Semiparametric Estimation of a Sample Selection Model in the Presence of Endogeneity

Jörg Schwiebert*

Abstract

In this paper, we derive a semiparametric estimation procedure for the sample selection model when some covariates are endogenous. Our approach is to augment the main equation of interest with a control function which accounts for sample selectivity as well as endogeneity of covariates. In contrast to existing methods proposed in the literature, our approach allows that the same endogenous covariates may enter the main *and* the selection equation. We show that our proposed estimator is $\sqrt{n}$-consistent and derive its asymptotic distribution. We provide Monte Carlo evidence on the small sample behavior of our estimator and present an empirical application. Finally, we briefly consider an extension of our model to quantile regression settings and provide guidelines for estimation.

Keywords: Sample selection model, semiparametric estimation, endogenous covariates, control function approach, quantile regression.
JEL codes: C21, C24, C26.

*Leibniz University Hannover, Institute of Labor Economics, Königsworther Platz 1, 30167 Hannover, Germany, Tel.: 0511/762-5657, E-mail: schwiebert@aoek.uni-hannover.de

1 Introduction

Sample selection bias is a problem frequently encountered in applied econometrics. Pioneered by the seminal paper of Heckman (1979), several authors have proposed methods to circumvent the problem in parametric and semi-nonparametric settings. A sample selection model typically consists of a main equation (of interest) and a selection equation. The variables in the main equation, however, can only be fully observed for a subset of the observations, where the probability of full observability is governed by the selection equation.

In this paper, we consider semiparametric estimation of a sample selection model when some explanatory variables are endogenous. As in many other econometric models, the endogeneity of explanatory variables causes parameter estimates to be biased. Hence, in order to obtain unbiased estimates of the parameters of interest, one needs an econometric model which not only accounts for sample selection issues but for endogeneity issues as well. In contrast to existing methods proposed in the literature, our approach allows that the same endogenous covariates may enter the main *and* the selection equation.

Sample selection models which also incorporate endogeneity issues have been previously studied by Wooldridge (2010), Das et al. (2003), Chib et al. (2009) and Semykina and Wooldridge (2010). Based on Heckman's (1979) original formulation of the sample selection model (using a joint normality assumption on the distribution of error terms), Wooldridge (2010) suggests estimating a probit model for the selection equation in the first step, and then to apply two stage least squares to the main equation including the inverse Mills ratio term (which controls for sample selectivity). Semykina and Wooldridge (2010) extend this approach to panel data sample selection models, and also consider semiparametric estimation based on a series expansion due to Newey (2009). Chib et al. (2009) impose a joint normality assumption not only on the error terms of main and selection equation, but on the endogenous covariates as well, and estimate the model using Bayesian techniques. Das et al. (2003) propose nonparametric estimation methods. That is, the main equation, the selection equation and the reduced form equations for the endogenous explanatory variables are being estimated nonparametrically using series methods.

Our approach to estimating a sample selection model with endogenous covariates differs from the parametric ones studied by Wooldridge (2010), Chib et al. (2009) and Semykina and Wooldridge (2010) since our estimation framework is semiparametric. That means, we do not have to rely on possibly too strong distributional assumptions like the joint normality assumption in Heckman's (1979) original formulation. In particular, imposing false distributional assumptions leads to biased estimates of the parameters of interest. However, a nonparametric framework as in Das et al. (2003) may suffer from the well known curse of dimensionality problem if the number of covariates if large.

We propose a semiparametric approach where we impose a set of linearity assumptions as it is common in semiparametric modeling. The main benefits of our semiparametric approach are that it is relatively simple and does not make strong parametric assumptions; furthermore, it helps avoid the curse of dimensionality problem raised in nonparametric settings and is easy to interpret.

Our estimation procedure relies on the Robinson (1988) estimator for partially linear models, which can be labeled a "kernel density" approach since it involves estimation of an unknown function using kernel weights. In particular, we follow Das et al. (2003) and expand the main equation with a "control function" which takes into account sample selection as well as endogeneity of covariates. A control function approach is convenient since it allows for some degree of conditional heteroskedasticity in the main equation (Newey et al., 1999) as well as nonlinearities in the endogenous explanatory variables. On the contrary, Semykina and Wooldridge (2010) use Newey's (2009) series expansion to control for sample selectivity in the main equation, and control for endogeneity by applying two stage least squares. Since we allow for some kind of conditional heteroskedasticity in the main equation, our approach is more general in this sense. Moreover, in series-based frameworks one has to specify basis functions to be used in the series expansions (e.g., polynomials, splines). Estimation results may be sensitive to this choice. On the other hand, a kernel-based approach requires instead the specification of bandwidth parameters. These are easier to alter, so that robustness checks may be obtained more easily.

Quite more important, our control function approach allows that the same endogenous covariates appear in the main equation *and* the selection equation as well. In many empirical applications there are common variables appearing in the main *and* the selection equation. Some of these might be endogenous. For instance, the classical application of the sample selection model is a wage regression for married women. In that case, both the wage (main equation) as well as the probability of labor force participation (selection equation) depend on educational attainment. Education, however, may be endogenous in both equations due to some underlying (and unobservable) factors summarized by the term "ability". Since ability is unobservable and correlated with the wage, the probability of labor force participation and educational attainment, we have a typical situation of endogeneity bias.

Wooldridge's (2010) approach (or Semykina and Wooldridge, 2010) only considers endogeneity in the main equation. However, if the selection equation also includes these endogenous covariates (which is quite realistic) one gets biased estimates even if the selection equation parameters have been estimated consistently. The reason is that the control function which depends on the selection index is itself endogenous through its dependence on the endogenous variables appearing in the selection equation. Moreover, the control function depends on both the selection index and the endogenous explanatory variables in general.

As a consequence, Wooldridge's method is only valid if we have an endogenous explanatory variable in the main equation which does not appear in the selection equation. On the contrary, our approach allows for the same endogenous covariates in both equations, and is thus more general.

The paper is organized as follows. In section 2, we set up the model and propose our estimation procedure. In section 3 we discuss the asymptotic properties of our estimator. In section 4 we give Monte Carlo Results on the small sample properties of our proposed estimator. We also compare our estimator to that of Wooldridge (2010). In section 5 we present an economic application in which we analyze how the income level affects the number of children of working women. Section 6 outlines an extension of our model to quantile regression settings. Finally, section 7 concludes the paper.

2 Model Setup and Estimation

The model we consider is given by

$$y_{1i}^* = x_i'\beta + \delta y_{2i} + \varepsilon_i \tag{1}$$

$$d_i^* = w_i'\gamma + u_i \tag{2}$$

$$d_i = 1(d_i^* > 0) \tag{3}$$

$$y_{1i} = \begin{cases} y_{1i}^* & \text{if } d_i = 1 \\ \text{“missing” otherwise} \end{cases} \tag{4}$$

$$y_{2i} = z_i'\alpha + v_i \text{ if } d_i = 1 \tag{5}$$

where $i = 1, \dots, N$ indexes individuals. The first equation is the main equation (of interest), where y_1^* is the latent dependent variable, x is a vector of exogenous explanatory variables, y_2 is an endogenous explanatory variable and ε is an error term. For simplicity, we consider only one endogenous explanatory variable, but an extension is straightforward. The second equation is the selection equation, where d^* is the latent dependent variable, w is a vector of exogenous explanatory variables and u is the error term. The third equation expresses that only the sign of d^* is observable. The fourth equation comprises the sample selection mechanism: y^* is only observable if the selection indicator d is equal to one. The fifth equation is the reduced form equation for the endogenous explanatory variable y_2, where z is a vector of exogenous explanatory (instrumental) variables and v is an error term.

Following Newey et al. (1999) and Das et al. (2003), we make the following assumptions:

Assumption 1: $E[\varepsilon_i|d_i = 1, w_i, x_i, z_i, v_i] = E[\varepsilon_i|w_i'\gamma, v_i] = g(w_i'\gamma, v_i)\ \forall i = 1, \ldots, N.$

Assumption 2: $Pr(x_i'\beta + \delta y_{2i} + \tilde{g}(w_i'\gamma, v_i) = 0|d_i = 1) = 1$ *implies there is a constant* c *with* $Pr(x_i'\beta + \delta y_{2i} = c|d_i = 1) = 1.$

The unknown function $g(\cdot)$ is our control function, which is assumed to depend only on the propensity score $w'\gamma$ and the reduced form error term v. Assumptions 1 implies that

$$E[y_{1i}|d_i = 1, w_i, x_i, z_i, v_i] = x_i'\beta + \delta y_{2i} + g(w_i'\gamma, v_i). \tag{6}$$

Note that Assumption 1 allows for conditional heteroskedasticity in the sense that ε may be heteroskedastic in $w'\gamma$ and v.

