Economics Letters 117 (2012) 762–766

Contents lists available at SciVerse ScienceDirect

Economics Letters

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

A practical comparison of the bivariate probit and linear IV estimators

Richard C. Chiburis [a,*], Jishnu Das [b,c], Michael Lokshin [b]

[a] Google Inc., 1600 Amphitheatre Pkwy, Mountain View, CA 94043, United States
[b] World Bank, United States
[c] Centre for Policy Research, India

ARTICLE INFO

Article history:
Received 27 March 2011
Received in revised form
6 July 2012
Accepted 22 August 2012
Available online 31 August 2012

Keywords:
Limited dependent variable
Dummy endogenous regressor
Treatment effects
Bootstrap
Goodness of fit

ABSTRACT

This paper compares asymptotic and finite sample properties of linear IV and bivariate probit in models with an endogenous binary treatment and binary outcome. The results provide guidance on the choice of model specification and help to explain large differences in the estimates depending on the specification chosen.

1. Introduction

This paper compares asymptotic and finite-sample properties of linear IV and bivariate probit in models with an endogenous binary treatment and binary outcome. A motivating example is the effect of private schooling on graduation rates. Here the "treatment" – private school attendance – and the "outcome" – graduation – can take one of two potential values. Comparing mean graduation rates of children in public and private schools likely yields a biased estimate of the causal effect of private schooling on graduation rates if omitted variables, such as ability, are correlated both with private school attendance and graduation rates.

There are two common approaches to estimating causal effects in such models. One approach disregards the binary structure of the outcome and treatment variables and presents linear instrumental variables (IV) estimates of the treatment effect; the second computes maximum-likelihood estimates of a bivariate probit (BP) model, which assumes that the outcome and treatment are each determined by latent linear index models with jointly normal error terms. Sometimes the two approaches can produce markedly different results. A persistent problem in the private schooling literature, for instance, is the large difference between linear IV and BP estimates of the treatment effect (Altonji et al., 2005).

The existing literature offers conflicting advice on the best course of action in empirical problems of this sort. Angrist (1991, 2001) stresses causal effects as opposed to structural parameters in these models and argues for the robustness of the simpler linear IV estimator to the distribution of the error terms. On the other hand, Bhattacharya et al. (2006) suggest that BP is slightly *more* robust than linear IV to non-normality of the error terms.

We compare BP and linear IV in terms of consistent estimation and mean-square error for the average treatment effect across a wide range of model specifications using Monte-Carlo simulations. Our three main contributions are (a) clarifying the relationship between the average treatment effect (ATE) obtained in the BP model with plim (Linear IV), which is also the LATE when there are no covariates in the model; (b) comparing the mean-square error and the actual size and power of tests based on these estimators across a wide range of parameter values; and (c) assessing the power of misspecification tests for BP models.

Based on our results, there are three main messages for practical applications. First, researchers can expect BP and linear IV estimates to differ substantially when treatment probabilities are low or sample sizes are below 5000. Second, confidence intervals should be recovered through bootstrapping for both BP and linear IV, particularly when sample sizes are below 10,000. Third, when BP is the preferred estimator, a score test should be used to check the goodness of fit.

* Corresponding author. Tel.: +1 512 475 8528; fax: +1 512 471 3510.
E-mail addresses: chiburis@austin.utexas.edu, chiburis@google.com (R.C. Chiburis), jdas1@worldbank.org (J. Das).

0165-1765/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2012.08.037

2. Model and estimators

Let $T \in \{0, 1\}$ be a potentially endogenous treatment, and let $Y \in \{0, 1\}$ be the outcome of interest. Let Y_1 be an individual's potential outcome had she received the treatment ($T = 1$), and let Y_0 be the individual's potential outcome had she not received the treatment ($T = 0$). Let $Z \in \{0, 1\}$ be an instrument for the treatment. Let T_1 be an individual's chosen treatment had she been given $Z = 1$, and let T_0 be an individual's chosen treatment had she been given $Z = 0$.

