Advance Access publication January 4, 2010 doi:10.1093/pan/mpp037

Endogeneity in Probit Response Models

David A. Freedman
Department of Statistics, University of California, Berkeley, CA 94720-3860
e-mail: freedman@stat.berkeley.edu

Jasjeet S. Sekhon
Department of Political Science, University of California, Berkeley, CA 94720-1950
e-mail: sekhon@berkeley.edu (corresponding author)

We look at conventional methods for removing endogeneity bias in regression models, including the linear model and the probit model. It is known that the usual Heckman two-step procedure should not be used in the probit model: from a theoretical perspective, it is unsatisfactory, and likelihood methods are superior. However, serious numerical problems occur when standard software packages try to maximize the biprobit likelihood function, even if the number of covariates is small. We draw conclusions for statistical practice. Finally, we prove the conditions under which parameters in the model are identifiable. The conditions for identification are delicate; we believe these results are new.

1 Introduction

Suppose a linear regression model describes responses to treatment and to covariates. If subjects self-select into treatment, the process being dependent on the error term in the model, endogeneity bias is likely. Similarly, we may have a linear model that is to be estimated on sample data; if subjects self-select into the sample, endogeneity becomes an issue.

Heckman (1978, 1979) suggested a simple and ingenious two-step method for taking care of endogeneity, which works under the conditions described in those papers. This method is widely used. Some researchers have applied the method to probit response models. However, the extension is unsatisfactory. The nonlinearity in the probit model is an essential difficulty for the two-step correction, which will often make bias worse. It is well-known that likelihood techniques are to be preferred—although, as we show here, the numerics are delicate.

In the balance of this article, we define models for (1) self-selection into treatment or control and (2) self-selection into the sample, with simulation results to delineate the statistical issues. In the simulations, the models are correct. Thus, anomalies in the behavior of estimators are not to be explained by specification error. Numerical issues are explored. We explain the motivation for the two-step estimator and draw conclusions for statistical

Authors' note: Derek Briggs, Allan Dafoe, Thad Dunning, Joe Eaton, Eric Lawrence, Walter Mebane, Jim Powell, Rocío Titiunik, and Ed Vytlacil made helpful comments. Errors and omissions remain the responsibility of the authors.

practice. We derive the conditions under which parameters in the models are identifiable; we believe these results are new. The literature on models for self-selection is huge, and so is the literature on probits; we conclude with a brief review of a few salient papers.

To define the models and estimation procedures, consider n subjects, indexed by $i = 1, \ldots, n$. Subjects are assumed to be independent and identically distributed. For each subject, there are two manifest variables X_i, Z_i and two latent variables U_i, V_i. Assume that (U_i, V_i) are bivariate normal, with mean 0, variance 1, and correlation ρ. Assume further that (X_i, Z_i) is independent of (U_i, V_i), that is, the manifest variables are exogenous. For ease of exposition, we take (X_i, Z_i) as bivariate normal, although that is not essential. Until further notice, we set the means to 0, the variances to 1, the correlation between X_i and Z_i to 0.40, and sample size n to 1000.

2 A Probit Response Model with an Endogenous Regressor

There are two equations in the model. The first is the selection equation:

$$C_i = 1 \text{ if } a + bX_i + U_i > 0, \text{else } C_i = 0. \tag{1}$$

In application, $C_i = 1$ means that subject i self-selects into treatment. The second equation defines the subject's response to treatment:

$$Y_i = 1 \text{ if } c + dZ_i + eC_i + V_i > 0, \text{else } Y_i = 0. \tag{2}$$

Notice that Y_i is binary rather than continuous. The data are the observed values of X_i, Z_i, C_i, Y_i. For example, the treatment variable C_i may indicate whether subject i graduated from college; the response Y_i, whether i has a full-time job.

Endogeneity bias is likely in (2). Indeed, C_i is endogenous due to the correlation ρ between the latent variables U_i and V_i. A two-step correction for endogeneity is sometimes used (although it should not be):

Step 1. Estimate the probit model (1) by likelihood techniques.

Step 2. To estimate (2), fit the expanded probit model

$$P(Y_i = 1|X_i, Z_i, C_i) = \Phi(c + dZ_i + eC_i + fM_i) \tag{3}$$

to the data, where

$$M_i = C_i\frac{\phi(a + bX_i)}{\Phi(a + bX_i)} - (1 - C_i)\frac{\phi(a + bX_i)}{1 - \Phi(a + bX_i)}. \tag{4}$$

Here, Φ is the standard normal distribution function with density $\varphi = \Phi'$. In application, a and b in (4) would be unknown. These parameters are replaced by maximum likelihood estimates (MLEs) obtained from Step 1. The motivation for M_i is explained in Section 6 below. Identifiability is discussed in Section 7: according to Proposition 1, parameters are identifiable unless $b = d = 0$.

The operating characteristics of the two-step correction were determined in a simulation study which draws 500 independent samples of size $n = 1000$. Each sample was constructed as described above. We set $a = 0.50$, $b = 1$, and $\rho = 0.60$. These choices create an environment favorable to correction.

