WPS5601

Policy Research Working Paper 5601

A Practical Comparison of the Bivariate Probit and Linear IV Estimators

Richard C. Chiburis
Jishnu Das
Michael Lokshin

The World Bank
Development Research Group
Poverty and Inequality Team
and Human Development and Public Services Team
March 2011

Policy Research Working Paper 5601

Abstract

This paper presents asymptotic theory and Monte-Carlo simulations comparing maximum-likelihood bivariate probit and linear instrumental variables estimators of treatment effects in models with a binary endogenous treatment and binary outcome. The three main contributions of the paper are (a) clarifying the relationship between the Average Treatment Effect obtained in the bivariate probit model and the Local Average Treatment Effect estimated through linear IV; (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 relative to the existing literature; and (c) assessing the performance of misspecification tests for bivariate probit models. The authors recommend two changes to common practices: bootstrapped confidence intervals for both estimators, and a score test to check goodness of fit for the bivariate probit model.

This paper is a product of the Poverty and Inequality Team, and Human Development and Public Services Team; Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at jdas1@worldbank.org and mlokshin@worldbank.org.

A Practical Comparison of the Bivariate Probit and Linear IV Estimators

Richard C. Chiburis* Jishnu Das† Michael Lokshin‡

*University of Texas at Austin
†Development Research Group, World Bank
‡Development Research Group, World Bank

1 Introduction

This paper examines estimation issues in empirical models with binary regressors and outcome variables. A motivating example is the effect of private schooling on graduation rates. Here the "treatment"—attending a private school—and the "outcome"—whether or not the individual graduated—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. In some cases, these differ by an order of magnitude with the linear IV estimates exhibiting larger coefficients and standard errors (Altonji, Elder, and Taber 2005).

Keeping aside the discussion on the relevance of reduced-form impacts versus structural parameters (Angrist 2001, Moffitt 2001), the existing literature sometimes offers conflicting advice on the best course of action in empirical problems of this sort. For instance, Angrist (1991, 2001) argues that the hard part of the empirical problem is finding a good instrument and that the "difficulties with endogenous variables in nonlinear limited dependent variables models are usually more apparent than real." This argument is supported by a stress on causal effects as opposed to structural parameters in these models and by Monte-Carlo simulations that argue for the robustness of the simpler linear estimator to the distribution of the error terms. On the other hand, Bhattacharya, Goldman, and McCaffrey (2006) present

simulations that suggest that BP is slightly more robust than IV to non-normality of the error terms. We show that both of these seemingly different viewpoints can be justified depending on the structure of the problem. The reconciliation is based on (a) distinguishing carefully between the Local Average Treatment Effect (LATE) estimated under the linear IV with the Average Treatment Effect (ATE) estimated under the BP model and (b) extending the parameter values in Monte-Carlo simulations to cover a far wider range of model specifications relative to the existing literature.

We present asymptotic and finite-sample Monte-Carlo simulation results for an extensive range of parameter values to help decide on a practical course of action when faced with a single dataset, a reliable instrument and (possibly) widely differing estimates of the treatment effect depending on the technique used. We focus on both the mean-square error of the estimators and on the size and power of hypothesis tests based on these estimators. We present simulations both with the BP model correctly specified and with misspecification due to non-normal error terms. Finally, we propose two straightforward additions to current practice that vastly improve the performance of the estimators and our confidence in the normality assumptions of the BP model.

Our first set of results assumes that the BP model is correctly specified. Under this assumption, when there are *no covariates*, BP tends to perform better than IV for smaller sample sizes (below 5000), with mixed results for larger samples. With a *continuous covariate*, the performance of BP dominates IV in all of our simulations, and BP performs especially well when the treatment probability is close to 0 or 1. For instance, when the treatment probability is 0.1, for all ranges of the outcome probabilities and even with sample sizes greater than 10,000 observations, the confidence intervals of the IV estimate remain too large for any meaningful hypothesis testing; in contrast, BP confidence intervals are much smaller.[1] Therefore, researchers should expect IV and BP coefficients to differ quite

[1] This particular finding explains the large differences between the IV and BP confidence intervals in the motivating example above, since the percentage of children in private schools in the United States is 10 percent or lower in most samples

substantially when treatment probabilities are low or when sample sizes are below 5000; linear IV estimates are particularly uninformative for hypothesis testing when treatment probabilities are low.

Further, as pointed out by Imbens and Angrist (1994) and others, the IV estimate is consistent for the local average treatment effect (LATE) and not the overall average treatment effect (ATE), which can be recovered from the maximum-likelihood BP estimate. That the estimators are estimating different effects accounts for a finding by Angrist (1991) that in some cases, the variance of the IV estimator is lower than that of the maximum-likelihood BP ATE estimator despite the well known efficiency of maximum likelihood.

As expected, across most parameters of our simulations, the BP estimator is not robust to misspecification of the BP model. Simulation results where the error terms exhibit excess skewness or excess kurtosis often lead to highly biased BP estimates, and tests based on BP estimates greatly overreject a true null hypothesis when the model is misspecified. Tests based on IV estimates are more robust in terms of size, but they are also less powerful. The results presented in Bhattacharya, Goldman, and McCaffrey (2006) on the robustness of the BP estimator to non-normal error terms arise only for specific combinations of the relevant parameters, as clarified by the extensive Monte-Carlo simulations considered here.

We propose two additional steps to recover better confidence intervals and check for model misspecification in BP estimators. For both BP and IV estimators, sample sizes have to exceed 10,000 before the coverage rates of the standard Wald-type confidence intervals approach the nominal coverage rate. In general, IV confidence intervals tend to be too conservative and BP confidence intervals are not conservative enough. We show that using bootstrapped confidence intervals (relative to analytical versions) improves the coverage rate of both IV and BP estimators for all parameter values. Second, we recommend running Murphy's (2007) score test to check the goodness-of-fit of the BP model; our simulations suggest that the score test is fairly good at picking up misspecifications arising from excess kurtosis or skewness in the error distributions.

The remainder of the paper is structured as follows. Section 2 reviews standard asymptotic results. Section 3 discusses the data generating process for the Monte-Carlo simulations and presents results. Section 4 concludes.

2 Asymptotic properties of IV and BP estimators

We first derive asymptotic results for the case of a single binary instrument and no covariates. The section also details the relationship between two commonly used treatment effects, the Average Treatment Effect (LATE) and the Local Average Treatment Effect (ATE).

Let $T \in \{0, 1\}$ be the 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

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

We think of individuals as being sampled from the joint distribution of the random variables (Z, T_1, T_0, Y_1, Y_0). For each individual i, 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 three commonly estimated treatment effects, defined as follows:

1. The average treatment effect (ATE) over the entire population is given by

$$\Delta_{ATE} = \mathrm{E}[Y_1] - \mathrm{E}[Y_0]. \tag{3}$$

2. The average treatment effect on the treated (ATT) is the average treatment effect only over those individuals who actually received the treatment:

$$\Delta_{ATT} = \mathrm{E}[Y_1 \mid T = 1] - \mathrm{E}[Y_0 \mid T = 1]. \tag{4}$$

3. The probability limit of the IV estimator is what Imbens and Angrist (1994) called the local average treatment effect (LATE):

$$\Delta_{LATE} = \frac{\mathrm{E}[Y \mid Z = 1] - \mathrm{E}[Y \mid Z = 0]}{\mathrm{E}[T \mid Z = 1] - \mathrm{E}[T \mid Z = 0]}. \tag{5}$$

Under the condition that $T_1 \geq T_0$ for all individuals, Imbens and Angrist show that Δ_{LATE} can be interpreted as the average treatment effect for the subpopulation that complies with the instrument, i.e. the subpopulation for which T would be equal to Z regardless of whether $Z = 0$ or $Z = 1$.

It is informative to compare these effects to the result we would obtain if we ignore that T is endogenous and run an OLS regression of Y on T and a constant.[2] The probability limit of such a regression is

$$\Delta_{OLS} = \mathrm{E}[Y_1 \mid T = 1] - \mathrm{E}[Y_0 \mid T = 0]. \tag{6}$$

If $\mathrm{E}[Y_1 \mid T]$ and $\mathrm{E}[Y_0 \mid T]$ are invariant to T, so that there is no selection bias, then $\Delta_{OLS} = \Delta_{ATE} = \Delta_{ATT}$. Note that this condition does not ensure that $\Delta_{LATE} = \Delta_{ATE}$.

[2]Since Y is binary, one might also consider running a probit of Y on T and a constant. With no covariates, running a probit and computing the treatment effect produces exactly the same result as OLS since both models have two parameters to fit two moments $\mathrm{E}[Y \mid T = 1]$ and $\mathrm{E}[Y \mid T = 0]$.

2.1 Bivariate probit model

Typically it is necessary to impose additional structure on the model in order to identify Δ_{ATE} and Δ_{ATT}.[3] One way to do this while still allowing the treatment to be endogenous is to assume a bivariate probit model, which is a linear index model with bivariate normal shocks (Heckman 1978):

$$\begin{aligned} T^* &= \alpha Z + \kappa_T + \varepsilon_1 \\ T &= \mathbf{1}\{T^* > 0\} \\ Y^* &= \gamma T + \kappa_Y + \varepsilon_2 \\ Y &= \mathbf{1}\{Y^* > 0\} \end{aligned} \tag{7}$$

with $(\varepsilon_1, \varepsilon_2)$ jointly distributed as standard bivariate normal with correlation ρ and independent of Z. Note that assumption (1) above follows from this independence condition, and that $\alpha \neq 0$ implies (2). When $\alpha > 0$, Δ_{LATE} (5) has the interpretation given by Imbens and Angrist (1994).

Define $p_T = \Pr[T = 1]$ and $p_Y = \Pr[Y = 1]$. Let Φ and ϕ be the standard normal distribution and density functions, respectively. Let $B(\cdot, \cdot; \rho)$ be the distribution function for the standard bivariate normal distribution with correlation ρ. The ATE (3) is

$$\Delta_{ATE} = \Phi(\kappa_Y + \gamma) - \Phi(\kappa_Y). \tag{8}$$

[3] Heckman and Vytlacil (1999) observe that if there exists a value z of the instrument such that $\mathrm{E}[T \mid Z = z] = 0$, then Δ_{ATT} is nonparametrically identified. If additionally there exists z' such that $\mathrm{E}[T \mid Z = z'] = 1$, then Δ_{ATE} is also identified. This is a type of "identification at infinity" result since it typically requires extreme values of Z to be observed. However, with a binary Z as we have in our simulations, these conditions are rarely satisfied.

A first-order Taylor approximation about $\gamma = 0$ is

$$\Delta_{ATE} \approx \gamma\phi(\Phi^{-1}(p_Y)). \tag{9}$$

The ATT (4) is given by

$$\begin{aligned} \Delta_{ATT} &= \Pr[Z=0]\frac{B(\kappa_T, \kappa_Y+\gamma;\rho) - B(\kappa_T,\kappa_Y;\rho)}{\Phi(\kappa_T)} \\ &\quad + \Pr[Z=1]\frac{B(\kappa_T+\alpha, \kappa_Y+\gamma;\rho) - B(\kappa_T+\alpha,\kappa_Y;\rho)}{\Phi(\kappa_T+\alpha)}. \end{aligned} \tag{10}$$

The LATE (5) can also be written as a function of the parameters in the bivariate probit model:

$$\Delta_{LATE} = \frac{[B(\kappa_T+\alpha,\kappa_Y+\gamma,\rho)+B(-(\kappa_T+\alpha),\kappa_Y,-\rho)]-[B(\kappa_T,\kappa_Y+\gamma,\rho)+B(-\kappa_T,\kappa_Y,-\rho)]}{\Phi(\kappa_T+\alpha)-\Phi(\kappa_T)}. \tag{11}$$

While all of the types of treatment effects are equal when $\rho = 0$, they can differ significantly for other values of ρ; in particular, the ordering of Δ_{ATE}, Δ_{ATT}, and Δ_{LATE} varies across parameter values. In Appendix A.2, we derive a Taylor approximation for the ratio of Δ_{LATE} to Δ_{ATE} as

$$\frac{\Delta_{LATE}}{\Delta_{ATE}} \approx 1 + \rho\Phi^{-1}(p_Y)\Phi^{-1}(p_T). \tag{12}$$

Since the probability limit of the IV estimator $\hat{\Delta}^{IV}$ is Δ_{LATE}, (12) can be used to obtain a quick and intuitive approximation of the bias of $\hat{\Delta}^{IV}$ for Δ_{ATE}. In general, $\hat{\Delta}^{IV}$ is most biased relative to Δ_{ATE} when $|\rho|$ is large, and p_T and p_Y are far from $\frac{1}{2}$. The sign of the bias depends on the signs of ρ, $\Phi^{-1}(p_Y)$, and $\Phi^{-1}(p_T)$. In Figure 1, we graph Δ_{ATE}, Δ_{ATT}, Δ_{LATE}, and Δ_{OLS} for model (7) with $\alpha = 0.3$, $\gamma = 0.4$, and different values of p_T, p_Y, and ρ. We vary p_T and p_Y by changing the constants κ_T and κ_Y. As suggested by the approximation (9) and given that γ is fixed, Δ_{ATE} changes with p_Y but is very little affected by p_T or ρ. When $\rho = 0$, all of the effects are equal. When $\rho > 0$, the OLS effect Δ_{OLS} (6) that ignores the endogeneity of T is biased upward relative to Δ_{ATE}, Δ_{ATT}, and Δ_{LATE}, which correctly take into account the endogeneity of the treatment. As predicted

by the approximation (12), Δ_{ATE} and Δ_{LATE} differ substantially when p_T and p_Y are far from $\frac{1}{2}$, and this difference increases with ρ. The effect on the treated Δ_{ATT} is naturally close to the overall average effect Δ_{ATE} when the probability of treatment p_T is high. When p_T is low, Δ_{ATT} is closer to Δ_{LATE} because in that case there is high overlap between the treated and complier populations. Figure 1 also highlights the limits of our approximation (12)—according to (12), Δ_{LATE} and Δ_{ATE} should the same when either p_Y or p_T is $\frac{1}{2}$, but at higher values of ρ, we see that the two effects diverge.

