A GMM approach for dealing with missing data on regressors and instruments*

Jason Abrevaya
Department of Economics
University of Texas

Stephen G. Donald
Department of Economics
University of Texas

This version: November 2014

Abstract

Missing data is one of the most common challenges facing empirical researchers. This paper presents a general GMM framework for dealing with missing data on explanatory variables or instrumental variables. For a linear-regression model with missing covariate data, an efficient GMM estimator under minimal assumptions on missingness is proposed. The estimator, which also allows for a specification test of the missingness assumptions, is compared to linear imputation methods and a dummy-variable approach commonly used in empirical research. For an instrumental-variables model with potential missingness of the instrument, the GMM framework suggests a rich set of instruments that can be used to improve efficiency. Simulations and empirical examples are provided to compare the GMM approach with existing approaches.

JEL Classification: C13, C30

Keywords: Missing observations, imputation, projections, GMM, instrumental variables.

*We are grateful to Shu Shen for excellent research assistance and to Garry Barrett, Robert Moffitt, and seminar participants at several institutions for their helpful comments. We especially would like to thank Yu-Chin Hsu, who pointed out a mistake in an earlier draft.

1 Introduction

A common feature of many data sets used in empirical research is that of missing information for certain variables. An example, and the focus of this paper, is the situation where an explanatory variable (or covariate or regressor) may be unavailable for large portions of the observational units. If the variable with missing observations is considered an important part of the model, simply omitting the variable from the analysis brings with it the possibility of substantial omitted variables bias. In practice as we document below, researchers deal with this problem using one of three methods. The first, and simplest, is to discard observations that have missing data on the covariate – we usually refer to this as the "complete data method". When data are missing for a large portion of observations this could result in a much smaller sample. A second common method, that we refer to as the "dummy variable method", is to enter a zero for the missing information and to include a dummy variable for "missingness" as an additional explanatory variable to "account" for the missing information. A third method is some form of imputation where the missing information is "estimated" based on the other available data. Linear imputation is a simple version of this approach whereby a linear regression using the complete case observations is used to "impute" the missing values.

To give a sense of the prevalence of missing data in empirical research as well as the popularity of these two methods of dealing with missing data, Table 1 provides some summary statistics for four top empirical economics journals (*Amerian Economic Review (AER)*, *Journal of Human Resources (JHR)*, *Journal of Labor Economics (JLE)*, and *Quarterly Journal of Economics (QJE)*) over a three-year period (2006-2008).[1] Over half of the empirical papers in JLE and QJE have a missing-data issue, and nearly 40% of all papers across the four journals having data missingness. Of the papers with missing data, a large majority (roughly 70%) report that they have dropped observations due to missing values and hence use the complete data method (usually OLS). Both the "dummy-variable method" and the "imputation method"

[1]To identify data missingness, we searched for the word "missing" within the full text of an article and, if found, read through the data description to check if the author(s) mentioned having observations with missing values. The method(s) used to deal with missingness were inferred from the data description and/or empirical results section.

Table 1: Data missingness in economics journals, 2006-2008

Journal	Empirical papers	Papers with missing data (% of empirical papers)	Method of handling missing data[a] (% of missing-data papers in parentheses)		
			Drop observations	Use indicator variables for missingness	Use an imputation method[b]
American Economic Review[c]	191	55 (28.8%)	40 (72.7%)	9 (16.4%)	14 (25.5%)
Journal of Human Resources	94	40 (42.6%)	26 (65.0%)	10 (25.0%)	6 (15.0%)
Journal of Labor Economics	52	26 (50.0%)	18 (69.2%)	4 (15.4%)	5 (19.2%)
Quarterly Journal of Economics	79	41 (51.9%)	29 (70.7%)	8 (19.5%)	10 (24.4%)
Total	416	162 (38.9%)	113 (69.8%)	31 (19.1%)	35 (21.6%)

[a] A given paper may use more than one method, so the percentages add up to more than 100%.
[b] This column includes any type of imputation methods (regression-based, using past/future values, etc.).
[c] Includes *Papers & Proceedings* issues.

are quite common approaches to handling missing data, with each being used in roughly 20% of the missing-data papers.

The choice of method comes down to considerations of bias and relative efficiency. If data are missing in a way that can be considered "missing at random", which we formalize below, then there is no bias from using just the complete data method[2]. When data are missing for a small portion of observations this may be the preferred method but if data are missing for many observations then one may want to somehow use the observations with missing data to improve estimation in terms of relative efficiency. A recent paper by Dardanoni et. al. (2011) has shown in a very general setting that there will generally be no gains relative to using the complete data unless certain types of restrictions are imposed. One approach aimed at improving precision of estimates is the linear imputation method. This method estimates a regression model of the covariate with missing values as a linear projection onto the covariates that are fully observed using the complete data. The estimated regression is then used to impute missing values. Again, some form of "missing at random" assumption is needed for the method to be consistent and

[2] We use the terms "complete data" and "non-missing data" interchangably.

the implicit restriction that makes this more efficient is that the regression model between the regressors is the same in the complete data observations as it is in the observations with missing regressor values. Two versions of this that have been discussed in econometrics are the weighted version (or essentially GLS version) of Dagenais (1973) and the unweighted version of Gourieroux and Monfort (1981). As noted below, even with the restriction the unweighted version of linear imputation is potentially but not necessarily more efficient than just using complete data. The other popular method for dealing with missing observations, the dummy variable method, is known to be generally inconsistent unless one is dealing with simple linear regression or else requires that the regressor with missing values is orthogonal to the remaining regressors (Jones (1996)). As noted in Dardononi et. al. (2011) this method generally does not fully "account" for missingness. Moreover, as we show below, there is not even a guarantee of improvements in terms of relative efficiency as compared to the complete data estimator.[3]

The issue of efficiency in missing data models has been considered in the theoretical statistics literature. Robins, Rotnizky and Zhao (1994) deal explicitly with the situation of missing regressors in regression analysis and under quite general assumptions derive the form of semiparametric efficient estimators (see also the more recent work by Chaudhuri and Guilkey (forthcoming)). Their framework also allows for missingness to depend on the dependent variable in the regression model, in which case inverse probability weighting of the score is needed to restore consistency. Their paper provides the optimal way to use non-missing observations and the observations with missing values. A recent paper by Graham, Pinto and Egel (2012) provides a similar estimator that is shown to have some improved higher order bias properties compared to the estimator proposed in Robins et. al. (1994). Even in the case of "missing at random" which is a special case, these methods requires the empirical researcher to deal with models for the missingness probability as well as certain conditional expectations of the score. While only one of these needs to be correctly specified in order to achieve consistency (a so called double robustness property), the methods appear to be difficult to implement in practice. In their most general form, the method requires estimation of several nonparametric objects

[3]A guess as to the reason for the popularity of the "dummy variable method" is that it was discussed in an early version of a standard texbook Greene (2008). There it was referred to as the "modified zero order regression" and was discussed in the context of a simple linear regression model where the method is consistent.

that are difficult to model and for which empirical researchers are unlikely to have much prior information. In our survey of empirical work in economics, we were unable to find applications of the methods.

In view of the apparent difficulty of implementing fully efficient methods, this paper focuses on the linear imputation type methods that lead to efficiency gains under "missing at random" assumptions. We revisit the linear imputation method and develop a one-step Generalized Method of Moments (GMM) procedure which involves the linear imputation model as a set of moment conditions. Using standard results for GMM estimation, we show that there are situations where the GMM estimator yields variance reductions relative to the OLS estimator using just complete data for a subset, and sometimes for all, of the coefficients of interest. As in the earlier linear imputation methods, the efficicncy gains are obtained using the restriction that the imputation regression in the complete data is the same as that in observations with missing regressor values. This is made explicit in the GMM framework and appears as a set of overidentifying restrictions. As a by-product we are able to test the validity of these restrictions using a standard overidentifying restrictions test. We also compare the GMM approach to the linear imputation methods proposed in Dagenais (1973) and Gourieroux and Monfort (1981) and show that the GMM estimator is generally at least as efficient as these earlier alternatives. In certain more restrictive situations, which essentially amount to homoskedasticity-type assumptions, we show that the GMM estimator is asymptotically equivalent to the weighted imputation estimator of Dagenais (1973). The unweighted linear imputation estimator discussed in Gourieroux and Monfort (1981) is not necessarily a more efficient estimator than OLS using complete data and is usually strictly dominated by the GMM estimator. Moreover, the apparent computational simplicity of computing this estimator is possibly offset by the fact that appropriate standard errors are not obtained from the usual OLS standard error formulae. We also examine the assumptions implicit in the dummy variable method and note that, as previously shown by Jones (1996), it is potentially inconsistent even under the assumption that the regressor is "missing at random." Moreover, even when the assumptions for consistency are met, the dummy variable method may in some cases actually be less efficient than the complete

data method. Our results provide insight into conditions that are needed for efficiency gains to be possible.

The paper is structured as follows. Section 2 introduces the model, notation, and assumptions. These involve the regression relationship of interest as well as the general linear projection relationship between the regressors. We then develop a set of moment conditions for the observed data and show that an optimally weighted GMM estimator that uses these conditions will bring efficiency gains in general. Section 3 compares the GMM estimator to estimators previously considered in the literature. Section 4 extends the GMM approach to situations where there are missing data in an instrumental variables model (either for the instrumental variable or the endogenous variable). The full set of instruments implied by the assumptions on missingness offer the possibility of efficiency gains. Section 5 considers a simulation study in which the GMM approach is compared to other methods in finite samples. Section 6 reports results from two empirical examples, the first a standard regression (with missing covariate) example and the second an instrumental-variables (with missing instrument) example. Section 7 concludes. Detailed proofs of the paper's results are provided in a Technical Appendix (supplemental materials).

2 Model Assumptions and Moment Conditions

Consider the following standard linear regression model

$$Y_i = X_i\alpha_0 + Z_i'\beta_0 + \varepsilon_i = W_i'\theta_0 + \varepsilon_i \qquad i = 1, ..., n \tag{2.1}$$

where X_i is a (possibly missing) scalar regressor, Z_i is a K-vector of (never missing) regressors, and $W_i \equiv (X_i, Z_i')'$. The first element of Z_i is 1; that is, the model is assumed to contain an intercept. We assume the residual only satisfies the conditions for (2.1) to be a linear projection, specifically

$$E(X_i\varepsilon_i) = 0 \text{ and } E(Z_i\varepsilon_i) = 0. \tag{2.2}$$

The variable m_i indicates whether or not X_i is missing for observational unit i:

$$m_i = \begin{cases} 1 & \text{if } X_i \text{ missing} \\ 0 & \text{if } X_i \text{ observed} \end{cases}$$

We assume the existence of a linear projection of X_i onto Z_i,

$$X_i = Z_i'\gamma_0 + \xi_i \text{ where } E(Z_i\xi_i) = 0. \tag{2.3}$$

Provided that X_i and the elements of Z_i have finite variances and that the variance-covariance matrix of (X_i, Z_i') is nonsingular, the projection in (2.3) is unique and completely general in the sense that it does not place any restrictions on the joint distribution of (X_i, Z_i'). Also, no homoskedasticity assumptions are imposed on ξ_i or ε_i, though the nature of the results under homoskedasticity is discussed below.

Observations with missing X_i are problematic since (2.1) cannot be used directly to construct moment conditions for estimating $\theta_0 \equiv (\alpha_0, \beta_0')'$ — all that we see for such observations is the combination (Y_i, Z_i'). Note, however, that (2.1) and (2.3) imply

$$Y_i = Z_i'(\gamma_0\alpha_0 + \beta_0) + \varepsilon_i + \xi_i\alpha_0 \stackrel{def}{=} Z_i'(\gamma_0\alpha_0 + \beta_0) + \eta_i. \tag{2.4}$$

For this relationship to be useful in estimation, we require an assumption on the missingness variable m_i. This is our version of the "missing at random" (MAR) assumption on m_i:

Assumption 1 *(i)* $E(m_iZ_i\varepsilon_i) = 0$*; (ii)* $E(m_iZ_i\xi_i) = 0$*; (iii)* $E(m_iX_i\varepsilon_i) = 0$.