Table 1 Simulation results

	c	d	e	ρ
True values				
	-1.0000	0.7500	0.5000	0.6000
Raw estimates				
Mean	-1.5901	0.7234	1.3285	
SD	0.1184	0.0587	0.1276	
Two-step				
Mean	-1.1118	0.8265	0.5432	
SD	0.1581	0.0622	0.2081	
MLE				
Mean	-0.9964	0.7542	0.4964	0.6025
SD	0.161	0.0546	0.1899	0.0900

Notes. Correcting endogeneity bias when the response is binary probit. There are 500 repetitions. The sample size is 1000. The correlation between latents is $\rho = 0.60$. The parameters in the selection equation (1) are set at $a = 0.50$ and $b = 1$. The parameters in the response equation (2) are set at $c = -1$, $d = 0.75$, and $e = 0.50$. The response equation includes the endogenous dummy C_i defined by (1). The correlation between the exogenous regressors is 0.40. MLE computed by VGAM 0.7-6.

Endogeneity is moderately strong: $\rho = 0.60$. So there should be some advantage to removing endogeneity bias. The dummy variable C_i is 1 with probability about 0.64, so it has appreciable variance. Furthermore, half the variance on the right hand side of (1) can be explained: $\text{var}(bX_i) = \text{var}(U_i)$. The correlation between the regressors is only 0.40: making that correlation higher exposes the correction to well-known instabilities.

The sample is large: $n = 1000$. Regressors are exogenous by construction. Subjects are independent and identically distributed. Somewhat arbitrarily, we set the true value of c in the response equation (2) to -1, whereas $d = 0.75$ and $e = 0.50$. As it turned out, these choices were favorable too.

Table 1 summarizes results for three kinds of estimates:

1. raw (ignoring endogeneity);
2. the two-step correction; and
3. full maximum likelihood.

For each kind of estimate and each parameter, the table reports the mean of the estimates across the 500 repetitions. Subtracting the true value of the parameter measures the bias in the estimator. Similarly, the standard deviation (SD) across the repetitions, also shown in the table, measures the likely size of the random error.

The "raw estimates" in Table 1 are obtained by fitting the probit model

$$P(Y_i = 1|X_i, Z_i, C_i) = \Phi(c + dZ_i + eC_i)$$

to the data, simply ignoring endogeneity. Bias is quite noticeable.

The two-step estimates are obtained via (3–4), with $\hat{a}$ and $\hat{b}$ obtained by fitting (1). We focus on d and e, as the parameters in equation (2) that may be given causal interpretations. Without correction, $\hat{d}$ averages about 0.72, with correction 0.83 (see Table 1). Correction doubles the bias. Without correction, $\hat{e}$ averages 1.33, with correction 0.54. Correction helps a great deal, but some bias remains.

With the two-step correction, the SD of $\hat{e}$ is about 0.21. Thus, random error in the estimates is appreciable, even with $n = 1000$. On the other hand, the standard error (SE)

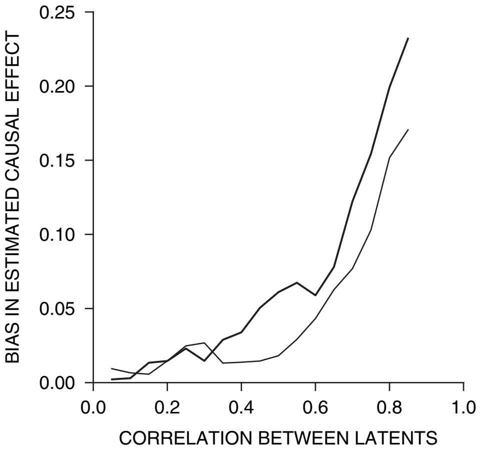

Fig. 1 The two-step correction. Graph of bias in $\hat{e}$ against ρ, the correlation between the latents. The light lower line sets the correlation between regressors to 0.40, and the heavy upper line sets the correlation to 0.60. Other parameters as for Table 1. Below 0.35, the lines crisscross.

across the 500 repetitions is $0.21/\sqrt{500} = 0.01$. The bias in $\hat{e}$ cannot be explained in terms of random error in the simulation: increasing the number of repetitions will not make any appreciable change in the estimated biases.

Heckman (1978) also suggested the possibility of fitting the full model—equations (1) and (2)—by maximum likelihood. The full model is a "bivariate probit" or "biprobit" model. Results are shown in the last two lines of Table 1. The MLE is essentially unbiased. The MLE is better than the two-step correction, although random error remains a concern.

We turn to some variations on the setup described in Table 1. The simulations reported there generated new versions of the regressors on each repetition. Freezing the regressors makes almost no difference in the results: SDs would be smaller in the third decimal place.

The results in Table 1 depend on ρ, the correlation between the latent variables in the selection equation and the response equation. If ρ is increased from 0.60 to 0.80, say, the performance of the two-step correction is substantially degraded. Likewise, increasing the correlation between the exogenous regressors degrades the performance.