2.2 Asymptotic variance of linear IV estimator

The asymptotic variance $\text{Avar}[\hat{\Delta}]$ of an estimator $\hat{\Delta}$ is defined such that

$$n \operatorname{Var}[\hat{\Delta}] \xrightarrow{p} \text{Avar}[\hat{\Delta}].$$

The asymptotic variance of the IV estimator $\hat{\Delta}^{IV}$ is

$$\text{Avar}[\hat{\Delta}^{IV}] = \frac{\frac{\Pr[Y=1|Z=1](1-\Pr[Y=1|Z=1])}{\Pr[Z=1]} + \frac{\Pr[Y=1|Z=0](1-\Pr[Y=1|Z=0])}{\Pr[Z=0]}}{\left(\Pr\left[T=1 \mid Z=1\right] - \Pr\left[T=1 \mid Z=0\right]\right)^2}. \tag{13}$$

To provide intuition for how $\text{Avar}[\hat{\Delta}^{IV}]$ changes with p_T, p_Y, and ρ, we derive in Appendix A.3 the following Taylor approximation of $\text{Avar}[\hat{\Delta}^{IV}]$ within the context of the bivariate probit model:

$$\text{Avar}[\hat{\Delta}^{IV}] \approx \frac{p_Y(1-p_Y)}{\alpha^2[\phi(\Phi^{-1}(p_T))]^2 \operatorname{Var}[Z]}. \tag{14}$$

The asymptotic variance of $\hat{\Delta}^{IV}$ increases as p_Y approaches $\frac{1}{2}$ and as p_T moves away from $\frac{1}{2}$. Furthermore, the approximation (14) of $\text{Avar}[\hat{\Delta}^{IV}]$ does not depend on ρ at all, and the exact $\text{Avar}[\hat{\Delta}^{IV}]$ (13) exhibits very little dependence on ρ, as illustrated in Figure 2, which plots $\text{Avar}[\hat{\Delta}^{IV}]$ for various parameter values.

2.3 Asymptotic variance of ML bivariate probit estimators

Let $\boldsymbol{\theta}$ denote the vector of the parameters of α, γ, κ_T, κ_Y, and ρ in the bivariate probit model (7). Maximum-likelihood estimates of $\boldsymbol{\theta}$ are obtained by selecting $\hat{\boldsymbol{\theta}}$ to maximize the log-likelihood function:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{n} \log L_i(\boldsymbol{\theta})$$

where

$$L_i(\boldsymbol{\theta}) = \begin{cases} B(\alpha Z_i + \kappa_T, \gamma T_i + \kappa_Y; \rho), \text{ if } T_i = 1 \text{ and } Y_i = 1; \\ B(\alpha Z_i + \kappa_T, -(\gamma T_i + \kappa_Y); -\rho), \text{ if } T_i = 1 \text{ and } Y_i = 0; \\ B(-(\alpha Z_i + \kappa_T), \gamma T_i + \kappa_Y; -\rho), \text{ if } T_i = 0 \text{ and } Y_i = 1; \\ B(-(\alpha Z_i + \kappa_T), -(\gamma T_i + \kappa_Y); \rho), \text{ if } T_i = 0 \text{ and } Y_i = 0. \end{cases} \tag{15}$$

Once we have estimates of the parameters, we can estimate most types of treatment effects, because they are functions of $\boldsymbol{\theta}$, by substituting the estimated parameters into the expressions (8) for the ATE, (10) for the ATT, and (11) for the LATE. We denote the respective ML estimators as $\hat{\Delta}^{BP}_{ATE}$, $\hat{\Delta}^{BP}_{ATT}$, and $\hat{\Delta}^{BP}_{LATE}$. Hence, if the bivariate probit model (7) is correctly specified, maximum-likelihood can be used to consistently estimate the ATE, ATT, or LATE, whereas the linear IV estimator $\hat{\Delta}^{IV}$ only consistently estimates the LATE (5).

The asymptotic variance of the ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is given by $\operatorname{Avar}[\hat{\boldsymbol{\theta}}] = I(\boldsymbol{\theta})^{-1}$, the inverse information matrix evaluated at the true $\boldsymbol{\theta}$. There are two common ways to calculate $I(\boldsymbol{\theta})$:

$$I_1(\boldsymbol{\theta}) = \left(\mathrm{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L_i(\boldsymbol{\theta}) \right) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L_i(\boldsymbol{\theta}) \right)' \right] \right) \tag{16}$$

and

$$I_2(\boldsymbol{\theta}) = -\,\mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log L_i(\boldsymbol{\theta}) \right]. \tag{17}$$

Using the delta method, we compute the asymptotic variance of any continuously differen-

tiable function f of $\boldsymbol{\theta}$ as

$$f'(\boldsymbol{\theta})' \operatorname{Avar}[\hat{\boldsymbol{\theta}}] f'(\boldsymbol{\theta}). \tag{18}$$

Since the ATE, ATT, and LATE are all functions of $\boldsymbol{\theta}$, we can compute the asymptotic variance of $\hat{\Delta}^{BP}_{ATE}$, $\hat{\Delta}^{BP}_{ATT}$, and $\hat{\Delta}^{BP}_{LATE}$ in this fashion. The results for $\hat{\Delta}^{BP}_{ATE}$ for model (7) with $\alpha = 0.3$, $\gamma = 0.4$, and many different values of p_T, p_Y, and ρ are shown in Figure 2. Note that the asymptotic variance of $\hat{\Delta}^{BP}_{ATE}$ is highly sensitive to ρ when p_T is far from $\frac{1}{2}$.

2.4 Comparing the asymptotic variances

Because linear IV only consistently estimates the LATE, the asymptotic variances of linear IV and maximum-likelihood BP are compared most fairly for estimation of the LATE. When the BP model (7) is correctly specified, maximum likelihood BP is asymptotically efficient for the LATE since it is asymptotically efficient for any smooth function of the parameters $\boldsymbol{\theta}$. Using the formulas (13) and (18), we compared the asymptotic variances of $\hat{\Delta}^{IV}$ and $\hat{\Delta}^{BP}_{LATE}$ for the LATE in model (7) with $\alpha = 0.3$, $\gamma = 0.4$, and across many different values of p_T, p_Y, and ρ. The asymptotic variance of $\hat{\Delta}^{BP}_{LATE}$ is always lower than that of $\hat{\Delta}^{IV}$, and on average, the variance of $\hat{\Delta}^{IV}$ is 28 percent higher than $\hat{\Delta}^{BP}_{LATE}$, or its standard deviation is 13 percent higher than $\hat{\Delta}^{BP}_{LATE}$. The efficiency gain from using $\hat{\Delta}^{BP}_{LATE}$ instead of $\hat{\Delta}^{IV}$ is far greater when a covariate is included in the model. In our models with a continuous covariate, on average the standard deviation of $\hat{\Delta}^{IV}$ is 150 percent higher than $\hat{\Delta}^{BP}_{LATE}$.

Angrist (1991) found that despite the efficiency of BP, the variance of $\hat{\Delta}^{BP}_{ATE}$ sometimes exceeds the variance of $\hat{\Delta}^{IV}$. Angrist's results follow from the asymptotics, as shown in Figure 2, where for certain values of p_T, p_Y, and ρ, the asymptotic variance of $\hat{\Delta}^{BP}_{ATE}$ is higher than that of $\hat{\Delta}^{IV}$. The key observation is that $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ are estimating two different things, ATE and 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 $\hat{\Delta}^{BP}_{ATE}$ will always be the better choice (when the BP

model is correct) except in rare cases in which $\hat{\Delta}^{IV}$ has lower asymptotic variance than $\hat{\Delta}^{BP}_{ATE}$ *and* the LATE happens to be close to the ATE.

3 Monte-Carlo simulations

To examine the properties of 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, Goldman, and McCaffrey (2006), and prove useful in understanding the performance of these estimators in practical applications. The wider range of parameters considered here qualitatively affects the nature of the recommendation. 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, Goldman, and McCaffrey (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. Table 1 compares the parameter ranges used in the different papers.

	Our simulations	Angrist (1991)	Bhattacharya et al. (2006)
p_T	$0.1 - 0.9$	$0.2 - 0.5$	0.5
p_Y	$0.1 - 0.9$	$0.5 - 0.9$	$0.0 - 0.7$
ρ	$0.0 - 0.7$	0.5	$0.1 - 0.5$
Δ_{ATE}	$0.05 - 0.16$	0.10	$0.00 - 0.42$
n	$400 - 30,000$	$400 - 800$	5000
Number of covariates	0 or 1	0	1

Table 1: Ranges of parameter values used in various studies.

3.1 Data-generating processes

Our data-generating processes (DGPs) are all based on the following latent-index model:

$$
\begin{aligned}
T_i^* &= \alpha Z_i + \beta_T X_i + \kappa_T + \varepsilon_{Ti} \\
T_i &= \mathbf{1}\{T_i^* > 0\} \\
Y_i^* &= \gamma T_i + \beta_Y X_i + \kappa_Y + \varepsilon_{Yi} \\
Y_i &= \mathbf{1}\{Y_i^* > 0\}
\end{aligned} \tag{19}
$$

where T_i^* and Y_i^* are latent continuous variables; α, γ, β_T, β_Y, κ_T, and κ_Y are parameters; X_i is an exogenous covariate; and Z_i is an instrumental dummy variable that is zero with probability $\frac{1}{2}$ and one with probability $\frac{1}{2}$.

Our DGPs are designed to mimic typical situations encountered in applied econometric applications. The values of the coefficients in the system (19) are chosen such that the true ATE is positive and falls in the range from 0.05 to 0.16 depending on the model specification. In all of the DGPs, the coefficients in (19) take the following values:

$$
\alpha = 0.3; \gamma = 0.4; \beta_T = 0.9; \beta_Y = 0.4 \tag{20}
$$

We consider two DGPs for X_i:

1. In the first DGP, $X_i = 0$ always, and hence we do not estimate β_T or β_Y.

2. In the second DGP, $X_i \sim \mathcal{N}(0, 1)$, and X_i is independent of Z_i.

The error terms ε_{Ti} and ε_{Yi} are always jointly independent of (X_i, Z_i) and can be generated according to any of six possible processes:

1. ε_T and ε_Y are jointly bivariate standard normal with correlation ρ taking on one of four possible values: $0, 0.3, 0.5, 0.7$.

2. Generate (u_T, u_Y) as bivariate normal with correlation 0.32. Then transform $\varepsilon_T = F(\Phi^{-1}(u_T))$ and $\varepsilon_Y = F(\Phi^{-1}(u_Y))$, where F is the CDF of a chi-square distribution

with 5 degrees of freedom. This results in skewed distributions for ε_T and ε_Y, and the bivariate probit model is misspecified.[4]

3. Generate (u_T, u_Y) as bivariate normal with correlation 0.32. Then transform $\varepsilon_T = F(\Phi^{-1}(u_T))$ and $\varepsilon_Y = F(\Phi^{-1}(u_Y))$, where F is the CDF of a t distribution with 4 degrees of freedom. This results in distributions for ε_T and ε_Y with high kurtosis, and the bivariate probit model is misspecified.

Furthermore, we also consider many values of the constants κ_T and κ_Y. They are chosen so that $p_T = \Pr[T = 1]$ and $p_Y = \Pr[Y = 1]$ each range separately over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$.

For each of the 300 combinations of possible DGPs for X_i, DGPs for $(\varepsilon_T, \varepsilon_Y)$, and values of p_T and p_Y specified above, we conduct Monte-Carlo simulations on samples of 400, 800, 1K, 2K, 3K, 5K, 8K, 10K, 15K, 20K, and 30K observations. We run 1000 simulations for each sample size. In each simulation we compute the IV estimate of the LATE and the maximum-likelihood BP estimates of the ATE. Greene (1998) observed that the endogeneity of T_i does not affect the form of the BP likelihood function (15), and hence BP estimates can be obtained directly from the bivariate probit routine available in many statistical software packages.

