0 \\ 0 & 2\sigma^4 \end{bmatrix}. \tag{D.0.29}$$

Further, let $\Omega_{rd} \equiv \mathbb{E}\{[s/p(z)]x_2'rd'\}$, $\Omega_{r^2d} \equiv \mathbb{E}\{[s/p(z)](r^2\sigma^{-2} - 1)d'\}$, $\Omega_{u_1r} \equiv \mathbb{E}\{[s/p(z)^2]$ $x_2'x_2u_1r\}$, $\Omega_{u_1r^2} \equiv \mathbb{E}\{[s/p(z)^2]x_2'u_1r^2\}$, and $\Omega_{u_1r^3} \equiv \mathbb{E}\{[s/p(z)^2]u_1r^3\}$. Then

$$F_{12}' = \begin{bmatrix} (\Omega'_{u_1r}\theta + \Omega'_{u_1r^2}) - [\Omega_{rd}[\mathbb{E}(dd')]^{-1}(\Omega'_{u_1d}\theta + \Omega'_{ru_1d})] & \Omega'_{u_1r} - \Omega_{rd}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1d} \\ (\Omega_{u_1r^2}\theta + \Omega'_{u_1r^3})\sigma^{-2} - \Omega_{r^2d}[\mathbb{E}(dd')]^{-1}(\Omega'_{u_1d}\theta + \Omega'_{ru_1d}) & \sigma^{-2}\Omega'_{u_1r} - \Omega_{r^2d}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1d} \end{bmatrix}. \tag{D.0.30}$$

Next note that $h_\beta D_{22}^{-1} = [h_\theta \sigma^2 \mathbb{E}(x_2' x_2)^{-1} \;\; 2h_{\sigma^2}\sigma^4]$. Using the definitions of F_{12} and D_{22}^{-1}, we get that the first column of $h_\beta D_{22}^{-1} F_{12}' D_{11}^{-1}$ is

$$
\begin{aligned}
Q_{31}^* \equiv\; & h_\theta \sigma^2 \mathbb{E}(x_2' x_2)^{-1} \Gamma_3^{-1} \\
& (\Omega'_{u_1 r^2} - \Omega_{rd}[\mathbb{E}(dd')]^{-1}\Omega'_{ru_1 d} - \{\Omega'_{u_1 r} - \Omega_{rd}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1 d}\}\Gamma_1^{-1}\Gamma_2) \\
& + 2h_{\sigma^2}\sigma^4 \Gamma_3^{-1} \\
& (\sigma^{-2}\Omega_{u_1 r^3} - \Omega_{r^2 d}[\mathbb{E}(dd')]^{-1}\Omega'_{ru_1 d} - \{\sigma^{-2}\Omega'_{u_1 r^2} - \Omega_{r^2 d}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1 d}\}\Gamma_1^{-1}\Gamma_2)
\end{aligned}
\tag{D.0.31}
$$

and the last k columns of $h_\beta D_{22}^{-1} F_{12}' D_{11}^{-1}$ are

$$
\begin{aligned}
Q_{32}^* \equiv\; & \sigma^2 \mathbb{E}(x_2' x_2)^{-1} ([-h_\theta \{\Omega'_{u_1 r^2} - \Omega_{rd}[\mathbb{E}(dd')]^{-1}\Omega'_{ru_1 d}\} \\
& - 2h_{\sigma^2}\sigma^2 \{\Omega'_{u_1 r^3}\sigma^{-2} - \Omega_{r^2 d}[\mathbb{E}(dd')]^{-1}\Omega'_{ru_1 d}\}\Gamma_3^{-1}(\theta' + \Gamma_2'\Gamma_1^{-1})] \\
& + \{(h_\theta \{\Omega'_{u_1 r} - \Omega_{rd}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1 d}\} \\
& - 2h_{\sigma^2}\sigma^2 \{\Omega'_{u_1 r} - \Omega_{r^2 d}[\mathbb{E}(dd')]^{-1}\Omega'_{u_1 d}\})\Gamma_1^{-1}[I_k + \Gamma_2\Gamma_3^{-1}(\theta' + \Gamma_2'\Gamma_1^{-1})]\})
\end{aligned}
\tag{D.0.32}
$$

Let $Q_{31} \equiv -\mathbb{E}(x_2' x_2 e_2) Q_{31}^*$ and $Q_{32} \equiv -\mathbb{E}(x_2' x_2 e_2) Q_{32}^*$. Then,

$$
D_{32} D_{22}^{-1} F_{12}' D_{11}^{-1} = [Q_{31} \quad Q_{32}] \tag{D.0.33}
$$

Clearly, neither Q_{31} nor Q_{32} is zero.

Thus,

$$
L_1 = Q_{11} + Q_{21} + Q_{31} \neq 0 \tag{D.0.34}
$$

$$
L_2 = Q_{12} + Q_{22} + Q_{32} \neq 0 \tag{D.0.35}
$$

which implies that it is possible to obtain strict efficiency gains for both α_1 and α_2.

Proof of Proposition E.1.

We first show that α_0 is a solution to $min_{\alpha \in \mathbb{A}} \mathbb{E}[s \cdot f_1(y, x, \alpha)]$. First note that for any $\alpha \in \mathbb{A}$,

$$
\mathbb{E}[s \cdot f_1(y, x, \alpha)] = \mathbb{E}\{\mathbb{E}[s \cdot f_1(y, x, \alpha)|x]\} = \mathbb{E}\{p(x_2)\, \mathbb{E}[f_1(y, x, \alpha)|x]\},
$$

where the second equality follows by iterated expectations.

$$\begin{aligned}\mathbb{E}[s \cdot f_1(y, x, \alpha)|x] &= \mathbb{E}\{\mathbb{E}[s \cdot f_1(y, x, \alpha)|y, x]|x\} = \mathbb{E}[\mathbb{E}(s|y, x) f_1(y, x, \alpha)|x] \\ &= \mathbb{E}[p(x_2) f_1(y, x, \alpha)|x] = p(x_2)\,\mathbb{E}[f_1(y, x, \alpha)|x],\end{aligned}$$

where the third equality follows from part 2 of Assumption E.2. Because $p(x_2) \geq 0\ \forall x_2 \in \mathbb{X}_2$, and α_0 minimizes $\mathbb{E}[f_1(y, x, \alpha)|x]$ for all $x \in \mathbb{X}$,

$$p(x_2)\,\mathbb{E}[f_1(y, x, \alpha_0)|x] \leq p(x_2)\,\mathbb{E}[f_1(y, x, \alpha)|x], \quad x \in \mathbb{X}, \quad \alpha \in \mathbb{A}.$$

The result follows from taking an expectation with respect to x.

