A Full Information Maximum Likelihood Approach to Estimating the Sample Selection Model with Endogenous Covariates[1]

Jörg Schwiebert

Gottfried Wilhelm Leibniz University Hannover

Department of Economics, Institute for Labor Economics

Königsworther Platz 1, 30167 Hannover, Germany

E-mail: schwiebert@aoek.uni-hannover.de

Keywords: Sample Selection Model; Endogeneity; Maximum Likelihood Estimation; Female Labor Supply

JEL classification: C31, C34, C36

Summary

In this paper we establish a full information maximum likelihood approach to estimating the sample selection model with endogenous covariates. We also provide a test for exogeneity which indicates whether endogeneity is in fact a matter or not. In contrast to other methods proposed in the literature which deal with sample selection and endogeneity, our approach is computationally simple and provides exact asymptotic standard errors derived from common maximum likelihood theory. A Monte Carlo study and an empirical example are presented which indicate that not accounting for endogeneity in sample selection models may lead to severely biased parameter estimates.

[1]I would like to thank Olaf Hübler, Patrick Puhani and my colleagues at the Institute for Labor Economics for helpful discussion and comments.

1. Introduction

The purpose of this paper is to establish a full information maximum likelihood (FIML) approach to estimating the sample selection model with endogenous covariates. Additionally, a test for exogeneity is provided which indicates whether endogeneity is in fact a matter or not.

Pioneered by Heckman (1979), the sample selection model, also known as Heckman model or Type II Tobit model (Amemiya, 1985), has been used as a state-of-the-art model for correcting ordinary least squares estimates for a potential selection bias. A leading example is given by wage regressions for women, where only a non-random part of the entire population of women is working and, thus, included in the sample. As it is well known, not accounting for the non-randomness of the sample induces biased parameter estimates. The most commonly employed methods to estimate these models are Heckman's two-step approach and maximum likelihood.[2] Both approaches involve the primary regression equation (the equation of interest) and a selection equation of the Probit type which controls for the sample selection mechanism.

However, in most studies using the sample selection model covariates are treated as exogenous. In the cross section case, few attempts have been made to account for possibly endogenous covariates. Exceptions are Wooldridge (2010, ch. 19) and Chib et al. (2009). Wooldridge (2010) essentially proposed a two-stage least squares approach, where fitted values from a first stage regression of the endogenous covariate(s) on instrumental variables are inserted into the primary regression equation (which includes the inverse Mill's ratio term). Semykina and Wooldridge (2010) used the same methodology when considering panel data models incorporating the simultaneous presence of endogeneity and sample selection. Further estimators for panel sample selection models with endogeneity have been proposed by Vella and Verbeek (1999) and Das, Newey and Vella (2003). While Vella and Verbeek (1999) considered conditional moment and conditional

[2] See Vella (1998) for an account of various methods to estimate models with sample selection bias. Puhani (2000) discusses the usefulness of the two-step approach.

maximum likelihood estimation, Das, Newey and Vella (2003) suggested nonparametric estimators. Back in the cross section setting, Chib et al. (2009) employed a full information maximum likelihood framework in a Bayesian setup, where estimation involves use of the Gibbs sampler.

In this paper, we will focus on the cross section case. The main advantage of Wooldridge's (2010) two stage least squares estimator is given by its computational simplicity. A drawback, however, is that due to the inclusion of the (estimated) inverse Mills ratio term the standard errors have to be adjusted, e.g. by applying bootstrapping techniques. As usual in Bayesian estimation, the estimator proposed by Chib et al. (2009) requires the incorporation of prior information and may, thus, be more appropriate in finite samples. Yet the disadvantage of this class of estimators is that they are computationally very demanding.

A common drawback of the approaches by Wooldridge (2010) and Chib et al. (2009) is that both fail to account for endogeneity in the selection equation as well. This may be a serious problem since in many applications of the sample selection model, most explanatory variables are included into the primary equation *and* into the selection equation. The importance of accounting for endogeneity in both equations will be further investigated in this paper by a series of Monte Carlo simulations.

Our proposed estimation framework generalizes the cross section approaches of Wooldridge (2010) and Chib et al. (2009) by accounting for endogeneity not only in the primary equation but in the selection equation as well. Furthermore, since we employ a full information maximum likelihood framework, our estimator is asymptotically efficient (provided that the distributional assumptions are correct) and we do not have to adjust standard errors (e.g., by bootstrapping techniques). This distinguishes our approach from Wooldridge (2010). Finally, our approach is less computationally demanding than that of Chib et al. (2009).

The estimation framework established in this paper is in the spirit of the estimators for the Tobit model with endogenous covariates as provided by Smith and Blundell

(1986) and the Probit model with endogenous covariates as provided by Rivers and Vuong (1988); see also Newey (1987). The proposed test for exogeneity is in line with Smith and Blundell (1986) and Rivers and Vuong (1988).

Besides developing the estimator and proposing some tests for exogeneity and for the absence of sample selection bias, we also provide Monte Carlo evidence on the consequences of (falsely) assuming exogeneity of covariates when, in fact, these are endogenous. The Monte Carlo simulations indicate that the bias may be substantial.