When $\rho = 0.80$ and the correlation between the regressors is 0.60, the bias in the two-step correction (3–4) for $\hat{d}$ is about 0.15; for $\hat{e}$, about 0.20. Figure 1 plots the bias in $\hat{e}$ against ρ, with the correlation between regressors set at 0.40 or 0.60, other parameters being fixed at their values for Table 1. The wiggles in the graph reflect variance in the Monte Carlo (there are "only" 500 replicates). The MLE is less sensitive to increasing correlations (data not shown).

Results are also sensitive to the distribution of the exogenous regressors. As the variance in the regressors goes down, bias goes up—in the two-step estimates and in the MLE. Furthermore, numerical issues become acute. There is some explanation: dividing the SD of X by 10, say, is equivalent to dividing b by 10 in equation (1); similarly for Z and d in equation (2). For small values of b and d, parameters are barely identifiable.

Figure 2 plots the bias in $\hat{e}$ against the common SD of X and Z, which is set to values ranging from 0.1 to 1.0 (other parameters are set as in Table 1). The light line represents the MLE. Some of the "bias" in the MLE is indeed small-sample bias—when the SD is 0.1, a sample with $n = 1000$ is a small sample. Some of the bias, however, reflects a tendency of likelihood maximizers to quit before finding the global maximum.

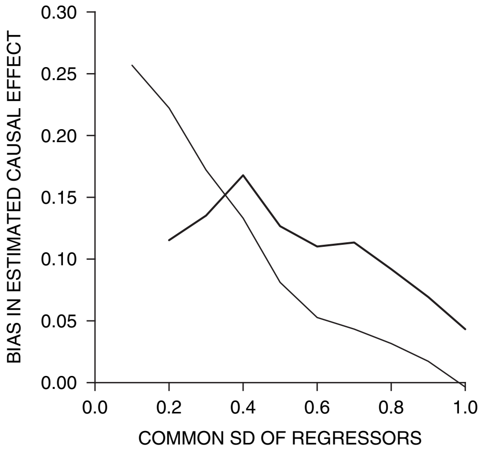

Fig. 2 Graph of bias in $\hat{e}$ against the common SD of the regressors *X* and *Z*. Other parameters as for Table 1. The light line represents the MLE, as computed by VGAM 0.7-6. The heavy line represents the two-step correction.

The heavy line represents the two-step correction. (With an SD of 0.1, data for the two-step correction are not shown, because there are huge outliers; even the median bias is quite changeable from one set of 500 repetitions to another, but 0.2 may be a representative figure.) Curiously, the two-step correction is better than the MLE when the SD of the exogenous regressors is set to 0.2 or to 0.3. This is probably due to numerical issues in maximizing the likelihood functions.

We believe the bias in the two-step correction (Figs 1 and 2) reflects the operating characteristics of the estimator, rather than operating characteristics of the software. Beyond 1.0, the bias in the MLE seems to be negligible. Beyond 1.5, the bias in the two-step estimator for *e* is minimal, but *d* continues to be a little problematic.

As noted above, changing the scale of *X* is equivalent to changing *b*. Similarly, changing the scale of *Z* is equivalent to changing *d* (see equations (1) and (2)). Thus, in Fig. 2, we could leave the SDs at 1 and run through a series of (*b*, *d*) pairs:

$$(0.1 \times b_0, 0.1 \times d_0), (0.2 \times b_0, 0.2 \times d), \ldots,$$

where $b_0 = 1$ and $d_0 = 0.75$ were the initial choices for Table 1.

The number of regressors should also be considered. With a sample size of 1000, practitioners would often use a substantial number of covariates. Increasing the number of regressors is likely to have a negative impact on performance.

3 A Probit Model with Endogenous Sample Selection

Consider next the situation where a probit model is fitted to a sample, but subjects self-select into the sample by an endogenous process. The selection equation is

$$C_i = 1 \text{ if } a + bX_i + U_i > 0, \text{else } C_i = 0. \tag{5}$$

(Selection means, into the sample.) The response equation is

$$Y_i = 1 \text{ if } c + dZ_i + V_i > 0, \text{else } Y_i = 0. \tag{6}$$

Table 2 Simulation results

	c	d	ρ
True values			
	−1.0000	0.7500	0.6000
Raw estimates			
Mean	−0.7936	0.7299	
SD	0.0620	0.0681	
Two-step			
Mean	−1.0751	0.8160	
SD	0.1151	0.0766	
MLE			
Mean	−0.9997	0.7518	0.5946
SD	0.0757	0.0658	0.1590

Notes. Correcting endogeneity bias in sample selection when the response is binary probit. There are 500 repetitions. The sample size is 1000. The correlation between latents is $\rho = 0.60$. The parameters in the selection equation (5) are set at $a = 0.50$ and $b = 1$. The parameters in the response equation (6) are set at $c = -1$, and $d = 0.75$. Response data are observed only when $C_i = 1$, as determined by the selection equation. This will occur for about 64% of the subjects. The correlation between the exogenous regressors is 0.40. MLE computed using *Stata* 9.2.