Several remarks are in order. First, the complete data estimator (defined explicitly below) also requires conditions (i) and (iii) (but not (ii)) of Assumption 1 in order to be consistent. Second, the conditions of Assumption 1 are weaker than assuming that m_i is independent of the unobserved variables and will be satisfied when $Z_i\varepsilon_i$, $Z_i\xi_i$, and $X_i\varepsilon_i$ are mean independent of m_i. Of course, assuming that m_i is statistically independent of $(X_i, Z_i, \varepsilon_i, \xi_i)$ will imply the conditions in Assumption 1; such an assumption is generally known as "missing completely at random" (MCAR). Assumption 1 allows for m_i to depend on the explanatory variables and other unobserved factors under certain conditions; for example, suppose that

$$m_i = 1(h(Z_i, v_i) > 0)$$

for some arbitrary function h so that missingness of X_i depends on the other explanatory variables as well as an unobserved factor v_i. For this missingness mechanism, Assumption 1 will be satisfied when v_i is independent of ε_i and ξ_i conditional on W_i, along with $E(\varepsilon_i|W_i) = 0$ and $E(\xi_i|Z_i) = 0$.[4]

We define a vector of moment functions based upon (2.2), (2.3), and (2.4),

$$g_i(\alpha, \beta, \gamma) = \begin{pmatrix} (1 - m_i)W_i(Y_i - X_i\alpha - Z_i'\beta) \\ (1 - m_i)Z_i(X_i - Z_i'\gamma) \\ m_i Z_i \left(Y_i - Z_i' \left(\gamma\alpha + \beta\right)\right) \end{pmatrix} = \begin{pmatrix} g_{1i}(\alpha, \beta, \gamma) \\ g_{2i}(\alpha, \beta, \gamma) \\ g_{3i}(\alpha, \beta, \gamma) \end{pmatrix}, \tag{2.5}$$

for which the following result holds:

Lemma 1 *Under Assumption 1,* $E(g_i(\alpha_0, \beta_0, \gamma_0)) = 0$.

Lemma 1 implies that the model and Assumption 1 generate a vector of $3K + 1$ moment conditions satisfied by the population parameter values $(\alpha_0, \beta_0, \gamma_0)$. Since there are $2K + 1$ parameters, there are K overidentifying restrictions — it is the availability of these overidentifying restrictions that provides a way of more efficiently estimating the parameters of interest. Indeed, as the following result shows, the use of a subset of the moment conditions that consists of g_{1i} and either g_{2i} or g_{3i} (but not both) results in an estimator for θ_0 that is identical to the "complete data estimator" using only g_{1i}, given by

$$\hat{\theta}_C = \left(\sum_{i=1}^{n}(1 - m_i)W_iW_i'\right)^{-1} \sum_{i=1}^{n}(1 - m_i)W_iY_i.$$

Lemma 2 *GMM estimators of* $\theta_0 = (\alpha_0, \beta_0')'$ *based on the moments* $(g_{1i}(\alpha, \beta, \gamma)', g_{2i}(\alpha, \beta, \gamma)')'$ *or the moments* $(g_{1i}(\alpha, \beta, \gamma)', g_{3i}(\alpha, \beta, \gamma)')'$ *are identical to the complete data estimator* $\hat{\theta}_C$.

This result and the general proposition that adding valid moment conditions cannot reduce asymptotic variance give rise to the possibility of efficiency gains from using the complete set of moment conditions.

Note that the result in Dardanoni et. al. (2011) concerning equivalence between general imputation methods and the complete data method is based on imputation coefficients possibly

[4]See Griliches (1986) for additional discussion on the relationship between m_i and the model.

differing between missing and non-missing observations, thus violating (ii) of Assumption 1. In their setup, one would allow for different γ vectors in the g_{2i} and g_{3i} moments, leading to GMM also being equivalent to the complete data estimator. Relative to Dardanoni et. al. (2011), the potential efficiency gains from GMM can be seen as coming from the restriction that γ is the same for missing and non-missing data (as implied by Assumption 1). As shown below, a byproduct of the GMM framework is a straightforward and robust overidentification test of this restriction.

We now consider the asymptotic variance of the standard optimally weighted GMM procedure. The optimal weight matrix for such a procedure is the inverse of the variance-covariance matrix of the moment function evaluated at the true values of the parameters,

$$\Omega = E(g_i(\alpha_0, \beta_0, \gamma_0) g_i(\alpha_0, \beta_0, \gamma_0)') = \begin{pmatrix} \Omega_{11} & \Omega_{12} & 0 \\ \Omega_{12}' & \Omega_{22} & 0 \\ 0 & 0 & \Omega_{33} \end{pmatrix}, \tag{2.6}$$

where

$$\Omega_{11} = E((1-m_i) W_i W_i' \varepsilon_i^2),\ \Omega_{22} = E((1-m_i) Z_i Z_i' \xi_i^2),$$
$$\Omega_{12} = E((1-m_i) W_i Z_i' \varepsilon_i \xi_i),\ \Omega_{33} = E(m_i Z_i Z_i' \eta_i^2).$$

The zero components in (2.6) follow from $m_i(1-m_i) = 0$.

To implement the optimally weighted GMM procedure, we take sample analogs of the three blocks and estimate the residuals using a preliminary consistent procedure:

$$\begin{aligned} \hat{\Omega}_{11} &= \frac{1}{n} \sum_i (1-m_i) W_i W_i' \hat{\varepsilon}_i^2,\ \hat{\Omega}_{22} = \frac{1}{n} \sum_i (1-m_i) Z_i Z_i' \hat{\xi}_i^2, \\ \hat{\Omega}_{12} &= \frac{1}{n} \sum_i (1-m_i) W_i Z_i' \hat{\varepsilon}_i \hat{\xi}_i,\ \hat{\Omega}_{33} = \frac{1}{n} \sum_i m_i Z_i Z_i' \hat{\eta}_i^2. \end{aligned}$$

In the simulations in Section 5, for instance, $\hat{\varepsilon}_i$ and $\hat{\xi}_i$ are estimated from the complete data regressions of Y_i on W_i and X_i on Z_i, respectively, and $\hat{\eta}_i$ is estimated from a regression of Y_i on Z_i using observations with missing X_i.

The optimal two-step GMM estimators, denoted by $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$, solve

$$\min_{\alpha,\beta,\gamma} \bar{g}(\alpha, \beta, \gamma)' \hat{\Omega}^{-1} \bar{g}(\alpha, \beta, \gamma), \tag{2.7}$$

where $\bar{g}(\alpha,\beta,\gamma) = n^{-1}\sum_{i=1}^{n} g_i(\alpha,\beta,\gamma)$ and $\hat{\Omega}$ is the estimator of Ω obtained by plugging $\hat{\Omega}_{11}$, $\hat{\Omega}_{22}$, $\hat{\Omega}_{12}$, and $\hat{\Omega}_{33}$ into (2.6). Although this GMM estimator is nonlinear in its parameters and therefore requires numerical methods, our simulations show that this type of problem is well-behaved and can be easily optimized in Stata (Version 11 or later) and other econometrics packages.

As is well known, and as stated in Proposition 1 below, the other component in the variance covariance matrix for the optimally weighted GMM is the gradient matrix corresponding to the moment functions. In this instance, this is given by

$$G = \begin{pmatrix} G_{11} & 0 \\ 0 & G_{22} \\ G_{31} & G_{32} \end{pmatrix} \tag{2.8}$$

where the components of the matrix are

$$\begin{aligned} G_{11} &= -E((1-m_i)W_iW_i'), \\ G_{22} &= -E((1-m_i)Z_iZ_i'), \\ G_{31} &= \begin{pmatrix} -E(m_iZ_iZ_i'\gamma_0) & -E(m_iZ_iZ_i') \end{pmatrix}, \\ G_{32} &= -E(m_iZ_iZ_i'\alpha_0). \end{aligned}$$

The first set of columns represent the expectation of the derivatives of the moment functions with respect to $(\alpha,\beta')'$ while the second set of columns are related to the derivatives with respect to γ. For purposes of inference, one can easily estimate the components of G by taking sample analogs evaluated at the GMM estimates of all the parameters.

Under standard regularity conditions, we have the following result:[5]

Proposition 1 *Under Assumption 1, the estimators $(\hat{\alpha},\hat{\beta},\hat{\gamma})$ are consistent and asymptotically normally distributed with asymptotic variance given by $(G'\Omega^{-1}G)^{-1}$. Moreover,*

$$n\bar{g}(\hat{\alpha}_G,\hat{\beta}_G,\hat{\gamma}_G)'\hat{\Omega}^{-1}\bar{g}(\hat{\alpha}_G,\hat{\beta}_G,\hat{\gamma}_G) \xrightarrow{d} \chi^2(K) \tag{2.9}$$

[5]By standard regularity conditions, we simply mean the finite variances and nonsingularity that allow for the unique representations in (2.1) and (2.3). These identification conditions will be implicitly assumed throughout the exposition below.

The behavior of the objective function in (2.9) gives rise to the possibility of testing the overidentifying restrictions imposed by Assumption 1. For instance, as discussed above, Assumption 1 would be violated when $W_i\varepsilon_i$ has nonzero expectation for observations where X_i is (non-)missing. Similarly, the assumption would be violated when $Z_i\xi_i$ has nonzero expectation for observations where X_i is (non-)missing. Since, as shown below, improvements in terms of efficiency can only be achieved when these restrictions are satisfied and imposed on estimation, the method provides a way of testing the validity of these assumptions.

Using the same notation, the asymptotic variance of the complete data estimator is given in the following result:

Lemma 3 *Under Assumption 1, the complete data estimator has asymptotic variance given by*

$$AVAR(\sqrt{n}(\hat{\theta}_C - \theta_0)) = (G_{11}\Omega_{11}^{-1}G_{11})^{-1} = G_{11}^{-1}\Omega_{11}G_{11}^{-1}.$$

As noted earlier, the fact that the complete data estimator $\hat{\theta}_C$ is also equivalent to the GMM estimator of θ_0 based on a subset of moment conditions (Lemma 2) implies that the optimally weighted GMM estimator using all $3K + 1$ moment conditions has a variance no larger than that given in Lemma 3. The remainder of this Section examines in detail when efficiency gains are possible with the GMM approach.

Some additional notation is needed. First, denote $\lambda = P(m_i = 0)$, so that in the asymptotic theory λ represents the asymptotic proportion of data that have non-missing X_i. Also, using the subscript m (or c) to denote an expectation for observations with missing (or non-missing) X_i values, define the following quantities:

$$\begin{aligned}
\Gamma_m = E(Z_iZ_i'|m_i = 1) \quad &\text{and} \quad \Gamma_c = E(Z_iZ_i'|m_i = 0) \\
\Omega_{\eta\eta m} = E(Z_iZ_i'\eta_i^2|m_i = 1) \quad &\text{and} \quad \Omega_{\eta\eta c} = E(Z_iZ_i'\eta_i^2|m_i = 0) \\
\Omega_{\varepsilon\eta m} = E(Z_iZ_i'\varepsilon_i\eta_i|m_i = 1) \quad &\text{and} \quad \Omega_{\varepsilon\eta c} = E(Z_iZ_i'\varepsilon_i\eta_i|m_i = 0) \\
\Lambda_{\varepsilon\eta\xi m} = E(Z_i\varepsilon_i\eta_i\xi_i|m_i = 1) \quad &\text{and} \quad \Lambda_{\varepsilon\eta\xi c} = E(Z_i\varepsilon_i\eta_i\xi_i|m_i = 0) \\
\sigma_{\xi m}^2 = E(\xi_i^2|m_i = 1) \quad &\text{and} \quad \sigma_{\xi c}^2 = E(\xi_i^2|m_i = 0)
\end{aligned}$$

The following general result characterizes the difference in asymptotic variances between the efficient GMM estimator and the complete data estimator.