In the simulations with nonzero covariates X_i, the ATE for the bivariate probit model is estimated as

$$\hat{\Delta}^{BP}_{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left(\Phi(\hat{\gamma} + \hat{\beta}_Y X_i + \hat{\kappa}_Y) - \Phi(\hat{\beta}_Y X_i + \hat{\kappa}_Y)\right).$$

The true ATE and LATE always lie in the interval $[-1, 1]$. While $\hat{\Delta}^{BP}_{ATE}$ will always fall in that interval, $\hat{\Delta}^{IV}$ is sometimes outside this interval, especially when the sample size is small.

[4]The correlation of 0.32 for (u_T, u_Y) was chosen so that the correlation of the transformed $(\varepsilon_T, \varepsilon_Y)$ is approximately 0.30, allowing for comparison to the bivariate normal simulations with $\rho = 0.30$.

3.2 Results

Our simulation results are presented in Figures 3 through 8. The first three are representationally similar: in every sub-figure, we plot the true Δ_{ATE} (the dotted line), the mean of $\hat{\Delta}^{BP}_{ATE}$ (the thick solid curve) and the mean of $\hat{\Delta}^{IV}$ (the thick dashed curve) against sample sizes between 400 and 30,000. We also show the range between the 5th and the 95th percentiles of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$. There are 9 sub-figures showing the behavior of the BP and IV estimators for different parameter values of p_T and p_Y and in every figure, we fix $\rho = 0.3$. Figure 3 presents simulations in estimations with no covariates, Figure 4 with covariates and Figure 5 examines departures from the BP model assumptions. In Appendix A.5, we provide tables of the root-mean-square error of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ for estimating Δ^{ATE} over a wider range of parameter values, which researchers can use with reference to the structure of their own particular problem.

There are several noteworthy features. First, when there are no covariates (Figure 3) the simulations match the asymptotics (Figures 1 and 2) fairly well in sample sizes larger than about 5,000. In sample sizes smaller than 5,000, $\hat{\Delta}^{BP}_{ATE}$ has *lower* variance than predicted by the asymptotics, because of mechanical bounds on the estimator ($\hat{\Delta}^{BP}_{ATE} \in [-1, 1]$). Second $\hat{\Delta}^{BP}_{ATE}$ can be *biased* in small samples, as often happens for maximum-likelihood estimators. Even when sample sizes are large, $\hat{\Delta}^{BP}_{ATE}$ can be biased under particular extreme combinations of p_T and p_Y—in our simulations, two particularly dramatic examples are ($p_T = 0.9, p_Y = 0.1$) and ($p_T = 0.1, p_Y = 0.9$).[5] Third, due to its relatively lower small-sample variance, $\hat{\Delta}^{BP}_{ATE}$ generally performs better than $\hat{\Delta}^{IV}$ in terms of RMSE for sample sizes smaller than about 5,000. For larger sample sizes, the efficiency of $\hat{\Delta}^{BP}_{ATE}$ relative to $\hat{\Delta}^{IV}$ is somewhat reduced and for extreme combinations of parameter values, $\hat{\Delta}^{IV}$ can be the better estimator in terms of RMSE.

[5] Firth (1993) describes several techniques for removing the first-order bias from maximum-likelihood estimates. We simulated both asymptotic first-order bias removal and bootstrap bias removal for the BP estimator but found that both techniques perform rather poorly, especially when $|\rho|$ is close to 1, since lower-order expansions poorly approximate the finite-sample bias.

Figure 4 shows that once we include covariates X in the BP model (19), $\hat{\Delta}^{BP}_{ATE}$ has much lower variance and outperforms $\hat{\Delta}^{IV}$ across *all* of our simulations in terms of RMSE for Δ_{ATE}. Indeed, in most cases, the IV standard errors are too large for meaningful hypothesis testing, a problem that is particularly severe when p_T is close to 0 or 1. These simulations highlight that the use of linear IV estimators with covariates can lead to extremely high standard errors and dramatic differences in $\hat{\Delta}^{BP}_{ATE}$ relative to $\hat{\Delta}^{IV}$.

An overarching theme thus far is that the BP estimators are generally more efficient than linear IV, especially when the model specification includes additional covariates. However, the gain in efficiency may be outweighed by the severe bias when the BP model is misspecified. Figure 5 examines departures from the BP model assumptions in the case with covariates.[6] In this case, $\hat{\Delta}^{BP}_{ATE}$ continues to have low variance but can be severely biased in some cases, with no clear guidance on the parameter values under which the expected bias will be worse. The evidence of bias presented here contrasts with the results of Bhattacharya, Goldman, and McCaffrey (2006), who suggest that BP is slightly more robust to non-normality than IV. As Figure 5 clarifies, Bhattacharya, Goldman, and McCaffrey's result is a direct consequence of their choice of parameters. In their simulations with non-normal errors, p_T is fixed at 0.5, p_Y ranges between 0.5 and 0.7, and ρ is 0.5. Our results in Figure 5 suggest that these happen to be values of p_T and p_Y for which $\hat{\Delta}^{BP}_{ATE}$ performs fairly well even when the BP model assumptions are violated.

[6] For simulations with skewness or excess kurtosis of the error terms, the results are similar because the BP estimates are still consistent, despite the misspecification. This is because with no covariates our misspecified DGP is observationally equivalent to a correctly specified bivariate probit model. Recall that (u_T, u_Y) are generated as bivariate normal with correlation ρ, and then $\varepsilon_T = f(u_T)$ and $\varepsilon_Y = f(u_Y)$ for some monotone function f. Let $\tilde{\kappa}_T = f^{-1}(\kappa_T)$, $\tilde{\alpha} = f^{-1}(\kappa_T + \alpha) - \tilde{\kappa}_T$, $\tilde{\kappa}_Y = f^{-1}(\kappa_Y)$, and $\tilde{\gamma} = f^{-1}(\kappa_Y + \gamma) - \tilde{\kappa}_Y$. Then a correctly specified bivariate probit model with coefficients $\tilde{\kappa}_T$, $\tilde{\alpha}$, $\tilde{\kappa}_Y$, $\tilde{\gamma}$, ρ produces the same distribution of observables as our DGP, and the values of all treatment effects are the same in both models. It would have been possible for the BP estimators to be inconsistent if we had modified the *joint* distribution of $(\varepsilon_T, \varepsilon_Y)$ rather than modifying the marginal distributions individually. With a nonzero covariate X_i, the assumption of normality will actually be restrictive because the transformation f^{-1} will be applied at more than two points and hence will no longer preserve linearity.

3.3 Coverage of confidence intervals

Our final simulation results examine the validity of confidence intervals generated by the various methods and the performance of goodness-of-fit tests for the BP model, which can help detect potential misspecification in the BP model. Figure 6 compares the nominal 95% confidence intervals based on $\hat{\Delta}^{IV}$ and $\hat{\Delta}^{BP}_{ATE}$ in terms of coverage of Δ_{ATE}. The standard error used to construct the confidence intervals for $\hat{\Delta}^{IV}$ is obtained using the sample analogue of the asymptotic variance (13). Results are shown in Figure 6 for a correctly specified model with $X_i \sim \mathcal{N}(0,1)$ and $\rho = 0.3$. As shown in the figure, 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 $\hat{\Delta}^{IV}$ converges to Δ_{LATE} rather than Δ_{ATE} (the dashed curve in the figure). The most common way to compute standard errors for the BP parameters $\boldsymbol{\theta}$ is by estimating the information matrix using the sample analogue of $I_2(\hat{\boldsymbol{\theta}})$ (17). We would then apply the delta method as in (18) to obtain standard errors for $\hat{\Delta}^{BP}_{ATE}$. BP confidence intervals for Δ_{ATE} computed in this way display significantly lower coverage than the nominal 95% for sample sizes below 5,000, even when the model is correctly specified, but coverage improves toward 95% in samples larger than 10,000 observations (the solid curve in the figure). Further investigation reveals that this undercoverage occurs because standard errors for the BP parameters are too small, and additional undercoverage is introduced in the delta-method step.[7] Alternatively, we tried estimating the information matrix using the sample analogue of $I_1(\hat{\boldsymbol{\theta}})$ (16), or standard errors can be estimated using the Huber-White sandwich (robust) estimator $\hat{I}_2(\hat{\boldsymbol{\theta}})^{-1}\hat{I}_1(\hat{\boldsymbol{\theta}})\hat{I}_2(\hat{\boldsymbol{\theta}})^{-1}$. These methods result in similar undercoverage of Δ_{ATE}.

Fortunately, bootstrapped confidence intervals appear to provide a simple fix for over and undercoverage in both the IV and BP models. In the bootstrap, we draw with replacement n observations from the data and estimate $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ using the new sample. This is

[7] Monfardini and Radice (2008) report a similar result that t-tests based on maximum-likelihood estimation of the BP model systematically overreject the hypothesis $\rho = 0$. Also as noted by Freedman and Sekhon (2010), part of the difficulty may be caused by numerical issues with Stata's implementation of likelihood maximization, as often the likelihood function is very flat and the algorithm fails to find the global maximum.

repeated many times, and the size-α confidence interval for Δ_{ATE} or Δ_{LATE} is reported as the interval between the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the simulated draws of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$.[8] By bootstrapping the entire procedure of calculating $\hat{\Delta}^{BP}_{ATE}$, we avoid using the delta method. Because we ran thousands of simulations we used 39 bootstrap replications in each of our simulations to save time (each bootstrap replication took about 15 seconds at $n = 30,000$), but we recommend at least 199 bootstrap replications in practice to reduce sampling noise. In addition, we simulated bootstrap results for only two sample sizes ($n = 400$ and $n = 3,000$) given the processing time involved.

The coverage rates of the bootstrapped BP confidence intervals for Δ_{ATE} are close to the nominal $1-\alpha$, as shown in Figures 6 and 7. 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 Figure 3 to be particularly problematic for BP. We therefore strongly recommend using bootstrapped confidence intervals for BP, whether one is estimating treatment effects or just the BP coefficients $\boldsymbol{\theta}$.[9] In addition, Figures 6 and 7 show 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 $\hat{\Delta}_{IV}$ is generally inconsistent for Δ_{ATE}.

3.4 Goodness-of-fit tests for bivariate probit

Figure 8 presents results of goodness-of-fit tests for the bivariate probit model. We compare the ability of two different goodness-of-fit tests to detect our non-normal data-generating processes. Our first test is an adaptation of the Hosmer and Lemeshow (1980) test to the bivariate probit model. This test divides the observations into subgroups and checks whether the frequencies of observed (y_i, t_i) match predicted frequencies given $\hat{\boldsymbol{\theta}}$ and the distribution

[8] Our simulations indicate that these quantile-based confidence intervals perform better than bootstrapping standard errors and then using a normal approximation to obtain confidence intervals.

[9] When the BP model is *misspecified*, the coverage rates of BP confidence intervals are severely affected by the misspecification when there are covariates X_i. The misspecification has a lesser impact on IV coverage rates, since IV standard errors tend to be larger and IV is generally not consistent even in the BP model. However, for the same reason, tests based on $\hat{\Delta}^{IV}$ are generally less powerful than tests based on $\hat{\Delta}^{BP}_{ATE}$.

of X_i and Z_i in each subgroup. The details of our adaptation of the Hosmer-Lemeshow test are given in Appendix A.4. The second goodness-of-fit test we use is a Rao score test developed by Murphy (2007).[10] This test embeds the bivariate normal distribution within a larger family of distributions by adding more parameters to the model and checks whether the additional parameters are all zeros using the score for the additional parameters at the BP estimate.[11] We set both tests to reject at a 5% significance level using asymptotic chi-square critical values.[12] In our simulations with a bivariate normal data generating process, both tests reject about 5% of the time, as expected. The score test performs much better than the Hosmer-Lemeshow test in detecting our non-normal data-generating processes, as shown in Figure 8. The results of this comparison of the two tests agree with those of Chiburis (2010) from simulations without an endogenous regressor.

4 Conclusion

We have derived asymptotic results and presented simulations comparing bivariate probit and linear IV estimators of the average treatment effect of a binary treatment on a binary outcome. Our simulation results provide some practical guidance on the choice of specification in practical problems with different parameter values and the presence/absence of covariates and can help explain widely differing results 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 coefficients to differ substantially when treatment probabilities are low or when sample sizes are below 5000. Linear IV estimates are particularly uninformative for hypothesis testing when treatment probabilities are low,

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

[11] Since T_i is endogenous, predicted probabilities of (y_i, t_i) used to calculate the test statistic are computed conditional on X_i and Z_i but not T_i.

[12] Murphy (2007) recommends bootstrapping the critical value of his test, but we find that the asymptotic critical values work well enough even at small sample sizes that the time-consuming bootstrap is not necessary.

a problem that is accentuated when there are covariates in the model. Table A5 in the Appendix provides the ATE, the ratio of LATE to ATE and the root mean square error for different values of p_T, p_Y and ρ for different sample sizes. These tables can be used as a guide for practical applications. 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.

- The difference between IV and BP estimates could also reflect differences between the LATE and ATE estimates recovered by the linear IV and BP procedures respectively. Again, Table A5 as well as our asymptotic ratio approximation provide a guide for the variance in these estimates.

- 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.

- As is well known, researchers should be aware that for a broad range of parameter values, misspecification of the BP model can lead to severe bias in BP estimates. This problem, however, does not arise in models with no covariates. In models with covariates, Murphy's goodness-of-fit score test (Murphy 2007, Chiburis 2010) can help detect misspecifications of the BP model.

