Economics Letters 167 (2018) 104–107

Contents lists available at ScienceDirect

Economics Letters

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

A note on the different interpretation of the correlation parameters in the Bivariate Probit and the Recursive Bivariate Probit

Massimo Filippini [a,b], William H. Greene [c], Nilkanth Kumar [a], Adan L. Martinez-Cruz [d,a,*]

[a] *Center of Economic Research (CER-ETH), ETH Zürich, Switzerland*
[b] *Università della Svizzera italiana, Switzerland*
[c] *Department of Economics, NYU Stern School of Business, United States*
[d] *Department of Economics, Centro de Investigación y Docencia Económicas, Región Centro (CIDE-RC), México*

HIGHLIGHTS

- The correlation in an RBP does not capture the correlation of the endogenous variables.
- No extra arguments are needed if an RBP and BP deliver correlations of opposite sign.
- A zero-correlation parameter in a BP may mask the presence of a recursive system.

ARTICLE INFO

Article history:
Received 25 September 2017
Received in revised form 29 January 2018
Accepted 18 March 2018
Available online 27 March 2018

JEL classification:
C35

Keywords:
Bivariate probit
Recursive Bivariate probit
Tetrachoric correlation
Monte Carlo simulation

ABSTRACT

This note shows that, if a Bivariate Probit (BP) model is estimated on data arising from a Recursive Bivariate Probit (RBP) process, the resulting BP correlation parameter is a weighted average of the RBP correlation parameter and the parameter associated to the endogenous binary variable. Two corollaries follow this proposition: i) the interpretation of the correlation parameter in the RBP is not the same as in the BP —i.e. the RBP correlation parameter does not necessarily reflect the correlation between the binary variables under study; and ii) a zero correlation parameter in a BP model, usually interpreted as evidence of independence between the binary variables under study, may actually mask the presence of an RBP process.

1. Introduction

In modelling two jointly determined binary choices, empirical researchers usually resort to either a Bivariate Probit (BP) or a Recursive Bivariate Probit (RBP). The BP is a system of two seemingly unrelated probit equations in which the correlation between the binary variables under analysis is captured by the conditional tetrachoric correlation of the error terms (Greene, 2018). The RBP is a system of two probit equations that allows the errors terms to be correlated, and the binary dependent choice in one equation to be an endogenous regressor in the other equation.

In this paper, assuming that a BP model is estimated on data *truly* arising from a RBP process, we show that the resulting BP correlation parameter is a weighted average of the RBP correlation parameter and the parameter associated to the endogenous binary variable in the RBP. We discuss two implications of this result.

First, the interpretation of the RBP correlation parameter does not follow the interpretation of the BP correlation parameter. That is, the RBP correlation parameter does not necessarily capture the correlation between the binary variables under analysis — i.e. once the effect of the endogenous variable is taken into account, the correlation between the errors terms is not necessarily of the same sign as the endogenous relationship. We have identified some confusion on this point in empirical applications — with cases in which researchers have used behavioural arguments when an RBP delivers a correlation parameter with the opposite sign to the one yielded by a BP model (e.g., Kassouf and Hoffmann, 2006; Gitto et al., 2006).

Second, a zero or close to zero BP correlation parameter, usually interpreted as evidence that the binary variables can be modelled as independent of each other (e.g., Humphreys et al., 2014), may not always imply independence of the binary variables under

* Corresponding author at: Centro de Investigación y Docencia Económicas, Región Centro (CIDE-RC), Aguascalientes, México.
E-mail addresses: mfilippini@ethz.ch (M. Filippini), wgreene@stern.nyu.edu (W.H. Greene), nkumar@ethz.ch (N. Kumar), adan.martinez@cide.edu (A.L. Martinez-Cruz).

https://doi.org/10.1016/j.econlet.2018.03.018
0165-1765/

analysis. In particular, a zero or close to zero correlation parameter may result from erroneously estimating a BP model on RBP data.

We present our result, and illustrate its two implications through a series of Monte Carlo simulations and an empirical application.

