A GMM APPROACH FOR DEALING WITH MISSING DATA ON REGRESSORS

Jason Abrevaya and Stephen G. Donald*

Abstract—Missing data are a common challenge facing empirical researchers. This paper presents a general GMM framework and estimator for dealing with missing values of an explanatory variable in linear regression analysis. The GMM estimator is efficient under assumptions needed for consistency of linear-imputation methods. The estimator, which also allows for a specification test of the missingness assumptions, is compared to existing linear imputation, complete data, and dummy variable methods commonly used in empirical research. The dummy variable method is generally inconsistent even when data are missing completely at random, and the dummy variable method, when consistent, can be less efficient than the complete data method.

I. Introduction

MISSING data are a common problem in empirical research in economics and other fields. An example, and the focus of this paper, is the situation where an explanatory variable ("covariate," "regressor") may be unavailable for many observational units. If this variable is an important part of the model, omitting it can lead to substantial omitted variables bias. In practice, as we document below, economists deal with this problem predominantly by using one of three methods. The first, and simplest, is the complete data method, where observations with missing covariate data are dropped. When data are missing for many observations, this approach can lead to a much smaller sample. A second common method, which we call the dummy variable method, is to enter a 0 for the missing information and include an additional dummy (indicator) variable that indicates "missingness." A third method is imputation, where the missing covariate is estimated based on the other available data. Linear imputation is a simple version of this approach in which a linear regression using the complete case observations is used to impute the missing values.

To give a sense of the prevalence of missing data in empirical research and the popularity of the different methods of dealing with missing data, we looked for empirical papers with missing values in four top empirical economics journals—*American Economic Review* (AER), *Journal of Human Resources* (JHR), *Journal of Labor Economics* (JLE), and *Quarterly Journal of Economics* (QJE)—over the three-year period 2006 to 2008.[1] Over half of the empirical papers in JLE and QJE have a missing data issue, and nearly 40% of all papers across the four journals have data missingness. Of the papers with missing data, a large majority (roughly 70%) report that they dropped observations due to missing values and hence use the complete data method (usually OLS). 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 papers with missing data.[2]

The choice of method comes down to considerations of bias and relative efficiency. If data are missing in a way that does not affect the OLS orthogonality conditions, then the complete data method is unbiased and consistent. While the complete data regression uses fewer observations, Dardanoni, Modica, and Peracchi (2011) have shown in a very general setting that efficiency gains require additional restrictions on the model. For instance, the linear imputation method of Dagenais (1973) and Gourieroux and Monfort (1981) is based on the assumption that the imputation equation holds in both the complete data and missing data subpopulations. Meanwhile, the dummy variable method is generally inconsistent unless the possibly missing covariate is orthogonal to the remaining covariates (Jones, 1996). Moreover, as we show, the dummy variable method does not even guarantee efficiency improvements over the complete data estimator.[3]

The issue of efficiency in missing data models has been considered previously in the literature. Robins, Rotnitzky, and Zhao (1994) deal explicitly with the situation of missing covariates in regression analysis and under quite general assumptions derive the form of semiparametric efficient estimators (see also Chaudhuri & Guilkey, 2016). Their framework also allows missingness to depend on the dependent variable, in which case inverse probability weighting of the score restores consistency and leads to an optimal estimator. Graham, Pinto, and Egel (2012) propose a similar estimator that exhibits improved higher-order bias properties. In general, this approach requires models for the missingness probability and certain conditional expectations of the score. While only one of these needs to be correctly specified in order to achieve consistency (a so-called double robustness property), the methods are somewhat difficult to implement in practice, requiring modeling and estimation of several nonparametric objects for which empirical researchers are unlikely to have much prior information. Perhaps for this reason, the approach does not seem to be used in empirical economics research.

Received for publication January 12, 2015. Revision accepted for publication July 20, 2016. Editor: Bryan S. Graham.

* University of Texas.

We are grateful to Shu Shen for excellent research assistance and to Garry Barrett, Yu-Chin Hsu, Robert Moffitt, seminar participants at several institutions, two anonymous referees, and the editor for their helpful comments.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/REST_a_00645.