A similar argument can be used to verify that β_0 solves $\min_{\beta \in \mathbb{B}} \mathbb{E}[s \cdot f_2(x_1, x_2, \beta)]$ and noting that $\mathbb{E}(s|x) = p(x_2)$ under part 2 of Assumption E.2. For the reduced form, part 1 of Assumption E.1 implies using iterated expectations that γ_0 minimizes $\mathbb{E}[f_3(y, x_2, \gamma)]$.

APPENDIX E

ASYMPTOTIC THEORY FOR UNWEIGHTED ESTIMATION

The notion of econometric models underlying the objective functions in (2.2.1)-(2.2.3) being correctly specified, and sample selection being based on x_2 is formalized in the following two assumptions.

Assumption E.1. *Assume that*

1. *For each* $x \in \mathbb{X}$, α_0 *solves* $\min_{\alpha\in\mathbb{A}} \mathbb{E}[f_1(y,x,\alpha)|x]$. *For each* $x_2 \in \mathbb{X}_2$, β_0 *and* γ_0 *solve* $\min_{\beta\in\mathbb{B}} \mathbb{E}[f_2(x_1,x_2,\beta)|x_2]$ *and* $\min_{\gamma\in\Gamma} \mathbb{E}[f_3(y,x_2,\gamma)|x_2]$ *respectively.*

2. α_0, β_0, *and* γ_0 *are the unique solutions to* $\min_{\alpha\in\mathbb{A}} \mathbb{E}[s \cdot f_1(y,x,\alpha)]$ *and* $\min_{\beta\in\mathbb{B}} \mathbb{E}[s \cdot f_2(x_1,x_2,\beta)]$ *respectively.*

Part 1 of this assumption practically means that the underlying model is correctly specified. Part 2 is needed to ensure that the selected subpopulation is sufficiently rich to identify the respective parameters. The notion that s depends on x_2 is formalized in part 2 of the following assumption.

Assumption E.2. *Assume that*

1. x_1 *is observed whenever* $s = 1$, (y, x_2) *are always observed.*

2. $P(s = 1|y, x_1, x_2) = P(s = 1|x_2) \equiv p(x_2)$.

It is simple to show that Assumptions E.1 and E.2 along with regularity conditions, imply consistency of $(\hat{\alpha}_{UJ}, \hat{\beta}_{UJ})$. I show that the following proposition holds.

Proposition E.1. *Under Assumptions E.1 and E.2,* α_0, β_0, *and* γ_0 *solve* $\min_{\alpha\in\mathbb{A}} \mathbb{E}[s \cdot f_1(y,x,\alpha)]$, $\min_{\beta\in\mathbb{B}} \mathbb{E}[s \cdot f_2(x_1,x_2,\beta)]$ *and* $\min_{\gamma\in\Gamma} \mathbb{E}[f_3(y,x_2,\gamma)]$ *respectively.*

The proof (given in Appendix C) simply follows from an iterated expectations argument and is an extension of that in Wooldridge (2002).

Theorem E.1. *Assume that*

1. $\{(y_i, x_i, s_i) : i = 1, \ldots, N\}$ *are random draws from the population satisfying Assumption E.2.*

2. *Assumption E.1 holds.*

3. *Parts 3 (except the assumptions on* Δ*), 4, and 6 of Theorem 2.4.1 hold.*

Then $(\hat{\alpha}_{UJ}, \hat{\beta}_{UJ}) \xrightarrow{p} (\alpha_0, \beta_0)$ *as* $N \to \infty$.

Once we verify that (α_0, β_0) are identified in the subpopulations defined by $s = 1$, the proof of Theorem E.1 is very similar to that of Theorem 2.4.1, and hence is omitted.

To derive the asymptotic distribution of $(\hat{\alpha}_{UJ}, \hat{\beta}_{UJ})$, we assume that $\mathbb{E}[g(\alpha, \beta)]$ is differentiable at (α_0, β_0) with the derivative defined as the following.

$$D_{U0} \equiv \mathbb{E}[\nabla_{(\alpha', \beta')'} g(\alpha, \beta)|_{(\alpha,\beta)=(\alpha_0,\beta_0)}] = \begin{bmatrix} D^0_{U11} & 0 \\ 0 & D^0_{U22} \\ D^0_{U31} & D^0_{U32} \end{bmatrix}, \tag{E.0.1}$$

where $D^0_{Uj1} = \partial g_j(\alpha, \beta)/\partial\alpha|_{(\alpha,\beta)=(\alpha_0,\beta_0)}$ and $D^0_{Uj2} = \partial g_j(\alpha, \beta)/\partial\beta|_{(\alpha,\beta)=(\alpha_0,\beta_0)}$, $j = 1, 2, 3$. Then the following theorem gives the $\sqrt{N}-$asymptotic normality result.

Theorem E.2.*(Asymptotic Normality): Assume that*

1. *The assumptions in Theorem E.1 hold*

2. $(\alpha_0, \beta_0) \in int(\mathbb{A} \times \mathbb{B})$.

3. $g(\alpha, \beta)$ *is twice continuously differentiable on* $int(\mathbb{A} \times \mathbb{B})$.

4. D_{U0} *is of full rank* $L_1 + L_2$.

Then,

$$\sqrt{N}[(\hat{\alpha}'_{UJ}, \hat{\beta}'_{UJ})' - (\alpha'_0, \beta'_0)'] \xrightarrow{d} Normal[0, (D'_{U0} C_0^{-1} D_{U0})^{-1}].$$

The proof follows in a straightforward manner from Theorem 3.4 of Newey and McFadden (1994) and hence is omitted.

APPENDIX F

TABLES FOR CHAPTER 2

Table F.1: Summary of missing data methods used in 5 highly ranked economics journals from 2018 to August 2020.

	Total	% Missingness	% CC	% DVM	% RI	% Other
American Economic Review	319	20.69	71.21	16.67	15.15	15.15
Quarterly Journal of Economics	109	28.44	74.19	9.68	9.68	29.68
Journal of Labor Economics	109	35.78	58.97	15.38	10.26	17.95
Journal of Human Resources	98	43.88	46.51	32.56	11.63	16.28
Journal of Political Economy	211	19.91	59.52	16.67	21.43	14.29
Total	846	26.12	62.44	18.55	14.03	17.65

[1] Column 1 shows the total number of papers published. Column 2 shows the percentage of papers that reported missing values. Columns 3-6 show the percentage of papers that used the complete cases estimator, the dummy variable method, the two-step regression imputation, and other methods respectively.

[2] The row percentages add to more than 100 because some papers use multiple methods.

