Econometric Reviews

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lecr20

Binary outcomes, OLS, 2SLS and IV probit

Chuhui Li, Donald S. Poskitt, Frank Windmeijer & Xueyan Zhao

To cite this article: Chuhui Li, Donald S. Poskitt, Frank Windmeijer & Xueyan Zhao (2022) Binary outcomes, OLS, 2SLS and IV probit, Econometric Reviews, 41:8, 859-876, DOI: 10.1080/07474938.2022.2072321

To link to this article: https://doi.org/10.1080/07474938.2022.2072321

Published online: 13 May 2022.

ECONOMETRIC REVIEWS
2022, VOL. 41, NO. 8, 859–876
https://doi.org/10.1080/07474938.2022.2072321

Taylor & Francis
Taylor & Francis Group

Binary outcomes, OLS, 2SLS and IV probit

Chuhui Li[a], Donald S. Poskitt[a], Frank Windmeijer[b], and Xueyan Zhao[a]

[a]Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia; [b]Department of Statistics and Nuffield College, University of Oxford, Oxford, UK

ABSTRACT

For a binary outcome Y, generated by a simple threshold crossing model with a single exogenous normally distributed explanatory variable X, the OLS estimator of the coefficient on X in a linear probability model is a consistent estimator of the average partial effect of X. Even in this very simple setting, we show that when allowing for X to be endogenously determined, the 2SLS estimator, using a normally distributed instrumental variable Z, does not identify the same causal parameter. It instead estimates the average partial effect of Z, scaled by the coefficient on Z in the linear first-stage model for X, denoted γ_1, or equivalently, it estimates the average partial effect of the population predicted value of X, $Z\gamma_1$. These causal parameters can differ substantially as we show for the normal Probit model, which implies that care has to be taken when interpreting 2SLS estimation results in a linear probability model. Under joint normality of the error terms, IV Probit maximum likelihood estimation does identify the average partial effect of X. The two-step control function procedure of Rivers and Vuong can also estimate this causal parameter consistently, but a double averaging is needed, one over the distribution of the first-stage error V and one over the distribution of X. If instead a single averaging is performed over the joint distribution of X and V, then the same causal parameter is estimated as the one estimated by the 2SLS estimator in the linear probability model. The 2SLS estimator is a consistent estimator when the average partial effect is equal to 0, and the standard Wald test for this hypothesis has correct size under strong instrument asymptotics. We show that, in general, the standard weak instrument first-stage F-test interpretations do not apply in this setting.

KEYWORDS

Binary outcomes; threshold crossing model; linear probability model; endogeneity; instrumental variables; two-stage least squares; weak instruments

JEL CLASSIFICATION

C13; C25; C26

1. Introduction

We consider estimation of model specifications for a binary dependent variable that is generated by the following simple threshold crossing specification

$$Y = 1(\beta_0 + \beta_1 X - U \geq 0),$$

where X is a normally distributed explanatory variable and U a continuous, zero mean error with pdf f_U. When X and U are independently distributed, then it is well known, following the results of Stoker (1986) (see also the discussion in (Wooldridge, 2010, p 579)), that the OLS estimator for δ_1 in the linear probability model (LPM)

$$Y = \delta_0 + \delta_1 X + \varepsilon,$$

CONTACT Frank Windmeijer frank.windmeijer@stats.ox.ac.uk Department of Statistics and Nuffield College, University of Oxford, Oxford, UK.

This article has been republished with minor changes. These changes do not impact the academic content of the article.

is a consistent and normal estimator of the average partial effect (APE) of X, which is denoted η_x and defined as

$$\eta_x = \beta_1 E_X[f_U(\beta_0 + \beta_1 X)].$$

Hence the OLS estimand in the LPM in this setup is a meaningful object with a clear interpretation.

Whilst this simple setup may have limited applicability, we use it to show that even in this setup one cannot simply extend this result to settings where X is endogenously determined so that $E[XU] \neq 0$, and where the parameters of the LPM are estimated by instrumental variable (IV) methods like two-stage least squares (2SLS), using a normally distributed IV Z. The 2SLS estimator for δ_1 in the linear probability model is then *not* a consistent estimator for the APE η_x, but, as detailed in Section 3, it estimates a different causal parameter, which is equal to the APE of Z, η_z, scaled by the linear effect γ_1 from the first-stage relationship, $X = Z\gamma_1 - V$. This causal parameter can also be interpreted as the APE of the population first-stage fitted value $\tilde{Z} = Z\gamma_1$. This highlights the fact that one cannot simply translate OLS and IV results developed for structural linear models to situations with a nonlinear data generating process, like the threshold model for binary dependent variables above. An exception is the case where $\beta_1 = 0$, as then the IV estimator is a consistent estimator of $\eta_x = 0$.

We apply and develop our findings further for the normal IV Probit model in Section 4. The model structure with normally distributed variables enables us to derive exact results. We establish in Section 4.1 that for the two-step control function estimation procedure of Rivers and Vuong (1988), a double averaging over the marginal distributions of the first-stage errors V and explanatory variable X leads to a consistent estimator of η_x. In contrast, a single averaging over the joint distribution of V and X, as for example proposed in (Wooldridge, 2010, p 589), is a consistent estimator of the same estimand as that of the 2SLS estimator, η_z/γ_1.

It is customary for applications of linear probability model IV regressions to report a robust F-statistic as a measure of instrument strength for the 2SLS estimation procedure, using critical values as tabulated by Stock and Yogo (2005) in relation to maximal relative bias of the 2SLS estimator, relative to that of the OLS estimator, see e.g., (Nunn and Qian, 2014, p. 1645). However, if a threshold crossing specification is appropriate for the data generating process, then OLS and 2SLS identify different causal parameters that can be very different from each other, and the Stock and Yogo (2005) weak instrument critical values for relative bias then do not apply. These critical values are further only valid under conditional homoskedasticity of both structural and first-stage errors. The robust F-statistic can then be used as a test for underidentification, but not for weak instruments, see also Andrews (2018) and Windmeijer (2021).

We present some Monte Carlo estimation results for the various estimators in Section 5, confirming our main findings. As the IV estimator is a consistent and normal estimator of η_x when $\eta_x = 0$, the 2SLS based Wald test for $H_0 : \delta_1 = 0$ has correct size under standard strong instrument asymptotics, and we present some power plots in Section 5.1. We further present some simulation results for the weak instrument size distortion of the IV Probit maximum likelihood estimator based Wald test in Section 5.2 and relate them to the values of the first-stage F statistic. This relationship is shown to be different from the standard linear model one as developed by Stock and Yogo (2005).

2. OLS in LPM for binary response model with exogenous normal explanatory variable

Following the setup and notation as in Arellano (2008), we consider a binary outcome Y, an explanatory variable X and unit variance continuous error U, related via a binary index model

$$Y = 1(\beta_0 + \beta_1 X - U \geq 0). \tag{1}$$

Potential outcomes, setting $X = x$, in this model are given by

$$Y(x) = 1(\beta_0 + \beta_1 x - U \geq 0)$$

and so the effect on the potential outcome of a change from x to x' for an individual with error U is given by

$$Y(x') - Y(x) = 1\big(\beta_0 + \beta_1 x' - U \geq 0\big) - 1(\beta_0 + \beta_1 x - U \geq 0).$$

The average effect, over the distribution of U, is then given by

$$E_U\big[Y(x') - Y(x)\big] = F_U\big(\beta_0 + \beta_1 x'\big) - F_U(\beta_0 + \beta_1 x),$$

where F_U is the cdf of U. For continuous X, the average marginal effect at $X = x$ is given by

$$\eta(x) = \frac{dE_U[Y(x)]}{dx} = \frac{dF_U(\beta_0 + \beta_1 x)}{dx} = \beta_1 f_U(\beta_0 + \beta_1 x),$$

where f_U is the pdf of U.

A potential object of interest is the average partial effect, which is the mean of the average marginal effects, taken over the distribution of X, and given by

$$\eta_x = \beta_1 E_X[f_U(\beta_0 + \beta_1 X)]. \tag{2}$$

Throughout we will investigate the properties of various estimators in this simple model design with a normally distributed X, as specified in the following assumption.

Assumption 1. $X \sim N(0, \sigma_x^2)$.