We further provide an empirical application to the estimation of a wage equation for married women. This is the classical example for sample selection bias, and has also been investigated by Chib et al. (2009) and Wooldridge (2010). In this example, we conjecture that the variable education may be endogenous, since it may be affected by unobserved variables such as ability, which itself affects the wage and the probability of labor market participation.

The paper is structured as follows. In section 2, the sample selection model with endogenous covariates is developed. Section 3 establishes the full information maximum likelihood estimation framework. In section 4, tests for exogeneity and for the absence of sample selection bias are proposed. Section 5 contains the results of the Monte Carlo simulations designed to indicate the bias of not accounting for endogeneity. Section 6 contains the empirical application to the estimation of a wage equation for married women. Finally, section 7 gives conclusions.

2. The Sample Selection Model with Endogenous Covariates

The model is given by

$$
\begin{aligned}
y_i^* &= X_{1i}\beta_1 + X_{2i}\beta_2 + C_i\beta_3 + u_i \equiv X_i\beta + u_i && (2.1)\\
z_i^* &= W_{1i}\gamma_1 + W_{2i}\gamma_2 + C_i\gamma_3 + v_i \equiv W_i\gamma + v_i && (2.2)\\
X_{2i} &= [X_{1i}, W_{1i}]\Delta_1 + Z_{1i}\Delta_2 + \varepsilon_{1i} \equiv \tilde{Z}_{1i}\Delta + \varepsilon_{1i} && (2.3)\\
W_{2i} &= [X_{1i}, W_{1i}]\Lambda_1 + Z_{2i}\Lambda_2 + \varepsilon_{2i} \equiv \tilde{Z}_{2i}\Lambda + \varepsilon_{2i} && (2.4)\\
C_i &= [X_{1i}, W_{1i}]\Upsilon_1 + Z_{3i}\Upsilon_2 + \varepsilon_{3i} \equiv \tilde{Z}_{3i}\Upsilon + \varepsilon_{3i} && (2.5)
\end{aligned}
$$

$$
\begin{aligned}
z_i &= \mathbf{1}(z_i^* > 0) && (2.6)\\
y_i &= y_i^*\mathbf{1}(z_i = 1) && (2.7)\\
i &= 1, \ldots, n.
\end{aligned}
$$

The first equation is the primary equation (equation of interest), where the latent dependent variable y_i^* is related to a $(1 \times K_1)$-vector of exogenous explanatory variables, X_{1i}, to a $(1 \times K_2)$-vector of endogenous explanatory variables only included in the primary equation but not in the selection equation, X_{2i}, and to a $(1 \times P)$-vector of endogenous explanatory variables included in the primary and the selection equation, C_i. The second equation is the selection equation, where the latent variable z_i^* is related to a $(1 \times L_1)$-vector of exogenous explanatory variables, W_{1i}, to a $(1 \times L_2)$-vector of endogenous explanatory variables, W_{2i} only included in the selection equation but not in the primary equation, and to C_i. In equations (2.3) to (2.5) it is assumed that the endogenous explanatory variables can be explained by a $(1 \times M_1)$-vector, a $(1 \times M_2)$-vector and a $(1 \times M_3)$-vector of instrumental variables, Z_{1i}, Z_{2i} and Z_{3i}, respectively. Equation (2.6) expresses that only the sign of z_i^* is observable. Finally, equation (2.7) comprises the selection mechanism, i.e. the latent variable y_i^* is only observed if the selection indicator z_i is equal to one. Equations

(2.1), (2.2), (2.6), and (2.7) build up the framework of the sample selection model without endogeneity as presented in many textbooks (e.g., Davidson and MacKinnon, 1993, pp. 542-543). The additional feature in equations (2.3) to (2.5) is that some of the covariates (X_2, W_2 and C) in the primary and the selection equation are endogenous, i.e. correlated with the error terms u and v. We assume that for each of these endogenous variables there exist instrumental variables Z_1, Z_2 and Z_3 which are not correlated with any error term in the model. For proper identification, the selection equation is supposed to contain at least one explanatory variable which is not included in the primary equation.

To complete the model, it is assumed that the vector of error terms $(u_i, v_i, \varepsilon'_{1i}, \varepsilon'_{2i}, \varepsilon'_{3i})'$ is distributed according to

$$\begin{pmatrix} u_i \\ v_i \\ \varepsilon'_{1i} \\ \varepsilon'_{2i} \\ \varepsilon'_{3i} \end{pmatrix} \sim \text{NID} \left(0, \begin{bmatrix} \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix} & \Omega' \\ \Omega_{(J\times 2)} & \Sigma_{(J\times J)} \end{bmatrix} \right), \tag{2.8}$$

where NID denotes "normally and independently distributed", $J \equiv K_2 + L_2 + P$, and the distribution should be interpreted as conditional on all exogenous variables (the conditioning has been omitted for the ease of notation). The covariance matrix of the error terms consists of four parts. The upper left part is the covariance matrix attributed to the error terms of the primary and selection equation, respectively, where σ_u^2 and σ_v^2 denote the variances of u and v, and ρ denotes the correlation coefficient. If there was no concern about endogeneity, inference would be based solely on this part of the covariance matrix, as it is common in the standard sample selection model. However, the (potential) presence of endogeneity is indicated by the $(J \times 2)$-matrix Ω, which captures the influence of unobserved factors which jointly affect the dependent variables in equation (2.1) and (2.2) and the endogenous explanatory variables. Note that endogeneity is absent if and only if Ω is equal to the null matrix. Finally, the error terms attributed to the endogenous

explanatory variables have covariance matrix Σ whose dimension is $(J \times J)$.