2. Correlation parameter in a BP when the data follow a RBP process

Consider a *true* data-generating process that follows the Bivariate Probit with the recursive structure proposed by Maddala (1986):

$$y^*_{1i} = \boldsymbol{\beta}'_1 x_{1i} + \upsilon_{1i}, \quad y_{1i} = 1 \text{ if } y^*_{1i} > 0,\ y_{1i} = 0 \text{ otherwise}, \tag{1}$$

$$y^*_{2i} = \delta y_{1i} + \boldsymbol{\beta}'_2 x_{2i} + \upsilon_{2i}, \quad y_{2i} = 1 \text{ if } y^*_{2i} > 0,\ y_{2i} = 0 \text{ otherwise}, \tag{2}$$

$$[\upsilon_{1i}, \upsilon_{2i}] \sim \Phi_2[(0,0),(1,1),\zeta], \quad \zeta \in [-1,1]$$

where i is the individual index; y^*_{1i} and y^*_{2i} are latent continuous variables for which only the binary variables y_{1i} and y_{2i} are observable; x_{1i} and x_{2i} are vectors of exogenous variables[1] ; and $(\upsilon_{1i}, \upsilon_{2i})'$ is a vector of error terms described by Φ_2 — a bivariate standard normal distribution with correlation ζ.[2]

Assume now that an empirical researcher estimates a Bivariate Probit that misses the recursive structure of Eqs. (1) and (2); i.e.

$$y^*_{1i} = \boldsymbol{\beta}'_1 x_{1i} + \upsilon_{1i}, \quad y_{1i} = 1 \text{ if } y^*_{1i} > 0,\ y_{1i} = 0 \text{ otherwise}, \tag{3}$$

$$y^*_{2i} = \boldsymbol{\beta}'_2 x_{2i} + \varepsilon_{2i}, \quad y_{2i} = 1 \text{ if } y^*_{2i} > 0,\ y_{2i} = 0 \text{ otherwise}, \tag{4}$$

$$[\upsilon_{1i}, \varepsilon_{2i}] \sim \Phi_2[(0,0),(1,1),\rho], \quad \rho \in [-1,1]$$

where ρ is the correlation between υ_{1i} and ε_{2i}.

If the BP defined by Eqs. (3) and (4) is estimated on the data generated by Eqs. (1) and (2), then the true recursive component is absorbed by the error term of Eq. (4) which implies that ρ is mechanically determined by ζ and δ; i.e.,

$$\begin{aligned}\rho \equiv corr(\upsilon_{1i}, \varepsilon_{2i}) &= corr(\upsilon_{1i}, \delta y_{1i} + \upsilon_{2i}) \\ &= \frac{cov(\upsilon_{1i}, \delta y_{1i} + \upsilon_{2i})}{\sqrt{var(\upsilon_{1i}) var(\delta y_{1i} + \upsilon_{2i})}} \\ &= \frac{cov(\upsilon_{1i}, \delta y_{1i}) + cov(\upsilon_{1i}, \upsilon_{2i})}{\sqrt{var(\delta y_{1i}) + var(\upsilon_{2i}) + 2cov(\delta y_{1i}, \upsilon_{2i})}} \\ &= \frac{\delta cov(\upsilon_{1i}, y_{1i}) + \zeta}{\sqrt{\delta^2 var(y_{1i}) + 2\delta cov(y_{1i}, \upsilon_{2i}) + 1}}. \end{aligned} \tag{5}$$

Not surprisingly, according to Eq. (5) if $\delta = 0$ then $\rho = \zeta$ — i.e. in the absence of a recursive structure, the RBP collapses to the BP. Also, Eq. (5) shows that ρ can plausibly take value zero, depending on the signs and relative magnitude of ζ and δ — i.e. a BP model estimated on RBP data can potentially deliver a zero correlation parameter which would erroneously be interpreted as evidence of independence between y_1 and y_2. Implicit in the previous statement (and in the setting of the RBP process), no restrictions are imposed on the signs of ζ and δ. For instance, ζ and δ may have opposite signs. While ζ captures the correlation between the error terms υ_1 and υ_2, it does not reflect the correlation between y_1 and y_2. Such correlation is subsumed into δ — which implies that the interpretation of the RBP correlation does not resemble the interpretation of the BP correlation parameter. Section 3 illustrates the implications from Eq. (5).