Note that setting $E[X] = 0$ is without loss of generality, as a constant is included in the model. Consider the linear probability model specification

$$Y = \delta_0 + \delta_1 X + \varepsilon. \tag{3}$$

We will first establish what quantity the OLS estimator of δ_1 estimates when the model for Y is given by (1), under the exogeneity assumption that U is independent of X.

Assumption 2. *Exogeneity,* $U \perp X$.

Under Assumption 2, if follows that $E[Y|X] = F_U(\beta_0 + \beta_1 X)$. The OLS estimand $\delta_{1,OLS}$ in (3) is then given by

$$\delta_{1,OLS} = \frac{Cov(X, Y)}{Var(X)} = \frac{E_X[XE[Y|X]]}{\sigma_x^2} = \frac{E_X[XF_U(\beta_0 + \beta_1 X)]}{\sigma_x^2}. \tag{4}$$

The equivalence of $\delta_{1,OLS}$ and η_x as defined in (2) under Assumptions 1 and 2 follows from the results in Stoker (1986), see also the discussion in (Wooldridge, 2010, p 579). We state this formally as a result here, with a proof provided in Appendix A for later reference.

Result 1. *Consider the binary outcome model (1) with X and U satisfying Assumptions 1 and 2. Then the OLS estimand* $\delta_{1,OLS}$ *as defined in (4) is equal to the average partial effect* η_x *as defined in (2),* $\delta_{1,OLS} = \eta_x$.

The implication is that the OLS estimator of δ_1 in the linear probability model (3) with normally distributed X is a consistent estimator of the APE η_x, a meaningful object. Note this is a robust finding, in the sense that it holds for any regular cdf F_U.

2.1. Consistency of method of moments

Following the results in Ruud (1983) and Stoker (1986), the robust consistency property of the OLS estimator for estimating η_x when X is normally distributed generalizes to the situation where

the researcher specifies a general regular cdf $G(.) \neq F_U(.)$, and when the parameters are estimated by the method of moments (MM). For a sample $\{Y_i, X_i\}_{i=1}^n$, the MM estimator solves $\sum_{i=1}^n (1, X_i)'(Y_i - G(\hat{\beta}_0 + \hat{\beta}_1 X_i)) = 0$. Following the results in (Ruud, 1983, pp. 226–227), the MM estimator converges to β_0^*, β_1^* when X is normally distributed, with

$$\begin{array}{ll} E_X\left[G\left(\beta_0^* + \beta_1^* X\right) - F_U(\beta_0 + \beta_1 X)\right] & = 0; \\ E_X\left[\left(G\left(\beta_0^* + \beta_1^* X\right) - F_U(\beta_0 + \beta_1 X)\right)X\right] & = 0, \end{array}$$

and then $\beta_1^* E_X(g(\beta_0^* + X\beta_1^*)) = \eta_x$, where $g(s) = dG(s)/ds$. This robustness property therefore holds for the Logit model with the parameters estimated by maximum likelihood (ML) as the Logit ML estimator is an MM estimator. For the Probit model it holds when estimated by MM, but not when estimated by ML, see Appendix B for further details.

The general result of Theorem 1 in (Stoker, 1986, p. 1465) states that $\eta_x = E[l(X)Y] = E_X[l(X)F_U(\beta_0 + \beta_1 X)]$ and $E_X\left[\frac{dG(\beta_0^* + \beta_1^* X)}{dX}\right] = E\left[l(X)G(\beta_0^* + \beta_1^* X)\right]$, where $l(X) = d\log f_X(X)/dX$, with $f_X(X)$ the marginal density of X. It follows that the method of moments estimation method for the APE described above results in a consistent estimator for η_x if $l(X) = aX$ with a some constant. For $X \sim N(0, \sigma_x^2)$, we have that $a = \sigma_x^{-2}$, satisfying this condition. It is clear from (4) that for the OLS estimator to be consistent for η_x it needs to be the case that $a = \sigma_x^{-2}$.

For the case of multiple explanatory variables, (Ruud, 1983, 1986) derived the conditions for the distribution of X under which the slope parameter vector β is consistently estimated up to scale for general user specified $G(.)$ when estimated by ML, implying that the ratios β_j/β_k are estimated consistently. These conditions are given by

$$E\left[X|X'\beta = c\right] = \theta_0 + \theta_1 c, \tag{5}$$

where θ_0 and θ_1 are k_x-vectors, with k_x the dimension of X. Estimating β consistently up to scale does not necessarily imply that the APEs are consistently estimated. This is clear from the result for the Probit ML estimator above, and is highlighted by (Stoker, 1986, p 1479) for the OLS estimator. The method of moments estimators will result in consistent estimators for the APEs if $l(X) = AX$ with A a constant matrix. For $X \sim N(0, \Sigma_X)$, we have that $A = \Sigma_X^{-1}$, which is again needed for the OLS estimator to be a consistent estimator for the APEs.

Ruud (1986) proposes data weighting schemes such that the distribution of the weighted regressors is close to satisfying condition (5), by weighting X such that the transformed data is approximately normal. Newey and Ruud (2005) propose density weighting schemes such that a weighted least squares (WLS) estimator is a consistent estimator for β up to scale. This is because the limit of the WLS estimator will behave as if X is distributed as multivariate normal. It follows then from the discussion above that the WLS estimator will be consistent for APEs, but these are the APEs for the weighted data, which may be difficult to interpret.

3. 2SLS in LPM for binary response model with normal instrument

Next, we allow for endogeneity, relaxing Assumption 2. We have a normally distributed instrument Z available, as specified in the following assumption.

Assumption 3. $Z \sim N(0, \sigma_z^2)$.

The first-stage relationship between X and Z is given by

$$X = \gamma_1 Z - V, \tag{6}$$

with $\gamma_1 \neq 0$, $Var(V) = \sigma_v^2$ and $Cov(U, V) = \sigma_{uv}$. Z is independent of both U and V:

Assumption 4. *Exogeneity of Z,* $Z \perp (U, V)$.

Then consider the IV or 2SLS estimator of δ_1 in the linear model (3). Its estimand is given by

$$\delta_{1,2sls} = \frac{Cov(Z,Y)}{Cov(Z,X)} = \frac{E_Z[ZE[Y|Z]]}{\gamma_1 \sigma_z^2}. \tag{7}$$

From (1) and (6) it follows that

$$\begin{aligned} Y &= 1(\beta_0 + \beta_1 X - U \geq 0) \\ &= 1(\beta_0 + \beta_1\gamma_1 Z - (U + \beta_1 V) \geq 0). \end{aligned}$$

Let $W_{\beta_1} = U + \beta_1 V$ with cdf $F_{W_{\beta_1}}$ and pdf $f_{W_{\beta_1}}$. We get for the potential outcomes, setting $Z = z$,

$$Y(z) = 1\big(\beta_0 + \beta_1\gamma_1 z - W_{\beta_1} \geq 0\big),$$

The average effect on the potential outcome of a change from z to z', over the distribution of W_{β_1}, is then given by

$$E_{W_{\beta_1}}\big[Y(z') - Y(z)\big] = F_{W_{\beta_1}}\big(\beta_0 + \beta_1\gamma_1 z'\big) - F_{W_{\beta_1}}(\beta_0 + \beta_1\gamma_1 z).$$

and the average marginal effect for z,

$$\eta(z) = \frac{dE_{W_{\beta_1}}[Y(z)]}{dz} = \frac{dF_{W_{\beta_1}}(\beta_0 + \beta_1\gamma_1 z)}{dz} = \beta_1\gamma_1 f_{W_{\beta_1}}(\beta_0 + \beta_1\gamma_1 z).$$

The mean of these average marginal effects over the distribution of Z is the average partial effect of Z, given by

$$\eta_z = \beta_1\gamma_1 E_Z\big[f_{W_{\beta_1}}(\beta_0 + \beta_1\gamma_1 Z)\big]. \tag{8}$$

The relationship between η_z and $\delta_{1,2sls}$ is given in the next proposition.

Proposition 1. *Consider the binary outcome model (1), with instrumental variable Z related to X as in (6) and satisfying Assumptions 3 and 4. Let η_z be as defined in (8) and the 2SLS estimand $\delta_{1,2sls}$ as in (7). Then $\delta_{1,2sls}$ is given by*

$$\delta_{1,2sls} = \beta_1 E_Z\big[f_{W_{\beta_1}}(\beta_0 + \beta_1\gamma_1 Z)\big] = \frac{\eta_z}{\gamma_1}.$$

Proof. Follows directly from expression (7) and the proof of Result 1, which establishes that $E_Z[ZE[Y|Z]]/\sigma_z^2 = \eta_z$.

