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: March 2010 (first draft: April 2009)

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 previous approaches in the literature (including imputation methods and a dummy-variable approach used in much recent 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 seminar participants at University of British Columbia, Northwestern University, and the Midwest Econometrics Group for their helpful comments.

1 Introduction

A common feature of many data sets used in empirical research is that of missing information for certain variables. For example, an explanatory variable 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." Alternatively, if the variable in question is considered important and to be "missing at random," then a simple way to deal with the problem is to omit the observations and estimate a model using observations with "complete" data. This method could, however, result in a much smaller sample size. Based on this concern, various procedures have been suggested and implemented to avoid the problem of "throwing the baby out with the bathwater." In linear regression models, earlier work has suggested an "imputation method" by which one imputes the missing values for the regressor by an auxiliary regression of the regressor on the other variables using the complete data. Then one uses the imputed values in place of the missing values in the regression of interest and performs OLS (see Gourieroux and Monfort (1981)) or a GLS procedure (see Dagenais (1973)). Another approach that we refer to as the "dummy variable method"[1] sets the missing value to zero and uses dummies or indicators for whether the regressor was missing for the observation.

Data missingness is a very common problem facing empirical researchers. To give a sense of its prevalence and also the methods used to deal 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).[2] 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%)

[1] Greene (2008) refers to this method, in the context of a simple regression, as the "modified zero order regression."

[2] 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.

report that they have dropped observations due to missing values. Both the "dummy-variable method" and the "imputation method" are quite common approaches to handling missing data, with each being used in roughly 20% of the missing-data papers. Except in the case of simple regression, the dummy-variable method is known to generally lead to biased and inconsistent estimation (Jones (1996)), yet Table 1 clearly indicates the method's prominence despite the inconsistency associated with it.

In this paper, we consider an approach to the missing regressor problem whereby we develop moment conditions involving the observed variables. These include moment conditions from the regression function for the complete data as well as moment conditions that are satisfied by the observed data when one of the regressors is missing. The latter conditions are general in the sense that they only use the linear projection from the regressor with missing values onto the other regressors and do not require any model such as a conditional mean. We show that efficiency gains are possible if we add to these conditions the moment conditions coming from the linear projection based on the complete data. Our method then involves a Generalized Method of Moments (GMM) procedure involving all the available moment conditions.

Our approach based on GMM and linear projections also sheds light on some of the advantages and disadvantages of the previous approaches. The two-step (unweighted) imputation method is shown to be not necessarily any more efficient than the "complete data method." Moreover, though the method is computationally straightforward, the usual standard errors generated by the second step regression would be inappropriate without being adjusted. By contrast, the GMM procedure is shown to be asymptotically at least as efficient as the complete data estimator under very general conditions. Moreover, the standard errors generated by the GMM procedure will be valid, and as a byproduct of our GMM procedure we can obtain an overidentifying restrictions test that will reject when the assumptions of the method are violated. We also show that, under an assumption of homoskedasticity of the various residuals, the GMM method has the same asymptotic variance as the Dagenais (1973) estimator. On the other hand, the GMM estimator involves optimal weighting regardless of the variance structure of the residuals and is asymptotically more efficient under heteroskedasticity. The "dummy variable" method in contrast, as previously shown by Jones (1996), 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 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.

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 and Z_i is a K-vector of (never missing) regressors. The first element of Z_i is 1 so that the model is assumed to contain an intercept. We assume that 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 that has the form

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

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, we will not impose any homoskedasticity assumptions on ξ_i or ε_i. An implication of (2.2) and (2.3) is that

$$E(\xi_i\varepsilon_i) = 0. \tag{2.4}$$

Observations with missing X_i are problematic since (2.1) cannot be used directly to construct moment conditions for estimating $(\alpha_0, \beta_0')'$ — all that we see for such observations is the

combination (Y_i, Z_i'). Note, however, that (2.1) and (2.3) imply

$$Y_i = Z_i'(\gamma_0\alpha_0 + \beta_0) + \varepsilon_i + \xi_i\alpha_0 \stackrel{def}{=} Z_i'(\gamma_0\alpha_0 + \beta_0) + \eta_i. \tag{2.5}$$

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" assumption on m_i:

Assumption 1 *(i)* $E(m_i Z_i \varepsilon_i) = 0$*; (ii)* $E(m_i Z_i \xi_i) = 0$*; (iii)* $E(m_i X_i \varepsilon_i) = 0$.

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" (or 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 $E(\xi_i|Z_i) = 0$.[3]

We define a vector of moment functions based upon (2.2), (2.3), and (2.5),

$$g_i(\alpha, \beta, \gamma) = \begin{pmatrix} (1 - m_i)W_i(Y_i - X_i\alpha - Z_i'\beta) \\ m_i Z_i (Y_i - Z_i'(\gamma\alpha + \beta)) \\ (1 - m_i)Z_i(X_i - Z_i'\gamma) \end{pmatrix} = \begin{pmatrix} g_{1i}(\alpha, \beta, \gamma) \\ g_{2i}(\alpha, \beta, \gamma) \\ g_{3i}(\alpha, \beta, \gamma) \end{pmatrix}, \tag{2.6}$$

for which the following result holds:

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

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

Lemma 1 implies that the model and Assumption 1 generate a vector of $(1 + 3K)$ moment conditions satisfied by the population parameter values $(\alpha_0, \beta_0, \gamma_0)$. Since there are $(1 + 2K)$ 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" (which utilizes only g_{1i}), given by

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

Lemma 2 *The GMM estimators of $\theta_0 = (\alpha_0, \beta_0')'$ based on the set of moments $(g_{1i}(\alpha, \beta, \gamma)', g_{2i}(\alpha, \beta, \gamma)')'$ or $(g_{1i}(\alpha, \beta, \gamma)', g_{3i}(\alpha, \beta, \gamma)')'$ 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. To examine this in further detail, we consider the asymptotic variance of the standard optimally weighted GMM procedure. Some additional notation will prove useful for describing the optimal weight matrix. Let $\lambda = P(m_i = 0)$, so that in the asymptotic theory λ represents the asymptotic proportion of data that have observations on X_i. Also, define the following matrices (all of which are $K \times K$):

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

so that the subscript "m" indicates an expectation for the group of observations with missing values and the subscript "c" represents the group without missing values. Note that, in the general case, we allow for the possibility that the second moment matrix of Z_i and the variance covariance matrices of the moment conditions can differ between observations that have missing

X_i and those that do not. If m_i is completely random (i.e., independent of all other variables), then $\Gamma_m = \Gamma_c$, $\Omega_{\varepsilon m} = \Omega_{\varepsilon c}$, and $\Omega_{\xi m} = \Omega_{\xi c}$.