We follow Imbens and Angrist (1994) in defining an instrument Z as satisfying the following conditions:

$$Z \text{ is independent of } (Y_0, Y_1, T_0, T_1) \tag{1}$$

and

$$\mathrm{E}[T \mid Z = 1] \neq \mathrm{E}[T \mid Z = 0]. \tag{2}$$

For each individual, we actually observe (Z, T, Y), where $T = T_Z$ and $Y = Y_T$. Suppose that we have an i.i.d. sample of n individuals. We focus on two commonly estimated treatment effects, defined as follows:

1. The average treatment effect (ATE) over the entire population is given by
$$\Delta_{\mathrm{ATE}} = \mathrm{E}[Y_1] - \mathrm{E}[Y_0]. \tag{3}$$
2. In the model with no covariates, the probability limit of the IV estimator, $\hat{\Delta}^{\mathrm{IV}}$, is what Imbens and Angrist (1994) called the local average treatment effect (LATE):
$$\begin{aligned} p\lim(\text{Linear IV}) &= \Delta_{\mathrm{LATE}} \\ &= \frac{\mathrm{E}[Y \mid Z = 1] - \mathrm{E}[Y \mid Z = 0]}{\mathrm{E}[T \mid Z = 1] - \mathrm{E}[T \mid Z = 0]}. \end{aligned} \tag{4}$$

2.1. Bivariate probit model

Typically it is necessary to impose additional structure on the model to identify Δ_{ATE}. One way to do this while allowing the treatment to be endogenous is to assume a bivariate probit model (Heckman, 1978):

$$\begin{aligned} T &= \mathbf{1}\{\alpha Z + \kappa_T + \varepsilon_1 > 0\} \\ Y &= \mathbf{1}\{\gamma T + \kappa_Y + \varepsilon_2 > 0\} \end{aligned} \tag{5}$$

with $(\varepsilon_1, \varepsilon_2)$ jointly distributed as standard bivariate normal with correlation ρ and independent of Z. Note that assumption (1) follows from this independence condition, and that $\alpha \neq 0$ implies (2). This model can be estimated using maximum likelihood (ML), and the estimated parameters can be plugged into formulas for Δ_{ATE} and Δ_{LATE} to obtain the ML estimators $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{ATE}}$ and $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{LATE}}$, respectively.

In fact, it can be shown (see Chiburis et al., 2011) that the ratio of Δ_{LATE} to Δ_{ATE} will depend on the $\Pr[T = 1]$, $\Pr[Y = 1]$ and the correlation ρ. Specifically, let $p_T = \Pr[T = 1]$, $p_Y = \Pr[Y = 1]$ and let Φ be the standard normal distribution function. Then, a Taylor approximation for the ratio of Δ_{LATE} to Δ_{ATE} in the BP model (see Chiburis et al., 2011) is given by:

$$\frac{\Delta_{\mathrm{LATE}}}{\Delta_{\mathrm{ATE}}} \approx 1 + \rho\Phi^{-1}(p_Y)\Phi^{-1}(p_T). \tag{6}$$

2.2. Asymptotic variances

Because linear IV only consistently estimates Δ_{LATE}, the asymptotic variances of linear IV and maximum-likelihood BP are compared most fairly for the estimation of Δ_{LATE}. When the BP model (5) is correctly specified, $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{LATE}}$ is asymptotically efficient for Δ_{LATE} since ML is asymptotically efficient for any smooth function of the parameters $\boldsymbol{\theta}$. In Chiburis et al. (2011) we compare the asymptotic variances of $\hat{\Delta}^{\mathrm{IV}}$ and $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{LATE}}$ across many different parameter values. The asymptotic variance of $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{LATE}}$ is always lower than that of $\hat{\Delta}^{\mathrm{IV}}$.