Equation (6) is the equation of primary interest; however, Y_i and Z_i are observed only when $C_i = 1$. Thus, the data are the observed values of (X_i, C_i) for all i, as well as (Z_i, Y_i) when $C_i = 1$. When $C_i = 0$, however, Z_i and Y_i remain unobserved. Notice that Y_i is binary rather than continuous. Notice too that C_i is omitted from (6); indeed, when (6) can be observed, $C_i \equiv 1$.

Fitting (6) to the observed data raises the question of endogeneity bias. Sample subjects have relatively high values of U_i; hence, high values of V_i. (This assumes $\rho > 0$.) Again, there is a proposed solution that involves two steps.

Step 1. Estimate the probit model (5) by likelihood techniques.

Step 2. Fit the expanded probit model

$$P(Y_i = 1|X_i, Z_i) = \Phi(c + dZ_i + fM_i) \tag{7}$$

to the data on subjects i with $C_i = 1$. This time,

$$M_i = \frac{\phi(a + bX_i)}{\Phi(a + bX_i)}. \tag{8}$$

Parameters in (8) are replaced by the estimates from Step 1. As before, this two-step correction doubles the bias in $\hat{d}$ (see Table 2). The MLE removes most of the bias. However, as for Table 1, the bias in the MLE depends on the SD of the regressors. Bias will be noticeable if the SDs are below 0.2. Some of this is small-sample bias in the MLE, and some reflects difficulties in numerical maximization.

Increasing the sample size from 1000 to 5000 in the simulations barely changes the averages, but reduces the SDs by a factor of about $\sqrt{5}$, as might be expected. This comment applies both to Tables 1 and 2 (data not shown), but not to the MLE results in Table 2. Increasing n would have made the STATA code prohibitively slow to run.

Many applications of Heckman's method feature a continuous response variable rather than a binary variable. Here, the two-step correction is on firmer ground, and parallel simulations (data not shown) indicate that the correction removes most of the endogeneity bias when the parameters are set as in Tables 1 and 2. However, residual bias is large when the SD of the regressors is set to 0.1 and the sample size is "only" 1000; the issues resolve when $n = 10,000$. The problem with $n = 1000$ is created by (1) large random errors in $\hat{b}$, coupled with (2) poorly conditioned design matrices. In more complicated situations, there may be additional problems.

4 Numerical Issues

Exploratory computations were done in several versions of *Matlab*, *R*, and *Stata*. In the end, to avoid confusion and chance capitalization, we redid the computations in a more unified way, with *R* 2.7 for the raw estimates, the two-step correction; VGAM 0.7-6 for the MLE in (1–2); and *Stata* 9.2 for the MLE in (5–6). Why do we focus on the behavior of *R* and *Stata*? *R* is widely used in the statistical community, and *Stata* is almost the lingua franca of quantitative social scientists.

Let b_0 and d_0 be the default values of b and d, namely, 1 and 0.75. As b and d decrease from the defaults, VGAM in *R* handled the maximization less and less well (Fig. 2). We believe VGAM had problems computing the Hessian, even for the base case in Table 1: its internally generated SEs were too small by a factor of about 2, for $\hat{c}, \hat{e}, \hat{\rho}$.

By way of counterpoint, *Stata* did somewhat better when we used it to redo the MLE in (1–2). However, if we multiply the default b_0 and d_0 by 0.3 or 0.4, bias in *Stata* becomes noticeable. If we multiply by 0.1 or 0.2, many runs fail to converge, and the runs that do converge produce aberrant estimates, particularly for a multiplier of 0.1. For multipliers of 0.2 to 0.4, the bias in $\hat{e}$ is upwards in *R* but downwards in *Stata*. In Table 2, *Stata* did well. However, if we scale b_0 and d_0 by 0.1 or 0.2, *Stata* has problems. In defense of *R* and *Stata*, we can say that they produce abundant warning messages when they get into difficulties.

In multidimensional problems, even the best numerical analysis routines find spurious maxima for the likelihood function. Our models present three kinds of problems: (1) flat spots on the log likelihood surface, (2) ill-conditioned maxima, where the eigenvalues of the Hessian are radically different in size, and (3) ill-conditioned saddle points with one small positive eigenvalue and several large negative eigenvalues. The maximizers in VGAM and *Stata* simply give up before finding anything like the maximum of the likelihood surface. This is a major source of the biases reported above.

The model defined by (1–2) is a harder challenge for maximum likelihood than (5–6), due to the extra parameter e. Our computations suggest that most of the difficulty lies in the joint estimation of three parameters, c, e, ρ. Indeed, we can fix a, b, d at the default values for Table 1 and maximize the likelihood over the remaining three parameters c, e, ρ. VGAM and *Stata* still have convergence issues. The problems are the same as with six parameters. For example, we found a troublesome sample where the Hessian of the log likelihood had eigenvalues 4.7, -1253.6, -2636.9. (We parameterize the correlation between the latents by $\log(1 + \rho) - \log(1 - \rho)$ rather than ρ, since that is how binom2.rho in VGAM does things.)