Assumption 2 is an identifying assumption which is needed to identify the parameters in β and δ. More precisely, there must not exist an exact functional relationship between the linear part of equation (6) and the unknown function $g(\cdot)$. A sufficient condition for this assumption to be fulfilled is that the selection equation as well as the reduced form equation for y_2 contain at least one variable which is exclusive in these equations. Note, however, that a constant term in x is not identified since it cannot be distinguished from the constant part of the unknown function $g(\cdot)$.

A convenient choice for estimating this model is the Robinson (1988) estimator for partially linear models. By equation (6), we can rewrite (the observable part of) the main equation as

$$y_{1i} = x_i'\beta + \delta y_{2i} + g(w_i'\gamma, v_i) + r_i, \quad i = 1, \ldots, n, \tag{7}$$

where $r_i \equiv y_{1i} - E[y_{1i}|d_i = 1, w_i, x_i, z_i, v_i]$ has a conditional mean of zero. Note that n denotes the number of individuals for which $d_i > 0$. Obviously the main equation consists of a linear part and the nonlinear function $g(\cdot)$. The idea of the Robinson estimator is to get rid of the unknown function $g(\cdot)$. To do this, take expectations of equation (7) conditional on $w_i'\gamma$ and v_i. This yields

$$E[y_{1i}|w_i'\gamma, v_i] = E[x_i|w_i'\gamma, v_i]'\beta + \delta E[y_{2i}|w_i'\gamma, v_i] + g(w_i'\gamma, v_i). \tag{8}$$

Subtracting (8) from (7) gives

$$y_{1i} - E[y_{1i}|w_i'\gamma, v_i] = (x_i - E[x_i|w_i'\gamma, v_i])'\beta + \delta(y_{2i} - E[y_{2i}|w_i'\gamma, v_i]) + r_i. \tag{9}$$

Since $E[y_{1i}|w_i'\gamma, v_i]$ and $E[x_i|w_i'\gamma, v_i]$ are unknown, they are replaced by kernel estimates such that

$$\hat{y}_{1i} \equiv \hat{E}[y_{1i}|w_i'\gamma, v_i] \equiv \frac{\frac{1}{n}\sum_{j=1}^n y_{1j}\frac{1}{h^2}K((w_j - w_i)'\gamma/h)K((v_j - v_i)/h)}{\frac{1}{n}\sum_{j=1}^n \frac{1}{h^2}K((w_j - w_i)'\gamma/h)K((v_j - v_i)/h)} \tag{10}$$

$$\hat{y}_{2i} \equiv \hat{E}[y_{2i}|w_i'\gamma, v_i] \equiv \frac{\frac{1}{n}\sum_{j=1}^n y_{2j}\frac{1}{h^2}K((w_j - w_i)'\gamma/h)K((v_j - v_i)/h)}{\frac{1}{n}\sum_{j=1}^n \frac{1}{h^2}K((w_j - w_i)'\gamma/h)K((v_j - v_i)/h)} \tag{11}$$

$$\hat{x}_i \equiv \hat{E}[x_i|w_i'\gamma, v_i] \equiv \frac{\frac{1}{n}\sum_{j=1}^n x_j\frac{1}{h^2}K((w_j - w_i)'\gamma/h)K((v_j - v_i)/h)}{\frac{1}{n}\sum_{j=1}^n \frac{1}{h^2}K((w_j - w_i)'\gamma/h)K((v_j - v_i)/h)}, \tag{12}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a kernel function (for example, the standard normal probability density function) and h is the bandwidth parameter satisfying $h \to 0$ as $n \to \infty$. For simplicity we have assumed the same kernel functions for $w'\gamma$ and v as well as the same bandwidths h. Let $q_i = (x_i', y_{2i})'$. The feasible Robinson estimator of $\theta = (\beta', \delta)'$ is then given by

$$\hat{\theta} = \left(\sum_{i=1}^n (q_i - \hat{q}_i)(q_i - \hat{q}_i)'d_i\right)^{-1} \sum_{i=1}^n \{(q_i - \hat{q}_i)(y_{1i} - \hat{y}_{1i})d_i\}, \tag{13}$$

where $\hat{q}_i = (\hat{x}_i, \hat{y}_{2i})$. Under some regularity conditions, it can be shown that the Robinson estimator is $\sqrt{n}$-consistent and has an asymptotic normal distribution.

However, for making inference we cannot use the asymptotic normality results of the Robinson estimator since we cannot observe γ and v. We thus have to replace these with "first stage" estimates $\hat{\gamma}$ and $\hat{v} \equiv y_2 - z'\hat{\alpha}$, respectively. We then have to replace the infeasible estimator in (13) with

$$\hat{\hat{\theta}} = \left(\sum_{i=1}^{n} (q_i - \hat{\hat{q}}_i)(q_i - \hat{\hat{q}}_i)' d_i \right)^{-1} \sum_{i=1}^{n} \{ (q_i - \hat{\hat{q}}_i)(y_{1i} - \hat{\hat{y}}_{1i}) d_i \}, \tag{14}$$

where $\hat{\hat{q}}_i = (\hat{\hat{x}}_i, \hat{\hat{y}}_{2i})$ and

$$\hat{\hat{y}}_{1i} \equiv \hat{E}[y_{1i} | w_i'\hat{\gamma}, \hat{v}_i] \equiv \frac{\frac{1}{n}\sum_{j=1}^{n} y_{1j} \frac{1}{h^2} K((w_j - w_i)'\hat{\gamma}/h) K((\hat{v}_j - \hat{v}_i)/h)}{\frac{1}{n}\sum_{j=1}^{n} \frac{1}{h^2} K((w_j - w_i)'\hat{\gamma}/h) K((\hat{v}_j - \hat{v}_i)/h)} \tag{15}$$

$$\hat{\hat{y}}_{2i} \equiv \hat{E}[y_{2i} | w_i'\hat{\gamma}, \hat{v}_i] \equiv \frac{\frac{1}{n}\sum_{j=1}^{n} y_{2j} \frac{1}{h^2} K((w_j - w_i)'\hat{\gamma}/h) K((\hat{v}_j - \hat{v}_i)/h)}{\frac{1}{n}\sum_{j=1}^{n} \frac{1}{h^2} K((w_j - w_i)'\hat{\gamma}/h) K((\hat{v}_j - \hat{v}_i)/h)} \tag{16}$$

$$\hat{\hat{x}}_i \equiv \hat{E}[x_i | w_i'\hat{\gamma}, \hat{v}_i] \equiv \frac{\frac{1}{n}\sum_{j=1}^{n} x_j \frac{1}{h^2} K((w_j - w_i)'\hat{\gamma}/h) K((\hat{v}_j - \hat{v}_i)/h)}{\frac{1}{n}\sum_{j=1}^{n} \frac{1}{h^2} K((w_j - w_i)'\hat{\gamma}/h) K((\hat{v}_j - \hat{v}_i)/h)}. \tag{17}$$

Unfortunately, replacing γ and v with estimates alters the asymptotic properties of the Robinson estimator, which will be further investigated in the next section.

3 Asymptotic Properties

Before we proceed with the asymptotic analysis, we make a comment on the interpretation of the sample size n. In our formulation, n refers to the number of non-missing observations in y_1. This interpretation of the sample size has been suggested by Powell (1987), for instance. This implies that the selection equation may be estimated at a faster rate than $\sqrt{n}$; for example a parametric estimation procedure such as probit or logit would yield a rate of $\sqrt{N}$, where N denotes the number of individuals for which the variables of the selection equation are fully observable. The fact that the selection equation (and possibly the reduced form equation for y_2) can be estimated at a faster rate than the main equation is crucial to our asymptotic analysis in the following, since in this case the asymptotic distribution of the Robinson estimator is unaffected by the fact that we use estimated values of $w'\gamma$ and v instead of the true values.

In order to derive the asymptotic properties of the estimator proposed in the last section, we briefly state the assumptions which guarantee the consistency and asymptotic normality properties of the ordinary (infeasible) Robinson estimator. We follow the exposition in Li and Racine (2009) which covers more general cases than in Robinson's original formulation (especially conditional heteroskedasticity).