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,

$$\Omega = E(g_i(\alpha_0, \beta_0, \gamma_0) g_i(\alpha_0, \beta_0, \gamma_0)') = \begin{pmatrix} \Omega_{11} & 0 & 0 \\ 0 & \Omega_{22} & 0 \\ 0 & 0 & \Omega_{33} \end{pmatrix}, \tag{2.7}$$

where

$$\Omega_{11} = E((1 - m_i) W_i W_i' \varepsilon_i^2) = E(W_i W_i' \varepsilon_i^2 | m_i = 0) P(m_i = 0) = \lambda \Omega_{\varepsilon c},$$
$$\Omega_{22} = E(m_i Z_i Z_i' \eta_i^2) = E(Z_i Z_i' \eta_i^2 | m_i = 1) P(m_i = 1) = (1 - \lambda) \Omega_{\eta m} = (1 - \lambda)(\Omega_{\varepsilon m} + \alpha_0^2 \Omega_{\xi m}),$$
$$\text{and } \Omega_{33} = E((1 - m_i) Z_i Z_i' \xi_i^2) = E(Z_i Z_i' \xi_i^2 | m_i = 0) P(m_i = 0) = \lambda \Omega_{\xi c}.$$

The block diagonal structure in (2.7) follows from $m_i(1 - m_i) = 0$ and $E(\xi_i \varepsilon_i) = 0$. To implement the optimally weighted GMM procedure, we take sample analogs of the three blocks and estimate the residuals using a preliminary consistent procedure:

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

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

The optimal two-step GMM procedure then solves the following problem,

$$\min_{\alpha, \beta, \gamma} \bar{g}(\alpha, \beta, \gamma)' \hat{\Omega}^{-1} \bar{g}(\alpha, \beta, \gamma), \tag{2.8}$$

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 $\hat{\Omega}_{11}$, $\hat{\Omega}_{22}$, and $\hat{\Omega}_{33}$ into (2.7). Before discussing the asymptotic properties, we note that there are alternatives to this standard GMM procedure. This GMM procedure requires numerical methods since there is a nonlinear restriction on the parameters. However, our simulations show that this type of problem is very well-behaved and can be easily optimized in Stata (Version 11 or later) and other econometrics packages. Alternatively, one could achieve the same asymptotic

results by taking one iteration of a Newton-type algorithm starting from an initial consistent estimator; this procedure is described in the Appendix.

To describe the properties of the GMM estimators, let G denote the gradient-matrix corresponding to the moment conditions in (2.6), specifically

$$G = \begin{pmatrix} G_{11} & 0 \\ G_{21} & G_{22} \\ 0 & G_{32} \end{pmatrix} \tag{2.9}$$

where the components of the matrix are

$$\begin{aligned} G_{11} &= -E((1-m_i)W_iW_i') \\ G_{21} &= \begin{pmatrix} -E(m_iZ_iZ_i'\gamma_0) & -E(m_iZ_iZ_i') \end{pmatrix} \\ G_{22} &= -E(m_iZ_iZ_i'\alpha_0) \\ G_{32} &= -E((1-m_i)Z_iZ_i'). \end{aligned}$$

Note that each component can be consistently estimated by plugging in consistent estimators of (α_0, γ_0) and taking sample averages.

Under standard regularity conditions, we have the following result:[4]

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

$$n\bar{g}(\hat{\alpha}_G, \hat{\beta}_G, \hat{\gamma}_G)'\hat{\Omega}^{-1}\bar{g}(\hat{\alpha}_G, \hat{\beta}_G, \hat{\gamma}_G) \xrightarrow{d} \chi^2(K) \tag{2.10}$$

The behavior of the objective function in (2.10) gives rise to the possibility of testing the overidentifying restrictions imposed by the assumptions of the model. The crucial assumptions that would result in a rejection are essentially those in Assumption 1. To interpret these conditions, note that for instance

$$E((1-m_i)W_i\varepsilon_i) = E(W_i\varepsilon_i) - E(m_iW_i\varepsilon_i)$$

[4]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.

so according to the model this would not equal zero whenever $W_i\varepsilon_i$ has nonzero expectation for observations where X_i is missing. This would occur when the regression relationship between Y_i and $(X_i, Z_i')'$ is different for observations with missing values on X_i compared to the complete data observations. Similarly, there would be a violation whenever the relationship between X_i and Z_i differs between the two sets of observations. Since, as we show 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.

In what follows, we show that the GMM procedure can potentially yield efficiency gains relative to the complete data estimator. For the efficiency comparisons below, the following additional notation will be useful:

$$\sigma^2_{\xi c} = E(\xi_i^2|m_i = 0),\ \omega_{\xi\varepsilon c} = E(\xi_i^2\varepsilon_i^2|m_i = 0),\ \sigma^2_\varepsilon = E(\varepsilon_i^2),\ \sigma^2_\xi = E(\xi_i^2).$$

Then we have the following result concerning the complete data estimator:

Lemma 3 *Under Assumption 1, the asymptotic variances for the complete data estimator are*

$$\begin{aligned} AVAR(\sqrt{n}(\hat{\alpha}_C - \alpha_0)) &= \frac{1}{\lambda}\left(\sigma^2_{\xi c}\right)^{-1}\omega_{\xi\varepsilon c}\left(\sigma^2_{\xi c}\right)^{-1} \\ AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) &= \frac{1}{\lambda}\left(\Gamma_c\Omega^{-1}_{\varepsilon c}\Gamma_c\right)^{-1} + \frac{1}{\lambda}\left(\sigma^2_{\xi c}\omega^{-1}_{\xi\varepsilon c}\sigma^2_{\xi c}\right)^{-1}\gamma_0\gamma_0' \end{aligned}$$

These asymptotic-variance expressions permit heteroskedasticity and allow for different variances between the complete data group and the group with missing data. For comparisons with the imputation methods and the dummy variable method in Section 3, it is useful to examine these formulae under the following stronger assumptions:

Assumption 2 *(i)* $\Gamma_m = \Gamma_c = \Gamma$, *(ii)* $\Omega_{\varepsilon m} = \Omega_{\varepsilon c} = \Omega_\varepsilon = \sigma^2_\varepsilon\Gamma$, *(iii)* $\Omega_{\xi m} = \Omega_{\xi c} = \sigma^2_\xi\Gamma$, *(iv)* $\omega_{\xi\varepsilon c} = \sigma^2_\xi\sigma^2_\varepsilon$.

These conditions will be satisfied when the the data on X_i are MCAR and when the residuals are homoskedastic (see, e.g., Gourieroux and Monfort (1981) or Nijman and Palm (1988)).