A Appendix

A.1 Stata commands

In this appendix we describe how to run our recommended BP and IV procedures for calculating treatment effects in Stata. The Stata commands `biprobittreat`, `scoregof`, and

`bphltest` are available for download at

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

Suppose we have a dataset with binary outcome `Y`, binary treatment `T`, instrument `Z`, and covariates `X1`, `X2`.

1. To compute $\hat{\Delta}^{IV}$ along with bootstrapped confidence intervals, type:
   - `ivregress 2sls Y X1 X2 (T=Z), vce(bootstrap, reps(199))`
   - `estat bootstrap, percentile`
2. To compute $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{BP}_{ATT}$ along with bootstrapped confidence intervals, type:
   - `bootstrap _b ate=r(ate) att=r(att), reps(199): biprobittreat (Y = T X1 X2) (T = Z X1 X2)`
   - `estat bootstrap, percentile`
3. To run the Murphy score and Hosmer-Lemeshow goodness-of-fit tests, type:
   - `biprobit (Y = T X1 X2) (T = Z X1 X2)`
   - `scoregof`
   - `bphltest`

A.2 Derivation of $\Delta_{LATE}/\Delta_{ATE}$ approximation (12)

Using (8) and (11), we compute a first-order Taylor approximation of $\frac{\Delta_{LATE}}{\Delta_{ATE}}$ about $\rho, \gamma, \alpha = 0$. Although the ratio is undefined for $\gamma = 0$ or $\alpha = 0$, the limit $\lim_{\alpha,\gamma,\rho \to 0} \frac{\Delta_{LATE}}{\Delta_{ATE}}$ exists and is equal to 1, so we can still compute the Taylor expansion around this point. The terms involving the derivatives with respect to γ and α are zero because $\frac{\Delta_{LATE}}{\Delta_{ATE}} = 1$ when $\rho = 0$, regardless of γ and α. This leaves us with

$$\frac{\Delta_{LATE}}{\Delta_{ATE}} \approx 1 + \rho \lim_{\alpha,\gamma \to 0} \frac{\phi\left(\sqrt{(\kappa_T+\alpha)^2+(\kappa_Y+\gamma)^2}\right)-\phi\left(\sqrt{(\kappa_T+\alpha)^2+\kappa_Y^2}\right)-\phi\left(\sqrt{\kappa_T^2+(\kappa_Y+\gamma)^2}\right)+\phi\left(\sqrt{\kappa_T^2+\kappa_Y^2}\right)}{\sqrt{2\pi}(\Phi(\kappa_T+\alpha)-\Phi(\kappa_T))(\Phi(\kappa_Y+\gamma)-\Phi(\kappa_Y))}.$$

The limit is indeterminate, so we apply L'Hôpital's rule twice to obtain

$$\frac{\Delta_{LATE}}{\Delta_{ATE}} \approx 1 + \rho\kappa_T\kappa_Y.$$

In order to write this in terms of p_T and p_Y, at $\alpha = \gamma = 0$ we approximate $\kappa_T \approx \Phi^{-1}(p_T)$ and $\kappa_Y \approx \Phi^{-1}(p_T)$, yielding

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

A.3 Derivation of $\text{Avar}[\hat{\Delta}^{IV}]$ approximation (14)

We can write (13) as

$$\text{Avar}[\hat{\Delta}^{IV}] = \frac{\frac{\Pr[Y=1|Z=1](1-\Pr[Y=1|Z=1])}{\Pr[Z=1]} + \frac{\Pr[Y=1|Z=0](1-\Pr[Y=1|Z=0])}{\Pr[Z=0]}}{\tilde{\alpha}^2},$$

where

$$\tilde{\alpha} = \Phi(\kappa_T + \alpha) - \Phi(\kappa_T)$$

is the probability limit of the coefficient on Z in the first stage of the IV regression. A Taylor approximation of $\tilde{\alpha}$ in terms of p_T is

$$\tilde{\alpha} \approx \alpha\phi(\Phi^{-1}(p_T))$$

since $\Phi^{-1}(p_T)$ is between κ_T and $\kappa_T+\alpha^*$. Furthermore, for reasonable values of the treatment effect we can use

$$\mathrm{E}\left[Y \mid Z=1\right]\left(1-\mathrm{E}\left[Y \mid Z=1\right]\right) \approx \mathrm{E}\left[Y \mid Z=0\right]\left(1-\mathrm{E}\left[Y \mid Z=0\right]\right)$$

to approximate

$$\begin{aligned}
\mathrm{Avar}[\hat{\Delta}^{IV}] &\approx \frac{p_Y(1-p_Y)}{\tilde{\alpha}^2 \Pr\left[Z=1\right]\left(1-\Pr\left[Z=1\right]\right)} \\
&= \frac{\mathrm{Var}[Y]}{\tilde{\alpha}^2 \mathrm{Var}[Z]} \\
&\approx \frac{p_Y(1-p_Y)}{\alpha^2[\phi(\Phi^{-1}(p_T))]^2 \mathrm{Var}[Z]}.
\end{aligned}$$

A.4 Adapted Hosmer-Lemeshow goodness-of-fit test for bivariate probit

The Hosmer and Lemeshow (1980) test statistic was developed to correct a problem with the simple Pearson test statistic. To compute the Pearson test statistic in the bivariate probit model with an endogenous regressor, we create cells for each unique value of (x, z) in the data and sort the observations into those cells. For each cell $c = 1, \ldots, C$, let O_{cyt} be the number of observations in cell c with $Y = y$ and $T = t$, and let E_{cyt} be the expected number of observations in cell c with $Y = y$ and $T = t$ according to the BP model. It is computed as $E_{cyt} = \sum_{i \in \text{cell } c} \hat{\pi}_{iyt}$, where $\hat{\pi}_{iyt}$ is the predicted probability of $(Y, T) = (y, t)$ given $(X, Z) = (x_i, z_i)$, evaluated using the BP model at the estimated parameters $\hat{\boldsymbol{\theta}}$. The Pearson test statistic is

$$\mathcal{X}^2 = \sum_{c=1}^{C} \sum_{y=0}^{1} \sum_{t=0}^{1} \frac{\left(O_{cyt} - E_{cyt}\right)^2}{E_{cyt}}.$$

When X and Z are discrete and the number of unique cells (x, z) is small relative to n, $\mathcal{X}^2$ is approximately distributed as chi-square with $3C - \dim(\boldsymbol{\theta})$ degrees of freedom under the null hypothesis that the true model is BP (Osius and Rojek 1992). We recommend the use

of the Pearson test statistic in such cases. However, when there are many unique values of (x, z) in the data, as is the case when X or Z is continuously distributed, Osius and Rojek (1992) show that this asymptotic approximation for $\mathcal{X}^2$ breaks down. They compute a better asymptotic distribution of $\mathcal{X}^2$ for the continuous case.

The method of Hosmer and Lemeshow (1980) and Fagerland, Hosmer, and Bofin (2008), which was originally developed for logistic models, is another way to modify the Pearson test for use with continuous X or Z. This test combines the observations into a smaller number of groups to ensure that the test statistic is well approximated by its asymptotic distribution.[13] To adapt the Hosmer and Lemeshow (1980) test to the bivariate probit model, we choose two constants G_1 and G_2. We first sort the observations into G_1 groups of roughly equal size based on $\Pr[T = 1 \mid \hat{\boldsymbol{\theta}}, X = x_i, Z = z_i]$. Within each of these groups, we then sort the observations into G_2 subgroups based on $\Pr[Y = 1 \mid \hat{\boldsymbol{\theta}}, X = x_i, Z = z_i]$. This results in a total of $G = G_1 G_2$ groups. For each of these groups g, let O_{gyt} be the number of observations in group g with $Y = y$ and $T = t$, and let $E_{gyt} = \sum_{i \in \text{group } g} \hat{\pi}_{iyt}$. The adapted Hosmer-Lemeshow test statistic is

$$C = \sum_{g=1}^{G} \sum_{y=0}^{1} \sum_{t=0}^{1} \frac{(O_{gyt} - E_{gyt})^2}{E_{gyt}}.$$

Under the null hypothesis that BP is the correct model, we expect C to be distributed approximately chi-square with $3(G - 2)$ degrees of freedom. This distribution was derived by Fagerland, Hosmer, and Bofin (2008) based on simulations. In our simulations, the Hosmer-Lemeshow test statistic C is computed with $G_1 = G_2 = 3$.[14]

[13] Pigeon and Heyse (1999) add a small modification to the Hosmer-Lemeshow statistic. Their statistic has a slightly different asymptotic distribution.

[14] This results in 9 total groups, which is in the range of 8 to 12 groups used by Fagerland, Hosmer, and Bofin (2008) in their simulations of the analogous test for multinomial logistic regressions.

A.5 Simulation root-mean-square error tables

				Root-mean-square error									
				$n = 400$		$n = 1,000$		$n = 3,000$		$n = 10,000$		$n = 30,000$	
p_T	p_Y	Δ_{ATE}	$\frac{\Delta_{LATE}}{\Delta_{ATE}}$	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**
0.1	0.1	0.08	1.00	0.38	2.57	0.35	1.53	0.27	0.22	0.16	0.12	0.09	0.06
0.1	0.3	0.15	1.00	0.36	3.46	0.33	1.99	0.26	0.35	0.16	0.19	0.10	0.10
0.1	0.5	0.16	1.00	0.36	4.24	0.32	3.14	0.25	0.39	0.16	0.21	0.10	0.12
0.1	0.7	0.13	1.00	0.38	4.08	0.34	3.14	0.24	0.36	0.14	0.19	0.08	0.11
0.1	0.9	0.06	1.00	0.43	2.58	0.38	2.94	0.27	0.23	0.15	0.12	0.06	0.07
0.3	0.1	0.08	1.00	0.28	1.93	0.22	0.21	0.13	0.11	0.07	0.06	0.04	0.03
0.3	0.3	0.14	1.00	0.31	1.91	0.24	0.31	0.16	0.16	0.09	0.09	0.05	0.05
0.3	0.5	0.16	1.00	0.31	2.40	0.24	0.34	0.15	0.18	0.10	0.10	0.06	0.06
0.3	0.7	0.13	1.00	0.31	1.23	0.22	0.31	0.14	0.17	0.08	0.09	0.05	0.05
0.3	0.9	0.06	1.00	0.29	2.29	0.22	0.21	0.12	0.11	0.06	0.06	0.03	0.03
0.5	0.1	0.07	1.00	0.24	0.54	0.18	0.17	0.10	0.09	0.05	0.05	0.03	0.03
0.5	0.3	0.14	1.00	0.29	0.96	0.22	0.27	0.13	0.14	0.08	0.08	0.04	0.04
0.5	0.5	0.16	1.00	0.30	0.81	0.22	0.30	0.14	0.15	0.09	0.09	0.05	0.05
0.5	0.7	0.14	1.00	0.29	0.64	0.21	0.27	0.14	0.15	0.08	0.08	0.05	0.05
0.5	0.9	0.07	1.00	0.25	0.49	0.18	0.17	0.10	0.09	0.05	0.05	0.03	0.03
0.7	0.1	0.06	1.00	0.27	1.13	0.21	0.20	0.12	0.11	0.05	0.06	0.03	0.03
0.7	0.3	0.13	1.00	0.31	1.70	0.24	0.32	0.14	0.17	0.08	0.09	0.05	0.05
0.7	0.5	0.16	1.00	0.32	1.15	0.24	0.34	0.15	0.18	0.09	0.10	0.06	0.06
0.7	0.7	0.14	1.00	0.30	0.88	0.24	0.31	0.16	0.17	0.09	0.09	0.06	0.05
0.7	0.9	0.08	1.00	0.28	0.47	0.22	0.21	0.13	0.11	0.07	0.06	0.04	0.03
0.9	0.1	0.06	1.00	0.44	2.45	0.37	0.83	0.27	0.23	0.15	0.12	0.06	0.06
0.9	0.3	0.13	1.00	0.39	5.02	0.35	0.91	0.26	0.35	0.15	0.18	0.08	0.10
0.9	0.5	0.16	1.00	0.38	5.09	0.34	1.26	0.25	0.37	0.16	0.21	0.10	0.12
0.9	0.7	0.15	1.00	0.36	4.03	0.33	0.85	0.26	0.35	0.17	0.19	0.10	0.11
0.9	0.9	0.08	1.00	0.37	3.65	0.35	0.68	0.26	0.24	0.16	0.12	0.09	0.06

Table 2: Root-mean-square error of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ for true ATE as a function of p_T and p_Y, in bivariate probit model simulations with no covariates and $\rho = 0$. For most values of p_T and p_Y, the RMSE of BP is much smaller than the RMSE for IV in the sample sizes below 3000, but the difference shrinks with larger sample sizes.

