

Chapter 2

Program Impact Estimation with Binary Outcome Variables: Monte Carlo Results for Alternative Estimators and Empirical Examples

David K. Guilkey and Peter M. Lance

2.1 Introduction

A common problem in program evaluation is measuring the impact of a binary program indicator on a binary outcome variable. For example, one of the most frequently used methods to promote contraceptive use in less developed countries is multi-media campaigns. Evaluation of such programs is complicated by the fact that, except in a very few cases, an experimental design is not used (Bauman et al. 1993; Mwaikambo et al. 2011) and the program implementers have little control over who is exposed to the campaign. The typical method that has been used to evaluate such programs relies on a cross sectional design where respondents are asked yes/no questions about program exposure and contraceptive use along with questions that solicit information about various other characteristics of the respondents that can serve as control variables in a multivariate analysis. In a systematic review of family planning interventions, Mwaikambo et al. (2011) found that two thirds of the 63 family planning interventions that were evaluated in the published literature between 1995 and 2005 involved this type of demand side intervention, although not all of them only considered binary outcomes.

Statistical methods used to measure program impact with this type of data have ranged from those that ignore the potential endogeneity of program recall, such as simple logit or probit regression (see Mwaikambo et al. 2011; Hutchinson and Wheeler 2006 for reviews) and propensity score matching (Babalola 2005), to

D.K. Guilkey (✉)

Department of Economics and the Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA

e-mail: dguilkey@email.unc.edu

P.M. Lance

Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516-3997, USA

e-mail: pmlance@email.unc.edu

estimators that correct for endogenous recall using linear or non-linear instrumental variables methods or some type of full information maximum likelihood method (Guilkey et al. 2006; Chen and Guilkey 2003; Guilkey and Hutchinson 2011). On the surface, it would seem that simple methods that do not correct for the potential endogeneity of program recall should inherently perform worse than those that do. However, it is possible to make a case for these simple approaches since methods that explicitly correct for endogeneity rely on the presence of valid exclusion restrictions – variables that affect program recall directly but only affect contraceptive use indirectly through the recall variable or, in some cases and as a last resort, the nonlinearity provided by parametric assumptions.

Unfortunately, there are typically few variables that are candidates for exclusion from the contraceptive use equation, and this is not a unique complication to the multimedia campaign impact evaluation literature: the paucity of potential credible, strong instruments is a widespread challenge in many other applications with different outcomes and endogenous regressors of interest but a similar behavioral structure. Instrumental variables methods as well as more complicated strategies such as full information maximum likelihood estimation can yield highly unstable results in the face of weak instruments. On the other hand, simple methods, even when inconsistent, could lead to results that capture more reliably true program effects (Bollen et al. 1995). In addition, some of the single and systems of equations estimators rely on the assumption of normally distributed error terms and there is evidence that when that assumption is violated, estimated impacts can be far from the truth (Mroz 1999; Chiburis et al. 2011).

The purpose of this paper is to provide the most comprehensive analysis to date of the finite sample performance of alternative methods to estimate program impact when both the treatment and the outcome variables are binary. We focus primarily on methods that can be implemented in STATA, a widely available statistical package, but we also evaluate a semi-parametric instrumental variables random effects model that is not available in STATA.¹ Much of the work to date has focused on a model in which either the treatment variable or the outcome variable is continuous while the other is binary (Guilkey et al. 1992; Bollen et al. 1995; Mroz 1999). Chiburis et al. (2011) do examine the finite sample performance of the bivariate probit estimator and several linear estimators for our case of interest; however, they focus on a model that is exactly identified case for linear models, which does not allow for the use of tests that require the model, at least in theory, to be overidentified. Further, they do not evaluate the wide range of estimators that are used in this setting, including semi-parametric models that are potentially robust to departures from normality. Our Monte Carlo data generation process is designed to mimic the type of data that has been used to evaluate the impact of program recall on contraceptive use in a developing country and we provide examples of the methods using data from Bangladesh and Tanzania. However, the methods have wide applicability beyond our specific examples given how often the basic behavioral structure behind them

¹The authors are currently writing STATA commands to implement this estimator.

appears in applied work. In this manuscript we restrict attention to the constant effect case, limiting somewhat the applicability of our findings to instances where, for instance, Local Average Treatment Effects are a concern.

This paper is organized as follows. In the next section, we lay out the statistical model and provide details on the alternative estimation and testing procedures that are evaluated. In Sect. 2.3, we detail the data generating process for the Monte Carlo experiment and the results of the experiment are presented in Sect. 2.4. Section 2.5 presents the empirical example and Sect. 2.6 concludes.

2.2 Model and Estimation Methods

We are concerned with a model of the following form:

$$Y_{i1}^* = X_i' \beta_1 + Z_i' \alpha + \epsilon_{i1} \quad (2.1)$$

$$Y_{i2}^* = X_i' \beta_2 + Y_{i1} \delta + \epsilon_{i2} \quad (2.2)$$

where there are $i = 1, 2, \dots, N$ observations and the dependent variables are latent variables. The observed dependent variables are binary indicators: $Y_{ij} = 1$ if $Y_{ij}^* > 0$ and $Y_{ij} = 0$ otherwise for $j = 1, 2$. X_i is a $k_X \times 1$ vector that represents variables that appear in both Eqs. (2.1) and (2.2) while Z_i is a $k_Z \times 1$ vector that represents a set of variables that are excluded from Eq. (2.2). The coefficients in the model are column vectors of appropriate dimension.

In our model, the observed binary indicator, Y_{i1} , is the right-hand-side endogenous explanatory variable, as opposed to the latent variable. It is well known for this case that there exist estimators that are technically identified without exclusion restrictions (α could be zero) due to functional form. However, the case that we are interested in this paper is the one in which there are at least two valid exclusion restrictions and so even the linear instrumental variables model would be over-identified. Our primary interest is the outcome in Eq. (2.2) with Eq. (2.1) specifying an endogenous treatment.

Several of the estimation methods that we compare assume that $[\epsilon_{i1}, \epsilon_{i2}]$ follows a bivariate normal distribution. To keep the notation simple, in this manuscript we capture this by assuming that $\text{var}(\epsilon_{ij}) = 1$ for $j = 1, 2$ and all i and that $E(\epsilon_{i1}, \epsilon_{i2}) = \rho$. The normalization that the error variances equal 1 means that the parameter estimates are only estimated to scale, as is common when the dependent variable is a binary indicator. However, the scale of the estimated parameters is of little concern in this paper since the most important basis of comparisons will be how well the various estimators approximate the population average treatment effect (ATE) defined as:

$$ATE = E(Y_2 | Y_1 = 1) - E(Y_2 | Y_1 = 0) \quad (2.3)$$

We now turn to a brief discussion of the estimators we consider in this manuscript.

2.2.1 Linear Probability Model (LPM)

Simple ordinary least squares estimation of Eq. (2.2) ignores the endogeneity of Y_{i1} and the binary nature of the dependent variable Y_{i2} . In this case, the estimated ATE is simply the estimate of δ and it will be a consistent estimator only if $E(Y_{i1}\epsilon_{i2}) = 0$.

2.2.2 Probit

From the class of single equation estimators for Eq. (2.2) that ignore the endogeneity of Y_{i1} , we also consider estimation of Eq. (2.2) by simple probit regression and then note that:

$$\hat{P}(Y_{i2} = 1) = \Phi(X_i\hat{\beta}_2 + Y_{i1}\hat{\delta}) \quad (2.4)$$

where $\Phi(\cdot)$ is the cumulative normal distribution function. We can now use (2.4) to obtain an estimate of the ATE:

$$\widehat{ATE} = \frac{1}{N} \sum_{i=1}^N \hat{P}(Y_{i2} = 1|Y_{i1} = 1) - \sum_{i=1}^N \hat{P}(Y_{i2} = 1|Y_{i1} = 0) \quad (2.5)$$

This will be a consistent estimator under the same conditions as presented for the OLS estimator.

2.2.3 Instrumental Variables

We compare three variants of linear instrumental variables: two-stage least squares (TSLS), limited information maximum likelihood (LIML) and generalized method of moments (GMM). In all cases, we use the default options in STATA for estimation per the `-ivreg-` command. We consider all three because they offer different estimation approaches within the context of linear instrumental variables and allow for different tests for endogeneity and identification. Tests for endogeneity are based on the Wu-Hausman (Wu 1974; Hausman 1978) and Durbin (1954) tests for TSLS, the standard Hausman test (Hausman 1978) for LIML, and a test referred to as the C statistic for GMM (Hayashi 2000). The identification tests considered for these estimation methods are: Sargon's test (Sargon 1958) for TSLS; Basman's test (Bassman 1960) for TSLS (specifically, Basman's χ^2 test) and LIML (Basman's F test); the Anderson-Rubin test (Anderson and Rubin 1950) for LIML; and Hansen's test (Hansen 1982) for GMM. Details regarding all tests can be found in the STATA reference manual and the cited references.

For all three estimators, we use the estimated δ as the estimate of the ATE. In general in linear instrumental variables models, what is actually estimated is a local average treatment effect (LATE) (Imbens and Angrist 1994; see Angrist and Pischke 2009 for an excellent and succinct review). However, the design of our experiment precludes the possibility of LATE, though it may be at play in the results from the two applied examples considered in this manuscript.

2.2.4 Linear Predictor and Residual Models

Terza et al. (2008) discuss two basic approaches commonly applied in the face of an endogenous regressor in a non-linear equation of interest and a possibly non-linear first stage for that endogenous regressor: first stage predictor substitution (which is essentially just the extension of linear two-stage least squares estimation to the nonlinear setting) and residual inclusion. The predictor substitution strategy is inconsistent whereas under very general conditions the residual inclusion strategy is consistent (Terza et al. 2008). Previous work has suggested that, in the setting of a second-stage binary dependent variable of interest and endogenous continuous regressor, residual inclusion should be consistent provided that the distribution of the unobservable determinants of the binary outcome and continuous endogenous regressor is jointly normal (Rivers and Vuong 1988; Bollen et al. 1995).

We consider two versions of the residual inclusion approach as adapted to the structure defined by the behavioral model in (2.1) and (2.2).² First, for the most obvious potential extension of Terza et al. (2008), Rivers and Vuong (1988) and Bollen et al. (1995) to the present setting, we estimate (2.1) by ordinary least squares (i.e. the linear probability model) and generate predicted residuals that are then included in probit regression of (2.2). In the results tables we refer to this estimator as Residual1. Second, we estimate (2.1) by probit and then calculate the generalized residuals using the following formula (Gourieroux et al. 1987):

$$\frac{(Y_{i1} - X_i' \beta_1 - Z_i' \alpha) \phi(Y_{i1} - X_i' \beta_1 - Z_i' \alpha)}{\Phi(Y_{i1} - X_i' \beta_1 - Z_i' \alpha) (1 - \Phi(Y_{i1} - X_i' \beta_1 - Z_i' \alpha))} \quad (2.6)$$

where $\Phi(\cdot)$ is the cumulative normal distribution function and $\phi(\cdot)$ is the normal density function. These residuals are then included in probit regression of (2.2). In the tables and text we refer to this estimator as Residual2.

²We did consider predictor substitution schemes as well but, as expected, they performed poorly and we do not include them in the comparisons.

2.2.5 *Bivariate Probit (BIPROBIT)*

Bivariate probit jointly estimates Eqs. (2.1) and (2.2) by maximum likelihood methods assuming bivariate normality for the error terms. The `-biprobit-` routine in STATA relies on standard Newton-Raphson estimation using a conventional approximation of the bivariate normal cumulative distribution function based on quadrature. We also considered including the `-mvprobit-` routine, which is not part of the basic STATA package but available as a user-written program (i.e., an .ado file). This routine is designed to allow for more than two binary outcome equations and uses Geweke-Hajivassilou smooth recursive conditioning simulator to approximate the bivariate cumulative normal density (see [Cappellari and Jenkins 2003](#)). In preliminary runs, we found that we needed to use far more than the default number of draws (five) in order to obtain accurate parameter estimates and so we dropped this estimator from consideration.

After the model is estimated, the treatment effect is calculated from the marginal probability distribution for the second outcome – using Eqs. (2.3) and (2.4) but with estimated coefficients obtained from the full information maximum likelihood estimator. An endogeneity test is simply a direct test of the null hypothesis that the error correlation across the two equations is zero. We also report an overidentification test that exploits the fact that this model is identified without exclusion restrictions by including the instruments as explanatory variables in Eq. (2.2) (adding the Z variables) and then performing a likelihood ratio test of the null hypothesis that the coefficients are jointly zero. Support for the null implies that these variables are in fact properly excluded.