Note that it is assumed that the distribution of the endogenous covariates can be reasonably approximated by a normal distribution, which favors continuous regressors and excludes binary regressors.

3. Full Information Maximum Likelihood Estimation

First, note that the conditional distribution of $(u_i, v_i)'$ given $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})$ is given by

$$\left[\begin{matrix} u_i \\ v_i \end{matrix}\right] \Bigg| \varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i} \sim \text{NID}\left(\Omega'\Sigma^{-1}\left[\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i}\right]', B\right) \tag{3.1}$$

where

$$B \equiv \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix} - \Omega'\Sigma^{-1}\Omega. \tag{3.2}$$

Define

$$\Psi \equiv \underset{(2\times J)}{\begin{pmatrix} \underset{(1\times K_2)}{\psi_{11}} & \underset{(1\times L_2)}{\psi_{12}} & \underset{(1\times P)}{\psi_{13}} \\ \underset{(1\times K_2)}{\psi_{21}} & \underset{(1\times L_2)}{\psi_{22}} & \underset{(1\times P)}{\psi_{23}} \end{pmatrix}} \equiv \Omega'\Sigma^{-1} \tag{3.3}$$

$$\Gamma \equiv \begin{pmatrix} \tilde{\sigma}^2 & \tilde{\rho}\tilde{\sigma} \\ \tilde{\rho}\tilde{\sigma} & 1 \end{pmatrix} \equiv \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix} - \Omega'\Sigma^{-1}\Omega, \tag{3.4}$$

where the lower right element of Γ has been set equal to unity due to normalization. Therefore, equation (3.1) can be recast as

$$\left[\begin{matrix} u_i \\ v_i \end{matrix}\right] \Bigg| \varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i} \sim \text{NID}\left(\begin{bmatrix} \psi_{11}\varepsilon_{1i}' + \psi_{12}\varepsilon_{2i}' + \psi_{13}\varepsilon_{3i}' \\ \psi_{21}\varepsilon_{1i}' + \psi_{22}\varepsilon_{2i}' + \psi_{23}\varepsilon_{3i}' \end{bmatrix}, \begin{pmatrix} \tilde{\sigma}^2 & \tilde{\rho}\tilde{\sigma} \\ \tilde{\rho}\tilde{\sigma} & 1 \end{pmatrix}\right), \tag{3.5}$$

which resembles the (unconditional) joint error distribution of the sample selection model without endogeneity (except for the non-zero means).[3]

Then, the likelihood function can be written as the product of a conditional distribution which resembles the (unconditional) likelihood function of the sample selection model without endogeneity and the joint distribution of the error terms $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$. Thus, the log-likelihood function is given by

$$\begin{aligned} l(\theta) = & \sum_{z_i=0} \log\{\Phi(-W_i\gamma - \psi_{21}\varepsilon'_{1i} - \psi_{22}\varepsilon'_{2i} - \psi_{23}\varepsilon'_{3i})\} \\ & + \sum_{z_i=1} \log\{\tilde{\sigma}^{-1}\phi(\tilde{\sigma}^{-1}(y_i - X_i\beta - \psi_{11}\varepsilon'_{1i} - \psi_{12}\varepsilon'_{2i} - \psi_{13}\varepsilon'_{3i}))\} \\ & + \sum_{z_i=1} \log\{\Phi((1-\tilde{\rho}^2)^{-1/2}[W_i\gamma + \psi_{21}\varepsilon'_{1i} + \psi_{22}\varepsilon'_{2i} + \psi_{23}\varepsilon'_{3i} \\ & \qquad + \tilde{\rho}\tilde{\sigma}^{-1}(y_i - X_i\beta - \psi_{11}\varepsilon'_{1i} - \psi_{12}\varepsilon'_{2i} - \psi_{13}\varepsilon'_{3i})])\} \\ & - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n} \begin{bmatrix} \varepsilon_{1i} & \varepsilon_{2i} & \varepsilon_{3i} \end{bmatrix} \Sigma^{-1} \begin{bmatrix} \varepsilon_{1i} & \varepsilon_{2i} & \varepsilon_{3i} \end{bmatrix}', \end{aligned} \tag{3.6}$$

where $\theta \equiv (\beta', \gamma', \tilde{\rho}, \tilde{\sigma}, \text{vec}(\Psi)', \text{vech}(\Sigma)', \text{vec}(\Delta)', \text{vec}(\Lambda)', \text{vec}(\Upsilon)')'$,

$$\varepsilon_{1i} = X_{2i} - \tilde{Z}_{1i}\Delta \tag{3.7}$$

$$\varepsilon_{2i} = W_{2i} - \tilde{Z}_{2i}\Lambda \tag{3.8}$$

$$\varepsilon_{3i} = C_i - \tilde{Z}_{3i}\Upsilon, \tag{3.9}$$

$\Phi(\cdot)$ denotes the standard normal cumulative distribution function and $\phi(\cdot)$ the standard normal probability density function.