Under Assumption 2, the asymptotic-variance expressions from Lemma 3 simplify to

$$\begin{aligned} AVAR(\sqrt{n}(\hat{\alpha}_C - \alpha_0)) &= \frac{1}{\lambda}\sigma_\varepsilon^2 \left(\sigma_\xi^2\right)^{-1} \\ AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) &= \frac{1}{\lambda}\sigma_\varepsilon^2 \Gamma^{-1} + \frac{1}{\lambda}\sigma_\varepsilon^2 \left(\sigma_\xi^2\right)^{-1} \gamma_0 \gamma_0'. \end{aligned}$$

Returning to the GMM estimator, the following lemma characterizes its asymptotic variance and provides comparisons to the complete data estimator:

Lemma 4 *Under Assumption 1,*

(i) the asymptotic variance for the GMM estimator of α_0 is

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

(ii) when $\alpha_0 = 0$ the asymptotic variance for the GMM estimator of β_0 is

$$\begin{aligned} AVAR(\sqrt{n}(\hat{\beta} - \beta_0)) &= \left(\lambda \Gamma_c \Omega_{\varepsilon c}^{-1} \Gamma_c + (1-\lambda)\Gamma_m \Omega_{\varepsilon m}^{-1} \Gamma_m\right)^{-1} + \frac{1}{\lambda}\left(\sigma_{\xi c}^2 \omega_{\xi\varepsilon c}^{-1} \sigma_{\xi c}^2\right)^{-1} \gamma_0 \gamma_0' \\ &\leq AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) \end{aligned}$$

(iii) when $\alpha_0 \neq 0$ the asymptotic variance of the GMM estimator of β_0 is

$$\begin{aligned} AVAR(\sqrt{n}(\hat{\beta} - \beta_0)) &= \left(\lambda' \Gamma_c \Omega_{\varepsilon c}^{-1} \Gamma_c + (1-\lambda) A\right)^{-1} + \frac{1}{\lambda}\left(\sigma_{\xi c}^2\right)^{-1} \gamma \omega_{\xi c} \gamma' \left(\sigma_{\xi c}^2\right)^{-1} \gamma_0 \gamma_0' \\ &\leq AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) \end{aligned}$$

where

$$A = \Gamma_m \Omega_{\eta m}^{-1} \Gamma_m \left\{ \left[\Gamma_m \Omega_{\eta m}^{-1} \Gamma_m\right]^{-1} - \left[\Gamma_m \Omega_{\eta m}^{-1} \Gamma_m + \frac{\lambda}{1-\lambda}\frac{1}{\alpha_0^2}\Gamma_c \Omega_{\xi c}^{-1} \Gamma_c\right]^{-1} \right\} \Gamma_m \Omega_{\eta m}^{-1} \Gamma_m \geq 0.$$

This result shows that the GMM estimator using the full set of moment conditions brings no improvement with respect to the estimation of the coefficient on the missing variable. There are improvements in the estimation of the coefficients on Z_i, except when $\lambda = 1$. The improvements will depend on the extent of the missingness as measured by $(1-\lambda)$. To provide further insight as to efficiency gains, we can show that under Assumption 2 the formulae in (ii) and (iii) simplify to

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

and

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

In the case where $\alpha_0 = 0$, the difference in the asymptotic variances is given by,

$$AVAR(\sqrt{n}(\hat{\beta}_C - \beta_0)) - AVAR(\sqrt{n}(\hat{\beta} - \beta_0)) = \left(\frac{1}{\lambda} - 1\right)\Gamma^{-1}$$

so that efficiency gains depend only on the amount of missingness. Of course in this situation one can do even better by imposing the restriction that X_i be eliminated and estimating the model using OLS on all observations which, under Assumption 2 would yield an estimator with asymptotic variance given by,

$$\sigma_\varepsilon^2 \Gamma^{-1} \leq AVAR(\sqrt{n}(\hat{\beta} - \beta_0))$$

In comparing GMM and the complete data estimator when $\alpha_0 \neq 0$ we find that,

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

so that the GMM estimator will be relatively more efficient in general and will be increasingly so the larger the amount of missing data and the larger the variance of ε relative to that of ξ.

2.1 Further Insights: A System-of-Equations Viewpoint

One can obtain further insight into the source of the efficiency gains by writing the model as a system of linear equations,

$$\begin{aligned} Y_i &= (1-m_i)X_i\alpha_0 + Z_i'\beta_0 + m_i Z_i'\gamma_0\alpha_0 + \varepsilon_i + m_i\xi_i\alpha_0 \qquad (2.11) \\ (1-m_i)X_i &= (1-m_i)Z_i'\gamma_0 + (1-m_i)\xi_i \end{aligned}$$

This can be thought of as a system of linear equations in the sense of seemingly unrelated regressions. In that context, one can see that the residuals in the two equations are uncorrelated. However, there are cross-equation nonlinear restrictions on the parameters that allow for efficiency improvements over the complete data estimator. These efficiency gains would be lost if the system were to be estimated equation by equation.

Lemma 5 *If OLS is used to estimate each equation in (2.11) then the estimates of* α_0 *and* β_0 *would be identical to* $\hat{\alpha}_C$ *and* $\hat{\beta}_C$*.*

This result implies that the efficiency gains for estimation of β_0 come from imposing the restrictions implied by the model. This could potentially be done using the formulation in (2.11) using a program that is capable of imposing nonlinear restrictions on SUR estimates. One should also note that even if the residuals ε_i and ξ_i are conditionally homoskedastic, the residual in the first equation is necessarily heteroskedastic due to the dependence on m_i except in the case where $\alpha_0 = 0$. This can be seen because under conditional homoskedasticity,

$$E((\varepsilon_i + m_i\xi_i\alpha_0)^2 \,|X_i, Z_i, m_i) = \sigma_\varepsilon^2 + \alpha_0^2 m_i \sigma_\xi^2$$

so that the variance for observations where X_i is missing is larger than the variance for observations where it is observed. As a result of this fact, applying OLS to the first equation would result in estimates for α_0 and β_0 that are identical to $\hat{\alpha}_C$ and $\hat{\beta}_C$, however the usual standard errors produced for OLS will no longer be valid even if ε_i and ξ_i are conditionally homoskedastic. One would expect the variance provided by OLS to be too large for the estimates of α_0 and β_0 since it would use as an estimate of the residual variance that is a weighted average of the residual variance for observations without missing values (which have variance σ_ε^2) and the residual variance for observations with missing values (whose residual has variance $\sigma_\varepsilon^2 + \alpha_0^2\sigma_\xi^2 \geq \sigma_\varepsilon^2$).

To obtain the full efficiency gains from estimating the model as a system one would need to take account of this heteroskedasticity. If one had access to a package that was capable of performing SUR but was not flexible enough to allow for heteroskedastic ε_i and ξ_i then a more appropriate formulation that could be used is,

$$\begin{aligned}
(1-m_i)Y_i &= (1-m_i)X_i\alpha_0 + (1-m_i)Z_i'\beta_0 + (1-m_i)\varepsilon_i \\
m_iY_i &= m_iZ_i'(\beta_0 + \gamma_0\alpha_0) + m_i(\varepsilon_i + \xi_i\alpha_0) \\
(1-m_i)X_i &= (1-m_i)Z_i'\gamma_0 + (1-m_i)\xi_i
\end{aligned}$$

If ε_i and ξ_i are homoskedastic, then the residuals in the three equations would be homoskedastic. Implementing this procedure would entail creating interactions of the dummies and the variables

in the model. Most programs will use equation by equation OLS to obtain estimates of the residual variances and covariances. Notice that, by construction, the residuals in the second equation are uncorrelated with those in either the first or third equation. Also, the least squares residual in the third equation will have a zero correlation with that in the first equation due to the fact that the former can be written

$$(1 - m_i)\hat{\xi}_i = (1 - m_i)X_i - (1 - m_i)Z_i'\hat{\gamma}$$

(where $\hat{\gamma}$ is the OLS estimator based on the third equation) and the fact that the OLS residual in the first equation is exactly orthogonal with $(1 - m_i)X_i$ and $(1 - m_i)Z_i$.