2.2.6 *Semi-parametric Maximum Likelihood Estimation (DFM)*

We consider a version of a semi-parametric estimator based on [Heckman and Singer \(1984\)](#) but using a non-linear extension proposed by [Mroz \(1999\)](#). To set up the likelihood function for this model, we adopt an error components approach to the unobservables and re-write Eqs. (2.1) and (2.2) as follows:

$$Y_{i1}^* = X_i' \beta_1 + Z_i' \alpha + \mu_{i1} + \epsilon_{i1}^* \quad (2.7)$$

$$Y_{i2}^* = X_i' \beta_2 + Y_{i1} \delta + \mu_{i2} + \epsilon_{i2}^* \quad (2.8)$$

where the correlation in the error terms is between the μ 's and $E(\epsilon_{i1}^*, \epsilon_{i2}^*) = 0$. The approach that we use for this estimator is based on the type-I Extreme Value distribution for the ϵ 's (leading to the logit model) instead of the normal distribution. However, the basis of comparison is the ATE as defined in Eq. (2.3) and not the estimated coefficients (which are well known to be different by a scale factor from

corresponding probit coefficients). Hence, the shift from the cumulative normal distribution to the logistic function still allows a simple comparison of results. We can then write:

$$P(Y_{i1}|\mu_{i1}) = \frac{e^{(X_i' \beta_1 + Z_i' \alpha + \mu_{i1})}}{1 + e^{(X_i' \beta_1 + Z_i' \alpha + \mu_{i1})}} \quad (2.9)$$

$$P(Y_{i2}|\mu_{i2}) = \frac{e^{(X_i' \beta_2 + Y_{i1} \delta + \mu_{i2})}}{1 + e^{(X_i' \beta_2 + Y_{i1} \delta + \mu_{i2})}} \quad (2.10)$$

The contribution to the likelihood function for observation i , conditional on the μ 's is:

$$\begin{aligned} L_i(\mu_{i1}, \mu_{i2}) &= [P(Y_{i1} = 1|\mu_{i1}) P(Y_{i2} = 1|\mu_{i2})]^{Y_{i1} Y_{i2}} \\ &\quad [P(Y_{i1} = 0|\mu_{i1}) P(Y_{i2} = 0|\mu_{i2})]^{(1-Y_{i1})(1-Y_{i2})} \\ &\quad [P(Y_{i1} = 1|\mu_{i1}) P(Y_{i2} = 0|\mu_{i2})]^{Y_{i1}(1-Y_{i2})} \\ &\quad [P(Y_{i1} = 0|\mu_{i1}) P(Y_{i2} = 1|\mu_{i2})]^{(1-Y_{i1})Y_{i2}} \end{aligned}$$

We assume that the distributions of the μ 's can be approximated by a step function with J steps for each of the μ 's and probability weights (w_j for $j = 1, 2, \dots, J$) that sum to one for the J steps. The unconditional contribution to the likelihood function for observation i can then be written:

$$L_i = \sum_{j=1}^J w_j L_i(\mu_{i1}, \mu_{i2}) \quad (2.11)$$

The likelihood function is simply the product of (2.11) over the N observations. In addition to the model's coefficients, one searches over $J - 1$ weights (since they sum to one) and $J - 1$ sets of the μ 's (since one of the μ 's must be set to zero if there is a constant term in the model). We call this the “discrete factor model” (and, for the sake of brevity, frequently refer to it as the ‘DFM’ in discussions below); (see [Mroz \(1999\)](#) for additional details). The estimated ATE can be obtained using Eq. (2.10) where the population parameters are replaced with estimates including the estimates for the weights and mass points (the μ 's).

In practice, one would add points of support to the heterogeneity distribution until there is no significant improvement in the likelihood function. However, this is not practical in a Monte Carlo experiment and so we simply set the number of points of support for the discrete distribution to four.

2.3 Data Generating Process

The basic logic behind the data generating process is straightforward: within each Monte Carlo experiment data are generated in a fashion that insures that the resulting estimation samples conform to the behavioral parameters of that experiment. Most of these behavioral parameters vary across Monte Carlo experiments (one was fixed across them). It is this variation in these parameters that allows examination of the comparative performance under alternative circumstances of the estimators considered in this study. The behavioral parameters that vary across experiments include the true (i.e. established by the design of the experiment): program effect ($E(Y_2|X, Y_1 = 1) - E(Y_2|X, Y_1 = 0)$); correlation of the errors $\{\epsilon_1, \epsilon_2\}$; average of the program outcome (Y_1) within the sample³; average of the outcome of interest (Y_2) within the sample; first stage strength of the instruments Z to explain Y_1 (as reflected in the χ^2 statistic emerging from a test of the joint significance of those instruments); and the bivariate error type (i.e. normal or non-normal errors).

In each experiment, the first step is to draw pseudo-randomly a sample of size N for the exogenous variables X , Z and ϵ (given the error correlation and type specified for that experiment). Given the draws from X and Z and this initial draw from ϵ , we then determine values for the system parameters β_1 , β_2 , α and δ from Eqs. (2.1) and (2.2) that insure that data generated conditional on those values for the system parameters and X and Z would conform to the remaining behavioral parameters. The experiment itself then involved replications (1,000 replications in the case of experiments involving 1,000 or 5,000 observations and 500 replications in the case of experiments involving 10,000 observations). In each, a new pseudo-random draw was made from the distribution of the error terms ϵ for each of the N observations and, conditional on that new draw, the draw from X and Z and the values for β_1 , β_2 , α and δ determined in the first step, new values for Y_1 and Y_2 were calculated for each observation. The performance of the various estimators considered in this manuscript was then recorded given “observed” data Y_1 , Y_2 , X and Z .

2.3.1 Sample Sizes and Behavioral Parameter Values

Our various Monte Carlo experiments are distinguished by the values of the behavioral parameters set for them, as well as the sample sizes involved. We consider many alternative combinations of these sample sizes and behavioral parameters. To begin with, three basic sample sizes are considered: 1,000, 5,000 and 10,000. These were selected based on a rough sense of the sort of ranges of sample sizes frequently encountered when estimating systems along the lines of Eqs. (2.1) and (2.2) using real world data.

³That is, the program enrollment prevalence within the sample.

For program participation prevalence and outcome prevalence we consider values of 0.5 and 0.25. These capture the cases of programs for which participation is comparatively common and less common, and outcomes of interest for which the same can be said.

For program impact (the true marginal effect of Y_1 on the probability of Y_2) we consider high (0.2) and modest (0.05) impact cases. The program impact levels reflect constant (as opposed to varying with observed or unobserved heterogeneity) effects.

The error terms ϵ are based on two basic bivariate distributions:

1. A bivariate standard normal distribution;
2. A non-normal distribution with a skewness of 1.5 and an excess kurtosis of 3.

The algorithm for drawing the non-normal errors is based on the method proposed by [Vale and Maurelli \(1983\)](#).⁴ The [Vale and Maurelli \(1983\)](#) approach involves a combination of [Fleishman's \(1978\)](#) procedure for generating non-normal random variables with a matrix decomposition method typically applied to the task of generating multivariate normal random variables ([Kaiser and Dickman 1962](#)). Two levels of error correlation are employed for these bivariate distributions: 0.1 and 0.3, allowing different degrees of endogeneity. Finally, we vary the first stage (Eq. (2.1)) explanatory power of the instruments as manifested by a χ^2 statistic resulting from a test of the joint significance of those instruments based on a probit regression of Y_1 on X and Z . We cover test statistic values of 15, 25, and 50, encompassing a range of instrument strength levels.

Overall explanatory power of Eqs. (2.1) and (2.2), as captured by the R^2 from ordinary least squares regression estimation of them, is fixed at 0.3 in both cases in order to reflect a degree of explanatory power more realistic to regression analyses using micro-level samples. This typically results in pseudo- R^2 values in the 0.15–0.25 range.

2.3.2 Drawing X and Z

The exogenous explanatory variables X and Z are pseudo-randomly drawn from the standard normal distribution. In this manuscript, four exogenous characteristics X (X_1 , X_2 , X_3 and X_4) and two instruments Z (Z_1 and Z_2) are drawn for each Monte Carlo experiment. Thus, in the terms of the discussion introducing Eqs. (2.1) and (2.2) in Sect. 2.2, $k = 4$ and $k_z = 2$.

⁴We are grateful to Stas Kolenikov for generously sharing a STATA .ado file that he wrote implementing that [Vale and Maurelli \(1983\)](#) procedure.

2.3.3 *The Mechanics of the Data Generating Process*

Each Monte Carlo experiment could be characterized by these behavioral parameters as applied to the system of equations (2.1) and (2.2). To begin with, the Monte Carlo experiments revolve around the latent variable equations

$$Y_{i1}^* = X_i' \beta_1 + Z_i' \alpha + \phi_1 \epsilon_{i1} \quad (2.12)$$

$$Y_{i2}^* = X_i' \beta_2 + Y_{i1} \delta + \phi_2 \epsilon_{i2} \quad (2.13)$$

which differ from (2.1) and (2.2) primarily by the coefficients ϕ on the error terms ϵ . (As will be seen below, these coefficients are placed on the errors to support the target R^2 of 0.3 in each equation.) Given the dimensionality of X and Z employed in this study, (2.12) and (2.13) are, effectively,

$$Y_{i1}^* = \beta_{10} + X_{1i} \beta_{11} + X_{2i} \beta_{12} + X_{3i} \beta_{13} + X_{4i} \beta_{14} + Z_{1i} \alpha_1 + Z_{2i} \alpha_2 + \phi_1 \epsilon_{i1} \quad (2.14)$$

$$Y_{i2}^* = \beta_{20} + X_{1i} \beta_{21} + X_{2i} \beta_{22} + X_{3i} \beta_{23} + X_{4i} \beta_{24} + Y_{i1} \delta + \phi_2 \epsilon_{i2} \quad (2.15)$$

These equations are used to generate the variables Y used for each experiment. To do this, specific values need to be assigned to the β 's, α 's, δ and the ϕ 's.

We begin with the β 's that served as coefficients for the four exogenous explanatory variables X_1, X_2, X_3 and X_4 . The values of these do not vary across experiments. For Eq. (2.1) these ($\beta_{11}, \beta_{12}, \beta_{13}$ and β_{14} , respectively) are set to $-0.5, 0.33, 0.57$ and -0.2 . The corresponding values for Eq. (2.2) are $-0.35, 0.33, 0.77$ and -0.18 . These values were randomly determined at the outset of the study.⁵

The remaining parameters of (2.14) and (2.15) are thus set at the outset of each experiment as follows:

1. N observations for X and Z are pseudo-randomly drawn from the multivariate standard normal distribution with zero correlation across X and Z ;
2. For each of these N observations, a pair of errors $\{\epsilon_1, \epsilon_2\}$ was drawn (either the bivariate normal distribution or via the [Vale and Maurelli \(1983\)](#) procedure, with correlation level indicated for that experiment);
3. The values for $\beta_{10}, \alpha_1, \alpha_2$ and ϕ_1 were set to guarantee the data generating process conformed to the program participation prevalence and first stage instrument strength indicated for that experiment as well as the explanatory power for Eq. (2.14) of $R^2 = 0.3$. This was done through an iterative search over candidate values for these four parameters as follows:

⁵Experimentation suggests that variation in the values assigned to these coefficient terms had very little impact on the statistics of interest in this study.

- (a) Set all four parameters to low initial values;
 - (b) Find the values for ϕ_1 and β_{10} that yield $R^2 = 0.3$ (from linear regression of Y_{i1}^* on X_i and Z_i) and the target program prevalence (with program participation Y_{1i} determined by whether Y_{i1}^* exceeds zero);
 - (c) Given these values, determine the χ^2 statistic resulting from a test of the joint significance of Z_1 and Z_2 based on a probit regression of Y_1 on the X 's and Z 's;
 - (d) If the χ^2 statistic value matched the target, the parameter value search was concluded. If not the values of α_1 and α_2 were increased incrementally and steps 3(b)–(d) were repeated.
4. Once the values for β_{10} , α_1 , α_2 and ϕ_1 had been found, Y_{1i} was determined by whether Y_{i1}^* exceeded zero given the draws for X , Z and ϵ_1 and those parameter values.
 5. The focus then shifted to Eq. (2.15), and a similar iterative process was used to find values for β_{20} , δ and ϕ_2 . It proceeded as follows:
 - (a) Set the three parameters to low initial values;
 - (b) Find values for β_{20} and ϕ_2 that yield $R^2 = 0.3$ (from linear regression of Y_{i2}^* on X_i and Y_{1i}) and the target prevalence for the outcome of interest;
 - (c) Given these values, determine the program effect according to

$$\Phi(\beta_{20} + X_{1i}\beta_{21} + X_{2i}\beta_{22} + X_{3i}\beta_{23} + X_{4i}\beta_{24} + \delta) \\ - \Phi(\beta_{20} + X_{1i}\beta_{21} + X_{2i}\beta_{22} + X_{3i}\beta_{23} + X_{4i}\beta_{24})$$

where $\Phi(\cdot)$ is the cumulative normal distribution function.

- (d) If the program impact matched the target parameter value, the search was concluded; if not δ was increased incrementally and steps 5(b)–(d) were repeated.
6. Once appropriate values for β_{20} , δ and ϕ_2 were found, Y_{2i} was determined by whether Y_{2i}^* exceeded zero given the draws for X and ϵ_2 as well as Y_{1i} and those parameter values.

The first phase of each Monte Carlo experiment thusly found values for the equation parameters that conformed to the behavioral parameters of that experiment.

The experiment then shifted to the empirical repetition phase. In each of the repetitions, the same sequence of events occurred:

1. A new draw for $\{\epsilon_1, \epsilon_2\}$ was made⁶;

⁶Step 1 was actually slightly more involved. It became apparent in early rounds of experiments that some behavioral parameters, particularly instrument strength, occasionally varied across replications to a degree with which the authors were not comfortable. In particular, the various replications from experiments involving first stage χ^2 statistics with target values of 15 and 25 occasionally produced overlapping ranges for the χ^2 statistic values actually generated across

2. Given the values assigned to $\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \alpha_1, \alpha_2$ and ϕ_1 , and $X_{1i}, X_{2i}, X_{3i}, X_{4i}, Z_{1i}, Z_{2i}$ and ϵ_{1i} , for each of the N observations Y_{i1} was set to 1 if Y_{i1}^* exceeded 0 and to 0 otherwise;
3. Given the values assigned to $\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \delta$ and ϕ_2 , and $X_{1i}, X_{2i}, X_{3i}, X_{4i}, Y_{i1}$ and ϵ_{2i} , for each of the N observations Y_{i2} was set to 1 if Y_{i2}^* exceeded 0 and to 0 otherwise.

The data X, Z, Y_1 and Y_2 so generated thus formed the empirical “observations” over which the performance of each of the estimators was then recorded for that repetition.

2.4 Monte Carlo Results

The results of the Monte Carlo experiments are presented in Tables 2.1–2.25. Tables 2.1–2.8 present mean absolute deviations between estimated and true ATE across either 1,000 (sample sizes 1,000 and 5,000) or 500 (sample size 10,000) replications of the each of the experiments. The experiments differ by their sample sizes or assumed behavioral parameters.⁷ Tables 2.9–2.16 present mean estimated ATE. Tables 2.17–2.21 present regression results summarizing the findings regarding ATE estimation. Tables 2.22–2.25 present a restricted set of results for the identification and endogeneity tests. Owing to space constraints, in Tables 2.1–2.16 and 2.22–2.25 we present only results for experiments in which the average frequencies for the two dependent variables Y_{i1} and Y_{i2} were both set to be the same at 0.25 or 0.5.