Proposition 2 *Under Assumption 1,*

$$AVAR(\sqrt{n}(\hat{\theta}_C - \theta_0)) - AVAR(\sqrt{n}(\hat{\theta} - \theta_0)) = \begin{pmatrix} A' \\ B' \end{pmatrix} D \begin{pmatrix} A & B \end{pmatrix} \geq 0,$$

where

$$\begin{aligned} A &= \Lambda_{\varepsilon\eta\xi c} \left(\sigma^2_{\xi c}\right)^{-1}, \\ B &= \Omega_{\varepsilon\eta c}\Gamma_c^{-1} - \Lambda_{\varepsilon\eta\xi c} \left(\sigma^2_{\xi c}\right)^{-1} \gamma_0', \\ D &= \frac{1}{\lambda^2}\Gamma_c^{-1}\left(\frac{1}{(1-\lambda)}\Gamma_m^{-1}\Omega_{\eta\eta m}\Gamma_m^{-1} + \frac{1}{\lambda}\Gamma_c^{-1}\Omega_{\eta\eta c}\Gamma_c^{-1}\right)^{-1}\Gamma_c^{-1}. \end{aligned}$$

Since the matrix D is positive definite under fairly general conditions, it becomes straightforward to consider separately situations in which there is a reduction in variance for estimation of α_0 and situations in which there is a reduction in variance for estimation of β_0. The difference corresponding to estimation of α_0 is given by

$$\left(\sigma^2_{\xi c}\right)^{-2} \Lambda'_{\varepsilon\eta\xi c} D \Lambda_{\varepsilon\eta\xi c} \geq 0$$

which is equal to zero if and only if $\Lambda_{\varepsilon\eta\xi c} = 0$. On the other hand, for estimation of β_0, the difference is given by $B'DB \geq 0$, which is equal to zero if and only if

$$B = \Omega_{\varepsilon\eta c}\Gamma_c^{-1} - \Lambda_{\varepsilon\eta\xi c} \left(\sigma^2_{\xi c}\right)^{-1} \gamma_0' = 0.$$

Given the definition of η_i, we can write

$$\begin{aligned} \Lambda_{\varepsilon\eta\xi c} &= E(Z_i\varepsilon_i^2\xi_i|m_i = 0) + \alpha_0 E(Z_i\varepsilon_i\xi_i^2|m_i = 0) \\ &\stackrel{def}{=} \Lambda_{\varepsilon\varepsilon\xi c} + \alpha_0\Lambda_{\varepsilon\xi\xi c} \end{aligned} \tag{2.10}$$

and

$$\begin{aligned} \Omega_{\varepsilon\eta c} &= E(Z_i Z_i'\varepsilon_i^2|m_i = 0) + \alpha_0 E(Z_i Z_i'\varepsilon_i\xi_i|m_i = 0) \\ &\stackrel{def}{=} \Omega_{\varepsilon\varepsilon c} + \alpha_0\Omega_{\varepsilon\xi c}. \end{aligned} \tag{2.11}$$

In contrast to the previous literature, this general result suggests that efficiency gains are possible for *both* α_0 and β_0 since the general assumptions do not imply either $\Lambda_{\varepsilon\eta\xi c} = 0$ or $B = 0$. The previous literature (eg Dagenais (1973) and Gourieroux and Monfort (1981)) either explicitly or implicitly imposed stronger assumptions on the model and revealed only the possibility of improvements for the estimation of β_0 using imputation methods discussed below. The above result indicates situations when improvements for α_0 will be possible and why, under more restrictive classical assumptions, improvements will not be available.

A simple more restrictive assumption is given by the following:

Assumption 2 *Suppose that* $E(\varepsilon_i|X_i, Z_i, m_i = 0) = 0$ *and* $E(\varepsilon_i^2|X_i, Z_i, m_i = 0) = \sigma_{\varepsilon c}^2$.

This assumption implies a classical regression model with mean zero and heteroskedastic residuals, as implied by the Gaussian-like assumptions made in the above-referenced literature. Under Assumption 2, based upon (2.10), note that $\Lambda_{\varepsilon\eta\xi c} = 0$ since

$$E(Z_i\varepsilon_i^2\xi_i|m_i = 0) = E(E(Z_i\varepsilon_i^2\xi_i|X_i, Z_i, m_i = 0)|m_i = 0) = \sigma_\varepsilon^2 E(Z_i\xi_i|m_i = 0) = 0$$

and

$$E(Z_i\varepsilon_i\xi_i^2|m_i = 0) = E(E(Z_i\varepsilon_i\xi_i^2|X_i, Z_i, m_i = 0)|m_i = 0) = E(Z_i\xi_i^2 E(\varepsilon_i|X_i, Z_i, m_i = 0)|m_i = 0) = 0.$$

Also, note that B simplifies in this case:

$$E(Z_iZ_i'\varepsilon_i^2|m_i = 0) = \sigma_{\varepsilon c}^2\Gamma_c$$
$$E(Z_iZ_i'\varepsilon_i\xi_i|m_i = 0) = E(E(Z_iZ_i'\varepsilon_i\xi_i|X_i, Z_i, m_i = 0)|m_i = 0) = E(Z_iZ_i'\xi_i E(\varepsilon_i|X_i, Z_i, m_i = 0)|m_i = 0) = 0$$

Therefore, under the more restrictive Assumption 2, the following results hold:

Lemma 4 *Under Assumptions 1 and 2,*

$$\begin{aligned} AVAR(\sqrt{n}(\hat{\alpha}_C - \alpha_0)) - AVAR(\sqrt{n}(\hat{\alpha} - \alpha_0)) &= 0 \\ AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) - AVAR(\sqrt{n}(\hat{\beta} - \beta_0)) &= \left(\sigma_{\varepsilon c}^2\right)^2 D > 0 \end{aligned}$$

Consistent with the previous literature, this result shows that under the classical assumptions on the residual ε_i there will be no gain in terms of estimating α_0 but generally there will be

a gain for β_0. As discussed below, additional assumptions made in the previous literature also simplify D relative to the general expression in Proposition 2.

Before considering the further restrictions that lead Lemma 4 to coincide with results in the previous literature, we return to the general case to explore situations where gains for both α_0 and β_0 are possible. To illustrate, we consider a specification for the residuals that permits flexible forms of scale heteroskedasticity:

Assumption 3 *$\varepsilon_i = \sigma_\varepsilon(X_i, Z_i)u_i$ and $\xi_i = \sigma_\xi(Z_i)v_i$, where (u_i, v_i) are jointly i.i.d. and independent of (X_i, Z_i), with u_i and v_i both having mean zero and variance one.*

This assumption implies Assumption 2 when $\sigma_\varepsilon(X_i, Z_i) = \sigma_\varepsilon$ and $E(u_i|v_i, Z_i, m_i = 0) = 0$. The two key terms with respect to estimation of α_0 are

$$\begin{aligned}\Lambda_{\varepsilon\varepsilon\xi c} &= E(Z_i\sigma_\varepsilon(Z_i'\gamma_0 + \sigma_\xi(Z_i)v_i, Z_i)^2 u_i^2 \sigma_\xi(Z_i)v_i|m_i = 0)\\ \Lambda_{\varepsilon\xi\xi c} &= E(Z_i\sigma_\varepsilon(Z_i'\gamma_0 + \sigma_\xi(Z_i)v_i, Z_i)u_i\sigma_\xi(Z_i)^2 v_i^2|m_i = 0)\end{aligned}$$

Neither term is zero without further restrictions. The second quantity, $\Lambda_{\varepsilon\xi\xi c}$, will equal zero with the additional condition $E(u_i|Z_i, v_i, m_i = 0) = 0$, which essentially corresponds to the condition that $E(\varepsilon_i|X_i, Z_i, m_i = 0) = 0$ as would be the case under Assumption 2. Otherwise, this term is not necessarily zero. Also, for the first quantity $\Lambda_{\varepsilon\varepsilon\xi c}$, if heteroskedasticity in ε_i is limited to dependence on just Z_i so that $\sigma_\varepsilon(X_i, Z_i) = \sigma_\varepsilon(Z_i)$, then

$$\begin{aligned}\Lambda_{\varepsilon\varepsilon\xi c} &= E(Z_i\sigma_\varepsilon(Z_i)^2 u_i^2 \sigma_\xi(Z_i)v_i|m_i = 0)\\ &= E(Z_i\sigma_\varepsilon(Z_i)^2\sigma_\xi(Z_i)E(u_i^2 v_i|Z_i, m_i = 0)|m_i = 0)\end{aligned}$$

which will be zero when $E(u_i^2 v_i|Z_i, m_i = 0) = 0$. For example, a bivariate normal distribution of (u_i, v_i) would satisfy this property.

Considering both $\Lambda_{\varepsilon\varepsilon\xi c}$ and $\Lambda_{\varepsilon\xi\xi c}$, the following result provides sufficient conditions for no efficiency gains with respect to α_0:

Lemma 5 *If Assumptions 1 and 3 hold and, furthermore, (i)* $\sigma_\varepsilon(X_i, Z_i) = \sigma_\varepsilon(Z_i)$*, (ii)* $E(u_i^2 v_i | m_i = 0) = 0$*, and (iii)* $E(u_i v_i^2 | m_i = 0) = 0$ *hold, then*

$$\begin{aligned} AVAR(\sqrt{n}(\hat{\alpha}_C - \alpha_0)) - AVAR(\sqrt{n}(\hat{\alpha} - \alpha_0)) &= 0 \\ AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) - AVAR(\sqrt{n}(\hat{\beta} - \beta_0)) &= \Gamma_c^{-1}\Omega_{\varepsilon\eta c} D \Omega_{\varepsilon\eta c}\Gamma_c^{-1} > 0 \end{aligned}$$

There are still generally improvements for estimation of β_0 with the main simplification coming from the fact that the second term in B is zero. Even when $\sigma_\varepsilon(Z_i)$ and $\sigma_\xi(Z_i)$ are constants (so that (i) holds), efficiency gains for α_0 are possible when the the joint third-moment conditions in (ii) and/or (iii) are violated.

For the sake of completeness and in order to compare the GMM method and the complete data method with alternative imputation methods proposed in Dagenias (1973) and Gourieroux and Monfort (1981), we derive explicit expressions for the asymptotic variances under the classical assumptions in those papers. To do this, define

$$\sigma_\varepsilon^2 = E(\varepsilon_i^2),\ \sigma_\xi^2 = E(\xi_i^2)$$

and let $\Omega_{\varepsilon\varepsilon m}$, $\Omega_{\xi\xi m}$ and $\Omega_{\xi\xi c}$ denote the matrix of moments analogous to the definitions of $\Omega_{\varepsilon\varepsilon c}$ and $\Omega_{\eta\eta m}$ above. With this notation, the following assumption is essentially equivalent to their assumptions:

Assumption 4 *(i) The residuals* ε_i *and* ξ_i *are conditionally (on* (X_i, Z_i, m_i) *and* (Z_i, m_i)*, respectively) mean zero and homoskedastic, (ii)* $\Gamma_m = \Gamma_c = \Gamma$*, (iii)* $\Omega_{\varepsilon\varepsilon m} = \Omega_{\varepsilon\varepsilon c} = \sigma_\varepsilon^2\Gamma$*, (iv)* $\Omega_{\xi\xi m} = \Omega_{\xi\xi c} = \sigma_\xi^2\Gamma$.

Conditions (ii)-(iv) will be satisfied, given (i), when X_i data are MCAR (see, e.g., Gourieroux and Monfort (1981) or Nijman and Palm (1988)). Under Assumption 4, the asymptotic-variance expressions for the complete data estimator simplify to

$$AVAR(\sqrt{n}(\hat{\alpha}_C - \alpha_0)) = \frac{1}{\lambda}\sigma_\varepsilon^2\left(\sigma_\xi^2\right)^{-1} \tag{2.12}$$

$$AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) = \frac{1}{\lambda}\sigma_\varepsilon^2\Gamma^{-1} + \frac{1}{\lambda}\sigma_\varepsilon^2\left(\sigma_\xi^2\right)^{-1}\gamma_0\gamma_0', \tag{2.13}$$

while for the GMM estimator,

$$AVAR(\sqrt{n}(\hat{\alpha}-\alpha_0)) = \frac{1}{\lambda}\sigma_\varepsilon^2\left(\sigma_\xi^2\right)^{-1} \tag{2.14}$$

$$AVAR(\sqrt{n}(\hat{\beta}-\beta_0)) = \sigma_\varepsilon^2\left(1+\frac{(1-\lambda)\sigma_\xi^2\alpha_0^2}{\lambda\left(\sigma_\varepsilon^2+\sigma_\xi^2\alpha_0^2\right)}\right)\Gamma^{-1}+\frac{1}{\lambda}\sigma_\varepsilon^2\left(\sigma_\xi^2\right)^{-1}\gamma_0\gamma_0'. \tag{2.15}$$

Comparing the first terms in (2.13) and (2.15), one can see the factors that affect the efficiency improvement from GMM estimation of β_0. The result in Lemma 5 regarding estimation of α_0 is also evident.