3 Comparison to Other Methods

It is of interest to compare the properties of the GMM estimator described in the previous section with other estimators suggested in the literature as well as those that appear to be in common use in empirical work. Section 3.1 considers the method known as linear imputation, whereby one uses observations on complete data to obtain predictions for X for the missing data based on a linear regression. Section 3.2 considers an approach that is based on the use of dummy variables to take "account" of the missingness. As we show in these sections, the linear imputation method is generally less efficient than the GMM method whereas the dummy variable method is potentially inconsistent and, even in situations where consistency is obtained, it is less efficient than the GMM procedure.

3.1 Linear Imputation

The linear imputation method proposed by Dagenais (1973) and discussed further by Gourieroux and Monfort (1981) can be thought of as a sequential approach to the estimation of (2.11).[5] First, one uses the second equation in (2.11) to estimate γ_0 by

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

[5]For related work on linear imputation, see also Conniffe (1983) and Lien and Rearden (1992).

and then one uses this estimate in the first equation which is then used to estimate α_0 and β_0. Note that when one makes this substitution one has

$$\begin{aligned} Y_i &= (1-m_i)X_i\alpha_0 + Z_i'\beta_0 + m_i Z_i'\hat{\gamma}\alpha_0 + \varepsilon_i + m_i\xi_i\alpha_0 + m_i Z_i'\left(\gamma_0 - \hat{\gamma}\right)\alpha_0 \qquad (3.12) \\ &= \left((1-m_i)X_i + m_i Z_i'\hat{\gamma}\right)\alpha_0 + Z_i'\beta_0 + \varepsilon_i + m_i\xi_i\alpha_0 + m_i Z_i'\left(\gamma_0 - \hat{\gamma}\right)\alpha_0. \end{aligned}$$

Note that the second line in (3.12) shows that with this substitution one has a regression model of Y_i on $\hat{X}_i = ((1-m_i)X_i + m_i Z_i'\hat{\gamma})$ and Z_i — that is, X_i is used if it is observed and otherwise one uses the linearly imputed value $Z_i'\hat{\gamma}$ in its place. Note also that the residual in the regression equation is necessarily heteroskedastic (unless $\alpha_0 = 0$) even if the components of the residuals are homoskedastic. Indeed, the conditional variance of the residual in this case (i.e., under homoskedasticity of ε_i and ξ_i) is given by

$$\sigma_\varepsilon^2 + m_i\sigma_\xi^2\alpha_0^2 + \sigma_\xi^2\alpha_0^2 m_i Z_i'\left(\sum_{i=1}^{n}(1-m_i)Z_i Z_i'\right)^{-1} Z_i, \qquad (3.13)$$

where the last term disappears asymptotically (since the elements of the matrix summation get arbitrarily large as n grows). Gourieroux and Monfort (1981) use the label "Ordinary Dagenais Estimator" for the OLS estimator for this model — based on that we use the notation $\hat{\theta}_{OD} = (\hat{\alpha}_{OD}, \hat{\beta}_{OD}')'$ and will also refer to it as the "unweighted linear imputation" estimator. In terms of computation, this method is appealing since it simply requires OLS to obtain γ_0 estimates and then one just predicts (or imputes) X_i for missing observations and uses these in place of the missing X_i values in another OLS regression. Though the estimation method is straightforward, care must be taken in the calculation of standard errors since, as (3.13) shows, the residual is necessarily heteroskedastic (unless $\alpha_0 = 0$) even if ε_i and ξ_i are homoskedastic. Moreover, the standard heteroskedasticity robust standard errors will also be invalid since the third component in the variance in (3.13) represents the effect of estimation error in estimating γ_0 on the variance and will not be captured in the robust variance computated by most packages. Appropriate standard errors that are also robust to heteroskedasticity in ε_i and ξ_i can be found using the methods described in Newey and McFadden (1994) or Wooldridge (2002).[6] If we let

[6] Alternatively, the bootstrap can be used to compute standard errors, as long as the bootstrap procedure replicates both steps of the procedure in order to account for the imputation noise.

$\hat{W}_i = (\hat{X}_i : Z_i')'$ and denote

$$\begin{aligned}\hat{e}_i &= Y_i - \hat{X}_i\hat{\alpha}_{OD} - Z_i'\hat{\beta}_{OD} \\ \hat{\xi}_i &= (1-m_i)X_i - (1-m_i)Z_i'\hat{\gamma}_{OD},\end{aligned}$$

then the variance-covariance matrix can be estimated by

$$\begin{aligned}\hat{V}\begin{pmatrix}\hat{\alpha}_{OD}\\ \hat{\beta}_{OD}\end{pmatrix} &= \left(\sum_i \hat{W}_i\hat{W}_i'\right)^{-1}(S_1+S_2)\left(\sum_i \hat{W}_i\hat{W}_i'\right)^{-1} \\ S_1 &= \sum_i \hat{W}_i\hat{W}_i'e_i^2 \\ S_2 &= \hat{\alpha}_{OD}^2\left(\sum_i m_i\hat{W}_iZ_i'\right)\hat{V}(\hat{\gamma})\left(\sum_i m_iZ_i\hat{W}_i'\right) \\ \hat{V}(\hat{\gamma}) &= \left(\sum_i (1-m_i)Z_iZ_i'\right)^{-1}\left(\sum_i (1-m_i)Z_iZ_i'\hat{\xi}_i^2\right)\left(\sum_i (1-m_i)Z_iZ_i'\right)^{-1}\end{aligned}$$

The term S_2 captures the effect of estimation error in the imputation. Were one to simply estimate (3.12) by OLS using heteroskedasticity consistent standard errors the variance would omit the term involving S_2 and would result in invalid standard errors except in the case where in fact $\alpha_0 = 0$. Practically speaking if $\hat{\alpha}$ is small then the usual robust standard errors would be very close to those given by the above expression.