Most tables presenting Monte Carlo experiment results cover a particular combination of target average treatment effect and error correlation. In all such tables, the columns provide results by the error type applied in the experiment (bivariate normal or bivariate non-normal) and, within each error type, instrument strength in terms of the χ^2 test statistic for the joint significance of the instruments in Eq. (2.1) as estimated by probit (e.g. $\chi^2 = 15, \chi^2 = 25$, etc.) for given values of Y_1 and Y_2 (where, for instance, $Y_1 = 0.25, Y_2 = 0.25$ indicates results for experiments

the replications for the two experiments. This muddled the waters somewhat for the purposes of making inferences about estimator performance differentials as instrument strength varied. To address this, we set tolerance bands for acceptable variation of such χ^2 values around their target for a given experiment. If, on a particular replication, a draw $\{\epsilon_1, \epsilon_2\}$ resulted in a χ^2 value outside of the tolerance range for that experiment, that draw was discarded and a new draw $\{\epsilon_1, \epsilon_2\}$ was made. This was done to insure that the replications within an experiment conformed to an acceptable degree to the parameters of that experiment.

⁷As explained in Sect. 2.3, the behavioral parameters are imposed by the design of the data generating process for each experiment and included the: program effect ($Pr(Y_2|X, Y_1 = 1) - Pr(Y_2|X, Y_1 = 0)$); correlation of the errors $\{\epsilon_1, \epsilon_2\}$; average of the program outcome (Y_1) within the sample; average of the outcome of interest (Y_2) within the sample; first stage strength of the instruments Z to explain Y_1 (as reflected in the χ^2 statistic emerging from a test of the joint significance of those instruments); and bivariate error type (i.e. normal or a non-normal errors).

Table 2.1 Mean absolute deviation of ATE for true ATE = 0.05, error correlation = 0.1, $Y_1 = 0.25$ and $Y_2 = 0.25$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000							
	LPM	0.0812	0.0805	0.0771	0.0797	0.0776	0.0719
	Probit	0.0597	0.0597	0.0592	0.0662	0.0647	0.0614
	TSLS	0.2099	0.1627	0.1076	0.2123	0.1599	0.1098
	LIML	0.2279	0.1692	0.1096	0.2314	0.1672	0.1121
	GMM	0.2111	0.1632	0.1076	0.2122	0.1600	0.1099
	Residual1	0.2043	0.1674	0.1132	0.1945	0.1572	0.1130
	Residual2	0.1431	0.1235	0.0946	0.1895	0.1564	0.1053
	BIPROBIT	0.1448	0.1251	0.0961	0.2319	0.1927	0.1227
	DFM	0.1175	0.1102	0.0943	0.0938	0.0919	0.0791
N = 5,000							
	LPM	0.0820	0.0824	0.0813	0.0848	0.0838	0.0834
	Probit	0.0588	0.0594	0.0588	0.0691	0.0684	0.0684
	TSLS	0.2208	0.1684	0.1141	0.2009	0.1574	0.1058
	LIML	0.2392	0.1766	0.1163	0.2143	0.1640	0.1081
	GMM	0.2209	0.1684	0.1142	0.2009	0.1574	0.1059
	Residual1	0.2130	0.1730	0.1209	0.1905	0.1576	0.1103
	Residual2	0.0854	0.0832	0.0713	0.3110	0.2591	0.1822
	BIPROBIT	0.0973	0.0900	0.0740	0.2862	0.2648	0.2149
	DFM	0.1198	0.1125	0.1050	0.0643	0.0648	0.0643
N = 10,000							
	LPM	0.0823	0.0808	0.0818	0.0821	0.0810	0.0804
	Probit	0.0594	0.0582	0.0593	0.0664	0.0654	0.0651
	TSLS	0.2054	0.1649	0.1097	0.2066	0.1643	0.1186
	LIML	0.2183	0.1697	0.1117	0.2223	0.1707	0.1205
	GMM	0.2051	0.1649	0.1109	0.2055	0.1643	0.1187
	Residual1	0.1989	0.1635	0.1129	0.2017	0.1650	0.1224
	Residual2	0.0716	0.0644	0.0622	0.3239	0.2951	0.2463
	BIPROBIT	0.0838	0.0771	0.0666	0.2774	0.2675	0.2466
	DFM	0.1310	0.1288	0.1076	0.0597	0.0597	0.0591

for which the average values of the endogenous variable Y_1 and the outcome of interest Y_2 are 0.25). Generally speaking, the rows of these tables provide statistics for the estimators considered in this manuscript at various sample sizes. Finally, to save space, the individual models are referred to in the rows of the tables by shorthand expressions: LPM for linear probability model (i.e. single equation OLS with no control for endogeneity); Probit for single equation probit regression; TSLS for two-stage least squares; LIML for the limited information linear instrumental variables estimator; GMM for the generalized method of moments implementation of the linear instrumental variables estimator; Residual1 and Residual2 for the two variants of the residual inclusion estimators; BIPROBIT for the bivariate probit estimator provided by the STATA `-biprobit-` command; and DFM for the discrete factor model.

Table 2.2 Mean absolute deviation of ATE for true ATE = 0.05, error correlation = 0.1, $Y_1 = 0.5$ and $Y_2 = 0.5$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000							
	LPM	0.0832	0.0832	0.0780	0.1001	0.0972	0.0919
	Probit	0.0749	0.0752	0.0712	0.0870	0.0846	0.0816
	TSLs	0.2133	0.1623	0.1109	0.2097	0.1624	0.1138
	LIML	0.2341	0.1688	0.1130	0.2279	0.1693	0.1161
	GMM	0.2136	0.1624	0.1113	0.2099	0.1628	0.1138
	Residual1	0.1986	0.1597	0.1112	0.1937	0.1568	0.1139
	Residual2	0.1666	0.1412	0.1043	0.1572	0.1304	0.1026
	BIPROBIT	0.1773	0.1447	0.1051	0.1269	0.1087	0.0935
	DFM	0.1420	0.1263	0.1045	0.1021	0.0833	0.0669
N = 5,000							
	LPM	0.0815	0.0813	0.0790	0.0998	0.0992	0.0999
	Probit	0.0738	0.0737	0.0718	0.0871	0.0867	0.0877
	TSLs	0.2077	0.1624	0.1116	0.2071	0.1654	0.1128
	LIML	0.2228	0.1688	0.1137	0.2220	0.1713	0.1146
	GMM	0.2079	0.1624	0.1117	0.2071	0.1655	0.1129
	Residual1	0.1906	0.1554	0.1116	0.1900	0.1589	0.1123
	Residual2	0.1286	0.1091	0.0907	0.1771	0.1454	0.0998
	BIPROBIT	0.1399	0.1184	0.0943	0.0983	0.0837	0.0630
	DFM	0.1438	0.1270	0.1043	0.0661	0.0587	0.0511
N = 10,000							
	LPM	0.0820	0.0829	0.0814	0.0993	0.1004	0.0989
	Probit	0.0740	0.0750	0.0737	0.0862	0.0872	0.0859
	TSLs	0.2067	0.1595	0.1077	0.2157	0.1633	0.1206
	LIML	0.2206	0.1663	0.1096	0.2307	0.1699	0.1228
	GMM	0.2049	0.1592	0.1076	0.2150	0.1632	0.1206
	Residual1	0.1861	0.1511	0.1071	0.1975	0.1560	0.1206
	Residual2	0.0997	0.0964	0.0814	0.2287	0.2072	0.1603
	BIPROBIT	0.1202	0.1094	0.0853	0.1081	0.0967	0.0747
	DFM	0.1471	0.1336	0.1127	0.0887	0.0877	0.0810

Before turning to the mean absolute deviation results, it is interesting to note that Tables 2.9–2.16 for mean estimated treatment effect indicate that there is typically, though not always, an upward bias to the estimated treatment effect even for estimators that correct for the endogeneity of the treatment effect. The bias, however, is typically smaller as one moves from a true treatment effect of 0.05–0.2.

The results in Tables 2.1–2.8 on mean absolute deviations are varied and difficult to summarize. A few broad trends seem to emerge. First, the bivariate probit model (BIPROBIT) appears to do well in general when the error terms are indeed jointly normally distributed. However, at sample size 1,000 it is frequently no better than DFM, especially when instrument strength is low and is sometimes

Table 2.3 Mean absolute deviation of ATE for true ATE = 0.05, error correlation = 0.3, $Y_1 = 0.25$ and $Y_2 = 0.25$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000							
	LPM	0.1945	0.1923	0.1835	0.2124	0.2084	0.1968
	Probit	0.1678	0.1658	0.1598	0.1929	0.1896	0.1814
	TSLs	0.2149	0.1612	0.1118	0.2146	0.1625	0.1061
	LIML	0.3353	0.1668	0.1136	0.2378	0.1686	0.1083
	GMM	0.2154	0.1617	0.1119	0.2144	0.1625	0.1065
	Residual1	0.2057	0.1605	0.1134	0.2020	0.1586	0.1099
	Residual2	0.1473	0.1281	0.0942	0.2115	0.1553	0.0980
	BIPROBIT	0.1540	0.1295	0.0923	0.2308	0.1696	0.0994
	DFM	0.1164	0.1155	0.0981	0.1125	0.1004	0.0742
N = 5,000							
	LPM	0.1986	0.1991	0.1977	0.2018	0.2012	0.1991
	Probit	0.1713	0.1720	0.1710	0.1854	0.1848	0.1831
	TSLs	0.2246	0.1673	0.1238	0.1969	0.1489	0.1021
	LIML	0.2421	0.1718	0.1255	0.2114	0.1558	0.1042
	GMM	0.2248	0.1675	0.1240	0.1970	0.1490	0.1020
	Residual1	0.2182	0.1695	0.1268	0.1801	0.1485	0.1064
	Residual2	0.0897	0.0899	0.0758	0.2005	0.1686	0.1173
	BIPROBIT	0.1051	0.0963	0.0782	0.2130	0.1777	0.1232
	DFM	0.1506	0.1443	0.1240	0.1207	0.1066	0.0924
N = 10,000							
	LPM	0.1948	0.1951	0.1942	0.2081	0.2085	0.2075
	Probit	0.1675	0.1678	0.1672	0.1909	0.1914	0.1903
	TSLs	0.1938	0.1598	0.1113	0.2015	0.1549	0.1068
	LIML	0.2100	0.1650	0.1134	0.2194	0.1621	0.1088
	GMM	0.1954	0.1598	0.1112	0.2013	0.1549	0.1066
	Residual1	0.1850	0.1566	0.1129	0.1891	0.1488	0.1114
	Residual2	0.0698	0.0658	0.0613	0.2661	0.2407	0.1901
	BIPROBIT	0.0956	0.0830	0.0684	0.2665	0.2428	0.1862
	DFM	0.1562	0.1447	0.1083	0.1053	0.1080	0.0881

worse than LPM and Probit when sample size is small and error correlation is low. In addition, the Residual2 estimator which uses a first stage probit regression to generate generalized residuals frequently has lower mean absolute deviation (MAD) than BIPROBIT. Whatever advantage BIPROBIT has when the true errors are normal disappears for non-normal errors. For non-normal errors, the DFM model typically performs the best. The linear instrumental variables estimators' performance increases significantly as instrument strength and sample size increases regardless of whether or not the true error distribution is normal or non-normal. That said, it is understandably difficult to grasp general patterns from the many cells of these tables.

Table 2.4 Mean absolute deviation of ATE for true ATE = 0.05, error correlation = 0.3, $Y_1 = 0.5$ and $Y_2 = 0.5$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000							
	LPM	0.2033	0.1994	0.1946	0.2308	0.2293	0.2211
	Probit	0.1948	0.1913	0.1871	0.2089	0.2076	0.2010
	TSLs	0.2192	0.1680	0.1162	0.2135	0.1633	0.1180
	LIML	0.2345	0.1726	0.1174	0.2305	0.1702	0.1188
	GMM	0.2193	0.1684	0.1166	0.2138	0.1643	0.1189
	Residual1	0.2047	0.1633	0.1162	0.1975	0.1576	0.1141
	Residual2	0.1805	0.1495	0.1093	0.1525	0.1263	0.0968
	BIPROBIT	0.1854	0.1494	0.1068	0.1331	0.1132	0.0909
	DFM	0.1411	0.1170	0.0969	0.1107	0.0933	0.0775
N = 5,000							
	LPM	0.2031	0.2020	0.2015	0.2269	0.2255	0.2249
	Probit	0.1953	0.1942	0.1939	0.2115	0.2101	0.2099
	TSLs	0.2092	0.1600	0.1139	0.2156	0.1706	0.1244
	LIML	0.2347	0.1657	0.1157	0.2288	0.1751	0.1250
	GMM	0.2092	0.1601	0.1140	0.2158	0.1707	0.1245
	Residual1	0.1893	0.1539	0.1132	0.1977	0.1638	0.1193
	Residual2	0.1248	0.1080	0.0934	0.2125	0.1742	0.1203
	BIPROBIT	0.1480	0.1180	0.0946	0.1177	0.0873	0.0657
	DFM	0.1491	0.1380	0.1070	0.0516	0.0454	0.0468
N = 10,000							
	LPM	0.2039	0.2024	0.2029	0.2316	0.2309	0.2297
	Probit	0.1956	0.1941	0.1947	0.2136	0.2130	0.2120
	TSLs	0.2186	0.1673	0.1142	0.2117	0.1599	0.1102
	LIML	0.2392	0.1768	0.1166	0.2273	0.1668	0.1116
	GMM	0.2184	0.1649	0.1138	0.2146	0.1606	0.1103
	Residual1	0.1869	0.1540	0.1127	0.1959	0.1547	0.1103
	Residual2	0.1077	0.0968	0.0746	0.1922	0.1683	0.1265
	BIPROBIT	0.1413	0.1186	0.0837	0.0725	0.0627	0.0516
	DFM	0.1557	0.1322	0.1102	0.0761	0.0689	0.0678

To perhaps provide a somewhat clearer overall picture, we consider a series of simple regression results. Tables 2.17–2.21 provide results for these regression analyses. These involve regressing mean absolute deviation estimates across the replications of our Monte Carlo experiments on dummy variables capturing the models that generated those mean absolute deviation estimates. The regression relies on a sample that has an observation for each mean absolute deviation estimate generated by each model considered in each Monte Carlo experiment (for instance, the typical Monte Carlo experiment will yield nine observations in the regression sample corresponding to the mean absolute deviation estimates generated by the various models). In Table 2.17, we present results across all experiments and a