An alternative way to interpret $\delta_{1,2sls}$ is that it is equal to the average partial effect of $\tilde{Z} = \gamma_1 Z$, the population fitted value of X given Z. This also follows directly from the results for the OLS estimator as presented in Section 2, because the 2SLS estimator is the OLS estimator in the linear probability model using the predicted values $\hat{X} = \hat{\gamma}_1 Z$, with $\hat{\gamma}_1$ the OLS estimator in the first-stage model 6. As $\hat{X} = \tilde{Z} + o_p(1)$, it follows that the OLS estimator estimates the average partial effect of $\hat{X}$ in large samples, which is the average partial effect of $\tilde{Z}$. It also follows from the discussion in Section 2.1 that the estimands for the Logit ML and Probit MM estimators using the predicted $\hat{X}$ as explanatory variable are equal to $\delta_{1,2sls}$.

It is clear that, in general, $\delta_{1,2sls} \neq \eta_x$, and we show in Fig. 1 in Section 4 below how different these estimands can be in a normal Probit specification as a function of σ_{uv}. When $\beta_1 = 0$, then $\delta_{1,2sls} = \eta_x$, as then $\eta_x = \eta_z = 0$. 2SLS estimation results can therefore be used to test the null hypothesis $H_0 : \delta_1 = 0$ which is equivalent to testing $H_0 : \eta_x = 0$.

It is also clear that the weak instruments testing procedure of Stock and Yogo (2005) based on the relative bias of the 2SLS estimator relative to that of the OLS estimator does not apply here, as the procedures estimate different causal parameters in general, with again the exception at

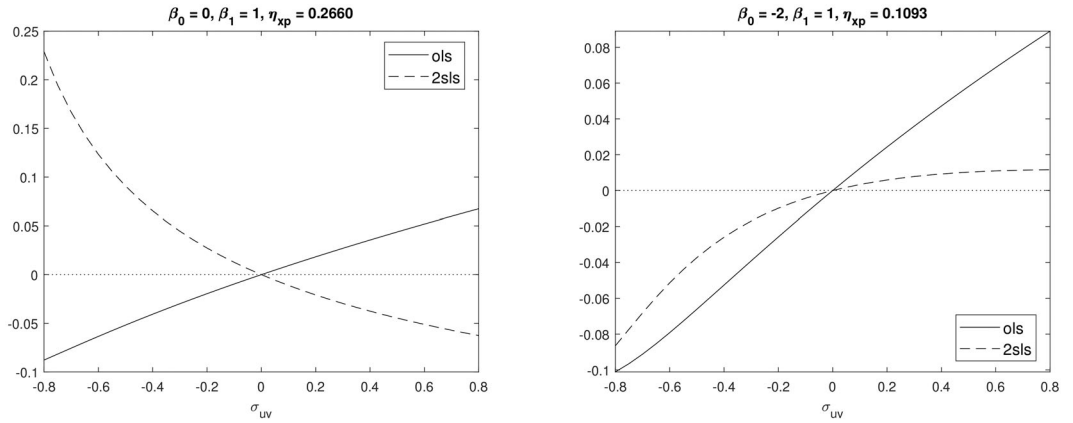

Figure 1. Bias $\delta_{1,ols} - \eta_{xp}$ and difference $\delta_{1,2sls} - \eta_{xp}$ as a function of σ_{uv}. $\sigma_x^2 = 1.25$.

$\beta_1 = 0$. In this simple one-variable model specification, when $\beta_1 = 0$ the model is a constant-only model, and hence the error in the LPM is then homoskedastic. The Stock and Yogo (2005) critical values for the first-stage F-statistic therefore apply in this case for both relative bias, when there are multiple instruments, and Wald test size distortion at $\beta_1 = 0$. However, when other exogenous variables are included in the model, the LPM error term is heteroskedastic and then these weak instruments critical values do not apply.

3.1. Consistent estimation of parameters up to scale

In a multivariate setting with X and Z multivariate normally distributed, it follows from the results as described in Section 2.1 for the consistent estimation of β up to scale, that the 2SLS estimator estimates the parameters consistently up to scale. This can be seen as follows. With the models specified, for $i = 1, ..., n$, as

$$Y_i = 1\left(\beta_0 + X_i'\beta - U_i \geq 0\right)$$
$$X_i = \Gamma' Z_i - V_i,$$

it follows that

$$Y_i = 1\left(\beta_0 + Z_i'\Gamma\beta - W_{\beta,i} \geq 0\right)$$

with $W_{\beta,i} = U_i + V_i'\beta$. As $Z_i \sim N(0, \Sigma_Z)$, it follows that $\Gamma' Z_i \sim N(0, \Gamma'\Sigma_Z\Gamma)$ and so for known Γ, the OLS estimator in the linear probability model using $\Gamma' Z_i$ as regressors is a consistent estimator of β up to scale. As Γ can be consistently estimated by the OLS estimator $\hat{\Gamma} = (Z'Z)^{-1}Z'X$, it follows that the OLS estimator, using $\widehat{X}_i = \hat{\Gamma}' Z_i = \Gamma' Z_i + o_p(1)$ as regressors, which is the 2SLS estimator, also consistently estimates β up to scale. The same then applies to the Probit and Logit ML estimators when using $\widehat{X}_i$ as the explanatory variables.

4. Probit model

In this section, we derive results that are specific to the Probit model, which specifies $U \sim N(0, 1)$. Under exogeneity Assumption 2, it follows that $E[Y|X] = \Phi(\beta_0 + \beta_1 X)$, where $\Phi(.)$ is the standard normal cdf. Because of symmetry, we can now write model (1) equivalently as $Y = 1(\beta_0 + \beta_1 X + U \geq 0)$.

The average partial effect of X in the Probit model is then given by

$$\eta_{xp} = \beta_1 E_X[\phi(\beta_0 + \beta_1 X)] \tag{9}$$

where $\phi(.)$ is the standard normal pdf.

For an iid sample $\{Y_i, X_i\}_{i=1}^n$, let $\hat{\beta}_{0p}$ and $\hat{\beta}_{1p}$ be the Probit ML estimators of β_0 and β_1. Then η_{xp} can be consistently estimated by

$$\hat{\eta}_{xp} = \hat{\beta}_{1p}\left(\frac{1}{n}\sum_{i=1}^{n}\phi\left(\hat{\beta}_{0p} + \hat{\beta}_{1p}X_i\right)\right). \tag{10}$$

With the density now fully specified in (9), we can further simplify the expression for η_{xp} using the following lemma:

Lemma 1. *Under the normality assumption of X, Assumption 1, it follows that*

$$E_X[\phi(\beta_0 + \beta_1 X)] = \frac{1}{\sqrt{1+\beta_1^2\sigma_x^2}}\phi\left(\frac{\beta_0}{\sqrt{1+\beta_1^2\sigma_x^2}}\right).$$

Proof. See Appendix A.

It follows then from Lemma 1 that

$$\eta_{xp} = \frac{\beta_1}{\sqrt{1+\beta_1^2\sigma_x^2}}\phi\left(\frac{\beta_0}{\sqrt{1+\beta_1^2\sigma_x^2}}\right). \tag{11}$$

Next, we allow for endogeneity. The relationship between X and instrument Z is given by

$$X = Z\gamma_1 + V, \tag{12}$$

with $\gamma_1 \neq 0$. The conditions for consistent and normal IV Probit maximum likelihood estimation are fulfilled:

Assumption 5.

$$\begin{pmatrix} U \\ V \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{uv} & 0 \\ \sigma_{uv} & \sigma_v^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix}\right).$$