The method proposed by Dagenais (1973) that Gourieroux and Monfort (1981) label "Generalized Dagenais Estimator" is a weighted least squares estimator that uses estimates of the variance in (3.13) to obtain weights for estimating the equation in (3.12). We denote this estimator by $\hat{\theta}_{GD} = (\hat{\alpha}_{GD}, \hat{\beta}_{GD}')'$ and will also refer to it as the "weighted linear imputation" estimator. The efficiency comparisons of the imputation (weighted and unweighted) estimators with the GMM estimator are summarized in the following proposition:

Proposition 2 *Under Assumption 1,*

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

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

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

This result shows that under the most general conditions there are no efficiency gains in using either the weighted or unweighted linear imputation estimators relative to the GMM procedure of the previous section. Generally speaking, there are never efficiency gains available in the estimation of α_0. For the estimation of β_0, one can gain intuition as to why the results in (ii) and (iii) of Proposition 2 hold by considering the linear imputation methods as GMM estimators. For (ii), the estimator $\hat{\theta}_{OD}$ can be thought of as a sequential GMM estimator with a suboptimal weight matrix in the second stage. That is, one first estimates γ_0 using the third moment function. This is then plugged into the first two moment functions so that the estimator can be found by solving,

$$\min_{\theta} \frac{1}{n}\sum_i \begin{pmatrix} (1-m_i)W_i(Y_i - X_i\alpha - Z_i'\beta) \\ m_i Z_i \left(Y_i - Z_i'\left(\hat{\gamma}\alpha + \beta\right)\right) \end{pmatrix} H'H \frac{1}{n}\sum_i \begin{pmatrix} (1-m_i)W_i(Y_i - X_i\alpha - Z_i'\beta) \\ m_i Z_i \left(Y_i - Z_i'\left(\hat{\gamma}\alpha + \beta\right)\right) \end{pmatrix}$$

with

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

Note that the weight matrix $H'H$ yields the least squares estimator since

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

which are the normal equations. For the weighted estimator, the result can be shown by noting that the estimator $(\hat{\alpha}_{GD}, \hat{\beta}_{GD}')'$ behaves asymptotically like the GMM estimator using the moment conditions (2.6) that uses a weight matrix that assumes the residuals ε_i and ξ_i are homoskedastic (and that uses the same estimates for the residual variances). Since such a GMM estimator is generally no more efficient than the estimator that uses the general weight matrix (2.7), the result holds and $\hat{\beta}_{GD}$ is no more efficient than $\hat{\beta}$.

To gain further insight, the following result gives the variance formulae and comparisons under the stronger homogeneity and homoskedasticity assumptions in Assumption 2. With these stronger conditions, the formula for the asymptotic variance simplifies to

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

Gourieroux and Monfort (1981) derived the same expression for this estimator. There are two things to note. Not surprisingly, in view of Proposition 2, this variance is larger than the variance for the GMM estimator since

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

The only instance in which the variances are the same is when $\alpha_0 = 0$ which of course is the situation where the variable with missing values can be dropped from the model. Otherwise one would expect the GMM estimator to be more efficient with efficiency gains more pronounced the larger is σ_ξ^2 relative to σ_ε^2. One can also compare the asymptotic variance of $\hat{\beta}_{OD}$ with that of $\hat{\beta}_C$ and in doing so one finds that the difference is,

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

so it is not necessarily the case that this unweighted method is any better for estimating β_0.[7] One can see why the unweighted estimator may be less efficient by examining the pieces that make up the variance. If γ_0 was known, then the variance would be,

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

which is clearly lower than the variance for the complete data estimator. There is an offsetting piece $\frac{(1-\lambda)\sigma_\xi^2\alpha_0^2}{\lambda\sigma_\varepsilon^2}\Gamma^{-1}$ that comes from the fact that γ_0 was estimated and if this is sufficiently large then it can wipe out the efficiency gains from using linear imputation and the whole data set. This is most likely to happen the larger is the variance of the imputation residual ξ_i and the more important the variable X_i.

For the weighted imputation estimator, the asymptotic variance simplifies to (Gourieroux and Monfort (1981))

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

[7] As noted by Griliches (1986), the claim by Gourierouux 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.

which is identical to that of the GMM estimator in this case. Indeed, as noted above, the weighted estimator behaves like a GMM estimator constructed with a weight matrix that is valid under homoskedasticity — when there is homoskedasticity the two estimators are asymptotically equivalent. In view of (3.13) one can see when the efficiency gains of the weighted estimator compared to the unweighted estimator. When $\alpha_0 = 0$ then the residual variance is homoskedastic and the weighted and unweighted estimators are equally efficient and more efficient than the complete data estimator. The gains for the weighted estimator relative to the unweighted estimator arise when $\alpha_0 \neq 0$ and will be larger the larger is the variance σ_ξ^2 — intuitively the larger this variance the more heteroskedastic the residual and the greater the advantage from the weighted estimator.

3.2 Dummy Variable Method

The method whereby one uses a zero (or some other value, like $\bar{X}$) for the missing value of X_i and compensates by including a dummy variable for "missingness" is what we refer to as the "dummy variable method." One can represent this method by using the formulation in (2.11) which we repeat here:

$$\begin{aligned} Y_i &= (1-m_i)X_i\alpha_0 + Z_i'\beta_0 + m_i Z_i'\gamma_0\alpha_0 + \varepsilon_i + m_i\xi_i\alpha_0 \\ &= (1-m_i)X_i\alpha_0 + Z_i'\beta_0 + m_i\gamma_{10}\alpha_0 + m_i Z_{2i}'\gamma_{20}\alpha_0 + \varepsilon_i + m_i\xi_i\alpha_0 \end{aligned} \tag{3.14}$$

In this equation, the intercept has been separated out from the other components of Z_i, with the latter denoted by the subvector Z_{2i}. The coefficient vector γ_0 is likewise partitioned into γ_{10} (for the intercept) and γ_{20} (for the subvector Z_{2i}). The dummy variable method amounts to running the regression without the regressors $m_i Z_{2i}'$. Let $\hat{\theta}_{DM} = (\hat{\alpha}_{DM}, \hat{\beta}_{DM})'$ denote the dummy variable estimator based on running the regression in (3.14). 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 3 *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 available data to regress Y_i on Z_i. The second condition requires that Z_{2i} is not useful for predicting X_i (using a linear predictor), which is unlikely to hold in practice except for randomized X_i quantities.

The variance of dummy variable estimator can be compared to that of other methods under different scenarios for which the dummy variable estimator is consistent. To make the comparisons clean, we also assume that

$$\Gamma = \begin{pmatrix} 1 & 0 \\ 0 & \Gamma_{22} \end{pmatrix} \tag{3.15}$$

Since the first element of Z_i is 1, this assumption 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 efficiency comparisons of the dummy-variable estimator with other approaches are summarized in the following proposition:

Proposition 4 *Under Assumptions 1 and 2 and the normalization in (3.15),*

(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 when $\gamma_{20} \neq 0$. Note that when in fact $\gamma_{20} = 0$ as well then there is no efficiency gain possible for estimating the coefficient of the missing variable. One can compare the variance of the estimate of α that would be possible if the X_i were fully observed — that variance under these conditions would be given by simply (use 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

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

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 and Z 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 under the strong assumptions it would be better to drop X completely and just use Z 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 larger than 1 in practice. The second part of (i) suggests that the estimator for the slopes in β_0 will be more efficiently estimated by the dummy method compared to the GMM method since,

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

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

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

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

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

which is equivalent to the condition regarding efficiency of 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 before it is less efficient than the GMM or weighted imputation estimator. Given that the complete data estimator can be obtained exactly by the regression in the first line of (2.11) it seems surprising that imposing the restriction that $\gamma_{20}=0$ would result in a larger variance when this is in fact a valid restriction. The reason seems to be related to the fact that when $\alpha_0\neq 0$ the residual in the model is heteroskedastic.