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

[2] Some papers used more than one method for dealing with missing observations, which is why the percentages add up to more than 100%.

[3] One possible reason for the popularity of the dummy variable method is that it has been discussed in the various editions of the popular econometrics textbook by Greene (2003). There it is referred to as the modified zero order regression in the context of a simple linear regression model where the method is consistent.

The Review of Economics and Statistics, October 2017, 99(4): 657–662

doi:10.1162/REST_a_00645

In line with common practice among empirical researchers, this paper focuses on models that do not allow covariate missingness to depend on the dependent variable. We revisit the linear imputation method and propose a generalized method of moments (GMM) procedure that incorporates the linear imputation model in the set of moment conditions. The GMM estimator yields efficiency gains relative to the complete data estimator for a subset, and sometimes all, of the parameters of interest. As with other linear imputation methods, the efficiency gains are obtained using the restriction that the imputation regression in the complete data is the same as that in the data with missing covariate values. In the GMM framework, this assumption is associated with a set of overidentifying restrictions that, as a by-product, allows a standard overidentification test of the validity of these restrictions. We show that the GMM estimator is at least as efficient as the two-step linear imputation estimators of Dagenais (1973) and Gourieroux and Monfort (1981). In comparing different approaches, we also examine the assumptions implicit in the dummy variable method. In addition to being potentially inconsistent even under an assumption of missing completely at random (MCAR; Jones, 1996), the dummy variable method can be less efficient than the complete data method even in cases when it is consistent.

The paper is structured as follows. Section II introduces the model and assumptions. The model includes a regression equation and a linear projection of the possibly missing covariate on the other covariates. Based on the missingness assumptions, a set of moment conditions for the observed data is developed, and an optimally weighted GMM estimator is proposed. Section III compares the GMM estimator to three approaches that empirical researchers commonly use: the complete data method, the dummy variable method, and the (non-GMM) linear imputation method. Section IV presents a brief Monte Carlo study. Section V concludes. Proofs of the theoretical results and more extensive Monte Carlo simulations are provided in the online technical appendix.

II. Model Assumptions, Moment Conditions, and GMM

Consider the following standard linear regression model,

$$y_i = \alpha_0 x_i + z_i'\beta_0 + \varepsilon_i = w_i'\theta_0 + \varepsilon_i \quad i = 1, \ldots, n, \tag{1}$$

where x_i is a (possibly missing) scalar regressor, z_i is a K-vector of (never missing) regressors, and $w_i \equiv (x_i, z_i')'$. The first element of z_i is 1, so the model contains an intercept. The residual ε_i is assumed to satisfy the conditions for equation (1) to be a linear projection, specifically

$$E(x_i\varepsilon_i) = 0 \text{ and } E(z_i\varepsilon_i) = 0. \tag{2}$$

The variable m_i indicates whether x_i is missing for observational unit i:

$$m_i = \begin{cases} 1 \text{ if } x_i \text{ missing} \\ 0 \text{ if } x_i \text{ observed} \end{cases}.$$

We assume the existence of a linear projection of x_i onto z_i:

$$x_i = z_i'\gamma_0 + \xi_i \text{ where } E(z_i\xi_i) = 0. \tag{3}$$

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 equation (3) is unique and completely general in the sense that it does not place any restrictions on the joint distribution of (x_i, z_i'). Also, no homoskedasticity assumptions are imposed on ε_i or ξ_i, although the nature of the results under homoskedasticity is discussed below.

Observations with missing x_i are problematic since equation (1) cannot be used directly to construct moment conditions for estimating $\theta_0 \equiv (\alpha_0, \beta_0')'$. However, equations (1) and (3) imply

$$y_i = z_i'(\gamma_0\alpha_0 + \beta_0) + \varepsilon_i + \xi_i\alpha_0 \overset{def}{=} z_i'(\gamma_0\alpha_0 + \beta_0) + \eta_i. \tag{4}$$

The following assumption on the missingness mechanism is the basis of the GMM approach.

Assumption 1. (a) $E(m_i z_i \varepsilon_i) = 0$; (b) $E(m_i z_i \xi_i) = 0$; (c) $E(m_i x_i \varepsilon_i) = 0$.