The FIML estimator of the sample selection model with endogenous covariates is thus given by

$$\hat{\theta} = \arg\max_{\theta} \; l(\theta). \tag{3.10}$$

[3]The approach undertaken here to accommodate the endogeneity problem has been called "control function approach" in the literature (see, e.g., Wooldridge, 2010, pp. 126-29).

4. Testing for Exogeneity and the Absence of Sample Selection Bias

Both the presence of exogeneity as well as the absence of sample selection bias can be tested relatively straightforwardly using Wald tests. For instance, if the null hypothesis claims that there is no endogeneity at all (the Ψ-matrix is the null matrix), the test statistic will be given by

$$W_\Psi = \text{vec}(\hat{\Psi})'(\text{Asy.Cov}[\text{vec}(\hat{\Psi})])^{-1}\text{vec}(\hat{\Psi}) \sim \chi^2(2J), \tag{4.1}$$

where Asy.Cov$[\text{vec}(\hat{\Psi})]$ denotes the asymptotic covariance matrix of $\text{vec}(\hat{\Psi})$. Under suitable regularity conditions (for instance, cf. Amemiya, 1985, pp. 120-127), this asymptotic covariance can be obtained by using the fact that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, -\mathscr{H}^{-1}), \tag{4.2}$$

where $\mathscr{H} = n^{-1}\text{E}\left(\frac{\partial^2 l(\theta_0)}{\partial\theta\partial\theta'}\right)$ and θ_0 is the true value of the parameter vector.

In a similar fashion it is possible to test significance of single elements of the Ψ-matrix as well as joint significance of some elements.

The Wald statistic for testing for the absence of sample selection bias ($\tilde{\rho} = 0$) is given by

$$W_{\tilde{\rho}} = \frac{\hat{\tilde{\rho}}^2}{\text{Asy.Var}(\hat{\tilde{\rho}})} \sim \chi^2(1), \tag{4.3}$$

which can be replaced by a simple significance test for $\tilde{\rho}$. The reason for employing $\tilde{\rho}$ instead of ρ to test for the absence of sample selection bias is motivated by the consideration that $\tilde{\rho}$ measures sample selection *after* controlling for endogeneity. Thus, $\tilde{\rho}$ only contains the part of the correlation between *u* and *v* which is *not* due to the endogeneity of some covariates.[4]

[4]This can also be deduced from the likelihood function (3.6). If $\tilde{\rho} = 0$, then consistent parameter estimation can be done by maximizing the likelihood function of both the primary equation and the selection equation separately, after the endogeneity corrections have been included into these equations. However,

5. Monte Carlo Results

In order to gauge the bias which occurs if one does not account for endogeneity, we conducted some Monte Carlo simulations whose results are presented in table 1.

The first column of table 1 contains the specification. We distinguish between four benchmark cases. In the first case, endogeneity is only present in the primary equation. In particular, it is assumed that

$$\begin{array}{lllllll} y_i^* & = .2 & +.4X_{1i} & +.9X_{2i} & & & +u_i \\ z_i^* & = 1 & & & +.7W_{1i} & & +v_i \\ X_{2i} & = .5 & +1.5X_{1i} & & -.2W_{1i} & +.7Z_{1i} & +\varepsilon_{1i} \end{array}$$

and

$$\text{Cov}[(u_i, v_i, \varepsilon_{1i})'] = \begin{pmatrix} 1 & & \\ .9 & 1 & \\ .5 & .4 & 2 \end{pmatrix}.$$

Note that we have assumed a relatively high correlation between the primary and the selection equation. Hence, we focus our attention on situations where sample selection bias is indeed a problem.

In the second case, endogeneity is only present in the selection equation:

$$\begin{array}{lllllll} y_i^* & = .2 & +.4X_{1i} & & & & +u_i \\ z_i^* & = 1 & +.7X_{1i} & +.3W_{2i} & & & +v_i \\ W_{2i} & = .5 & +1.5X_{1i} & & +.7Z_{2i} & & +\varepsilon_{2i} \end{array}$$

unless the matrix Σ has a special structure, this approach will be inefficient in general.

and

$$\mathrm{Cov}[(u_i, v_i, \varepsilon_{2i})'] = \begin{pmatrix} 1 & & \\ .9 & 1 & \\ .5 & .4 & 2 \end{pmatrix}.$$

In the third case, there is one common variable in both equations which is endogenous:

$$\begin{array}{lllllll} y_i^* & = .2 & +.4X_{1i} & & +.9C_i & & +u_i \\ z_i^* & = 1 & & +.7W_{1i} & +.3C_i & & +v_i \\ C_i & = .5 & +1.5X_{1i} & -.2W_{1i} & & +.7Z_{3i} & +\varepsilon_{3i} \end{array}$$

and

$$\mathrm{Cov}[(u_i, v_i, \varepsilon_{3i})'] = \begin{pmatrix} 1 & & \\ .9 & 1 & \\ .5 & .4 & 2 \end{pmatrix}.$$