				Root-mean-square error									
				$n = 400$		$n = 1,000$		$n = 3,000$		$n = 10,000$		$n = 30,000$	
p_T	p_Y	Δ_{ATE}	$\frac{\Delta_{LATE}}{\Delta_{ATE}}$	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**
0.1	0.1	0.08	1.47	0.38	3.03	0.32	1.88	0.23	0.22	0.11	0.12	0.06	0.07
0.1	0.3	0.15	1.12	0.36	4.93	0.33	4.81	0.25	0.34	0.15	0.18	0.09	0.10
0.1	0.5	0.16	0.90	0.37	6.65	0.34	4.22	0.26	0.38	0.17	0.20	0.10	0.11
0.1	0.7	0.12	0.70	0.41	5.30	0.37	2.99	0.28	0.34	0.19	0.19	0.11	0.11
0.1	0.9	0.05	0.47	0.32	3.40	0.41	2.67	0.40	0.23	0.25	0.12	0.16	0.07
0.3	0.1	0.07	1.19	0.27	1.38	0.20	0.21	0.11	0.11	0.06	0.06	0.03	0.04
0.3	0.3	0.14	1.11	0.31	1.58	0.24	0.32	0.15	0.16	0.09	0.09	0.05	0.05
0.3	0.5	0.16	1.02	0.32	2.16	0.24	0.34	0.16	0.18	0.09	0.10	0.05	0.05
0.3	0.7	0.13	0.91	0.31	1.78	0.24	0.31	0.15	0.16	0.09	0.09	0.05	0.05
0.3	0.9	0.06	0.73	0.32	1.29	0.26	0.21	0.16	0.11	0.08	0.06	0.04	0.04
0.5	0.1	0.06	0.96	0.24	0.50	0.18	0.18	0.10	0.09	0.05	0.05	0.03	0.03
0.5	0.3	0.14	1.03	0.29	0.89	0.22	0.28	0.13	0.14	0.08	0.08	0.04	0.04
0.5	0.5	0.16	1.05	0.30	1.08	0.23	0.29	0.15	0.15	0.08	0.08	0.05	0.05
0.5	0.7	0.14	1.03	0.29	0.83	0.21	0.26	0.13	0.14	0.08	0.08	0.04	0.04
0.5	0.9	0.06	0.96	0.25	0.54	0.18	0.17	0.10	0.09	0.05	0.05	0.03	0.03
0.7	0.1	0.06	0.73	0.32	1.19	0.25	0.23	0.15	0.11	0.08	0.06	0.04	0.04
0.7	0.3	0.13	0.91	0.31	1.82	0.25	0.33	0.15	0.16	0.09	0.09	0.05	0.05
0.7	0.5	0.16	1.02	0.32	1.31	0.25	0.37	0.16	0.18	0.09	0.10	0.05	0.05
0.7	0.7	0.14	1.11	0.31	1.25	0.24	0.31	0.15	0.16	0.09	0.09	0.05	0.05
0.7	0.9	0.07	1.19	0.27	0.72	0.20	0.21	0.11	0.11	0.06	0.06	0.03	0.04
0.9	0.1	0.05	0.47	0.34	2.73	0.41	1.28	0.40	0.24	0.27	0.12	0.15	0.07
0.9	0.3	0.12	0.70	0.42	4.12	0.37	1.90	0.29	0.35	0.19	0.18	0.11	0.10
0.9	0.5	0.16	0.90	0.38	4.81	0.34	2.07	0.26	0.38	0.17	0.20	0.10	0.11
0.9	0.7	0.15	1.12	0.38	4.24	0.34	2.85	0.25	0.35	0.16	0.18	0.09	0.10
0.9	0.9	0.08	1.47	0.39	3.12	0.34	2.45	0.23	0.22	0.12	0.12	0.07	0.08

Table 3: Root-mean-square error of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ for true ATE as a function of p_T and p_Y, in bivariate probit model simulations with no covariates and $\rho = 0.3$.

				Root-mean-square error									
				$n = 400$		$n = 1,000$		$n = 3,000$		$n = 10,000$		$n = 30,000$	
p_T	p_Y	Δ_{ATE}	$\frac{\Delta_{LATE}}{\Delta_{ATE}}$	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**
0.1	0.1	0.08	1.90	0.39	2.55	0.30	2.33	0.19	0.22	0.09	0.13	0.05	0.10
0.1	0.3	0.15	1.17	0.39	3.58	0.33	5.51	0.24	0.33	0.15	0.18	0.09	0.10
0.1	0.5	0.16	0.75	0.38	4.92	0.34	4.74	0.28	0.37	0.19	0.20	0.12	0.12
0.1	0.7	0.12	0.44	0.38	3.78	0.42	3.10	0.34	0.35	0.23	0.19	0.15	0.12
0.1	0.9	0.05	0.18	0.18	2.82	0.24	3.12	0.35	0.23	0.41	0.13	0.31	0.08
0.3	0.1	0.07	1.31	0.26	1.82	0.19	0.21	0.10	0.11	0.05	0.06	0.03	0.04
0.3	0.3	0.14	1.25	0.31	2.01	0.22	0.29	0.14	0.16	0.08	0.09	0.05	0.06
0.3	0.5	0.16	1.06	0.31	2.12	0.24	0.33	0.16	0.17	0.09	0.09	0.05	0.06
0.3	0.7	0.13	0.82	0.31	2.30	0.24	0.32	0.16	0.16	0.09	0.09	0.05	0.06
0.3	0.9	0.06	0.47	0.30	1.78	0.32	0.21	0.23	0.11	0.11	0.07	0.06	0.04
0.5	0.1	0.06	0.84	0.26	0.50	0.18	0.17	0.10	0.09	0.05	0.05	0.03	0.03
0.5	0.3	0.13	1.09	0.29	0.98	0.21	0.26	0.13	0.14	0.07	0.07	0.04	0.04
0.5	0.5	0.16	1.15	0.29	1.11	0.22	0.27	0.14	0.15	0.08	0.08	0.05	0.05
0.5	0.7	0.13	1.09	0.28	0.81	0.21	0.25	0.12	0.13	0.07	0.08	0.04	0.04
0.5	0.9	0.06	0.84	0.26	0.49	0.18	0.17	0.11	0.09	0.05	0.05	0.03	0.03
0.7	0.1	0.06	0.47	0.31	1.03	0.31	0.21	0.23	0.11	0.12	0.07	0.06	0.04
0.7	0.3	0.13	0.82	0.31	2.61	0.25	0.31	0.16	0.16	0.09	0.09	0.05	0.05
0.7	0.5	0.16	1.06	0.32	1.32	0.25	0.33	0.16	0.17	0.09	0.09	0.05	0.05
0.7	0.7	0.14	1.25	0.31	0.87	0.23	0.29	0.14	0.16	0.08	0.09	0.05	0.06
0.7	0.9	0.07	1.31	0.26	0.50	0.17	0.20	0.09	0.11	0.05	0.06	0.03	0.04
0.9	0.1	0.05	0.18	0.19	2.37	0.25	0.96	0.36	0.23	0.44	0.13	0.31	0.08
0.9	0.3	0.12	0.44	0.39	4.19	0.41	1.20	0.34	0.35	0.24	0.19	0.15	0.12
0.9	0.5	0.16	0.75	0.40	4.66	0.35	1.25	0.28	0.37	0.19	0.20	0.12	0.12
0.9	0.7	0.15	1.17	0.39	3.94	0.34	1.35	0.24	0.34	0.15	0.18	0.09	0.11
0.9	0.9	0.08	1.90	0.37	2.76	0.30	0.58	0.19	0.22	0.09	0.13	0.05	0.09

Table 4: Root-mean-square error of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ for true ATE as a function of p_T and p_Y, in bivariate probit model simulations with no covariates and $\rho = 0.5$.

				Root-mean-square error									
				$n = 400$		$n = 1,000$		$n = 3,000$		$n = 10,000$		$n = 30,000$	
p_T	p_Y	Δ_{ATE}	$\frac{\Delta_{LATE}}{\Delta_{ATE}}$	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**
0.1	0.1	0.08	2.60	0.40	3.06	0.29	2.23	0.16	0.24	0.07	0.16	0.04	0.14
0.1	0.3	0.15	1.12	0.41	4.07	0.34	3.47	0.25	0.33	0.15	0.17	0.09	0.10
0.1	0.5	0.16	0.46	0.40	4.25	0.39	3.95	0.33	0.38	0.24	0.21	0.16	0.14
0.1	0.7	0.12	0.15	0.25	3.57	0.29	3.02	0.35	0.37	0.37	0.21	0.29	0.15
0.1	0.9	0.05	0.02	0.09	2.93	0.09	2.18	0.10	0.23	0.15	0.13	0.21	0.09
0.3	0.1	0.06	1.34	0.26	1.10	0.17	0.23	0.08	0.11	0.04	0.06	0.02	0.04
0.3	0.3	0.14	1.54	0.30	1.61	0.21	0.31	0.12	0.17	0.07	0.11	0.04	0.09
0.3	0.5	0.16	1.11	0.31	2.65	0.24	0.33	0.15	0.17	0.09	0.09	0.05	0.06
0.3	0.7	0.12	0.60	0.32	2.93	0.28	0.34	0.18	0.17	0.11	0.10	0.06	0.07
0.3	0.9	0.05	0.16	0.17	1.17	0.19	0.22	0.25	0.12	0.28	0.07	0.21	0.06
0.5	0.1	0.06	0.53	0.23	0.50	0.22	0.19	0.17	0.10	0.08	0.06	0.04	0.04
0.5	0.3	0.13	1.16	0.27	0.76	0.20	0.26	0.12	0.14	0.06	0.08	0.04	0.05
0.5	0.5	0.16	1.39	0.28	0.87	0.20	0.26	0.12	0.15	0.07	0.10	0.04	0.08
0.5	0.7	0.13	1.16	0.26	0.70	0.19	0.24	0.11	0.14	0.06	0.07	0.04	0.05
0.5	0.9	0.06	0.53	0.23	0.48	0.22	0.18	0.17	0.10	0.08	0.06	0.04	0.04
0.7	0.1	0.05	0.16	0.18	0.81	0.20	0.23	0.24	0.12	0.28	0.07	0.20	0.06
0.7	0.3	0.12	0.60	0.32	1.40	0.28	0.32	0.18	0.17	0.11	0.10	0.06	0.07
0.7	0.5	0.16	1.11	0.31	1.53	0.23	0.33	0.15	0.17	0.08	0.09	0.05	0.06
0.7	0.7	0.14	1.54	0.29	1.32	0.20	0.29	0.12	0.17	0.07	0.11	0.04	0.09
0.7	0.9	0.06	1.34	0.26	0.67	0.16	0.20	0.08	0.10	0.04	0.06	0.02	0.04
0.9	0.1	0.05	0.02	0.09	2.43	0.08	0.72	0.09	0.24	0.14	0.13	0.22	0.08
0.9	0.3	0.12	0.15	0.24	4.51	0.29	0.94	0.35	0.36	0.39	0.20	0.30	0.14
0.9	0.5	0.16	0.46	0.40	4.53	0.39	1.37	0.33	0.38	0.23	0.21	0.16	0.14
0.9	0.7	0.15	1.12	0.42	4.73	0.34	1.66	0.25	0.35	0.16	0.17	0.09	0.10
0.9	0.9	0.08	2.60	0.39	3.45	0.27	0.92	0.15	0.23	0.06	0.16	0.03	0.14

Table 5: Root-mean-square error of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ for true ATE as a function of p_T and p_Y, in bivariate probit model simulations with no covariates and $\rho = 0.7$.

				Root-mean-square error									
				$n = 400$		$n = 1,000$		$n = 3,000$		$n = 10,000$		$n = 30,000$	
p_T	p_Y	Δ_{ATE}	$\frac{\Delta_{LATE}}{\Delta_{ATE}}$	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**
0.1	0.1	0.08	1.35	0.20	5.87	0.14	5.89	0.08	0.30	0.04	0.15	0.03	0.09
0.1	0.3	0.14	1.09	0.25	8.04	0.17	21.61	0.10	0.47	0.06	0.23	0.03	0.13
0.1	0.5	0.15	0.92	0.26	10.12	0.18	18.59	0.11	0.49	0.06	0.24	0.03	0.14
0.1	0.7	0.12	0.77	0.27	11.15	0.18	34.14	0.10	0.48	0.05	0.23	0.03	0.14
0.1	0.9	0.05	0.56	0.21	3.06	0.18	19.59	0.09	0.31	0.04	0.16	0.02	0.10
0.3	0.1	0.07	1.13	0.18	3.19	0.13	0.39	0.08	0.14	0.05	0.07	0.03	0.04
0.3	0.3	0.13	1.07	0.24	8.31	0.18	1.04	0.11	0.22	0.06	0.11	0.03	0.06
0.3	0.5	0.15	1.01	0.25	25.13	0.17	0.98	0.10	0.23	0.06	0.12	0.03	0.07
0.3	0.7	0.12	0.93	0.23	42.05	0.15	0.47	0.09	0.23	0.05	0.11	0.03	0.07
0.3	0.9	0.06	0.79	0.18	39.50	0.10	0.30	0.05	0.15	0.03	0.08	0.02	0.05
0.5	0.1	0.06	0.97	0.18	4.46	0.12	0.93	0.07	0.12	0.04	0.07	0.02	0.04
0.5	0.3	0.13	1.01	0.26	18.72	0.18	1.64	0.11	0.19	0.06	0.10	0.04	0.05
0.5	0.5	0.15	1.02	0.28	45.66	0.20	1.82	0.12	0.20	0.07	0.10	0.04	0.06
0.5	0.7	0.13	1.01	0.25	12.23	0.19	1.24	0.11	0.19	0.06	0.10	0.04	0.06
0.5	0.9	0.06	0.96	0.17	10.47	0.11	0.85	0.07	0.13	0.04	0.07	0.02	0.04
0.7	0.1	0.06	0.80	0.19	24.85	0.11	0.82	0.06	0.14	0.03	0.08	0.02	0.04
0.7	0.3	0.12	0.93	0.23	36.60	0.15	1.09	0.08	0.22	0.05	0.11	0.03	0.06
0.7	0.5	0.15	1.02	0.26	49.65	0.18	1.09	0.10	0.23	0.05	0.12	0.03	0.07
0.7	0.7	0.13	1.07	0.25	56.82	0.18	0.83	0.11	0.22	0.06	0.11	0.03	0.07
0.7	0.9	0.07	1.13	0.18	41.57	0.13	0.90	0.08	0.15	0.05	0.08	0.03	0.05
0.9	0.1	0.05	0.58	0.23	4.47	0.20	1.85	0.11	0.31	0.04	0.15	0.02	0.09
0.9	0.3	0.12	0.77	0.27	7.58	0.17	2.96	0.10	0.47	0.05	0.24	0.03	0.13
0.9	0.5	0.15	0.94	0.26	10.60	0.18	2.96	0.10	0.51	0.06	0.24	0.03	0.14
0.9	0.7	0.14	1.11	0.25	9.90	0.18	2.62	0.11	0.47	0.06	0.23	0.03	0.14
0.9	0.9	0.08	1.36	0.21	7.01	0.14	2.21	0.08	0.31	0.04	0.15	0.02	0.09

Table 6: Root-mean-square error of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ for true ATE as a function of p_T and p_Y, in bivariate probit model simulations with covariate X and $\rho = 0$.