Several remarks are in order. First, the complete data estimator, defined as

$$\hat{\theta}_C = \left(\sum_{i=1}^{n}(1 - m_i)w_i w_i'\right)^{-1} \sum_{i=1}^{n}(1 - m_i)w_i y_i, \tag{5}$$

also requires conditions a and c (but not b) of assumption 1 for consistency. Second, assumption 1 amounts to imposing the restrictions that both the regression model, equation (1), and the imputation model, equation (3), are the same for the missing data and complete data observations. Third, the conditions of assumption 1 are weaker than assuming that m_i is independent of the unobserved variables. For instance, the following conditional mean assumptions are sufficient for assumption 1 to hold:

$$E(\varepsilon_i|m_i, x_i, z_i) = 0 \text{ and } E(z_i\xi_i|m_i) = 0,$$

with the former sufficient for parts a and c and the latter sufficient for part b of assumption 1. Focusing on the latter condition, which underlies the possible efficiency gains for imputation methods, one would expect it to hold when the x-on-z projection model, equation (3), is true whether x is missing or not. As an example, in a wage regression with x being IQ score and z being education level, part b of assumption 1 allows the missingness of IQ to be related to education but restricts the linear prediction of IQ given education to be the same whether IQ is observed or not; that is, for a given education level, true IQ values should not be systematically different for IQ-missing and IQ-observed individuals. Whether this condition holds within a particular application should be considered carefully by the researcher.

Of course, the stronger assumption that m_i is statistically independent of $(x_i, z_i, \varepsilon_i, \xi_i)$ will imply the conditions in assumption 1. Such an assumption is generally known as missing completely at random (MCAR), and assumption 1 is clearly weaker than MCAR since m_i is allowed to depend on z_i in arbitrary ways. Assumption 1, however, is stronger than the missing at random (MAR) assumption that is commonly used in the statistics literature. The MAR assumption allows missingness to depend on the completely observed z_i as well as the dependent variable (and thus ε_i), but not the partially observed regressor x_i. Under this more general assumption, the complete data estimator and linear imputation approaches are generally inconsistent. Estimators that utilize propensity score weighting, such as Horvitz and Thompson (1952) and Robins et al. (1994), deliver consistency. Even the MAR assumptions, however, are not necessarily more general than the assumptions needed for the complete data method. To summarize, MAR does not allow (conditional) dependence of missingness on x_i (while the complete data method does), the complete data method does not allow (conditional) dependence of missingness on y_i (while MAR does), and the full assumption 1 does not allow (conditional) dependence of missingness on x_i or y_i.

In contrast to assumption 1, Dardanoni et al. (2011) allow the linear projection of x_i onto z_i to depend on m_i. Under their weaker assumptions, one is led back to an estimator that is identical to the complete data estimator. Thus, any efficiency gains over the complete data estimator arise from imposing the restriction that the linear projection, equation (3), is the same for missing data and complete data observations. As a diagnostic for the validity of this assumption, the GMM procedure that follows leads to a test of the overidentifying restrictions embodied in assumption 1.

To develop the GMM estimator, define the vector of moment functions based on equations (2), (3), and (4),

$$g_i(\alpha,\beta,\gamma) = \begin{pmatrix} (1-m_i)w_i(y_i - \alpha x_i - z_i'\beta) \\ (1-m_i)z_i(x_i - z_i'\gamma) \\ m_i z_i\left(y_i - z_i'(\gamma\alpha+\beta)\right) \end{pmatrix} = \begin{pmatrix} g_{1i}(\alpha,\beta,\gamma) \\ g_{2i}(\alpha,\beta,\gamma) \\ g_{3i}(\alpha,\beta,\gamma) \end{pmatrix}, \tag{6}$$

and note that under assumption 1,

$$E(g_i(\alpha_0,\beta_0,\gamma_0)) = 0.$$

Therefore, one has $3K+1$ moment conditions available to estimate the $2K+1$ vector of parameters $(\alpha_0, \beta_0, \gamma_0)$, yielding K overidentifying restrictions. It is easy to show that the just-identified GMM estimator that uses g_{1i} and either g_{2i} or g_{3i} (but not both) is equivalent to the complete data estimator. In this setup, then, it is the restriction on the linear projection discussed above, and given by assumption 1b, that allows efficiency gains.

