Working Paper 99-83
Statistics and Econometrics Series 32
November 1999

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax 34 - 91- 624.9849

SEMIPARAMETRIC THREE STEP ESTIMATION METHODS IN LABOR SUPPLY MODELS

Juan M.Rodríguez-Póo, Stefan Sperlich and Ana I.Fernández*

Abstract

The aim of this paper is to provide an alternative way of specification and estimation of a labor supply model. The proposed estimation procedure can be included in the so called predicted wage methods and its main interest is twofold. First, under standard assumptions in studies of labor supply, the estimator based on predicted wages is shown to be consistent and asymptotically normal. Moreover, we propose also a consistent estimator of the asymptotic covariance matrix. In the main part of the paper we introduce a semiparametric estimator based on marginal integration techniques that allows for nonlinear relationships between the labor supply variable and its covariates. We show that also the wage equation could be modeled nonparametrically. The asymptotic properties of the estimators are given. Finally, in a detailed application we compare the results empirically against those obtained in standard three step estimators based on predicted wages.

Key Words
Semiparametric regression; Heckman three step estimator; additive Models; marginal integration; predicted wage methods.

* Rodríguez-Póo, Departamento de Economía, Universidad de Cantabria, Spain; Sperlich, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Spain e-mail: stefan@est-econ.uc3m.es; Fernández, Departamento de Econometría y Estadística, Universidad del País Vasco, Spain. This research was financially supported by SFB 373, Deutsche Forschungsgemeinschaft, at Humboldt-Universität zu Berlin, Dirección General de Enseñanza Superior del Ministerio de Educación y Ciencia under research grants PB95-0346, PB96-1469-C05-03, and University of the Basque Country research grant UPV-038.321-G55/98. We thank Enno Mammen and four anonymous referees for helpful discussion and finding a fatal error in an earlier version.

1 Introduction

The purpose of this paper is to provide an alternative way of specification and estimation of a standard model of labor supply. Our interest involves a structural labor supply model in which hours of work depend on the wage rate and other explanatory variables. The difficulty in estimating such system occurs because first, information is not available on the wage rate for those who do not work, and second, the wage rate is determined endogeneously. To avoid the first problem, the estimation method must take into account the sample selection bias and the second problem is solved by specifying a relationship that considers the wage rate as a endogeneous variable.

The estimation procedure proposed in this paper is a three step method based on the ideas developed by Heckman (1979) and it can be included in the so called predicted wage methods (Wales and Woodland, 1980). Its main interest is twofold. First, in the standard econometric model that is traditionally assumed in studies of labor supply, the three step estimator based on predicted wages is shown to be consistent and asymptotically normal. Moreover, we provide also a consistent estimator of the asymptotic variance covariance matrix for this three step estimator. Second, the classical assumption of linearity between working hours and other explanatory variables is relaxed allowing for a semiparametric partial additive relationship. It is also possible to identify and estimate the model when both, the hours of work (or participation) and the (log) wage equation are non- or semiparametric. These relationships have been traditionally assumed to be linear, but as Blundell and Meghir (1986) pointed out there exist very few theoretical foundations to support this hypothesis. A natural way to extend the classic (log) linear models is to relax the additive components from linear to arbitrary but smooth functional forms. Further, additive models or modeling has a long history and strong foundation in economic theory, see e.g. the standard work of Deaton and Muellbauer (1980). Beside this, also in nonparametric regression additive models are quite popular for their dimension reduction and interpretability properties, see e.g. Stone (1985). Here, the nonparametric additive components are estimated according to the method developed in Härdle, Huet, Mammen and Sperlich (1998). It is based on quasi Likelihood estimation (Severini and Staniswalis, 1994) and marginal integration techniques (Linton and Nielsen, 1995). The resulting estimators turn out to be semiparametric ones, in the sense that the distribution of the random errors is assumed to be known (gaussian in this paper) but the index function is not specified.

Within this set up, we suggest a root-n consistent estimator for the (log) wage equation, derive the asymptotic distribution and provide a easy to calculate consistent estimator of the asymptotic variance covariance matrix. This is of outstanding interest for the empirical economic studies, and in this paper enables us to make comparisons between the three step fully parametric estimator and the semiparametric one. This certainly holds also for analyzing the possibly nonlinear relationships in the considered models.