Table 2.5 Mean absolute deviation of ATE for true ATE = 0.2, error correlation = 0.1, $Y_1 = 0.25$ and $Y_2 = 0.25$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000							
	LPM	0.0972	0.0976	0.0939	0.0577	0.0569	0.0565
	Probit	0.0711	0.0715	0.0700	0.0426	0.0424	0.0432
	TSLS	0.2150	0.1650	0.1145	0.2036	0.1595	0.1051
	LIML	0.2350	0.1711	0.1167	0.2204	0.1663	0.1071
	GMM	0.2160	0.1652	0.1148	0.2044	0.1603	0.1053
	Residual1	0.2312	0.1897	0.1372	0.2173	0.1820	0.1270
	Residual2	0.1684	0.1457	0.1112	0.1974	0.1690	0.1093
	BIPROBIT	0.1741	0.1500	0.1111	0.2482	0.2076	0.1318
	DFM	0.1324	0.1264	0.1143	0.1332	0.1328	0.1250
N = 5,000							
	LPM	0.0919	0.0919	0.0906	0.0734	0.0735	0.0706
	Probit	0.0639	0.0639	0.0633	0.0555	0.0558	0.0536
	TSLS	0.2245	0.1621	0.1155	0.2026	0.1485	0.1078
	LIML	0.2446	0.1690	0.1172	0.2175	0.1539	0.1097
	GMM	0.2246	0.1621	0.1157	0.2027	0.1484	0.1077
	Residual1	0.2385	0.1892	0.1375	0.2161	0.1714	0.1312
	Residual2	0.0950	0.0929	0.0825	0.3361	0.2919	0.2127
	BIPROBIT	0.1071	0.1004	0.0849	0.3154	0.2988	0.2520
	DFM	0.0977	0.0933	0.0924	0.0616	0.0576	0.0542
N = 10,000							
	LPM	0.0960	0.0962	0.0958	0.0577	0.0584	0.0577
	Probit	0.0673	0.0676	0.0673	0.0405	0.0411	0.0407
	TSLS	0.1935	0.1508	0.1062	0.1974	0.1516	0.1093
	LIML	0.2066	0.1571	0.1081	0.2147	0.1572	0.1116
	GMM	0.1933	0.1507	0.1064	0.1968	0.1518	0.1093
	Residual1	0.2190	0.1793	0.1293	0.2154	0.1741	0.1292
	Residual2	0.0746	0.0790	0.0663	0.3123	0.2908	0.2362
	BIPROBIT	0.0869	0.0902	0.0725	0.2798	0.2707	0.2444
	DFM	0.1351	0.1341	0.1157	0.0414	0.0433	0.0412

stratification by error type (bivariate normal versus bivariate non-normal). The omitted category among the regressors (which are dummy variables indicating the model behind the mean absolute deviation estimate in a particular observation) is the linear probability model (LPM). Thus, a negative number means that the model outperforms the omitted category model (the LPM) while a positive number means that it performed more poorly than that omitted category model. For these tables we used all of the experiments (i.e. we did not confine ourselves to cases where program participation prevalence and outcome prevalence were both 0.25 or 0.5).

From Table 2.17 it is clear that, across all Monte Carlo experiments, only simple Probit and DFM perform slightly better than LPM while all other estimators perform

Table 2.6 Mean absolute deviation of ATE for true ATE = 0.2, error correlation = 0.1, $Y_1 = 0.5$ and $Y_2 = 0.5$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000							
	LPM	0.0809	0.0814	0.0816	0.1071	0.1062	0.1035
	Probit	0.0728	0.0737	0.0746	0.0888	0.0886	0.0891
	TSLs	0.2147	0.1704	0.1173	0.2233	0.1676	0.1193
	LIML	0.2328	0.1780	0.1192	0.2447	0.1739	0.1209
	GMM	0.2157	0.1706	0.1175	0.2240	0.1684	0.1197
	Residual1	0.2080	0.1752	0.1259	0.2210	0.1732	0.1263
	Residual2	0.1801	0.1548	0.1137	0.1859	0.1450	0.1100
	BIPROBIT	0.1879	0.1591	0.1144	0.1362	0.1126	0.0962
	DFM	0.1430	0.1405	0.1216	0.1671	0.1634	0.1555
N = 5,000							
	LPM	0.0759	0.0747	0.0756	0.1003	0.0997	0.1006
	Probit	0.0681	0.0670	0.0680	0.0831	0.0829	0.0841
	TSLs	0.2017	0.1592	0.1176	0.2119	0.1647	0.1208
	LIML	0.2182	0.1654	0.1195	0.2243	0.1708	0.1225
	GMM	0.2018	0.1592	0.1177	0.2121	0.1648	0.1209
	Residual1	0.2001	0.1641	0.1252	0.2091	0.1713	0.1283
	Residual2	0.1307	0.1164	0.1000	0.2655	0.2196	0.1481
	BIPROBIT	0.1443	0.1252	0.1023	0.1519	0.1224	0.0851
	DFM	0.1273	0.1185	0.1060	0.1026	0.1046	0.0983
N = 10,000							
	LPM	0.0826	0.0816	0.0809	0.0939	0.0930	0.0921
	Probit	0.0744	0.0734	0.0728	0.0764	0.0756	0.0748
	TSLs	0.2023	0.1590	0.1140	0.2204	0.1684	0.1187
	LIML	0.2173	0.1645	0.1159	0.2361	0.1764	0.1209
	GMM	0.2034	0.1606	0.1143	0.2205	0.1673	0.1192
	Residual1	0.2010	0.1649	0.1231	0.2178	0.1740	0.1298
	Residual2	0.1051	0.0998	0.0880	0.3435	0.3172	0.2567
	BIPROBIT	0.1227	0.1102	0.0946	0.2020	0.1792	0.1398
	DFM	0.1462	0.1352	0.1160	0.0693	0.0676	0.0694

slightly worse. For Monte Carlo experiments involving normal errors, BIPROBIT and Residual2 perform slightly better than LPM while DFM and Probit perform about the same as LPM. This result for BIPROBIT is not surprising since it is the asymptotically efficient estimator, given that it is based on a joint distributional assumption for the errors that happens to exactly match the actual error distribution behind the data generating process. The other four estimators perform worse. Although their point estimates are small, they are significantly different from zero in all four cases. For non-normal errors, no estimator performs better than LPM except for DFM and the two worst performing estimators are Residual2 and BIPROBIT. This is not surprising since these two estimators rely heavily on a normality assumption for the error term.

Table 2.7 Mean absolute deviation of ATE for true ATE = 0.2, error correlation = 0.3, $Y_1 = 0.25$ and $Y_2 = 0.25$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000							
	LPM	0.1985	0.2020	0.1918	0.1951	0.1883	0.1747
	Probit	0.1779	0.1820	0.1718	0.1806	0.1741	0.1630
	TSLS	0.2144	0.1643	0.1098	0.2126	0.1599	0.1116
	LIML	0.2295	0.1701	0.1108	0.2302	0.1658	0.1137
	GMM	0.2157	0.1648	0.1100	0.2134	0.1602	0.1116
	Residual1	0.2322	0.1898	0.1348	0.2290	0.1894	0.1419
	Residual2	0.1565	0.1472	0.1130	0.2076	0.1661	0.1137
	BIPROBIT	0.1647	0.1469	0.1086	0.2401	0.1867	0.1180
	DFM	0.1432	0.1415	0.1208	0.1757	0.1546	0.1407
N = 5,000							
	LPM	0.1957	0.1955	0.1936	0.1602	0.1606	0.1575
	Probit	0.1749	0.1748	0.1733	0.1487	0.1490	0.1461
	TSLS	0.2196	0.1737	0.1211	0.2033	0.1526	0.1068
	LIML	0.2346	0.1788	0.1224	0.2197	0.1598	0.1094
	GMM	0.2198	0.1739	0.1212	0.2034	0.1526	0.1067
	Residual1	0.2381	0.2004	0.1422	0.2208	0.1829	0.1471
	Residual2	0.0962	0.0928	0.0853	0.1612	0.1429	0.0988
	BIPROBIT	0.1144	0.1041	0.0883	0.1907	0.1648	0.1160
	DFM	0.1093	0.1098	0.0978	0.0966	0.0925	0.0782
N = 10,000							
	LPM	0.1963	0.1969	0.1960	0.1694	0.1692	0.1701
	Probit	0.1748	0.1754	0.1747	0.1571	0.1570	0.1577
	TSLS	0.1784	0.1465	0.1072	0.1987	0.1503	0.1087
	LIML	0.1895	0.1517	0.1083	0.2153	0.1581	0.1115
	GMM	0.1791	0.1452	0.1085	0.1989	0.1505	0.1087
	Residual1	0.2090	0.1787	0.1380	0.2216	0.1827	0.1484
	Residual2	0.0809	0.0692	0.0680	0.1978	0.1720	0.1366
	BIPROBIT	0.1051	0.0869	0.0793	0.2025	0.1780	0.1439
	DFM	0.1381	0.1244	0.1060	0.1056	0.0916	0.0727

The results presented above likely mask some important variations in the performance of the estimators for different configurations of the data generating process. In Table 2.18 we present results based on further stratification of the simple regression by error correlation. For normal errors and the lower error correlation level of 0.1, no estimator has a lower MAD than LPM but the Probit estimator's MAD is not significantly different from that for the LPM. However, it is interesting to note that the relative performance of the estimators is completely different with normal errors and error correlation 0.3. Now Residual2 and BIPROBIT are the dominant estimators followed closely by DFM while the other estimators are not much different in terms of MAD from LPM. For non-normal errors, the results

Table 2.8 Mean absolute deviation of ATE for true ATE = 0.2, error correlation = 0.3, $Y_1 = 0.5$ and $Y_2 = 0.5$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000							
	LPM	0.1757	0.1706	0.1712	0.2202	0.2199	0.2151
	Probit	0.1715	0.1662	0.1669	0.2027	0.2025	0.1990
	TSLs	0.2177	0.1627	0.1134	0.2101	0.1670	0.1324
	LIML	0.2342	0.1698	0.1156	0.2252	0.1712	0.1335
	GMM	0.2182	0.1636	0.1138	0.2105	0.1679	0.1334
	Residual1	0.2162	0.1728	0.1258	0.2183	0.1787	0.1372
	Residual2	0.1862	0.1549	0.1172	0.1891	0.1507	0.1195
	BIPROBIT	0.1920	0.1535	0.1125	0.1487	0.1226	0.1049
	DFM	0.1428	0.1385	0.1199	0.1607	0.1634	0.1512
N = 5,000							
	LPM	0.1750	0.1745	0.1732	0.1934	0.1932	0.1926
	Probit	0.1707	0.1703	0.1691	0.1835	0.1832	0.1830
	TSLs	0.2058	0.1630	0.1143	0.2038	0.1582	0.1135
	LIML	0.2201	0.1700	0.1166	0.2188	0.1641	0.1150
	GMM	0.2059	0.1631	0.1144	0.2037	0.1583	0.1136
	Residual1	0.2074	0.1777	0.1292	0.2130	0.1738	0.1272
	Residual2	0.1324	0.1179	0.0992	0.3206	0.2778	0.2028
	BIPROBIT	0.1526	0.1310	0.1020	0.2706	0.2052	0.1334
	DFM	0.1211	0.1090	0.0992	0.0905	0.0924	0.0889
N = 10,000							
	LPM	0.1799	0.1797	0.1793	0.2073	0.2063	0.2068
	Probit	0.1753	0.1751	0.1748	0.1939	0.1930	0.1936
	TSLs	0.2101	0.1485	0.1133	0.2052	0.1657	0.1058
	LIML	0.2290	0.1549	0.1154	0.2200	0.1721	0.1076
	GMM	0.2116	0.1483	0.1132	0.2052	0.1657	0.1085
	Residual1	0.2166	0.1623	0.1305	0.2141	0.1792	0.1249
	Residual2	0.1065	0.1052	0.0907	0.3346	0.3080	0.2551
	BIPROBIT	0.1393	0.1226	0.0991	0.2238	0.1890	0.1397
	DFM	0.1349	0.1364	0.1157	0.0504	0.0521	0.0485

are quite different. We see that for error correlation 0.1, only PROBIT and DFM perform as well as LPM, with all other methods performing significantly worse. At the error correlation level of 0.3, DFM dominates all other estimators.

We also consider stratification of the summary regression by instrument strength. Results for this are presented in Tables 2.19 and 2.20. We consider only two instrument strength levels (as manifested by the size of the χ^2 statistic obtained from a test of the joint significance of the instruments in a probit regression with Y_2 as the dependent variable): $\chi^2 = 15$ and $\chi^2 = 50$. Not surprisingly, the estimators most affected by instrument strength are the linear instrumental variables methods. They perform quite poorly compared with the LPM at instrument strength $\chi^2 = 15$.

Table 2.9 Mean ATE for true ATE = 0.05, error correlation = 0.1, $Y_1 = 0.25$ and $Y_2 = 0.25$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
Obs = 1,000							
	LPM	0.1306	0.1299	0.1261	0.1291	0.1269	0.1212
	Probit	0.1084	0.1086	0.1077	0.1155	0.1139	0.1106
	TSLS	0.0755	0.0770	0.0654	0.0184	0.0444	0.0486
	LIML	0.0739	0.0745	0.0642	0.0074	0.0400	0.0469
	GMM	0.0761	0.0773	0.0657	0.0181	0.0443	0.0488
	Residual1	0.1089	0.0980	0.0677	0.0527	0.0613	0.0516
	Residual2	0.0896	0.0852	0.0701	0.1755	0.1410	0.0913
	BIPROBIT	0.0811	0.0808	0.0735	0.1675	0.1465	0.0943
	DFM	0.1175	0.1076	0.0952	0.1076	0.1045	0.0819
Obs = 5,000							
	LPM	0.1315	0.1319	0.1309	0.1343	0.1334	0.1329
	Probit	0.1083	0.1089	0.1083	0.1186	0.1179	0.1179
	TSLS	0.0948	0.0880	0.0821	0.0477	0.0480	0.0545
	LIML	0.0913	0.0858	0.0810	0.0424	0.0443	0.0530
	GMM	0.0949	0.0883	0.0823	0.0481	0.0481	0.0547
	Residual1	0.1122	0.0967	0.0793	0.0655	0.0532	0.0483
	Residual2	0.0705	0.0726	0.0699	0.3602	0.3070	0.2279
	BIPROBIT	0.0524	0.0596	0.0640	0.3292	0.3034	0.2495
	DFM	0.1186	0.1082	0.1105	0.1105	0.1097	0.1096
Obs = 10,000							
	LPM	0.1318	0.1303	0.1313	0.1316	0.1305	0.1300
	Probit	0.1089	0.1077	0.1088	0.1159	0.1149	0.1146
	TSLS	0.0837	0.0663	0.0727	0.1136	0.0944	0.0981
	LIML	0.0813	0.0647	0.0719	0.1125	0.0927	0.0979
	GMM	0.0837	0.0664	0.0801	0.1158	0.0941	0.0982
	Residual1	0.0878	0.0656	0.0669	0.1160	0.0915	0.0883
	Residual2	0.0708	0.0560	0.0730	0.3734	0.3446	0.2958
	BIPROBIT	0.0532	0.0399	0.0613	0.3269	0.3171	0.2940
	DFM	0.1227	0.1229	0.1007	0.1088	0.1085	0.1079

However, even at instrument strength $\chi^2 = 50$, they do not perform any better than the LPM model (or at least they do not do so to a statistically significant degree). For BIPROBIT and Residual2, we see improved performance as instrument strength increases for both normal and non-normal errors. Finally, DFM improves with increasing instrument strength for normal errors but does roughly equally well at the two instrument strengths when the errors are non-normal.