Finally, in the fourth case it is assumed that both equations include an endogenous variable which is exclusive for each equation:

$$\begin{array}{llllllll} y_i^* & = .2 & +.4X_{1i} & +.9X_{2i} & & & & +u_i \\ z_i^* & = 1 & +.7X_{1i} & & +.3W_{2i} & & & +v_i \\ X_{2i} & = .5 & +1.5X_{1i} & & & +.7Z_{1i} & & +\varepsilon_{1i} \\ W_{2i} & = -2 & +1.8X_{1i} & & & & +.6Z_{2i} & +\varepsilon_{2i} \end{array}$$

and

$$\mathrm{Cov}[(u_i, v_i, \varepsilon_{1i}, \varepsilon_{2i})'] = \begin{pmatrix} 1 & & & \\ .9 & 1 & & \\ .5 & .4 & 2 & \\ .4 & .5 & 1 & 2 \end{pmatrix}.$$

Throughout, X_{1i}, Z_{1i}, Z_{2i} and Z_{3i}, $i = 1, \ldots, n$, are scalars which have been simulated from a standard normal distribution. For each of the four cases, these random numbers have been drawn once and kept fixed during simulation. In total, each simulation encompasses 1000 repetitions in which parameter estimates have been computed. Table 1 presents the mean of these estimates over the repetitions, along with the corresponding standard deviations.

In order to gauge the finite-sample performance of the estimator outlined in section 3, table 1 contains simulation results for different sample sizes. For each sample size, table 1 displays the results for the FIML estimator presented in section 3 ("IV") and contrasts these results with those obtained when using the ordinary estimator for the sample selection model which does not account for endogeneity ("non-IV"). To save space, only the estimates for the parameters of the primary equation and selection equation are presented.

In specification (i) where there is only one endogenous variable included in the primary equation, the IV estimator performs well with respect to the estimates of the primary equation, even for $n = 100$. However, the estimates for the selection equation are upward biased in finite samples; this property is common in all specifications (i)-(iv). In specification (ii) where there is only one endogenous variable in the selection equation, the estimator for the primary equation does well for $n \geq 200$. This is also true for specification (iii) with a common endogenous variable in both equations. When each equation contains an exclusive endogenous variable (specification (iv)), good results are obtained for $n \geq 500$.

On the contrary, in most cases the non-IV estimator yields severely biased estimates of the parameters of the primary equation among all specifications. For instance, for a sample size of $n = 1000$ the bias ranges from 13 to 248.1 percent. However, the estimates of the selection equation are sometimes relatively close to their true values (specifications (i) and (iii)). This notwithstanding, note especially that the estimates of the parameters of the main equation are severely biased even if endogeneity is only present in the selection equation (specification (ii)). This result, which is due to the nonlinearity of the underlying

model, has not gained much attention in the literature yet.

Overall, the results show that the FIML-IV estimator from section 3 outperforms the ordinary estimator for the sample selection model, especially with respect to the parameters in the primary equation and in case of large sample sizes. Moreover, the results indicate that the bias in the parameter estimates may be substantial if one does not account for endogeneity.

6. Empirical Application

In the following, we employ the labor supply data used by Mroz (1987) as well as data from the German Socio-Economic Panel (GSOEP) of 2008 to give two examples where the methods developed in this paper can be applied.

For both data sets, we estimated a wage equation for married women. However, as a wage equation can only be fitted to the subsample of women who are actually working, a simple regression with the women's wage as the dependent variable may yield inconsistent parameter estimates due to the possibility of sample selection. Hence, the appropriate model to estimate the wage equation should be a sample selection model. A variable which is commonly included as an explanatory variable is education. However, there might be some background variables like ability which cannot be observed and, thus, are captured within the error terms. These variables are likely to affect not only wages and labor force participation, but education as well. Therefore, *a priori* education should not be regarded as exogenous. The consequences of falsely treating an endogenous variable like education as exogenous have been illustrated in the preceding section; hence, estimates from the ordinary sample selection model may be severely biased.

We estimated the following model: The primary equation contains the natural logarithm of the hourly wage as its dependent variable; explanatory variables are experience, experience squared and education. The selection equation includes experience, experience squared, non-wife income, age, number of children aged until 6 years of age in the household, number of children aged 6 years or older in the household and education.

Since education is treated as endogenous, instrumental variables are needed for estimation. Following Wooldridge (2010), we chose mother's education, father's education and husband's education as instrumental variables for education.[5] Means and standard deviations of these variables are presented in table 2.

Results for the data of Mroz are shown in table 3, while table 4 presents the results for the GSOEP data. In both tables, estimation results for the ordinary sample selection model ("non-IV") and the sample selection model with endogeneity ("IV") are given. The first part of these tables contains the parameter estimates for the variables of the primary equation, as well as estimates of the selection parameter $\tilde{\rho}$ and the endogeneity parameter ψ_{11}. This last parameter indicates whether endogeneity of education is relevant in the primary equation. The second part presents the parameter estimates for the selection equation. Additionally included is the endogeneity parameter ψ_{21}, which indicates whether endogeneity of education is relevant in the selection equation. Finally, the third part includes the parameter estimates of the exogenous variables and instrumental variables with respect to education. In analogy with the instrumental variables terminology, this part has been labeled "first stage".