Due to the endogeneity problem, $\sigma_{uv} \neq 0$, the Probit estimator $\hat{\eta}_{xp}$ as defined in (10) and the OLS estimator $\hat{\delta}_{1,OLS}$ are no longer consistent estimators of η_{xp}. From (12) and Assumption 5 it follows that

$$\begin{pmatrix} X \\ U \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{uv} \\ \sigma_{uv} & 1 \end{pmatrix}\right)$$

and so

$$U = \psi X + C,$$

where $\psi = \sigma_{uv}/\sigma_x^2$ and $C \sim N(0, 1-\tau^2)$, with $\tau = \sigma_{uv}/\sigma_x$. Therefore

$$E[Y|X] = \Phi\left(\frac{\beta_0 + (\beta_1 + \psi)X}{\sqrt{1-\tau^2}}\right) = \Phi\left(\tilde{\beta}_0 + \tilde{\beta}_1 X\right),$$

where $\tilde{\beta}_0 = \beta_0/\sqrt{1-\tau^2}$ and $\tilde{\beta}_1 = (\beta_1 + \psi)/\sqrt{1-\tau^2}$. It follows then from Lemma 1 that the standard Probit specification in this design with endogeneity estimates as average partial effect

$$\tilde{\eta}_{xp} = \tilde{\beta}_1 E_X\left[\phi(\tilde{\beta}_0 + \tilde{\beta}_1 X)\right] = \frac{\tilde{\beta}_1}{\sqrt{1+\tilde{\beta}_1^2\sigma_x^2}}\phi\left(\frac{\tilde{\beta}_0}{\sqrt{1+\tilde{\beta}_1^2\sigma_x^2}}\right), \tag{13}$$

which is the estimand of the Probit based estimator $\hat{\eta}_{xp}$. It follows directly from Result 1 that then also $\delta_{1,ols} = \tilde{\eta}_{xp}$. Clearly, $\tilde{\eta}_{xp} \neq \eta_{xp}$ unless $\sigma_{uv} = 0$, as then $\psi = \tau = 0$.

For this Probit specification with endogenous X and instrumental variable Z, we get the following expression for the 2SLS estimand $\delta_{1,2sls}$.

Proposition 2. *For model specifications* (1, 12) *and under Assumption 5, the 2SLS estimand* $\delta_{1,2sls}$ *for the 2SLS estimator of* δ_1 *in the linear probability model* (3) *as defined in* (7) *is given by*

$$\delta_{1,2sls} = \frac{\beta_1}{\sqrt{1+\beta_1^2\sigma_x^2 + 2\beta_1\sigma_{uv}}}\phi\left(\frac{\beta_0}{\sqrt{1+\beta_1^2\sigma_x^2 + 2\beta_1\sigma_{uv}}}\right). \tag{14}$$

Hence $\delta_{1,2sls} \neq \eta_{xp}$ *unless* $\beta_1 = 0$ *and/or* $\sigma_{uv} = 0$, *as the difference with the expression for* η_{xp} *as given in* (11) *is the additional term* $2\beta_1\sigma_{uv}$ *in the square-root denominator terms.*

Proof. See Appendix A.

As an illustration, Fig. 1 plots the bias $\delta_{1,ols} - \eta_{xp}$ and difference $\delta_{1,2sls} - \eta_{xp}$ as a function of σ_{uv}, for three different parameter settings with $\sigma_x^2 = 1.25$. The choice for σ_x^2 here is the same as that for the simulation results presented in Table 1 in Section 5 below. In the first plot, $\beta_0 = 0$ and $\beta_1 = 1$, resulting in $\eta_{xp} = 0.266$. For this design, the OLS bias is negative for negative values of σ_{uv} and positive for positive values of σ_{uv}. This is the opposite for the 2SLS difference, which has a positive (negative) difference for negative (positive) values of σ_{uv}. The difference of $\delta_{1,2sls}$ from η_{xp} is especially large at the more negative values of σ_{uv}. In the second plot we set the value of $\beta_0 = -2$, keeping the value of $\beta_1 = 1$. Due to the shift in the distribution, this results in smaller value $\eta_{xp} = 0.109$. The bias pattern of OLS is similar to that in the first plot, but now the 2SLS difference is in the same direction as the OLS bias, negative for negative values of σ_{uv} and positive for positive values of σ_{uv}. The 2SLS difference is substantially smaller than the OLS bias for positive values of σ_{uv}, but the two are quite similar for negative values of σ_{uv}. Note that $E[Y] = 0.5$ in the first plot, for all values of σ_{uv}, whereas the second plot refers to a rarer outcome, with $E[Y] = 0.006$ at $\sigma_{uv} = -0.8$, $E[Y] = 0.08$ at $\sigma_{uv} = 0$, and $E[Y] = 0.13$ at $\sigma_{uv} = 0.8$.

4.1. ML and two-step estimation

For an iid sample $\{Y_i, X_i, Z_i\}_{i=1}^n$, the IV Probit maximum likelihood estimator is consistent and normal under Assumption 5. The ML estimator for $\theta = (\beta_0, \beta_1, \gamma_0, \gamma_1, \rho, \sigma_v)$ is given by

$$\hat{\theta}_{ml} = \arg\min_{\theta} \sum_{i=1}^{n} \log\left(\Phi(G_i(\theta))^{Y_i}(1-\Phi(G_i(\theta)))^{1-Y_i}\frac{1}{\sigma_v}\phi\left(\frac{X_i-\gamma_0-\gamma_1 Z_i}{\sigma_v}\right)\right),$$

where

$$G_i(\theta) = \frac{\beta_0 + \beta_1 X_i + (\rho/\sigma_v)(X_i - \gamma_0 - \gamma_1 Z_i)}{\sqrt{(1-\rho^2)}},$$

with $\rho = \sigma_{uv}/\sigma_v$.

The causal average marginal effect at $X=x$ is therefore consistently estimated by

$$\hat{\eta}_{ml}(x) = \hat{\beta}_{1,ml}\phi\left(\hat{\beta}_{0,ml} + \hat{\beta}_{1,ml}x\right),$$

and the average partial effect by

$$\hat{\eta}_{x,ml} = \hat{\beta}_{1,ml}\left(\frac{1}{n}\sum_{i=1}^{n}\phi\left(\hat{\beta}_{0,ml} + \hat{\beta}_{1,ml}X_i\right)\right). \tag{15}$$

A popular alternative estimation method is the two-step control function approach of Rivers and Vuong (1988). From Assumption 5 it follows that

$$U = \omega V + W,$$

where $\omega = \sigma_{uv}/\sigma_v^2 = \rho/\sigma_v$, and $W \sim N(0, 1-\rho^2)$.

Therefore,

$$\begin{aligned} E[Y|X,V] &= \Phi\left(\frac{\beta_0 + \beta_1 X + \omega V}{\sqrt{1-\rho^2}}\right) \\ &= \Phi\left(\beta_{0\rho} + \beta_{1\rho}X + \omega_\rho V\right), \end{aligned}$$

where e.g., $\beta_{1\rho} = \beta_1/\sqrt{1-\rho^2}$.

Following (Wooldridge, 2010, p. 588), the average marginal effects at $X=x$ are obtained by taking derivatives of the average structural function $s(x)$, given by

$$s(x) = E_V\left[\Phi\left(\beta_{0\rho} + \beta_{1\rho}x + \omega_\rho V\right)\right].$$

Let $\hat{V}_i = X_i - \hat{\gamma}_0 - Z_i\hat{\gamma}_1$ be the OLS residual, then a consistent estimator of $s(x)$ is given by

$$\hat{s}_{2s}(x) = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\hat{\beta}_{0\rho} + \hat{\beta}_{1\rho}x + \hat{\omega}_\rho\hat{V}_i\right),$$

where $\hat{\beta}_{0\rho}$, $\hat{\beta}_{1\rho}$ and $\hat{\omega}_\rho$ are the standard Probit ML estimators in a Probit model for Y_i with X_i and $\hat{V}_i$ as explanatory variables. It therefore follows that a consistent estimator for the average marginal effect $\eta_p(x)$ at $X=x$ is given by

$$\hat{\eta}_{p,2s}(x) = \hat{\beta}_{1\rho}\left(\frac{1}{n}\sum_{i=1}^{n}\phi\left(\hat{\beta}_{0\rho} + \hat{\beta}_{1\rho}x + \hat{\omega}_\rho\hat{V}_i\right)\right). \tag{16}$$

A consistent estimator of the average partial effect η_{xp} is then given by

$$\hat{\eta}_{xp,2s} = \hat{\beta}_{1\rho}\left(\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\phi\left(\hat{\beta}_{0\rho} + \hat{\beta}_{1\rho}X_j + \hat{\omega}_\rho\hat{V}_i\right)\right). \tag{17}$$

In contrast, (Wooldridge, 2010, p. 589), proposes to estimate an average partial effect as

$$\hat{\alpha}_{2s} = \hat{\beta}_{1\rho}\left(\frac{1}{n}\sum_{i=1}^{n}\phi\left(\hat{\beta}_{0\rho} + \hat{\beta}_{1\rho}X_i + \hat{\omega}_\rho\hat{V}_i\right)\right). \tag{18}$$

This estimator $\hat{\alpha}_{2s}$ is a a consistent estimator of α, given by

$$\alpha = \beta_{1\rho}E_{X,V}\left[\phi\left(\beta_{0\rho} + \beta_{1\rho}X + \omega_\rho V\right)\right]. \tag{19}$$

The next proposition shows that α is equal to the 2SLS estimand $\delta_{1,2sls}$.