3 Comparison to Other Methods

3.1 Linear Imputation

This section compares the general GMM procedure with the linear imputation method proposed by Dagenais (1973) and discussed further by Gourieroux and Monfort (1981). First, note that plugging the projection (2.3) into the regression model (2.1) yields

$$Y_i = \left((1-m_i)X_i + m_iZ_i'\gamma_0\right)\alpha_0 + Z_i'\beta_0 + \varepsilon_i + m_i\xi_i\alpha_0, \tag{3.16}$$

where by Assumption 1 the composite residual $\varepsilon_i + m_i\xi_i\alpha_0$ is orthogonal to the regressors $((1-m_i)X_i, Z_i', m_iZ_i')$. The methods discussed in these earlier papers can be thought of as a sequential approach where one first estimates γ_0 using the second moment condition by

$$\hat{\gamma} = \left(\sum_{i=1}^n (1-m_i)Z_iZ_i'\right)^{-1}\sum_{i=1}^n (1-m_i)Z_iX_i$$

and then substituting into (3.16) and using regression based methods.

Given an estimate $\hat{\gamma}$ version of γ_0, the operational version of (3.16) is

$$Y_i = \left((1-m_i)X_i + m_iZ_i'\hat{\gamma}\right)\alpha_0 + Z_i'\beta_0 + \varepsilon_i + m_i\xi_i\alpha_0 + m_iZ_i'\left(\gamma_0-\hat{\gamma}\right)\alpha_0. \tag{3.17}$$

Equation (3.17) is essentially a regression model with $\hat{X}_i = ((1-m_i)X_i + m_iZ_i'\hat{\gamma})$ and Z_i as regressors — that is, X_i is used if it is observed and the linearly imputed value $Z_i'\hat{\gamma}$ is used otherwise. OLS estimation can be used, but even under homoskedasticity assumptions on ε_i and

ξ_i the residual will be necessarily heteroskedastic due to: (i) $m_i\xi_i\alpha_0$ appearing for missing X_i, and (ii) estimation error in using $\hat{\gamma}$ in place of γ_0. Under the conditions of Dagenais (1973) and Gourieroux and Monfort (1981), which are essentially equivalent to Assumption 4, the residual variance is

$$\sigma_\varepsilon^2 + m_i\sigma_\xi^2\alpha_0^2 + \sigma_\xi^2\alpha_0^2 m_i Z_i' \left(\sum_{\ell=1}^{n}(1-m_\ell)Z_\ell Z_\ell'\right)^{-1} Z_i. \tag{3.18}$$

and the covariance across residuals for observations $i \neq j$ given by

$$\sigma_\xi^2\alpha_0^2 m_i m_j Z_i' \left(\sum_{\ell=1}^{n}(1-m_\ell)Z_\ell Z_\ell'\right)^{-1} Z_j. \tag{3.19}$$

Dagenais (1973) proposed a FGLS procedure for the estimation of (3.17) where one estimates (3.18) and (3.19) using preliminary estimates $\hat{\sigma}_\varepsilon^2$, $\hat{\sigma}_\xi^2$, and $\hat{\alpha}$ that would be consistent under the homoskedasticity assumptions. Gourieroux and Monfort (1981) use the label "Generalized Dagenais (GD) estimator" for this FGLS procedure and the label "Ordinary Dagenais (OD) estimator" for the standard OLS estimator.

The following result facilitates comparisons between GD and GMM by showing that GD behaves asymptotically like a GMM estimator with a particular weight matrix:

Proposition 3 (GD and GMM) *The GD estimator is asymptotically equivalent to GMM using all the moments in (2.5) and an estimated weight matrix based on (2.6) where* $\hat{\Omega}_{12} = 0$, *and*

$$\hat{\Omega}_{11} = \hat{\sigma}_\varepsilon^2 \frac{1}{n}\sum_i (1-m_i)W_i W_i', \; \hat{\Omega}_{22} = \hat{\sigma}_\xi^2 \frac{1}{n}\sum_i (1-m_i)Z_i Z_i', \; \hat{\Omega}_{33} = \left(\hat{\sigma}_\varepsilon^2 + \hat{\alpha}^2\hat{\sigma}_\xi^2\right)\frac{1}{n}\sum_i m_i Z_i Z_i'$$

using estimates $\hat{\sigma}_\varepsilon^2$, $\hat{\sigma}_\xi^2$ *and* $\hat{\alpha}$ *that have the same limits as those used in GD.*

This result suggests that GD behaves like GMM using a weight matrix that is optimal under the conditions in Assumption 4. When these conditions are satisfied, GMM and GD have the same asymptotic variance as given in (2.14) and (2.15). When they are violated, as would occur under heteroskedasticity, the general results for GMM imply that the general version of GMM laid out in the previous section will be at least as efficient as GD. We explore some of these gains in simulations that follow.

One can also compare GD to a sequential GMM procedure where one first estimates γ_0 using the second set of moment functions and then uses the other two moment functions with γ_0 replaced by $\hat{\gamma}$ in the third moment function. Specifically, the second stage uses the moment vector

$$\frac{1}{n}\sum_i \begin{pmatrix} (1-m_i)W_i(Y_i - X_i\alpha - Z_i'\beta) \\ m_i Z_i \left(Y_i - Z_i'\left(\hat{\gamma}\alpha + \beta\right)\right) \end{pmatrix} \stackrel{def}{=} \bar{g}_S(\alpha, \beta, \hat{\gamma}) \tag{3.20}$$

Under conditional homoskedasticity assumptions on ε_i and ξ_i, the optimal weight matrix for this GMM procedure can be estimated by

$$\begin{pmatrix} \hat{\Omega}_{11} & 0 \\ 0 & \hat{\Omega}_{33} + \hat{\alpha}^2\hat{\sigma}_\xi^2 \left(\frac{1}{n}\sum_i m_i Z_i Z_i'\right)\left(\frac{1}{n}\sum_i (1-m_i) Z_i Z_i'\right)^{-1}\frac{1}{n}\sum_i m_i Z_i Z_i' \end{pmatrix}^{-1} \stackrel{def}{=} \hat{\Omega}_S^{-1} \tag{3.21}$$

where $\hat{\Omega}_{11}$, $\hat{\Omega}_{33}$, $\hat{\alpha}^2$, and $\hat{\sigma}_\xi^2$ are as in Proposition 3. Note the second term in the lower right corner comes from the fact that one has used an estimate of γ_0.

Proposition 4 (GDE-SGMM) *The GD estimator is asymptotically equivalent to sequential GMM using the moments $\bar{g}_S(\alpha, \beta, \hat{\gamma})$and the weight matrix $\hat{\Omega}_S^{-1}$ that uses estimates $\hat{\sigma}_\varepsilon^2$, $\hat{\sigma}_\xi^2$ and $\hat{\alpha}$ that have the same limits as those used in GD.*

Again, the weight matrix for this sequential GMM procedure will generally be optimal only when the homoskedasticity assumptions are valid. This result facilitates a comparison with the OD estimator since that procedure is numerically identical to a sequential GMM procedure that uses as a weight matrix (in place of $\hat{\Omega}_S^{-1}$) $H'H$ where

$$H = \begin{pmatrix} 1 & 0 & \hat{\gamma}' \\ 0 & I & I \end{pmatrix}.$$

Note that

$$H\bar{g}_S(\alpha, \beta, \hat{\gamma}) = \frac{1}{n}\sum_i \begin{pmatrix} \hat{X}_i(Y_i - X_i\alpha - Z_i'\beta) \\ Z_i\left(Y_i - Z_i'\left(\hat{\gamma}\alpha + \beta\right)\right) \end{pmatrix},$$

which are the normal equations for OLS estimation of (3.17). Even under homoskedasticity assumptions on ε_i and ξ_i, the usual OLS standard errors are invalid due to the presence of pre-estimation error in the residual. Since, as we argue below, there is little reason to use the OD estimator, we omit the details of how one would properly estimate its standard errors.

Standard results for FGLS and GMM imply that the optimal GMM estimator will be at least as efficient as OD and that, under the stronger homoskedasticity assumptions, GD will be at least as efficient as OD. The efficiency comparisons are summarized in the following proposition:

Proposition 5 *Under Assumption 1,*

(i) $AVAR(\sqrt{n}(\hat{\alpha}-\alpha_0)) \leq AVAR(\sqrt{n}(\hat{\alpha}_{OD}-\alpha_0)) = AVAR(\sqrt{n}(\hat{\alpha}_{GD}-\alpha_0)) = AVAR(\sqrt{n}(\hat{\alpha}_C-\alpha_0))$

(ii) $AVAR(\sqrt{n}(\hat{\beta}-\beta_0)) \leq AVAR(\sqrt{n}(\hat{\beta}_{OD}-\beta_0))$

(iii) $AVAR(\sqrt{n}(\hat{\beta}-\beta_0)) \leq AVAR(\sqrt{n}(\hat{\beta}_{GD}-\beta_0))$

(iv) $AVAR(\sqrt{n}(\hat{\beta}_{GD}-\beta_0)) \leq AVAR(\sqrt{n}(\hat{\beta}_{OD}-\beta_0))$ *if Assumption 4(i) holds*

Neither the OD or GD estimators bring any improvements with respect to estimating α_0 whereas the results of Section 2 indicate that this is possible in some cases using GMM. With respect to β_0, one can show that even under the stronger homoskedasticity assumptions the OD estimator, unlike the GD estimator, is not guaranteed to bring about efficiency improvements relative to the complete data estimator. The asymptotic variance for $\hat{\beta}_{OD}$ under the stronger homogeneity and homoskedasticity assumptions in Assumption 4 is given by

$$\sigma_\varepsilon^2\left(1+\frac{(1-\lambda)\sigma_\xi^2\alpha_0^2}{\lambda\sigma_\varepsilon^2}\right)\Gamma^{-1}+\frac{1}{\lambda}\sigma_\varepsilon^2\left(\sigma_\xi^2\right)^{-1}\gamma_0\gamma_0'.$$

This asymptotic variance is at least as large as (2.15) (which is also the variance of GD in this case) since the difference is

$$\sigma_\varepsilon^2\left(1+\frac{(1-\lambda)\sigma_\xi^2\alpha_0^2}{\lambda\sigma_\varepsilon^2}\right)-\sigma_\varepsilon^2\left(1+\frac{(1-\lambda)\sigma_\xi^2\alpha_0^2}{\lambda\left(\sigma_\varepsilon^2+\sigma_\xi^2\alpha_0^2\right)}\right)=\frac{\sigma_\varepsilon^2(1-\lambda)\sigma_\xi^2\alpha_0^2}{\lambda}\frac{\sigma_\xi^2\alpha_0^2}{\sigma_\varepsilon^2\left(\sigma_\varepsilon^2+\sigma_\xi^2\alpha_0^2\right)}\geq 0.$$

To see that the asymptotic variance for $\hat{\beta}_{OD}$ is not even guaranteed to be lower than that of

$\hat{\beta}_C$, note that[6]

$$\begin{aligned} AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) - AVAR(\sqrt{n}(\hat{\beta}_{OD} - \beta_0)) &= \left(\sigma_\varepsilon^2 \frac{1}{\lambda} - \sigma_\varepsilon^2 \left(1 + \frac{(1-\lambda)\sigma_\xi^2 \alpha_0^2}{\lambda \sigma_\varepsilon^2}\right)\right) \Gamma^{-1} \\ &= \sigma_\varepsilon^2 \left(\frac{(1-\lambda)}{\lambda \sigma_\varepsilon^2}\left(\sigma_\varepsilon^2 - \sigma_\xi^2 \alpha_0^2\right)\right) \Gamma^{-1} \gtrless 0. \end{aligned}$$

Therefore, despite its convenience, the OD estimator is an inferior alternative. It is less efficient than GMM and not even guaranteed to improve upon the complete data estimator. The simulations in Section 5 provide numerical evidence on these points.