[3] The articles that do not explicitly mention the method of imputation are included in the two-step regression imputation category since this is the most frequently used method within the imputation category.

Table F.2: Effect of grade variance on probability of having a 4 year college degree.

	Complete cases	Joint GMM	% ↓ in s.e.	Plug-in	DVM
Log(income)	0.148	0.148		0.149	0.150
	(0.042)	(0.041)	2.38	(0.042)	(0.042)
GSD	-0.146	-0.140		-0.138	-0.139
	(0.039)	(0.035)	10.26	(0.037)	(0.035)
GPA	0.329	0.331		0.338	0.339
	(0.049)	(0.043)	12.24	(0.043)	(0.043)
Black	0.413	0.407		0.386	0.395
	(0.128)	(0.116)	9.38	(0.114)	(0.114)
Hispanic	0.539	0.445		0.404	0.419
	(0.147)	(0.135)	8.16	(0.138)	(0.135)
Live in south	0.140	0.149		0.144	0.137
	(0.065)	(0.058)	10.77	(0.057)	(0.057)
Lived in urban area	0.093	0.080		0.082	0.083
	(0.068)	(0.061)	10.29	(0.062)	(0.060)
Mother's education	0.060	0.057		0.055	0.056
	(0.015)	(0.014)	6.67	(0.014)	(0.014)
Father's education	0.063	0.063		0.062	0.062
	(0.011)	(0.010)	9.09	(0.010)	(0.010)
Female	-0.128	-0.132		-0.138	-0.146
	(0.059)	(0.053)	10.17	(0.052)	(0.052)
Cognitive skills	0.436	0.420		0.400	0.404
	(0.050)	(0.044)	12	(0.044)	(0.044)
Non-Cognitive skills	0.012	0.015		0.018	0.016
	(0.030)	(0.027)	10	(0.028)	(0.027)
N	3219	3942		3942	3942
p-value for J stat		0.590			

APPENDIX G

PROOFS FOR CHAPTER 3

Proof of Lemma 3.4.1

Starting with $f_{1i}(.)$, we want to show that $\mathbb{E}\left(\sum_{t=1}^{T} s_{it}\ddot{x}'_{it}\ddot{u}_{it}\right) = 0$. Since $\sum_{t=1}^{T} s_{it}\ddot{x}'_{it}\ddot{u}_{it} = \sum_{t=1}^{T} s_{it}\ddot{x}'_{it}u_{it}$, we want to show that $\mathbb{E}\left(\sum_{t=1}^{T} s_{it}\ddot{x}'_{it}u_{it}\right) = 0$.

First, Assumption 3.2.1 implies by the law of iterated expectations (LIE) that $\mathbb{E}\left(u_{it}|x_i, s_i\right) = 0$. Now, $\forall\ t = 1, \ldots, T$

$$E(s_{it}\ddot{x}'_{it}u_{it}) = \mathbb{E}[\mathbb{E}\left(s_{it}\ddot{x}'_{it}u_{it}|x_i, s_i\right)] = \mathbb{E}[s_{it}\ddot{x}'_{it}\,\mathbb{E}\left(u_{it}|x_i, s_i\right)] = 0.$$

Therefore, $\mathbb{E}\left(\sum_{t=1}^{T} s_{it}\ddot{x}'_{it}u_{it}\right) = 0$.

Using a similar argument for $f_{2i}(.)$, we want to show that $\mathbb{E}\left(\sum_{t=1}^{T} s_{it}\ddot{x}'_{2it}r_{it}\right) = 0$. Now, $\forall\ t = 1, \ldots, T$

$$\mathbb{E}(s_{it}\ddot{x}'_{2it}r_{it}) = \mathbb{E}[\mathbb{E}\left(s_{it}\ddot{x}'_{2it}r_{it}|x_{2i}, s_i\right)] = \mathbb{E}[s_{it}\ddot{x}'_{2it}\,\mathbb{E}\left(r_{it}|x_{2i}, s_i\right)] = 0.$$

The last equality follows from $\mathbb{E}\left(r_{it}|x_{2i}, s_i\right) = 0$ which follows from Assumption 3.2.2 and LIE. Therefore, $\mathbb{E}\left(\sum_{t=1}^{T} s_{it}\ddot{x}'_{2it}r_{it}\right) = 0$.

For $f_{3i}(.)$, we want to show that $\mathbb{E}[\sum_{t=1}^{T}(1 - s_{it})\dot{x}'_{2it}v_{it}] = 0$. First, note that using the LIE, Assumption 3.2.1 implies that $\mathbb{E}\left(u_{it}|x_{2i}, s_i\right) = 0$. This combined with $\mathbb{E}\left(r_{it}|x_{2i}, s_i\right) = 0$ implies that $\mathbb{E}\left(v_{it}|x_{2i}, s_i\right) = \mathbb{E}\left(\beta_1 r_{it} + u_{it}|x_{2i}, s_i\right) = 0$. Now, $\forall\ t = 1, \ldots, T$

$$\mathbb{E}[(1 - s_{it})\dot{x}'_{2it}v_{it}] = \mathbb{E}\{\mathbb{E}[(1 - s_{it})\dot{x}'_{2it}v_{it}|x_{2i}, s_i]\} = \mathbb{E}[(1 - s_{it})\dot{x}'_{2it}\,\mathbb{E}(v_{it}|x_{2i}, s_i)] = 0.$$

and hence $\mathbb{E}[\sum_{t=1}^{T}(1 - s_{it})\dot{x}'_{2it}v_{it}] = 0$.

Proof of Proposition 3.4.2.1

$\hat{\beta}_D$ is obtained by estimating the parameters in equation (3.4.3) using POLS. POLS will be consistent if

$$\mathbb{E}\begin{bmatrix} g_{1i}(.) \\ g_{2i}(.) \\ g_{3i}(.) \end{bmatrix} \equiv \mathbb{E}\begin{bmatrix} \sum_{t=1}^{T} \grave{s}_{it}\grave{x}'_{1it}\ \grave{e}_{it} \\ \sum_{t=1}^{T}(1 - \grave{s}_{it})\ \grave{e}_{it} \\ \sum_{t=1}^{T} \grave{x}'_{2it}\ \grave{e}_{it} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{F.1}$$

We are going to show that each of these holds true iff either $\beta_1 = 0$ or $\pi_2 = d_i = 0 \; \forall \; i$.

