# Semiparametric Estimation of a Binary Choice Model with Sample Selection

Jörg Schwiebert*

## Abstract

In this paper we provide semiparametric estimation strategies for a sample selection model with a binary dependent variable. To the best of our knowledge, this has not been done before. We propose a control function approach based on two different identifying assumptions. This gives rise to semiparametric estimators which are extensions of the Klein and Spady (1993), maximum score (Manski, 1975) and smoothed maximum score (Horowitz, 1992) estimators. We provide Monte Carlo evidence and an empirical example to study the finite sample properties of our estimators. Finally, we outline an extension of these estimators to the case of endogenous covariates.

**Keywords:** Sample selection model, binary dependent variable, semiparametric estimation, control function approach, endogenous covariates.
**JEL codes:** C21, C24, C25, 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

Since the seminal work of Heckman (1979), the sample selection model has become a standard tool in applied econometrics. Its objective is to obtain consistent estimates of the parameters of interest by removing a potential sample selection bias. In most cases, the sample selection model consists of a main equation with a continuous dependent variable (which is only partially observable) and a binary selection equation determining whether the dependent variable of the main equation is observed or not.

In this paper, we consider semiparametric estimation of a binary choice model with sample selection. That means, we do not assume a continuous dependent variable in the main equation but a binary one instead, taking only the values one or zero. Parametric estimation typically involves an assumption on the distribution of error terms (e.g., bivariate normal) and the setup of an appropriate likelihood function which is then maximized to obtain parameter estimates. However, as in the ordinary sample selection model originated by Heckman (1979), a parametric assumption on the joint distribution of error terms gives inconsistent parameter estimates if these assumptions are not fulfilled.

For the same reasons, several authors have analyzed semi-nonparametric methods to estimate the *ordinary* sample selection model which assumes a continuous dependent variable; examples include Gallant and Nychka (1987), Powell (1987), Ahn and Powell (1993), Das et al. (2003) and Newey (2009). However, to the best of our knowledge, there has not been developed a semiparametric estimation approach for a sample selection model with a binary dependent variable yet. This paper closes the gap.

In particular, we propose two different estimation strategies based on two distinct assumptions on the sample selection mechanism. Both strategies may be associated with what has been called the "control function approach". Our first estimation strategy is an extension of the Klein and Spady (1993) semiparametric estimation procedure for binary choice models. More specifically, our approach closely resembles the one of Rothe (2009), who extended the Klein and Spady estimator to a binary choice model with endogenous covariates. We can follow Rothe's approach since handling endogeneity and

sample selectivity is conceptually similar.

Our second estimation strategy is based on augmenting the main equation with a "control function" term which accounts for sample selectivity. This term is simply a generalization of the inverse Mills ratio term in the ordinary sample selection model. We will show how combining "similar" observations makes it possible to get rid of the unknown control function, so that the resulting model can be estimated by known techniques. In particular, we employ the maximum score estimator due to Manski (1975) and the smoothed maximum score estimator due to Horowitz (1992). This approach is conceptually similar to Powell (1987).

A sample selection models for a binary dependent variable was first considered by van de Ven and van Praag (1981). They simply augmented a probit model with an inverse Mills ratio term and estimated the model by maximum likelihood. The authors proposed to consider these probit estimates as approximative since the probit specification is inappropriate (as the error term after including the inverse Mills ratio term is not normally distributed even if the original error term is normally distributed). However, van de Ven and van Praag (1981) also provide the "true" likelihood function (based on a joint normality assumption).[1] The reason why the authors considered the approximative probit model with the inverse Mills ratio term included instead of the true likelihood function was due to the computational costs of maximizing the true likelihood function at that time.

The van de Ven and van Praag (1981) model has often been employed in empirical research. Van de Ven and van Praag (1981) used their model to analyze empirically the demand for deductibles in private health insurance. Further examples of application of the model include, for instance, Boyes et al. (1989), Greene (1992) and Mohanty (2002). While Boyes et al. (1989) and Greene (1992) used the model to analyze loan default probabilities, Mohanty (2002) employed the model to study teen employment differentials in Los Angeles county.

[1] Meng and Schmidt (1985) also analyzed this model and provided the likelihood function.

However, the van de Ven and van Praag (1981) model is parametric since it relies on a joint normality assumption on the error terms in the (latent) main equation and the selection equation. As raised above, parametric estimation leads to inconsistent parameter estimates if the parametric assumptions are not fulfilled.

We will investigate the consequences of estimating a misspecified parametric model in a small Monte Carlo study, in which we will also investigate the finite sample properties of our proposed semiparametric estimators. We also provide an empirical example in which we apply parametric and semiparametric estimators to study the determinants which lead women to work from home. In this example, we show how semmiparametric estimates may indicate that parametric estimates are subjected to misspecification.

The remainder of this paper is organized as follows. In section 2 we set up the econometric model. In section 3 we review parametric estimation of the model, and in section 4 we propose our semiparametric estimation strategies. In section 5, we conduct a small Monte Carlo study to compare the performance of the parametric and semiparametric estimators in small samples. Section 6 contains an empirical example where we apply our estimators to real data. In section 7, we extend our model to the case where explanatory variables are allowed to be endogenous. Finally, section 8 concludes the paper.

# 2 The Model

The model we consider is given by

$$y_i^* = x_i'\beta + \varepsilon_i \tag{1}$$

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

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

$$y_i = \begin{cases} 1(y_i^* > 0) \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases}, \tag{4}$$

where $i = 1, \ldots, N$ indexes individuals. The first equation is the main equation of interest, where $y^*$ is the latent dependent variable, $x$ is a vector of exogenous explanatory variables and $\varepsilon$ is an error term. 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 an error term. The third equation expresses that only the sign of $d^*$ is observable. By equation (4), the same is true for $y^*$, but only if $d$ is equal to one. Otherwise, $y^*$ cannot be observed ("missing"). This model differs from the ordinary sample selection model by the fact that the dependent variable of the outcome equation is binary, taking only the values one or zero.

Now we make three assumption which are assumed to hold irrespective of whether the model is estimated by parametric or semiparametric techniques. The first assumption is standard in sample selection modeling and is needed to identify the parameters of our model:

ASSUMPTION 1: *$w$ contains at least one variable which is not included in $x$.*

Assumption 1 is a well-known exclusion restriction on the variables appearing in the main equation. It says that there is at least one variable included in the selection equation which can be excluded from the main equation (i.e., a variable that has no direct impact on the dependent variable).

Our next assumption is on the sampling process:

ASSUMPTION 2: *$\{y_i^*, x_i, d_i^*, w_i\}_{i=1}^N$ is an i.i.d. sample from some underlying distribution. $y_i \equiv 1(y_i^* > 0)$ is observable if and only if $d_i \equiv 1(d_i^* > 0) = 1$.*

We further require that there is no "multicollinearity":

ASSUMPTION 3: *$x$ and $w$ are not contained in any proper linear subspace of $\mathbb{R}^K$ and $\mathbb{R}^L$, respectively, where $K$ and $L$ denote the dimension of $x$ and $w$, respectively.*

This is again a standard assumption which is needed to identify the model parameters.

Having made these basic assumptions, we proceed to consider parametric and semiparametric estimation of our model.

# 3 Parametric Estimation

We briefly consider parametric estimation of the model set up in the last section, as proposed by van de Ven and van Praag (1981).[2] To do this, we have to make an assumption on the joint distribution of the error terms of main and selection equation.