3.2 Dummy Variable Method

We can define the "dummy variable method" using (3.16) and separating the intercept from the other variables in Z_i so that $Z_i = (1, Z_{2i}')$ and also $\gamma_0 = (\gamma_{10}, \gamma_{20}')$ so that

$$Y_i = (1 - m_i) X_i \alpha_0 + Z_i' \beta_0 + m_i \gamma_{10} \alpha_0 + m_i Z_{2i}' \gamma_{20} \alpha_0 + \varepsilon_i + m_i \xi_i \alpha_0. \tag{3.22}$$

Given Assumption 1, equation (3.22) is a valid regression model in the sense that the residual $\varepsilon_i + m_i \xi_i \alpha_0$ is orthogonal to the regressors. The dummy variable method amounts to running the regression without the regressors $m_i Z_{2i}'$. Let $\hat{\theta}_{DM} \equiv (\hat{\alpha}_{DM}, \hat{\beta}_{DM}')'$ denote the dummy variable estimator based on running the regression in (3.22) omitting the regressors $m_i Z_{2i}'$.[7] The following proposition (see also Jones (1996)) formally states the result that the dummy variable method will be subject to omitted-variables bias (and inconsistency) unless certain restrictions are satisfied:

Proposition 6 *The estimators* $(\hat{\alpha}_{DM}, \hat{\beta}_{DM}')'$ *are biased and inconsistent unless (i)* $\alpha_0 = 0$ *or (ii)* $\gamma_{20} = 0$.

The first condition is that X_i is an irrelevant variable in the regression of interest (2.1), in which case the best solution to the missing-data problem is to drop X_i completely and use all

[6] As pointed out by Griliches (1986), the claim by Gourieroux and Monfort (1981) that the unweighted estimator for β_0 is at least as efficient as the complete data estimator is in fact an error. The error is the result of a slight mistake in the algebra — that error that has been corrected by hand in the version that can be found in JSTOR.

[7] The regression also yields an estimate of $\gamma_{10}\alpha_0$.

available data to regress Y_i on Z_i. The second condition requires that either Z_{2i} is non-existent in the model in the first place so that the original model (2.1) is a simple linear regression model or else Z_{2i} is not useful for predicting X_i. If Z_{2i} is non-existent, the dummy-variable estimator is equivalent to the complete-data estimator.

Even if the conditions of Proposition 6 are met and the dummy method is consistent, it is still in general difficult to compare the variance with that of the complete data estimator. Although the complete data estimator results from estimating the unrestricted version of (3.22) by OLS, dropping the irrelevant regressor $m_i Z_{2i}'$ does not necessarily result in efficiency gains when there is conditional heteroskedasticity. To facilitate comparisons (under either (i) or (ii) of Proposition 6) consider the restrictive situation where there is homoskedasticity and homogeneity (Assumption 4) and also the normalization

$$\Gamma = \begin{pmatrix} 1 & 0 \\ 0 & \Gamma_{22} \end{pmatrix}. \tag{3.23}$$

Since the first element of Z_i is 1, (3.23) amounts to assuming that the regressors in Z_{2i} are mean zero and will not alter any of the slope coefficients; the intercept in the model (i.e., the first element of β_0) will be altered by this normalization. The asymptotic variance of the dummy variable estimator, along with efficiency comparisons to the complete data estimator, are summarized in the following proposition:

Proposition 7 *Under Assumptions 1 and 4 and the normalization in (3.23),*

(i) when $\alpha_0 = 0$ *we have,*

$$\begin{aligned} AVAR(\sqrt{n}(\hat{\alpha}_{DM} - \alpha_0)) &= \frac{\sigma_\varepsilon^2}{\lambda\left(\sigma_\xi^2 + (1-\lambda)\gamma_{20}'\Gamma_{22}\gamma_{20}\right)} \leq AVAR(\sqrt{n}(\hat{\alpha}_C - \alpha_0)) \\ AVAR(\sqrt{n}(\hat{\beta}_{DM} - \beta_0)) &= \sigma_\varepsilon^2 \begin{pmatrix} \frac{1}{\lambda} & 0 \\ 0 & \Gamma_{22}^{-1} \end{pmatrix} \\ &\quad + \sigma_\varepsilon^2 \frac{\lambda}{\left(\sigma_\xi^2 + (1-\lambda)\,\gamma_{20}'\Gamma_{22}\gamma_{20}\right)} \begin{pmatrix} \lambda^{-2}\gamma_{10}^2 & \lambda^{-1}\gamma_{10}\gamma_{20}' \\ \lambda^{-1}\gamma_{10}\gamma_{20} & \gamma_{20}\gamma_{20}' \end{pmatrix} \\ &\leq AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) \end{aligned}$$

(ii) when $\gamma_{20} = 0$ we have,

$$\begin{aligned} AVAR(\sqrt{n}(\hat{\alpha}_{DM} - \alpha_0)) &= \frac{\sigma_\varepsilon^2}{\lambda\sigma_\xi^2} = AVAR(\sqrt{n}(\hat{\alpha}_C - \alpha_0)) \\ AVAR(\sqrt{n}(\hat{\beta}_{DM} - \beta_0)) &= \begin{pmatrix} \sigma_\varepsilon^2 \frac{1}{\lambda} & 0 \\ 0 & \left(\sigma_\varepsilon^2 + (1-\lambda)\sigma_\xi^2\alpha_0^2\right)\Gamma_{22}^{-1} \end{pmatrix} + \sigma_\varepsilon^2 \frac{1}{\lambda\sigma_\xi^2} \begin{pmatrix} \gamma_{10}^2 & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

Result (i) says that the estimator for α_0 will be more efficient than the complete data estimator when $\alpha_0 = 0$ and $\gamma_{20} \neq 0$. (When $\gamma_{20} = 0$ as well, there is no efficiency gain possible for estimating the coefficient of the missing variable.) One can compare the variance of the estimator of α_0 that would be possible if the X_i were fully observed — the variance under these conditions would be given by (using FO subscript to denote the estimator with fully observed data)

$$AVAR(\sqrt{n}(\hat{\alpha}_{FO} - \alpha_0)) = \frac{\sigma_\varepsilon^2}{\sigma_\xi^2}$$

so then the relative efficiency would be

$$\begin{aligned} \frac{AVAR(\sqrt{n}(\hat{\alpha}_{DM} - \alpha_0))}{AVAR(\sqrt{n}(\hat{\alpha}_{FO} - \alpha_0))} &= \left(\frac{\sigma_\varepsilon^2}{\lambda\left(\sigma_\xi^2 + (1-\lambda)\gamma_{20}'\Gamma_{22}\gamma_{20}\right)}\right)\left(\frac{\sigma_\varepsilon^2}{\sigma_\xi^2}\right)^{-1} \\ &= \frac{\sigma_\xi^2}{\lambda\left(\sigma_\xi^2 + (1-\lambda)\gamma_{20}'\Gamma_{22}\gamma_{20}\right)} \\ &= \left(\lambda + (1-\lambda)\lambda\frac{\gamma_{20}'\Gamma_{22}\gamma_{20}}{\sigma_\xi^2}\right)^{-1} \end{aligned}$$

which depends on λ as well as the signal to noise ratio in the relationship between X_i and Z_i. It is possible that the dummy variable method could be more efficient than the full data method when $\gamma_{20}'\Gamma_{22}\gamma_{20}$ is large relative to σ_ξ^2. Intuitively when this occurs there is a strong relationship between X_i and Z_i and it is hard to estimate α_0. When $\alpha_0 = 0$, it does not matter whether X_i is observed or not and apparently there are some gains from entering it as a zero and using a missing indicator in its place. This result should not be pushed too far, however, since it only occurs when $\alpha_0 = 0$ — in this instance, it would be better to drop X_i completely and just use Z_i in the regression. Also, one suspects that in most applications in economics the noise σ_ξ^2 is likely to be large relative to the signal $\gamma_{20}'\Gamma_{22}\gamma_{20}$ so that the ratio of variances is likely to be larger than one in practice. The second part of (i) suggests that the estimator of the slopes in

β_0 will be more efficient than GMM since[8]

$$\frac{\lambda}{(\sigma_\xi^2 + (1-\lambda)\,\gamma_{20}'\Gamma_{22}\gamma_{20})}\gamma_{20}\gamma_{20}' \leq \frac{1}{\lambda\sigma_\xi^2}\gamma_{20}\gamma_{20}'.$$

When $\gamma_{20} = 0$, result (ii) implies that the dummy method is no more efficient than the complete data estimator for α_0 and could be more or less efficient than the complete data estimator of the β_0 slopes since

$$\frac{\sigma_\varepsilon^2}{\lambda} - \left(\sigma_\varepsilon^2 + (1-\lambda)\sigma_\xi^2\beta_1^2\right) = \frac{(1-\lambda)}{\lambda}\left(\sigma_\varepsilon^2 - \sigma_\xi^2\beta_1^2\right) \lesseqgtr 0.$$

This comparison is equivalent to the efficiency comparison between the unweighted imputation estimator and the complete data estimator. In this instance, in fact, the dummy method has the same asymtptotic variance as the unweighted imputation estimator and, as such, is less efficient than the GMM or weighted imputation estimator.

The results of this section suggest little to recommend the dummy variable method for dealing with missingness. It raises the possibility of bias and inconsistency. As a practical matter, one may be willing to live with this bias if the method had a lower variance but even this is not guaranteed. The only situation where one does not sacrifice bias in exchange for variance improvements is precisely the case where the missing variable can be eliminated completely.

4 Missing Data in Instrumental Variable Models

The GMM framework to handle missingness can easily be modified to handle other models for which GMM estimators are commonly used. In this section, we consider extending the methodology to the case of instrumental-variables models. Section 4.1 considers the case where the instrumental variable may be missing, whereas Section 4.2 considers the case where the

[8] Also for the intercept (under the reparameterization) there will be an efficiency gain relative to GMM since

$$\frac{\lambda\lambda^{-2}\gamma_{10}^2}{(\sigma_\xi^2 + (1-\lambda)\,\gamma_{20}'\Gamma_{22}\gamma_{20})} = \frac{\gamma_{10}^2}{\lambda(\sigma_\xi^2 + (1-\lambda)\,\gamma_{20}'\Gamma_{22}\gamma_{20})} \leq \frac{\gamma_{10}^2}{\lambda\sigma_\xi^2}.$$

endogenous variable may be missing. As the latter case turns out to be very similar to the situation considered in Section 2, the discussion in Section 4.2 will be somewhat brief.

4.1 Missing Instrument Values

This section considers a situation in which an instrumental variable has potentially missing values. An example of this occurs in Card (1995), where IQ score is used as an instrument for the "Knowledge of the World of Work" (KWW) test score in a wage regression; IQ score is missing for about 30% of the sample, and Card (1995) simply omits the observations with missing data in the IV estimation. Other authors (see, for example, Dahl and DellaVigna (2009)) have used a dummy variable approach to deal with missing values for an instrument — instead of dropping observations with missing values, one enters a zero for the missing value and "compensates" by using dummies for "missingness."

Consider a simple situation in which there is a single instrument for a single endogenous regressor and where the instrument may be missing. The model consists of the following "structural" equation

$$Y_{1i} = Y_{2i}\delta_0 + Z_i'\beta_0 + \varepsilon_i, \quad E(Z_i\varepsilon_i) = 0, \; E(Y_{2i}\varepsilon_i) \neq 0, \tag{4.24}$$

where all the variables are observed, and a reduced form (linear projection) for the endogenous regressor Y_{2i},

$$Y_{2i} = X_i\pi_0 + Z_i'\Pi_0 + v_i, \; E(X_iv_i) = 0, \; E(Z_iv_i) = 0. \tag{4.25}$$

Missingness of the instrumental variable X_i is denoted by the indicator variable m_i (equal to one if X_i missing). The missing-at-random assumptions required in this context are[9]

$$E(m_iX_i\varepsilon_i) = E(m_iX_iv_i) = E(m_iX_i\xi_i), \tag{4.26}$$

where ξ_i is the projection error from the projection of X_i onto Z_i (as in (2.3)). Also, X_i is assumed to be a valid and useful instrument in the sense that

$$E(X_i\varepsilon_i) = 0 \text{ and } \pi_0 \neq 0.$$

[9]See Mogstad and Wiswall (2010) for a treatment of missing instrumental variables under alternative assumptions.

For this model, one is primarily interested in the estimation of the parameters of (4.24) so that efficient estimation of the parameters of (4.25) is not of paramount concern. As in Section 2, we assume that the linear projection in (2.3) exists. The complete data method in this context amounts to using only the observations for which $m_i = 0$ (X_i not missing) — this is the approach in Card (1995). Given the missing-at-random assumption (4.26), this is a consistent approach that asymptotically uses a proportion of data represented by $\lambda = P(m_i = 0)$.