3. Illustration

The Monte Carlo simulations in this section are designed to illustrate how the sign of $\hat{\rho}$ depends on the signs and values of both ζ and δ — i.e. we illustrate that ρ might be estimated at zero or close to zero even if ζ is not zero.

Also, we borrow data from Blasch et al. (2017) to illustrate that, in empirical applications, ζ does not necessarily reflect the direction of the correlation between y_1 and y_2.

3.1. Monte Carlo simulations

A pseudo-population of 100,000 individuals has been simulated according to the following recursive data-generating process:

$$y^*_{1i} = -2.00 + 0.10 z_{1i} + 0.90 z_{2i} + \upsilon_{1i}, \quad y_{1i} = 1 \text{ if } y^*_{1i} > 0,\ y_{1i} = 0 \text{ otherwise}, \tag{6}$$

$$y^*_{2i} = \delta y_{1i} - 1.00 + 1.20 z_{1i} - 0.20 z_{2i} + \upsilon_{2i}, \quad y_{2i} = 1 \text{ if } y^*_{2i} > 0,\ y_{2i} = 0 \text{ otherwise}, \tag{7}$$

where z_1 and z_2 are two exogenous variables. Realizations of z_1 are drawn from a binomial distribution with probability of success of 0.5; and realizations of z_2 are drawn from a normal distribution with mean 2 and unitary standard deviation.

Results reported in Table 1 illustrate the values of $\hat{\rho}$ that arise from erroneously estimating a BP on RBP data generated according to Eqs. (6) and (7). Each set of results in Table 1 arise from 1000 Monte Carlo replications.

The first panel of Table 1 illustrates that when no correlation between unobservables is present in the true RBP process ($\zeta = 0$),[3] then the correlation parameter in a BP model takes value zero when the parameter associated to the endogenous variable takes value zero — i.e. the sign and magnitude of $\hat{\rho}$ is determined by the sign and magnitude of δ. In the six scenarios of the first panel, ζ is assumed to be zero; from left to right δ takes values 2.00, 1.50, 0.40, 0.00, −1.50, and −2.00, respectively. Consistently with Eq. (5), $\hat{\rho}$ is positive when δ is positive; zero when $\delta = 0$; and negative when δ is negative. We also report $\hat{\zeta}$ and $\hat{\delta}$ to document that the correct RBP model yields estimates that reflect the true parameters.

The second panel of Table 1 illustrates that when ζ is large and positive (e.g., $\zeta = 0.80$), then $\hat{\rho}$ is large and positive when δ is also positive. However, $\hat{\rho}$ can take a value close to zero and even shift its sign if δ is negative and relatively large — in this scenario, $\delta \approx -1.50$ provokes the shift in sign. In the six scenarios of the second panel, $\zeta = 0.80$, and δ goes from 2.00 to −2.00 in a similar fashion as in the previous panel of results. Notice that $\hat{\rho} = 0.99$ when $\delta = 2.00$, and $\hat{\rho} = 0.90$ even when $\delta = 0.40$. A δ near −1.50 is needed to shift the sign of $\hat{\rho}$ — i.e. under this scenario, δ is required to be negative and large for a shift in sign to occur.

Following a similar reasoning as in the second panel, the third panel of Table 1 illustrates that a large and negative correlation between unobservables in the RBP process (e.g., $\zeta = -0.80$) translates into a large and negative $\hat{\rho}$ unless δ is large and positive.

[1] We have assumed that there is exogenous variation in both x_{1i} and x_{2i}, and that there is at least one exclusion. The model described by Eqs. (1) and (2) is generally identified even if $x_{1i} = x_{2i} = x_i$, granted enough variation is provided by the exogenous covariates in the model. In the narrow case in which $x_{1i} = x_{2i} =$ a dummy variable, the absence of an exclusion restriction may result in failure of identification. A detailed discussion of this case and on identification in the RBP model more generally may be found in Wilde (2000), Mourifié and Méango (2014) and Han and Vytlacil (2017). Further discussion on identification goes beyond the scope of this paper.

[2] A recursive structure is logically consistent with, for instance, the health production model in which individuals first engage in a healthy behaviour (y^*_1) in order to produce health (y^*_2) (see Humphreys et al., 2014).

[3] This scenario is not unseen in empirical applications. For instance, Greene (1998)'s RBP yields a zero correlation parameter.