First, note that

$$\begin{aligned}\grave{e}_{it} = e_{it} - \bar{e}_i =&[(1 - s_{it})x_{22it} - \overline{(1 - s_i)x_{22i}}]\pi_2\beta_1 + [(1 - s_{it})d_i - \overline{(1 - s_i)d_i}]\beta_1 \\ &+ [(1 - s_{it})r_{it} - \overline{(1 - s_i)r_i}]\beta_1 + [u_{it} - \bar{u}_i],\end{aligned}$$

where

$$\overline{(1 - s_i)x_{22i}} = T_i^{-1} \sum_{q=1}^{T} (1 - s_{iq})x_{22iq}$$

$$\overline{(1 - s_i)d_i} = T_i^{-1} \sum_{q=1}^{T} (1 - s_{iq})d_i = (1 - T^{-1}T_i)d_i$$

$$\overline{(1 - s_i)r_i} = T_i^{-1} \sum_{q=1}^{T} (1 - s_{iq})r_{iq}.$$

Starting with g_{1i}, the first term is

$$\mathbb{E}\{\sum_{t=1}^{T} \grave{s}_{it}\grave{x}'_{1it}[(1-s_{it})x_{22it} - \overline{(1 - s_i)x_{22i}}]\pi_2\beta_1\} = \sum_{t=1}^{T} \mathbb{E}\{\grave{s}_{it}\grave{x}'_{1it}[(1-s_{it})x_{22it} - \overline{(1 - s_i)x_{22i}}]\pi_2\beta_1\}.$$

Consider this expectation for each t separately. It is 0 iff either $\pi_2 = 0$ or $\beta_1 = 0$ or both are 0. If neither of these conditions holds, then this term will be a non-zero number, except by fluke. For the second term,

$$\mathbb{E}\{\grave{s}_{it}\grave{x}'_{1it}[(1 - s_{it}) - \overline{(1 - s_i)d_i}]d_i\beta_1\}$$

is zero $\forall \; t$ iff $\beta_1 = 0$ or $d_i = 0 \; \forall \; i$ or both. For the third term,

$$\mathbb{E}\{\grave{s}_{it}\grave{x}'_{1it}[(1 - s_{it})r_{it} - \overline{(1 - s_i)r_i}]\beta_1\}$$

is zero $\forall \; t$ iff $\beta_1 = 0$. For the fourth term,

$$\mathbb{E}[\grave{s}_{it}\grave{x}'_{1it}(u_{it} - \bar{u}_i)]$$

is zero $\forall \; t$ under Assumption 3.2.1.

Moving on to g_{2i}, for the first term

$$\mathbb{E}\{[(1 - s_{it}) - (1 - T^{-1}T_i)][(1 - s_{it})x_{22it} - \overline{(1 - s_i)x_{22i}}]\pi_2\beta_1\}$$

is zero $\forall\ t$ iff $\pi_2 = 0$ or $\beta_1 = 0$ or both. For the second term,

$$\mathbb{E}\{[(1 - s_{it}) - (1 - T^{-1}T_i)]^2 d_i \beta_1\}$$

is zero $\forall\ t$ iff $\beta_1 = 0$ or $d_i = 0\ \forall\ i$ or both. For the third term,

$$\mathbb{E}\{[(1 - s_{it}) - (1 - T^{-1}T_i)][(1 - s_{it})r_{it} - \overline{(1 - s_i)r_i}]\beta_1\}$$

is zero under Assumption 3.2.2. For the fourth term,

$$\mathbb{E}[[(1 - s_{it}) - (1 - T^{-1}T_i)](u_{it} - \bar{u_i})]$$

is zero $\forall\ t$ under Assumption 3.2.1.

Moving on to g_{3i}, for the first term

$$\mathbb{E}\{\grave{x}'_{2it}[(1 - s_{it})x_{22it} - \overline{(1 - s_i)x_{22i}}]\pi_2\beta_1\}$$

is zero $\forall\ t$ iff $\pi_2 = 0$ or $\beta_1 = 0$ or both. For the second term,

$$\mathbb{E}\{\grave{x}'_{2it}[(1 - s_{it}) - (1 - T^{-1}T_i)]d_i\beta_1\}$$

is zero $\forall\ t$ iff $\beta_1 = 0$ or $d_i = 0\ \forall\ i$ or both. For the third term,

$$\mathbb{E}\{\grave{x}'_{2it}[(1 - s_{it})r_{it} - \overline{(1 - s_i)r_i}]\beta_1\}$$

is zero under Assumption 3.2.2. For the fourth term,

$$\mathbb{E}[\grave{x}'_{2it}(u_{it} - \bar{u_i})]$$

is zero $\forall\ t$ under Assumption 3.2.1.

Thus, for each of the moment conditions in (F.1) to be zero, we need either $\beta_1 = 0$ or $\pi_2 = d_i = 0$ $\forall\ i$.

Proof of Proposition 3.4.3.1

Let the error $\beta_1(1 - s_{it})[\ddot{x}_{2it}(\pi - \hat{\pi}_{Imp}) + \ddot{r}_{it}] + \ddot{u}_{it} \equiv \ddot{e}_{it}$ and let the set of regressors $[s_{it}\ddot{x}_{1it} + (1 - s_{it})\ddot{x}_{2it}\hat{\pi}_{Imp} \quad \ddot{x}_{2it}] \equiv \ddot{z}_{it}$

The POLS estimator is

$$\hat{\beta}_{Imp} = \Big(\sum_i \sum_t \ddot{z}'_{it} \ddot{z}_{it} \Big)^{-1} \sum_i \sum_t \ddot{z}'_{it} \ddot{y}_{it} \tag{F.2}$$

$$= \beta + \Big(N^{-1} \sum_i \sum_t \ddot{z}'_{it} \ddot{z}_{it} \Big)^{-1} N^{-1} \sum_i \sum_t \ddot{z}'_{it} \ddot{e}_{it} \tag{G.0.1}$$

Consider the probability limit of the term $N^{-1} \sum_i \sum_t \ddot{z}'_{it} \ddot{e}_{it}$. Plugging in the definitions of $\ddot{z}_{it}$ and $\ddot{e}_{it}$, the first term is

$$plim \sum_t N^{-1} \sum_i s_{it} \ddot{x}_{1it} \ddot{u}_{it} = \sum_t \mathbb{E}(s_{it} \ddot{x}_{1it} \ddot{u}_{it}) = 0$$