Similar to the arguments in Section 2, one can use (2.3) to write a reduced form linear projection that is satisfied for the entire sample as

$$Y_{2i} = (1 - m_i)X_i\pi_0 + Z_i'\Pi_0 + m_i Z_i'\gamma_0\pi_0 + v_i + \pi_0 m_i\xi_i. \tag{4.27}$$

Then, partitioning $Z_i = (1, Z_{2i}')'$, the full-sample reduced form in (4.27) suggests that one use an instrument set that consists of $((1 - m_i)X_i, Z_i, m_i, m_i Z_{2i})$ when estimating (4.24) based on the entire sample. On the other hand, the dummy variable approach amounts to using the subset $((1 - m_i)X_i, Z_i, m_i)$. The interesting question in this case is whether there are benefits from using the entire sample with either set of instruments relative to just omitting observations with missing values for the instrument. Intuitively, based on standard results, one cannot imagine that the "dummy approach" could be better than the approach based on the full set of instruments $((1-m_i)X_i, Z_i, m_i, m_i Z_{2i})$. To address this question, we compare the properties of the IV estimators. Clearly each estimator is consistent so it comes down to relative variances. We use similar notation to previous sections so that $(\hat{\delta}_C, \hat{\beta}_C)$ denotes the IV estimator using complete data where X_i is used to instrument Y_{2i}. Similarly $(\hat{\delta}_D, \hat{\beta}_D)$ is the 2SLS estimator that uses the instrument set $((1 - m_i)X_i, Z_i, m_i)$ and the entire sample. The 2SLS estimator using the full instrument set (based on (4.27)) $((1-m_i)X_i, Z_i, m_i, m_i Z_{2i}')$ and the entire sample is denoted by $(\hat{\delta}_F, \hat{\beta}_F)$.

It seems intuitively clear that $(\hat{\delta}_F, \hat{\beta}_F)$ will be at least as efficient as $(\hat{\delta}_D, \hat{\beta}_D)$ — what is less clear is how these estimators perform relative to the complete data IV estimator. The following result shows that with respect to estimation of δ_0 there is no advantage from using the 2SLS methods and the entire sample and in the case of $\hat{\delta}_D$ one may actually be worse off (relative to the complete data estimator) in terms of asymptotic variance.

Proposition 8 *If ε_i, v_i, and ξ_i are conditionally homoskedastic and $E(Z_i Z_i') = I$, then*

$$AVAR(\sqrt{n}(\hat{\delta}_C - \delta_0)) = AVAR(\sqrt{n}(\hat{\delta}_F - \delta_0)) = \sigma_\varepsilon^2 \left(\pi_0^2 \lambda \sigma_\xi^2\right)^{-1}$$

$$AVAR(\sqrt{n}(\hat{\delta}_D - \delta_0)) = \sigma_\varepsilon^2 \left(\pi_0^2 \lambda \sigma_\xi^2\right)^{-1} \left(1 + \frac{(1-\lambda)\gamma_{20}'\gamma_{20}}{\sigma_\xi^2}\right).$$

The full instrument estimator and the complete data estimator have the same asymptotic variance while the dummy method is less efficient by an amount that depends on the amount of missing data as well as the coefficient γ_{20} — when this is zero, the dummy variable method has the same asymptotic variance as the other two estimators. This result suggests that the method of dealing with missingness in instruments by using zeros for the missing value and m_i alone to compensate is likely to be inferior compared to the method that drops observations with missing values for the instrument. To reach the same level of efficiency, one must add to the instrument set the interactions of m_i with all the elements of Z_i. The following proposition shows that the latter described method does bring some improvements with respect to the estimation of β_0:

Proposition 9 *If ε_i, v_i, and ξ_i are conditionally homoskedastic and $E(Z_i Z_i') = I$, then*

$$\begin{aligned}
AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) &= \sigma_\varepsilon^2 \left(\frac{1}{\lambda} I + \left(\pi_0^2 \lambda \sigma_\xi^2\right)^{-1} (\gamma_0 \pi_0 + \Pi_0)(\gamma_0 \pi_0 + \Pi_0)'\right) \\
AVAR(\sqrt{n}(\hat{\beta}_F - \beta_0)) &= \sigma_\varepsilon^2 \left(I + \left(\pi_0^2 \lambda \sigma_\xi^2\right)^{-1} (\gamma_0 \pi_0 + \Pi_0)(\gamma_0 \pi_0 + \Pi_0)'\right) \\
AVAR(\sqrt{n}(\hat{\beta}_D - \beta_0)) &= \sigma_\varepsilon^2 \left(I + \left(\pi_0^2 \lambda \sigma_\xi^2\right)^{-1} \left(1 + \frac{(1-\lambda)\gamma_{20}'\gamma_{20}}{\sigma_\xi^2}\right)(\gamma_0 \pi_0 + \Pi_0)(\gamma_0 \pi_0 + \Pi_0)'\right).
\end{aligned}$$

Comparing these asymptotic variances, the full instrument 2SLS estimator has the lowest asymptotic variance — it is unequivocally more efficient than the complete IV estimator when there is a non-negligible portion of missing data. The full instrument estimator is more efficient than the dummy estimator except when $\gamma_{20} = 0$, in which case the additional instruments in the full set are useless for Y_{2i}. The comparison between the dummy method and the complete IV estimator depends on the various parameters of the model — large γ_{20} tends to make the complete estimator more efficient, while small λ tends to make the dummy more efficient.

These results have a simple implication. For missing instrument values (where missingness satisfies our assumptions), the method that is guaranteed to deliver asymptotic efficiency is the 2SLS estimator with a full set of instruments obtained from interactions of m_i and Z_i. Compensating for missingness by simply using the dummy alone is not a good idea unless one believes the instrument is uncorrelated with the other exogenous variables in the model. In general, while using the dummy alone may bring a benefit for some coefficients, it may also come at a cost for the other coefficients.

4.2 Missing Endogenous-Variable Values

We now consider the case where the endogenous regressor Y_{2i} may be missing, and let m_i denote the indicator variable for the missingness of Y_{2i}. Otherwise, we consider the same structural and reduced-form models as in (4.24) and (4.25), respectively:[10]

$$Y_{1i} = Y_{2i}\delta_0 + Z_i'\beta_0 + \varepsilon_i, \quad E(Z_i\varepsilon_i) = 0, \; E(Y_{2i}\varepsilon_i) \neq 0$$

$$Y_{2i} = X_i\pi_0 + Z_i'\Pi_0 + v_i, \; E(X_iv_i) = 0, \; E(Z_iv_i) = 0.$$

In comparing these two equations to their counterparts in the missing-exogenous-variable model of Section 2 (see (2.1) and (2.3), respectively), there are two key differences: (i) the RHS variable Y_{2i} is not orthogonal to the first-equation error, and (ii) an additional exogenous variable (X_i) is orthogonal to both the first- and second-equation errors. It is straightforward to incorporate both of these differences into the GMM framework.

Let $W_i = (X_i, Z_i')'$ denote the full vector of exogenous variables in the model. Then, the appropriate vector of moment functions (analogous to (2.5)) is given by

$$h_i(\delta,\beta,\pi,\Pi) = \begin{pmatrix} (1-m_i)W_i(Y_{1i} - Y_{2i}\delta - Z_i'\beta) \\ m_iW_i\left(Y_{1i} - X_i\pi\delta - Z_i'\left(\Pi\delta + \beta\right)\right) \\ (1-m_i)W_i(Y_{2i} - X_i\pi - Z_i'\Pi) \end{pmatrix} = \begin{pmatrix} h_{1i}(\delta,\beta,\pi,\Pi) \\ h_{2i}(\delta,\beta,\pi,\Pi) \\ h_{3i}(\delta,\beta,\pi,\Pi) \end{pmatrix}. \tag{4.28}$$

Note that consistency of the GMM estimator requires the following missing-at-random assumption:[11]

$$E(m_iW_i\varepsilon_i) = E(m_iW_iv_i) = 0.$$

[10] Although this formulation restricts the instrumental variable X_i to be scalar, it is trivial to extend the GMM estimator below to the case of vector X_i.

[11] Consistency of the complete-data estimator requires $E(m_iW_i\varepsilon_i) = 0$.

There are a total of $3K+3$ moments in (4.28) and $2K+2$ parameters in $(\delta_0, \beta_0, \pi_0, \Pi_0)$, so that the optimal GMM estimator would yield a test of overidentifying restrictions, analogous to Proposition 1, with $\chi^2(K+1)$ limiting distribution.

5 Monte Carlo Experiments

In this section, we conduct several simulations to examine the small-sample performance of the various methods considered in Sections 2 and 3 under different data-generating processes. We consider a very simple setup with $K = 2$:

$$\begin{aligned} Y_i &= X_i\alpha_0 + \beta_1 + \beta_2 Z_{2i} + \sigma_\varepsilon(X_i, Z_i)u_i \\ X_i &= \gamma_1 + \gamma_2 Z_{2i} + \sigma_\xi(Z_i)v_i. \\ \sigma_\varepsilon(X_i, Z_i) &= \sqrt{\theta_0 + \theta_1 X_i^2 + \theta_2 Z_i^2} \\ \sigma_\xi(Z_i) &= \sqrt{\delta_0 + \delta_1 Z_i^2} \\ Z_i &\sim N(0,1) \end{aligned}$$

We fix $(\beta_1, \beta_2, \gamma_1, \gamma_2) = (1,1,1,1)$ throughout the experiments. In all but one of the designs α_0 is set to 1; in one design, it is set to 0.1. For each of the designs, we consider a simple missingness mechanism in which exactly half of the X_i's are missing completely at random ($\lambda = 1/2$). We consider a total of eight different designs, with the first five based upon

$$u_i, v_i \sim N(0,1),\ v_i \perp u_i$$

and the following parameter values:

Design 1: $\alpha_0 = 1$, $\theta_0 = \delta_0 = 10$, $\theta_1 = \theta_2 = \delta_1 = 0$

Design 2: $\alpha_0 = 0.1$, $\theta_0 = \delta_0 = 10$, $\theta_1 = \theta_2 = \delta_1 = 0$

Design 3: $\alpha_0 = 1$, $\theta_0 = 1$, $\delta_0 = 10$, $\theta_1 = \theta_2 = \delta_1 = 0$

Design 4: $\alpha_0 = 1$, $\theta_0 = \delta_0 = \theta_2 = \delta_1 = 1$, $\theta_1 = 0$

Design 5: $\alpha_0 = 1$, $\theta_0 = \delta_0 = \theta_1 = \theta_2 = \delta_1 = 1$

Designs 1–3 have homoskedastic residuals and are used to illustrate the effect of different values for the variances and the importance of the missing variable itself. Designs 4 and 5 introduce heteroskedasticity, with conditional variances dependent on Z in Design 4 and the main-equation variance also dependent on X in Design 5. These latter two designs are meant to examine the potential efficiency gains for α_0 indicated by the result in Proposition 2 and the special case following that proposition.

The remaining three designs are based on

$$\begin{aligned} u_i &\sim N(0,1) \\ v_i &= u_i^2 - 1. \end{aligned}$$

Note that u_i and v_i are independent (and mean zero) conditional on X_i and Z_i and also that $E(u_i v_i) = 0$ but $E(u_i^2 v_i) = 2$. This setup is meant to investigate the relevance of the third-moment condition for efficiency gains for α_0 discussed after Proposition 2. The parameters for Designs 6 and 7 are as follows:

Design 6: $\alpha_0 = 1$, $\theta_0 = \delta_0 = 1$, $\theta_1 = \theta_2 = \delta_1 = 0$

Design 7: $\alpha_0 = 1$, $\theta_0 = \delta_0 = \theta_2 = \delta_1 = 1$, $\theta_1 = 0$

For Design 8, we consider $\alpha_0 = 1$ and the following exponential forms for the residual standard deviations:

$$\sigma_\epsilon(X_i, Z_i) = \exp\left(\sqrt{0.1 + 0.2Z_i^2 + 0.1X_i^2}\right)$$

$$\sigma_\xi(X_i, Z_i) = \exp\left(\sqrt{0.1 + 0.2Z_i^2}\right)$$

This last design illustrates more dramatic efficiency gains that are possible with the GMM estimator.

For all simulations, a sample size of $n = 400$ is used. The results are reported in Tables 2–9. For a set of 1000 replications for each design, these tables report the bias, variance, and overall MSE for the estimators of the parameters $(\alpha_0, \beta_1, \beta_2)$. The estimators considered are those discussed in Sections 2 and 3, namely (i) the complete case estimator, (ii) the dummy variable

estimator, (iii) the unweighted imputation (OD) estimator, (iv) the weighted imputation (GD) estimator, and (v) the optimal GMM estimator.