In the next section of the paper we specify the simultaneous equation structural model of labor supply. We also recall the basic principles of three step estimation methods based on

predicted wages and we establish the main statistical results in the fully parametric context. In Section 3 we introduce the semiparametric three step estimator and we establish its main asymptotic properties. Here we separate mainly in the cases of modelling the influence of wages linear or non linear since this can produce serious identification problems. Section 4 presents an extensive application based on a Spanish labor force data. In Section 5 we finally conclude. In Appendix I we give the assumptions, in Appendix II we prove the main results and finally in section III of the Appendix we provide an algorithm to compute the proposed estimator. The quoting and discussion of further decisive or important literature is always given in the corresponding sections, especially references to other semiparametric approaches in this field.

2 The Structural Model of Labor Supply

We start by considering a structural econometric model of labor supply. To this end, we previously specify the relationship among wages, hours of work (or participation) and other explanatory variables, and next, we will introduce formally the assumptions that are necessary to obtain the statistical properties of the three step estimator based on predicted wages. We will finally propose a consistent estimator of the asymptotic variance covariance matrix of the previous estimator.

Let us consider a labor supply in which both the wage rate and hours of work are endogeneously determined. The extended model can be expressed as follows. Hours of work are a function of a vector of explanatory variables x which includes the log of the wage rate as its first element $\{w\}$. In addition it is assumed that the log of the wage rate w is a function of another vector of exogeneous variables z. Thus, we have for the individual i we have the following structural model of labor supply

- the hours equation,

$$(1) \qquad y_i = \begin{cases} l_1(w_i, x_i) + u_{1i} & if \quad l_1(w_i, x_i) + u_{1i} > 0 \\ 0 & otherwise \end{cases}$$

- and the log wage equation

$$(2) \qquad w_i = l_2(z_i) + u_{2i} \; .$$

Instead of an hour equation like (1), we could also consider another standard labor participation model in which the endogeneous variable is having a paid job or not (see section 4), so we estimate a probit instead a tobit model. In equation (2) is important to note that we are interested in the unconditional expectation of (log) wages, i.e. not conditioned on 'having a job', though we certainly observe only wages for people with job. So we have

$$\begin{aligned} E[w_i|z_i] &= l_2(z_i) \; , \qquad \text{unconditioned, but} \\ (3) \qquad E[w_i|z_i, y_i > 0] &= l_2(z_i) + E[u_{2i}|y_i > 0] \; = \; l_2(z_i) + E[u_{2i}|\zeta_i = 1] \; , \end{aligned}$$

where ζ_i is a binary response variable indicating whether person i has a job 1 or not 0.

As it has been remarked in Wales and Woodland (1980), there exist two problems in the estimation of the structural parameters of the previous simultaneous equation system. First, both equations are subject to sample selection bias and second, the model is a set of two simultaneous equations in which the wage rate, which is an explanatory variable in the hours equation, is correlated with the hours equation disturbance.

In order to solve these problems and estimate the structural parameters of the previous model, some further hypotheses are needed.

(A.1) The values $\{w_i, x_i, z_i\}_{i=1}^N$ are realizations from i.i.d random variables, where $W \in I\!R$, $X \in I\!R^{p+d}$ and $Z \in I\!R^r$. Moreover, $\{y_i\}_{i=1}^N$ are realizations from a truncated random variable, and ζ is a binary variable that takes the value 1 as $y > 0$ and 0 otherwise.

(A.2) The indicator ζ has bounded support, and X and Z have finite sixth order moments.

(A.3) The data satisfy the restrictions (1) and (2).

(A.4) Moreover, for the linear model case we have

$$\begin{aligned} l_1(w_i, x_i) &= \beta_w w_i + x_i^T \beta, \\ l_2(z_i) &= z_i^T \gamma_1 \end{aligned} \tag{4}$$

where the vector z contains at least one variable not contained in x.