In Table 2.20, rather than stratifying by error distribution, we stratify by error correlation and then instrument strength. This table clearly isolates the cases in which the linear instrumental variables estimators perform relatively well. We see that all three linear instrumental variables estimators are inferior to LPM when the error correlation is low regardless of instrument strength. At the lower value

Table 2.10 Mean ATE for true ATE = 0.05, error correlation = 0.1, $Y_1 = 0.5$ and $Y_2 = 0.5$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
Obs = 1,000							
	LPM	0.1327	0.1326	0.1273	0.1496	0.1467	0.1414
	Probit	0.1243	0.1245	0.1204	0.1365	0.1341	0.1310
	TSLS	0.0826	0.0818	0.0773	0.0923	0.0874	0.0812
	LIML	0.0743	0.0804	0.0763	0.0884	0.0843	0.0800
	GMM	0.0832	0.0819	0.0778	0.0926	0.0881	0.0817
	Residual1	0.0823	0.0787	0.0719	0.0813	0.0760	0.0707
	Residual2	0.0804	0.0783	0.0746	0.0100	0.0290	0.0589
	BIPROBIT	0.0703	0.0779	0.0761	0.0271	0.0420	0.0654
	DFM	0.1224	0.1131	0.0922	0.1215	0.0982	0.0808
Obs = 5,000							
	LPM	0.1310	0.1308	0.1285	0.1493	0.1487	0.1494
	Probit	0.1233	0.1232	0.1213	0.1366	0.1362	0.1372
	TSLS	0.0770	0.0680	0.0639	0.1066	0.0896	0.0906
	LIML	0.0724	0.0653	0.0626	0.1054	0.0872	0.0895
	GMM	0.0771	0.0680	0.0639	0.1068	0.0897	0.0908
	Residual1	0.0751	0.0631	0.0568	0.1032	0.0835	0.0818
	Residual2	0.0745	0.0689	0.0613	−0.1167	−0.0774	−0.0217
	BIPROBIT	0.0538	0.0563	0.0567	−0.0320	−0.0098	0.0230
	DFM	0.1315	0.1148	0.0978	0.1074	0.0998	0.0918
Obs = 10,000							
	LPM	0.1315	0.1324	0.1309	0.1488	0.1499	0.1484
	Probit	0.1235	0.1245	0.1232	0.1357	0.1367	0.1354
	TSLS	0.0293	0.0377	0.0518	0.0605	0.0592	0.0721
	LIML	0.0242	0.0347	0.0503	0.0544	0.0562	0.0705
	GMM	0.0314	0.0362	0.0519	0.0575	0.0562	0.0726
	Residual1	0.0569	0.0494	0.0568	0.0905	0.0799	0.0830
	Residual2	0.0742	0.0652	0.0580	−0.1792	−0.1570	−0.1088
	BIPROBIT	0.0454	0.0486	0.0525	−0.0566	−0.0451	−0.0186
	DFM	0.1296	0.1185	0.1045	0.1360	0.1348	0.1285

for instrument strength, the linear instrumental variables estimators still offer no improvement over LPM when error correlation is 0.1. However, they offer substantial improvement over LPM when error correlation is 0.3 and instrument strength is high. We do not display results for instrument strength 25 but in this case, the linear instrumental variables estimators offer slight improvement over LPM with the higher error correlation. This relatively strong performance for the linear instrumental variables methods is robust to a further stratification by error distribution (results not displayed). When error correlation is 0.3 and instrument strength is 50, there is no difference in the level of improvement over LPM for normal or non-normal errors. This is reassuring given that the linear instrumental approach has been recommended in this setting (e.g. Angrist and Krueger 2001).

Table 2.11 Mean ATE for true ATE = 0.05, error correlation = 0.3, $Y_1 = 0.25$ and $Y_2 = 0.25$

	Normal errors			Non-normal errors		
	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
Obs = 1,000						
LPM	0.2440	0.2418	0.2330	0.2619	0.2579	0.2463
Probit	0.2173	0.2153	0.2093	0.2424	0.2391	0.2309
TSLs	0.0965	0.0886	0.0742	0.0768	0.0613	0.0538
LIML	-0.0203	0.0812	0.0706	0.0683	0.0533	0.0499
GMM	0.0960	0.0889	0.0745	0.0765	0.0610	0.0537
Residual1	0.1086	0.0827	0.0531	0.0856	0.0543	0.0302
Residual2	0.1071	0.0983	0.0728	0.2143	0.1460	0.0825
BIPROBIT	0.0969	0.0955	0.0751	0.1886	0.1280	0.0722
DFM	0.1306	0.1298	0.1059	0.1380	0.1230	0.0835
Obs = 5,000						
LPM	0.2481	0.2486	0.2472	0.2513	0.2507	0.2486
Probit	0.2209	0.2215	0.2205	0.2349	0.2343	0.2326
TSLs	0.1234	0.1226	0.1071	0.0254	0.0442	0.0453
LIML	0.1148	0.1168	0.1040	0.0109	0.0359	0.0414
GMM	0.1242	0.1232	0.1076	0.0256	0.0446	0.0455
Residual1	0.1229	0.1076	0.0802	0.0303	0.0298	0.0152
Residual2	0.0907	0.0970	0.0866	0.2387	0.2068	0.1512
BIPROBIT	0.0638	0.0810	0.0770	0.1984	0.1731	0.1202
DFM	0.1742	0.1675	0.1497	0.1665	0.1511	0.1353
Obs = 10,000						
LPM	0.2444	0.2446	0.2438	0.2576	0.2580	0.2570
Probit	0.2170	0.2173	0.2167	0.2404	0.2409	0.2398
TSLs	0.0552	0.0694	0.0730	0.0571	0.0566	0.0566
LIML	0.0401	0.0635	0.0694	0.0409	0.0480	0.0519
GMM	0.0532	0.0697	0.0732	0.0568	0.0561	0.0563
Residual1	0.0459	0.0485	0.0376	0.0495	0.0364	0.0210
Residual2	0.0839	0.0803	0.0748	0.3151	0.2896	0.2390
BIPROBIT	0.0410	0.0465	0.0530	0.2984	0.2745	0.2219
DFM	0.1764	0.1647	0.1309	0.1546	0.1571	0.1361

However, the linear instrumental variables estimator never performs as well as DFM for any of these stratifications.

Finally, in Table 2.21 we add to the basic summary models presented in Table 2.17 controls for the behavioral parameters of the Monte Carlo experiment. Among the sample size regressors, the omitted category is experiments with 1,000 observations. The omitted instrument strength is $\chi^2 = 15$ (the lowest). The comparison values for error correlation and treatment effect are 0.1 and 0.05, respectively. Finally, for both program (i.e. program enrollment) and treatment prevalence the comparison value is 0.5.

There appears to be a clear performance improvement at the larger sample sizes with normal errors but performance actually deteriorates as sample size

Table 2.12 Mean ATE for true ATE = 0.05, error correlation = 0.3, $Y_1 = 0.5$ and $Y_2 = 0.5$

	Normal errors			Non-normal errors		
	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
Obs = 1,000						
LPM	0.2528	0.2489	0.2441	0.2803	0.2788	0.2706
Probit	0.2444	0.2408	0.2366	0.2584	0.2571	0.2505
TSLs	0.1183	0.0991	0.0997	0.1100	0.0998	0.1123
LIML	0.1084	0.0932	0.0970	0.0932	0.0902	0.1085
GMM	0.1185	0.0996	0.1001	0.1119	0.1013	0.1142
Residual1	0.1108	0.0890	0.0839	0.0851	0.0695	0.0784
Residual2	0.1138	0.0948	0.0903	0.0251	0.0368	0.0717
BIPROBIT	0.1087	0.0956	0.0933	0.0531	0.0624	0.0896
DFM	0.1471	0.1224	0.1006	0.1382	0.1168	0.1035
Obs = 5,000						
LPM	0.2526	0.2515	0.2510	0.2765	0.2750	0.2744
Probit	0.2448	0.2437	0.2434	0.2610	0.2596	0.2594
TSLs	0.0375	0.0667	0.0735	0.1199	0.1332	0.1153
LIML	0.0131	0.0597	0.0700	0.1091	0.1278	0.1122
GMM	0.0374	0.0668	0.0734	0.1202	0.1333	0.1154
Residual1	0.0402	0.0577	0.0565	0.1069	0.1115	0.0892
Residual2	0.0729	0.0737	0.0712	-0.1611	-0.1201	-0.0576
BIPROBIT	0.0380	0.0548	0.0644	-0.0585	-0.0189	0.0212
DFM	0.1577	0.1424	0.1145	0.0952	0.0890	0.0909
Obs = 10,000						
LPM	0.2534	0.2519	0.2524	0.2811	0.2804	0.2792
Probit	0.2451	0.2436	0.2442	0.2631	0.2625	0.2615
TSLs	-0.0107	-0.0095	0.0325	0.0611	0.0701	0.0788
LIML	-0.0313	-0.0211	0.0285	0.0445	0.0602	0.0742
GMM	-0.0105	-0.0080	0.0329	0.0641	0.0685	0.0803
Residual1	0.0166	-0.0015	0.0261	0.0818	0.0742	0.0712
Residual2	0.0694	0.0526	0.0601	-0.1421	-0.1172	-0.0732
BIPROBIT	0.0161	0.0156	0.0394	-0.0086	0.0083	0.0318
DFM	0.1674	0.1426	0.1111	0.1253	0.1183	0.1172

increases for non-normal errors. Interestingly, however, the effects of instrument strength and error correlation do not differ substantially by error type. Increasing the instrument strength always reduces MAD while increasing the error correlation always increases it. The true treatment effect has a small but significant effect on performance (with performance deteriorating as true treatment effect increases). Program participation and outcome prevalence have substantial, highly significant effects, but in opposite directions. Performance clearly worsens with non-normal errors.

Before proceeding, it is worth reflecting on the generally poor performance of several estimators in the case of non-normal errors. This is very concerning when one considers that, in many respects, the non-normal error distribution considered in

Table 2.13 Mean ATE for true ATE = 0.2, error correlation = 0.1, $Y_1 = 0.25$ and $Y_2 = 0.25$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
Obs = 1,000							
	LPM	0.2965	0.2970	0.2933	0.2564	0.2559	0.2556
	Probit	0.2694	0.2704	0.2691	0.2395	0.2399	0.2410
	TSLS	0.2403	0.2468	0.2399	0.1717	0.2068	0.1941
	LIML	0.2315	0.2437	0.2387	0.1624	0.2043	0.1928
	GMM	0.2404	0.2469	0.2402	0.1714	0.2066	0.1941
	Residual1	0.2336	0.2325	0.2209	0.1688	0.1966	0.1766
	Residual2	0.2223	0.2349	0.2228	0.3155	0.2953	0.2319
	BIPROBIT	0.2115	0.2326	0.2246	0.3170	0.3030	0.2375
	DFM	0.1704	0.1715	0.1846	0.1981	0.1949	0.1900
Obs = 5,000							
	LPM	0.2914	0.2914	0.2901	0.2730	0.2730	0.2701
	Probit	0.2634	0.2634	0.2628	0.2550	0.2553	0.2531
	TSLS	0.2522	0.2404	0.2507	0.2097	0.2061	0.2032
	LIML	0.2504	0.2377	0.2502	0.2043	0.2039	0.2021
	GMM	0.2527	0.2406	0.2510	0.2102	0.2061	0.2033
	Residual1	0.2378	0.2237	0.2293	0.1968	0.1870	0.1761
	Residual2	0.2167	0.2147	0.2208	0.5348	0.4901	0.4086
	BIPROBIT	0.1996	0.2035	0.2167	0.5092	0.4914	0.4409
	DFM	0.2120	0.2024	0.2199	0.2074	0.2053	0.2033
Obs = 10,000							
	LPM	0.2956	0.2957	0.2953	0.2572	0.2579	0.2572
	Probit	0.2668	0.2671	0.2668	0.2400	0.2406	0.2402
	TSLS	0.2417	0.2283	0.2309	0.2236	0.2338	0.2192
	LIML	0.2380	0.2254	0.2295	0.2231	0.2333	0.2182
	GMM	0.2419	0.2283	0.2310	0.2183	0.2338	0.2192
	Residual1	0.2124	0.1928	0.1942	0.1981	0.2105	0.1939
	Residual2	0.2133	0.2122	0.2092	0.5118	0.4902	0.4354
	BIPROBIT	0.1977	0.1975	0.2006	0.4792	0.4698	0.4432
	DFM	0.2560	0.2504	0.2378	0.2165	0.2164	0.2136

this study represents a rather forgiving departure from joint normality. For instance, it still involves unimodal marginal distributions for the errors and a unimodal surface for the joint density of the errors in \mathbf{R}^3 . This may indeed be too generous from the standpoint of accurately reflecting conditions likely to be encountered in actual applied microeconomic settings.