Angrist (1991) found that despite the efficiency of BP, the variance of $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{ATE}}$ sometimes exceeds the variance of $\hat{\Delta}^{\mathrm{IV}}$. Angrist's results actually follow from the asymptotics, since for certain values of p_T, p_Y, and ρ, the asymptotic variance of $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{ATE}}$ is higher than that of $\hat{\Delta}^{\mathrm{IV}}$. The key observation is that $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{ATE}}$ and $\hat{\Delta}^{\mathrm{IV}}$ are estimating two different quantities, Δ_{ATE} and Δ_{LATE} respectively, and depending on parameters one can be estimated more precisely than the other. If we are interested in minimizing mean-square error for estimating the ATE, then $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{ATE}}$ will be the better choice (when the BP model is correct) except in rare cases in which $\hat{\Delta}^{\mathrm{IV}}$ has lower asymptotic variance than $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{ATE}}$ *and* Δ_{LATE} happens to be close to Δ_{ATE}.

2.3. Covariates

The IV and BP estimators also easily extend to the case in which we add exogenous regressors X into both equations of (5). Thus far, we assumed that Y is independent of Z. In many applications, the independence assumption may be valid only when additional covariates are included in the model. In this case, the weakest assumption is that Y is independent of Z conditional on X. Then, we can compute LATE(x) for each individual x separately, and the overall LATE is the average of LATE(x) over the distribution of x. However, the LATE thus computed is generally not equal to the plim (linear IV). Since our main contribution is to compare the BP and linear IV in Monte-Carlo simulations to the ATE, we use the terminology plim (linear IV) in the subsequent discussion that includes covariates.

3. Finite-sample simulations

To compare the BP and IV estimators in finite samples and with misspecifications, we conducted Monte-Carlo simulations across a range of parameter values. These parameter values represent a wider selection compared to those used in previous work by Angrist (1991) and Bhattacharya et al. (2006), and prove useful in understanding the performance of these estimators in practical applications. For instance, we find that for some combinations of p_T and p_Y, deviations from normality in the BP model result in significant bias, in contrast to the results of Bhattacharya et al. (2006) over more limited simulations. Also, Angrist's (1991) finding of near-efficiency of IV disappears when we add an exogenous covariate to the model. See Chiburis et al. (2011) for details of the simulation.

3.1. Small-sample bias and variance

Fig. 1 presents simulations of (3), and Fig. 2 presents simulations with a covariate X in both equations. Each figure contains nine subfigures representing different values of $p_T = \Pr[T = 1]$ and $p_Y = \Pr[Y = 1]$. In each subfigure we plot the true Δ_{ATE}, the mean of $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{ATE}}$ and the mean of $\hat{\Delta}^{\mathrm{IV}}$ against sample sizes between 400 and 30,000. We also show the range between the 5th and the 95th percentiles of $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{ATE}}$ and $\hat{\Delta}^{\mathrm{IV}}$. Our figures show results for $\rho = 0.3$, but these results are representative of our simulations with other values of ρ.

Fig. 1 has several noteworthy features. First, $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{ATE}}$ can be *biased* in small samples, as often happens for maximum-likelihood estimators. Even when sample sizes are large, $\hat{\Delta}^{\mathrm{BP}}_{\mathrm{ATE}}$ can be biased under particular extreme combinations of p_T and p_Y—two examples