This last equality is due to Assumption 3.2.1. The second term is

$$plim \sum_t N^{-1} \sum_i (1 - s_{it}) \hat{\pi}'_{Imp} \ddot{x}'_{2it} \ddot{u}_{it} = \sum_t \mathbb{E}[(1 - s_{it}) \pi' \ddot{x}'_{2it} \ddot{u}_{it}] = 0$$

where the last equality again holds because of Assumption 3.2.1. The third term is

$$plim \sum_t N^{-1} \sum_i (1 - s_{it}) \hat{\pi}'_{Imp} \ddot{x}'_{2it} [\ddot{r}_{it} + \ddot{x}_{2it} (\pi - \hat{\pi})] \beta_1$$

$$= \sum_t \mathbb{E}[(1 - s_{it}) \pi' \ddot{x}'_{2it} \ddot{r}_{it}] \beta_1 = 0$$

The second equality here holds because $\hat{\pi}_{Imp}$ is a consistent estimator of π, and the third holds due to Assumption 3.2.2. The fourth term is

$$plim \sum_t N^{-1} \sum_i \ddot{x}'_{2it} \ddot{u}_{it} = \sum_t \mathbb{E}(\ddot{x}'_{2it} \ddot{u}_{it}) = 0$$

where the last equality holds due to Assumption 3.2.1. Finally,

$$plim \sum_t N^{-1} \sum_i \ddot{x}'_{2it} (1 - s_{it}) [\ddot{r}_{it} + \ddot{x}_{2it} (\pi - \hat{\pi})] \beta_1$$

$$= \sum_t \mathbb{E}[(1 - s_{it}) \ddot{x}'_{2it} \ddot{r}_{it}] \beta_1 = 0$$

as proved above.

Since $plim \ N^{-1} \sum_i \sum_t \ddot{z}'_{it} \ddot{e}_{it} = 0$, from (F.2), $plim \ \hat{\beta}_{Imp} = \beta$.

Proof of Lemma 3.6.1

(i) Start with $\mathbb{E}[m_{1i}(\beta, \pi)] = 0$. This will hold true if

$$\mathbb{E}\begin{bmatrix} s_{ip}x_{1ip}s_{it}\tilde{u}_i(t) \\ x'_{2ip}s_{it}\tilde{u}_i(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Now, we can write

$$\begin{aligned}
\mathbb{E}[s_{ip}x_{1ip}s_{it}\tilde{u}_i(t)] &= \mathbb{E}(s_{ip}x_{1ip}s_{it}u_{it}) - \mathbb{E}\left[s_{ip}x_{1ip}s_{it}T_i(t)^{-1}\sum_{q=t+1}^{T} s_{iq}u_{iq}\right] \\
&= \mathbb{E}(s_{ip}x_{1ip}s_{it}u_{it}) - \sum_{q=t+1}^{T}\mathbb{E}[s_{ip}x_{1ip}s_{it}T_i(t)^{-1}s_{iq}u_{iq}] \\
&= \mathbb{E}(s_{ip}s_{it})\,\mathbb{E}(x_{1ip}u_{it}) - \sum_{q=t+1}^{T}\mathbb{E}[s_{ip}s_{it}T_i(t)^{-1}s_{iq}]\,\mathbb{E}(x_{1ip}u_{iq}) \\
&= 0
\end{aligned}$$

The first equality follows from just the definition of $u_i(t)$, the third follows from $\mathbf{s}_i \perp\!\!\!\perp (\mathbf{x}_i, \mathbf{u}_i, \mathbf{r}_i)$ and the last one follows from Assumption 3.6.1. Similarly, $\mathbb{E}[x'_{2ip}s_{it}\tilde{u}_i(t)] = 0$.

Moving on to $\mathbb{E}[m_{2i}(\beta, \pi)] = 0$, we need $\mathbb{E}[x'_{2ip}s_{it}\tilde{r}_{it}] = 0$. We can write

$$\begin{aligned}
\mathbb{E}[x'_{2ip}s_{it}\tilde{r}_i(t)] &= \mathbb{E}(x'_{2ip}s_{it}r_{it}) - \mathbb{E}\left[x'_{2ip}s_{it}T_i(t)^{-1}\sum_{q=t+1}^{T} s_{iq}r_{iq}\right] \\
&= \mathbb{E}(x'_{2ip}s_{it}r_{it}) - \sum_{q=t+1}^{T}\mathbb{E}\left[x'_{2ip}s_{it}T_i(t)^{-1}s_{iq}r_{iq}\right] \\
&= \mathbb{E}(s_{it})\,\mathbb{E}(x'_{2ip}r_{it}) - \sum_{q=t+1}^{T}\mathbb{E}[s_{it}T_i(t)^{-1}s_{iq}]\,\mathbb{E}(x'_{2ip}r_{iq}) \\
&= 0
\end{aligned}$$

The third equality follows from $\mathbf{s}_i \perp\!\!\!\perp (\mathbf{x}_i, \mathbf{u}_i, \mathbf{r}_i)$ and the last one follows from Assumption 3.6.2. Finally we consider the third set of moment conditions $\mathbb{E}[m_{2i}(\beta, \pi)] = 0$, for which we need

$\mathbb{E}[x'_{2ip}(1-s_{it})\check{v}_{it}] = 0$. We can write $\mathbb{E}[x'_{2ip}(1-s_{it})\check{v}_i(t)]$ equals

$$\begin{aligned}
&\mathbb{E}[x'_{2ip}(1-s_{it})v_{it}] - \mathbb{E}[x'_{2ip}(1-s_{it})(T-t-T_i(t))^{-1}\sum_{q=t+1}^{T}(1-s_{iq})v_{iq}] \\
=&\mathbb{E}[x'_{2ip}(1-s_{it})v_{it}] - \sum_{q=t+1}^{T}\mathbb{E}[x'_{2ip}(1-s_{it})(T-t-T_i(t))^{-1}(1-s_{iq})v_{iq}] \\
=&\mathbb{E}(1-s_{it})\,\mathbb{E}(x'_{2ip}v_{it}) - \sum_{q=t+1}^{T}\mathbb{E}[(1-s_{it})(T-t-T_i(t))^{-1}(1-s_{iq})]\,\mathbb{E}(x'_{2ip}v_{iq}) \\
=&0
\end{aligned}$$

where the last equality follows from $\mathbb{E}(x'_{2ip}v_{iq}) = 0$ which follows from Assumptions 3.6.1 and 3.6.2.