(A.5) Define the parameter vector $\theta = (\beta_w, \beta, \gamma_1)$ and the parameter space $\Theta = I\!B_w \times I\!B \times \Gamma$. Then $\theta \in \Theta$, Θ is a compact set, and θ_0 is an interior point of Θ.

(A.6) $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim N(0, \Sigma)$ where $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

Taking into account these assumptions, in order to estimate all structural parameters of the labor supply model several procedures have been proposed in the literature (see Wales and Woodland, 1980). Among them, the methods based on predicted wage rates have been followed by several authors (see Boskin, 1973; Hall, 1973 and Rosen, 1976). Traditionally, this method has consisted on a three step procedure.

In the first step we estimate the parameters of the reduced form model for the hours equation by using a probit maximum likelihood procedure. For the ease of notation we assume that z contains no variable nor a linear combination of the variables in x. Notice that this is not necessary for the suggested procedures but would otherwise complicate the presentation a lot. The reduced form model is given by

$$\zeta_i = \begin{cases} 1 & if \quad z_i^T \alpha + x_i^T \beta + v_{1i} > 0 \\ 0 & otherwise \end{cases} . \tag{5}$$

indicating whether somebody has a paid job or not.

The relationship between the structural and the reduced form parameters is $\alpha = \beta_w \gamma$ and $v_{1i} = \beta_w u_{2i} + u_{1i}$. The variable ζ_i is equal to 1 iff $y_i > 0$ and 0 otherwise, and therefore, under assumptions (A.1) to (A.6) the likelihood function has the form

$$(6) \qquad L(\alpha, \beta) = \prod_{i=M+1}^{N} \left(1 - F\left(\frac{z_i^T \alpha + x_i^T \beta}{\sigma_v}\right)\right) \prod_{i=1}^{M} F\left(\frac{z_i^T \alpha + x_i^T \beta}{\sigma_v}\right),$$

where M is the number of individuals for those who both wages and number of working hours are observed, N is the number of individuals in the sample, $F(\bullet)$ is the cumulative normal distribution function and $\sigma_v^2 = \beta_w^2 \sigma_2^2 + \sigma_1^2 2\beta_w \rho \sigma_1 \sigma_2$. Maximum likelihood estimates of the reduced form parameters $\hat{\alpha}$ and $\hat{\beta}$ can be estimated by introducing the identifying restriction $\sigma_v = 1$. For only determining the bias correcting factor, i.e. the Mills ratio, we certainly do not need this restriction. The estimators $\hat{\alpha}$ and $\hat{\beta}$ must fulfill the following restriction

$$(7) \qquad \frac{1}{N} \sum_{i=1}^{N} m(\tilde{x}_i; \delta) = \frac{1}{N} \sum_{i=1}^{N} \frac{\zeta_i - F\left(\tilde{x}_i^T \delta\right)}{F\left(\tilde{x}_i^T \delta\right)\left(1 - F\left(\tilde{x}_i^T \delta\right)\right)} f\left(\tilde{x}_i^T \delta\right) \tilde{x}_i = 0 \,,$$

where $\tilde{x}_i^T = \begin{pmatrix} x_i^T & z_i^T \end{pmatrix}$, $\delta = \begin{pmatrix} \alpha^T & \beta^T \end{pmatrix}^T$ and $f(\bullet)$ stands for the gaussian density.

In the second step, the log wage equation is estimated by least square methods correcting the sample selection bias by the Mill's ratio. More exactly, the selection bias corrected equation will be

$$(8) \qquad w_i = z_i^T \gamma_1 + \gamma_2 \lambda\left(z_i^T \alpha + x_i^T \beta\right) + v_{2i}, \quad i = 1, \cdots, M.$$

Here, v_{2i} is the error in the conditional equation. The vector estimate $\hat{\gamma} = \begin{pmatrix} \hat{\gamma}_1^T & \hat{\gamma}_2 \end{pmatrix}^T$ must fulfill the condition

$$(9) \qquad \frac{1}{N} \sum_{i=1}^{N} g\left(\tilde{z}_i; \hat{\delta}, \gamma\right) = -\frac{2}{N} \sum_{i=1}^{N} \zeta_i \tilde{z}_i \left(w_i - \tilde{z}_i^T \gamma\right) = 0 \,,$$

where $\tilde{z}_i^T = \left[z_i^T \quad \lambda\left(\tilde{x}_i^T \hat{\delta}\right)\right]$ and $\lambda(\bullet) = f(\bullet)/F(\bullet)$ is the inverse of Mill's ratio.