First, consider table 3. The results show significance of education in the primary and the selection equation. Moreover, the instrumental variables for education employed in the "first stage" are highly significant. The remaining variables possess the expected signs. However, the estimates of $\tilde{\rho}$, ψ_{11} and ψ_{21} are not significantly different from zero, indicating that there is neither a selection bias nor an endogeneity bias present.[6] These results are in line with those obtained by Wooldridge (2010). In this case, therefore, applying OLS to the wage equation would be sufficient.

We now turn to the results for the GSOEP data (table 4). In this example, there is indeed evidence for sample selection bias. In both the "non-IV" and "IV" setting, the estimate of $\tilde{\rho}$ is substantial in absolute value and highly significant. Moreover, after endogeneity in the education variable has been controlled for, the coefficient of education

[5] For the appropriateness of these instrumental variables, cf. the discussion in Card (1999), pp. 1822-26.
[6] In addition, joint significance of ψ_{11} and ψ_{21} is rejected as well (p-value of 0.1907).

in the primary equation changes from 0.069 to 0.1051. In the selection equation, the coefficient of education becomes insignificant. In this case, the endogeneity parameters ψ_{11} and ψ_{21} are highly significant, thus providing evidence for the endogeneity of education.

To summarize, the estimation results for the GSOEP data clearly show the importance of controlling not only for sample selection bias but for the endogeneity of covariates as well. In this example, the returns to education would have been severely underestimated if one did only control for sample selection bias but not for endogeneity, thus confirming the results of section 5.

7. Conclusion

In this paper we have developed a full information maximum likelihood estimation framework for the sample selection model with endogenous covariates. Moreover, we have established straightforward tests for exogeneity and the absence of sample selection bias.

Drawbacks of this estimation framework are that it crucially depends on the normality assumption and that it, therefore, does not encompass binary endogenous covariates. However, modifications of normality would complicate the structure of the likelihood function and thus take away some simplicity of the approach undertaken here, so these issues have been ignored.

The main benefits of the framework are its simplicity and its asymptotic efficiency, provided the distributional assumptions are satisfied. Moreover, in contrast to two stage least squares approaches, no standard error adjustment is necessary.

As the Monte Carlo results of section 5 and the empirical examples from section 6 have shown, falsely ignoring endogeneity of covariates in sample selection models leads to severely biased parameter estimates. This underlines the necessity to employ appropriate econometric models to account for these issues. This paper is an attempt to do so.

References

Amemiya T. 1985. *Advanced Econometrics*. Oxford: Basil Blackwell.

Card D. 1999. The causal effect of education on earnings. In Ashenfelter O, Card D (eds.) *Handbook of Labor Economics*, volume 3 of *Handbook of Labor Economics*, chapter 30. Elsevier, 1801–1863.
URL http://ideas.repec.org/h/eee/labchp/3-30.html

Chib S, Greenberg E, Jeliazkov I. 2009. Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics* **18**: 321–348.
URL doi:10.1198/jcgs.2009.07070

Das M, Newey WK, Vella F. 2003. Nonparametric estimation of sample selection models. *Review of Economic Studies* **70**: 33–58.
URL http://ideas.repec.org/a/bla/restud/v70y2003i1p33-58.html

Davidson R, MacKinnon JG. 1993. *Estimation and Inference in Econometrics*. New York, NY: Oxford University Press.

Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* **47**: 153–61.
URL http://ideas.repec.org/a/ecm/emetrp/v47y1979i1p153-61.html

Mroz TA. 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* **55**: 765–99.
URL http://ideas.repec.org/a/ecm/emetrp/v55y1987i4p765-99.html

Newey WK. 1987. Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics* **36**: 231–250.
URL http://ideas.repec.org/a/eee/econom/v36y1987i3p231-250.html

Puhani PA. 2000. The heckman correction for sample selection and its critique. *Journal of Economic Surveys* **14**: 53–68.
URL http://ideas.repec.org/a/bla/jecsur/v14y2000i1p53-68.html

Rivers D, Vuong QH. 1988. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics* **39**: 347–366.
URL http://ideas.repec.org/a/eee/econom/v39y1988i3p347-366.html

Semykina A, Wooldridge JM. 2010. Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics* **157**: 375–380.
URL http://ideas.repec.org/a/eee/econom/v157y2010i2p375-380.html

Smith RJ, Blundell RW. 1986. An exogeneity test for a simultaneous equation tobit model with an application to labor supply. *Econometrica* **54**: 679–85.
URL http://ideas.repec.org/a/ecm/emetrp/v54y1986i3p679-85.html

Vella F. 1998. Estimating models with sample selection bias: A survey. *Journal of Human Resources* **33**: 127–169.
URL http://ideas.repec.org/a/uwp/jhriss/v33y1998i1p127-169.html

Vella F, Verbeek M. 1999. Two-step estimation of panel data models with censored endogenous variables and selection bias. *Journal of Econometrics* **90**: 239–263.
URL http://ideas.repec.org/a/eee/econom/v90y1999i2p239-263.html

Wooldridge JM. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press, 2nd edition.