Proposition 3. *For the model specifications (1), (12) and under Assumption 5, let α be as defined in (19) and $\delta_{1,2sls}$ as in (14), then $\alpha = \delta_{1,2sls}$.*

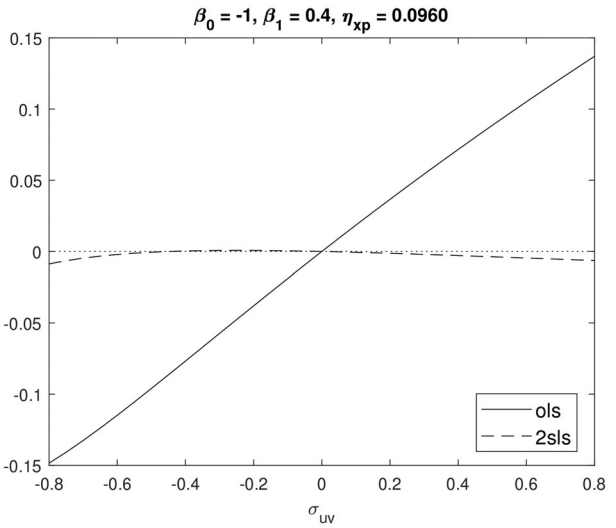

Figure 2. Bias $\delta_{1,ols} - \eta_{xp}$ and difference $\delta_{1,2sls} - \eta_{xp}$ as a function of σ_{uv}. $\sigma_x^2 = 1.25$.

Proof. See Appendix A.

There are parameter values for which the difference between η_{xp} and $\delta_{1,2sls}$ is small for the range of σ_{uv} values as considered above in Fig. 1. One example is given in Fig. 2, where $\sigma_x^2 = 1.25$, $\beta_0 = -1$ and $\beta_1 = 0.4$, resulting in a value $\eta_{xp} = 0.096$. As the value of $E[Y]$ changes in general with the value of σ_{uv}, everything else constant, it is not straightforward to determine whether one is in a small difference regime as a function of the values $E[Y]$, σ_x^2 and $\delta_{1,2sls}$, which can be estimated from the data. In the specific case considered here of the Probit model with normally distributed X and Z, an indicator for the bias of 2SLS is the difference between the values of $\hat{\delta}_{1,2sls}$ and $\hat{\eta}_{xp,2s}$, as further illustrated in Table 1 in Section 5 below.

5. Some Monte Carlo results

To illustrate the findings, we present results from a Monte Carlo exercise. The data are generated as,

$$\begin{pmatrix} U_i \\ V_i \\ Z_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right);$$
$$X_i = \gamma_1 Z_i + V_i;$$
$$Y_i = 1(\beta_0 + \beta_1 X_i + U_i \geq 0).$$

We draw samples of size $n = 500$, and set $\gamma_1 = 0.5$, and hence $\sigma_x^2 = 1.25$, $\beta_0 = 0$ and $\beta_1 = 1$. At these parameter values, the same as those for the first plot in Fig. 1, the average partial effect is given by $\eta_{xp} = 0.266$. We vary $\rho = -0.75, -0.5, ..., 0.75$, resulting in values for $\delta_{1,ols}$ ranging from 0.184 at $\rho = -0.75$ to 0.330 at $\rho = 0.75$, and values of $\delta_{1,2sls}$ ranging from 0.461 at $\rho = -0.75$, to 0.206 at $\rho = 0.75$. Table 1 presents estimation results for OLS, $\hat{\delta}_{1,ols}$; $\hat{\eta}_{xp}$ based on the standard Probit estimator, ignoring any endogeneity of X; 2SLS, $\hat{\delta}_{1,2sls}$; and $\hat{\alpha}_{2s}$ and $\hat{\eta}_{xp,2s}$ as defined in (18) and (16) based on the two-step Probit estimation results. The results clearly confirm the theoretical results obtained above. The observation that the standard deviation of $\hat{\eta}_{xp,2s}$ increases with the value of ρ is in line with the findings of (Zhang et al., 2020, pp. 11–12), who show that for a positive treatment effect β_1, the instrument identification power decreases with increasing values of ρ, everything else constant.

Table 1. Estimation results, $\eta_{xp} = 0.266$, $n = 500$.

ρ	−0.75	−0.5	−0.25	0	0.25	0.5	0.75
$\delta_{1,ols}$	0.184	0.214	0.241	0.266	0.289	0.310	0.330
$\hat{\delta}_{1,ols}$	0.184	0.214	0.242	0.267	0.289	0.310	0.330
	(0.016)	(0.014)	(0.013)	(0.012)	(0.011)	(0.011)	(0.011)
$\hat{\eta}_{xp}$	0.184	0.214	0.241	0.266	0.289	0.310	0.329
	(0.015)	(0.014)	(0.012)	(0.011)	(0.011)	(0.011)	(0.014)
$\delta_{1,2sls}$	0.461	0.357	0.302	0.266	0.241	0.221	0.206
$\hat{\delta}_{1,2sls}$	0.462	0.359	0.303	0.267	0.242	0.223	0.206
	(0.047)	(0.040)	(0.038)	(0.036)	(0.035)	(0.034)	(0.034)
$\hat{\alpha}_{2s}$	0.462	0.359	0.303	0.266	0.242	0.222	0.205
	(0.048)	(0.040)	(0.037)	(0.035)	(0.034)	(0.032)	(0.030)
$\hat{\eta}_{xp,2s}$	0.266	0.265	0.264	0.264	0.264	0.263	0.262
	(0.009)	(0.011)	(0.015)	(0.020)	(0.024)	(0.028)	(0.032)

Notes: Means and (standard deviations) of 1000 MC replications.

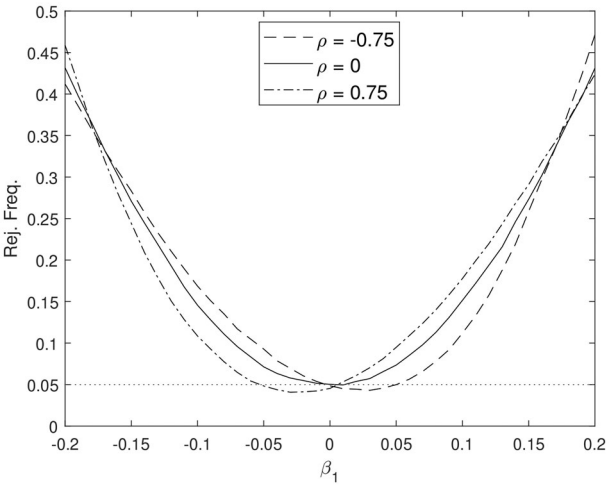

Figure 3. Rejection frequency of Wald test for $H_0 : \delta_1 = 0$ at 5% level.

Although our derivations are exact for models without other explanatory variables, it is quite common in practice for the OLS and Probit estimates of the APE to be the same for continuous, approximately normally distributed explanatory variables, also in the presence of additional explanatory variables that are possibly correlated with X, see e.g., the discussion in (Wooldridge, 2010, p. 579). We present some Monte Carlo estimation results in Appendix C for a model that further includes a binary and a χ^2_1 distributed explanatory variables, that are correlated with each other and with X. The results in terms of directions and magnitudes of bias/difference from η_{xp} are very similar to the ones presented here in Table 1.