				Root-mean-square error									
				$n = 400$		$n = 1,000$		$n = 3,000$		$n = 10,000$		$n = 30,000$	
p_T	p_Y	Δ_{ATE}	$\frac{\Delta_{LATE}}{\Delta_{ATE}}$	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**
0.1	0.1	0.07	1.75	0.22	3.79	0.15	10.98	0.09	0.29	0.05	0.15	0.03	0.10
0.1	0.3	0.14	1.17	0.27	6.61	0.19	11.02	0.12	0.47	0.06	0.22	0.04	0.13
0.1	0.5	0.15	0.81	0.29	9.62	0.20	5.82	0.12	0.51	0.07	0.24	0.04	0.15
0.1	0.7	0.12	0.52	0.28	8.79	0.22	39.55	0.13	0.48	0.07	0.23	0.04	0.14
0.1	0.9	0.05	0.27	0.13	4.58	0.14	11.49	0.13	0.32	0.09	0.16	0.04	0.10
0.3	0.1	0.06	1.29	0.20	22.22	0.14	0.32	0.09	0.14	0.04	0.07	0.03	0.05
0.3	0.3	0.13	1.21	0.27	7.64	0.19	0.90	0.12	0.21	0.06	0.11	0.04	0.07
0.3	0.5	0.15	1.06	0.27	14.39	0.18	0.74	0.11	0.24	0.06	0.12	0.04	0.07
0.3	0.7	0.12	0.84	0.23	20.86	0.15	0.50	0.09	0.22	0.05	0.11	0.03	0.07
0.3	0.9	0.05	0.53	0.17	37.83	0.12	0.29	0.06	0.15	0.03	0.08	0.02	0.05
0.5	0.1	0.06	0.89	0.19	18.25	0.12	1.31	0.07	0.12	0.03	0.06	0.02	0.04
0.5	0.3	0.12	1.08	0.26	23.20	0.19	0.67	0.12	0.18	0.06	0.10	0.04	0.06
0.5	0.5	0.15	1.11	0.29	8.86	0.21	1.72	0.13	0.20	0.07	0.10	0.04	0.06
0.5	0.7	0.13	1.07	0.26	16.75	0.19	0.87	0.12	0.18	0.06	0.09	0.04	0.06
0.5	0.9	0.06	0.86	0.17	1.47	0.11	0.28	0.06	0.13	0.03	0.07	0.02	0.04
0.7	0.1	0.05	0.56	0.19	14.48	0.13	0.67	0.06	0.14	0.03	0.08	0.02	0.05
0.7	0.3	0.12	0.86	0.24	40.01	0.15	1.47	0.09	0.22	0.05	0.11	0.03	0.07
0.7	0.5	0.15	1.05	0.26	46.67	0.18	0.92	0.11	0.22	0.06	0.12	0.04	0.07
0.7	0.7	0.13	1.20	0.27	49.33	0.19	1.78	0.12	0.22	0.06	0.11	0.04	0.07
0.7	0.9	0.06	1.26	0.19	9.03	0.14	1.55	0.08	0.15	0.05	0.08	0.03	0.05
0.9	0.1	0.05	0.28	0.16	4.39	0.17	1.58	0.15	0.30	0.09	0.15	0.04	0.09
0.9	0.3	0.12	0.52	0.29	10.31	0.22	3.26	0.13	0.47	0.07	0.24	0.04	0.14
0.9	0.5	0.15	0.81	0.28	10.45	0.20	4.10	0.12	0.51	0.07	0.24	0.04	0.15
0.9	0.7	0.14	1.17	0.27	10.56	0.20	2.45	0.12	0.49	0.07	0.22	0.04	0.14
0.9	0.9	0.07	1.76	0.21	8.89	0.15	2.13	0.09	0.34	0.05	0.16	0.03	0.10

Table 7: Root-mean-square error of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ for true ATE as a function of p_T and p_Y, in bivariate probit model simulations with covariate X and $\rho = 0.3$.

				Root-mean-square error									
				$n = 400$		$n = 1,000$		$n = 3,000$		$n = 10,000$		$n = 30,000$	
p_T	p_Y	Δ_{ATE}	$\frac{\Delta_{LATE}}{\Delta_{ATE}}$	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**
0.1	0.1	0.07	2.13	0.24	10.15	0.16	13.46	0.09	0.29	0.05	0.16	0.03	0.12
0.1	0.3	0.14	1.19	0.29	9.15	0.20	10.95	0.13	0.47	0.07	0.23	0.04	0.13
0.1	0.5	0.15	0.65	0.31	15.02	0.23	10.69	0.14	0.51	0.08	0.24	0.05	0.15
0.1	0.7	0.12	0.32	0.27	7.99	0.24	22.77	0.18	0.50	0.10	0.24	0.06	0.15
0.1	0.9	0.05	0.15	0.10	4.87	0.09	14.08	0.10	0.32	0.12	0.16	0.11	0.10
0.3	0.1	0.06	1.38	0.22	10.52	0.15	0.31	0.09	0.14	0.04	0.07	0.02	0.05
0.3	0.3	0.13	1.36	0.28	17.03	0.19	0.66	0.12	0.21	0.06	0.12	0.04	0.08
0.3	0.5	0.15	1.10	0.26	35.91	0.18	0.54	0.11	0.23	0.06	0.11	0.04	0.07
0.3	0.7	0.12	0.75	0.23	9.86	0.15	0.49	0.09	0.23	0.05	0.12	0.03	0.07
0.3	0.9	0.05	0.34	0.15	24.94	0.12	0.29	0.09	0.15	0.04	0.09	0.02	0.05
0.5	0.1	0.05	0.75	0.18	17.90	0.12	0.88	0.06	0.12	0.03	0.07	0.02	0.04
0.5	0.3	0.12	1.13	0.26	9.02	0.19	1.10	0.11	0.18	0.06	0.09	0.04	0.06
0.5	0.5	0.15	1.23	0.29	8.71	0.21	1.26	0.13	0.19	0.07	0.10	0.04	0.07
0.5	0.7	0.12	1.12	0.25	7.44	0.18	0.84	0.11	0.18	0.06	0.10	0.03	0.06
0.5	0.9	0.05	0.72	0.18	1.66	0.11	0.51	0.06	0.13	0.03	0.07	0.02	0.04
0.7	0.1	0.05	0.36	0.15	4.83	0.13	0.42	0.08	0.14	0.04	0.08	0.02	0.05
0.7	0.3	0.12	0.76	0.23	19.62	0.15	0.99	0.09	0.22	0.05	0.11	0.03	0.07
0.7	0.5	0.15	1.10	0.26	9.00	0.18	1.87	0.11	0.22	0.06	0.11	0.04	0.07
0.7	0.7	0.13	1.35	0.27	28.15	0.19	2.07	0.12	0.21	0.07	0.11	0.04	0.08
0.7	0.9	0.06	1.34	0.21	6.61	0.15	1.70	0.08	0.15	0.04	0.08	0.02	0.05
0.9	0.1	0.05	0.17	0.10	4.84	0.10	2.63	0.11	0.30	0.11	0.16	0.10	0.10
0.9	0.3	0.11	0.33	0.26	10.61	0.24	3.27	0.18	0.50	0.10	0.24	0.06	0.15
0.9	0.5	0.15	0.66	0.30	12.18	0.23	3.01	0.14	0.54	0.08	0.25	0.05	0.15
0.9	0.7	0.14	1.19	0.28	10.92	0.20	4.55	0.12	0.54	0.07	0.22	0.04	0.13
0.9	0.9	0.07	2.15	0.23	7.39	0.17	3.29	0.09	0.34	0.05	0.17	0.03	0.12

Table 8: Root-mean-square error of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ for true ATE as a function of p_T and p_Y, in bivariate probit model simulations with covariate X and $\rho = 0.5$.

				Root-mean-square error									
				$n = 400$		$n = 1,000$		$n = 3,000$		$n = 10,000$		$n = 30,000$	
p_T	p_Y	Δ_{ATE}	$\frac{\Delta_{LATE}}{\Delta_{ATE}}$	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**	**BP**	**IV**
0.1	0.1	0.07	2.75	0.25	5.99	0.15	8.13	0.08	0.29	0.04	0.18	0.03	0.15
0.1	0.3	0.14	1.12	0.31	17.42	0.22	18.43	0.13	0.44	0.07	0.23	0.04	0.12
0.1	0.5	0.15	0.42	0.35	12.13	0.29	15.32	0.18	0.53	0.10	0.27	0.06	0.17
0.1	0.7	0.12	0.15	0.22	20.78	0.21	9.53	0.21	0.50	0.18	0.25	0.11	0.17
0.1	0.9	0.05	0.13	0.06	7.74	0.05	13.87	0.05	0.32	0.06	0.17	0.08	0.10
0.3	0.1	0.06	1.34	0.24	2.31	0.16	0.86	0.07	0.14	0.03	0.07	0.02	0.05
0.3	0.3	0.13	1.61	0.27	6.40	0.19	0.63	0.11	0.20	0.06	0.13	0.03	0.10
0.3	0.5	0.15	1.13	0.24	15.60	0.16	0.54	0.10	0.23	0.05	0.12	0.03	0.07
0.3	0.7	0.12	0.56	0.23	38.01	0.16	0.64	0.09	0.23	0.05	0.12	0.03	0.09
0.3	0.9	0.05	0.16	0.11	5.50	0.09	0.39	0.08	0.15	0.07	0.09	0.04	0.06
0.5	0.1	0.05	0.48	0.16	17.87	0.11	0.63	0.06	0.12	0.03	0.07	0.02	0.05
0.5	0.3	0.12	1.17	0.23	5.10	0.16	1.32	0.09	0.17	0.05	0.09	0.03	0.06
0.5	0.5	0.15	1.44	0.28	14.33	0.20	1.18	0.12	0.19	0.07	0.12	0.04	0.09
0.5	0.7	0.12	1.18	0.24	11.22	0.16	0.99	0.09	0.18	0.05	0.09	0.03	0.06
0.5	0.9	0.05	0.48	0.16	5.67	0.10	0.46	0.06	0.13	0.03	0.07	0.02	0.05
0.7	0.1	0.05	0.16	0.10	4.21	0.09	0.46	0.09	0.15	0.07	0.09	0.05	0.06
0.7	0.3	0.12	0.56	0.22	7.76	0.16	0.86	0.09	0.22	0.05	0.12	0.03	0.08
0.7	0.5	0.15	1.12	0.24	4.12	0.17	0.88	0.10	0.22	0.06	0.12	0.03	0.07
0.7	0.7	0.13	1.59	0.27	9.49	0.19	0.78	0.11	0.21	0.06	0.13	0.04	0.10
0.7	0.9	0.06	1.35	0.24	3.27	0.15	0.64	0.07	0.14	0.03	0.08	0.02	0.05
0.9	0.1	0.05	0.14	0.06	15.44	0.06	12.17	0.06	0.29	0.07	0.17	0.07	0.10
0.9	0.3	0.11	0.15	0.21	15.21	0.21	12.75	0.21	0.52	0.17	0.25	0.10	0.16
0.9	0.5	0.14	0.41	0.32	18.70	0.28	8.28	0.19	0.56	0.11	0.27	0.06	0.17
0.9	0.7	0.14	1.09	0.31	26.71	0.21	14.96	0.13	0.49	0.07	0.22	0.04	0.13
0.9	0.9	0.07	2.78	0.25	11.85	0.17	5.20	0.09	0.33	0.05	0.19	0.03	0.15

Table 9: Root-mean-square error of $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}^{IV}$ for true ATE as a function of p_T and p_Y, in bivariate probit model simulations with covariate X and $\rho = 0.7$.