ASSUMPTION H: *$(\varepsilon, u)$ has a bivariate standard normal distribution with correlation coefficient $\rho$, i.e. $Pr(\varepsilon_i < a, u_i < b|x_i, w_i) = \Phi_2(a, b; \rho) \quad \forall i = 1, \ldots, N$, where $\Phi_2(\cdot, \cdot; \rho)$ denotes the bivariate standard normal c.d.f. with correlation coefficient $\rho$.*

The log-likelihood function for this model is given by

$$\log L(\beta, \gamma) = \sum_{i=1}^{N} \log(1 - \Phi(w_i'\gamma))1(d_i = 0) + \sum_{i=1}^{N} \log(\Phi_2(x_i'\beta, w_i'\gamma; \rho))1(d_i = 1, y_i = 1) \tag{5}$$

$$+ \sum_{i=1}^{N} \log(\Phi_2(-x_i'\beta, w_i'\gamma; -\rho))1(d_i = 1, y_i = 0), \tag{6}$$

where $\Phi(\cdot)$ denotes the univariate standard normal c.d.f. Maximization of the log-likelihood function can be carried out as usual, giving estimates of $\beta$ and $\gamma$ which are consistent, asymptotically normal and asymptotically efficient (provided Assumption H holds). Formally, we establish Theorem H:

THEOREM H: *Let $\theta = (\hat{\beta}', \hat{\gamma}')'$. Under assumptions 1, 2, H and standard regularity conditions as in Amemiya (1985, Theorems 4.1.2 and 4.1.3), we have that (a) $\hat{\theta} - \theta = o_p(1)$ and (b) $\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$, where $I(\theta) = N^{-1}E\left[\frac{\partial L}{\partial \theta}\frac{\partial L}{\partial \theta'}\right]$.*

PROOF: *Follows from standard maximum likelihood theory; see Amemiya (1985), chapter 4.*

□

We will denote the (parametric) maximum likelihood estimator of $\beta$ by $\hat{\beta}_H$, where the "H" is a shortcut for "Heckprob", named after the STATA command for estimating a probit model with sample selection.

[2] Also see Greene (2008), pp. 895-897.

As already raised in the introduction, the "Heckprob" estimator loses its (asymptotic) optimality properties if the assumptions on the distribution of the error terms are not satisfied. In the next section, we will consider semiparametric estimation procedures which do not rely on strong parametric assumptions.

# 4 Semiparametric Estimation

In order to estimate the model set up in section 2 semiparametrically, we first have to make an identifying assumption. Assumption 1 from section 2 is a necessary assumption to identify the model parameters but it is not sufficient.[3] Here we give two identifying assumptions which give rise to different estimation strategies.

ASSUMPTION 4: *Either*

*(a)* $Pr(y_i = 1|d_i = 1, x_i, w_i) = E[1(\varepsilon_i > -x_i'\beta)|w_i'\gamma] = G(x_i'\beta, w_i'\gamma)$ *with* $\frac{\partial G(u,v)}{\partial u} > 0 \quad \forall i = 1, \ldots, N$ *or*

*(b)* $median[\varepsilon_i|d_i = 1, x_i, w_i] = median[\varepsilon_i|w_i'\gamma] = g(w_i'\gamma) \quad \forall i = 1, \ldots, N$

*holds with probability one.*

Assumption 4 (a) allows to estimate the model parameters by semiparametric maximum likelihood. In particular, we propose to estimate $\beta$ by Rothe's (2009) extension of the Klein and Spady (1993) semiparametric estimation procedure for binary choice models. Note that the log-likelihood function of our observed sample is given by

$$\log L(\beta|\gamma = \hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} y_i \log(G(x_i'\beta, w_i'\hat{\gamma})) + (1 - y_i)\log(1 - G(x_i'\beta, w_i'\hat{\gamma})), \qquad (7)$$

where $n < N$ is the number of observations for which the $y$ is observable. Note that we used a preliminary estimate of $\gamma$ in the log likelihood function. In principle, we could estimate the parameters of main and selection equation simultaneously which would

[3]Of course, Assumption 3 is needed for identification as well. We highlight Assumption 1 because it is specific to sample selection models, whereas Assumption 3 is a more general assumption which is usually required to hold in any point-identified econometric model.

be efficient. However, two-stage estimators are often preferred due to a reduction of dimensionality and computational issues regarding the stability of numerical optimization routines. Consequently, we assume that the parameters in $\gamma$ can be consistently estimated by some first-stage estimation procedure:

ASSUMPTION 5: *For the first-stage estimator of $\gamma$, it holds that $\hat{\gamma} - \gamma = o_p(1)$.*

However, the log-likelihood function cannot simply be maximized in order to yield estimates of $\beta$ since the function $G(\cdot)$ is unknown. Klein and Spady (1993) and Rothe (2009) suggest to replace this function by kernel density estimates. More specifically,

$$\hat{G}(x_i'\beta, w_i'\hat{\gamma}) = \frac{\frac{1}{n}\sum_{j\neq i}^{n} y_j \frac{1}{h_x h_w} K(x_i'\beta/h_x) K(w_i'\hat{\gamma}/h_w)}{\frac{1}{n}\sum_{j\neq i}^{n} \frac{1}{h_x h_w} K(x_i'\beta/h_x) K(w_i'\hat{\gamma}/h_w)}, \tag{8}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a univariate kernel density function (e.g., the standard normal probability density function) and $h_x$ and $h_w$ are bandwidth parameters satisfying $h_x \to 0$ and $h_w \to 0$ as $n \to \infty$. Then, estimation can be performed as usual with $G(\cdot)$ in (7) replaced by (8), i.e.,

$$\hat{\beta}_{KS} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} y_i \log(\hat{G}(x_i'\beta, w_i'\hat{\gamma})) + (1 - y_i) \log(1 - \hat{G}(x_i'\beta, w_i'\hat{\gamma})). \tag{9}$$

Since the coefficients of a binary choice model are only identified up to scale, we have to put a restriction on $\beta$. A common choice is to set the first component of $\beta$ equal to one, i.e., $\beta = (1, \tilde{\beta}')'$.

In order to prevent the log-likelihood function from becoming unbounded, one could multiply the contribution of a single observation in the log-likelihood function with a trimming factor $\tau_i$, which excludes observations for which $G(x_i'\beta, w_i'\hat{\gamma})$ is close to one or zero. Introducing trimming facilitates the derivation of the asymptotic distribution of the estimator, but is usually ignored in practical applications.

Note that the selection index $w'\gamma$ is estimated over the full set of observations $N$, whereas the log-likelihood function for the Klein and Spady estimator only includes $n < N$, i.e., the selected observations. Typically, the rate of convergence of the first-stage

estimator is faster than the convergence rate of the second-stage estimator, since the former is based on more observations. This means we can asymptotically ignore the fact that $w'\gamma$ has been estimated. Formally, we strengthen Assumption 5:

ASSUMPTION 5': $\sqrt{n}(\hat{\gamma} - \gamma) = o_p(1)$.