The theory for the GMM estimator is standard, but it is instructive to present the asymptotic variance to enable comparisons to the complete data estimator. First, write the variance-covariance matrix of the moment function evaluated at the true values of the parameters,

$$\Omega = E(g_i(\alpha_0,\beta_0,\gamma_0)g_i(\alpha_0,\beta_0,\gamma_0)') = \begin{pmatrix} \Omega_{11} & \Omega_{12} & 0 \\ \Omega_{12}' & \Omega_{22} & 0 \\ 0 & 0 & \Omega_{33} \end{pmatrix}, \tag{7}$$

where[4]

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

The zero components in equation (7) follow from $m_i(1-m_i)=0$. The standard two-step optimally weighted GMM estimator $(\hat\alpha, \hat\beta, \hat\gamma)$ solves

$$\min_{\alpha,\beta,\gamma} \bar g(\alpha,\beta,\gamma)'\hat\Omega^{-1}\bar g(\alpha,\beta,\gamma), \tag{8}$$

where $\bar g(\alpha,\beta,\gamma) = n^{-1}\sum_{i=1}^n g_i(\alpha,\beta,\gamma)$ and $\hat\Omega$ is an estimated version of Ω using preliminary first-step consistent residual estimates and sample analogs:

$$\hat\Omega_{11} = \frac{1}{n}\sum_i (1-m_i)w_iw_i'\hat\varepsilon_i^2,\ \hat\Omega_{22} = \frac{1}{n}\sum_i (1-m_i)z_iz_i'\hat\xi_i^2,$$
$$\hat\Omega_{12} = \frac{1}{n}\sum_i (1-m_i)w_iz_i'\hat\varepsilon_i\hat\xi_i,\ \hat\Omega_{33} = \frac{1}{n}\sum_i m_iz_iz_i'\hat\eta_i^2.$$

For instance, $\hat\varepsilon_i$ and $\hat\xi_i$ can be estimated from the complete data regressions of y_i on w_i and x_i on z_i, respectively, and $\hat\eta_i$ from a regression of y_i on z_i for the $m_i=1$ subsample. Although this GMM estimator is nonlinear in its parameters and therefore requires numerical methods, our experience is that the optimization problem is very well behaved and can be easily implemented in Stata (Version 11 or later) and other econometrics packages.[5]

The other component in the variance-covariance matrix for the optimal GMM estimator is the gradient matrix corresponding to the moment functions, given here by

$$G = \begin{pmatrix} G_{11} & 0 \\ 0 & G_{22} \\ G_{31} & G_{32} \end{pmatrix}, \tag{9}$$

[4] In an earlier version of this paper, we had incorrectly claimed that $\Omega_{12}=0$. We thank Yu-Chin Hsu for pointing out to us that this is not implied by assumption 1. The fact that Ω_{12} may be nonzero is what allows us to show potential efficiency gains for estimating the coefficient on the missing regressor.

[5] A Stata do file that implements the procedure for a sample data set is available online at http://www.utexas.edu/cola/economics/faculty/ja8294#code-and-data.

where the components of the matrix are

$$G_{11} = -E((1-m_i)w_i w_i'), G_{22} = -E((1-m_i)z_i z_i'),$$
$$G_{31} = (-E(m_i z_i z_i'\gamma_0) - E(m_i z_i z_i')), G_{32} = -E(m_i z_i z_i'\alpha_0).$$

The first set of columns represents the expectation of the derivatives of the moment functions with respect to $(\alpha, \beta')'$, while the second set of columns is related to the derivatives with respect to γ. As shown in the technical appendix, G can be expressed entirely in terms of the parameters α_0 and γ_0, the two second-moment matrices,

$$\Gamma_c = E((1-m_i)z_i z_i') \text{ and } \Gamma_m = E(m_i z_i z_i'),$$

and the quantity $\sigma^2_{\xi c} \equiv E((1-m_i)\xi_i^2)$. For purposes of inference, the components of G are estimated by taking sample analogs evaluated at the GMM estimates.