5.1. 2SLS Wald test for $H_0 : \eta_x = 0$

We next analyze the behavior of the 2SLS Wald test for testing $H_0 : \eta_x = 0$, which is equivalent to testing $H_0 : \delta_1 = 0$. This Wald test is given by

$$W_{2sls} = \frac{\hat{\delta}^2_{1,2sls}}{\hat{Var}_r\left(\hat{\delta}_{1,2sls}\right)},$$

where $\hat{Var}_r(\hat{\delta}_{1,2sls})$ is a heteroskedasticity robust variance estimator. Fig. 3 shows the rejection frequencies of this test at the 5% level as a function of the value of β_1 for values of $\rho = -0.75; 0;$

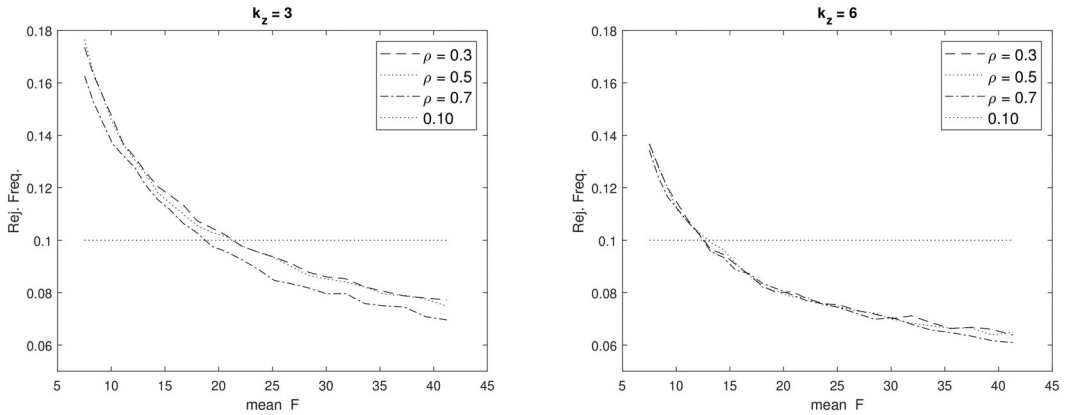

Figure 4. Rejection frequencies of Wald test for true $H_0 : \eta_x = \eta_{x0}$ at 5% level.

0.75. The other values of the parameters are as above, including the sample size of $n = 500$. The number of Monte Carlo replications for each value of β_1 is equal to 10, 000. We see that the test has correct size, but that power is affected by the bias of the 2SLS under the alternative. For example, for $\rho = -0.75$, there is a negative bias as shown in Table 1 and power is less than nominal size for small positive values of β_1.

5.2. ML estimator and weak instruments

Whilst the results obtained above for the 2SLS Wald test are limited to the value $\delta_1 = 0$, or $\eta_x = 0$, they are useful in practice as the hypothesis $H_0 : \eta_x = 0$ is often the main hypothesis of interest. For other values of η_x we need to consider different estimators and we next consider the ML IV estimator $\hat{\eta}_{x,ml}$ as defined in (15) that is a consistent and normal estimator of η_x if Assumption 1 for the Probit IV model holds.

The design is as in Table 1, with $\beta_1 = 1$, but here $n = 1000$ and $k_z = 3$ or $k_z = 6$ independent standard normally distributed instrumental variables. The first-stage is given by

$$X = cZ'\iota_{k_z} + V \tag{20}$$

where ι_{k_z} is a k_z-vector of ones. We vary the information content by varying the value of c in (20). We consider 3 values of $\rho = 0.3, 0.5$ and 0.7. The number of Monte Carlo replications is equal to 20, 000, for each value of c.

Fig. 4 shows the size behavior of the Wald test as a function of the mean value of the F-test statistic, testing $H_0 : \eta_x = \eta_{x0}$, for $k_z = 3$ in the left panel, and $k_z = 6$ in the right panel. Unlike the standard linear 2SLS case, the behavior of the test is not adversely affected by increasing values of ρ, and an increase in the number of instruments does improve the behavior of the test for the same value of the first-stage F-statistic. The critical values of the F-statistic for a weak instrument Wald test size of 10% at the 5% level is here approximately 29.40 for $k_z = 3$ and 17.30 for $k_z = 6$, thus bearing no relationship with the Stock and Yogo (2005) critical values for the standard linear 2SLS case of respectively 22.30 and 29.18, or the LIML critical values of respectively 9.61 and 5.61.

6. Conclusions

Binary outcome models are frequently used in empirical studies, and linear probability models are often specified and estimated by the two-stage least squares instrumental variables estimator

when the continuous treatment variable is endogenous. When the issue of weak instruments is a potential concern, the Stock and Yogo (2005) test, suitable for linear models, is then applied to detect a weak instrument problem. This article presents both theoretical and Monte Carlo results to show the implications and consequences of this popular linearization of binary models when the data generating process (DGP) is a threshold crossing latent equation model such as Probit.

From Stoker (1986) we have the result that when a normally distributed treatment variable X is exogenous, and the true DGP is an additive threshold crossing model, the OLS estimator for the coefficient of X in the LPM is a consistent estimator for η_x, the average partial effect (APE) of X, defined as the average marginal effect of X, averaged over the distribution of X.

The key results from the article can be summarized as follows:

1. When X is endogenous, the estimand $\delta_{1,2sls}$ for the 2SLS estimator of the LPM with normally distributed instrument Z is not the same as η_x. The estimand of the 2SLS estimator is in fact the APE of the population fitted value of X, $Z\gamma_1$.
2. For a normal Probit specification, the 2SLS estimand is shown to have a connection with the Rivers and Vuong (1988) two-step control function estimator. If the Rivers and Vuong (1988) two-step control function estimator is used, a single averaging over all observed pairs of X_i and estimated residual $\hat{V}_i$ gives a consistent estimate of the 2SLS estimand $\delta_{1,2sls}$, whilst double averaging over $\hat{V}_i$ and X_j estimates the APE of X, η_x.

We also present some results for the special case of $\beta_1 = 0$ and so the true APE of X is equal to zero, which has important implications for empirical researchers. We show that if $\beta_1 = 0$, $\delta_{1,2sls} = \eta_x = 0$, and so 2SLS is a consistent estimator for the APE of X. The 2SLS based Wald test for $H_0 : \delta_1 = 0$ then has correct size under standard strong instrument asymptotics.

These results have important implications for empirical practitioners, where the hypothesis the researchers often wish to test is $H_0 : \beta_1 = 0$, or there is a zero treatment effect of X. If they wish to use LPM-2SLS to estimate a threshold crossing model, and instruments are not weak, then the size of the Wald test is controlled. So if $H_0 : \beta_1 = 0$ is rejected, one can be relatively confident that the treatment variable X has a non-zero effect on outcome Y, but one cannot be certain about the estimated magnitude of the estimated APE of X, as when $\beta_1 \neq 0$, the 2SLS estimator and inference no longer apply to η_x in general. Therefore, the message for researchers is clear. When an endogenous continuous treatment variable is present for binary outcome variable models, the 2SLS estimator in the LPM does not estimate the APE of the treatment in general if the true model is a nonlinear threshold crossing model. The Stock and Yogo (2005) weak instrument tests also do not apply. See Frazier et al. (2019) for a proposed weak IV test for IV Probit type models, and Magnusson (2010) for weak IV robust inference.

Appendix A

Proofs

Proof. Result 1

$$\begin{aligned}
\delta_{1,OLS} &= \int_{-\infty}^{\infty} xF_U(\beta_0+\beta_1 x)f_X(x)dx/\sigma_x^2 \\
&= \frac{1}{\sqrt{2\pi\sigma_x^2}}\int_{-\infty}^{\infty} F_U(\beta_0+\beta_1 x)\frac{x}{\sigma_x^2}\exp\left(-\frac{x^2}{2\sigma_x^2}\right)dx \\
&= -\frac{1}{\sqrt{2\pi\sigma_x^2}}\int_{-\infty}^{\infty} F_U(\beta_0+\beta_1 x)\frac{d}{dx}\left(\exp\left(-\frac{x^2}{2\sigma_x^2}\right)\right)dx \\
&= -f_X(x)F_U(\beta_0+\beta_1 x)\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} f_X(x)\frac{d}{dx}(F_U(\beta_0+\beta_1 x))dx \\
&= \beta_1\int_{-\infty}^{\infty} f_U(\beta_0+\beta_1 x)f_X(x)dx \\
&= \beta_1 E_X[f_U(\beta_0+\beta_1 X)] \\
&= \eta_x.
\end{aligned}$$