Assumption 3: *$(y_{1i}, q_i, w_i, z_i), i = 1, \ldots, N$ are i.i.d. observations, $(w'\gamma, v)$ has a density which is bounded from above, $g(\cdot) \in \mathcal{G}_\nu^4$, and $E[q|w'\gamma, v] \in \mathcal{G}_\nu^4$, where $\nu \geq 2$ is an integer.*

In this formulation, $\mathcal{G}_\nu^4$ denotes a class of functions which are ν times differentiable and satisfy Lipschitz conditions such as $|g(z) - g(z')| = H_g(z)||z' - z||$, where $H_g(z)$ is a continuous function with $E|H_g(z)|^4 < \infty$ and $|| \cdot ||$ denotes the Euclidean norm. Hence, Assumption 3 puts some restrictions on the smoothness of the unknown function $g(\cdot)$.

Next, we make some assumptions on the moments of r and q.

Assumption 4: *$E[r|q, w'\gamma, v] = 0$, $E[r^2|q, w'\gamma, v] = \sigma^2(w'\gamma, v)$ is continuous in $(w'\gamma, v)$, $E|q|^4 < \infty$ and $E|r|^4 < \infty$.*

The next two assumptions deal with the kernel function $K(\cdot)$ and the bandwidth parameter h.

Assumption 5: *$K(\cdot)$ is a bounded νth order kernel with $K(t) = O(1/[1 + |t|]^{\nu+1})$.*

Assumption 6: *$nh^4 \to \infty$ and $nh^{4\nu} \to 0$ as $n \to \infty$.*

The conditions on the bandwidth parameter in Assumption 6 ensure that replacing the conditional expectations in (9) with kernel estimates does not alter the asymptotic distribution.

Under Assumptions 1-6, it follows from Robinson (1988) that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Phi^{-1}\Psi\Phi^{-1})$ for the *infeasible* estimator, where $\hat{\Phi} = E[(q_i - q_i^*)(q_i - q_i^*)']$, $\hat{\Psi} = E[\sigma^2(q_i, w_i'\gamma, v_i)(q_i - q_i^*)(q_i - q_i^*)']$ and $q_i^* = E[q_i|w_i'\gamma, v_i]$.

In order to establish asymptotic properties of our *feasible* estimator $\hat{\hat{\theta}}$, we make the following assumption on the first stage estimators $\hat{\gamma}$ and $\hat{\alpha}$.

Assumption 7: *(i) γ and α each lie in the interior of a compact set; (ii) $\hat{\gamma} - \gamma = O_p(n^{-p_1})$ and $\hat{\alpha} - \alpha = O_p(n^{-p_2})$, respectively, with $p_1 > 1/2$ and $p_2 \geq 1/2$.*

Note that Assumption 7 implies that the selection equation is estimated at a faster rate than $\sqrt{n}$, while the reduced form equation for y_2 may be estimated at the same rate.

3.1 Consistency

The Robinson estimator is consistent in general. This property does also hold true when we use estimates of γ and v rather than the true values, provided that these estimates are consistent themselves. We can, therefore, establish the following lemma:

Lemma 1: *Under Assumptions 1-7,* $\hat{\hat{\theta}} - \theta = o_p(1)$.

Proof: Given the consistency of the feasible Robinson estimator, consistency of the infeasible Robinson estimator (which depends on first-stage estimates) can be easily established using the consistency proof of the ordinary Robinson estimator and an application of Lebesgue's dominated convergence theorem (see Billingsley, 1995, Theorem 16.4).

3.2 Asymptotic normality

The infeasible Robinson estimator has an asymptotic normal distribution which is achieved at a $\sqrt{n}$-rate. However, using estimates of γ and v instead of their true values alters the asymptotic distribution. In estimation problems where a preliminary "first stage" estimator is included, a common strategy to derive the asymptotic distribution is to assume that the preliminary estimator converges at a faster rate to its true value than the actual ("second stage") estimator does. For example, Kyriazidou (1997) proceeds in this way. As mentioned above, the selection equation can typically be estimated at a faster rate than $\sqrt{n}$. If this is also true for the reduced form equation for y_2 (for instance, if the variables in this equation are fully observable and the error term is not correlated with the selection effect), we can establish the following theorem:

Theorem 1: *Under Assumptions 1-7, and if* $p_1 > 1/2$ *and* $p_2 > 1/2$, *then* $\sqrt{n}(\hat{\hat{\theta}} - \theta) \xrightarrow{d} \mathcal{N}(0, \Phi^{-1}\Psi\Phi^{-1})$, *where* Φ *and* Ψ *are defined as before.*

Proof: See appendix.

However, if the reduced form equation for y_2 can only be estimated at rate $\sqrt{n}$, we have the following result:

Theorem 2: *Under Assumptions 1-7, and if* $\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, \Gamma)$*, then* $\sqrt{n}(\hat{\hat{\theta}} - \theta) \xrightarrow{d} \mathcal{N}(0, \Phi^{-1}(\Psi + \Omega)\Phi^{-1})$*, where* Φ *and* Ψ *are defined as before and* Ω *is given by*

$$\Omega = \left(E\left[(q_i - q_i^*) \left(\frac{\partial g_i(w_i'\gamma, v_i)}{\partial v_i} \right) z_i' \right] \right) \Gamma \left(E\left[(q_i - q_i^*) \left(\frac{\partial g_i(w_i'\gamma, v_i)}{\partial v_i} \right) z_i' \right] \right)' . \qquad (18)$$

Proof: See appendix.

Hence, in this case the asymptotic distribution of the feasible estimator depends on the asymptotic properties of the reduced form estimator $\hat{\alpha}$ through the additional matrix Ω in the variance term. Since Ω depends on a partial derivative of the unknown function $g(\cdot)$, one needs to estimate the partial derivative in order to calculate the finite-sample analog to equation (18). Another possibility is to use bootstrap standard errors. Let $\hat{\hat{\theta}}_l$ be the bootstrap estimate of the lth replication. An estimator of the variance of $\hat{\hat{\theta}}$ is then given by

$$\frac{1}{L} \sum_{l=1}^{L} \left(\hat{\hat{\theta}}_l - \hat{\hat{\theta}} \right)^2 , \qquad (19)$$

where L denotes the total number of replications.

4 Monte Carlo Evidence

In this section, we provide some (limited) Monte Carlo evidence on the small sample properties of our proposed estimator. Our simulated model is given by

$$y_{1i}^* = \beta_1 x_i + \beta_2 y_{2i} + \varepsilon_i \tag{20}$$

$$d_i^* = x_i + w_i + u_i \tag{21}$$

$$y_{2i} = x_i + w_i + z_i + v_i \tag{22}$$

$$\varepsilon_i = \eta_i + \nu_{\varepsilon,i} \tag{23}$$

$$u_i = \eta_i + \nu_{u,i} \tag{24}$$

$$v_i = \eta_i + \nu_{v,i} \tag{25}$$

$$d_i = 1(d_i^* > 0) \tag{26}$$

$$y_{1i} = \begin{cases} y_{1i}^* & \text{if } d_i = 1 \\ \text{“missing” otherwise} \end{cases}, \tag{27}$$

$i = 1, \ldots, N$, where $\beta_1 = \beta_2 = 1$, $x \sim U_{[0,1]}$, $w \sim \mathcal{N}(1,1)$, $z \sim \mathcal{N}(1,1)$, $\eta \sim \mathcal{N}(0,1)$, $\nu_\varepsilon \sim \mathcal{N}(0,1)$, $\nu_u \sim \mathcal{N}(0,1)$ and $\nu_v \sim \mathcal{N}(0,1)$.

We also consider a slightly different design where our endogenous explanatory variable y_2 also enters the selection equation. In this case, we replace equation (21) with

$$d_i^* = y_{2i} + w_i + u_i. \tag{28}$$

Note that the three error terms ε, u and v are correlated since they depend on the common factor η. Such a situation may be quite realistic; for instance, when measuring the returns to education for married women it is likely that the wage level, education, and the probability of labor force participation all depend on a common factor like “ability”.

In our Monte Carlo experiments, we simulated data for sample sizes of 250, 500 and 1,000. Each experiment encompasses 1,000 replications. For different estimators and sample sizes, we computed the mean of the estimates of the main equation parameters as well as the root mean squared error and the empirical sizes of t tests with a nominal size of 5%. These t tests test the hypothesis $H_0 : \beta_1 = 1$ and $H_0 : \beta_2 = 1$, respectively. The empirical sizes of the t tests will show how well the asymptotic distribution approximates the finite sample distribution of the estimators we consider. Since we assume that the reduced form equation for y_2 is observable for all individuals and is independent of selection effects, we can rely on theorem 1 of section 3 and apply the asymptotic theory of the ordinary Robinson estimator when performing t tests based on our proposed estimator. For the kernel $K(\cdot)$, we used the standard normal p.d.f., and chose a bandwidth of $h = n^{-1/6.5}$, which is in accordance with Assumption 6 as the standard normal p.d.f. is a second order kernel.