The results of this section suggest that there is not much 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 for this method. 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. In certain situations, it could actually have a larger variance than the complete data estimator. For the estimates of the coefficients on Z_i, the result in (i) shows that the dummy variable estimator is potentially more efficient than the complete data estimator. Comparisons with the GMM estimator depend on the values of the parameters.

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 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,

$$Y_{1i} = Y_{2i}\delta_0 + Z_i'\beta_0 + \varepsilon_i, \quad E(Z_i\varepsilon_i) = 0, \; E(Y_{2i}\varepsilon_i) \neq 0 \tag{4.16}$$

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

$$Y_{2i} = X_i\pi_0 + Z_i'\Pi_0 + v_i, \; E(X_iv_i) = 0, \; E(Z_iv_i) = 0. \tag{4.17}$$

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

$$E(m_i X_i \varepsilon_i) = E(m_i X_i v_i) = E(m_i X_i \xi_i), \tag{4.18}$$

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

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

For this model, one is primarily interested in the estimation of the parameters of (4.16) so that efficient estimation of the parameters of (4.17) 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 (4.18), 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

$$Y_{2i} = (1 - m_i) X_i \pi_0 + Z_i' \Pi_0 + m_i Z_i' \gamma_0 \pi_0 + v_i + \pi_0 m_i \xi_i. \tag{4.19}$$

Then, partitioning $Z_i = (1, Z_{2i}')'$, the full-sample reduced form in (4.19) 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 (4.16) 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 (4.19)) $((1-m_i)X_i, Z_i, m_i, m_i Z_{2i}')$ 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 5 *If ε_i, v_i and ξ_i are conditionally homoskedastic and $E(Z_i Z_i') = I$, then:*

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

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

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 result shows that the latter described method does bring some improvements with respect to the estimation of β_0.

Proposition 6 *If ε_i, v_i and ξ_i are conditionally homoskedastic and $E(Z_i Z_i') = I$, then:*

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

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 (asymptotically) to deliver 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 (4.16) and (4.17), respectively:[8]

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

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

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.

[8] 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.

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 (2.6)) is given by

$$h_i(\delta, \beta, \pi, \Pi) = \begin{pmatrix} (1-m_i)W_i(Y_{1i} - Y_{2i}\delta - Z_i'\beta) \\ m_i W_i \left(Y_{1i} - X_i\pi\delta - Z_i'\left(\Pi\delta + \beta\right)\right) \\ (1-m_i)W_i(Y_{2i} - X_i\pi - Z_i'\Pi) \end{pmatrix} = \begin{pmatrix} h_{1i}(\delta, \beta, \pi, \Pi) \\ h_{2i}(\delta, \beta, \pi, \Pi) \\ h_{3i}(\delta, \beta, \pi, \Pi) \end{pmatrix}. \tag{4.20}$$

Note that consistency of the GMM estimator requires the following missing-at-random assumption:[9]

$$E(m_i W_i \varepsilon_i) = E(m_i W_i v_i) = 0.$$

There are a total of $3K + 3$ moments in (4.20) 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 $K = 2$,

$$\begin{aligned} Y_i &= X_i\alpha + \beta_1 + \beta_2 Z_{2i} + \varepsilon_i \\ X_i &= \gamma_1 + \gamma_2 Z_{2i} + \xi_i. \end{aligned}$$

We fix $(\beta_1, \beta_2, \gamma_1) = (1, 1, 1)$ throughout the experiments but consider different combinations of values for γ_2 and α (either 0.1 or 1). We generate $Z_{2i} \sim N(0,1)$, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and $\xi_i \sim N(0, \sigma_\xi^2)$ where we consider the cases where the variance parameters are either 1 or 10. At a variance value of 10, the goodness of fit of the models is roughly of an order found in empirical research in economics. We also consider experiments where we allow ε_i to be heteroskedastic with variance function given by,

$$V(\varepsilon_i | X_i, Z_{2i}) = \exp(0.5(\beta_1 + \beta_2 Z_{2i})).$$

For most of the designs, we consider a simple missingness mechanism in which exactly half of the X_i's are missing completely at random. In one design, however, we consider an extreme case

[9] Consistency of the complete-data estimator requires $E(m_i W_i \varepsilon_i) = 0$.

in which the X_i's are missing when Z_{2i} observations are below the median value (an extreme form of missingness allowed by Assumption 1). We consider a total of ten different designs, summarized as follows:

Design 1: $\sigma_\epsilon^2 = 1$, $\sigma_\xi^2 = 1$, $\alpha = 1$, $\gamma_2 = 1$

Design 2: $\sigma_\epsilon^2 = 1$, $\sigma_\xi^2 = 10$, $\alpha = 1$, $\gamma_2 = 1$

Design 3: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 1$, $\alpha = 1$, $\gamma_2 = 1$

Design 4: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 10$, $\alpha = 1$, $\gamma_2 = 1$

Design 5: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 10$, $\alpha = 0.1$, $\gamma_2 = 1$

Design 6: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 10$, $\alpha = 1$, $\gamma_2 = 0.1$

Design 7: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 10$, $\alpha = 0.1$, $\gamma_2 = 0.1$

Design 8: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 10$, $\alpha = 1$, $\gamma_2 = 1$

Design 9: $Var(\epsilon|X_i, Z_i) = \exp(0.5(\beta_1 + \beta_2 Z_{2i}))$, $\sigma_\xi^2 = 10$, $\alpha = 0.1$, $\gamma_2 = 0.1$

Design 10: $Var(\epsilon|X_i, Z_i) = \exp(0.5(\beta_1 + \beta_2 Z_{2i}))$, $\sigma_\xi^2 = 10$, $\alpha = 0.1$, $\gamma_2 = 0.1$, missingness based on Z_{2i}

For all simulations, a sample size of $n = 200$ is used. The results are reported in Tables 2–11. For a set of 1000 replications for each design, these tables report the bias, variance, and overall MSE for the estimators of the parameters $(\alpha, \beta_1, \beta_2)$. 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 estimator, (iv) the weighted imputation estimator, and (v) the (efficient) 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 α in Designs 1 and 3 where the relationship between X and Z is strongest (i.e., σ_ξ^2 is small and γ_2 is large). In these cases, the estimates of the β's are also quite badly biased for this estimator. The estimate of β_2 is also biased in

Designs 4 and 8 where there is a fair amount of noise in both equations. In Design 10, where missingness is based on the value of Z_2, the dummy estimates have a very pronounced bias. Even in cases where the bias is not as pronounced, there are a number of instances where the dummy estimator has substantially larger variance than the other estimators. Only in cases where there is a lot of noise and where α and/or γ_2 are small is there a benefit in terms of variance for some parameters that are sufficient to result in MSE improvements — for instance in Designs 5 and 6 for a subset of the parameters.