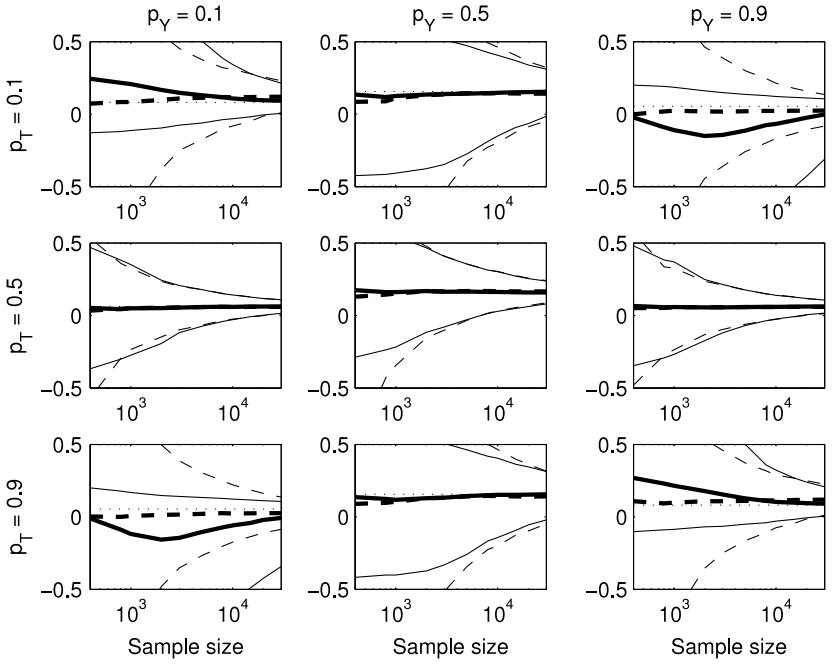

Fig. 1. Spread of BP and IV estimates in simulations with no covariates and $\rho = 0.3$. The area between the thin solid curves represents the range between the 5th and 95th percentiles of the BP estimator, and the area between the thin dashed curves represents the same range for the IV estimator. The thick solid curve is the mean BP estimate, the thick dashed curve is the mean IV estimate, and the dotted line is the true ATE.

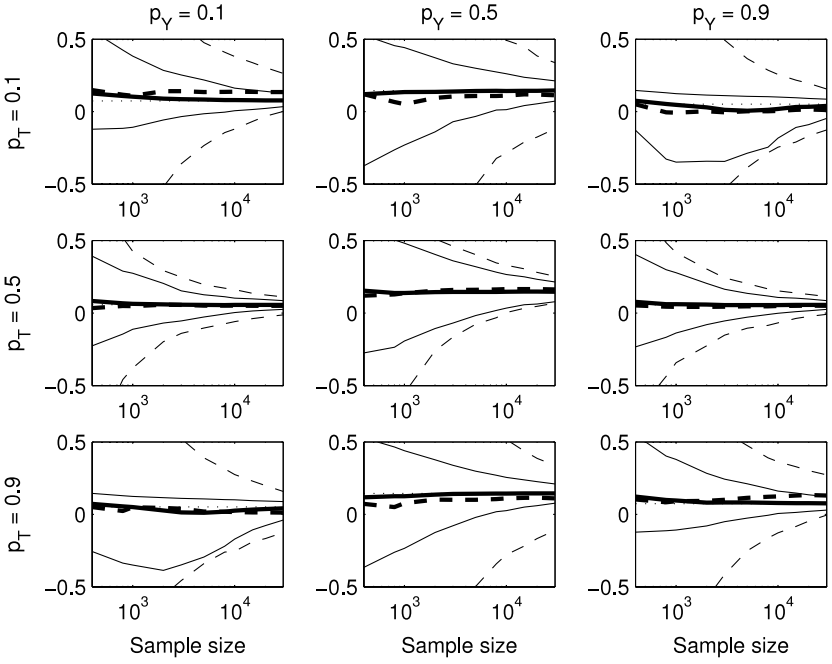

Fig. 2. Spread of BP and IV estimates in simulations with covariate X and $\rho = 0.3$. The area between the thin solid curves represents the range between the 5th and 95th percentiles of the BP estimator, and the area between the thin dashed curves represents the same range for the IV estimator. The thick solid curve is the mean BP estimate, the thick dashed curve is the mean IV estimate, and the dotted line is the true ATE.

are ($p_T = 0.9, p_Y = 0.1$) and ($p_T = 0.1, p_Y = 0.9$). Second, $\hat{\Delta}^{\text{IV}}$ often goes outside the feasible range $[-1, 1]$ in small samples, but $\hat{\Delta}^{\text{BP}}_{\text{ATE}}$ mechanically stays within this range. Hence, $\hat{\Delta}^{\text{BP}}_{\text{ATE}}$ generally outperforms $\hat{\Delta}^{\text{IV}}$ in terms of MSE for Δ_{ATE} for sample sizes under about 5000. For larger samples, either estimator can have lower MSE, depending on the parameters.