In table 1 we collected estimation results for our simulated model when using what we labeled "naive" estimators. The first naive estimator is a simple OLS estimator which neither controls for endogeneity nor sample selection bias. We see that this estimator is biased and that t tests almost always reject the null hypothesis. The next naive estimator is a two stage least squares (2SLS) estimator which controls for the endogeneity in y_2 (instruments are x, w and z, of course) but not for sample selectivity. Compared to the OLS estimator, we have a smaller bias and smaller sizes of t tests. The bias in both β_1 and β_2 occurs because the fitted values from the first stage regression in 2SLS, which are inserted into the main equation, are - like x - correlated with the omitted selectivity correction term. The relatively good performance of the 2SLS estimator may be due to the fact that we have a relatively small fraction of non-missing observations in y_1 which is on the order of 20%, so that the selectivity effect plays a minor role. Finally, we estimated the model by Heckman's two step method for the sample selection model which controls for selectivity but not for endogeneity. This means, we augmented the main equation with an inverse Mills ratio term (as ε and u are normally distributed). We see that the estimator of β_1 has less bias compared to OLS and more adequate empirical sizes of t tests. However, regarding the estimator of β_2 there is a similar bias and unreliable outcomes of

t tests.

The main conclusion that should be drawn from table 1 is that conventional estimators are biased if they are not able to control for sample selectivity *and* endogeneity. In table 2, we present results for two estimators which indeed control for both types of specification errors. The first estimator is the infeasible Robinson estimator from section 2. Hence, we did not estimate γ and α in the first stage but assumed their true values to be known instead. We contrast the estimation results from our proposed estimator to those obtained by the estimator suggested by Wooldridge (2010) (or Semykina and Wooldridge, 2010, respectively). As mentioned in the introduction, Wooldridge's strategy is to augment the main equation with a selection correction term and then to apply 2SLS to this augmented equation (using all exogenous variables and the inverse Mills ratio term as instruments). As the error terms of main and selection equation are normally distributed, we augmented the main equation with the inverse Mills ratio term. We assumed that γ is known in advance, and so the inverse Mills ratio term is known. However, since the instruments for y_2 suggested by Wooldridge also contain the inverse Mills ratio term (whose influence on y_2 is not known in advance), we *estimated* the first stage for y_2 instead of assuming the true values of these coefficients were known in advance. Hence, the Wooldridge estimator may not be fully comparable to our proposed estimator as it depends on estimated values. With these caveats in mind, we nevertheless see that both estimators perform well for all sample sizes. The RMSE of the Wooldridge estimator is slightly larger which may be due to the estimation of the first stage. However, the empirical sizes of the t tests are close to their nominal sizes, which suggests that tests based on asymptotic distribution theory are valid for both estimators.

As noted in the introduction, if the selection equation also contains y_2, the Wooldridge estimator is biased. The reason is that the inverse Mills ratio term includes y_2 and is, thus, endogenous as well, even if the parameters of the selection equation were known in advance (or have been estimated consistently). On the contrary, our proposed estimator can deal with endogeneity in the selection equation as well. To illustrate the different behavior of both estimators in the presence of endogeneity in the selection equation, we replaced equation (21) above with equation (28) and obtained Monte Carlo results for this slightly modified model. Table 3 gives the results. As we can see, our proposed estimator performs well, while the Wooldridge estimator is biased in β_2. However, the Wooldridge estimator yields unbiased estimates of β_1. This is due to the fact that correct estimation of β_1 depends on a correct specification of the selectivity correction term. But as the selection equation parameters were known in advance, the selection effect is correctly accounted for by the inverse Mills ratio term. Hence, the endogeneity of the selectivity correction term affects only estimation of the parameters belonging to the endogenous explanatory variables, but not estimation of the remaining parameters (provided that the selectivity correction term has been correctly specified).

The next issue to be considered is the effect of estimating the parameters of the selection equation and the reduced form equation for y_2 in a "first stage", rather than assuming these values to be known in advance. Given our assumptions on the reduced form equation for y_2, especially its linearity, the natural semiparametric first stage estimator for this equation is OLS. However, concerning the selection equation several semiparametric estimators have been proposed (Manski, 1975; Han, 1987; Ichimura, 1993; Horowitz, 1992; Klein and Spady, 1993). We consider two of them, the Klein and Spady (1993) estimator and the smoothed maximum score estimator due to Horowitz (1992). The Klein and Spady (1993) estimator estimates not only the parameters in the selection equation, but the c.d.f. of u as well, and it attains the semiparametric efficiency bound. Moreover, it is $\sqrt{N}$ consistent under appropriate regularity conditions. The Horowitz (1992) estimator is a smoothed version of Manksi's (1975) maximum score estimator with a rate of convergence which is slower than $\sqrt{N}$ but which can be made arbitrarily close to $\sqrt{N}$. Like Manski's (1975) estimator, this estimator is robust to heteroskedasticity of an unknown form. However, in contrast to Manski's estimator the smoothed maximum score estimator has a faster rate of convergence.

In addition to equation (24), we consider three further distributions of u or ν_u, respectively, in the selection equation in order to show how these estimators perform under different distributions. These distributions are very close to those used in Horowitz (1992).

(i) $\nu_u \sim$ uniform with zero mean and unit variance

(ii) $\nu_u \sim$ Student's t with 3 degrees of freedom and normalized to have unit variance

(iii) $\nu_u = 0.25(1 + 2a^2 + a^4)\nu$, where $a = x + w$ and $\nu \sim \mathcal{N}(0, 1)$

(In the case of endogeneity in the selection equation, a in (iii) is replaced by $a = y_2 + w$.)

As both estimators require the specification of kernel-type functions and bandwidths, we made the following choices: For the Klein and Spady (1993) estimator, we selected the standard normal p.d.f. and a bandwidth of $h = n^{-1/6.5}$, while for the Horowitz (1992) estimator we selected the standard normal c.d.f. and a bandwidth of $h = n^{-1/6.5}$. Furthermore, both estimators require a normalization as the parameters of a binary choice model are only identified up to scale. We chose that the coefficient of x in the selection equation be equal to unity (which is indeed its true value). Since the objective function for obtaining the smoothed maximum score estimates is difficult to maximize numerically, we performed a grid search over the interval [-1,3] with a step length of 0.005. Note that both the Klein and Spady and the Horowitz estimator are robust to the kind of conditional heteroskedasticity in distribution (iii).

We not only considered these two semiparametric estimators for estimating the selection equation parameters, but we also analyzed the performance of the OLS estimator (which, in this context, is known as the linear probability model (LPM)) and the two well-known parametric estimators for binary choice models, logit and probit. The idea why we consider these alternative estimators is that in large samples the Klein and Spady estimator requires long computation times, which limits its applicability especially in the case of bootstrap-based inference. On the other hand, the smoothed maximum score estimator requires sophisticated optimization routines in the case of many explanatory variables, as there is the possibility of finding a local rather than a global maximum when maximizing the objective function. For these practical reasons, it may be convenient to stick to a conventional model such as the LPM, logit or probit to estimate the selection equation. We seek to investigate how this possibly wrong model specification for the selection equation affects the outcomes of our feasible Robinson estimator.

The Monte Carlo results for the feasible Robinson estimator based on our five selection equation estimators are presented in table 4. Note that these results are also based on preliminary (OLS) estimates of the reduced form equation for y_2. As we can see, the Robinson estimator based on the Klein and Spady estimates works very well even in small sample sizes and for all considered distributions of u, with acceptable empirical sizes of t tests. The Robinson estimator based on the smoothed maximum score estimates performs similarly with respect to β_2, but exhibits some (small) bias regarding β_1. Since the correct estimation of β_1 relies heavily on a correct estimation of the selection equation (as these estimates are used to eliminate the selection effect which directly affects estimation of β_1 but not β_2), we may conclude that the smoothed maximum score estimator performs not sufficiently well in small samples, a conclusion which has also been raised by Kyriazidou (1997) and attributed to the relatively large finite sample bias of this estimator. Hence, one may conclude that the Klein and Spady estimator should be preferred in applications, unless one suspects that the selection equation is contaminated by a significant amount of conditional heteroskedasticty of unknown form.