Several things stand out in the results. With the exception of the dummy variable method, none of the methods have much bias for any of the parameters. The dummy variable method can be very biased. This is most pronounced for β_2 in all cases except for the Design 2 where X is relatively unimportant. There is also substantial bias for both α_0 and β_1, especially in Designs 4–8. The variance of the dummy variable estimator can be larger or smaller than the complete data estimator, but the overall MSE for this estimator is never smaller than the complete data estimator with the exception of the case where α_0 is small. There is little evidence to suggest the dummy method has any advantage except in cases where X can be dropped from the analysis.

Regarding the other estimators, the results for Designs 1–3 support the theoretical results: the weighted imputation and GMM estimators are roughly equally efficient and do substantially better for estimating β_1 and β_2 except in Design 2 where X is relatively unimportant in the main regression model. The unweighted imputation estimator is relatively inefficient in Designs 1 and 3 for estimating β_1 and β_2. In Design 2, where α_0 is small, the unweighted estimator does as well as the weighted estimator and GMM but this is the only instance where this occurs. For estimating α_0, the unweighted and weighted imputation estimators appear identical.

The results in Design 4–8 show that the GMM estimator is generally the estimation method with the lowest variance and overall MSE. There are also gains in terms of estimating α_0 in Designs 5–8. The result in Design 4 is consistent with the discussion following Proposition 2, which suggested that when the conditional variance only depended on Z that there would not be any improvement for estimating α_0.

The GMM estimator itself seems to be very well behaved from a numerical standpoint. Using standard Gauss-type iterations, the estimates were found with a very tight convergence criterion in no more than around ten iterations in any case. The experiments all ran very quickly (i.e., a few seconds on a standard desktop computer in *GAUSS8.0*) despite the fact that

Table 2: Monte Carlo simulations, Design 1

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α_0	0.001	1.84	0.005
	β_1	-0.001	22.76	0.057
	β_2	-0.001	22.26	0.056
Dummy-variable method	α_0	-0.047	1.88	0.007
	β_1	0.049	23.44	0.061
	β_2	0.535	15.88	0.326
Unweighted imputation	α_0	0.001	1.84	0.005
	β_1	0.002	22.63	0.057
	β_2	0.006	22.25	0.056
Weighted imputation	α_0	0.001	1.84	0.005
	β_1	0.000	17.65	0.044
	β_2	0.003	16.93	0.042
GMM (efficient)	α_0	0.004	1.87	0.005
	β_1	-0.003	17.79	0.044
	β_2	0.000	17.10	0.043

Design 1: $\alpha_0 = 1$, $\theta_0 = \delta_0 = 10$, $\theta_1 = \theta_2 = \delta_1 = 0$

a GMM estimate with nonlinearity had to be computed via numerical methods 1000 times for each experiment.

Overall, the simulation results suggest the following. First, if one believes the data are homoskedastic, then the two-step linear imputation method is preferred as long as one uses the weighting suggested by Dagenais (1973). Second, the dummy variable method, though convenient, has little else to recommend it. It can be substantially biased except in cases where one could actually just toss out X completely and do a regression on Z — moreover, it also does not necessarily bring about variance improvements which is the whole purpose of imputation in the first place. Third, the GMM estimators seem to be numerically stable, bring about variance gains in a variety of cases including homoskedastic and heteroskedastic cases, and, as an added bonus, give rise to the possibility of testing the restrictions on the models that bring about the possibility of efficiency gains.

Table 3: Monte Carlo simulations, Design 2

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α_0	0.001	1.84	0.005
	β_1	-0.001	22.76	0.057
	β_2	-0.001	22.26	0.056
Dummy-variable method	α_0	-0.004	1.77	0.004
	β_1	0.005	22.66	0.057
	β_2	0.060	10.51	0.030
Unweighted imputation	α_0	0.001	1.84	0.005
	β_1	-0.001	13.56	0.034
	β_2	0.006	11.97	0.030
Weighted imputation	α_0	0.001	1.84	0.005
	β_1	-0.001	13.53	0.034
	β_2	0.006	11.99	0.030
GMM (efficient)	α_0	0.002	1.89	0.005
	β_1	-0.002	13.60	0.034
	β_2	0.006	12.22	0.031

Design 2: $\alpha_0 = 0.1$, $\theta_0 = \delta_0 = 10$, $\theta_1 = \theta_2 = \delta_1 = 0$

Table 4: Monte Carlo simulations, Design 3

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α_0	0.000	0.18	0.000
	β_1	0.000	2.28	0.006
	β_2	0.000	2.23	0.006
Dummy-variable method	α_0	-0.048	0.28	0.003
	β_1	0.049	2.97	0.010
	β_2	0.530	6.95	0.298
Unweighted imputation	α_0	0.000	0.18	0.000
	β_1	0.003	10.68	0.027
	β_2	0.001	11.92	0.030
Weighted imputation	α_0	0.000	0.18	0.000
	β_1	0.000	2.17	0.005
	β_2	0.000	2.12	0.005
GMM (efficient)	α_0	0.001	0.19	0.000
	β_1	-0.001	2.18	0.005
	β_2	-0.001	2.12	0.005

Design 3: $\alpha_0 = 1$, $\theta_0 = 1$, $\delta_0 = 10$, $\theta_1 = \theta_2 = \delta_1 = 0$

Table 5: Monte Carlo simulations, Design 4

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α_0	0.004	2.83	0.007
	β_1	-0.004	7.18	0.018
	β_2	-0.002	10.85	0.027
Dummy-variable method	α_0	-0.200	3.09	0.048
	β_1	0.201	8.11	0.061
	β_2	0.608	7.67	0.389
Unweighted imputation	α_0	0.004	2.83	0.007
	β_1	-0.004	7.11	0.018
	β_2	0.000	10.87	0.027
Weighted imputation	α_0	0.004	2.83	0.007
	β_1	-0.004	6.12	0.015
	β_2	-0.001	8.83	0.022
GMM (efficient)	α_0	0.005	2.83	0.007
	β_1	-0.005	6.10	0.015
	β_2	-0.002	8.89	0.022

Design 4: $\alpha_0 = 1$, $\theta_0 = \delta_0 = \theta_2 = \delta_1 = 1$, $\theta_1 = 0$

Table 6: Monte Carlo simulations, Design 5

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α_0	0.011	13.93	0.035
	β_1	-0.012	19.34	0.049
	β_2	-0.002	22.65	0.057
Dummy-variable method	α_0	-0.194	13.94	0.073
	β_1	0.195	19.89	0.088
	β_2	0.612	14.63	0.411
Unweighted imputation	α_0	0.011	13.93	0.035
	β_1	-0.010	19.33	0.048
	β_2	-0.001	22.46	0.056
Weighted imputation	α_0	0.011	13.93	0.035
	β_1	-0.010	17.78	0.045
	β_2	0.002	19.46	0.049
GMM (efficient)	α_0	0.006	12.53	0.031
	β_1	-0.007	16.27	0.041
	β_2	-0.002	18.35	0.046

Design 5: $\alpha_0 = 1$, $\theta_0 = \delta_0 = \theta_1 = \theta_2 = \delta_1 = 1$

Table 7: Monte Carlo simulations, Design 6

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α_0	0.003	5.09	0.013
	β_1	-0.005	2.71	0.007
	β_2	-0.001	7.24	0.018
Dummy-variable method	α_0	-0.204	4.30	0.052
	β_1	0.200	2.88	0.047
	β_2	0.601	3.94	0.371
Unweighted imputation	α_0	0.003	5.09	0.013
	β_1	-0.002	5.91	0.015
	β_2	-0.001	8.46	0.021
Weighted imputation	α_0	0.003	5.09	0.013
	β_1	-0.001	3.09	0.008
	β_2	-0.001	6.96	0.017
GMM (efficient)	α_0	0.004	4.88	0.012
	β_1	-0.007	2.72	0.007
	β_2	-0.003	6.74	0.017

Design 6: $\alpha_0 = 1$, $\theta_0 = \delta_0 = 1$, $\theta_1 = \theta_2 = \delta_1 = 0$

Table 8: Monte Carlo simulations, Design 7

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α_0	0.002	6.54	0.016
	β_1	-0.005	3.99	0.010
	β_2	-0.001	14.50	0.036
Dummy-variable method	α_0	-0.114	5.97	0.028
	β_1	0.110	4.46	0.023
	β_2	0.556	10.96	0.337
Unweighted imputation	α_0	0.002	6.54	0.016
	β_1	-0.002	9.65	0.024
	β_2	0.002	18.74	0.047
Weighted imputation	α_0	0.002	6.54	0.016
	β_1	0.000	4.42	0.011
	β_2	0.000	13.28	0.033
GMM (efficient)	α_0	0.001	6.13	0.015
	β_1	-0.005	3.72	0.009
	β_2	0.001	12.83	0.032

Design 7: $\alpha_0 = 1$, $\theta_0 = \delta_0 = \theta_2 = \delta_1 = 1$, $\theta_1 = 0$

Table 9: Monte Carlo simulations, Design 8

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α_0	0.029	148.90	0.373
	β_1	-0.034	112.47	0.282
	β_2	-0.001	41.23	0.103
Dummy-variable method	α_0	-0.112	156.16	0.403
	β_1	0.110	116.13	0.302
	β_2	0.574	54.31	0.465
Unweighted imputation	α_0	0.029	148.90	0.373
	β_1	-0.035	131.35	0.330
	β_2	-0.018	96.21	0.241
Weighted imputation	α_0	0.029	148.90	0.373
	β_1	-0.033	124.48	0.312
	β_2	-0.006	69.65	0.174
GMM (efficient)	α_0	0.002	18.45	0.046
	β_1	-0.006	22.74	0.057
	β_2	0.003	24.19	0.060

Design 8: $\alpha_0 = 1$, exponential variance form (see text)

6 Empirical examples

This section considers application of the GMM and other missing-covariate estimation methods to datasets with a large amount of missing data on variables of interest. Section 6.1 considers the estimation of regression models using data from the Wisconsin Longitudinal Study, where a covariate of interest is observed for only about a quarter of sampled individuals. Section 6.2 considers the Card (1995) data (from the National Longitudinal Survey of Young Men) mentioned in Section 4.1; for this data, we estimate instrumental-variables regressions where one of the instrumental variables (IQ score) is missing for about one-third of the observations.

6.1 Regression example — Wisconsin Longitudinal Study

The Wisconsin Longitudinal Study (WLS) has followed a random sample of individuals who graduated from Wisconsin high schools in 1957. In addition to the original survey, several follow-up surveys have been used to gather longitudinal information for the sample. For this example, we focus on a specific data item that is available for only a small fraction of the overall sample.

Specifically, we look at BMI (body mass index) ratings based upon high-school yearbook photos of the individuals. For several reasons, this high-school BMI rating variable is observed for only about a quarter of individuals.[12] The variable should not be considered missing completely at random (MCAR) since observability depends on whether a school's yearbook is available or not and, therefore, could be related to variables correlated with school identity. The high-school BMI rating is based on the independent assessment of six individuals, and the variable that we use should be viewed as a proxy for BMI (or, more accurately, perceived BMI) in high school.[13]

We consider two regression examples where the high-school BMI rating variable is used as an explanatory variable for future outcomes. The dependent variables that we consider are (i) completed years of schooling (as of 1964) and (ii) adult BMI (as reported in 1992-1993). For the schooling regression, IQ score (also recorded in high school) is used as an additional covariate; for the adult BMI regression, IQ score and completed years of schooling are used as additional covariates. The results are reported in Table 10, where estimation is done separately for men and women and three different methods are considered: (i) the complete-case method, (ii) the dummy-variable method, and (iii) the GMM method. The schooling regression results are in the top panel (Panel A), and the adult-BMI regression results are in the bottom panel (Panel B).

There is a lot of missing data associated with the high-school BMI rating variable. In the schooling regression, high-school BMI rating is observed for only 888 of 3,969 men (22.4%) and 1,107 of 4,276 women (25.9%). As a result, we see that the complete-case method results in much higher standard errors for the other covariates (e.g., the standard errors on the IQ variable in Panel A are roughly twice as large as those from either the dummy-variable or GMM methods). While the dummy-variable method is not guaranteed to be consistent, it gives quite similar results on the high-school BMI rating coefficient to the other methods; as the theory of Sections 2 and 3 has suggested, there is little difference in the standard errors for this covariate

[12] According to the WLS website, yearbooks were available and coded for only 72% of graduates; in addition, in the release of the data, ratings had not been completed for the full set of available photos.