Proof. Lemma 1

A standard integration result is

$$\int_{-\infty}^{\infty}\exp\left(-(bx+ax^2)\right) = \sqrt{\frac{\pi}{a}}\exp\left(\frac{b^2}{4a}\right),$$

with $a>0$. The result then follows as

$$\begin{aligned}
E_X[\phi(\beta_0+\beta_1 X)] &= \frac{1}{2\pi\sigma_x}\int_{-\infty}^{\infty}\exp\left(-\frac{(\beta_0+\beta_1 x)^2}{2}\right)\exp\left(-\frac{x^2}{2\sigma_x^2}\right) \\
&= \frac{1}{2\pi\sigma_x}\exp\left(-\frac{\beta_0^2}{2}\right)\int_{-\infty}^{\infty}\exp\left(-\left(\beta_0\beta_1 x+\left(\frac{1+\beta_1^2\sigma_x^2}{2\sigma_x^2}\right)x^2\right)\right) \\
&= \frac{1}{2\pi\sigma_x}\sqrt{\frac{\pi}{\frac{1+\beta_1^2\sigma_x^2}{2\sigma_x^2}}}\exp\left(-\frac{\beta_0^2}{2}\right)\exp\left(\frac{(\beta_0\beta_1)^2}{4\left(\frac{1+\beta_1^2\sigma_x^2}{2\sigma_x^2}\right)}\right) \\
&= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{1+\beta_1^2\sigma_x^2}}\exp\left(-\frac{\beta_0^2}{2\left(1+\beta_1^2\sigma_x^2\right)}\right) = \frac{1}{\sqrt{1+\beta_1^2\sigma_x^2}}\phi\left(\frac{\beta_0}{\sqrt{1+\beta_1^2\sigma_x^2}}\right).
\end{aligned}$$

Proof. Proposition 2

As the distribution of U is symmetric, we write the model as

$$\begin{aligned}
Y &= 1(\beta_0+\beta_1 X+U\geq 0) \\
&= 1(\beta_0+\beta_1\gamma_1 Z+(U+\beta_1 V)\geq 0).
\end{aligned}$$

It follows that

$$E[Y|Z] = \Phi\left(\frac{\beta_0+\beta_1\gamma_1 Z}{\sqrt{1+\beta_1^2\sigma_v^2+2\beta_1\sigma_{uv}}}\right) = \Phi\left(\beta_0^*+\beta_1^*\gamma_1 Z\right),$$

with

$$\beta_0^* = \frac{\beta_0}{\sqrt{1+\beta_1^2\sigma_v^2+2\beta_1\sigma_{uv}}};\quad \beta_1^* = \frac{\beta_1}{\sqrt{1+\beta_1^2\sigma_v^2+2\beta_1\sigma_{uv}}} \tag{A.1}$$

Therefore, from the proof of Proposition 1 and Lemma 1 it follows that

$$\begin{aligned}\delta_{1,2sls} &= \frac{E_Z\left[Z\Phi\left(\beta_0^* + \beta_1^*\gamma_1 Z\right)\right]}{\gamma_1\sigma_z^2} \\ &= \beta_1^* E_Z\left[\phi\left(\beta_0^* + \beta_1^*\gamma_1 Z\right)\right] \\ &= \frac{\beta_1^*}{\sqrt{1+\beta_1^{*}\gamma_1^2\sigma_z^2}}\phi\left(\frac{\beta_0^*}{\sqrt{1+\beta_1^{*2}\gamma_1^2\sigma_z^2}}\right).\end{aligned}$$

Then the result follows, as

$$\begin{aligned}\frac{\beta_1^*}{\sqrt{1+\beta_1^{*2}\gamma_1^2\sigma_z^2}} &= \frac{\beta_1}{\sqrt{1+\beta_1^2\sigma_v^2+2\beta_1\sigma_{uv}+\gamma_1^2\beta_1^2\sigma_z^2}} = \frac{\beta_1}{\sqrt{1+\beta_1^2\left(\gamma_1^2\sigma_z^2+\sigma_v^2\right)+2\beta_1\sigma_{uv}}} \\ &= \frac{\beta_1}{\sqrt{1+\beta_1^2\sigma_x^2+2\beta_1\sigma_{uv}}}\end{aligned}$$

and

$$\frac{\beta_0^*}{\sqrt{1+\beta_1^{*2}\gamma_1^2\sigma_z^2}} = \frac{\beta_0}{\sqrt{1+\beta_1^2\sigma_x^2+2\beta_1\sigma_{uv}}}.$$

Proof. Proposition 3

We can write

$$\begin{aligned}E_{X,V}\left[\phi\left(\beta_{0\rho}+\beta_{1\rho}X+\omega_\rho V\right)\right] &= E_{Z,V}\left[\phi\left(\beta_{0\rho}+\beta_{1\rho}\gamma_1 Z+\left(\beta_{1\rho}+\omega_\rho\right)V\right)\right] \\ &= E_Z\left[E_{V|Z}\left[\phi\left(\beta_{0\rho}+\beta_{1\rho}\gamma_1 Z+\left(\beta_{1\rho}+\omega_\rho\right)V\right)|Z\right]\right]\end{aligned}$$

From Lemma 1 it follows that

$$E_{V|Z}[\phi\left(\beta_{0\rho}+\beta_{1\rho}\gamma_1 Z+\left(\beta_{1\rho}+\omega_\rho\right)V\right)|Z] = \frac{1}{\sqrt{1+\left(\beta_{1\rho}+\omega_\rho\right)^2\sigma_v^2}}\phi\left(\frac{\beta_{0\rho}+\beta_{1\rho}\gamma_1 Z}{\sqrt{1+\left(\beta_{1\rho}+\omega_\rho\right)^2\sigma_v^2}}\right).$$

Further,

$$\begin{aligned}\left(\beta_{\rho 1}+\omega_\rho\right)^2\sigma_v^2 &= \left(\frac{\beta_1+\omega}{\sqrt{1-\rho^2}}\right)^2\sigma_v^2 = \frac{\beta_1^2+2\beta_1\omega+\omega^2}{1-\rho^2}\sigma_v^2 \\ &= \frac{\beta_1^2\sigma_v^2+2\beta_1\sigma_{uv}+\rho^2}{1-\rho^2},\end{aligned}$$

and so

$$1+\left(\beta_{1\rho}+\omega_\rho\right)^2\sigma_v^2 = \frac{1+\beta_1^2\sigma_v^2+2\beta_1\sigma_{uv}}{1-\rho^2}.$$

Therefore,

$$\beta_{1\rho}E_{V|Z}[\phi\left(\beta_{0\rho}+\beta_{1\rho}\gamma_1 Z+\left(\beta_{1\rho}+\omega_\rho\right)V\right)|Z] = \frac{\beta_1}{\sqrt{1+\beta_1^2\sigma_v^2+2\beta_1\sigma_{uv}}}\phi\left(\frac{\beta_0+\beta_1\gamma_1 Z}{\sqrt{1+\beta_1^2\sigma_v^2+2\beta_1\sigma_{uv}}}\right)$$

and hence

$$\alpha = \frac{\beta_1^*}{\gamma_1}E_Z\left[\phi\left(\beta_0^*+\beta_1^* Z\right)\right] = \delta_{1,2sls},$$

with β_0^* and β_1^* as defined in (A.1).

Appendix B

Robustness property of method of moments

From the model (1) specification we have the $E[Y|X] = F_U(\beta_0 + \beta_1 X)$. Following Ruud (1983), let $G(.)$ denote a regular cdf that the researcher assumes, then the ML estimator converges to β_0^*, β_1^* when X is normally distributed, with

$$E_X\left[h\left(\beta_0^* + \beta_1^* X\right)\left(G\left(\beta_0^* + \beta_1^* X\right) - F_U(\beta_0 + \beta_1 X)\right)\right] = 0 \tag{B.1}$$

$$E_X\left[h\left(\beta_0^* + \beta_1^* X\right)\left(G\left(\beta_0^* + \beta_1^* X\right) - F_U(\beta_0 + \beta_1 X)\right)X\right] = 0, \tag{B.2}$$

where

$$h\left(\beta_0^* + \beta_1^* X\right) = \frac{g\left(\beta_0^* + \beta_1^* X\right)}{G\left(\beta_0^* + \beta_1^* X\right)\left(1 - G\left(\beta_0^* + \beta_1^* X\right)\right)},$$

with $g(s) = dG(s)/ds$, see (Ruud, 1983, p. 227, Eqs. (12) and (13)).