The OLS estimator is also a semiparametric estimator. Its use for binary dependent variables has often been criticized because one may obtain predicted probabilities which lie outside of the $[0, 1]$ interval. However, since the OLS estimator only serves for estimating the parameters in the selection equation as a first step estimator, this argument loses some of its validity. As we can see from table 4, the Robinson estimator based on the LPM estimates performs well except for the case were u is heteroskedastic. Like the logit and the probit model, the LPM is not robust against conditional heteroskedasticity. However, for the remaining distributions we obtain sensible results even for small sample sizes, which suggests that the LPM can well be used as a first stage estimator for the selection equation.

Regarding the Robinson estimator based on probit and logit estimates of the selection equation parameters, we see that the estimates are very similar to the estimates based on the LPM, even if the distribution of u does not fit the model assumptions (this latter point has also been recognized by Horowitz (1992)). This suggests that, as long as u is not conditionally heteroskedastic, one can obtain good estimation results for the main equation parameters by applying either the LPM, logit or probit to the selection equation. This is especially important for large sample sizes combined with bootstrap-based inference, as the computational burden is considerably lower if one uses one of these three estimators to estimate the selection equation (as opposed to using the Klein and Spady or smoothed maximum score estimators).

Finally, we considered the performance of the feasible Robinson estimator when there is endogeneity in the selection equation. For simplicity, we estimated the selection equation with 2SLS and did not consider extended versions of, e.g., the Klein and Spady estimator which also account for endogeneity (see Blundell and Powell, 2004; Rothe, 2009). Table 5 contains the results. As we can see, the Robinson estimator yields good results for β_1 which are similar to using the LPM in the first stage. For β_2 there is some (small) bias which may be due to the fact that estimation of β_2 is affected by endogeneity *and* selection effects, which in turn requires a larger sample size to obtain "good" results.

5 Empirical Application

In this section, we present an economic application of our estimator. We seek to study the effect of the wage rate on the number of children of married women who work full time. From an economic perspective, the higher the (potential) wage of a woman, the higher are the opportunity costs of having children and, thus, the lower the probability of having *many* children. We thus expect a negative impact of the wage variable in a regression of the number of children on the wage rate (and other covariates).

We restrict our analysis to full time working women because these might differ from part time working women in some respects (and, of course, women who are not working). For instance, part time working women may have a higher preference for having children.

Restricting our sample to a specific group of women may cause a sample selection bias. Moreover, the women's wage may not be exogenous since wages as well as the number of children are determined by preferences and, thus, past human capital investments of women. In order to account for this potential endogeneity, we use industry dummies as instrumental variables for a woman's wage. We argue that industry choices are unrelated to unobserved factors (such as preferences) affecting the number of children.

In our empirical specification, the main equation has the number of children (*nchild*) as its dependent variable. Explanatory variables are the (log) annual wage (*lwage*), educational attainment (in years, *educ*), the (log) husband's wage (*lhuswage*), age (*age*) and age squared (*age2*). The selection equation (governing the decision to work full time, *ftwork*) includes the metropolitan status (*metropolitan*, =1 if woman lives in metropolitan area), educational attainment, the (log) husband's wage, age and age squared.

The data are from the American Community Survey 2010 sample.[1] We restrict our sample to white married women who were born in the U.S. and between 25 and 40 years of age. Descriptive statistics of our variables are given in table 6 (except for the industry dummies). We have a total of 71,730 women of whom 44,639 are working full time (that is, equal to or more than 36 hours per week), which corresponds to a fraction of women working full time of 62.2 percent.

In order to implement our feasible Robinson estimator, we have to obtain "first stage" estimates of γ and v. The natural candidate for obtaining consistent estimates of v or α, respectively, is OLS, as pointed out in the last section. Consequently, we regressed the log wage on our instrumental variables and the remaining exogenous variables. We performed this regression for the "selected" subsample only as the reduced form equation for the log wage may be contaminated by selection effects. The estimation results are presented in table 7. Note that the t statistics of our instruments are relatively large,

[1] We obtained our data files from the IPUMS-USA database (Ruggles et al., 2010).

thus indicating that our instrumental variables have a sufficiently high impact on our endogenous explanatory variable. The F-statistic of joint significance of our instrumental variables also supports this view with a value of 280.23.

Regarding estimation of the selection equation, we refer to the conclusions raised in the last section. Since the sample size is relatively large, we selected the LPM, the logit and the probit model. Estimation results for the selection equation based on these three models are given in table 8. Note that the variable *metropolitan* which is exclusive in the selection equation has a significant impact on the probability of working full time. Educational attainment and the husband's wage also have strong impacts (with plausible signs), while age and age squared seem to be irrelevant. As usual, the coefficient estimates of these models are not directly comparable, as the estimates reflect the ratio of a coefficient and the square root of a variance parameter. Instead, we can compare coefficient ratios across the three models. The ratio of the coefficient of *metropolitan* and the coefficient of *educ*, for example, is -0.383 for logit, -0.399 for probit and -0.402 for the LPM. The remaining coefficient ratios are also similar across the three models, which indicates that estimation results are not crucially affected by the choice of the model.

After having obtained first stage estimates of γ and v, we can employ the feasible Robinson estimator to get estimates of the parameters of the main equation. Since we estimated v only for the selected sample, however, we cannot rely on Theorem 1 from section 3 in order to obtain standard errors. As raised above, we use bootstrapping instead. We did this by sampling with replacement from the original sample for a total of 100 bootstrap replications. For each replication, we computed the feasible Robinson estimator and computed the variance of our estimated coefficients according to the formula in equation (19). Table 9 gives the results of the feasible Robinson estimator in dependency of the three models which were used to estimate the selection equation. The coefficient of *lwage* is around -0.25, meaning that a doubling of wages (increase by 100 percent) decreases the number of children by 0.25 on average. All estimated coefficients are relatively similar across the three models used to estimate the selection equation (maybe except for *lhuswage*), which is in accordance with the results from the last section.

To gain further insights into our data set and for a comparison of our proposed estimator with other estimators, we estimated the wage equation with OLS, 2SLS and the maximum likelihood estimator of the Heckman sample selection model. In the last section, we labeled these estimators the "naive" estimators as neither controls for the joint presence of sample selectivity and endogeneity. When applying OLS, we see from table 9 that the coefficient of *lwage* is smaller (in absolute terms) than the estimate from the Robinson estimator and that the husband's wage is insignificant. In order to control for endogeneity of the (log) wage, we computed the IV estimator using the same instrumental variables as for the Robinson estimator. From table 9 we see that the coefficient of *lwage* increases dramatically (in absolute terms and compared to OLS), while the effects of *age* and *age2* are similar. However, the coefficients of *educ* and *lhuswage* change considerably. Finally we employed the Heckman estimator in order to control for sample selection bias. Recall that this estimator is based on a joint normality assumption concerning the error terms in main and selection equation. From table 9 we see that the estimates are very close to the OLS estimates. Nevertheless, the estimated correlation coefficient between the error terms of main and selection equation is about -0.106 and significantly different from zero at the 1%-level, so that we have to reject the hypothesis that there is no sample selection bias.

The three Robinson estimators support the evidence from OLS, IV and the Heckman model. The IV results suggest the presence of a remarkable endogeneity bias in the log wage. In this sense, the Robinson estimates point into the right direction as their absolute values are larger than those of OLS. Put differently, the Robinson estimators yield plausible results. However, the discrepancy between the Robinson results and the IV results suggests that selectivity bias is a concern as well. This confirms the Monte Carlo results from the last section that it is not sufficient to control for endogeneity alone.

The coefficients of *educ* and *lhuswage* are very different across the estimators listed in table 9 and are sometimes insignificant, whereas they are very large in case of the Robinson estimator. To get some further insights, we estimated the coefficients again for different choices of the bandwidth parameter h. In addition to our initial bandwidth choice of $h = n^{-1/6.5}$, we selected bandwidths of $h = n^{-1/6}$, $h = n^{-1/7}$ and $h = n^{-1/8}$. Results (based on a probit estimation of the selection equation) are given in table 10. While the estimated coefficients of *lwage*, *age* and *age2* are very close over the bandwidths, the coefficients of *educ* and *lhuswage* exhibit a lot of variation. Hence, these coefficient estimates are rather unstable and not robust against variations of the bandwidth parameter. Consequently, interpretation of the effects of educational attainment and the husband's wage on the number of children should be done with caution. The remaining variables, however, possess numerically robust coefficients across the three estimators, so that one can have some confidence that these estimates measure the respective effects correctly.