For instance, in the real world the joint distribution of the error term from a particular application involving a system along the lines of Eqs. (2.1) and (2.2) is likely often to involve multi-modality: the joint distribution of the unobservables for the error term in many settings is likely to reflect substantial mass for extreme (in terms of behavior) types of individuals that would be difficult to accommodate accurately with unimodal joint distributions under which such varied and extreme

Table 2.14 Mean ATE for true ATE = 0.2, error correlation = 0.1, $Y_1 = 0.5$ and $Y_2 = 0.5$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
Obs = 1,000							
	LPM	0.2803	0.2809	0.2809	0.3066	0.3057	0.3030
	Probit	0.2722	0.2731	0.2737	0.2883	0.2880	0.2886
	TSLS	0.2511	0.2310	0.2291	0.2714	0.2665	0.2611
	LIML	0.2471	0.2289	0.2279	0.2646	0.2642	0.2601
	GMM	0.2525	0.2316	0.2297	0.2729	0.2676	0.2618
	Residual1	0.2305	0.2154	0.2147	0.2383	0.2410	0.2459
	Residual2	0.2307	0.2167	0.2183	0.1206	0.1569	0.2122
	BIPROBIT	0.2262	0.2162	0.2197	0.1575	0.1826	0.2237
	DFM	0.1932	0.1802	0.1781	0.2455	0.2402	0.2563
Obs = 5,000							
	LPM	0.2754	0.2742	0.2751	0.2998	0.2992	0.3001
	Probit	0.2676	0.2665	0.2676	0.2826	0.2824	0.2836
	TSLS	0.2061	0.2139	0.2202	0.2811	0.2649	0.2628
	LIML	0.2016	0.2116	0.2192	0.2799	0.2633	0.2621
	GMM	0.2063	0.2140	0.2203	0.2814	0.2649	0.2630
	Residual1	0.1858	0.1968	0.2044	0.2575	0.2433	0.2437
	Residual2	0.2070	0.2054	0.2061	−0.0613	−0.0140	0.0676
	BIPROBIT	0.1901	0.1941	0.2026	0.0547	0.0859	0.1351
	DFM	0.2132	0.2075	0.2031	0.1920	0.1947	0.2053
Obs = 10,000							
	LPM	0.2821	0.2811	0.2804	0.2934	0.2925	0.2916
	Probit	0.2739	0.2729	0.2723	0.2759	0.2751	0.2743
	TSLS	0.1942	0.2002	0.2089	0.2346	0.2432	0.2289
	LIML	0.1878	0.1977	0.2076	0.2294	0.2404	0.2273
	GMM	0.1937	0.2000	0.2103	0.2348	0.2418	0.2289
	Residual1	0.1869	0.1938	0.2013	0.2339	0.2372	0.2224
	Residual2	0.2126	0.2031	0.2041	−0.1440	−0.1177	−0.0572
	BIPROBIT	0.1898	0.1888	0.1984	−0.0025	0.0203	0.0598
	DFM	0.2488	0.2482	0.2307	0.1748	0.1718	0.1780

combinations are typically found only with much lower probability. If Y_1 were smoking and Y_2 were obesity, for example, one could easily imagine a significant proportion of the population with combinations of strong unobserved tendencies toward and away from smoking and obesity that are hard to accommodate with unimodal (in terms of marginal errors or density surface in \mathbf{R}^3) errors, let alone joint normality. However, it is also hard to believe that the performance of many models (such as those based on joint normality) would improve from what is presented in this manuscript once the departure from joint normality involved relaxing the assumption of unimodality.

In Tables 2.22–2.25, we examine a limited set of results for the endogeneity tests and the identification tests considered in our Monte Carlo experiments. To begin

Table 2.15 Mean ATE for true ATE = 0.2, error correlation = 0.3, $Y_1 = 0.25$ and $Y_2 = 0.25$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
Obs = 1,000							
	LPM	0.3981	0.4015	0.3913	0.3946	0.3878	0.3742
	Probit	0.3774	0.3815	0.3713	0.3801	0.3736	0.3625
	TSLS	0.2668	0.2543	0.2463	0.2156	0.2264	0.1967
	LIML	0.2561	0.2474	0.2431	0.2005	0.2190	0.1930
	GMM	0.2676	0.2549	0.2468	0.2154	0.2267	0.1970
	Residual1	0.2317	0.2095	0.1909	0.1826	0.1830	0.1460
	Residual2	0.2301	0.2257	0.2129	0.3335	0.2908	0.2075
	BIPROBIT	0.2189	0.2221	0.2173	0.3056	0.2755	0.1984
	DFM	0.1947	0.1892	0.1794	0.2499	0.2228	0.1860
Obs = 5,000							
	LPM	0.3952	0.3950	0.3931	0.3597	0.3601	0.3570
	Probit	0.3744	0.3743	0.3728	0.3482	0.3485	0.3456
	TSLS	0.2850	0.2826	0.2625	0.1494	0.1651	0.1624
	LIML	0.2770	0.2779	0.2590	0.1358	0.1573	0.1587
	GMM	0.2855	0.2832	0.2629	0.1498	0.1655	0.1625
	Residual1	0.2464	0.2408	0.2081	0.1246	0.1236	0.1066
	Residual2	0.2164	0.2219	0.2175	0.3340	0.3097	0.2487
	BIPROBIT	0.1905	0.2048	0.2092	0.2971	0.2809	0.2194
	DFM	0.2483	0.2534	0.2409	0.2604	0.2486	0.2191
Obs = 10,000							
	LPM	0.3958	0.3964	0.3955	0.3689	0.3687	0.3696
	Probit	0.3743	0.3749	0.3742	0.3566	0.3565	0.3572
	TSLS	0.2316	0.2282	0.2297	0.1636	0.1702	0.1713
	LIML	0.2227	0.2211	0.2260	0.1480	0.1609	0.1670
	GMM	0.2306	0.2265	0.2288	0.1630	0.1699	0.1712
	Residual1	0.1730	0.1621	0.1588	0.1274	0.1185	0.1098
	Residual2	0.2176	0.2123	0.2032	0.3943	0.3681	0.3288
	BIPROBIT	0.1836	0.1817	0.1827	0.3824	0.3548	0.3092
	DFM	0.2762	0.2559	0.2367	0.2920	0.2794	0.2546

with, in each of these tables we list models with the specific test associated with that model in parentheses. In both tables we present proportions of p-values that exceed or fall below some important threshold. We begin with the overidentification tests in Tables 2.22 and 2.23, for which the null is that the overidentifying restrictions are valid (i.e. that the specification considered is valid).⁸ Since the identifying

⁸Recall that the overidentification test statistic for the bivariate probit model is simply the χ^2 statistic for a test of the joint significance of the instruments in the marginal probit equation for Y_2 under the “just identified” specification under which the instruments appear in both marginal probit equations and identification rests on nonlinearity from functional form (i.e. joint normality) alone. The null hypothesis of such a test is that the instruments are not jointly significant regressors

Table 2.16 Mean ATE for true ATE = 0.2, error correlation = 0.3, $Y_1 = 0.5$ and $Y_2 = 0.5$

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
Obs = 1,000							
	LPM	0.3752	0.3701	0.3707	0.4197	0.4194	0.4146
	Probit	0.3710	0.3657	0.3665	0.4022	0.4020	0.3985
	TSLS	0.2350	0.2126	0.2262	0.2729	0.2709	0.2804
	LIML	0.2255	0.2051	0.2229	0.2584	0.2639	0.2769
	GMM	0.2357	0.2135	0.2269	0.2747	0.2729	0.2823
	Residual1	0.2155	0.1894	0.2006	0.2235	0.2195	0.2345
	Residual2	0.2185	0.1941	0.2034	0.0842	0.1343	0.1948
	BIPROBIT	0.2140	0.1949	0.2085	0.1373	0.1765	0.2229
	DFM	0.1853	0.1671	0.1668	0.2149	0.2113	0.2298
Obs = 5,000							
	LPM	0.3745	0.3740	0.3727	0.3929	0.3927	0.3921
	Probit	0.3702	0.3698	0.3686	0.3830	0.3827	0.3825
	TSLS	0.1848	0.1780	0.2060	0.2153	0.2274	0.2348
	LIML	0.1716	0.1707	0.2026	0.2031	0.2206	0.2317
	GMM	0.1849	0.1780	0.2059	0.2151	0.2277	0.2350
	Residual1	0.1651	0.1532	0.1786	0.1800	0.1896	0.1963
	Residual2	0.1855	0.1813	0.1892	−0.1205	−0.0773	0.0004
	BIPROBIT	0.1597	0.1647	0.1842	−0.0701	−0.0033	0.0731
	DFM	0.2231	0.2071	0.1977	0.1265	0.1244	0.1334
Obs = 10,000							
	LPM	0.3794	0.3792	0.3788	0.4068	0.4058	0.4063
	Probit	0.3748	0.3746	0.3743	0.3934	0.3925	0.3931
	TSLS	0.1182	0.1547	0.1773	0.2124	0.2121	0.2165
	LIML	0.1038	0.1462	0.1734	0.1984	0.2039	0.2125
	GMM	0.1193	0.1534	0.1773	0.2125	0.2123	0.2175
	Residual1	0.1069	0.1300	0.1491	0.1916	0.1857	0.1810
	Residual2	0.1804	0.1819	0.1804	−0.1350	−0.1085	−0.0555
	BIPROBIT	0.1377	0.1580	0.1664	−0.0241	0.0108	0.0610
	DFM	0.2372	0.2437	0.2148	0.1561	0.1563	0.1609

restrictions in our Monte Carlo experiments are indeed valid by construction, large test statistics (and accompanying low p-values) would be cause for concern. Tables 2.22 and 2.23 thus present the percentage of p-values that are in the concerning range (i.e. below the conventional cutoff level of 0.1). Interestingly, here all three linear instrumental variables models appear to do quite well.

The same cannot be said for the bivariate probit model, which produces large test statistics (as evidenced by low p-values) alarmingly often. Its performance appears

in marginal probit equation for Y_2 (i.e. that they are legitimately excluded from the marginal probit equation for Y_2).

Table 2.17 Basic summary regressions

Method	All models		Normal errors		Non-normal errors	
	Coeff.	T-statistic	Coeff.	T-statistic	Coeff.	T-statistic
Probit	−0.0038	−0.71	−0.0054	−1.05	−0.0023	−0.28
TSLS	0.0361	6.76	0.0397	7.81	0.0325	4.03
LIML	0.0450	8.42	0.0490	9.64	0.0410	5.07
GMM	0.0363	6.79	0.0399	7.85	0.0326	4.04
Residual1	0.0361	6.76	0.0393	7.72	0.0330	4.08
Residual2	0.0409	7.64	−0.0201	−3.96	0.1018	12.6
BIPROBIT	0.0402	7.52	−0.0128	−2.52	0.0933	11.54
DFM	−0.0242	−4.52	−0.0041	−0.81	−0.0442	−5.47
N	2,592		1,296		1,296	

to improve considerably with larger sample sizes and joint normality of errors. On the whole, however, using an overidentification test which relies on the non-linearity of the bivariate probit model is of limited usefulness.

Turning to Tables 2.24 and 2.25 and the endogeneity test results, we now consider the proportion of the time that the test statistic yields a large p-value. This is once again natural and fitting since the null hypothesis in these tests is exogeneity. A small test statistic (and accompanying large p-value) would thus be cause for concern since endogeneity is present in our models by design (and therefore the null should be rejected). We consider the proportion of p-values that exceed 0.1. Here the results are generally far less reassuring. The performance of endogeneity tests in the linear instrumental variables models is poor,⁹ with particularly misleading results in the case of the Hausman test. The performance of the bivariate probit model, Residual1 and Residual2 are not much better. In general, these results suggest that conventional endogeneity tests are more or less completely unreliable, at least in terms of conventional benchmark p-value thresholds when both dependent variables are binary.

2.5 Empirical Examples

We present two empirical examples based on data sets from Bangladesh and Tanzania that have been previously analyzed by [Chen and Guilkey \(2003\)](#) and [Guilkey and Hutchinson \(2011\)](#). The models that we use in this paper are highly simplified compared to those presented in the original papers. However, they are sufficiently detailed to provide a good comparison of the methods and to demonstrate the pitfalls that one might encounter in analyzing similar problems.

⁹We refer to the Wu-Hausman test ([Wu 1974](#); [Hausman 1978](#)) simply as “Wu” in Tables 2.24 and 2.25.

Table 2.18 Summary regressions stratified by error correlation

Method	Normal errors				Non-normal errors			
	Error correlation 0.1		Error correlation 0.3		Error correlation 0.1		Error correlation 0.3	
	Coeff.	T-statistic	Coeff.	T-statistic	Coeff.	T-statistic	Coeff.	T-statistic
Probit	-0.0063	-1.1	-0.0044	-0.72	-0.0029	-0.29	-0.0016	-0.15
TSLs	0.0906	15.77	-0.0111	-1.81	0.0867	8.55	-0.0216	-1.96
LIML	0.0992	17.28	-0.0012	-0.19	0.0950	9.37	-0.0131	-1.19
GMM	0.0908	15.8	-0.0109	-1.78	0.0867	8.54	-0.0214	-1.94
Residual1	0.0895	15.58	-0.0109	-1.79	0.0856	8.44	-0.0197	-1.79
Residual2	0.0304	5.3	-0.0707	-11.58	0.1711	16.87	0.0326	2.96
BIPROBIT	0.0372	6.47	-0.0628	-10.29	0.1623	16	0.0242	2.2
DFM	0.0453	7.89	-0.0536	-8.77	0.0034	0.33	-0.0918	-8.34
N	648		648		648		648	

Table 2.19 Summary regressions stratified by instrument strength

Method	Normal errors				Non-normal errors			
	Instrument strength 15		Instrument strength 50		Instrument strength 15		Instrument strength 50	
	Coeff.	T-statistic	Coeff.	T-statistic	Coeff.	T-statistic	Coeff.	T-statistic
Probit	-0.0055	-0.73	-0.0052	-0.88	-0.0023	-0.18	-0.0021	-0.17
TSLS	0.0878	11.64	-0.0085	-1.44	0.0792	6.16	-0.0131	-1.06
LIML	0.1070	14.19	-0.0065	-1.11	0.0958	7.45	-0.0110	-0.89
GMM	0.0881	11.68	-0.0084	-1.42	0.0792	6.16	-0.0129	-1.04
Residual1	0.0788	10.45	-0.0028	-0.47	0.0707	5.5	-0.0057	-0.46
Residual2	-0.0073	-0.97	-0.0352	-5.96	0.1377	10.71	0.0599	4.84
BIPROBIT	0.0046	0.61	-0.0328	-5.55	0.1250	9.72	0.0562	4.54
DFM	0.0059	0.79	-0.0168	-2.84	-0.0406	-3.16	-0.0486	-3.93
N	432		432		432		432	