(ii) Starting with $\mathbb{E}[m_{1i}(\beta,\pi)] = 0$, we can first write

$$\begin{aligned}
\mathbb{E}[s_{ip}x_{1ip}s_{it}\tilde{u}_i(t)] &= \mathbb{E}(s_{ip}x_{1ip}s_{it}u_{it}) - \sum_{q=t+1}^{T}\mathbb{E}[s_{ip}x_{1ip}s_{it}T_i(t)^{-1}s_{iq}u_{iq}] \\
&= \mathbb{E}[\mathbb{E}(s_{ip}x_{1ip}s_{it}u_{it}|\mathbf{x}_i^t,\mathbf{s}_i)] - \sum_{q=t+1}^{T}\mathbb{E}\{\mathbb{E}[s_{ip}x_{1ip}s_{it}T_i(t)^{-1}s_{iq}u_{iq}|\mathbf{x}_i^t,\mathbf{s}_i]\} \\
&= \mathbb{E}[s_{ip}x_{1ip}s_{it}\,\mathbb{E}(u_{it}|\mathbf{x}_i^t,\mathbf{s}_i)] - \sum_{q=t+1}^{T}\mathbb{E}\{s_{ip}x_{1ip}s_{it}T_i(t)^{-1}s_{iq}\,\mathbb{E}[u_{iq}|\mathbf{x}_i^t,\mathbf{s}_i]\} \\
&= 0
\end{aligned}$$

The second equality follows from the LIE and the fourth follows from Assumption 3.6.1'. This is because using the LIE, Assumption 3.6.1' implies that $\mathbb{E}(u_{it}|\mathbf{x}_i^t,\mathbf{s}_i) = 0$ for every $t = 1,\ldots,T$. Moreover, since $\mathbb{E}(u_{iq}|\mathbf{x}_i^q,\mathbf{s}_i) = 0$ for $q = t+1,\ldots,T$, using the LIE implies that $\mathbb{E}(u_{iq}|\mathbf{x}_i^t,\mathbf{s}_i) = 0$ for any $t < q$. Similarly, $\mathbb{E}[x'_{2ip}s_{it}\tilde{u}_i(t)] = 0$.

We can write a similar proof for $\mathbb{E}[m_{2i}(\beta,\pi)] = 0$ using the LIE and Assumption 3.6.2'. For

$\mathbb{E}[m_{3i}(\beta,\pi)] = 0$, write

$$\begin{aligned}
\mathbb{E}[x'_{2ip}(1-s_{it})\check{v}_i(t)] &= \mathbb{E}[x'_{2ip}(1-s_{it})v_{it}] - \sum_{q=t+1}^{T} \mathbb{E}[x'_{2ip}(1-s_{it})(T-t-T_i(t))^{-1}(1-s_{iq})v_{iq}] \\
&= \mathbb{E}\{\mathbb{E}[x'_{2ip}(1-s_{it})v_{it}|x^t_{2i},\mathbf{s}_i]\} \\
&\quad - \sum_{q=t+1}^{T} \mathbb{E}\{\mathbb{E}[x'_{2ip}(1-s_{it})(T-t-T_i(t))^{-1}(1-s_{iq})v_{iq}|x^t_{2i},\mathbf{s}_i]\} \\
&= \mathbb{E}\{x'_{2ip}(1-s_{it})\,\mathbb{E}[v_{it}|x^t_{2i},\mathbf{s}_i]\} \\
&\quad - \sum_{q=t+1}^{T} \mathbb{E}\{x'_{2ip}(1-s_{it})(T-t-T_i(t))^{-1}(1-s_{iq})\,\mathbb{E}[v_{iq}|x^t_{2i},\mathbf{s}_i]\} \\
&= 0
\end{aligned}$$

where the second equality follows from the LIE and the fourth from Assumptions 3.6.1' and 3.6.2'. This is because using the LIE and the fact that $v_{it} = \beta_1 r_{it} + u_{it}$, Assumptions 3.6.1' and 3.6.2' imply that $\mathbb{E}(v_{it}|\mathbf{x}^t_{2i},\mathbf{s}_i) = 0$ for every $t = 1,\ldots,T$. Moreover, since $\mathbb{E}(v_{iq}|\mathbf{x}^q_{2i},\mathbf{s}_i) = 0$ for $q = t+1,\ldots,T$, using the LIE implies that $\mathbb{E}(v_{iq}|\mathbf{x}^t_{2i},\mathbf{s}_i) = 0$ for any $t < q$.

APPENDIX H

EXTENSIONS TO CHAPTER 3

H.1 Missing vectors

In the model of interest (3.2.1), we assumed that x_{1it} is a scalar. We can extend this framework to the case where x_{1it} is a $m \times 1$ vector, all elements of which are missing at the same time. In other words, if one element of x_{1it} is missing for observation i at time t, then so are all the other elements of x_{1it}. This does not fundamentally change the analysis and the single missing data indicator s_{it} is still sufficient to characterize missingness.

The population model is given by

$$y_{it} = x_{1it}\beta_1 + x_{2it}\beta_2 + c_i + u_{it} \equiv x_{it}\beta + c_i + u_{it}, \quad t = 1, \ldots, T, \tag{H.1.1}$$

which is the same as equation (3.2.1) except x_{1it} is a $1 \times m$ vector now. The imputation equations are a set of m equations (one for each element in x_{1it}).

$$x_{1it} = x_{2it}\Pi + d_i + r_{it} \tag{H.1.2}$$

where Π is a $k \times m$ matrix and d_i is a $1 \times m$ vector. The reduced form is

$$y_{it} = (x_{2it}\Pi + d_i + r_{it})\beta_1 + x_{2it}\beta_2 + c_i + u_{it} \equiv x_{2it}\gamma + h_i + v_{it}, \tag{H.1.3}$$

where $\gamma \equiv \Pi\beta_1 + \beta_2$, $h_i \equiv d_i\beta_1 + c_i$, and $v_{it} \equiv r_{it}\beta_1 + u_{it}$.

Since all elements of x_{1it} are missing at the same time, the definition of the missing data indicator given in section 3 is still sufficient to characterize missingness. That is, $s_{it} = 1$ if x_{1it} is observed and 0 otherwise. Then the joint GMM is based on the following set of moment functions.

$$f_i(\beta, \Pi) = \begin{bmatrix} \sum_{t=1}^{T} s_{it}\ddot{x}_{it}'(\ddot{y}_{it} - \ddot{x}_{1it}\beta_1 - \ddot{x}_{2it}\beta_2) \\ \sum_{t=1}^{T} s_{it}\ddot{x}_{2it}' \otimes (\ddot{x}_{1it} - \ddot{x}_{2it}\Pi)' \\ \sum_{t=1}^{T} (1 - s_{it})\dot{x}_{2it}'\big(\dot{y}_{it} - \dot{x}_{2it}(\beta_1\Pi + \beta_2)\big) \end{bmatrix} \equiv \begin{bmatrix} f_{1i}(\beta, \Pi) \\ f_{2i}(\beta, \Pi) \\ f_{3i}(\beta, \Pi) \end{bmatrix} \tag{H.1.4}$$

This is a set of $k(2+m)+m$ moment conditions with $k(1+m)+m$ parameters to estimate. Thus the number of over-identifying restrictions still equals k. Note that $f_{2i}(.)$ is still a set of exactly identified moment functions, and hence Lemma 3.4.2 is still valid. The rest of the GMM estimation proceeds the same way as in Section 4, except the matrices C and D are now based on the moment conditions in (G.4).