6 Extension to Quantile Regression Settings

In this section, we briefly consider an extension of our model to quantile regression settings and provide guidelines for estimation. Let $Q_\tau(y|x)$, $0 < \tau < 1$, be the τth conditional quantile of y given x. Then, combining the approaches in Buchinsky (1998) and Lee (2007), we may write

$$Q_\tau(y_i|d_i = 1, w_i, x_i, z_i, v_i) = q_i'\theta + Q_\tau(\varepsilon_i|d_i = 1, w_i, x_i, z_i, v_i) \tag{29}$$

$$= x_i'\beta + h(w_i'\gamma, v_i). \tag{30}$$

Hence, we again employ a control function approach where the expectation operator from Assumption 1 is replaced by a quantile operator. As pointed out in Huber and Melly (2011), a setup of the model as in equations (29) and (30) requires conditional independence of ε and x, so that any quantile regression yields the same slope coefficients. Hence, one of the benefits of quantile regression, namely that coefficients vary over the distribution, must be sacrificed. However, quantile regression may nevertheless be useful as it is more robust than mean regression. Huber and Melly (2011) also provide a test on the validity of the conditional independence assumption.

Estimation can be carried out as in Lee (2007). In particular, the unknown control function $h(w'\gamma, v)$ can be approximated by a series expansion, where one may use power series or splines, for instance. Let $P_K(q, w'\gamma, v) = (q, p_1(w'\gamma, v), \ldots, p_K(w'\gamma, v))$, where $p_k(\cdot), k = 1, \ldots, K$, denote known basis functions and $K \to \infty$ as $n \to \infty$. Furthermore, let $\zeta = (\theta', a_1, \ldots, a_K)'$, where $a_1, \ldots, a_K$ are the series coefficients which must be estimated along with the parameters of interest. Then, the estimator of ζ is given by

$$\hat{\zeta} = \arg\min_{\zeta} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - P_k(q, \hat{w}'\gamma, \hat{v})\zeta), \tag{31}$$

where $\rho_\tau(\cdot)$ is the "check function" defined as $\rho_\tau(u) = |u| + (2\tau - 1)u$ for $0 < \tau < 1$.

Lee (2007) shows that this estimator is $\sqrt{n}$ consistent under appropriate regularity conditions. Note that, once again, first stage estimates of $\hat{\gamma}$ and $\hat{v}$ enter the objective function. If these first stage estimators converge at a faster rate than $\sqrt{n}$, we can repeat the statement from section 3 that, in this case, it does not matter asymptotically whether $\hat{\gamma}$ and $\hat{v}$ are estimated or one uses the true values.

7 Conclusion

In this paper, we derived a semiparametric estimation procedure for the sample selection model which also controls for endogeneity of covariates. We presented some Monte Carlo results and demonstrated that our proposed estimator performs well in finite samples. In contrast to existing approaches raised in the literature, our approach is able to handle situations in which the same endogenous covariates enter the main as well as the selection equation.

We also extended our model to quantile regression settings. Quantile regression methods have become a popular tool in applied econometrics as they allow to obtain heterogeneous effects over the entire distribution of the dependent variable. As noted above, however, this feature is not possible in the sample selectivity and endogeneity case, at least if the model is specified as in section 6. However, quantile regression may nevertheless be useful due to its robustness properties.

Since its popularization by Heckman in 1979, many researcher have used the sample selection model to control for sample selection bias. On the other hand, the exogeneity assumption on the covariates has only seldom been challenged. But, and this is the crucial result of this paper, if sample selectivity and endogeneity of covariates are jointly present, econometric models *should* account for both types of specification error jointly if one wishes to obtain consistent estimates of the parameters of interest.

References

Ahn, H. and Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1-2):3–29.

Billingsley, P. (1995). *Probability and Measure*. Wiley, New York, NY, 3rd edition.

Blundell, R. W. and Powell, J. L. (2004). Endogeneity in semiparametric binary response models. *Review of Economic Studies*, 71:655–679.

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the usa: a quantile regression approach. *Journal of Applied Econometrics*, 13(1):1–30.

Chib, S., Greenberg, E., and Jeliazkov, I. (2009). Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics*, 18(2):321–348.

Das, M., Newey, W. K., and Vella, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies*, 70(1):33–58.

Han, A. K. (1987). A non-parametric analysis of transformations. *Journal of Econometrics*, 35(2-3):191–209.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–61.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60(3):505–31.

Huber, M. and Melly, B. (2011). Quantile regression in the presence of sample selection. Discussion Paper no. 2011-09, University of St. Gallen.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 61(2):387–421.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica*, 65(6):1335–1364.

Lee, S. (2007). Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics*, 141(2):1131 – 1158.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice.* Princeton Univ. Press, Princeton, NJ.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3(3):205–228.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *Econometrics Journal*, 12:S217–S229.

Newey, W. K., Powell, J. L., and Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603.

Powell, J. L. (1987). Semiparametric estimation of bivariate limited dependent variable models. Manuscript, University of California, Berkeley.

Robinson, P. M. (1988). Root- n-consistent semiparametric regression. *Econometrica*, 56(4):931–54.

Rothe, C. (2009). Semiparametric estimation of binary response models with endogenous regressors. *Journal of Econometrics*, 153(1):51–64.

Ruggles, S., Alexander, J. T., Genadek, K., Goeken, R., Schroeder, M. B., and Sobek, M. (2010). Integrated public use microdata series: Version 5.0 [machine-readable database]. Minneapolis: University of Minnesota.

Semykina, A. and Wooldridge, J. M. (2010). Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics*, 157(2):375–380.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* The MIT Press, Cambridge, MA, 2nd edition.

Appendix

Proof of theorems 1 and 2.

The Robinson estimator has the following linear representation:

$$\hat{\hat{\theta}} = \theta + \left(\sum_{i=1}^{n} (q_i - \hat{\hat{q}}_i)(q_i - \hat{\hat{q}}_i)' \right)^{-1} \sum_{i=1}^{n} \{ (q_i - \hat{\hat{q}}_i)(g_i - \hat{\hat{g}}_i + r_i - \hat{\hat{r}}_i) \}. \tag{32}$$

Let

$$q_i^* = E[q_i | w_i'\gamma, v_i], \quad g_i^* = E[g_i | w_i'\gamma, v_i], \quad r_i^* = E[r_i | w_i'\gamma, v_i] \tag{33}$$

and

$$\tilde{q}_i = E[q_i | w_i'\hat{\gamma}, \hat{v}_i], \quad \tilde{g}_i = E[g_i | w_i'\hat{\gamma}, \hat{v}_i], \quad \tilde{r}_i = E[r_i | w_i'\hat{\gamma}, \hat{v}_i]. \tag{34}$$

Then, eq. (1) can be augmented as

$$\begin{aligned} \hat{\theta} = \theta + & \left(\sum_{i=1}^{n} (q_i - \hat{\hat{q}}_i)(q_i - \hat{\hat{q}}_i)' \right)^{-1} \sum_{i=1}^{n} \{ (q_i - q_i^* + q_i^* - \tilde{q}_i + \tilde{q}_i - \hat{\hat{q}}_i) \\ & \times (g_i - g_i^* + g_i^* - \tilde{g}_i + \tilde{g}_i - \hat{\hat{g}}_i + r_i - r_i^* + r_i^* - \tilde{r}_i + \tilde{r}_i - \hat{\hat{r}}_i) \} \qquad (35) \\ & = \left(\sum_{i=1}^{n} (q_i - \hat{\hat{q}}_i)(q_i - \hat{\hat{q}}_i)' \right)^{-1} \sum_{i=1}^{n} \{ (q_i - q_i^* + q_i^* - \tilde{q}_i + \tilde{q}_i - \hat{\hat{q}}_i) \\ & \times (g_i^* - \tilde{g}_i + \tilde{g}_i - \hat{\hat{g}}_i + r_i - \hat{\hat{r}}_i) \} \qquad (36) \end{aligned}$$

Under regularity conditions, the analysis of Robinson (1988) implies that

$$\begin{aligned} \hat{\theta} = \theta + \left(\sum_{i=1}^{n} (q_i - \hat{\hat{q}}_i)(q_i - \hat{\hat{q}}_i)' \right)^{-1} & \sum_{i=1}^{n} \{ (q_i - q_i^* + q_i^* - \tilde{q}_i) \\ & \times (g_i^* - \tilde{g}_i + r_i) \} + o_p(n^{-1/2}). \end{aligned} \tag{37}$$

Hence,

$$\sqrt{n}(\hat{\theta}-\theta) = \left(\frac{1}{n}\sum_{i=1}^{n}(q_i-\hat{q}_i)(q_i-\hat{q}_i)'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{(q_i-q_i^*+q_i^*-\tilde{q}_i)(g_i^*-\tilde{g}_i+r_i)\}\right)+o_p(1) \tag{38}$$

$$= \hat{C}^{-1}(A_1+A_2+A_3+A_4)+o_p(1), \tag{39}$$

where

$$A_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(q_i-q_i^*)r_i \tag{40}$$

$$A_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(q_i-q_i^*)(g_i^*-\tilde{g}_i) \tag{41}$$

$$A_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(q_i^*-\tilde{q}_i)r_i \tag{42}$$

$$A_4 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(q_i^*-\tilde{q}_i)(g_i^*-\tilde{g}_i) \tag{43}$$

Since $E[r|q, w'\gamma, v]=0$, it follows from the central limit theorem that $A_1 \xrightarrow{d} \mathcal{N}(0,\Psi)$.