For the Logit model, we have $G_L(s) = \exp(s)/(1 + \exp(s))$, so $h_L(s) = 1$, and hence

$$E_X\left[XG_L\left(\beta_0^* + \beta_1^* X\right)\right] = E_X[XF_U(\beta_0 + \beta_1 X)].$$

Following Result 1 and its proof in Appendix A, it follows that then

$$\begin{aligned} E_X[XF_U(\beta_0 + \beta_1 X)] &= \sigma_x^2 \eta_x = \sigma_x^2 \beta_1 E_X[f_U(\beta_0 + \beta_1 X)] \\ E_X\left[XG_L\left(\beta_0^* + \beta_1^* X\right)\right] &= \sigma_x^2 \beta_1^* E_X\left[g_L\left(\beta_0^* + \beta_1^* X\right)\right] \end{aligned}$$

and so

$$\beta_1^* E_X\left[g_L\left(\beta_0^* + \beta_1^* X\right)\right] = \eta_x.$$

Hence the standard Logit estimator of the APE is consistent.

For the Probit model, as

$$h_P\left(\beta_0^* + \beta_1^* X\right) = \frac{\phi\left(\beta_0^* + \beta_1^* X\right)}{\Phi\left(\beta_0^* + \beta_1^* X\right)\left(1 - \Phi\left(\beta_0^* + \beta_1^* X\right)\right)},$$

such robustness result cannot be obtained for the Probit ML procedure.

As the Logit ML estimator is equivalent to a method of moments (MM) estimator, we can extend its robustness property to the general MM estimation procedure. For any user specified regular cdf $G(.)$, it follows from the results in Ruud (1983), and analogously to (B.1) and (B.2), that the MM estimator that solves $\sum_{i=1}^n (1, X_i)'(Y_i - G(\hat{\beta}_0 + \hat{\beta}_1 X_i)) = 0$, converges to β_0^*, β_1^* when X is normally distributed, with

$$\begin{aligned} E_X\left[G\left(\beta_0^* + \beta_1^* X\right) - F_U(\beta_0 + \beta_1 X)\right] &= 0; \\ E_X\left[\left(G\left(\beta_0^* + \beta_1^* X\right) - F_U(\beta_0 + \beta_1 X)\right)X\right] &= 0, \end{aligned}$$

where the notation β_0^*, β_1^* is generic. It then follows from the exposition above that

$$E_X[XG\left(\beta_0^* + \beta_1^* X\right)] = \sigma_x^2 \beta_1^* E_X[g\left(\beta_0^* + \beta_1^* X\right)] = \sigma_x^2 \eta_x,$$

or

$$\beta_1^* E_X\left[g\left(\beta_0^* + \beta_1^* X\right)\right] = \eta_x.$$

We illustrate this with a small Monte Carlo exercise. We specify f_U as the student t-distribution with 3 degrees of freedom. Table B.1 shows the estimation results for η_x for OLS, Logit and Probit, with for the latter separate entries for the ML and MM estimators. The entries in the column labeled η_x are the mean and standard deviation of $\beta_1(\frac{1}{n}\sum_{i=1}^n f_U(\beta_0 + \beta_1 X_i))$. Only the Probit ML procedure results in a small downward bias.

Table B.1. Estimation results, $n = 2000$, $U \sim t_3$.

	η_x	OLS	Logit	Probit	
				ML	MM
mean	0.2840	0.2841	0.2840	0.2815	0.2840
st dev	0.0056	0.0051	0.0059	0.0058	0.0057

Notes: 10, 000 MC replications, $\beta_0 = 0$, $\beta_1 = 2$, $\sigma_x^2 = 1.5$.

Table C.1. Estimation results, $\eta_{xp} \approx 0.249$, $n = 500$.

ρ	−0.75	−0.5	−0.25	0	0.25	0.5	0.75
$\hat{\delta}_{1,\,ols}$	0.160	0.195	0.224	0.251	0.275	0.296	0.317
	(0.015)	(0.014)	(0.013)	(0.012)	(0.012)	(0.011)	(0.011)
$\hat{\eta}_{xp}$	0.160	0.194	0.224	0.250	0.274	0.296	0.316
	(0.015)	(0.014)	(0.013)	(0.012)	(0.012)	(0.012)	(0.015)
$\hat{\delta}_{1,\,2sls}$	0.401	0.324	0.280	0.251	0.230	0.212	0.197
	(0.049)	(0.041)	(0.038)	(0.035)	(0.034)	(0.033)	(0.032)
$\hat{\alpha}_{2s}$	0.400	0.323	0.279	0.250	0.229	0.211	0.197
	(0.046)	(0.040)	(0.036)	(0.034)	(0.032)	(0.030)	(0.028)
$\hat{\eta}_{xp,\,2s}$	0.237	0.241	0.245	0.248	0.250	0.251	0.251
	(0.011)	(0.013)	(0.016)	(0.020)	(0.023)	(0.028)	(0.030)

Notes: Means and (standard deviations) of 1000 MC replications.

Appendix C

Some Monte Carlo results with additional explanatory variables

Table C.1 presents some estimation results for a specification that includes additional, correlated, explanatory variables. One, D_i is a binary variable, the other C_i is χ^2_1 distributed, with its mean shifted if $D_i = 1$. The data generating process is given by

$$\begin{pmatrix} U_i \\ V_i \\ Z_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right);$$

$$P(D_i = 1) = 0.4$$

$$C_i = 0.2D_i + W_i^2; \; \square \; \square \; \mathrm{W_i} \sim \mathrm{N}(0, 1)$$

$$X_i = 0.5Z_i + 0.2D_i + 0.2C_i + V_i;$$

$$Y_i = 1(-0.5 + X_i + 0.2C_i + 0.2D_i + U_i \geq 0).$$

Estimation results for $n = 500$ for 1000 Monte Carlo replications are presented in Table C.1. The sample average partial effects of X, calculated as $\frac{1}{n}\sum_{i=1}^{n} \phi(-0.5 + X_i + 0.2C_i + 0.2D_i)$, for these samples has a mean of 0.249 (sd 0.006).

The results in terms of directions and magnitudes of bias/difference from η_{xp} are very similar to those presented in Table 1 for the model without additional explanatory variables. A difference is that $\hat{\eta}_{xp,\,2s}$ now has a small negative bias for $\rho < 0$, the bias increasing in absolute value as ρ becomes more negative.

Acknowledgments

We would like to thank Esfandiar Maasoumi, the editor, an anonymous associate editor and anonymous reviewer for their helpful comments. We acknowledge research funding from the Australian Research Council Discovery Grants DP140102345 and DP210103094.

References

Andrews, I. (2018). Valid two-step identification-robust confidence sets for GMM. *The Review of Economics and Statistics* 100(2):337–348. doi:10.1162/REST_a_00682

Arellano, M. (2008). *Binary models with endogenous explanatory variables*. Technical Report, CEMFI.

Frazier, D., Renault, E., Zhang, L., Zhao, X. (2019). Weak instrument test in discrete choice models. Paper presented at the *International Association for Applied Econometrics Conference*, June 25–28, Nicosia, Cyprus.

Magnusson, L. (2010). Inference in limited dependent variable models robust to weak identification. *The Econometrics Journal* 13(3):S56–S79. doi:10.1111/j.1368-423X.2009.00309.x

Newey, W. K., Ruud, P. A. (2005). Density weighted linear least squares. In D. W. K. Andrews and J. H. Stock, eds., *Identification and inference for econometric models*, Cambridge: Cambridge University Press, pp. 554–573.

Nunn, N., Qian, N. (2014). US food aid and civil conflict. *American Economic Review* 104(6):1630–1666. doi:10.1257/aer.104.6.1630

Rivers, D., Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics* 39(3):347–366. doi:10.1016/0304-4076(88)90063-2

Ruud, P. A. (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica* 51(1):225. doi:10.2307/1912257

Ruud, P. A. (1986). Consistent estimation of limited dependent variable models despite misspecification of distribution. *Journal of Econometrics* 32(1):157–187. doi:10.1016/0304-4076(86)90017-5

Stock, J. H., Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews and J. H. Stock, eds., *Identification and inference for econometric models*, Cambridge: Cambridge University Press, pp. 80–108.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* 54(6):1461. doi:10.2307/1914309

Windmeijer, F. (2021). Testing underidentification in linear models, with applications to dynamic panel and asset pricing models. *Journal of Econometrics*. doi:10.1016/j.jeconom.2021.03.007

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. 2nd ed. Cambridge: The MIT Press.

Zhang, L., Frazier, D. T., Poskitt, D. S., Zhao, X. (2020). Decomposing identification gains and evaluating instrument identification power for partially identified average treatment effects. Working Paper. arXiv:2009.02642v1.