Fig. 2 shows that with a covariate in the model $\hat{\Delta}^{\text{BP}}_{\text{ATE}}$ significantly outperforms $\hat{\Delta}^{\text{IV}}$ across *all* of our parameter values. The IV standard errors are often too large for meaningful hypothesis testing, especially when p_T is close to 0 or 1.

3.2. Validity of confidence intervals

Fig. 3 compares nominal 95% confidence intervals based on $\hat{\Delta}^{\text{IV}}$ and $\hat{\Delta}^{\text{BP}}_{\text{ATE}}$ in terms of coverage of Δ_{ATE}. The IV coverage tends to be too high (greater than 95%) for small samples but slowly deteriorates towards zero as the sample size increases and $\hat{\Delta}^{\text{IV}}$ converges to plim (linear IV) rather than Δ_{ATE} (the dashed curve in the figure). Standard BP confidence intervals for Δ_{ATE} display significantly lower coverage than the nominal 95% for sample sizes below 5000, even when the model is correctly specified.

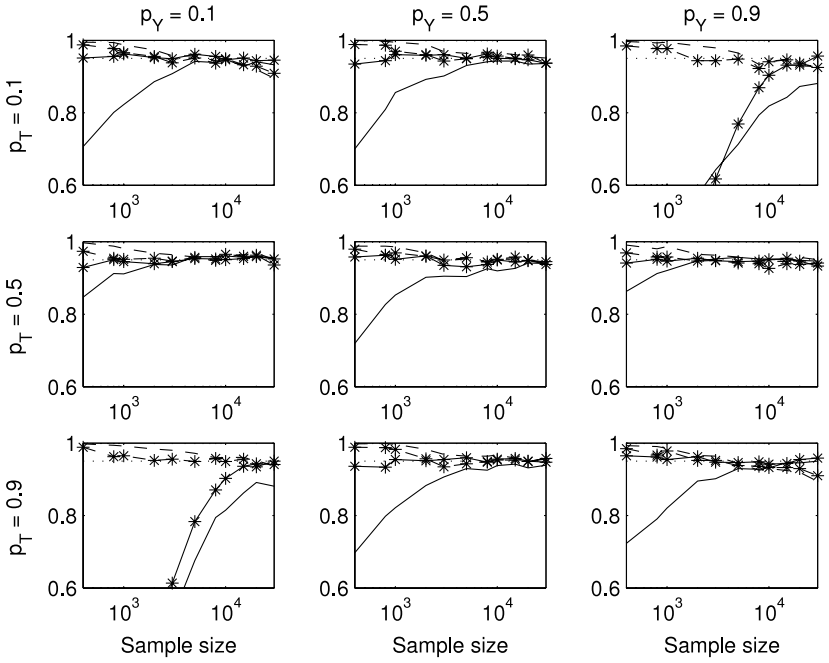

Fig. 3. Coverage of the true Δ_{ATE} for nominal 95% confidence intervals in simulations with normally distributed covariate X_i and $\rho = 0.3$. The solid and dashed curves correspond to the size of tests based on $\hat{\Delta}^{\text{BP}}_{\text{ATE}}$ and $\hat{\Delta}_{\text{IV}}$, respectively. The poor coverage can be improved by bootstrapping the critical values, and the size from bootstrapping for $\hat{\Delta}^{\text{BP}}_{\text{ATE}}$ and $\hat{\Delta}_{\text{IV}}$ is shown by the starred solid and starred dashed curves, respectively.

Fortunately, bootstrapped confidence intervals provide a simple fix for over- and undercoverage in the IV and BP models, respectively. In the bootstrap, we draw with replacement n observations from the data and estimate $\hat{\Delta}^{\text{BP}}_{\text{ATE}}$ and $\hat{\Delta}^{\text{IV}}$ using the new sample. This is repeated many times, and the size-α confidence interval for Δ_{ATE} or plim (linear IV) is reported as the interval between the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the simulated draws of $\hat{\Delta}^{\text{BP}}_{\text{ATE}}$ and $\hat{\Delta}^{\text{IV}}$.