Under standard regularity conditions, we have the following result:

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

$$n\bar{g}(\hat{\alpha}, \hat{\beta}, \hat{\gamma})'\hat{\Omega}^{-1}\bar{g}(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \xrightarrow{d} \chi^2(K). \quad (10)$$

The behavior of the objective function in equation (10) gives rise to the possibility of testing the overidentifying restrictions imposed by assumption 1. In practice, a researcher using this GMM approach should always report the p-value associated with the test based on equation (10), that is, the overidentification specification test of assumption 1.

III. Comparisons to Other Missing Covariate Approaches

A. Comparison to the Complete Data Method

Using the notation of section II, the asymptotic variance of the complete data estimator $\hat{\theta}_C$ is

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

Since $\hat{\theta}_C$ is equivalent to a GMM estimator that uses a subset of the moment conditions in equation (6), the GMM estimator $\hat{\theta}$ is clearly at least as efficient as $\hat{\theta}_C$. Examination of the efficiency gains for $\hat{\theta}$ requires additional notation. Specifically, define

$$\Omega_{\varepsilon\eta c} = E((1-m_i)z_i z_i'\varepsilon_i\eta_i)$$
$$\Lambda_{\varepsilon\eta\xi c} = E((1-m_i)z_i\varepsilon_i\eta_i\xi_i),$$

where the subscript c denotes an expectation for complete data observations.

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

Proposition 2. *Under assumption 1,*

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

where $A = \Lambda_{\varepsilon\eta\xi c}\left(\sigma^2_{\xi c}\right)^{-1}$, $B = \Omega_{\varepsilon\eta c}\Gamma_c^{-1} - \Lambda_{\varepsilon\eta\xi c}\left(\sigma^2_{\xi c}\right)^{-1}\gamma_0'$, and D is a positive-definite matrix defined in the appendix.

The difference corresponding to estimation of α_0 is given by

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

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

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

Thus, in contrast to the previous literature on linear imputation, proposition 2 implies that efficiency gains are possible for both α_0 and β_0 since assumption 1 does not imply either $\Lambda_{\varepsilon\eta\xi c} = 0$ or $B = 0$.

B. Comparison to the Dummy Variable Method

The dummy variable method is not necessarily consistent under assumption 1 or even under MCAR. Note that model (1) and linear imputation (4) imply

$$y_i = (1-m_i)\alpha_0 x_i + z_i'\beta_0 + m_i z_i'\gamma_0\alpha_0 + \varepsilon_i + m_i\xi_i\alpha_0. \quad (11)$$

Under assumption 1, equation (11) is a valid regression model with the residual $\varepsilon_i + m_i\xi_i\alpha_0$ orthogonal to the regressors. In fact, OLS on this model yields estimates that are identical to the complete data method. GMM's efficiency gains come from the additional orthogonality conditions from

$$(1-m_i)x_i = (1-m_i)z_i'\gamma + (1-m_i)\xi_i, \quad (12)$$

by assumption 1. Indeed, one can easily show that GMM based on orthogonality of regressors and residuals in equations (11) and (12) is equivalent to the GMM estimator of section II.

The dummy variable method imposes zero restrictions on some of the parameters in equation (11). To make this explicit, write $z_i = (1, z_{2i}')$ and $\gamma_0 = (\gamma_{10}, \gamma_{20}')$ so that equation (11) becomes

$$y_i = (1-m_i)x_i\alpha_0 + z_i'\beta_0 + m_i\gamma_{10}\alpha_0 + m_i z_{2i}'\gamma_{20}\alpha_0 + \varepsilon_i + m_i\xi_i\alpha_0. \quad (13)$$

The dummy variable method omits the regressors $m_i z_{2i}'$, running OLS on

$$y_i = (1-m_i)x_i\alpha_0 + z_i'\beta_0 + m_i\gamma_{10}\alpha_0 + error. \quad (14)$$

Since equation (13) is the correct model, the dummy variable method is subject to omitted-variables bias and inconsistency unless certain restrictions are satisfied, as shown by Jones (1996):

Proposition 3. *The OLS estimators of* (α_0, β_0)*, from the regression in equation (14), are biased and inconsistent unless (a)* $\alpha_0 = 0$ *or (b)* $\gamma_{20} = 0$.