Table 1: Monte Carlo Results

Spec.	Param.	$n = 100$		$n = 200$		$n = 500$		$n = 1000$	
		IV	non-IV	IV	non-IV	IV	non-IV	IV	non-IV
	$\beta_1 = .2$	.2397	.1409	.2031	.0934	.2028	.1168	.2014	.0988
		(.1500)	(.1498)	(.0968)	(.0887)	(.0556)	(.0529)	(.0416)	(.0381)
	$\beta_2 = .4$	.4019	−.0191	.3947	.0396	.4023	.0338	.3988	.0379
		(.2439)	(.1535)	(.1532)	(.0983)	(.0945)	(.0664)	(.0621)	(.0413)
(i)	$\beta_3 = .9$	.8991	1.1570	.9020	1.1412	.8978	1.1415	.9007	1.1404
		(.1396)	(.0781)	(.0933)	(.0525)	(.0567)	(.0347)	(.0381)	(.0220)
	$\gamma_1 = 1$	1.1316	1.0201	1.1043	1.0101	1.1016	1.0086	1.0995	1.0087
		(.2492)	(.1993)	(.1467)	(.1270)	(.0867)	(.0758)	(.0625)	(.0553)
	$\gamma_2 = .7$	.8567	.7483	.7895	.7067	.7724	.6744	.7688	.6707
		(.2445)	(.2169)	(.1337)	(.1264)	(.0815)	(.0795)	(.0574)	(.0564)
	$\beta_1 = .2$	.3068	.6661	.2234	.6784	.2000	.6719	.2001	.6962
		(.2070)	(.2250)	(.1203)	(.1531)	(.0597)	(.1178)	(.0395)	(.0642)
	$\beta_2 = .4$	.3082	.0520	.3818	.0181	.4009	.0340	.4000	.0128
		(.1726)	(.1892)	(.1170)	(.1426)	(.0561)	(.1012)	(.0411)	(.0584)
(ii)	$\gamma_1 = 1$	1.1567	.9346	1.1254	.8766	1.1021	.8544	1.0967	.8541
		(.2989)	(.2554)	(.1853)	(.1623)	(.1085)	(.1093)	(.0743)	(.0690)
	$\gamma_2 = .7$	.8226	.2775	.7896	.2177	.7743	.2391	.7708	.2292
		(.5229)	(.3628)	(.3142)	(.2517)	(.1624)	(.1646)	(.1143)	(.0994)
	$\gamma_3 = .3$	.3685	.6418	.3451	.6291	.3316	.5854	.3250	.5851
		(.3325)	(.2152)	(.1895)	(.1403)	(.0897)	(.0826)	(.0672)	(.0513)
	$\beta_1 = .2$	.2681	.1575	.2113	.0981	.2010	.0825	.2005	.0863
		(.1695)	(.1742)	(.0987)	(.1015)	(.0588)	(.0570)	(.0431)	(.0392)
	$\beta_2 = .4$	.3874	.0147	.4091	.0145	.4007	.0327	.4012	.0348
		(.2270)	(.1553)	(.1554)	(.1031)	(.0963)	(.0631)	(.0635)	(.0440)
(iii)	$\beta_3 = .9$	.8858	1.1484	.8893	1.1739	.8992	1.1724	.8977	1.1664
		(.1339)	(.0829)	(.0957)	(.0588)	(.0592)	(.0346)	(.0403)	(.0238)
	$\gamma_1 = 1$	1.1446	1.0109	1.1222	.9984	1.1044	.9923	1.0987	.9819
		(.2707)	(.2044)	(.1637)	(.1346)	(.0969)	(.0861)	(.0630)	(.0561)
	$\gamma_2 = .7$	.8557	.7658	.8053	.7422	.7760	.7292	.7711	.7180
		(.2600)	(.2334)	(.1556)	(.1520)	(.0877)	(.0872)	(.0582)	(.0576)
	$\gamma_3 = .3$	.3569	.4696	.3380	.4160	.3324	.4256	.3286	.4216
		(.1622)	(.1385)	(.0834)	(.0756)	(.0501)	(.0455)	(.0349)	(.0313)
	$\beta_1 = .2$	.4320	.3423	.2554	.2899	.1995	.2248	.1988	.2260
		(.3394)	(.2752)	(.2044)	(.1967)	(.0835)	(.0876)	(.0601)	(.0649)
	$\beta_2 = .4$	.2738	.0267	.3687	.0735	.4053	.1103	.3994	.1036
		(.3803)	(.2147)	(.2173)	(.1532)	(.1219)	(.0819)	(.0818)	(.0603)
(iv)	$\beta_3 = .9$	.8887	1.0489	.8965	1.0462	.8983	1.0516	.9010	1.0514
		(.1856)	(.0747)	(.1063)	(.0480	(.0651)	(.0304)	(.0429)	(.0209)
	$\gamma_1 = 1$	1.2063	1.5246	1.1415	1.5172	1.0920	1.4562	1.0882	1.4517
		(.5953)	(.39175)	(.4180)	(.2665)	(.2316)	(.1525)	(.1597)	(.1111)
	$\gamma_2 = .7$	.8397	.4488	.7793	.4218	.7665	.4216	.7599	.4254
		(.5378)	(.2963)	(.3654)	(.1890)	(.2137)	(.1099)	(.1391)	(.0805)
	$\gamma_3 = .3$	.3724	.5504	.3450	.5326	.3281	.5056	.3278	.5041
		(.2849)	(.1572)	(.1935)	(.1060)	(.1062)	(.0604)	(.0719)	(.0426)