For A_2, a Taylor series expansion yields

$$A_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(q_i-q_i^*)\left(\frac{\partial\tilde{g}_i(w_i'\gamma,v_i)}{\partial\hat{v}_i}z_i'(\hat{\alpha}-\alpha)\right)+o_p(1) \tag{44}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n}(q_i-q_i^*)\left(\frac{\partial\tilde{g}_i(w_i'\gamma,v_i)}{\partial\hat{v}_i}\right)z_i'\right)\sqrt{n}(\hat{\alpha}-\alpha)+o_p(1) \tag{45}$$

Note that

$$\frac{1}{n}\sum_{i=1}^{n}(q_i-q_i^*)\left(\frac{\partial\tilde{g}_i(w_i'\gamma,v_i)}{\partial\hat{v}_i}\right)z_i' \xrightarrow{p} E\left[(q_i-q_i^*)\left(\frac{\partial g_i(w_i'\gamma,v_i)}{\partial v_i}\right)z_i'\right]. \tag{46}$$

Hence, if $\sqrt{n}(\hat{\alpha}-\alpha) \xrightarrow{d} \mathcal{N}(0,\Gamma)$, then $A_2 \xrightarrow{d} \mathcal{N}(0,\Omega)$, where

$$\Omega = \left(E\left[(q_i-q_i^*)\left(\frac{\partial g_i(w_i'\gamma,v_i)}{\partial v_i}\right)z_i'\right]\right)\Gamma\left(E\left[(q_i-q_i^*)\left(\frac{\partial g_i(w_i'\gamma,v_i)}{\partial v_i}\right)z_i'\right]\right)'. \tag{47}$$

On the contrary, if $\sqrt{n}(\hat{\alpha} - \alpha) = o_p(1)$, then $A_2 = o_p(1)$.

Regarding A_3, we have that

$$A_3 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} r_i \left(\frac{\partial \tilde{g}_i(w_i'\gamma, v_i)}{\partial \hat{v}_i} z_i'(\hat{\alpha} - \alpha) \right) + o_p(1) \tag{48}$$

$$= \left(\frac{1}{n} \sum_{i=1}^{n} r_i \left(\frac{\partial \tilde{g}_i(w_i'\gamma, v_i)}{\partial \hat{v}_i} \right) z_i' \right) \sqrt{n}(\hat{\alpha} - \alpha) + o_p(1) \tag{49}$$

Note that

$$\frac{1}{n} \sum_{i=1}^{n} r_i \left(\frac{\partial \tilde{g}_i(w_i'\gamma, v_i)}{\partial \hat{v}_i} \right) z_i' \xrightarrow{p} E\left[r_i \left(\frac{\partial g_i(w_i'\gamma, v_i)}{\partial v_i} \right) z_i' \right] = 0 \tag{50}$$

since r is uncorrelated with the exogenous variables. Hence, $A_3 = o_p(1)$.

For A_4, we have that

$$A_4 \leq \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} ||(q_i^* - \tilde{q}_i)|| \, |(g_i^* - \tilde{g}_i)| \leq \sqrt{n} ||\hat{\xi} - \xi||^2 = o_p(1). \tag{51}$$

By a law of large numbers argument and the dominated convergence theorem, it follows that $\hat{C} \xrightarrow{p} \Phi$, where $\Phi = E[(q_i - q_i^*)(q_i - q_i^*)']$. Therefore, if $\sqrt{n}(\hat{\alpha} - \alpha) = o_p(1)$, we obtain that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Phi^{-1}\Psi\Phi^{-1}); \tag{52}$$

if $\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, \Gamma)$, we have

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Phi^{-1}(\Psi + \Omega)\Phi^{-1}). \tag{53}$$

□

Table 1: Naive estimators

		OLS			IV			Heckman		
		Mean	RMSE	Size	Mean	RMSE	Size	Mean	RMSE	Size
N=250	$\beta_1 = 1$	0.728	0.291	0.742	0.948	0.130	0.073	0.917	0.144	0.103
	$\beta_2 = 1$	1.221	0.226	0.995	0.947	0.093	0.089	1.274	0.278	1.000
N=500	$\beta_1 = 1$	0.722	0.287	0.968	0.941	0.101	0.098	0.911	0.119	0.185
	$\beta_2 = 1$	1.221	0.224	1.000	0.945	0.076	0.159	1.274	0.276	1.000
N=1000	$\beta_1 = 1$	0.723	0.281	0.999	0.941	0.082	0.165	0.912	0.105	0.319
	$\beta_2 = 1$	1.221	0.223	1.000	0.947	0.064	0.268	1.274	0.275	1.000

Table 2: Robinson vs. Wooldridge

		Robinson			Wooldridge		
		Mean	RMSE	Size	Mean	RMSE	Size
N=250	$\beta_1 = 1$	1.000	0.120	0.044	1.009	0.128	0.050
	$\beta_2 = 1$	1.025	0.082	0.058	0.996	0.092	0.043
N=500	$\beta_1 = 1$	1.000	0.081	0.045	1.001	0.087	0.044
	$\beta_2 = 1$	1.015	0.057	0.056	0.995	0.067	0.056
N=1000	$\beta_1 = 1$	0.999	0.059	0.055	1.001	0.061	0.048
	$\beta_2 = 1$	1.010	0.041	0.047	0.999	0.046	0.043

Table 3: Robinson vs. Wooldridge - endogeneity in the selection equation

		Robinson			Wooldridge		
		Mean	RMSE	Size	Mean	RMSE	Size
N=250	$\beta_1 = 1$	0.992	0.113	0.046	1.010	0.138	0.045
	$\beta_2 = 1$	1.004	0.143	0.028	0.883	0.189	0.049
N=500	$\beta_1 = 1$	0.998	0.077	0.043	1.006	0.094	0.038
	$\beta_2 = 1$	0.995	0.107	0.032	0.881	0.160	0.149
N=1000	$\beta_1 = 1$	1.000	0.056	0.051	1.003	0.067	0.039
	$\beta_2 = 1$	0.991	0.080	0.040	0.888	0.133	0.292