Furthermore, we restate in slightly modified form the assumptions in Rothe (2009) used to establish the consistency and asymptotic normality of his estimator. We summarize these assumptions in Assumption 6:

ASSUMPTION 6:

*a)* *There exists a unique interior point $\tilde{\beta} \in \mathcal{B}$ such that the relationship $E[y|x, w, d = 1] = E[y|x'\beta, w'\gamma]$ holds for $(x, w) \in \mathcal{A}$, a set with positive probability.*

*b)* *The parameter space $\mathcal{B}$ is a compact subset of $\mathbb{R}^{K-1}$ and $\tilde{\beta}$ is an element of its interior.*

*c)* *(i) For all $\tilde{\beta} \in \mathcal{B}$, the distribution of the random vector $(x'\beta, w'\gamma)$ admits a density function $f(x'\beta, w'\gamma)$ with respect to Lebesgue measure.*

*(ii) For all $\tilde{\beta} \in \mathcal{B}$, $f(x'\beta, w'\gamma)$ is $r$ times continuously differentiable in its arguments and the derivatives are uniformly bounded.*

*(iii) For all $\tilde{\beta} \in \mathcal{B}$, $G(x'\beta, w'\gamma)$ is $r$ times continuously differentiable in its arguments and the derivatives are uniformly bounded.*

*(iv) $f(x'\beta, w'\gamma)$ and $G(x'\beta, w'\gamma)$ are twice continuously differentiable in $\tilde{\beta}$.*

*d)* *For $\mathcal{X}$ a compact subset of the support of $(x, w)$, define $T(\mathcal{X}) = \{t \in \mathbb{R}^2 : \exists (x, w) \in \mathcal{X}, \tilde{\beta} \in \mathcal{B}$ s.t. $t = (x'\beta, w'\gamma)\}$. Then $\mathcal{X}$ is chosen such that:*

*(i)* $\inf_{t \in T(\mathcal{X}), \tilde{\beta} \in \mathcal{B}} f(x'\beta, w'\gamma) > 0$

*(ii)* $\inf_{t \in T(\mathcal{X}), \tilde{\beta} \in \mathcal{B}} G(x'\beta, w'\gamma) > 0$ *and* $\sup_{t \in T(\mathcal{X}), \tilde{\beta} \in \mathcal{B}} G(x'\beta, w'\gamma) < 1$.

*e) The matrix*

$$\Sigma = E\left[\frac{\tau(\partial G(x'\beta, w'\hat{\gamma})/\partial\tilde{\beta})(\partial G(x'\beta, w'\hat{\gamma})/\partial\tilde{\beta})'}{G(x'\beta, w'\hat{\gamma})(1 - G(x'\beta, w'\hat{\gamma}))}\right]$$

*is positive definite.*

*f) The kernel function $K : \mathbb{R} \to \mathbb{R}$ satisfies (i) $\int K(z)dz = 1$, (ii) $\int K(z)z^{\mu}dz = 0$ for all $\mu = 1, \ldots, r-1$, (iii) $\int |K(z)z^{\mu}|dz < \infty$ for $\mu = r$, (iv) $K(z) = 0$ if $|z| > 1$, (v) $K(z)$ is $r$ times continuously differentiable.*

*g) The bandwidths $h_x$ and $h_w$ satisfy: $h = cn^{-\delta}, h \in \{h_x, h_w\}$ for some constant $c > 0$ and $\delta$ such that $1/(2r) < \delta < 1/8$.*

We can now establish the following theorem:

THEOREM 1: *Under Assumptions 1-3, 4 (a), 5' and 6, we have that (a) $\hat{\tilde{\beta}}_{KS} - \tilde{\beta} = o_p(1)$ and (b) $\sqrt{n}(\hat{\tilde{\beta}} - \tilde{\beta}) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1}\Psi_1\Sigma^{-1})$, where*

$$\Sigma = E\left[\frac{\tau(\partial G(x'\beta, w'\hat{\gamma})/\partial\tilde{\beta})(\partial G(x'\beta, w'\hat{\gamma})/\partial\tilde{\beta})'}{G(x'\beta, w'\hat{\gamma})(1 - G(x'\beta, w'\hat{\gamma}))}\right] \tag{10}$$

*and*

$$\begin{aligned}\Psi_1 =& E[\{[\tau(\partial G(x'\beta, w'\hat{\gamma})/\partial\tilde{\beta}) - E[\tau(\partial G(x'\beta, w'\hat{\gamma})/\partial\tilde{\beta})|x'\beta, w'\hat{\gamma}]] \\ & [\tau(\partial G(x'\beta, w'\hat{\gamma})/\partial\tilde{\beta}) - E[\tau(\partial G(x'\beta, w'\hat{\gamma})/\partial\tilde{\beta})|x'\beta, w'\hat{\gamma}]]\}/(G(x'\beta, w'\hat{\gamma})(1 - G(x'\beta, w'\hat{\gamma}))].\end{aligned} \tag{11}$$

PROOF: *Our estimation approach is conceptually the same as in Rothe (2009). The difference is that Rothe proposes a control function approach to control for endogeneity of covariates instead of sample selectivity. In his derivations, a reduced form error term (resulting from the reduced form equation of the endogenous explanatory variable) plays the same role as $w'\gamma$ does for our estimator. We can thus follow the arguments in Rothe (2009), who derives consistency and asymptotic normality of his estimator (by checking*

*whether the conditions in Chen, Linton and van Keilegom, 2003, are fulfilled). The only modification which must be made for our estimator is that the estimation error due to estimation of $w'\gamma$ vanishes asymptotically since $\sqrt{n}(\hat{\gamma}-\gamma) = o_p(1)$. Therefore, the asymptotic covariance matrix of our estimator does not include an additional term which captures the uncertainty due to the estimation of $w'\gamma$, i.e., the $\Sigma^{-1}\Psi_2\Sigma^{-1}$ term in Theorem 3 of Rothe (2009) vanishes.*

□

As the asymptotic distribution provided in the Theorem requires knowledge of unknown derivatives, an alternative way to conduct inference is to use the bootstrap. Rothe (2009) argues in the same way.

Now we consider estimation when Assumption 4 (b) is valid. Assumption 4 (b) is on the conditional median of $\varepsilon$. It allows to rewrite the (observable part of the) main equation as

$$y_i^* = x_i'\beta + g(w_i'\gamma) + v_i, \quad i = 1, \ldots, n, \tag{12}$$

where $v_i \equiv \varepsilon_i - median[\varepsilon_i|w_i'\gamma]$. Since, by construction, $v$ has a conditional median of zero, we could apply Manski's (1975) maximum score estimator to obtain parameter estimates. Again, this is not feasible as the function $g(\cdot)$ is unknown. However, suppose we have two individuals with the same value of $w'\gamma$. In that case, we can subtract equation (12) for individual $i$ from the equation for individual $j$, i.e.,

$$y_i^* - y_j^* = (x_i - x_j)'\beta + g(w_i'\gamma) - g(w_j'\gamma) + v_i - v_j \tag{13}$$

$$= (x_i - x_j)'\beta + v_i - v_j. \tag{14}$$

The differencing in equations (13) and (14) resembles the underlying idea of Manski's (1987) conditional maximum score approach for binary panel data. In the panel data approach, an individual specific "fixed effect" is removed by differencing over time for a

given individual, while in our case we have a cross sectional data set and use differencing to remove an unknown function.