One of us has an improved likelihood maximizer called GENOUD (Sekhon and Mebane 1998). GENOUD seems to do much better at the maximization, and its internally generated SEs are reasonably good. Results for GENOUD and *Stata* not reported here and are available from the authors.

5 Implications for Practice

There are two main conclusions from the simulations and the analytic results.

1. Under ordinary circumstances, the two-step correction should not be used in probit response models. In some cases, the correction will reduce bias, but in many other cases, the correction will increase bias.
2. If the bivariate probit model is used, special care should be taken with the numerics. Conventional likelihood maximization algorithms produce estimates that are far away from the MLE. Even if the MLE has good operating characteristics, the "MLE" found by the software package may not. Results from VGAM 0.7-6 should be treated with caution. Results from *Stata* 9.2 may be questionable for various combinations of parameters.

The models analyzed here are very simple, with one covariate in each of (1–2) and (5–6). In real examples, the number of covariates may be quite large, and numerical behavior will be correspondingly more problematic.

Of course, there is a question more salient than the numerics: what is it that justifies probit models and the like as descriptions of behavior? For additional discussion, see Freedman (2005), which has further cites to the literature on this point.

6 Motivating the Estimator

Consider (1–2). We can represent V_i as $\rho U_i + \sqrt{1-\rho^2}W_i$, where W_i is an $N(0, 1)$ random variable, independent of U_i. Then

$$\begin{aligned} E\left\{V_i \middle| X_i = x, C_i = 1\right\} &= E\left\{\rho U_i + \sqrt{1-\rho^2}W_i \middle| U_i > -a-bx_i\right\} \\ &= \rho E\{U_i | U_i > -a-bx_i\} \\ &= \rho \frac{1}{\Phi(a+bx_i)} \int_{-a-bx_i}^{\infty} x\phi(x)\mathrm{d}x \\ &= \rho \frac{\phi(a+bx_i)}{\Phi(a+bx_i)} \end{aligned} \tag{9}$$

because $P\{U_i > -a-bx_i\} = P\{U_i < a+bx_i\} = \Phi(a+bx_i)$. Likewise,

$$E\left\{V_i \middle| X_i = x, C_i = 0\right\} = -\rho \frac{\phi(a+bx_i)}{1-\Phi(a+bx_i)}. \tag{10}$$

In (2), therefore, $E\{V_i - \rho M_i \mid X_i, C_i\} = 0$. If (2) were a linear regression equation, then OLS estimates would be unbiased, the coefficient of M_i being nearly ρ. (These remarks take a and b as known, with the variance of the error term in the linear regression normalized to 1.) However, (2) is not a linear regression equation: (2) is a probit model. That is the source of the problem.

7 Identifiability

Identifiability means that parameters are determined by the joint distribution of the observables; parameters that are not identifiable cannot be estimated. In the model defined by (1–2), the parameters are a, b, c, d, e and the correlation ρ between the latents; the observables are X_i, Z_i, C_i, and Y_i. In the model defined by (5–6), the parameters are a, b, c, d and the correlation ρ between the latents; observables are $X_i, C_i, \tilde{Z}_i, \tilde{Y}_i$, where $\tilde{Z}_i = Z_i$ and $\tilde{Y}_i = Y_i$

when $C_i = 1$, whereas $\tilde{Z}_i = \tilde{Y}_i = M$ when $C_i = 0$. Here, M is just a special symbol that denotes "missing."

Results are summarized as Propositions 1 and 2. The statements involve the sign of d, which is $+1$ if $d > 0$, 0 if $d = 0$, and -1 if $d < 0$. Since subjects are independent and identically distributed, only $i = 1$ need be considered. The variables (X_1, Z_1) are taken as bivariate normal, with a correlation strictly between -1 and $+1$. This assumption is discussed below.

Proposition 1. Consider the model defined by (1–2). The parameters a and b in (1) are identifiable, and the sign of d in (2) is identifiable. If $b \neq 0$, the parameters c, d, e, ρ in (2) are identifiable. If $b = 0$ but $d \neq 0$, the parameters c, d, e, ρ are still identifiable. However, if $b = d = 0$, the remaining parameters c, e, ρ are not identifiable.

Proposition 2. Consider the model defined by (5–6). The parameters a and b in (5) are identifiable, and the sign of d in (6) is identifiable. If $b \neq 0$, the parameters c, d, ρ in (6) are identifiable. If $b = 0$ but $d \neq 0$, the parameters c, d, ρ are still identifiable. However, if $b = d = 0$, the remaining parameters c, ρ are not identifiable.

Proof of Proposition 1. Clearly, the joint distribution of C_1 and X_1 determines a and b, so we may consider these as given. The distributions of X_1 and Z_1 are determined (this is not so helpful). We can take the conditional distribution of Y_1 given $X_1 = x$ and $Z_1 = z$ as known. In other words, suppose (U, V) are bivariate normal with mean 0, variance 1, and correlation ρ.