For the other estimators, there is no improvement in terms of estimating α as one might expect from the results in Sections 2 and 3. In fact, it appears that the weighted and unweighted imputation estimates of α are numerically identical to the complete data estimates. The GMM estimates are slightly different but biases and variances are very similar in all experiments. In terms of estimating β, the weighted imputation and GMM estimators are similar with the weighted imputation estimator slightly preferred in the homoskedastic cases. In the heteroskedastic cases, the GMM estimates have lower variances than the weighted imputation estimator except when missingness is based on Z_2 in Design 10. The weighted imputation and GMM estimators generally have lower variances than the complete data estimator. In Design 2, where the imputation equation is relatively noisy, there is much less improvement in terms of the variance. When α is small, the gains in terms of variance can be quite substantial.

The unweighted imputation estimator can also bring variance gains, as in Designs 3 and 5. On the other hand, it can also be much worse than the complete data estimator as in Design 2. Not surprisingly, the unweighted estimator is outperformed by the weighted version and GMM except in Design 7, when the weighting implied by (3.13) is approximately constant. This latter result is probably the result of some finite-sample variability induced by using estimates of the weights in the weighted and GMM estimators.

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	0.006	2.1	0.011
	β_1	-0.003	3.9	0.020
	β_2	-0.002	4.1	0.021
Dummy-variable method	α	-0.327	1.9	0.116
	β_1	0.330	4.1	0.129
	β_2	0.668	2.6	0.459
Unweighted imputation	α	0.006	2.1	0.011
	β_1	-0.003	4.3	0.021
	β_2	-0.005	4.2	0.021
Weighted imputation	α	0.006	2.1	0.011
	β_1	-0.003	3.6	0.018
	β_2	-0.003	3.7	0.018
GMM (efficient)	α	0.010	2.1	0.011
	β_1	-0.008	3.6	0.018
	β_2	-0.007	3.6	0.018

Design 1: $\sigma_\epsilon^2 = 1$, $\sigma_\xi^2 = 1$, $\alpha = 1$, $\gamma_2 = 1$. $n = 200$. 1000 simulations.

a GMM estimate with nonlinearity had to be computed via numerical methods 1000 times for each experiment (plus all the other calculations).

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	0.000	0.2	0.001
	β_1	-0.004	2.2	0.011
	β_2	0.004	2.3	0.012
Dummy-variable method	α	-0.049	0.3	0.004
	β_1	0.044	2.8	0.016
	β_2	0.532	8.0	0.323
Unweighted imputation	α	0.000	0.2	0.001
	β_1	-0.001	11.8	0.059
	β_2	-0.011	11.4	0.057
Weighted imputation	α	0.000	0.2	0.001
	β_1	-0.004	2.2	0.011
	β_2	0.002	2.2	0.011
GMM (efficient)	α	0.001	0.2	0.001
	β_1	-0.005	2.2	0.011
	β_2	0.002	2.2	0.011

Design 2: $\sigma_\epsilon^2 = 1$, $\sigma_\xi^2 = 10$, $\alpha = 1$, $\gamma_2 = 1$. $n = 200$. 1000 simulations.

Table 4: Monte Carlo simulations, Design 3

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α	-0.010	18.7	0.094
	β_1	0.014	39.7	0.199
	β_2	0.012	38.2	0.191
Dummy-variable method	α	-0.345	14.1	0.189
	β_1	0.349	34.2	0.293
	β_2	0.679	14.7	0.535
Unweighted imputation	α	-0.010	18.7	0.094
	β_1	0.016	30.4	0.152
	β_2	0.019	30.7	0.154
Weighted imputation	α	-0.010	18.7	0.094
	β_1	0.015	30.4	0.152
	β_2	0.019	30.5	0.153
GMM (efficient)	α	-0.002	18.9	0.095
	β_1	0.010	30.8	0.154
	β_2	0.008	30.5	0.152

Design 3: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 1$, $\alpha = 1$, $\gamma_2 = 1$. $n = 200$. 1000 simulations.

Table 5: Monte Carlo simulations, Design 4

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α	-0.001	2.1	0.011
	β_1	0.009	21.0	0.105
	β_2	0.015	24.0	0.120
Dummy-variable method	α	-0.049	2.1	0.013
	β_1	0.055	21.5	0.110
	β_2	0.543	17.3	0.381
Unweighted imputation	α	-0.001	2.1	0.011
	β_1	-0.017	22.8	0.115
	β_2	0.024	24.1	0.121
Weighted imputation	α	-0.001	2.1	0.011
	β_1	-0.004	16.5	0.082
	β_2	0.020	18.9	0.095
GMM (efficient)	α	0.003	2.1	0.011
	β_1	-0.008	16.7	0.084
	β_2	0.016	19.2	0.096

Design 4: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 10$, $\alpha = 1$, $\gamma_2 = 1$. $n = 200$. 1000 simulations.

Table 6: Monte Carlo simulations, Design 5

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α	0.002	2.0	0.010
	β_1	-0.006	22.9	0.114
	β_2	-0.011	21.7	0.109
Dummy-variable method	α	-0.004	1.9	0.010
	β_1	0.000	22.6	0.113
	β_2	0.052	11.0	0.058
Unweighted imputation	α	0.002	2.0	0.010
	β_1	0.007	13.3	0.067
	β_2	-0.003	12.5	0.063
Weighted imputation	α	0.002	2.0	0.010
	β_1	0.007	13.3	0.067
	β_2	-0.003	12.5	0.062
GMM (efficient)	α	0.002	2.1	0.010
	β_1	0.003	13.6	0.068
	β_2	-0.002	12.8	0.064

Design 5: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 10$, $\alpha = 0.1$, $\gamma_2 = 1$. $n = 200$. 1000 simulations.

Table 7: Monte Carlo simulations, Design 6

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α	0.001	2.0	0.010
	β_1	-0.015	21.8	0.109
	β_2	-0.002	20.8	0.104
Dummy-variable method	α	0.001	2.0	0.010
	β_1	-0.015	21.5	0.108
	β_2	0.045	15.8	0.081
Unweighted imputation	α	0.001	2.0	0.010
	β_1	-0.011	22.2	0.111
	β_2	-0.001	23.0	0.115
Weighted imputation	α	0.001	2.0	0.010
	β_1	-0.013	16.6	0.083
	β_2	-0.001	16.4	0.082
GMM (efficient)	α	0.006	2.1	0.011
	β_1	-0.020	17.0	0.085
	β_2	-0.002	16.6	0.083

Design 6: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 10$, $\alpha = 1$, $\gamma_2 = 0.1$. $n = 200$. 1000 simulations.

Table 8: Monte Carlo simulations, Design 7

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α	-0.004	2.2	0.011
	β_1	0.005	20.8	0.104
	β_2	0.012	22.0	0.110
Dummy-variable method	α	-0.004	2.1	0.011
	β_1	0.005	20.6	0.103
	β_2	0.007	10.1	0.050
Unweighted imputation	α	-0.004	2.2	0.011
	β_1	0.004	12.4	0.062
	β_2	0.003	10.3	0.052
Weighted imputation	α	-0.004	2.2	0.011
	β_1	0.004	12.3	0.062
	β_2	0.003	10.4	0.052
GMM (efficient)	α	-0.004	2.2	0.011
	β_1	0.004	12.6	0.063
	β_2	0.003	11.2	0.056

Design 7: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 10$, $\alpha = 0.1$, $\gamma_2 = 0.1$. $n = 200$. 1000 simulations.