Moreover, Powell (1987) used the same strategy to estimate an ordinary sample selection with a continuous dependent variable. He also augmented the main equation with a control function, which is a generalization of the inverse Mills ratio term occurring in the ordinary Heckman selection model with normally distributed error terms. As in our approach, Powell then combined "similar" observations, differenced the main equations, thereby eliminating the unknown control function, and estimated the model parameters using the differenced variables.[4]

Note that despite of the model transformation in equations (13) and (14) due to differencing we are still able to identify the parameters in $\beta$. We simply combine only observations for which $y_i \neq y_j$. Then, we have the following correspondence:

$$y_i^* - y_j^* \begin{cases} > 0 \text{ if } y_i = 1 \wedge y_j = 0 \\ \\ < 0 \text{ if } y_i = 0 \wedge y_j = 1 \end{cases}, \tag{15}$$

which implies that the transformed model using only observations with $y_i \neq y_j$ is again a binary choice model. Since the conditional median of the differenced error terms is zero, we can apply the maximum score estimator to the transformed model in order to obtain an estimate of $\beta$.

In general, however, $w'\gamma$ will assume a continuum of values rather than a finite number. Hence, it will be nearly impossible to find and combine observations with the same value of the selection index $w'\gamma$. Instead, one may combine individuals with a "similar" index value. This yields a maximum score estimator which puts most weight on pairs of observations which have "close" selection indexes. More precisely, our proposed estimator

[4]This strategy has also been used by Ahn and Powell (1993). In their case, the control function depends on the probability of being selected. On the contrary, in our and Powell's (1987) approach, the control function depends on the selection index $w'\gamma$. A further application of the strategy has been provided by Kyriazidou (1997), who considered semiparametric estimation of a panel data sample selection model.

of $\beta$ is given by

$$\hat{\beta}_{MS} = \arg\max_{\beta} -\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |\tilde{y}_{ij} - 1(\tilde{x}'_{ij}\beta > 0)| \frac{1}{h} K(\tilde{w}'_{ij}\hat{\gamma}/h) 1(y_i \neq y_j), \qquad (16)$$

where $\tilde{y}_{ij} = 1(y_i^* - y_j^* > 0)$, $\tilde{x}_{ij} = x_i - x_j$, $\tilde{w}_{ij} = w_i - w_j$, $K : \mathbb{R} \to \mathbb{R}$ is a univariate kernel density function which is bounded, absolutely integrable and symmetric about zero, and $h$ is a bandwidth parameter which converges to zero when the sample size approaches infinity. Note that the minimization problem in equation (16) uses only observations for which $y_i \neq y_j$, and, for the same reasons as given above, preliminary estimates of $\gamma$.

Note further that $K(\cdot)$ serves as a weighting function. In particular, pairs of observations who are very similar in their selection index $w'\gamma$ receive a relatively large weight, whereas pairs of observations who differ substantially in $w'\gamma$ take a weight which is close to zero. In the limit, only pairs of observations with very close selection indexes receive a positive weight. So in the limit it is possible to base estimation on pairs of observations with roughly the same selection index, so that the impact of the control function vanishes (since it is completely differenced out) and we can consistently estimate the model parameters.

However, since the objective function in (16) is not differentiable it may be difficult to obtain parameter estimates. Horowitz (1992) proposes a smoothed maximum score estimator which features a smooth objective function. Using that estimator, our estimation problem may be written as

$$\hat{\beta}_{SMS} = \arg\max_{\beta} \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (2\tilde{y}_{ij} - 1) \frac{1}{h_x} \Phi(\tilde{x}'_{ij}\beta/h_x) \frac{1}{h_w} K(\tilde{w}'_{ij}\hat{\gamma}/h_w) 1(y_i \neq y_j), \qquad (17)$$

where $\Phi(\cdot)$ is a smooth function satisfying $\lim_{u\to-\infty} \Phi(u) = 0$ and $\lim_{u\to\infty} \Phi(u) = 1$, and $h_x$ is a bandwidth parameter which converges to zero when the sample size approaches infinity.

Note again that both the maximum score and the smoothed maximum score estimator

estimate $\beta$ only up to scale. We will set the same identifying assumption as in the case of the Klein and Spady estimator, hence $\beta = (1, \tilde{\beta}')'$.

In order to establish consistency of $\hat{\tilde{\beta}}_{MS}$ and $\hat{\tilde{\beta}}_{SMS}$ we need some further assumptions which lead to consistency of the maximum score and smoothed maximum score estimators in general, i.e. without sample selectivity. We take these assumptions from Horowitz (1992) and summarize them in Assumption 6:

ASSUMPTION 7:

*a)* $0 < Pr(\tilde{y} = 1|\tilde{x}, \tilde{w}'\gamma = 0) < 1$ *for almost every* $\tilde{x}$.

*b)* $\beta_1 \neq 0$, *and for almost every* $(\tilde{x}_2, \ldots, \tilde{x}_K)$, *the distribution of* $\tilde{x}_1$ *conditional on* $(\tilde{x}_2, \ldots, \tilde{x}_K)$ *and* $\tilde{w}'\gamma = 0$ *has everywhere positive density with respect to Lebesgue measure.*

*c)* $\beta_1 = 1$ *and* $\tilde{\beta}$ *is contained in a compact subset of* $\mathbb{R}^{K-1}$.

Moreover, we need an assumption on the marginal distribution of $\tilde{w}'\gamma$, which is taken from Assumption R4 in Kyriazidou (1997):

ASSUMPTION 8: *The marginal distribution of* $W \equiv \tilde{w}'\gamma$ *is absolutely continuous, with density function* $f_W$ *which is bounded from above on its support and strictly positive at zero, i.e.* $f_W(0) > 0$.

We establish the following theorem:

THEOREM 2: *Under Assumptions 1-3, 4 (b), 5, 7 and 8 we have that* $\hat{\tilde{\beta}}_{MS} - \tilde{\beta}_{MS} = o_p(1)$ *and* $\hat{\tilde{\beta}}_{SMS} - \tilde{\beta}_{SMS} = o_p(1)$.

PROOF: *First, let*

$$S_{MS} = -\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |\tilde{y}_{ij} - 1(\tilde{x}'_{ij}\beta > 0)| \frac{1}{h} K(\tilde{w}'_{ij}\hat{\gamma}/h) 1(y_i \neq y_j)$$

*and*

$$S_{SMS} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (2\tilde{y}_{ij} - 1) \frac{1}{h_x} \Phi(\tilde{x}'_{ij}\beta/h_x) \frac{1}{h_w} K(\tilde{w}'_{ij}\hat{\gamma}/h_w) 1(y_i \neq y_j).$$