The joint distribution of the observables determines a, b, and two functions ψ_0, ψ_1 of x, z:

$$\psi_0(x, z) = P(a + bx + U < 0 \text{ and } c + dz + V > 0),$$

$$\psi_1(x, z) = P(a + bx + U > 0 \text{ and } c + dz + e + V > 0). \tag{11}$$

There is no additional information about the parameters.

Fix x at any convenient value, and consider $z > 0$. Then $z \to \psi_0(x, z)$ is strictly decreasing, constant, or strictly increasing, according as $d < 0$, $d = 0$, or $d > 0$. The sign of d is therefore determined. The rest of proof, alas, consists of a series of cases.

The case $b \neq 0$ and $d > 0$: Let $u = -a - bx$, $v = -z$, $\xi = U$, and $\zeta = (V + c)/d$. Then (ξ, ζ) are bivariate normal, with unknown correlation ρ. We know ξ has mean 0 and variance 1. The mean and variance of ζ are unknown, being c/d and $1/d^2$, respectively. But

$$P(\xi < u \text{ and } \zeta > v) \tag{12}$$

is known for all (u, v). Does this determine ρ, c, d? Plainly so, because (12) determines the joint distribution of ξ, ζ. We can then compute ρ, $d = 1\big/\sqrt{\text{var}(\zeta)}$, and $c = dE(\zeta)$. Finally, ψ_1 in (11) determines e. This completes the argument for the case $b \neq 0$ and $d > 0$.

The case $b \neq 0$ and $d < 0$ is the same, except that $d = -1\big/\sqrt{\text{var}(\zeta)}$.

The case $b \neq 0$ and $d = 0$: Here, we know

$$P(U < u \text{ and } c + V > 0) \text{ for all } u. \tag{13}$$

Let $u \to \infty$: the marginal distribution of V determines c. Furthermore, from (13), we can compute $P(V > -c \mid U = u)$ for all u. Given $U = u$, we know that V is distributed as $\rho u + \sqrt{1-\rho^2}W$, where W is $N(0, 1)$. If $\rho = \pm 1$, then

$$P(V > -c | U = u) = 1 \text{ if } \rho u > -c$$

$$= 0 \text{ if } \rho u < -c$$

If $-1 < \rho < 1$, then

$$P\{V > -c | U = u\} = P\left\{W > -\frac{c + \rho u}{\sqrt{1-\rho^2}}\right\} = \Phi\left(\frac{c + \rho u}{\sqrt{1-\rho^2}}\right). \tag{14}$$

So we can determine whether $\rho = \pm 1$, and if so, which sign is right. Suppose $-1 < \rho < 1$. Then (14) determines $\left(c + \rho u\right) \big/ \sqrt{1-\rho^2}$. Differentiate with respect u to see that (14) determines $\rho \big/ \sqrt{1-\rho^2}$. This is a $1-1$ function of ρ. Thus, ρ can be determined, and then c; finally, e is obtained from ψ_1 in (11). This completes the argument for the case $b \neq 0$ and $d = 0$.

The case $b = 0$ and $d > 0$*:* As above, let W be independent of U and $N(0, 1)$; represent V as $\rho U + \sqrt{1-\rho^2}W$. Let $G = \{U < -a\}$. From ψ_0 and a, we compute

$$\begin{aligned} P\left\{V > -c - dz \middle| G\right\} &= P\left\{\rho U + \sqrt{1-\rho^2}W > -c - dz \middle| G\right\} \\ &= P\left\{\frac{\rho}{d}U + \frac{\sqrt{1-\rho^2}}{d}W + \frac{c}{d} > -z | G\right\}. \end{aligned} \tag{15}$$

Write U_a for U conditioned so that $U < -a$. The right hand side of (15), as a function of z, determines the distribution function of the sum of three terms: two independent random variables, U_a and $\sqrt{1-\rho^2}W \big/ d$, where W is standard normal, plus the constant c/d. This distribution is therefore known, although it depends on the three unknowns, c, d, ρ.

Write Λ for the log Laplace transform of U_a. This is a known function. Now compute the log Laplace transform of the distribution in (15). This is

$$t \to \Lambda\left(\frac{\rho}{d}t\right) + \frac{1-\rho^2}{d^2}t^2 + \frac{c}{d}t. \tag{16}$$

Again, this function is known, although c, d, and ρ are unknown. Consider the expansion of (16) as a power series near 0, of the form $\kappa_1 t + \kappa_2 t^2/2! + \kappa_3 t^3/3! + \cdots$. The κ's are the "cumulants" or "semi-invariants" of the distribution in (15). These are known quantities because the function in (16) is known: κ_1 is the mean of the distribution given by (15), whereas κ_2 is the variance, and κ_3 is the central third moment.