This framework can further be extended to the case where the elements of x_{1it} are not missing at the same time. Although it leads to loss of some information in this case, it is still more efficient than using the complete case analysis. For instance, consider the case where in equation (3.2.1), $x_{1it} = [w_{it} \;\; w_{i,t-1}]$, where w_{it} is a policy variable. If w_{it} contains missing values, then so does $w_{i,t-1}$. In this case, the missingness cannot be entirely characterized with a single missing data indicator as w_{it} and $w_{i,t-1}$ are missing in different time periods for observation i. We define the selection indicators as the following.

$$s_{1it} = \begin{cases} 1 & \text{if both } w_{it} \text{ and } w_{i,t-1} \text{ are observed} \quad t = 1, ..., T \\ 0 & \text{otherwise} \end{cases}$$

$$s_{2it} = \begin{cases} 1 & \text{if neither } w_{it} \text{ nor } w_{i,t-1} \text{ is observed} \quad t = 1, ..., T \\ 0 & \text{otherwise} \end{cases}$$

Thus, the complete cases are those time periods for individual i for which w_i is observed in both the current and the previous period, and are characterized by $s_{1it} = 1$. One option in this case is to estimate β using the complete cases fixed effects, as discussed in Section 4.

However, we can also use the joint GMM by utilizing the observations for which $s_{2it} = 1$. Note that s_{2it} does not characterize all the incomplete cases. It is equal to 1 only for the observations for which neither w_{it} nor $w_{i,t-1}$ is observed, and 0 for both the complete cases as well as the observations for which either w_{it} or $w_{i,t-1}$ is observed. It thus does not make use of the observations for which both s_{1it} and s_{2it} are 0.

We impose the following assumption on the population distribution.

Assumption G.1 For every $t = 1, \ldots, T$, (i) $\mathbb{E}(s_{1it}\ddot{x}'_{it}u_{it}) = 0$ (ii) $\mathbb{E}(s_{1it}\ddot{x}'_{2it}r_{it}) = 0$ (iii) $\mathbb{E}(s_{2it}\dot{x}'_{2it}v_{it}) = 0$

The joint GMM is then based on the following moment functions.

$$f_i(\beta, \pi) = \begin{bmatrix} \sum_{t=1}^{T} s_{1it}\ddot{x}'_{it}(\ddot{y}_{it} - \ddot{x}_{1it}\beta_1 - \ddot{x}_{2it}\beta_2) \\ \sum_{t=1}^{T} s_{1it}\ddot{x}'_{2it}(\ddot{x}_{1it} - \ddot{x}_{2it}\pi) \\ \sum_{t=1}^{T} s_{2it}\dot{x}'_{2it}\left(\dot{y}_{it} - \dot{x}_{2it}(\beta_1\pi + \beta_2)\right) \end{bmatrix} \equiv \begin{bmatrix} f_{1i}(\beta, \Pi) \\ f_{2i}(\beta, \Pi) \\ f_{3i}(\beta, \Pi) \end{bmatrix} \tag{H.1.5}$$

where

$$\ddot{x}_{it} = x_{it} - (\sum_{q=1}^{T} s_{1it})^{-1} \sum_{q=1}^{T} s_{1iq}x_{iq}$$

$$\ddot{y}_{it} = y_{it} - (\sum_{q=1}^{T} s_{1it})^{-1} \sum_{q=1}^{T} s_{1iq}y_{iq}$$

$$\dot{x}_{it} = x_{it} - (\sum_{q=1}^{T} s_{2it})^{-1} \sum_{q=1}^{T} s_{2iq}x_{iq}$$

$$\dot{y}_{it} = y_{it} - (\sum_{q=1}^{T} s_{2it})^{-1} \sum_{q=1}^{T} s_{2iq}y_{iq}$$

That is, for $f_{1i}(.)$ and $f_{2i}(.)$, the variables are still time demeaned using the complete cases, but for $f_{3i}(.)$, they are time demeaned using only the observations for which neither w_{it} nor $w_{i,t-1}$ is observed. Note that the moment functions $f_{2i}(.)$ imply that both w_{it} and $w_{i,t-1}$ will be imputed using the same covariates x_{2it}

The rest of the GMM estimation proceeds in the usual fashion using the moment functions in (G.5).

In order to utilize all the incomplete cases, we can further extend this framework by introducing a separate selection indicator for w_{it} and $w_{i,t-1}$ and writing a separate imputation equation (with different sets of covariates) for each of these.

H.2 Time varying unobserved heterogeneity

We can extend the basic model in Section 2 to allow for the unobserved heterogeneity to vary over time. So instead of equation (3.2.1), our model of interest is now

$$y_{it} = x_{it}\beta + \eta_t c_i + u_{it}, \quad t = 1, \ldots, T. \tag{H.2.1}$$

The coefficients of c_i are now η_t which are time-varying parameters to be estimated. We also allow for time-varying heterogeneity in the imputation model. The new model is

$$x_{1it} = x_{2it}\pi + \zeta_t d_i + r_{it}, \quad t = 1, \ldots, T. \tag{H.2.2}$$

The reduced form then becomes

$$y_{it} = x_{2it}\gamma + h_{it} + v_{it}, \quad t = 1, \ldots, T \tag{H.2.3}$$

where $\gamma \equiv \beta_1\pi + \beta_2$, $h_{it} \equiv \beta_1\zeta_t d_i + \eta_t c_i$, and $v_{it} \equiv \beta_1 r_{it} + u_{it}$.[1]

The question we consider here is that under what assumptions will the joint GMM defined in Section 3 consistently estimates β and π. Starting with equation (G.6), if we time demean using the complete cases, we get

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{\eta}_t c_i + \ddot{u}_{it}, \quad t = 1, \ldots, T, \tag{H.2.4}$$

where $\ddot{y}_{it}$, $\ddot{x}_{it}$, and $\ddot{u}_{it}$ are defined in the same way as in Section 3. But now, this transformation does not eliminate c_i. Therefore, for the moment conditions $\mathbb{E}[f_{1i}(\beta)] = 0$ in (3.4.10) to be valid, we need for every $t = 1, \ldots, T$

$$\mathbb{E}[s_{it}\ddot{x}'_{it}(\ddot{\eta}_t c_i + \ddot{u}_{it})] = 0. \tag{H.2.5}$$

We know that for every $t = 1, \ldots, T$

$$\mathbb{E}(s_{it}\ddot{x}'_{it}\ddot{u}_{it}) = 0 \tag{H.2.6}$$

under Assumption 3.3.2. We additionally need that for every $t = 1, \ldots, T$

$$\mathbb{E}(s_{it}\ddot{x}'_{it}\ddot{\eta}_t c_i) = 0. \tag{H.2.7}$$

[1]Note that the definitions of γ and v_{it} are the same as those in Section 2. Only the unobserved heterogeneity has changed.