*denote the objective function whose maximization yields* $\hat{\tilde{\beta}}_{MS}$ *and* $\hat{\tilde{\beta}}_{SMS}$*, respectively. Combining Lemma A1 of Kyriazidou (1997) with a law of large numbers for U-statistics (see Serfling, 1980, Theorem A, p. 190) and Lebesgue's dominated convergence theorem (see Billingsley (1995), Theorem 16.4.) to handle* $\hat{\gamma}$*, we obtain that*

$$S_{MS} \xrightarrow{p} S^* \quad \textit{uniformly}$$

*and*

$$S_{SMS} \xrightarrow{p} S^* \quad \textit{uniformly,}$$

*where* $S^* = -f_W(0)E\left[|\tilde{y} - 1(\tilde{x}'\beta > 0)|1(y_i \neq y_j)|\tilde{w}'\gamma = 0\right] \int K(v)dv$*. Uniform convergence follows from the boundedness of the objective functions. The implied equivalence of the probability limits of the maximum score and smoothed maximum score objective functions has been proven by Horowitz (1992). To prove consistency of* $\hat{\tilde{\beta}}_{MS}$ *and* $\hat{\tilde{\beta}}_{SMS}$*, respectively, it remains to show that* $S^*$ *is uniquely maximized at* $\tilde{\beta}$*. To do this, we just have to consider the expectation term in* $S^*$ *as the remaining terms are independent of* $\tilde{\beta}$*, so* $S^*$ *is maximized when the expectation is minimized. Since the expectation in* $S^*$ *is conditional on* $\tilde{w}'\gamma = 0$*, we just have the situation of an "ordinary" binary choice model where there is no unknown function* $g(\cdot)$*. We just have a binary dependent variable* $\tilde{y}$ *and a set of covariates* $\tilde{x}$*. Hence, the same arguments which are needed to show point-identification of the maximum score estimator can be applied (see Manski, 1985, or Newey and McFadden, 1994, p.2139) to show point-identification of* $\tilde{\beta}$*, which in connection with the uniform convergence of* $S_{MS}$ *and* $S_{SMS}$ *towards* $S^*$ *implies convergence in probability of* $\hat{\tilde{\beta}}_{MS}$ *and* $\hat{\tilde{\beta}}_{SMS}$ *towards* $\tilde{\beta}$*.*

□

We do not provide asymptotic distribution theory for these estimators since in case of the maximum score estimator the form of the asymptotic distribution is very complicated

and not suitable for practical inference; as an alternative, Manski and Thompson (1986) examined the performance of the bootstrap and found encouraging results. In case of the smoothed maximum score estimator Horowitz (1992) derived the asymptotic distribution and reported a relatively weak finite sample performance of the asymptotic theory, hence he also proposes to use the bootstrap.

We follow these lines of reasoning and propose to use the bootstrap for obtaining standard errors, too; for instance, the standard errors in our empirical example in section 6 have been obtained in that way.

# 5 Monte Carlo Evidence

In this section, we provide some (limited) Monte Carlo evidence on the finite sample performance of our proposed estimators. We not only consider the behavior of the semi-parametric estimators from section 3, but also the behavior of the parametric "Heckprob" estimator from section 2. Our simulated model is given by

$$y_i^* = \beta_1 q_i + \beta_2 x_i + \varepsilon_i \tag{18}$$

$$d_i^* = x_i + w_i + u_i \tag{19}$$

$$\varepsilon_i = u_i + \nu_i \tag{20}$$

$$d_i = 1(d_i^* > 0) \tag{21}$$

$$y_i = \begin{cases} 1(y_i^* > 0) \text{ if } d_i = 1 \\ \text{"missing" otherwise} \end{cases}, \tag{22}$$

$i = 1, \ldots, N$, where $\beta_1 = \beta_2 = 1$, $x \sim U_{[0,1]}$, $q \sim \mathcal{N}(1,1)$ and $w \sim \mathcal{N}(1,1)$.

For $u$ and $\nu$, we consider the following distributions:

(i) $u \sim \mathcal{N}(0,1), \nu \sim \mathcal{N}(0,5)$

(ii) $u \sim \mathcal{N}(0,1), \nu \sim 0.8\mathcal{N}(-1,0.6) + 0.2\mathcal{N}(4,2)$

(iii) $u \sim \mathcal{N}(0, \exp(0.1 + 0.5(x + w))), \nu \sim \mathcal{N}(0, 5)$

(iv) $u \sim \mathcal{N}(0, 1), \nu \sim \mathcal{N}(0, \exp(0.1 + 0.5(q + x)))$.

Except for distribution (iii), these distributions have been taken from Rothe (2009). In case of distribution (i) we have a normal distribution for which the parametric "Heckprob" estimator should yield consistent estimates. Distribution (ii) is a mixture of two normal distributions. Its density is skewed to the right and bimodal (see Rothe, 2009). Distribution (iii) aims to consider the effects of conditional heteroskedasticity in the selection equation. In this case, all three semiparametric estimation procedures should yield consistent estimates. On the other hand, distribution (iv) implies conditional heteroskedasticity in the main equation only. In this specification, only the Klein and Spady estimator should yield consistent estimates.

Note that our proposed estimators each require a normalization. We implemented such a normalization by setting $\beta_1$ equal to its true value of one. That means, the only parameter to be estimated in the main equation is $\beta_2$.

For all our proposed estimators, we have to specify kernel-type functions and bandwidths. We made the following choice: For the Klein and Spady estimator (KS), we chose the standard normal p.d.f. as the kernel function. Instead of specifying bandwidths in advance, we follow Rothe (2009) and let the bandwidth choice be a part of the optimization problem. Put differently, our optimization routine simultaneously seeks for the optimal parameter values *and* the optimal bandwidth values. Advantages of this procedure are that (a) there is no subjectivity in bandwidth choice and (b) a very large value of $h_w$ would indicate that sample selection bias is not relevant (see Rothe, 2009).

In case of the maximum score estimator (MS), we chose the standard normal p.d.f. as the kernel function and selected a bandwidth according to the rule $h = n^{-1/6.5}$. For the smoothed maximum score estimator (SMS) we chose the standard normal c.d.f. for $\Phi(\cdot)$ and the standard normal p.d.f. for $K(\cdot)$. We set $h_x = h_w = n^{-1/6.5}$. We also normalized the arguments of the kernel functions to have unit variance, which justifies the choice of the same bandwidth rule for both kernel functions. In contrast to the

Klein and Spady estimator, the bandwidths are given ad hoc rather than being part of the optimization problem. We did this because computation of the maximum score and smoothed maximum score estimator is relatively difficult due to the presence of local optima. Instead, we specified the bandwidths in advance so that there is only one parameter, i.e. $\beta_2$, over which optimization is performed. To find the optimal value of $\hat{\beta}_2$, we performed a grid search over the interval $[-1, 3]$ with a step length of 0.005.

We performed the Monte Carlo simulations for sample sizes of $N \in \{250, 500, 1000\}$ and used 100 replications. For each simulation we computed the mean of the estimates over the replications, as well as the standard deviation and the root mean squared error (RMSE). These measures of estimator performance are typically used in Monte Carlo studies and should help to gauge the performance of the estimators under consideration.

At first we seek to analyze the performance of our three proposed estimators independently of the first-stage estimation of the selection index $w'\gamma$. Recall that each of our semiparametric estimators relies on first-stage estimates of the selection index. In principle, we could use any first-stage estimator provided we use the same estimator for all three second-stage estimators (so that we can reasonably compare the second-stage estimates). We, however, refrain for the moment from estimating the selection index and consider how the estimators perform in an "ideal" situation where the selection index is known, so that estimation results of the second stage are not contaminated by estimation error in the first stage.

Table 1 contains the results for distribution (i) and a known selection index. We see from Table 1 that, in terms of RMSE, the estimators perform better as the sample size increases (as expected). However, we also see that the mean of the estimates differs slightly from the true value of one even for the relatively large sample size of $N = 1000$. The reason is that the estimates exhibit a lot of variation, as indicated by the standard deviations. Among the three estimators, the maximum score and the smoothed maximum score estimator have lower RMSE's then the Klein and Spady estimator due to lower standard deviations, which means that these estimators seem to be slightly more precise.

We will investigate if this property also holds true for the remaining distributions and in case that the selection index is estimated rather than known in advance.