Condition a requires that x_i is an irrelevant variable in model (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. Condition b requires that z_{2i} is not useful for predicting x_i. (If z_{2i} has no elements, then equation [1] is a simple linear regression model, and the dummy variable estimator is equivalent to the complete data estimator.)

Presumably, one reason that researchers use the dummy variable method is the desire to achieve efficiency gains through the use of a larger sample. To evaluate this motivation, it is easiest to assume that the conditions of proposition 3 hold and, further, that the residuals ε_i and ξ_i are homoskedastic. If $\alpha_0 = 0$, the dummy variable (restricted) estimator of equation (14) is at least as efficient as the complete data (unrestricted) estimator of equation (11). Of course, in this case, it would be best to drop x_i completely. If $\gamma_{20} = 0$ and $\alpha_0 \neq 0$, the residual in equation (11) is heteroskedastic even when ε_i and ξ_i are homoskedastic; since the residual depends on m_i, its variance differs for missing covariate and nonmissing observations. Then, arguments along the lines of Hansen (2015) can be used to show that the restricted least squares estimator of equation (13), with the valid restriction $\alpha_0\gamma_{20} = 0$, does not necessarily result in a more efficient estimator than the complete data (unrestricted) estimator of equation (13). As such, even when it does not suffer from inconsistency, the dummy variable method does not guarantee efficiency gains and therefore has little justification under assumption 1.

C. *Comparison to the Linear Imputation Methods*

One can also make some observations about the linear imputation methods in the context of equations (11) and (12). These estimators proceed in a sequential manner and first estimate by OLS, equation (12), and then plug the estimated $\hat{\gamma}$ into equation (11). Gourieroux and Monfort (1981) do this using a second-step OLS, which is computationally convenient. Note, however, that the resulting residual from this procedure is

$$\varepsilon_i + m_i\xi_i\alpha_0 + m_i z_i'(\gamma_0 - \hat{\gamma})\,\alpha_0.$$

Thus, unless $\alpha_0 = 0$, there is heteroskedasticity and also cross-residual nonzero covariances, meaning that OLS will be relatively inefficient even under homoskedasticity assumptions on ε_i and ξ_i. Indeed, one can also show that the estimator provides no efficiency gains for estimation of α_0 and also does not necessarily bring improvement in estimating β_0 compared to the complete data estimator. Thus, apart from computational convenience, there is no practical advantage in using this approach.

To deal with the heteroskedasticity and nonzero covariances in the residuals above, Dagenais (1973) proposed a feasible generalized least squares (FGLS) estimator under the assumption that the residuals ε_i and ξ_i are homoskedastic. This FGLS estimator is asymptotically equivalent to a version of the GMM estimator that uses a weight matrix that is optimal under homoskedasticity assumptions and an assumption that $\Omega_{12} = 0$. Under these conditions, FGLS (or GMM) delivers efficiency gains over the complete data method for estimation of β_0, but no efficiency gains for estimation of α_0. More generally (under heteroskedasticity or $\Omega_{12} \neq 0$), the FGLS estimator is relatively inefficient compared to the GMM estimator.

IV. Monte Carlo Simulations

In this section, we illustrate some of the theoretical results with Monte Carlo simulations. (Interested readers can refer to the online appendix and Abrevaya & Donald, 2014, for a more extensive set of Monte Carlo simulations.)[6] Here, we consider a simple setup with $K = 2$:

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

All regression and conditional variance coefficients are set at 1, and we assume that half of the x_i's are missing completely at random. For this data-generating process, one can show that $\Lambda_{\varepsilon\eta\xi c} \neq 0$ so that the theory predicts efficiency gains even for estimation of α_0. Results are presented for the complete data method, the Dagenais (FGLS) method, the dummy variable method, and the GMM method. For each method, the bias, variance, and overall MSE for the estimators of the parameters $(\alpha_0, \beta_1, \beta_2)$ are calculated. The results are reported in table 1 for 1,000 replications using a sample size of $n = 400$.