Table 2: Descriptive Statistics

	Mroz		GSOEP 2008	
Variable	Mean	Std.dev.	Mean	Std.dev.
log wage	4.1777	3.3103	2.1304	0.4775
exper	10.6308	8.0691	17.1184	8.6946
educ	12.2869	2.2802	12.8439	2.6118
nwifeinc	20.1290	11.6348	33.9573	20.6718
age	42.5379	8.0726	43.8712	7.4560
kidslt6	0.2377	0.5240	0.2142	0.5043
kidsge6	1.3533	1.3199	0.5847	0.8524
motheduc	9.2510	3.3675	9.4564	0.9952
fatheduc	8.8088	3.5723	9.7009	1.3346
huseduc	12.4914	3.0208	13.0674	2.8157
Sample size	753		2143	
No. of obs. with wage>0	428		1561	

Table 3: Estimation of a Wage Equation for Married Women - Mroz

	non-IV		IV	
Primary Equation				
const	-0.5527^{**}	(0.2604)	-0.2786	(0.3139)
exper	0.0428^{***}	(0.0149)	0.0449^{***}	(0.0151)
expersq	-0.00008^{**}	(0.0004)	-0.0009^{**}	(0.0004)
educ	0.1084^{***}	(0.0149)	0.0849^{***}	(0.0218)
$\tilde{\rho}$	0.0141	(0.1491)	0.0248	(0.1492)
ψ_{11}			0.0413	(0.0290)
Selection Equation				
const	0.2664	(0.5090)	0.6084	(0.6522)
exper	0.1233^{***}	(0.0187)	0.1261^{***}	(0.0191)
expersq	-0.0019^{***}	(0.0006)	-0.0019^{***}	(0.0006)
nwifeinc	-0.0121^{**}	(0.0049)	-0.0105^{*}	(0.0053)
age	-0.0528^{***}	(0.0085)	-0.0543^{***}	(0.0087)
kidslt6	-0.8674^{***}	(0.1187)	-0.8620^{***}	(0.1190)
kidsge6	0.0359	(0.0435)	0.0316	(0.0438)
educ	0.1313^{***}	(0.0254)	0.1046^{**}	(0.0406)
ψ_{21}			0.0425	(0.0502)
"First Stage"				
const			5.3947^{***}	(0.5826)
exper			0.0577^{***}	(0.0219)
expersq			-0.0008	(0.0007)
nwifeinc			0.0147^{**}	(0.0058)
age			-0.0051	(0.0098)
kidslt6			0.1269	(0.1298)
kidsge6			-0.0700	(0.0511)
motheduc			0.1307^{***}	(0.0224)
fatheduc			0.0951^{***}	(0.0212)
huseduc			0.3489^{***}	(0.0233)

*, ** and *** indicate significance at 1%, 5% and 10%, respectively. Standard errors in parentheses.

Table 4: Estimation of a Wage Equation for Married Women - GSOEP 2008

	non-IV		IV	
		Primary Equation		
const	1.3279***	(0.1066)	0.8969***	(0.1212)
exper	−0.0048	(0.0069)	−0.0083	(0.0067)
expersq	0.0002	(0.0002)	0.0003*	(0.0002)
educ	0.0690***	(0.0046)	0.1051***	(0.0075)
$\tilde{\rho}$	−0.6512***	(0.0710)	−0.6890***	(0.0578)
ψ_{11}			−0.0598***	(0.0099)
		Selection Equation		
const	0.4917	(0.3240)	1.3767***	(0.3844)
exper	0.1728***	(0.0160)	0.1714***	(0.0161)
expersq	−0.0019***	(0.0004)	−0.0019***	(0.0004)
nwifeinc	0.0127***	(0.0017)	0.0156***	(0.0019)
age	−0.0834***	(0.0074)	−0.0802***	(0.0073)
kidslt6	−0.5693***	(0.0827)	−0.5040***	(0.0810)
kidsge6	0.0193	(0.0380)	0.0222	(0.0377)
educ	0.1032***	(0.0145)	0.0183	(0.0261)
ψ_{21}			0.1259***	(0.0315)
		"First Stage"		
const			1.1016*	(0.5962)
exper			0.0264	(0.0203)
expersq			−0.0011**	(0.0005)
nwifeinc			0.0002***	(0.0000)
age			0.0037	(0.0087)
kidslt6			0.3709***	(0.1041)
kidsge6			−0.1051*	(0.0546)
motheduc			0.2870***	(0.0531)
fatheduc			0.3442***	(0.0399)
huseduc			0.3664***	(0.0179)

*, ** and *** indicate significance at 1%, 5% and 10%, respectively. Standard errors in parentheses.