Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Missing data, imputation, and endogeneity

Ian K. McDonough [a], Daniel L. Millimet [b,c,*]

[a] University of Nevada, Las Vegas, United States
[b] Southern Methodist University, United States
[c] IZA, Germany

ARTICLE INFO

Article history:
Available online xxxx

JEL classification:
C36
C51
J13

Keywords:
Imputation
Missing data
Instrumental variables
Birth weight
Childhood development

ABSTRACT

Bassmann (1957, 1959) introduced two-stage least squares (2SLS). In subsequent work, Basmann et al. (1971) investigated its finite sample performance. Here we build on this tradition focusing on the issue of 2SLS estimation of a structural model when data on the endogenous covariate is missing for some observations. Many such imputation techniques have been proposed in the literature. However, there is little guidance available for choosing among existing techniques, particularly when the covariate being imputed is endogenous. Moreover, because the finite sample bias of 2SLS is not monotonically decreasing in the degree of measurement accuracy, the most accurate imputation method is not necessarily the method that minimizes the bias of 2SLS. Instead, we explore imputation methods designed to increase the first-stage strength of the instrument(s), even if such methods entail lower imputation accuracy. We do so via simulations as well as with an application related to the medium-run effects of birth weight.

1. Introduction

Basmann (1957) introduces Two-Stage Least Squares (2SLS) as a means of estimating structural models that suffer from endogeneity when exclusion restrictions are available. In particular, the estimator allows one to take advantage of having more instrumental variables than endogenous regressors, in which case researchers are able to conduct tests of overidentifying restrictions (Sargan, 1958; Basmann, 1960; Hansen, 1982). In subsequent work, Basmann et al. (1971) investigate the finite sample performance of the 2SLS estimator. Because of this research, and the future research it spurred (e.g., Stock et al., 2002; Flores-Lagunes, 2007), the properties of 2SLS are well understood in many settings. However, one setting that has been inadequately addressed to date pertains to 2SLS estimation of a structural model when data on the endogenous covariate(s) are missing for some observations.[1]

Dealing with missing data is a frequent challenge confronted by empirical researchers. Ibrahim et al. (2005) note that medical researchers analyzing clinical trials often face the problem of missing data for various reasons, including survey nonresponse, loss of data, human error, and failing to meet protocol standards in follow up visits. Burton and Altman (2004), reviewing 100 articles across seven cancer journals, found that 81 of the 100 articles involve analyses with missing covariate data. Empirical researchers in economics face similar challenges. Abrevaya and Donald (2013), surveying four of the top empirical economics journals over a recent three-year period (2006–2008), find that nearly 40% of papers inspected had to confront missing data.[2]

Given the pervasive nature of missing data in empirical research, the literature on handling missing data is vast. Unfortunately, the literature tends to ignore the distinction between exogenous and endogenous covariates (i.e., whether the covariate is endogenous in the absence of missing data). As we discuss below, this distinction is likely to be salient as the 'optimal' method for dealing with missing data on an exogenous covariate may not be 'optimal' for an endogenous covariate. Specifically, the finite sample performance of various approaches for dealing with a missing covariate may differ when the resulting model is estimated via 2SLS as opposed to Ordinary Least Squares (OLS). This is the subject we investigate here.

Methods for dealing with (exogenous) missing covariates can be divided into two broad categories: *ad hoc* approaches and *imputation* approaches. The most widely used methods for dealing with missing covariate data are considered ad hoc by many researchers despite their popularity. These ad hoc approaches include so-called

* Correspondence to: Department of Economics, Box 0496, Southern Methodist University, Dallas, TX 75275-0496, United States; Fax: +214 768-1821.
E-mail address: millimet@smu.edu (D.L. Millimet).

[1] In complementary work, Feng (2016) consider the problem of missing data on the instrument for some observations.

[2] The journals inspected in Abrevaya and Donald (2013) inlcude *American Economic Review*, *Journal of Human Resources*, *Journal of Labor Economics*, and *Quarterly Journal of Economics*. See Table 1 in Abrevaya and Donald (2013) for more details.

http://dx.doi.org/10.1016/j.jeconom.2017.05.006
0304-4076/

complete case analysis and variations on missing-indicator methods (Schafer and Graham, 2002; Burton and Altman, 2004; Dardanoni et al., 2011; Abrevaya and Donald, 2013). Popular imputation approaches include regression (conditional mean) imputation and variants of nearest neighbor matching (Allison, 2001; Rosenbaum, 2002; Mittinty and Chacko, 2005). Multiple imputation methods, with the advancement of computational power, have also become more widely used in empirical research (Rubin, 1987).

Complete case analysis, as the name suggests, uses only observations without missing data. With this approach, efficiency losses can be substantial and bias may be introduced depending on the nature of the missingness (Pigott, 2001; Schafer and Graham, 2002; Horton and Kleinman, 2007). The missing-indicator method, in the context of continuous variables, entails creation of a binary indicator of missingness and replacement of the missing values with some common value. The created indicator variable and covariate imputed with some common value (usually the mean) are included, along with their interaction, in the estimating equation. With missing categorical variables, an indicator for a 'missing' category is added to the model. Although widely used and convenient, this method has been severely criticized (Jones, 1996; Schafer and Graham, 2002; Dardanoni et al., 2011).