Table 2.20 Summary regressions by error correlation and instrument strength

Method	Error correlation 0.1			Error correlation 0.3		
	Instrument strength 15		Instrument strength 50	Instrument strength 15		Instrument strength 50
	Coeff.	T-statistic		Coeff.	T-statistic	
Probit	-0.0048	-0.42	-0.0044	-0.0030	-0.27	-0.0029
TSLS	0.1362	11.84	0.0414	0.0309	2.69	-0.0630
LIML	0.1530	13.3	0.0434	0.0498	4.35	-0.0609
GMM	0.1361	11.84	0.0415	0.0312	2.72	-0.0627
Residual1	0.1263	10.98	0.0468	0.0232	2.03	-0.0553
Residual2	0.1259	10.95	0.0711	0.0044	0.39	-0.0464
BIPROBIT	0.1227	10.67	0.0724	0.0068	0.6	-0.0489
DFM	0.0301	2.61	0.0168	-0.0647	-5.65	-0.0822
N	432		432	432		432

Table 2.21 Summary regressions with controls for all experimental features

Method	All models		Normal errors		Non-normal errors	
	Coeff.	T statistic	Coeff.	T statistic	Coeff.	T statistic
Probit	−0.0038	−0.84	−0.0054	−1.39	−0.0023	−0.33
TSLS	0.0361	7.95	0.0397	10.32	0.0325	4.82
LIML	0.0450	9.9	0.0490	12.73	0.0410	6.07
GMM	0.0363	7.99	0.0399	10.37	0.0326	4.84
Residual1	0.0361	7.95	0.0393	10.2	0.0330	4.89
Residual2	0.0409	8.99	−0.0201	−5.23	0.1018	15.1
BIPROBIT	0.0402	8.85	−0.0128	−3.33	0.0933	13.82
DFM	−0.0242	−5.32	−0.0041	−1.07	−0.0442	−6.55
Sample size 5,000	−0.0019	−0.74	−0.0099	−4.47	0.0061	1.55
Sample size 10,000	−0.0013	−0.48	−0.0154	−6.93	0.0129	3.31
Instrument strength 25	−0.0271	−10.31	−0.0257	−11.55	−0.0284	−7.3
Instrument strength 50	−0.0577	−21.98	−0.0546	−24.55	−0.0607	−15.59
Error correlation 0.3	0.0213	9.93	0.0252	13.86	0.0174	5.47
Treatment effect 0.2	0.0038	1.75	0.0001	0.04	0.0074	2.34
Program prevalence 0.25	0.0251	11.7	0.0115	6.35	0.0386	12.14
Outcome prevalence 0.25	−0.0265	−12.35	−0.0164	−9.02	−0.0365	−11.49
Non-normal errors	0.0239	11.14				
N	2,592		1,296		1,296	

That said, an important consideration to remember now that we have moved from simulations (for which we control all parameters) to applications with real world samples is that heterogeneous treatment effects (which were not considered in the simulations) may be at play and driving differences in estimates.

In Bangladesh, we use data that was gathered to examine how self-exposure to the Smiling Sun multimedia communication campaign in rural Bangladesh impacted women's use of modern contraception (more details and more extensive models are to be found in [Guilkey and Hutchinson \(2011\)](#)). The Smiling Sun communication program, launched in Bangladesh in 2001, was a multi-channel campaign with the objectives of establishing the Smiling Sun symbol, disseminating important health-related messages, and promoting health services in urban and rural areas at Paribarik Shastha Clinics (Family Health Clinics) operated by the NGO Service Delivery Program (for which the Smiling Sun served as a logo). The campaign involved a 26-episode television drama serial 'Eyi Megh Eyi Roudro' ("Now cloud, now sunshine"), television advertisements, radio spots, posters, billboards, press ads in daily newspapers and local publicity efforts.

The data were collected roughly at the beginning of the Smiling Sun campaign in 2001 and then again 2 years later. Questions were asked of women of reproductive age about whether they had seen the Smiling Sun logo and, if so, whether they had seen it in a television drama, in a television advertisement, on the radio, on a billboard, at a signboard at a clinic, or elsewhere. In the original paper, we examined the impact of recall for each source separately. In the simplified model used here, all sources are combined into a single binary indicator for exposure to the program.

Table 2.22 Identification tests for true ATE = 0.2, error correlation = 0.3, $Y_1 = 0.25$ and $Y_2 = 0.25$: proportion of times that the p-value for the test statistic is less than 0.1

	Normal errors			Non-normal errors		
	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000						
TSLs_Sargan	0.0900	0.1010	0.0980	0.1010	0.0950	0.0940
TSLs_Basermann	0.0900	0.1000	0.0980	0.1010	0.0950	0.0930
LIML_AndersonRubin	0.0880	0.0990	0.0980	0.1010	0.0950	0.0950
LIML_Basermann	0.0880	0.0970	0.0980	0.0950	0.0930	0.0920
GMM_Hansen	0.0840	0.1000	0.0980	0.0990	0.0920	0.0880
BIPROBIT	0.2748	0.2890	0.2821	0.2624	0.3361	0.3810
N = 5,000						
TSLs_Sargan	0.1080	0.0910	0.1190	0.0930	0.1040	0.0990
TSLs_Basermann	0.1080	0.0910	0.1190	0.0930	0.1040	0.0990
LIML_AndersonRubin	0.1030	0.0910	0.1190	0.0910	0.1020	0.0990
LIML_Basermann	0.1010	0.0910	0.1180	0.0910	0.1020	0.0990
GMM_Hansen	0.1070	0.0890	0.1200	0.0930	0.1030	0.0970
BIPROBIT	0.1416	0.1249	0.1263	0.2385	0.2613	0.3340
N = 10,000						
TSLs_Sargan	0.1100	0.1080	0.1180	0.1000	0.1240	0.1200
TSLs_Basermann	0.1100	0.1080	0.1180	0.1000	0.1240	0.1200
LIML_AndersonRubin	0.1080	0.1060	0.1180	0.0980	0.1200	0.1200
LIML_Basermann	0.1080	0.1060	0.1180	0.0980	0.1200	0.1200
GMM_Hansen	0.1080	0.0960	0.1200	0.1000	0.1240	0.1180
BIPROBIT	0.1172	0.0984	0.1240	0.2360	0.3026	0.4140

We pool the data from 2001 and 2003 in the analysis. Descriptive statistics and variable definitions are presented in Table 2.26. There are three exclusion restrictions in the current use of contraception equation: the last three variables that indicate the number of Smiling Sun posters in clinics that are within 1 km of the sample cluster and whether or not the household owns a TV and radio (two separate indicators).

In Tanzania, we use data gathered over a 9-year period for the purpose of evaluating that nation's National Population Policy (NPP). The NPP began in 1992 and was developed to address a very high total fertility rate of about 6.3 children (Ngallaba et al. 1993) and an under five mortality rate of 141 per 1,000 live births. The NPP had substantial funding from donor agencies including the United States Agency for International Development (USAID).

The main USAID program in Tanzania for family planning was the Family Planning Support System (FPSS) project. The major components of the program were to train health providers in the provision of family planning, to provide logistical support for the provision of family planning supplies and to develop an information, education and communication (IEC) program to promote family planning. This program ended in 1999 and cross sectional data were gathered in 1991, 1994, 1996, and 1999 to evaluate its impact. Chen and Guilkey (2003)

Table 2.23 Identification tests for true $ATE = 0.2$, error correlation = 0.3, $Y_1 = 0.5$ and $Y_2 = 0.5$: proportion of times that the p-value for the test statistic is less than 0.1

	Normal errors			Non-normal errors		
	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000						
TSLS_Sargan	0.0880	0.1000	0.0900	0.0910	0.1070	0.1210
TSLS_Basmann	0.0860	0.1000	0.0890	0.0910	0.1070	0.1200
LIML_AndersonRubin	0.0840	0.1000	0.0900	0.0910	0.1070	0.1210
LIML_Basmann	0.0820	0.0990	0.0880	0.0900	0.1060	0.1200
GMM_Hansen	0.0900	0.0970	0.0890	0.0950	0.1050	0.1190
BIPROBIT	0.2791	0.3252	0.3439	0.3141	0.3033	0.2464
N = 5,000						
TSLS_Sargan	0.0970	0.0890	0.0990	0.0790	0.0820	0.0900
TSLS_Basmann	0.0970	0.0890	0.0990	0.0790	0.0810	0.0900
LIML_AndersonRubin	0.0930	0.0890	0.0990	0.0780	0.0800	0.0900
LIML_Basmann	0.0920	0.0870	0.0990	0.0780	0.0800	0.0880
GMM_Hansen	0.0990	0.0900	0.1010	0.0790	0.0840	0.0920
BIPROBIT	0.1518	0.1436	0.1506	0.4598	0.5172	0.5518
N = 10,000						
TSLS_Sargan	0.0960	0.0980	0.1060	0.0860	0.1040	0.0900
TSLS_Basmann	0.0960	0.0980	0.1060	0.0860	0.1040	0.0900
LIML_AndersonRubin	0.0940	0.0980	0.1040	0.0860	0.1040	0.0900
LIML_Basmann	0.0940	0.0980	0.1040	0.0860	0.1040	0.0900
GMM_Hansen	0.1000	0.1000	0.1020	0.0860	0.1040	0.1060
BIPROBIT	0.1443	0.1403	0.1506	0.2385	0.2780	0.3560

provide a comprehensive evaluation of the program's impact. In this example, we estimate the impact of having heard a family message from any source on current contraceptive use. The summary statistics for the sample are found in Table 2.27.

We estimated the two equation models, one for self-reported exposure to a message and one for current contraceptive use, using all nine methods that were evaluated in the Monte Carlo experiments. The results of the Monte Carlo experiments suggest that the overidentification tests were reasonably reassuring while the endogeneity tests were highly inaccurate. The overidentification tests in Bangladesh for 2SLS, LIML, GMM all fail to reject the null hypothesis that the exclusion restrictions are valid, the desired result. The results for these tests for Tanzania were mixed: the p-values were 0.09, 0.09, and 0.21 for 2SLS, LIML, and GMM respectively and so there is weak evidence to support the null. When we included the excluded variables in the BIPROBIT models and tested to see if these variables had direct effects on contraceptive use, we found that two of three exclusion restrictions were valid for Bangladesh while both exclusion restrictions were valid for Tanzania. Since the DFM is also identified without exclusion restrictions, we performed the same test using DFM and found that none of the excluded variables had direct effects on contraceptive use. Thus, the evidence seems to suggest that the models are identified.

Table 2.24 Endogeneity tests for true ATE = 0.2, error correlation = 0.3, $Y_1 = 0.25$ and $Y_2 = 0.25$: proportion of times that the p-value for the test statistic is greater than 0.1

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000							
	TSLS_Durban	0.8630	0.8010	0.7140	0.8180	0.7930	0.6280
	TSLS_Wu	0.8630	0.8020	0.7160	0.8180	0.7950	0.6280
	LIML_Hausman	1.0000	1.0000	0.9900	1.0000	1.0000	0.9810
	GMM_Hayashi	0.8610	0.8000	0.7060	0.8180	0.7880	0.6240
	Residual1	0.8710	0.8180	0.7300	0.8230	0.8000	0.6440
	Residual2	0.8330	0.7770	0.6860	0.8970	0.8740	0.7180
	BIPROBIT	0.8499	0.7658	0.6690	0.7886	0.8006	0.6640
N = 5,000							
	TSLS_Durban	0.8750	0.8420	0.7490	0.7860	0.7290	0.5740
	TSLS_Wu	0.8750	0.8430	0.7510	0.7860	0.7290	0.5740
	LIML_Hausman	1.0000	1.0000	0.9940	1.0000	0.9970	0.9750
	GMM_Hayashi	0.8780	0.8390	0.7490	0.7830	0.7280	0.5720
	Residual1	0.8770	0.8410	0.7690	0.7910	0.7280	0.5820
	Residual2	0.6410	0.6140	0.5590	0.9010	0.8840	0.7870
	BIPROBIT	0.7570	0.7090	0.6020	0.8400	0.8220	0.7230
N = 10,000							
	TSLS_Durban	0.8500	0.7680	0.6740	0.7780	0.7280	0.5860
	TSLS_Wu	0.8500	0.7680	0.6740	0.7780	0.7300	0.5860
	LIML_Hausman	1.0000	0.9980	0.9780	1.0000	0.9920	0.9620
	GMM_Hayashi	0.8460	0.7600	0.6660	0.7720	0.7220	0.5840
	Residual1	0.8460	0.7780	0.6780	0.7740	0.7260	0.6020
	Residual2	0.4940	0.4660	0.3600	0.8720	0.8940	0.8980
	BIPROBIT	0.6300	0.5680	0.4300	0.8040	0.8540	0.8480

We also performed tests of the null hypothesis that having heard a family planning message is exogenous. The results across the two data sets were consistent for 2SLS, GMM, Residual1, Residual2, and BIPROBIT: the null hypothesis was strongly rejected for Tanzania and the tests failed to reject the null hypothesis in Bangladesh. We did not perform a formal endogeneity test for DFM. However, for both data sets, the DFM yielded highly significant heterogeneity parameters using a four point of support model (the same number of points of support employed in the Monte Carlo experiments) and, as can be seen in the tables below, in both samples the point estimate of the ATE is quite different for the DFM and models that do not correct for endogeneity.

Table 2.28 presents the estimated ATE's across all nine methods along with standard errors. The ATE's and standard errors are drawn directly from the regression results for the linear models while the STATA margins command was used to obtain the ATE and standard errors for all non-linear models except DFM. The standard errors for the DFM model were obtained by using a parametric bootstrap procedure with 10,000 replications using a FORTRAN program.