Table 9: Monte Carlo simulations, Design 8

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α	-0.011	9.2	0.046
	β_1	0.021	101.0	0.506
	β_2	0.021	188.2	0.941
Dummy-variable method	α	-0.056	9.3	0.050
	β_1	0.065	93.8	0.473
	β_2	0.526	100.7	0.780
Unweighted imputation	α	-0.011	9.2	0.046
	β_1	0.028	67.5	0.338
	β_2	0.006	112.8	0.564
Weighted imputation	α	-0.011	9.2	0.046
	β_1	0.025	65.6	0.329
	β_2	0.007	111.1	0.555
GMM (efficient)	α	-0.003	9.1	0.046
	β_1	0.009	64.6	0.323
	β_2	-0.011	102.2	0.511

Design 8: $\sigma_\epsilon^2 = 10$, $\sigma_\xi^2 = 10$, $\alpha = 1$, $\gamma_2 = 1$. $n = 200$. 1000 simulations.

Table 10: Monte Carlo simulations, Design 9

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α	-0.014	9.6	0.048
	β_1	-0.023	102.6	0.514
	β_2	-0.088	184.8	0.932
Dummy-variable method	α	-0.015	9.6	0.048
	β_1	-0.018	101.4	0.507
	β_2	-0.018	78.9	0.395
Unweighted imputation	α	-0.014	9.6	0.048
	β_1	0.017	55.6	0.278
	β_2	-0.020	79.2	0.396
Weighted imputation	α	-0.014	9.6	0.048
	β_1	0.016	55.7	0.279
	β_2	-0.021	79.3	0.397
GMM (efficient)	α	-0.013	9.7	0.049
	β_1	0.019	58.1	0.291
	β_2	-0.011	75.5	0.378

Design 9: $Var(\epsilon|X_i, Z_i) = \exp(0.5(\beta_1 + \beta_2 Z_{2i}))$, $\sigma_\xi^2 = 10$, $\alpha = 0.1$, $\gamma_2 = 0.1$.
$n = 200$. 1000 simulations.

Table 11: Monte Carlo simulations, Design 10

Estimation method	Parameter	Bias	n*Var	MSE
Complete-case method	α	-0.006	9.4	0.047
	β_1	-0.092	259.4	1.306
	β_2	0.075	258.3	1.297
Dummy-variable method	α	-0.005	9.3	0.046
	β_1	0.765	99.9	1.085
	β_2	-0.999	93.8	1.468
Unweighted imputation	α	-0.006	9.4	0.047
	β_1	-0.008	57.0	0.285
	β_2	-0.007	48.4	0.242
Weighted imputation	α	-0.006	9.4	0.047
	β_1	-0.009	57.2	0.286
	β_2	-0.006	48.5	0.243
GMM (efficient)	α	-0.006	9.3	0.046
	β_1	-0.013	57.6	0.288
	β_2	0.001	49.7	0.248

Design 10: $Var(\epsilon|X_i, Z_i) = \exp(0.5(\beta_1 + \beta_2 Z_{2i}))$, $\sigma_\xi^2 = 10$, $\alpha = 0.1$, $\gamma_2 = 0.1$, missingness based on Z_{2i}. $n = 200$. 1000 simulations.

6 Empirical examples

This section considers application of the GMM (and other) 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.[10] 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.[11]

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 12, 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

[10] 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.

[11] 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 12: 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

Sections 2 and 3 has suggested, there is little difference in the standard errors for this covariate 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 12 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.[12] 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

[12] 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%.

indicator variable. IQ score is used as an instrument (X) for KWW, but IQ data is missing 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 13 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 13 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 13: 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)
Observations	2040	2963	2963	2040	2963	2963

7 Conclusion

This paper has considered a variety of 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."

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 stronger parametric assumptions on the relationship between missing and non-missing covariates are required (see, for example, Conniffe and O'Neill (2009)).

Appendix

<< **Proofs to be added here** >>

The one-step GMM estimator: As discussed in Section 2, an alternative GMM estimator that is efficient and does not require non-linear optimization is the one-step GMM estimator.

Let $\hat{\gamma}_C$ denote the complete data estimator of γ based on (2.3) and define

$$\hat{G} = \begin{pmatrix} \hat{G}_{11} & 0 \\ \hat{G}_{21} & \hat{G}_{22} \\ 0 & \hat{G}_{32} \end{pmatrix}$$

where

$$\begin{aligned} \hat{G}_{11} &= -\frac{1}{n}\sum_i (1-m_i) W_i W_i', \quad \hat{G}_{21} = \begin{pmatrix} -\frac{1}{n}\sum_i m_i Z_i Z_i' \hat{\gamma}_C & -\frac{1}{n}\sum_i m_i Z_i Z_i' \end{pmatrix} \\ \hat{G}_{22} &= -\frac{1}{n}\sum_i m_i Z_i Z_i' \hat{\alpha}_C, \quad \hat{G}_{32} = -\frac{1}{n}\sum_i (1-m_i) Z_i Z_i' \end{aligned}$$

Then one Newton step starting at the complete data estimator would involve the following calculation,

$$\begin{pmatrix} \hat{\theta}^{(1)} \\ \hat{\gamma}^{(1)} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_C \\ \hat{\gamma}_C \end{pmatrix} - (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}\hat{G}'\hat{\Omega}^{-1}\bar{g}(\hat{\alpha}_C, \hat{\beta}_C, \hat{\gamma}_C).$$

The standard one-step results, along the lines of those in Newey and McFadden (1994), imply that the estimator so obtained will have the same asymptotic distribution as the GMM estimator that solves (2.8). The properties of this one-step estimator are the same as those given in Proposition 1.

References

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behavior: Essay in Honour of John Vanderkamp.* Ed. L.N. Christophedes, E.K. Grant and R. Swidinsky, pp. 102-220. Toronto: University of Toronto Press.

Conniffe, D. (1983), "Small-Sample Properties of Estimators of Regression Coefficients Given a Common Pattern of Missing Data," *The Review of Economic Studies* 50(1), pp. 111-120.

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 Movie Violence," *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.

Gourieroux, C. and A. Monfort (1981), "On the Problem of Missing Observations in Linear Models," *Review of Economic Studies* 48(4), pp. 579-586.

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

Griliches, Z., B. H. Hall, and J. A. Hausman (1978), "Missing data and self-selection in large panels," *Annales de l'INSEE* 30/31, pp. 137-176.

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

Lien, D. and D. Rearden (1992), "A note on estimating regression coefficients with missing data," *Econometric Reviews* 11(1), pp. 119-122.

Newey, W. K. and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in Engle, R. F. and McFadden, D., eds., *Handbook of Econometrics Vol IV*, Amsterdam: New Holland.

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

Wooldridge, J. M. (2002), *Econometric analysis of cross section and panel data*, MIT Press.