[13] Each rater assigned a relative body mass score from 1 (low) to 11 (high). The variable that we use, named `srbmi` in the public-release WLS dataset, is a standardized variable calculated separately for male and female photos. According to the WLS documentation, this variable is calculated by generating rater-specific z scores, summing the z scores for a given photo, and dividing by the number of raters.

Table 10: Regression examples, Wisconsin Longitudinal Study data

Panel A	Dependent variable = years of education					
	Men			Women		
	Complete-case method	Dummy-variable method	GMM method	Complete-case method	Dummy-variable method	GMM method
High-school BMI rating	0.0878	0.0889	0.0826	-0.2748	-0.2727	-0.2650
	(0.0757)	(0.0757)	(0.0751)	(0.0603)	(0.0600)	(0.0602)
IQ	0.0644	0.0673	0.0674	0.0473	0.0486	0.0476
	(0.0041)	(0.0019)	(0.0019)	(0.0034)	(0.0017)	(0.0018)
Missing-BMI indicator		-0.0174			-0.0867	
		(0.0729)			(0.0557)	
Constant	7.3962	7.1067	7.0781	8.6023	8.4702	8.5001
	(0.4023)	(0.1915)	(0.1805)	(0.3287)	(0.1697)	(0.1701)
Test statistic (d.f. 2)			0.866			3.208
(p-value)			(0.649)			(0.201)
Observations	888	3969	3969	1107	4276	4276
Panel B	Dependent variable = adult BMI					
	Men			Women		
	Complete-case method	Dummy-variable method	GMM method	Complete-case method	Dummy-variable method	GMM method
High-school BMI rating	1.5504	1.5345	1.5577	1.9491	1.9213	2.0204
	(0.1693)	(0.1699)	(0.1675)	(0.2213)	(0.2196)	(0.2101)
IQ	0.0221	-0.0066	0.0002	0.0092	0.0007	0.0051
	(0.0100)	(0.0058)	(0.0062)	(0.0130)	(0.0070)	(0.0075)
Years of education	-0.1780	-0.1590	-0.1698	-0.1817	-0.2224	-0.1389
	(0.0662)	(0.0394)	(0.0421)	(0.0973)	(0.0552)	(0.0616)
Missing-BMI indicator		-0.1032			0.1666	
		(0.1583)			(0.1947)	
Constant	27.8288	30.4810	29.8730	27.4363	28.8575	27.4230
	(1.0309)	(0.5839)	(0.6218)	(1.5291)	(0.8467)	(0.9356)
Test statistic (d.f. 3)			10.316			1.776
(p-value)			(0.016)			(0.620)
Observations	698	2587	2587	873	2917	2917

across the three methods. While the coefficient estimates for the methods are quite similar in the education regressions, there are a few differences in the adult BMI regressions. For instance, the estimated effect of education on adult BMI is -0.2224 (s.e. 0.0552) for the dummy-variable method and -0.1389 (s.e. 0.0616) for the GMM method.

The dummy-variable method is, of course, not guaranteed to even be consistent under the missingness assumptions that yield GMM consistency. Moreover, the GMM method allows for an overidentification test of the assumptions being made. The overidentification test statistics are reported in Table 10 and, under the null hypothesis of correct specification, have limiting distributions of χ^2_2 and χ^2_3 for the education and adult BMI regressions, respectively. The test for the adult-BMI regression on the male sample has a p-value of 0.016, casting serious doubt on the missingness assumptions.[14] For this regression, note that the complete-case method estimate for the IQ coefficient was positive (0.0221) and statistically significant at a 5% level (s.e. 0.0100); in contrast, the GMM estimate for the IQ coefficient is very close to zero in magnitude and statistically insignificant. The result of the overidentification test, however, would caution a researcher against inferring too much from this difference. The test indicates that the assumptions needed for consistency of GMM are not satisfied; it is also possible that complete-case method estimator is itself inconsistent here (e.g., if Assumption (i) and/or (iii) are violated).

6.2 Instrumental variable example — Card (1995)

This section considers IV estimation of a log-wage regression using the data of Card (1995). The sample consists of observations on male workers from the National Longitudinal Survey of Young Men (NLSYM) in 1976. The endogenous variable (Y_2) is KWW (an individual's score on the "Knowledge of the World of Work" test), with exogenous variables (Z_2) including years of education, years of experience (and its square), an $SMSA$ indicator variable (1 if living in an SMSA in 1976), a $South$ indicator variable (1 if living in the south in 1976), and a black-race indicator variable. IQ score is used as an instrument (X) for KWW, but IQ data are missing

[14] The probability of seeing one of the four p-values less than 0.016 (under correct specification for all four cases) is roughly 6.2%.

for 923 of the 2,963 observations. The complete-data sample, where IQ and the other variables are non-missing, has 2,040 observations. An additional specification, in which education is also treated as endogenous, is considered; for this specification, an indicator variable for living in a local labor market with a 4-year college is used as an additional instrumental variable (and is always observed, unlike IQ score).

Table 11 reports the IV estimation results. Three estimators are considered: (i) the complete-data IV estimator (2,040 observations), (ii) the dummy-variable IV estimator (2,963 observations, using the missingness indicator as an additional instrument), and (iii) the full IV estimator (2,963 observations, using the missingness indicator and its interactions with Z_2 as additional instruments). The first three columns of Table 11 use IQ as an instrument for KWW, and the second three columns also use the near-4-year-college indicator variable as an instrument for education.

The results clearly illustrate the greater efficiency associated with the full IV approach. In the first specification, the complete-data and dummy-IV estimators have very similar standard errors, whereas the full IV estimator provides efficiency gains (roughly 10-15%) for the coefficient estimates of both the endogenous (KWW) variable and the exogenous variables. The efficiency gains in the second specification are far more dramatic. For the KWW coefficient, the full-instrument standard error is 0.0097 as compared to the complete-data and dummy-IV standard errors of 0.0218 and 0.0146, respectively. For the education coefficient, the full-instrument standard error is 0.0356 as compared to the complete-data and dummy-IV standard errors of 0.0946 and 0.0528, respectively. Thus, the standard errors on the endogenous variable are roughly a third lower for the full-IV estimator as compared to the dummy-IV estimator. For the exogenous variables, the full-IV standard errors are uniformly lower, with the largest efficiency gains evident for the experience variables and the black indicator.

Table 11: Instrumental variable examples, Card (1995) NLSYM data

	Dependent variable = ln(weekly wage)					
	IQ score as instrument for KWW			IQ score as instrument for KWW and near-4-year-college indicator as instrument for years of education		
	Complete Data	Dummy Instrument	Full Instrument	Complete Data	Dummy Instrument	Full Instrument
KWW	0.0191	0.0189	0.0204	0.0034	0.0202	0.0278
	(0.0051)	(0.0059)	(0.0046)	(0.0218)	(0.0146)	(0.0097)
Education	0.0367	0.0313	0.0280	0.1061	0.0274	0.0053
	(0.0116)	(0.0136)	(0.0109)	(0.0946)	(0.0528)	(0.0356)
Experience	0.0606	0.0525	0.0503	0.1075	0.0501	0.0363
	(0.0126)	(0.0113)	(0.0099)	(0.0647)	(0.0316)	(0.0219)
Experience squared	-0.0019	-0.0016	-0.0016	-0.0030	-0.0016	-0.0013
	(0.0005)	(0.0004)	(0.0004)	(0.0015)	(0.0006)	(0.0004)
Black	-0.0633	-0.0683	-0.0590	-0.1247	-0.0612	-0.0184
	(0.0385)	(0.0412)	(0.0342)	(0.0910)	(0.0752)	(0.0523)
SMSA	0.1344	0.1317	0.1295	0.1400	0.1303	0.1216
	(0.0201)	(0.0181)	(0.0173)	(0.0214)	(0.0202)	(0.0186)
South	-0.0766	-0.1106	-0.1095	-0.0810	-0.1100	-0.1061
	(0.0184)	(0.0159)	(0.0158)	(0.0193)	(0.0162)	(0.0163)
Constant	4.7336	4.8681	4.8773	4.0223	4.8932	5.0284
	(0.0945)	(0.0783)	(0.0751)	(0.9699)	(0.4490)	(0.3171)
J-statistic			16.8			10.2
(p-value)			(0.0184)			(0.1189)
Observations	2040	2963	2963	2040	2963	2963

7 Conclusion

This paper has considered several methods that avoid the problem of dropping observations in the face of missing data on explanatory variables. We proposed a GMM procedure based on set of moment conditions in the context of a regression model with a regressor that has missing values. The moment conditions were obtained with minimal additional assumptions on the data, and the method was shown to provide efficiency gains for some of the parameters. The GMM approach was compared to some well known linear imputation methods and shown to be equivalent to an optimal version of such methods under stronger assumptions than used to justify the GMM approach and potentially more efficient under the more general conditions. The (sub-optimal) unweighted linear imputation and the commonly used dummy method were found to potentially provide a "cure that is worse than the disease." As noted in the Introduction, the GMM approach does not give rise to semiparametric efficiency in the sense of Robins et. al. (1994), although the proposed approach is more likely to be appealing to empirical researchers in economics. In future work, it would be interesting to examine the extent to which the GMM is less efficient than the semiparametric efficient estimator in in cases where one has little information about some of the objects required to implement the approach of Robins et. al. (1994).

The paper also shows how the GMM approach can be extended to other settings where estimation can be naturally cast into a method-of-moments framework. In Section 4, for instance, the GMM approach was used to provide estimators for cases in which an instrument or an endogenous regressor might have missing values. In ongoing work, we are considering the application of GMM methods to linear panel-data models with missing covariate data. For non-linear models, where the projection approach is no longer applicable, it appears that strong parametric assumptions on the relationship between missing and non-missing covariates are required (see, for example, Conniffe and O'Neill (2009)). Also, the theoretical development here has focused upon the case of a single missing covariate. The idea of efficient GMM estimation can be extended to additional missing covariates, as in Muris (2011).

References

Card, D. (1995), "Using geographic variation in college proximity to estimate the return to schooling," in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp.* Ed. L.N. Christophedes, E.K. Grant and R. Swidinsky, pp. 102-220. Toronto: University of Toronto Press.

Chaudhuri, S. and D. Guilkey (forthcoming), "GMM with multiple missing variables," *Journal of Applied Econometrics.*

Conniffe, D. and D. O'Neill (2009), "Efficient probit estimation with partially missing covariates," IZA Discussion Paper No. 4081.

Dahl, G. and S. DellaVigna (2009), "Does movie violence increase violent crime?" *Quarterly Journal of Economics* 124, pp. 677-734.

Dagenais, M. C. (1973), "The use of incomplete observations in multiple regression analysis: a generalized least squares approach," *Journal of Econometrics* 1, pp. 317-328.

Dardanoni, V., S. Modica and F. Peracchi (2011), "Regression with imputed covariates: A generalized missing-indicator approach," *Journal of Econometrics* 162, pp. 362-368.

Gourieroux, C. and A. Monfort (1981), "On the problem of missing observations in linear models," *Review of Economic Studies* 48(4), pp. 579-586.

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

Griliches, Z. (1986), "Economic data issues," in Griliches, Z. and Intrilligator, M., eds., *Handbook of Econometrics Vol III*, Amsterdam: New Holland.

Jones, M. P. (1996), "Indicator and stratification methods for missing explanatory variables in multiple linear regression," *Journal of the American Statistical Association* 91(433), pp. 222-230.

Mogstad, M. and M. Wiswall (2010), "Instrumental variables estimation with partially missing instruments," IZA Discusion Paper No. 4689.

Muris, C. (2011), "Efficient GMM estimation with a general missing data pattern," mimeo, Simon Fraser University.

Nijman, T. and F. Palm (1988), "Efficiency gains due to using missing data procedures in regression models," *Statistical Papers* 29, pp. 249-256.

Robins, J.M., A. Rohitzky and L. P. Zhao (1994), "Estimation of regression coefficients when some regressors are not always observed" *Journal of the American Statistical Association* 89, pp.846-866.

Wooldridge, J. (2007) "Missing Data" http://www.nber.org/WNE/lect_12_missing.pdf – part of the NBER Lecture Series "What's New in Econometrics" with Guido Imbens.