In Table 2, we reconsider distribution (i) but now the selection index is estimated. For obtaining these estimates, we used the same type of estimator in the first stage as in the second stage. That means, for the Klein and Spady estimator we used a Klein and Spady estimator in the first stage, for the maximum score estimator we used a maximum score estimator in the first stage and for the smoothed maximum score estimator we used a smoothed maximum score estimator in the first stage. The idea is that in practice it would seem a bit uncommon to use one semiparametric estimator in the first stage and to use a different semiparametric estimator in the second stage, at least in principle. In the empirical example in section 6 we will, however, provide a practical reason why using different estimators in first and second stage might be sensible.

Note that Table 2 also contains results for the parametric "Heckprob" model from section 2. Since distribution (i) implies a normal distribution of the error terms in main and selection equations, one might expect that the "Heckprob" model should perform quite well. Table 2 confirms this conjecture. We see that the estimators perform relatively similar with respect to the standard deviation. The differing means are again a result of the relatively great deal of variation of the estimators. When comparing these results to those from Table 1 we see that there is not much difference in standard deviations. Hence we may conclude that using the same type of estimator for first and second stage does not lead to stark distortions between the estimators.

In Table 3 we consider the mixed normal distribution (ii). We can see that the "Heckprob" estimator performs surprisingly well, having the least bias and the least RMSE among all estimators and for all sample sizes. The standard deviations of the estimators are generally lower when compared to distribution (i), which is due to the fact that the error term variance is smaller for distribution (ii). Among the semiparametric estimators, the maximum score estimator has the least bias but the largest RMSE.

Table 4 contains results for distribution (iii) where we have conditional heteroskedas-

ticity in the selection equation but not in the main equation. In this case, all three semiparametric estimators are consistent, whereas the "Heckprob" estimator is not. However, from Table 4 we see that the "Heckprob" estimator performs very well. All estimators exhibit a great deal of variation, which again explains the slight biases of these estimators.

Finally, we consider distribution (iv), where we have conditional heteroskedasticity in the main equation but not in the selection equation. In this case, only the Klein and Spady estimator is consistent. From Table 5 we see that not only the Klein and Spady estimator but also the remaining semiparametric estimators perform relatively well. The "Heckprob" estimator, however, exhibits a larger bias than one might have expected. Nevertheless, the "Heckprob" estimator has the smallest RMSE among all estimators.

From these results, we can draw two major conclusions. First, in all considered designs the estimators exhibit a lot of variation (as indicated by the standard deviations). Moreover, we also experienced considerable variation between the estimators. Hence, the first major conclusion is that one needs substantial sample sizes to obtain precise estimates. Second, the parametric "Heckprob" estimator performs relatively well even in situations where it should be biased. Of course, these results may be an artifact of our simulation designs and need not hold in general. However, especially in small sample sizes the parametric estimator may be a sensible alternative due to its favorable RMSE properties. At least one could test the parametric estimator against a semiparametric alternative (at least in a heuristic way, e.g. by considering whether the confidence intervals overlap). When considering the standard deviations of the semiparametric estimators over the simulations, it seems relatively likely that results based on the parametric estimator would not be rejected empirically. For large sample sizes, however, a semiparametric estimator should be preferred as it relies on considerably fewer assumptions than the parametric estimator. Put differently, the larger the sample size the more obvious it should be when the parametric assumptions are not fulfilled.

# 6 Empirical Example

In this section, we present an empirical example in order to illustrate the applicability of our proposed estimators to real data. In this example, we seek to analyze whether the number of children has an effect on a woman's probability of (partly) working from home. We are thus concerned with a situation where we have a binary dependent variable (working from home: yes/no) which is only observable for women who are working. This fact may constitute a sample selection bias.

Now we describe our empirical specification. Our main equation contains the number of children and education attainment as explanatory variables. With regard to our dependent variable, we expect the following effects: We conjecture that the number of children has a positive effect on the probability of working from home, since a larger number of children requires a higher amount of child care services. We also expect a positive effect of education, since a better education may be correlated with "technology-affine" jobs in which it is possible to work from home. For instance, working from home may require the capability of getting along with electronic equipment (e.g., personal computers).

Since our dependent variable is only observable for those women who are working, we have to specify a selection equation which governs the probability of working. We selected the following explanatory variables: the number of children, education, age and age squared. Since the selection equation contains more variables than the main equation, we suppose that the exclusion restriction from Assumption 1 is satisfied.

Our data is taken from the German Socio-Economic Panel (GSOEP) for the year 2002. Our sample consists of 989 married women aged 25 to 35 with German nationality. From these women, 565 are working (57.1 %). Summary statistics of the variables are given in Table 6.

We specify our estimators as in the last section. That means, in case of the Klein and Spady estimator we selected the standard normal p.d.f. as the kernel function and let the optimal bandwidth be obtained simultaneously with the parameters of interest; in case of the maximum score estimator, we chose the standard normal p.d.f. as the kernel function

and selected a bandwidth according to the rule $h = n^{-1/6.5}$; for the smoothed maximum score estimator we chose the standard normal c.d.f. for $\Phi(\cdot)$ and the standard normal p.d.f. for $K(\cdot)$. We set $h_x = h_w = n^{-1/6.5}$.

However, for the estimation of the selection equation we employed the Klein and Spady estimator irrespective of the second-stage estimator. The reason is that we have four covariates. In this case, using the maximum score or smoothed maximum score estimator is rather complicated since one needs a suitable optimization routine and optimization results may be contaminated by the presence of local maxima. For these reasons, the maximum score and the smoothed maximum score estimator have only seldom been used in applied econometrics. On the contrary, the Klein and Spady estimator works well if the number of covariates is moderate. Since semiparametric estimation of the selection equation requires a normalization, we set the coefficient of education equal to one.

Table 7 contains the Klein and Spady estimates of the selection equation parameters. As expected, the number of children has a negative impact on the probability of working. For a woman's age we get a U-shaped pattern which is plausible for the sample under consideration, since women start working when they are young, then leave the labor market to raise their children and return thereafter. Standard errors of these estimates have been obtained by performing 100 bootstrap replications.

Table 8 contains the second-stage results for the Klein and Spady estimator (KS), the maximum score estimator (MS) and the smoothed maximum score (SMS) estimator. The coefficient of education has been set equal to one due to normalization. We also provide estimates using the "Heckprob" estimator. Standard errors are again based on 100 bootstrap replications. As can be seen from Table 8, the coefficient of the number of children is positive over all estimators. However, only in case of the "Heckprob" and smoothed maximum score estimator the coefficient estimate is significantly different from zero. We get the same picture as in the Monte Carlo simulations from the last section: The semiparametric estimates exhibit a lot of variation and relatively large standard errors. However, the semiparametric estimates also indicate that the effect of the number of

children on the probability of working from home may be larger than the estimate of the "Heckprob" model. Although it is unlikely that the parametric "Heckprob" model would be rejected by the data when compared to one of these semiparametric alternatives, the semiparametric estimates at least hint that the parametric estimates may be biased, i.e. that the effect of the number of children is larger than the parametric estimate indicates.

Finally, we conducted a small robustness check. While in case of the Klein and Spady estimator the bandwidth is selected optimally by being part of the optimization problem, the bandwidths for the maximum score and smoothed maximum score estimator have been selected ad hoc. We, thus, provide some robustness analysis by varying these bandwidths. From Table 9 we see that variations of the bandwidths alter the estimates for the maximum score and smoothed maximum score estimator to some extent, but the differences are relatively small. We conclude that estimation results are not very sensitive with respect to bandwidth choice.