In the third step the structural parameters of the hours equation are estimated. In order to do this, we construct the predicted wages unconditionally for all individuals in the sample

$$(10) \qquad \hat{w}_i = z_i^T \hat{\gamma}_1 \qquad i = 1, \cdots, N.$$

Recall that $\hat{\gamma}_1, \hat{\gamma}$ are either O.L.S. or feasible G.L.S. of the log wage equation. Then, by substituting the predicted wages in the hours equation (1) it is possible to estimate the structural parameters by tobit maximum likelihood. For this unconditional predicted wages the likelihood function has the expression

$$(11) \qquad L(\beta_w, \beta) = \prod_{i=1}^{M} \frac{1}{\sigma_1} f\left(\frac{y_i - \beta_w \hat{w}_i - x_i^T \beta}{\sigma_1}\right) \prod_{i=M+1}^{N} \left(1 - F\left(\frac{\beta_w \hat{w}_i + x_i^T \beta}{\sigma_1}\right)\right).$$

If the structural model is recursive ($\rho = 0$), then the predicted wages for participants are the observed ones, and for nonparticipants, the predicted ones. The problem is that if $\rho \neq 0$, then $\hat{w}$ is endogeneous and therefore the tobit maximum likelihood estimators are inconsistent. This is also the problem when the predicted wages are generated conditionally by using the inverse of Mill's ratio. In this case, the estimators of the hours equation are also inconsistent since the criteria that determines the truncation of both structural equations is the same. The estimators derived from the maximization of (11) respectively its logarithm will fulfill the following equation condition, see Olsen (1978),

$$(12)\quad \frac{1}{N}\sum_{i=1}^{N} h\left(\mathbf{x}_i; \hat{\delta}, \hat{\gamma}, \tau, \sigma_1\right) = \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \zeta_i \left(\frac{1}{\sigma_1} y_i - \tau^T \mathbf{x}_i\right) \mathbf{x}_i - (1-\zeta_i) \frac{f\left(\mathbf{x}_i^T \tau\right)}{1-F\left(\mathbf{x}_i^T \tau\right)} \mathbf{x}_i \\ \zeta_i \left\{\sigma_1 - \left(\frac{1}{\sigma_1} y_i - \tau^T \mathbf{x}_i\right) y_i\right\} \end{pmatrix}$$

where $\mathbf{x}_i^T = \left[z_i^T \hat{\gamma}_1 \quad x_i^T\right]$ and $\tau = \begin{pmatrix} \frac{\beta_w}{\sigma_1} & \left(\frac{\beta}{\sigma_1}\right)^T \end{pmatrix}^T$.

As it was remarked before, in the presence of simultaneity, tobit maximum likelihood estimates are consistent when real wages are replaced by unconditionally predicted wages. However, the usual standard errors obtained from the tobit equation are not appropriated. The reason is that we have to take into account the parameter estimates from the previous steps. Whereas statistical properties of two step estimators are well known in the literature of labor supply functions (see Vella, 1998), very little is known about predicted wage methods. In what follows, under very general assumptions, we show the consistency and asymptotic normality of the three step estimator based on predicted wages. Moreover, we provide a consistent estimator of the asymptotic variance-covariance matrix. Before giving the theoretical results, we will introduce some notation. Let

$$m(x,z) = m(x,z;\delta_0), \quad g(x,z) = g(x,z;\delta_0,\gamma_0), \quad h(x,z) = h(x,z;\delta_0,\gamma_0,\tau_0),$$

$$M_\delta = E\left[\nabla_\delta m(x,z)\right], \quad G_\delta = E\left[\nabla_\delta g(x,z)\right], \quad G_\gamma = E\left[\nabla_\gamma g(x,z)\right],$$

$$H_\gamma = E\left[\nabla_\gamma h(x,z)\right], \quad H_\tau = E\left[\nabla_\tau h(x,z)\right], \quad \text{and } \psi(x,z) = -M_\delta^{-1} m(x,z).$$