References

Altonji, J., T. Elder, and C. Taber (2005). "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113(1): 151–184.

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.

Angrist, J., and J. Pischke (2009). *Mostly Harmless Econometrics*. Princeton University Press, Princeton.

Bhattacharya, J., D. Goldman, and D. McCaffrey (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.

Fagerland, M. W., D. W. Hosmer, and A. M. Bofin (2008). "Multinomial Goodness-of-fit Tests for Logistic Regression Models," *Statistics in Medicine*, 27(21): 4238–4253.

Firth, D. (1993). "Bias Reduction of Maximum Likelihood Estimates," *Biometrika*, 80(1): 27–38.

Freedman, D. A., and J. S. Sekhon (2010). "Endogeneity in Probit Response Models," *Political Analysis*, 18(2): 138–150.

Greene, W. (1998). "Gender Economics Courses in Liberal Arts Colleges: Further Results," *Journal of Economic Education*, 29(4): 291–300.

Heckman, J. J. (1978). "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46(6): 931–959.

Heckman, J. J., and E. J. Vytlacil (1999). "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96(8): 4730–4734.

Hosmer, D. W., and S. Lemeshow (1980). "Goodness of Fit Tests for the Multiple Logistic Regression Model," *Communications in Statistics*, 9(10): 1043–1069.

Imbens, G., and J. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2): 467–475.

Moffitt, R. A. (2001). "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice: Comment," *Journal of Business and Economic Statistics*, 19(1): 20–23.

Monfardini, C., and R. Radice (2008). "Testing Exogeneity in the Bivariate Probit Model: A Monte Carlo Study," *Oxford Bulletin of Economics and Statistics*, 70(2): 271–282.

Murphy, A. (2007). "Score Tests of Normality in Bivariate Probit Models," *Economics Letters*, 95(3): 374–379.

Osius, G., and D. Rojek (1992). "Normal Goodness-of-fit Tests for Multinomial Models with Large Degrees of Freedom," *Journal of the American Statistical Association*, 87(420): 1145–1152.

Pigeon, J. G., and J. F. Heyse (1999). "An Improved Goodness of Fit Statistic for Probability Prediction Models," *Biometrical Journal*, 41(1): 71–82.

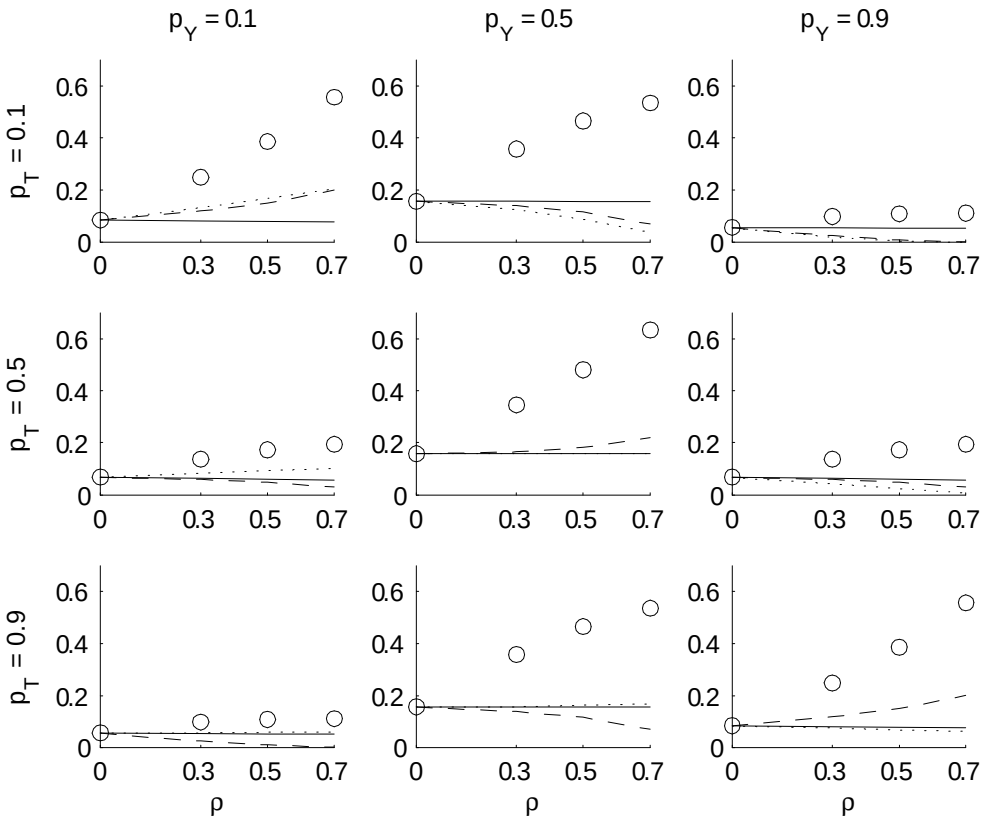

Figure 1: Δ_{ATE} (solid lines), Δ_{LATE} (long dashed lines), and Δ_{ATT} (dotted lines), for the bivariate probit model (7) with $\alpha = 0.3$, $\gamma = 0.4$, and several values of p_T, p_Y, and ρ. The circles denote Δ_{OLS}, the probability limit of an OLS regression of Y on T.

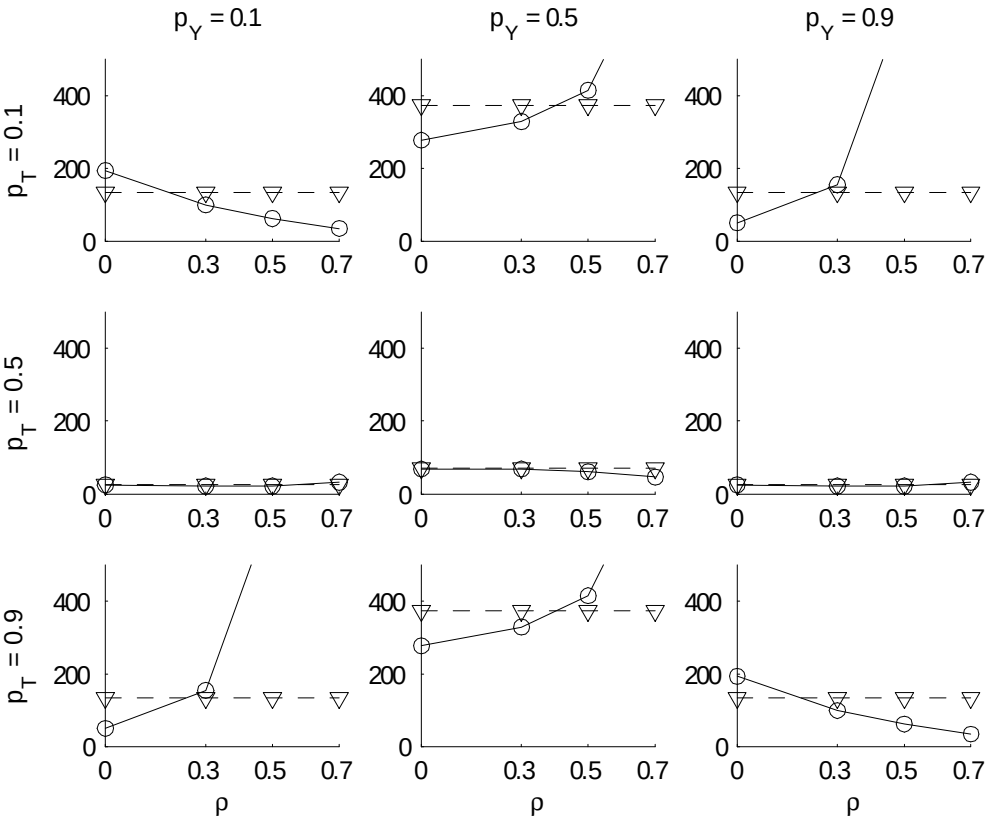

Figure 2: Asymptotic variance of $\hat{\Delta}_{BP}^{ATE}$ (solid lines with circles) and $\hat{\Delta}_{IV}$ (dashed lines with triangles), for various values of ρ, p_T, and p_Y. For example, an asymptotic variance of 200 means that at a given sample size n, the variance of the estimator is approximately $200/n$.

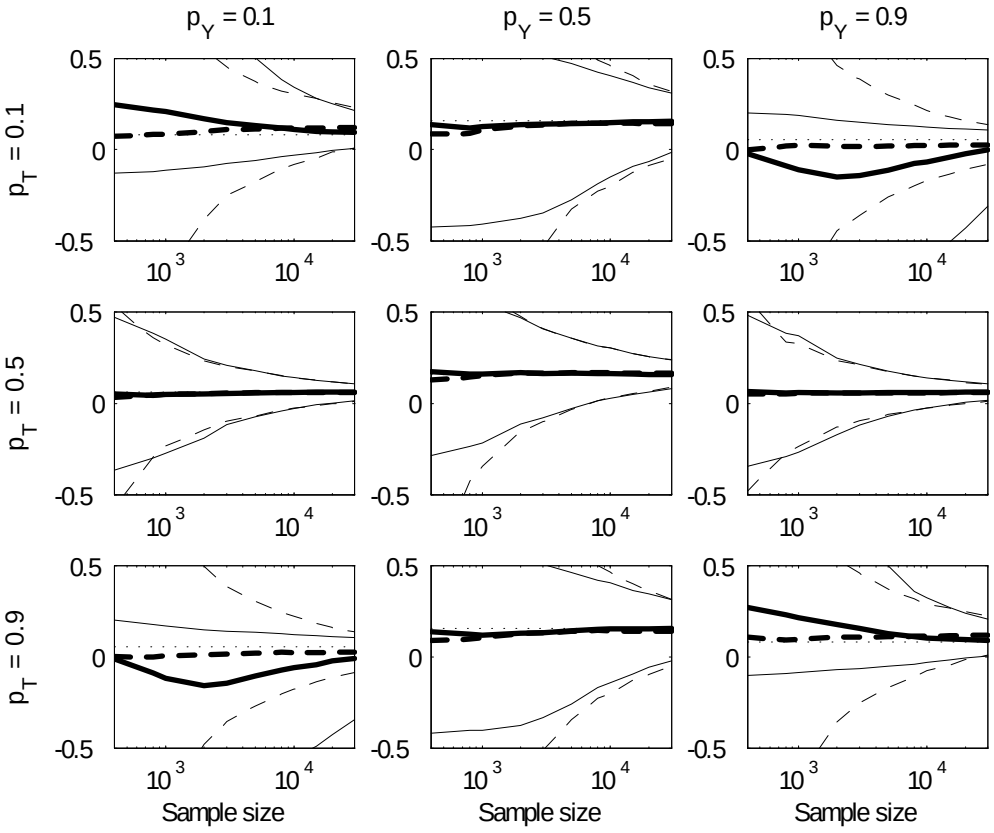

Figure 3: Spread of BP and IV estimates in simulations with no covariates and $\rho = 0.3$. The area between the thin solid curves represents the range between the 5th and 95th percentiles of the BP estimator, and the area between the thin dashed curves represents the same range for the IV estimator. 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.

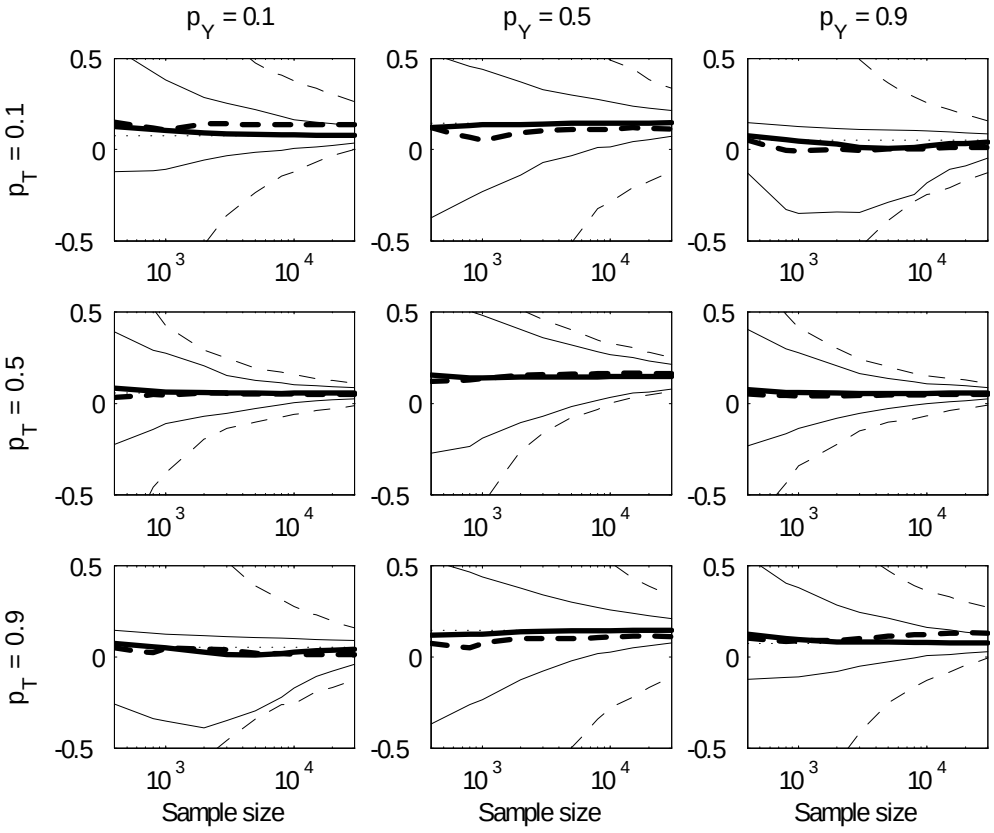

Figure 4: Spread of BP and IV estimates in simulations with covariate X and $\rho = 0.3$. The area between the thin solid curves represents the range between the 5th and 95th percentiles of the BP estimator, and the area between the thin dashed curves represents the same range for the IV estimator. The 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.