# 7 Endogenous Covariates

In empirical applications, one may often be confronted with variables in the main and selection equation which may be endogenous. In that case, our proposed estimators are inconsistent in general. However, our control function framework easily allows to take endogeneity of covariates into account. To see this, let $x^e$ be an endogenous explanatory variable appearing in the main equation and possibly in the selection equation, too. Moreover, let the reduced form equation for $x^e$ be

$$x_i^e = z_i'\alpha + \eta_i, \tag{23}$$

where $z$ is a vector of instrumental variables and $\eta$ is an error term. We can now modify Assumption 4 to take the endogeneity into account:

ASSUMPTION 4': *Either*

*(a)* $Pr(y_i = 1|d_i = 1, x_i, w_i, z_i, \eta_i) = E[1(\varepsilon_i > -x_i'\beta)|w_i'\gamma, \eta_i] = G(x_i'\beta, w_i'\gamma, \eta_i)$ *with* $\frac{\partial G(u,v,w)}{\partial u} >$

$0 \quad \forall i = 1, \ldots, N$ *or*

*(b)* $median[\varepsilon_i | d_i = 1, x_i, w_i, z_i, \eta_i] = median[\varepsilon_i | w_i'\gamma, \eta_i] = g(w_i'\gamma, \eta_i) \quad \forall i = 1, \ldots, N$

*holds with probability one.*

We can once again implement the estimators proposed above. In case of Assumption 4' (a), we choose a modified Klein and Spady estimator such that

$$\hat{\beta}^e_{KS} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} y_i \log(\hat{G}(x_i'\beta, w_i'\hat{\gamma}, \hat{\eta}_i)) + (1 - y_i) \log(1 - \hat{G}(x_i'\beta, w_i'\hat{\gamma}, \hat{\eta}_i)), \qquad (24)$$

where

$$\hat{G}(x_i'\beta, w_i'\hat{\gamma}, \hat{\eta}_i) = \frac{\frac{1}{n}\sum_{j \neq i}^{n} y_j \frac{1}{h_x h_w h_\eta} K(x_i'\beta/h_x) K(w_i'\hat{\gamma}/h_w) K(\hat{\eta}_i/h_\eta)}{\frac{1}{n}\sum_{j \neq i}^{n} \frac{1}{h_x h_w h_\eta} K(x_i'\beta/h_x) K(w_i'\hat{\gamma}/h_w) K(\hat{\eta}_i/h_\eta)}. \qquad (25)$$

Note that the only difference between equation (23) and equation (8) above is that we have to take the (estimated) reduced form error term of our endogenous variable into account, so that we need an additional kernel function. It is obvious that augmenting the function $G(\cdot)$ with more kernel functions requires large sample sizes to produce reliable estimation results. This problem is even more severe when we have several endogenous explanatory variables. In that case, estimation results might be contaminated by the curse of dimensionality.

If Assumption 4' (b) is true, we can again choose between the maximum score estimator and the smoothed maximum score estimator. In the first case, our proposed estimator of $\beta$ is given by

$$\hat{\beta}^e_{MS} = \arg\min_{\beta} -\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |\tilde{y}_{ij} - 1(\tilde{x}_{ij}'\beta > 0)| \frac{1}{h} K(\tilde{w}_{ij}'\hat{\gamma}/h) \frac{1}{h_\eta} K(\tilde{\hat{\eta}}_{ij}/h_\eta) 1(y_i \neq y_j), \qquad (26)$$

while in the second case

$$\hat{\beta}^{e}_{SMS} = \arg\max_{\beta} \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (2\tilde{y}_{ij} - 1) \frac{1}{h_x} \Phi(\tilde{x}'_{ij}\beta / h_x) \frac{1}{h_w} K(\tilde{w}'_{ij}\hat{\gamma} / h_w) \frac{1}{h_\eta} K(\tilde{\hat{\eta}}_{ij} / h_\eta) 1(y_i \neq y_j), \tag{27}$$

where $\tilde{\hat{\eta}}_{ij} = \hat{\eta}_i - \hat{\eta}_j$ and $h_\eta$ is a bandwidth parameter which converges to zero as the sample size tends to infinity.

Note, once again, that these estimators are based on first-stage estimates not only of the selection index, but of the reduced form error term as well. The reduced form error term can be naturally obtained by an ordinary least squares regression of the endogenous explanatory variable on the instrumental variables. For a consistent estimation of the selection index, it matters whether the endogenous explanatory variable is included in the selection equation as well. If not, the selection index can be estimated as before, using one of the available semiparametric procedures already considered in this paper. However, if the endogenous covariate is included in the selection equation, an application of these procedures would produce inconsistent estimates as the endogeneity is not taken into account. In that case, one must apply estimators for binary choice models which control for endogeneity. Such estimators have been proposed by Blundell and Powell (2004) and Rothe (2009), for instance.

# 8 Conclusion

In this paper, we proposed three semiparametric estimators to estimate a sample selection model with a binary dependent variable. We conducted some Monte Carlo simulations and found that estimates based on these estimators exhibit a lot of variation and come along with large root mean squared errors. On the contrary, the parametric "Heckprob" estimator which is based on a joint normality assumption performs quite well and has sometimes relatively low root mean squared errors.

The conclusions from these findings are that (a) one should use the semiparametric estimators in case of large sample sizes and (b) in small samples, the parametric estimator may be preferred if it is successfully tested against a semiparametric alternative. The reason for preferring parametric estimates is that coefficient estimates, especially in small samples, are estimated with higher precision. However, in large samples it may become obvious that the parametric model is misspecified, hence a semiparametric estimation procedure should be chosen.

As our empirical example has shown, semiparametric estimates, though subjected to a lot of variability, can nevertheless be used to gauge and to improve on parametric estimates. More specifically, our example indicates that the effect of the number of children on the probability of working from home is underestimated if one chooses the parametric "Heckprob" estimator. Indeed, if sample sizes become larger, a semiparametric estimator should clearly be preferred in order to avoid inconsistencies resulting from a misspecified parametric model.

We also outlined an extension of our semiparametric estimators to the case of endogenous covariates. Endogeneity may be a concern in many empirical applications, and not accounting for endogeneity will lead to inconsistent parameter estimates in general. Extending our estimators to handle endogenous covariates is quite straightforward. However, given the variability of the semiparametric estimators shown in section 4 (which do not control for endogeneity), we conjecture that this problem may be even more severe if our estimation procedures also have to account for endogeneity. This indicates that one needs even larger sample sizes to obtain reliable estimates.

From the three proposed semiparametric estimators, the Klein and Spady estimator is the most promising and most likely to be used in applications. This is due to the fact that the maximum score and smoothed maximum score estimator require a rather complicated optimization procedure which should also account for the presence of potentially many local maxima. On the other hand, the Klein and Spady estimator can be obtained quite easily (if the number of covariates is moderate) and has already been used successfully in

applied econometrics in order to estimate binary choice models.

# 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.

Amemiya, T. (1985). *Advanced Econometrics*. Basil Blackwell, Oxford.

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.

Boyes, W. J., Hoffman, D. L., and Low, S. A. (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, 40(1):3–14.

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608.

Greene, W. H. (1992). A statistical model for credit scoring. Working Paper No. EC-95-6, Department of Economics, Stern School of Business, New York University.

Greene, W. H. (2008). *Econometric Analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, 6th edition.