Coverage rates of the bootstrapped BP confidence intervals for Δ_{ATE} are close to the nominal 95%, as shown in Fig. 3. The only exceptions are in small samples in the extreme cases ($p_T = 0.1, p_Y = 0.9$) and ($p_T = 0.9, p_Y = 0.1$), which have been shown in Fig. 1 to be particularly problematic for BP. In addition, Fig. 3 shows that bootstrapping also reduces the overcoverage of IV confidence intervals that we saw in small samples, although it does not prevent undercoverage of IV in large samples because $\hat{\Delta}_{\text{IV}}$ is generally inconsistent for Δ_{ATE}.

3.3. Testing for bivariate probit misspecification

A theme thus far is that the BP estimators are generally more efficient than linear IV, especially when the model specification includes additional covariates. However, when the BP model is misspecified, BP continues to have a lower standard error than IV but can sometimes be severely biased. Fortunately, a Rao score test developed by Murphy (2007) performs well in detecting when the BP model is misspecified and hence BP estimation is inconsistent.[1] We recommend running this test before using BP results. More details on the misspecification bias of BP and simulation results of the specification test are given in Chiburis et al. (2011).

4. Conclusion

We have derived asymptotic results and presented simulations comparing the bivariate probit and linear IV estimators of the average treatment effect of a binary treatment on a binary outcome. Our simulation results provide guidance on the choice of estimation method in practical problems and help explain the difference in the estimated treatment effects depending on the specification chosen. Our findings can be summarized as four main messages for practical applications in empirical models with binary regressors and binary outcome variables:

- Researchers should expect IV and BP estimates to differ substantially when treatment probabilities are close to 0 or 1, or when sample sizes are below 5000. Linear IV estimates are particularly uninformative when treatment probabilities are low, a problem that is accentuated when there are covariates in the model. One recommendation is to present both linear IV and BP estimates when there are covariates in the model, and for the ranges of p_T and p_Y where IV confidence intervals are large.
- Differences between IV and BP estimates can also reflect differences between plim (Linear IV) and ATE.
- Confidence intervals recovered through bootstrapping are a must in these models when sample sizes are below 10,000 and should be preferred to analytical standard errors for all applications.
- Misspecification of the BP model can lead to bias in BP estimates. A goodness-of-fit score test (Murphy, 2007; Chiburis, 2010) can help detect misspecifications of the BP model.

Appendix

The Stata commands `biprobittreat`, `scoregof`, and `bphltest` are available for download at

https://webspace.utexas.edu/rcc485/www/code.html.

Instructions are given in Chiburis et al. (2011).

References

Altonji, J., Elder, T., Taber, C., 2005. Selection on observed and unobserved variables: assessing the effectiveness of catholic schools. Journal of Political Economy 113 (1), 151–184.

[1] See Chiburis (2010) for corrections to several errors in Murphy (2007) and an alternative derivation of the test.

Angrist, J., 1991. Instrumental variables estimation of average treatment effects in econometrics and epidemiology. NBER Technical Working Paper No. 0115.
Angrist, J., 2001. Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. Journal of Business and Economic Statistics 19 (1), 2–16.
Bhattacharya, J., Goldman, D., McCaffrey, D., 2006. Estimating probit models with self-selected treatments. Statistics in Medicine 25 (3), 389–413.
Chiburis, R.C., 2010. Score tests of normality in bivariate probit models: comment. Working Paper, University of Texas at Austin.
Chiburis, R.C., Das, J., Lokshin, M., 2011. A practical comparison of the bivariate probit and linear IV estimators. World Bank Policy Research Working Paper, #5601.
Heckman, J.J., 1978. Dummy endogenous variables in a simultaneous equation system. Econometrica 46 (6), 931–959.
Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475.
Murphy, A., 2007. Score tests of normality in bivariate probit models. Economics Letters 95 (3), 374–379.