Theorem 1 *Assume conditions (A.1) to (A.6) hold, then*

$$\sqrt{N}\left(\hat{\tau} - \tau_0\right) \xrightarrow{D} N\left(0, V\left(\tau_0\right)\right)$$

where

$$\begin{aligned} V(\tau_0) &= H_\tau^{-1} E\left[h(x,z) + H_\gamma G_\gamma^{-1}\left(g(x,z) + G_\delta \psi(x,z)\right)\right] \\ &\quad \times \left[h(x,z) + H_\gamma G_\gamma^{-1}\left(g(x,z) + G_\delta \psi(x,z)\right)\right]^T H_\tau^{-1}. \end{aligned}$$

Moreover, if

$$\hat{V}(\hat{\tau}) = \hat{B}_N(x_i,z_i)^{-1} \hat{A}_N(x_i,z_i) \hat{B}_N(x_i,z_i)^{-1},$$

where

$$\hat{A}_N(x_i, z_i) = \frac{1}{N}\sum_{i=1}^{N}\left\{h\left(x_i, z_i, \hat{\delta}, \hat{\gamma}, \hat{\tau}\right) + \hat{A}_{1N}(x_i, z_i)\right\}\left\{h\left(x_i, z_i, \hat{\delta}, \hat{\gamma}, \hat{\tau}\right) + \hat{A}_{1N}(x_i, z_i)\right\}^T,$$

$$\hat{A}_{1N}(x_i, z_i) = \left[\frac{1}{N}\sum_{i=1}^{N}\nabla_\gamma g\left(x_i, z_i, \hat{\delta}, \hat{\gamma}\right)\right]^{-1}\left\{g\left(x_i, z_i, \hat{\delta}, \hat{\gamma}\right) + \left[\frac{1}{N}\sum_{i=1}^{N}\nabla_\delta g\left(x_i, z_i, \hat{\delta}, \hat{\gamma}\right)\right]\hat{A}_{2N}(x_i, z_i)\right\},$$

$$\hat{A}_{2N}(x_i, z_i) = -\left[\frac{1}{N}\sum_{i=1}^{N}\nabla_\delta m\left(x_i, z_i, \hat{\delta}\right)\right]^{-1} m\left(x_i, z_i; \hat{\delta}\right)$$

and

$$\hat{B}_N(x_i, z_i) = \frac{1}{N}\sum_{i=1}^{N}\nabla_\gamma h\left(x_i, z_i, \hat{\delta}, \hat{\gamma}, \hat{\tau}\right),$$

then

$$\hat{V}(\hat{\tau}) \xrightarrow{P} V(\tau_0)$$

as N tends to infinity.

In order to estimate the parameters that are contained in the structural model described in (1) and (2) it has been necessary to introduce some additional restrictions that were considered in assumptions (A.1) to (A.6). As it was pointed out in Mroz (1987), some of these restrictions are hardly supported by economic theory and they are only introduced on the grounds of convenience. Among them, distributional assumptions and linear prespecified relationships between different variables appear to be the assumptions that present the weakest justification. The impact of misspecification on distributional assumptions in this type of models has been studied in depth by Vijverberg (1991), and Blundell and Meghir (1986) discuss some models proposed by economic theory that do not imply linearity between hours of work, log-wages and other explanatory variables of the hours equation. In the next Section, we will relax some of the previous assumptions and we will provide alternative methods to estimate the structural model of labor supply defined by equations (1) and (2).

3 A Semiparametric Approach to the Heckman Estimator

In this section we are particularly interested in relaxing assumption (A.4). That is, instead of imposing linear relationships among the different variables, we will keep additivity but allowing for unknown functional forms in the relationship between these variables. Assumption (A.6) concerning distributional assumptions is not relaxed in this work. One reason is, when dealing with empirical research, the normality assumption does not appear to be a very restrictive one (see Fernández and Rodríguez-Póo, 1997), in other words, a misspecification of the error distribution turns does by far not effect the estimation results inferences as a

misspecification of the functional forms. Additionally, for estimating a generalized additive partial linear model some distribution assumptions are needed in practice anyway.