Table 4: Estimator performance for different selection equation estimators

			Klein & Spady			Smoothed Max. Score			LPM			Probit			Logit		
			Mean	RMSE	Size	Mean	RMSE	Size	Mean	RMSE	Size	Mean	RMSE	Size	Mean	RMSE	Size
N=250	normal	$\beta_1 = 1$	1.005	0.142	0.086	0.987	0.134	0.086	1.001	0.139	0.084	1.001	0.139	0.090	1.001	0.139	0.092
		$\beta_2 = 1$	1.022	0.084	0.071	1.022	0.085	0.079	1.022	0.084	0.069	1.022	0.084	0.070	1.022	0.084	0.071
	uniform	$\beta_1 = 1$	0.995	0.138	0.094	0.983	0.135	0.111	0.993	0.138	0.091	0.993	0.137	0.090	0.993	0.137	0.092
		$\beta_2 = 1$	1.023	0.092	0.109	1.023	0.092	0.105	1.023	0.093	0.107	1.023	0.093	0.109	1.023	0.093	0.109
	t	$\beta_1 = 1$	0.996	0.144	0.086	0.978	0.138	0.116	0.989	0.140	0.094	0.989	0.140	0.094	0.989	0.140	0.096
		$\beta_2 = 1$	1.021	0.089	0.095	1.022	0.090	0.088	1.022	0.090	0.095	1.022	0.090	0.094	1.022	0.090	0.093
	het	$\beta_1 = 1$	0.998	0.198	0.079	0.976	0.160	0.090	0.904	0.261	0.092	0.901	0.261	0.093	0.902	0.259	0.093
		$\beta_2 = 1$	1.025	0.096	0.068	1.028	0.098	0.074	1.023	0.093	0.071	1.023	0.094	0.075	1.023	0.094	0.075
N=500	normal	$\beta_1 = 1$	0.999	0.096	0.079	0.988	0.092	0.084	0.998	0.094	0.076	0.997	0.094	0.079	0.998	0.094	0.080
		$\beta_2 = 1$	1.012	0.062	0.085	1.013	0.063	0.083	1.012	0.062	0.082	1.012	0.062	0.083	1.012	0.062	0.085
	uniform	$\beta_1 = 1$	0.998	0.097	0.081	0.988	0.094	0.097	0.996	0.095	0.089	0.996	0.096	0.086	0.997	0.096	0.086
		$\beta_2 = 1$	1.019	0.063	0.083	1.019	0.064	0.082	1.019	0.063	0.084	1.019	0.063	0.083	1.019	0.063	0.083
	t	$\beta_1 = 1$	1.003	0.100	0.089	0.987	0.094	0.087	0.999	0.097	0.076	1.000	0.099	0.088	1.000	0.099	0.091
		$\beta_2 = 1$	1.011	0.062	0.084	1.012	0.063	0.090	1.012	0.063	0.091	1.012	0.063	0.085	1.012	0.063	0.086
	het	$\beta_1 = 1$	1.007	0.126	0.078	0.987	0.110	0.082	0.922	0.220	0.115	0.924	0.223	0.110	0.925	0.223	0.111
		$\beta_2 = 1$	1.016	0.071	0.069	1.018	0.072	0.073	1.012	0.070	0.076	1.012	0.070	0.080	1.012	0.070	0.081
N=1000	normal	$\beta_1 = 1$	1.000	0.065	0.073	0.992	0.065	0.084	0.999	0.065	0.073	0.999	0.065	0.072	0.999	0.065	0.071
		$\beta_2 = 1$	1.009	0.043	0.072	1.010	0.044	0.073	1.009	0.043	0.072	1.009	0.043	0.067	1.009	0.043	0.067
	uniform	$\beta_1 = 1$	0.997	0.071	0.098	0.989	0.070	0.111	0.997	0.070	0.099	0.997	0.070	0.094	0.997	0.070	0.095
		$\beta_2 = 1$	1.010	0.047	0.091	1.011	0.047	0.097	1.010	0.047	0.096	1.010	0.047	0.092	1.010	0.047	0.093
	t	$\beta_1 = 1$	1.000	0.071	0.098	0.988	0.070	0.121	0.998	0.072	0.105	0.998	0.070	0.097	0.998	0.070	0.097
		$\beta_2 = 1$	1.009	0.045	0.083	1.010	0.046	0.086	1.010	0.046	0.086	1.010	0.045	0.082	1.010	0.045	0.082
	het	$\beta_1 = 1$	1.001	0.086	0.096	0.984	0.079	0.100	0.935	0.189	0.172	0.936	0.186	0.163	0.937	0.185	0.164
		$\beta_2 = 1$	1.010	0.050	0.054	1.010	0.050	0.059	1.007	0.049	0.059	1.007	0.049	0.060	1.007	0.049	0.059

Table 5: Estimator performance in case of endogeneity in the selection equation

			2SLS		
			Mean	RMSE	Size
N=250	normal	$\beta_1 = 1$	0.996	0.122	0.072
		$\beta_2 = 1$	0.986	0.136	0.028
	uniform	$\beta_1 = 1$	0.991	0.126	0.083
		$\beta_2 = 1$	0.985	0.153	0.063
	t	$\beta_1 = 1$	0.988	0.127	0.093
		$\beta_2 = 1$	0.982	0.151	0.046
	het	$\beta_1 = 1$	0.963	0.160	0.083
		$\beta_2 = 1$	0.973	0.133	0.070
N=500	normal	$\beta_1 = 1$	0.999	0.085	0.062
		$\beta_2 = 1$	0.980	0.113	0.038
	uniform	$\beta_1 = 1$	0.997	0.084	0.064
		$\beta_2 = 1$	0.983	0.111	0.042
	t	$\beta_1 = 1$	1.002	0.088	0.076
		$\beta_2 = 1$	0.976	0.116	0.061
	het	$\beta_1 = 1$	0.980	0.116	0.084
		$\beta_2 = 1$	0.963	0.105	0.077
N=1000	normal	$\beta_1 = 1$	1.002	0.061	0.068
		$\beta_2 = 1$	0.982	0.084	0.056
	uniform	$\beta_1 = 1$	0.998	0.065	0.085
		$\beta_2 = 1$	0.985	0.089	0.056
	t	$\beta_1 = 1$	1.001	0.065	0.084
		$\beta_2 = 1$	0.982	0.090	0.068
	het	$\beta_1 = 1$	0.979	0.083	0.080
		$\beta_2 = 1$	0.962	0.087	0.102

Table 6: Descriptive statistics

Variable	Mean	Std. Dev.
nchild (ftwork=1)	1.317	1.126
lwage (ftwork=1)	10.572	0.635
educ	14.498	2.036
lhuswage	10.700	0.825
age	33.195	4.482
ftwork	0.622	0.485
metropolitan	0.765	0.424
no. obs.		71,730
no. obs. with ftwork=1		44,639

Table 7: Selection equation estimates

Variable	Logit	Probit	LPM
metropolitan	-0.0745	-0.0475	-0.0178
	(0.0189)	(0.0116)	(0.0043)
educ	0.1948	0.1192	0.0442
	(0.0041)	(0.0025)	(0.0009)
lhuswage	-0.2894	-0.1709	-0.0617
	(0.0108)	(0.0063)	(0.0024)
age	0.0104	0.0055	0.0020
	(0.0283)	(0.0174)	(0.0064)
age2	-0.0001	0.0000	0.0000
	(0.0004)	(0.0003)	(0.0001)
cons	0.5996	0.3251	0.6072
	(0.4645)	(0.2847)	(0.1058)

Note: Standard errors in parentheses.

Table 8: Reduced form estimates

Variable	Coeff.	S.E.	t-Stat.
agriculture	0.2290189	0.0562054	4.07
mining	0.8409302	0.0840846	10
construction	0.5408305	0.0345118	15.67
manufacturing	0.7347648	0.0227991	32.23
wholesale	0.70541	0.0304954	23.13
retail	0.2717634	0.0212436	12.79
transport	0.5706506	0.0328031	17.4
utilities	0.8415711	0.0530471	15.86
communication	0.6425299	0.0295484	21.74
finance	0.7127673	0.0210127	33.92
management	0.6403309	0.0210235	30.46
social	0.3649633	0.0186949	19.52
arts	-0.0589073	0.0232935	-2.53
public	0.6772269	0.0235076	28.81
armed	0.9888614	0.068932	14.35
educ	0.1527763	0.0017988	84.93
lhuswage	0.0476687	0.0043233	11.03
age	0.0949767	0.0121166	7.84
age2	-0.0012707	0.0001833	-6.93
cons	5.371472	0.2001243	26.84

Table 9: Main equation estimates

Variable	OLS	IV	Heckman	Robinson: Logit	Robinson: Probit	Robinson: LPM
lwage	-0.1558	-0.8386	-0.1552	-0.2536	-0.2538	-0.2541
	(0.0088)	(0.0473)	(0.0088)	(0.0236)	(0.0237)	(0.0237)
educ	-0.0815	0.0069	-0.0888	-0.4909	-0.4823	-0.4796
	(0.0028)	(0.0067)	(0.0037)	(0.0404)	(0.0408)	(0.0403)
lhuswage	-0.0101	0.0698	-0.0002	0.6140	0.5798	0.5609
	(0.0065)	(0.0088)	(0.0072)	(0.0632)	(0.0612)	(0.0583)
age	0.5326	0.6067	0.5323	0.5197	0.5232	0.5234
	(0.0178)	(0.0196)	(0.0178)	(0.0622)	(0.0615)	(0.0608)
age2	-0.0068	-0.0077	-0.0068	-0.0068	-0.0068	-0.0068
	(0.0003)	(0.0003)	(0.0003)	(0.0009)	(0.0009)	(0.0009)
cons	-5.7703	-2.1403	-5.7006	-	-	-
	(0.2958)	(0.4000)	(0.2971)			

Note: Standard errors in parentheses.

Table 10: Robustness check: varying the bandwidth

Variable	$h = n^{-1/6.5}$	$h = n^{-1/6}$	$h = n^{-1/7}$	$h = n^{-1/8}$
lwage	-0.2538	-0.2552	-0.2516	-0.2452
educ	-0.4823	-0.5278	-0.4365	-0.3560
lhuswage	0.5798	0.6477	0.5112	0.3902
age	0.5232	0.5212	0.5251	0.5282
age2	-0.0068	-0.0068	-0.0068	-0.0068