Imputation approaches augment the original estimating equation with an imputation model in order to predict values of the missing data. Once the missing data are replaced with their predicted values, the original model is estimated using the full sample. Regression imputation obtains predicted values for the missing data by utilizing data on observations with complete data to obtain an estimated regression function with the covariate containing missing values as the dependent variable. The estimated regression function is then used to impute missing values with the predicted conditional mean. Nearest neighbor matching is done by replacing missing data with the values from observations with complete data deemed to be 'closest' according to some metric. Common univariate distance metrics include the Mahalanobis measure or the absolute difference in propensity scores, where the propensity score is the predicted probability that an observation has missing data (Mittinty and Chacko, 2005; Gimenez-Nadal and Molina, 2016). Matching methods are a variant of so-called hot deck imputation where the 'deck' in this case is just a single nearest neighbor (Andridge and Little, 2010). Multiple imputation methods specify *multiple* (M, where $M > 1$) imputation models, rather than just a single imputation model. As such, M complete data sets are obtained by imputing the missing values M times. Common methods for imputing the M data sets are extensions of the regression and nearest neighbor matching methods described above. Using each of the imputed data sets, the analysis of interest is carried out M times with the M estimates being combined into a single result.

Despite this robust literature on missing data methods, there is a lack of guidance for applied researchers in dealing with missingness in endogenous covariates. As stated in Schafer and Graham (2002, p. 149), the goal of a statistical procedure is to make "valid and efficient inferences about a population of interest" irrespective of whether any data are missing. In our case, the statistical procedure is 2SLS and we wish to make inferences about some population parameter(s), θ . As such, any treatment of missing data should be evaluated in terms of the properties of the resulting estimate of θ, $\widehat{\theta}$. It is well known that the finite sample properties of 2SLS are complex even in the absence of missing data. Complete case analysis may introduce additional complexities due to nonrandom selection depending on the nature of the missingness. The missing-indicator approach introduces an additional endogenous covariate (due to the interaction term between the missingness indicator and the endogenous covariate), as well as measurement error in the already endogenous covariate due to the replacement of the missing data with an arbitrary value. Finally, any imputation procedure almost surely introduces measurement error in the endogenous covariate. Thus, understanding the implications of handling missing data in the specific context of 2SLS seems necessary. In the context of imputation, this point is made even more salient since the finite sample bias of 2SLS is not monotonically decreasing in the degree of measurement, or imputation, accuracy (Millimet, 2015). Furthermore, the finite sample bias depends on the strength of the instruments, which may be impacted by the imputation method. As such, and perhaps counter to intuition, the most accurate imputation method may not be the method that minimizes the finite sample bias of 2SLS.

In light of this, we investigate the finite sample performance of several approaches to dealing with missing covariate data when the covariate is endogenous even in the absence of any missingness. Specifically, we focus on imputation approaches and discuss the finite sample properties of OLS and 2SLS when one imputes an endogenous covariate prior to estimation. Then, we assess the finite sample performance of various imputation approaches in a Monte Carlo study. For comparison, we also examine the performance of the complete case and missing-indicator approaches. Finally, we illustrate the different approaches with an application to the causal effect of birth weight on the cognitive development of children in low-income households using data from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-2011 (ECLS-K:2011). In our sample, birth weight is missing for roughly 16% of children. Moreover, because birth weight is likely to be endogenous, we utilize instruments based on state-level regulations that affect participation in the Supplemental Nutrition Assistance Program (SNAP) similar to Meyerhoefer and Pylypchuk (2008). SNAP (formerly known as the Food Stamp Program) has been shown to affect the health of low-income pregnant women and, hence, affect pregnancy outcomes (Baum, 2012).

The Monte Carlo results suggest that imputation methods that incorporate the instruments along with other exogenous covariates generally produce the smallest finite sample bias of the 2SLS estimator. This is attributable, at least in part, to the improved instrument strength in the resulting first-stage estimation, as well as the improved imputation accuracy since the endogenous covariate is a function of the instruments (assuming they are valid). Among the ad hoc approaches, the complete case approach often does surprisingly well, while the missing-indicator approach does not. In terms of our application, however, we find surprisingly little substantive difference across the various estimators in terms of the point estimates, although the estimators that incorporate the instruments into the imputation model do lead to better instrument strength. Nonetheless, we do find some statistically and economically significant evidence that birth weight has an impact on math achievement at the beginning of kindergarten. This result is driven entirely by non-white male children.

The remainder of the paper is organized as follows. Section 2 sets up the structural model and discusses different methods for handling missing covariate data. Section 3 describes the Monte Carlo Study. Section 4 contains the application. Finally, Section 5 concludes.

2. Model

2.1. Setup

We consider the following structural model

$$y = x_1\beta_1 + \beta_2 x_2^* + \varepsilon \tag{1}$$

$$x_2^* = x_1\pi_1 + z\pi_2 + \eta \tag{2}$$

where y is a $N \times 1$ vector of an outcome variable, x_1 is a $N \times K$ matrix of exogenous covariates with the first element equal to one, x_2^* is a $N \times 1$ continuous endogenous covariate vector, β_1 is a

Please cite this article in press as: McDonough, I.K., Millimet, D.L., Missing data, imputation, and endogeneity. Journal of Econometrics (2017), http://dx.doi.org/10.1016/j.jeconom.2017.05.006.