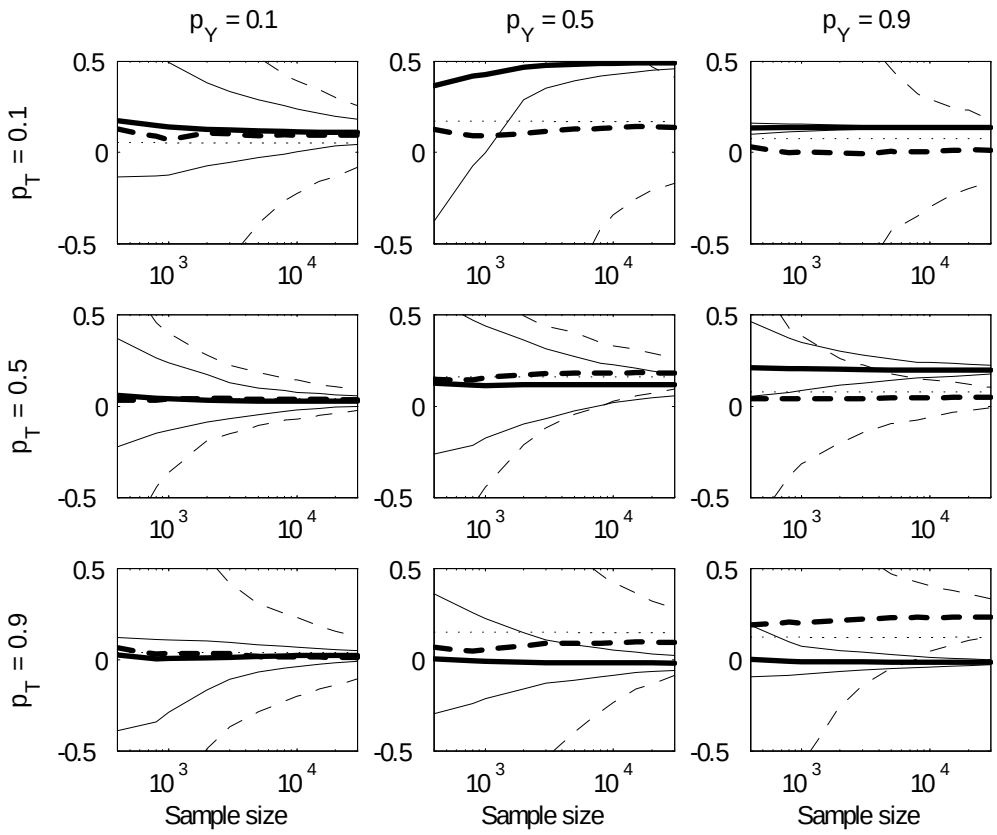

Figure 5: Spread of BP and IV estimates in simulations with covariate X and $\rho = 0.3$ and skewed error terms. The area between the thin solid curves represents the range between the 5th and 95th percentiles of the BP estimator, and the area between the thin dashed curves represents the same range 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.

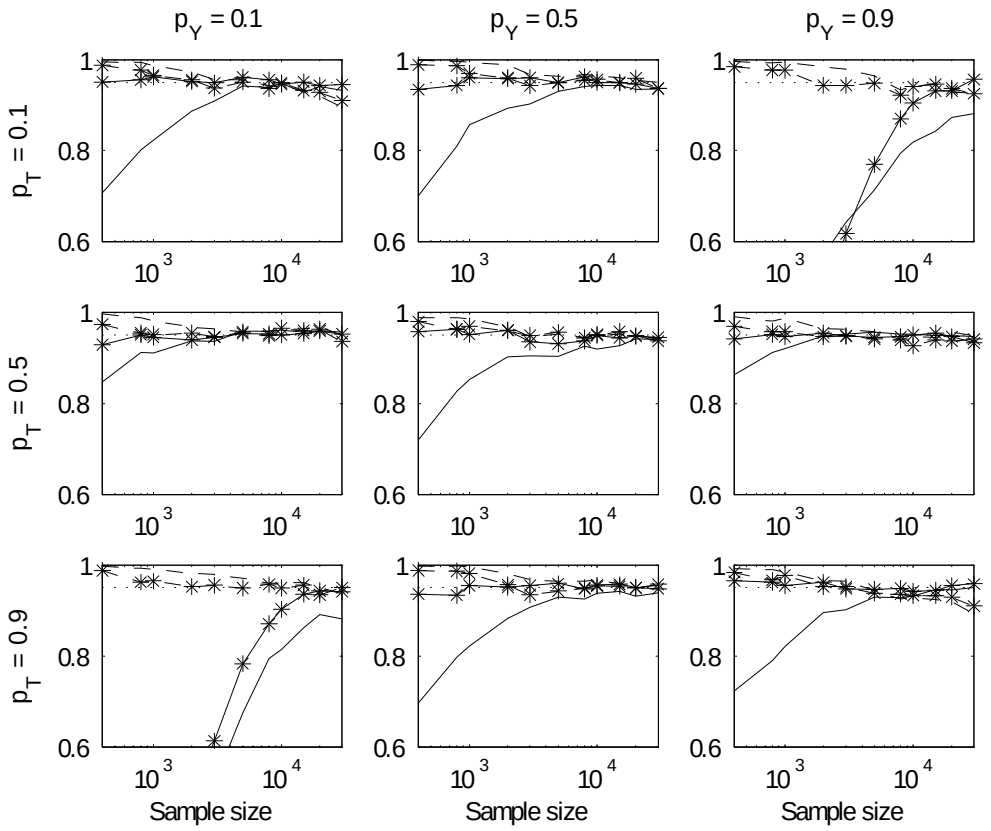

Figure 6: Coverage of the true Δ_{ATE} for nominal 95% confidence intervals in simulations with normally distributed covariate X_i and $\rho = 0.3$. The solid and dashed curves correspond to the size of tests based on $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}_{IV}$, respectively. The poor coverage can be improved by bootstrapping the critical values, and the size from bootstrapping for $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}_{IV}$ is shown by the starred solid and starred dashed curves, respectively.

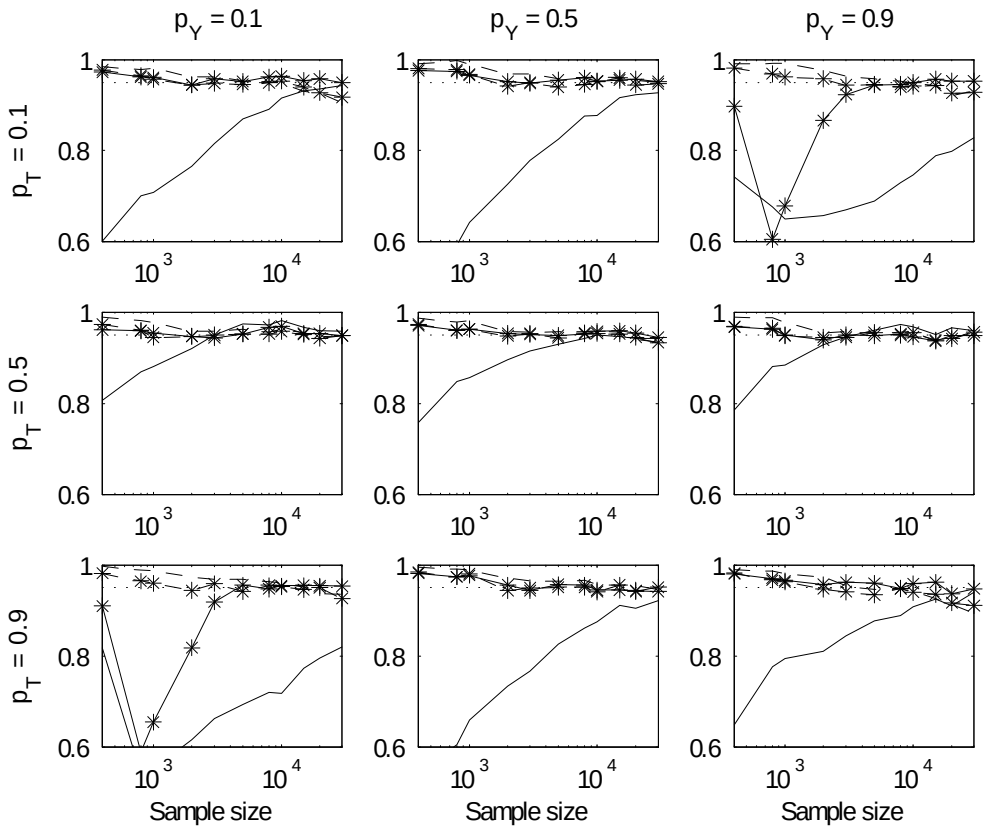

Figure 7: Coverage of the true Δ_{ATE} for nominal 95% confidence intervals in simulations with no covariates and $\rho = 0.3$. The solid and dashed curves correspond to the size of tests based on $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}_{IV}$, respectively. The poor coverage can be improved by bootstrapping the critical values, and the size from bootstrapping for $\hat{\Delta}^{BP}_{ATE}$ and $\hat{\Delta}_{IV}$ is shown by the starred solid and starred dashed curves, respectively.

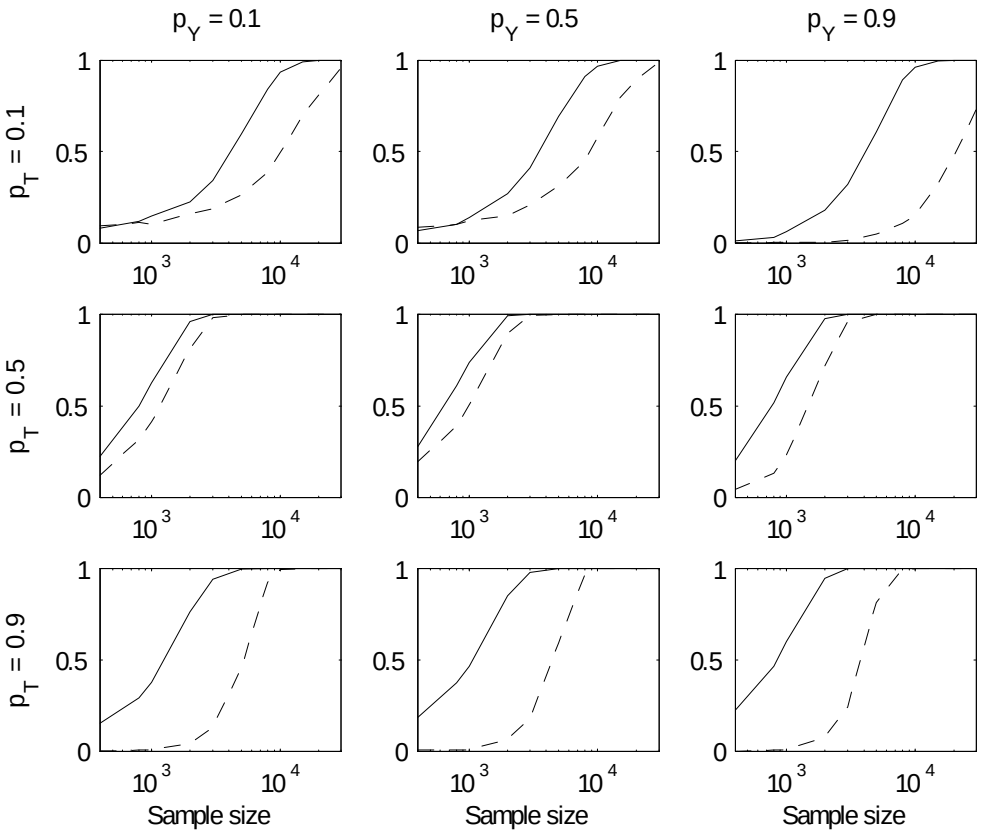

Figure 8: Power (rejection probability) of 5%-level Murphy score (solid curves) and adapted Hosmer-Lemeshow (dashed curves) goodness-of-fit tests for normality in simulations with covariate X and $\rho = 0.3$ and skewed error terms.