Table 2.25 Endogeneity tests for true ATE = 0.2, error correlation = 0.3, $Y_1 = 0.5$ and $Y_2 = 0.5$: proportion of times that the p-value for the test statistic is greater than 0.1

		Normal errors			Non-normal errors		
		$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$	$\chi^2 = 15$	$\chi^2 = 25$	$\chi^2 = 50$
N = 1,000							
	TSLS_Durban	0.8500	0.7980	0.7040	0.8620	0.8110	0.7330
	TSLS_Wu	0.8510	0.8040	0.7060	0.8630	0.8130	0.7340
	LIML_Hausman	1.0000	0.9980	0.9920	1.0000	0.9980	0.9900
	GMM_Hayashi	0.8430	0.8020	0.7080	0.8600	0.8080	0.7300
	Residual1	0.8440	0.7980	0.7080	0.8510	0.8090	0.7380
	Residual2	0.8410	0.7820	0.6870	0.5760	0.5860	0.5680
	BIPROBIT	0.8510	0.7800	0.6680	0.7287	0.6740	0.6210
N = 5,000							
	TSLS_Durban	0.8010	0.7340	0.6510	0.8240	0.7850	0.6910
	TSLS_Wu	0.8010	0.7350	0.6510	0.8240	0.7850	0.6920
	LIML_Hausman	1.0000	0.9960	0.9840	1.0000	1.0000	0.9820
	GMM_Hayashi	0.8000	0.7280	0.6540	0.8250	0.7840	0.6870
	Residual1	0.8020	0.7350	0.6510	0.8220	0.7790	0.6930
	Residual2	0.6870	0.6450	0.5670	0.0110	0.0230	0.0350
	BIPROBIT	0.7400	0.6760	0.5760	0.0864	0.0590	0.0720
N = 10,000							
	TSL_Durban	0.6960	0.7080	0.5680	0.8040	0.7220	0.6060
	TSLS_Wu	0.6960	0.7080	0.5680	0.8040	0.7240	0.6060
	LIML_Hausman	1.0000	0.9960	0.9680	1.0000	0.9940	0.9720
	GMM_Hayashi	0.7000	0.7080	0.5640	0.8060	0.7240	0.6140
	Residual1	0.6980	0.7280	0.5740	0.8340	0.7440	0.6380
	Residual2	0.6060	0.5280	0.4280	0.0000	0.0000	0.0020
	BIPROBIT	0.6640	0.5960	0.4540	0.0020	0.0040	0.0040

There is a fairly wide range in estimated ATEs across methods. For Bangladesh, the DFM has the largest estimated ATE but also the largest standard error. We also see that LPM and all the methods that assume normality give similar estimated ATE's while the three instrumental variables methods give results between these methods and DFM. For Tanzania, DFM and the three instrumental variables methods give very consistent results with estimated ATE's approximately double what is found for the two methods that do not correct for endogeneity (LPM and Probit). The residual inclusion methods yield similar point estimates for the ATE as BIPROBIT which falls above the methods that do not correct for endogeneity and below the DFM and the instrumental variables methods.

Given the results of the Monte Carlo experiments, one would probably place the most confidence in the results obtained for the DFM followed by the instrumental variables methods. None of these methods rely on the assumption of normality for the error distributions in models and the results of these methods are highly consistent for Tanzania and least somewhat consistent for Bangladesh.

Table 2.26 Descriptive statistics for Bangladesh ($N = 21,472$)

Variable	Mean	Standard dev.
Endogenous variables		
Current user of contraception	0.458	0.498
Recall smiling sun message	0.223	0.417
Exogenous variables		
Woman age 20–24	0.179	0.383
Woman age 25–29	0.178	0.383
Woman age 30–34	0.169	0.375
Woman age 35–39	0.140	0.347
Woman age 40–44	0.111	0.314
Woman age 45–49	0.072	0.259
Woman has primary education	0.248	0.432
Woman has secondary education	0.180	0.385
Husband has primary education	0.190	0.392
Husband has secondary education	0.243	0.429
Husband has college education	0.020	0.141
Sum of the number of contraceptive methods available within 1 km	1.318	2.333
Indicator for 2003 survey	0.406	0.491
Number of facilities within 1 km with smiling sun posters	0.305	0.542
Household has a radio	0.305	0.461
Household has a television	0.142	0.349

2.6 Conclusion

We conclude with some thoughts regarding the pattern of results presented in Sect. 2.4. We first note that, when error correlation and instrument strength are low, the models that we consider that attempt explicitly to correct for the endogeneity of a binary regressor do not seem to perform as well as alternatives that simply ignore potential endogeneity. Even BIPROBIT, for which identification ultimately rests on the assumption of jointly normal errors in Eqs. (2.1) and (2.2) does not perform as well as LPM under circumstances of weak error correlation and weak instruments, even when the true error distribution is bivariate normal.

As either instrument strength or error correlation increases, our findings suggest that the researcher has attractive options relative to the simple methods. As expected, BIPROBIT performs well under these circumstances when the true error distribution is bivariate normal. However, Residual2 performs as well or is even slightly better than BIPROBIT. In addition, even when the true error distribution is bivariate normal, DFM represents a significant improvement over LPM and performs only slightly worse than Residual2 and BIPROBIT. When the true error distribution is non-normal, DFM dominates all other estimators. The only estimation methods that come close are the linear instrumental variables estimators, which are also robust to non-normal errors. However, these estimators only approach but do not equal the

Table 2.27 Descriptive statistics for Tanzania ($N = 17,724$)

Variable	Mean	Standard dev.
Endogenous variables		
Current user of contraception	0.115	0.319
Recall family planning message	0.387	0.487
Exogenous variables		
Woman age 15–19	0.221	0.415
Woman age 20–24	0.196	0.397
Woman age 25–29	0.174	0.379
Woman age 30–34	0.132	0.338
Woman age 35–39	0.113	0.317
Woman age 40–44	0.086	0.280
Woman 1–6 years of education	0.219	0.413
Woman 7 years of education	0.411	0.492
Woman 8 or more years of education	0.017	0.128
Partner 1–6 years of education	0.170	0.376
Partner 7 years of education	0.274	0.446
Partner 8 or more years of education	0.042	0.201
Number of contraceptive methods seen in stock in facilities within 5 km	1.229	1.702
Household owns a radio	0.347	0.476
Household owns a television	0.002	0.044

Table 2.28 Estimated average treatment effects and standard errors for the two empirical examples

Method	Bangladesh		Tanzania	
	ATE	SE	ATE	SE
LPM	0.0669	0.0082	0.0700	0.0050
Probit	0.0699	0.0082	0.0676	0.0050
TSLS	0.0843	0.0376	0.1327	0.0224
LIML	0.0843	0.0376	0.1329	0.0224
GMM	0.0841	0.0377	0.1320	0.0236
Residual1	0.0840	0.0375	0.1231	0.0243
Residual2	0.0508	0.0358	0.1181	0.0236
BIPROBIT	0.0508	0.0358	0.1178	0.0229
DFM	0.1188	0.0545	0.1361	0.0440
N	21,472		17,724	

performance of DFM when both error correlation and instrument strength are high. Nonetheless, they are a reasonable option for researchers using standard statistical packages (at least until an implementation of the DFM becomes available within one of these packages, a project on which the authors have now embarked with STATA).

The superior performance of the DFM and, to some extent, the linear instrumental variables estimators when the true error distribution is non-normal is even more impressive when one considers that the design of our experiments

involving non-normal errors was likely comparatively favorable to models that assume normality compared with real world circumstances: our approach to non-normal errors still retained the unimodality of the joint distribution of the errors and of the surface of its joint distribution in \mathbf{R}^3 . In some sense this likely gave even those models explicitly motivated by joint normality some fighting chance for reasonable fit to the data. Real world circumstances will likely involve multi-modal distributions, reflecting the presence of combinations of pronounced “types” within the population. Nonetheless, our results suggest that methods that rely on the assumption of normally distributed errors are a poor choice relative to the more robust methods considered in this paper even in the unimodal case. In that sense they echo the concerns about the fragility of identification by functional form that have been in the literature in various contexts for nearly three decades (e.g. [LaLonde 1986](#); [Manning et al. 1987](#)).

In terms of practical advice for applied researchers, our results thus do suggest some guidelines. First, less parametric methods, including linear instrumental variables models if not the DFM itself (the estimation of which is, for the moment, impractical for most) are preferable to methods that rely on more parametric assumptions for the joint error distribution: joint normality assumptions work out particularly well only when the errors are indeed jointly normal (and even then only when instrument strength and error correlation were high), and this is likely a heroic assumption in many applied microeconomic applications. Even when the bivariate probit performs well, it does not necessarily significantly outperform simpler methods (such as `Residual2`) that also implicitly rely on joint normality. Put slightly differently, it is not clear that the explicit functional form assumption of the bivariate probit model is buying the user much in terms of performance, even under ideal circumstances.

Second, and perhaps intuitively unsurprisingly in light of the evidence regarding weak instruments in the setting of continuous outcomes of interest and endogenous variables (e.g. [Stock and Staiger 1997](#); [Bound et al. 1995](#)), instrument strength matters. Indeed, even in the case of models relying on joint normality assumptions for the errors when the errors are actually jointly non-normal, increases in instrument strength yield performance benefits. Moreover, it is straightforward to assess instrument strength.

Overidentification tests proved reasonably reliable in the binary outcome and binary endogenous variable setting even with linear instrumental variables based tests. We can offer far less guidance regarding the other key indicator of likely model performance, tests of endogeneity. Unfortunately, we cannot even say with any confidence that formal endogeneity tests are, in this setting and with the currently available set of conventional tests, necessarily any more reliable than informed theoretical assumptions by applied researchers.

As for future work in the methodological arena, it is clear that nonparametric full-information maximum likelihood approaches hold great potential promise. Much remains to be done in this area, including the introduction of routines for estimating these models as part of standard statistical packages such as STATA, continued

improvement of estimation methodology and the development and refinement of tests (such as a formal endogeneity test). Finally, model performance in circumstances of heterogeneous treatment effects is now under consideration by the authors in a follow-up to this manuscript.

References

- Anderson T, Rubin H (1950) The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Ann Math Stat* 21:570–582
- Angrist J, Krueger A (2001) Instrumental variables and the search for identification: from supply and demand to natural experiments. *J Econ Perspect* 15:69–85
- Angrist J, Pischke J (2009) Mostly harmless econometrics: an empiricist's companion. Princeton University Press, Princeton
- Babalola S (2005) Communication, ideation and contraceptive use in Burkina Faso: an application of the propensity score matching method. *J Fam Plan Reprod Health Care* 31:207–212
- Bassman R (1960) On finite sample distributions of generalized classical linear identifiability test statistics. *J Am Stat Assoc* 55:650–659
- Bauman K, Viadro C, Tsui A (1993) Family planning program effects in developing countries: conclusions and related considerations. The evaluation project working paper IM-03-03
- Bollen K, Guilkey D, Mroz T (1995) Binary outcomes and endogenous explanatory variables: tests and solutions with an application to the demand for contraceptive use in Tunisia. *Demography* 32:111–131
- Bound J, Jaeger D, Baker R (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc* 90:443–450
- Cappellari L, Jenkins S (2003) Multivariate probit regression using simulated maximum likelihood. *STATA J* 3:278–294
- Chen S, Guilkey D (2003) Determinants of contraceptive method choice in rural Tanzania between 1991 and 1999. *Stud Fam Plan* 34:263–276
- Chiburis RC, Das J, Lokshin M (2011) A practical comparison of the bivariate probit and linear IV estimators. The World Bank Policy research working paper 5601
- Durbin J (1954) Errors in variables. *Rev Int Stat Inst* 22:23–32
- Fleishman A (1978) A method for simulating nonnormal distributions. *Psychometrika* 43:521–532
- Gourieroux C, Monfort A, Renault E, Trognon A (1987) Generalized residuals. *J Econom* 34:5–32
- Guilkey D, Hutchinson P (2011) Overcoming methodological challenges in evaluating health communication campaigns: evidence from rural Bangladesh. *Stud Fam Plan* 42:93–106
- Guilkey D, Mroz T, Taylor L (1992) Estimation and testing in simultaneous equations models with discrete outcomes using cross section data. UNC-CH Department of Economics working paper
- Guilkey D, Hutchinson P, Lance P (2006) Cost effectiveness analysis for health communications programs. *J Health Commun* 11:47–67
- Hansen L (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–1054
- Hausman J (1978) Specification tests in econometrics. *Econometrica* 46:1251–1271
- Hayashi F (2000) *Econometrics*. Princeton University Press, Princeton
- Heckman J, Singer B (1984) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52:271–320
- Hutchinson P, Wheeler J (2006). The cost effectiveness of health communication programs: what do we know? *J Health Commun* 11:7–45
- Imbens G, Angrist J (1994) Identification and estimation of local average treatment effects. *Econometrica* 62:467–475

- Kaiser H, Dickman K (1962) Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika* 27:179–182
- LaLonde R (1986) Evaluating the econometric evaluations of training programs with experimental data. *Am Econ Rev* 76:604–620
- Manning W, Duan N, Rogers W (1987) Monte Carlo evidence on the choice between sample selection and two-part models. *J Econom* 35:59–82
- Mwaikambo L, Speizer I, Schurmann A, Morgan G, Fikree F (2011) What works in family planning interventions: a systematic review. *Stud Fam Plan* 42:67–82
- Mroz T (1999) Discrete factor approximations in simultaneous equations models: estimating the impact of a dummy endogenous variable on a continuous outcome. *J Econom* 92:233–274
- Ngallaba S, Kapiga S, Ruyoba I, Boerma J (1993) Tanzania demographic and health survey 1991/1992. Macro International Inc., Columbia
- Rivers D, Vuong Q (1988) Limited information estimators and exogeneity tests for simultaneous probit models. *J Econom* 39:347–366
- Sargon J (1958) The estimation of economic relationships using instrumental variables. *Econometrica* 26:393–415
- Stock J, Staiger D (1997) Instrumental variables regression with weak instruments. *Econometrica* 65:557–586
- Terza J, Basu A, Rathouz P (2008) Two-stage residual inclusion estimation: addressing endogeneity in health econometric modelling. *J Health Econ* 27:531–543
- Vale C, Maurelli V (1983) Simulating multivariate nonnormal distributions. *Psychometrika* 48:465–471
- Wu D (1974) Alternative tests of independence between stochastic regressors and disturbances: finite sample results. *Econometrica* 42:529–546

<http://www.springer.com/978-1-4899-8007-6>

Festschrift in Honor of Peter Schmidt

Econometric Methods and Applications

Sickles, R.; Horrace, W.C. (Eds.)

2014, XII, 409 p. 16 illus., 9 illus. in color., Hardcover

ISBN: 978-1-4899-8007-6