A sufficient condition for this to hold is that for every $t = 1, \ldots, T$

$$\mathbb{E}(c_i | \ddot{x}_{it}, s_i) = 0. \tag{H.2.8}$$

This says that at time t, the unobserved heterogeneity c_i is mean independent of the time deviated x_{it} and selection in all time periods. This is clearly stronger than Assumption 3.3.2 which did not put any restriction on the relationship between s_i and c_i. However, it is weaker than assuming c_i is mean independent of x_{it}. We are only assuming that it is mean independent of the time deviated x_{it}, that is $\ddot{x}_{it}$.

Similarly, when we time demean the new imputation model (G.7), we get

$$\ddot{x}_{1it} = \ddot{x}_{2it}\pi + \ddot{\zeta}_t d_i + \ddot{r}_{it}, \quad t = 1, \ldots, T. \tag{H.2.9}$$

For the moment conditions $\mathbb{E}[f_{2i}(\pi)] = 0$ in (3.4.10) to be valid, we need that for every $t = 1, \ldots, T$

$$\mathbb{E}[s_{it}\ddot{x}'_{2it}(\ddot{\zeta}_t d_i + \ddot{r}_{it})] = 0 \tag{H.2.10}$$

for which we need to assume that for every $t = 1, \ldots, T$

$$\mathbb{E}(d_i | \ddot{x}_{2it}, s_i) = 0 \tag{H.2.11}$$

in addition to Assumption 3.3.2. Similarly, the time deviated reduced form is

$$\ddot{y}_{it} = \ddot{x}_{2it}\gamma + \ddot{h}_{it} + \ddot{v}_{it}, \quad t = 1, \ldots, T. \tag{H.2.12}$$

It is easy to see that given equation (G.17), Assumptions (G.13) and (G.16) along with Assumption 3.3.2 are sufficient for the moment conditions $\mathbb{E}[f_{3i}(\beta, \pi)] = 0$ in (3.4.10) to be valid.

REFERENCES

REFERENCES

Abrevaya, J., & Donald, S. G. (2011). A gmm approach for dealing with missing data on regressors and instruments. *Unpublished manuscript*.

Abrevaya, J., & Donald, S. G. (2017). A gmm approach for dealing with missing data on regressors. *Review of Economics and Statistics*, *99*(4), 657–662.

Ahu, S. C., & Schmidt, P. (1995). A separability result for gmm estimation, with applications to gls prediction and conditional moment tests. *Econometric Reviews*, *14*(1), 19–34.

Angrist, J. D., & Krueger, A. B. (1992). The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American statistical Association*, *87*(418), 328–336.

Angrist, J. D., & Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, *13*(2), 225–235.

Arellano, M., & Meghir, C. (1992). Female labour supply and on-the-job search: an empirical model estimated using complementary data sets. *The Review of Economic Studies*, *59*(3), 537–559.

Card, D. (1993). *Using geographic variation in college proximity to estimate the return to schooling*. National Bureau of Economic Research Cambridge, Mass., USA.

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.

Devereux, P. J., & Hart, R. A. (2010). Forced to be rich? returns to compulsory schooling in britain. *The Economic Journal*, *120*(549), 1345–1364.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

Hellerstein, J. K., & Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics*, *81*(1), 1–14.

Hentschel, J., Lanjouw, J. O., Lanjouw, P., & Poggi, J. (2000). Combining census and survey data to trace the spatial dimensions of poverty: A case study of ecuador. *The World Bank Economic Review*, *14*(1), 147–165.

Inoue, A., & Solon, G. (2005). Two-sample instrumental variables estimators. *NBER Working Paper*(t0311).

Inoue, A., & Solon, G. (2010). Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, *92*(3), 557–561.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, *91*(433), 222–230.

Klevan, S., Weinberg, S. L., & Middleton, J. A. (2016). Why the boys are missing: Using social capital to explain gender differences in college enrollment for public high school students. *Research in Higher Education*, *57*(2), 223–257.

Klevmarken, N. A. (1982). *Missing variables and two-stage least-squares estimation from more than one data set* (Tech. Rep.). Research Institute of Industrial Economics.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, *87*(420), 1227–1237.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (Vol. 333). John Wiley & Sons.

Loureiro, M. L., & Nayga Jr, R. M. (2006). Obesity, weight loss, and physician's advice. *Social science & medicine*, *62*(10), 2458–2468.

Loureiro, M. L., & Nayga Jr, R. M. (2007). Physician's advice affects adoption of desirable dietary behaviors. *Review of Agricultural Economics*, *29*(2), 318–330.

MaCurdy, T., Mroz, T., & Gritz, R. M. (1998). An evaluation of the national longitudinal survey on youth. *The Journal of Human Resources*, *33*(2), 345–436.

McDonough, I. K., & Millimet, D. L. (2017). Missing data, imputation, and endogeneity. *Journal of econometrics*, *199*(2), 141–155.

Mogstad, M., & Wiswall, M. (2012). Instrumental variables estimation with partially missing instruments. *Economics Letters*, *114*(2), 186–189.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

Ortega-Sanchez, R., Jimenez-Mena, C., Cordoba-Garcia, R., Muñoz-Lopez, J., Garcia-Machado, M. L., & Vilaseca-Canals, J. (2004). The effect of office-based physician's advice on adolescent exercise behavior. *Preventive medicine*, *38*(2), 219–226.

Pacini, D., & Windmeijer, F. (2016). Robust inference for the two-sample 2sls estimator. *Economics letters*, *146*, 50–54.

Prokhorov, A., & Schmidt, P. (2009). Gmm redundancy results for general missing data problems. *Journal of Econometrics*, *151*(1), 47–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1987). *Multiple imputation for non-response in surveys*. John Wiley & Sons.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, *8*(1), 3–15.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Secker-Walker, R. H., Solomon, L. J., Flynn, B. S., Skelly, J. M., & Mead, P. B. (1998). Reducing smoking during pregnancy and postpartum: physician's advice supported by individual counseling. *Preventive medicine*, *27*(3), 422–430.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, *30*(4), 377–399.

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

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.