The dummy variable method exhibits considerable bias for estimation of all parameters, whereas the other methods are unbiased. The results show that the GMM method is most efficient for estimation of α_0 (an MSE that is roughly 90% of the MSE for either complete data or FGLS), while the MSE for the complete data and FGLS methods is comparable. For estimation of the β parameters, GMM achieves efficiency gains on the order of 5% to 10% relative to either the complete data method or the FGLS estimator. As predicted by

[6] This paper is available online at http://www.utexas.edu/cola/economics/faculty/ja8294#code-and-data.

TABLE 1.—MONTE CARLO SIMULATION RESULTS

Estimation Method	Parameter	Bias	n×Var	MSE
Complete case method	α_0	0.011	13.93	0.035
	β_1	−0.012	19.34	0.049
	β_2	−0.002	22.65	0.057
Dummy variable method	α_0	−0.194	13.94	0.073
	β_1	0.195	19.89	0.088
	β_2	0.612	14.63	0.411
Dagenais (FGLS)	α_0	0.011	13.93	0.035
	β_1	−0.010	17.78	0.045
	β_2	0.002	19.46	0.049
GMM	α_0	0.006	12.53	0.031
	β_1	−0.007	16.27	0.041
	β_2	−0.002	18.35	0.046

theory, the FGLS estimator provides efficiency gains relative to the complete data estimator. We briefly summarize some of the other simulation results contained within the online appendix and Abrevaya and Donald (2014). When conditional variances are made more heteroskedastic (e.g., using exponential functions rather than quadratics), there is a more pronounced gain in efficiency for estimation of α_0. When the conditional variance of ε depends only on x (i.e., $\theta_2 = 0$), then none of the methods provides an efficiency gain for estimation of α_0. (When $\theta_2 = 0$, one can show that $\Lambda_{\varepsilon\eta\xi c} = 0$.) For the dummy variable method, the variance is larger than the other methods for β_1 and smaller for β_2; it is noteworthy that this method is quite biased and results in the overall MSE being larger, and in some cases substantially so, than all the other methods. The only situation where the dummy method has good performance is when α_0 is small. In designs with homoskedasticity, the FGLS estimator performs very well, sometimes leading to MSE values that are just below the GMM estimator (despite their asymptotic equivalence). The unweighted (non-FGLS) two-step linear imputation estimator, on the other hand, sometimes does better than the complete data estimator but sometimes does worse; across all designs, this estimator was less efficient than FGLS or GMM.

V. Conclusion

This paper has considered several methods of dealing with missing values for an explanatory variable. A GMM procedure was proposed using the basic assumptions that underlie existing linear imputation methods. Under these assumptions, the GMM estimate is more efficient than the complete data method, the dummy variable method, or the linear imputation methods. The GMM estimator also leads to a natural overidentification test of the assumptions. The computationally convenient unweighted linear imputation method and the commonly used dummy method were found to potentially provide a "cure that is worse than the disease." As noted in section I, the GMM approach does not give rise to semiparametric efficiency in the sense of Robins et al. (1994), although the proposed approach is straightforward to use by empirical researchers. We should also note that we have intentionally simplified our theoretical analysis by focusing on the case of a single missing covariate (see Muris, 2013, or Chaudhuri & Guilkey, 2016, for more on this issue.)

REFERENCES

Abrevaya, J., and S. G. Donald, "A GMM Approach for Dealing with Missing Data on Regressors and Instruments," University of Texas at Austin working paper (2014).

Chaudhuri, S., and D. Guilkey, "GMM with Multiple Missing Variables," *Journal of Applied Econometrics* 31 (2016), 678–706.

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

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

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

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

Greene, W. H., *Econometric Analysis*, 5th ed. (Upper Saddle River, NJ: Prentice Hall, 2003).

Hansen, B. E., *Econometrics* (2015). http://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf.

Horvitz, D. G., and D. J. Thompson, "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association* 47 (1952), 663–685.

Jones, M. P., "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression," *Journal of the American Statistical Association* 91 (1996), 222–230.

Muris, C., "Efficient GMM Estimation with a General Missing Data Pattern," Simon Fraser University mimeograph (2013).

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