Of course, $\Lambda'(0) = E(U_a) = -\varphi(-a)/\Phi(-a)$. Thus, $\kappa_1 = -\varphi(-a)/\Phi(-a) + c/d$, which determines c/d. Next, $\Lambda''(0) = \text{var}(U_a)$, so $\kappa_2 = (\rho/d)^2\text{var}(U_a) + (1 - \rho^2)/d^2$ is determined. Finally, $\kappa_3 = \Lambda'''(0)$ is the third central moment of U_a. Since U_a has a skewed distribution, $\Lambda'''(0) \neq 0$. We can compute $(\rho/d)^3$ from κ_3, and then ρ/d. Next, we get $1/d^2$ from κ_2, and then $1/d$. (We are looking at the case $d > 0$.) Finally, c comes from κ_1. Thus, c, d, ρ are determined, and e comes from ψ_1 in (11). This completes the argument for the case $b = 0$ and $d > 0$.

The case $b = 0$ and $d < 0$ follows by the same argument.

The case $b = d = 0$: The three remaining parameters, c, e, and ρ, are not identifiable.

For simplicity, take $a = 0$, although this is not essential. Suppose

$$P(U<0 \text{ and } V>-c)=\alpha \tag{17}$$

is given, with $0 < \alpha < 1/2$. Likewise,

$$P(U>0 \text{ and } V>-c-e)=\beta \tag{18}$$

is given, with $0 < \beta < 1/2$. The joint distribution of the observables contains no further information about the remaining parameters c, e, ρ. Choose any particular ρ with $-1 \leqslant \rho \leqslant 1$. Choose c so that (17) holds and e so that (18) holds. The upshot: there are infinitely many c, e, ρ triplets yielding the same joint distribution for the observables. This completes the argument for the case $b = d = 0$, and so for Proposition 1.

Proof of Proposition 2. Here, we know the joint distribution of (X_1, C_1), which determines a, b. We also know the joint distribution of (X_1, Z_1, Y_1) given $C_1 = 1$; we do not know this joint distribution given $C_1 = 0$. As in (11), suppose (U, V) are bivariate normal with mean 0, variance 1 and correlation ρ. The joint distributions of the observables determine a, b and the function

$$\psi_1(x,z)=P(a+bx+U>0 \text{ and } c+dz+V>0). \tag{19}$$

There is no other information in the system; in particular, we do not know the analog of ψ_0. Most of the argument is the same as before, or even a little easier. We consider in detail only one case.

The case $b = d = 0$: The two remaining parameters, c, ρ are not identifiable. Again, take $a = 0$. Fix any α with $0 < \alpha < 1/2$. Suppose

$$P(U>0 \text{ and } V>-c)=\alpha \tag{20}$$

is given. There is no other information to be had about c, ρ. Fix any ρ with $-1 \leqslant \rho \leqslant 1$ and solve (20) for c. There are infinitely many c, ρ pairs giving the same joint distribution for the observables when $b = d = 0$. This completes our discussion of Proposition 2.

Remarks

1. The random variable U_a was defined in the course of proving Proposition 1. If desired, the moments of U_a can be obtained explicitly in terms of φ and Φ, using repeated integration by parts.
2. The Laplace transform of U_a is easily obtained by completing the square, and

$$t\to\frac{1}{\sqrt{2\pi}}\exp\left(\frac{1}{2}t^2\right)\frac{\Phi(-a-t)}{\Phi(-a)}. \tag{21}$$

The third derivative of the log-Laplace transform can be computed from (21), but it is painful.

3. The argument for the case $b = 0$ and $d > 0$ in Proposition 1 is somewhat intricate, but it actually covers all values of b, whether zero or non-zero. The argument shows that for any particular real α, the values of c, d, ρ are determined by the number $P(\alpha + U < 0)$ and the function

$$z \rightarrow P(\alpha + U < 0 \text{ and } c + dz + V > 0).$$

4. Likewise, the argument for the case $b \neq 0$ and $d = 0$ proves more. If we know $P(U < u)$ and $P(U < u \quad \text{and} \quad \gamma + V > 0)$ for all real u, that determines γ and ρ.
5. In (17), for example, if $\alpha = 1/2$, then $\rho = -1$; but c can be anywhere in the range $[0, \infty)$.
6. The propositions can easily be extended to cover vector-valued exogenous variables.
7. Our proof of the propositions really does depend on the assumption of an imperfect correlation between X_i and Z_i. We hope to consider elsewhere the case where $Z_i \equiv X_i$. The assumption of normality is not material; it is enough if the joint distributions have full support, although positive densities are probably easier to think about.
8. The assumption of bivariate normality for the latent variables is critical. If this is wrong, estimates are likely to be inconsistent.
9. Suppose (U, V) are bivariate normal with correlation ρ, and $-1 < \rho < 1$. Then

$$\rho \rightarrow P(U > 0 \text{ and } V > 0)$$

is strictly monotone. This is Slepian's theorem: see Tong (1980). If the means are 0 and the variances are 1, numerical calculations suggest this function is convex on $(-1, 0)$ and concave on $(0, 1)$.