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.

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):pp. 1335–1364.

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

Manski, C. F. (1985). Semiparametric analysis of discrete response : Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27(3):313–333.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, 55(2):357–62.

Manski, C. F. and Thompson, T. (1986). Operational characteristics of maximum score estimation. *Journal of Econometrics*, 32(1):85 – 108.

Meng, C.-L. and Schmidt, P. (1985). On the cost of partial observability in the bivariate probit model. *International Economic Review*, 26(1):71–85.

Mohanty, M. S. (2002). A bivariate probit approach to the determination of employment: a study of teen employment differentials in los angeles county. *Applied Economics*, 34(2):143–156.

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

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume 4 of *Handbook of Econometrics*, chapter 36, pages 2111–2245. Elsevier.

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

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

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York, NY.

Van de Ven, W. P. M. M. and Van Praag, B. M. S. (1981). The demand for deductibles in private health insurance : A probit model with sample selection. *Journal of Econometrics*, 17(2):229–252.

# Appendix

Table 1: Design I - normal + known index

| | | Mean | Std.dev. | RMSE |
|---|---|---|---|---|
| N=250 | KS | 0.9549 | 0.9384 | 0.9395 |
| | MS | 1.1426 | 0.9181 | 0.9292 |
| | SMS | 1.1029 | 0.8803 | 0.8864 |
| N=500 | KS | 0.8715 | 0.6840 | 0.6961 |
| | MS | 0.9535 | 0.6789 | 0.6806 |
| | SMS | 0.9938 | 0.6774 | 0.6774 |
| N=1000 | KS | 0.9704 | 0.5877 | 0.5885 |
| | MS | 1.0264 | 0.5306 | 0.5312 |
| | SMS | 1.0448 | 0.5298 | 0.5317 |

Table 2: Design I - normal + unknown index

| | | Mean | Std.dev. | RMSE |
|---|---|---|---|---|
| N=250 | KS | 0.9935 | 0.9049 | 0.9050 |
| | MS | 1.0921 | 0.9315 | 0.9361 |
| | SMS | 1.1539 | 0.8183 | 0.8328 |
| | Heckprob | 1.2045 | 0.9478 | 0.9699 |
| N=500 | KS | 1.0829 | 0.6524 | 0.6577 |
| | MS | 0.9542 | 0.7621 | 0.7635 |
| | SMS | 1.0918 | 0.7194 | 0.7252 |
| | Heckprob | 1.0415 | 0.7188 | 0.7200 |
| N=1000 | KS | 0.9235 | 0.5576 | 0.5629 |
| | MS | 1.0349 | 0.5536 | 0.5547 |
| | SMS | 1.1381 | 0.5437 | 0.5611 |
| | Heckprob | 1.1075 | 0.5278 | 0.5387 |

Table 3: Design II - mixed normal

| | | Mean | Std.dev. | RMSE |
|---|---|---|---|---|
| N=250 | KS | 0.9582 | 0.7122 | 0.7135 |
| | MS | 1.0270 | 0.7139 | 0.7144 |
| | SMS | 1.2323 | 0.6327 | 0.6744 |
| | Heckprob | 1.0819 | 0.6016 | 0.6072 |
| N=500 | KS | 0.8736 | 0.4902 | 0.5064 |
| | MS | 0.9370 | 0.4960 | 0.5000 |
| | SMS | 1.1107 | 0.4486 | 0.4621 |
| | Heckprob | 1.0143 | 0.4181 | 0.4184 |
| N=1000 | KS | 0.9061 | 0.3551 | 0.3675 |
| | MS | 0.9591 | 0.3853 | 0.3875 |
| | SMS | 1.0873 | 0.3199 | 0.3317 |
| | Heckprob | 1.0112 | 0.3009 | 0.3011 |

Table 4: Design III - heteroskedasticity in selection equation

| | | Mean | Std.dev. | RMSE |
|---|---|---|---|---|
| N=250 | KS | 0.9161 | 0.9395 | 0.9432 |
| | MS | 0.9237 | 1.0451 | 1.0479 |
| | SMS | 0.9356 | 0.8893 | 0.8916 |
| | Heckprob | 0.9739 | 1.1096 | 1.1099 |
| N=500 | KS | 0.8329 | 0.8355 | 0.8522 |
| | MS | 1.0303 | 0.8617 | 0.8623 |
| | SMS | 1.0601 | 0.8266 | 0.8288 |
| | Heckprob | 0.9152 | 0.8227 | 0.8271 |
| N=1000 | KS | 0.9278 | 0.6498 | 0.6539 |
| | MS | 0.9673 | 0.5557 | 0.5567 |
| | SMS | 1.0467 | 0.5290 | 0.5311 |
| | Heckprob | 0.9637 | 0.5093 | 0.5106 |

Table 5: Design IV - heteroskedasticity in main equation

| | | Mean | Std.dev. | RMSE |
|---|---|---|---|---|
| N=250 | KS | 0.8518 | 0.8785 | 0.8910 |
| | MS | 0.9777 | 0.9164 | 0.9167 |
| | SMS | 1.1839 | 0.8249 | 0.8453 |
| | Heckprob | 0.8827 | 0.7572 | 0.7664 |
| N=500 | KS | 0.8211 | 0.6351 | 0.6601 |
| | MS | 0.9931 | 0.8260 | 0.8261 |
| | SMS | 1.1176 | 0.7913 | 0.8001 |
| | Heckprob | 0.7617 | 0.5831 | 0.6304 |
| N=1000 | KS | 0.9288 | 0.5548 | 0.5594 |
| | MS | 1.0700 | 0.6632 | 0.6670 |
| | SMS | 1.1745 | 0.5800 | 0.6059 |
| | Heckprob | 0.7868 | 0.4517 | 0.5000 |

Table 6: Summary statistics

| | Mean | Std. | Min | Max |
|---|---|---|---|---|
| hoffice | 0.156 | 0.363 | 0 | 1 |
| children | 1.499 | 1.068 | 0 | 5 |
| educ | 12.213 | 2.272 | 7 | 18 |
| age | 31.624 | 2.848 | 25 | 35 |
| No. of obs. | | | | 989 |
| No. of obs. working | | | | 565 |

Table 7: Estimates of selection equation parameters

| | |
|---|---|
| children | -0.7721 |
| | (0.2027) |
| age | -0.6471 |
| | (0.6392) |
| age2 | 0.0117 |
| | (0.0104) |
| educ | 1 |
| | (-) |

Note: Standard errors in parentheses. Standard errors are based on 100 bootstrap replications.

Table 8: Estimation results

| | Heckprob | KS | MS | SMS |
|---|---|---|---|---|
| children | 0.4565 | 0.9059 | 0.835 | 1.725 |
| | (0.0477) | (3.0815) | (1.4042) | (0.4826) |
| educ | 0.0441 | 1 | 1 | 1 |
| | (0.0253) | | | |
| const | -1.2384 | - | - | - |
| | (0.4013) | | | |

Note: Standard errors in parentheses. Standard errors are based on 100 bootstrap replications.

Table 9: Varying the bandwidth

| | $h = n^{-1/6.5}$ | $h = n^{-1/6}$ | $h = n^{-1/7}$ | $h = n^{-1/8}$ |
|---|---|---|---|---|
| ms | 0.835 | 0.835 | 0.875 | 0.9 |
| sms | 1.725 | 1.61 | 1.825 | 2 |