8 Some Relevant Literature

Cumulants are discussed by Rao (1973, 101). The ratio φ/Φ in (8) is usually called the "inverse Mills ratio," in reference to Mills (1926)—although Mills tabulates $[1 - \Phi(x)]/\varphi(x)$ for $x \geqslant 0$. Heckman (1978, 1979) proposes the use of M_i to correct for endogeneity and selection bias in the linear case, with a very clear explanation of the issues. He also describes potential use of the MLE. Meng and Schmidt (1985) discuss cases where the bivariate probit MLE is fragile. Rivers and Vuong (1988) propose an interesting alternative to the Heckman estimator. Their estimator (perhaps confusingly) is also called a two-step procedure. It seems most relevant when the endogenous variable is continuous; ours is binary.

For other estimation strategies and discussion, see Angrist (2001). Bhattacharya, Goldman, and McCaffrey (2006) discuss several "two-step" algorithms, including a popular Instrumental-Variables Least Squares regression estimator that turns out to be inconsistent; they do not seem to consider the particular two-step estimator of concern in our paper. Muthen (1979) discusses identifiability in a model with latent causal variables. The VGAM manual (Yee 2007) notes difficulties in computing standard errors. According to Stata (2005), its maximum likelihood routine "provides consistent, asymptotically efficient estimates for all the parameters in [the] models."

Van de Ven and Van Praag (1981) found little difference between the MLE and the two-step correction; the difference doubtless depends on the model under consideration. Instabilities in the two-step correction are described by Winship and Mare (1992), Copas and Li (1997), and Briggs (2004), among others. For additional citations, see Dunning and Freedman (2007). Ono (2007) uses the two-step correction with probit response in a study of the Japanese labor market; X and Z are multidimensional. The sample size is 10,000, but only 300 subjects select into the treatment condition. Bushway, Johnson, and Slocum (2007) describe many overenthusiastic applications of the two-step correction in the criminology literature: binary response variables are among the least of the sins.

We do not suggest that finding the true maximum of the likelihood function guarantees the goodness of the estimator, because there are situations where the MLE performs rather badly. Freedman (2007) has a brief review of the literature on this topic. However, we would suggest that spurious maxima are apt to perform even less well, particularly with the sort of models considered here.

References

Angrist, J. D. 2001. Estimation of limited-dependent variable models with binary endogenous regressors: simple strategies for empirical practice. *Journal of Business and Economic Statistics* 19:2–28(with discussion).

Bhattacharya, J., D. Goldman, and D. McCaffrey. 2006. Estimating probit models with self-selected treatments. *Statistics in Medicine* 25:389–413.

Briggs, D. C. 2004. Causal inference and the Heckman model. *Journal of Educational and Behavioral Statistics* 29:397–420.

Bushway, S., B. D. Johnson, and L. A. Slocum. 2007. Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. *Journal of Quantitative Criminology* 23:151–78.

Copas, J. B., and H. G. Li. 1997. Inference for non-random samples. *Journal of the Royal Statistical Society, Series B* 59:55–77.

Dunning, T., and D. A. Freedman. 2007. Modeling selection effects. In *The handbook of social science methodology*, eds. Steven Turner and William Outhwaite, 225–31. London: Sage.

Freedman, D. A. 2005. *Statistical models: theory and practice*. New York: Cambridge University Press.

———. 2007. How can the score test be inconsistent? *The American Statistician* 61:291–95.

Heckman, J. J. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46:931–59.

———. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–61.

Meng, C., and P. Schmidt. 1985. On the cost of partial observability in the bivariate probit model. *International Economic Review* 26:71–85.

Mills, J. P. 1926. Table of the ratio: area to boundary ordinate, for any portion of the normal curve. *Biometrika* 18:395–400.

Muthen, B. 1979. A structural probit model with latent variables. *Journal of the American Statistical Association* 74:807–11.

Ono, H. 2007. Careers in foreign-owned firms in Japan. *American Sociological Review* 72:267–90.

Rao, C. R. 1973. *Linear statistical inference*. 2nd ed. New York: Wiley.

Rivers, D., and Q. H. Vuong. 1988. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics* 39:347–66.

Sekhon, J. S., and W. R. Mebane Jr. 1998. Genetic optimization using derivatives: theory and application to nonlinear models. *Political Analysis* 7:189–213.

Stata. 2005. *Stata base reference manual*. Stata Statistical Software. Release 9. Vol. 1. College Station, TX: StataCorp LP.

Tong, Y. L. 1980. *Probability inequalities in multivariate distributions*. New York: Academic Press.

Van de Ven, W. P. M. M., and B. M. S. Van Praag. 1981. The demand for deductibles in private health insurance: a probit model with sample selection. *Journal of Econometrics* 17:229–52.

Winship, C., and R. D. Mare. 1992. Models for sample selection bias. *Annual Review of Sociology* 18:327–50.

Yee, T. W. 2007. The VGAM package. http://www.stat.auckland.ac.nz/ỹee/VGAM