Since we handle with nested equations, nonparametric and even nonlinear modelling can cause serious identification problems. Remember that in the first step we have to estimate

$$E[\zeta|X=x,Z=z]=F[l_1(w,x)/\sigma_v]=F[l_1\{l_2(z)+u_{2i},x\}/\sigma_v]\ , \tag{13}$$

with l_1, l_2 arbitrary smooth functions and x, z not necessarily different in all entrances. Fortunately, in the first step we need only the index to be estimated and not explicitly l_1, l_2. So they are mainly three ways to proceed: First, replacing the wages w by their expectations to get rid of u_{i2} and estimating the index with a multidimensional nonparametric smoother; second, replacing w in step one by instrumental variables arguing that the wage would be endogenous here as Ahn and Powell (1993) or Mroz (1987) do; or third, to impose linearity for the influence of wages in l_1. Note that from a statistical point of view the second and third version result in the same estimation procedures and differ only in notation. This will be seen better inside of Section 3.1. In the next section we consider respectively linearity and non-linearity in wages.

3.1 Linearity in wages

Following the previous reasoning, $l_1(w,x)$ and $l_2(z)$ are alternatively specified as

$$\begin{aligned} l_1(w_i,x_i) &= \theta+\beta_w w_i+\beta^T x_i^d+\textstyle\sum_{j=1}^d \eta_j(x_{ij}^c), \\ l_2(z_i) &= \gamma_0+z_i^T\gamma_1 \end{aligned} \tag{14}$$

The η_j's are unknown functions that need to be estimated jointly with the parameters θ, β_w and the vectors β and γ_1. For reasons we discuss later, we put and treat this time the intercept (γ_0) in the log-wage equation explicitly. Note that the functional relationship between hours and log wages is kept linear. This will be relaxed in Section 3.2, in the paper. Moreover, the wage equation is also assumed to be linear. This is done without loss of generality. In fact, $l_2(\bullet)$ could be taken to be nonparametric, and this would not affect the results stated in the paper.

Taking into account the structural model of labor supply that was introduced in the previous section (equations (1) and (2)), now, we re-develop the predicted wage estimation method considering the model that was introduced in equation (14). To this end, we will distinguish three subsections that will describe each of the steps that were previously introduced for the fully parametric model.

3.1.1 First step: Estimation of the reduced form parameters

This subsection is devoted to the estimation of the reduced form parameters of the hours equation. In order to do this, in (14), if we substitute the wage equation (2) into the hours

equation (1) and using the normality assumption then the reduced form model is

$$E[\zeta|Z=z,X=x]=F\left\{\frac{\beta^T x^d+\theta+z^T\alpha+\sum_{j=1}^{d}\eta_j(x_j^c)}{\sigma_v}\right\}. \tag{15}$$

Where x^c is a vector of continuous variables and x^d is a vector of discrete variables. This expression falls within the class of the so called Generalized Additive Partial Linear Model (GAPLM), where F is the cumulative normal distribution function. In analogy to the parametric case we take it as being the gaussian normal distribution function. Again, under the identifying condition $\sigma_v = 1$ and all elements of z different from x, we could estimate β, $\alpha = \beta_w\gamma_1$, and the η_j's uniquely, but do not need this restriction to get the inverse of Mill' s ratio for the second step. Nevertheless, for the ease of notation we set $\sigma_v = 1$.

The parameters β, θ and α, and the additive components $\{\eta_j(x^c)\}_{j=1}^{d}$ are estimated by a method proposed by Härdle, Huet, Mammen and Sperlich (1998). To make clear that there can be an overlapping of entrances in z and x, in equation (15) we rewrite $\beta^T x^d+\sum_{j=1}^{d}\eta_j(x_j^c)$ as $\beta^T v^d+\sum_{l=1}\psi_l(v_l^c)$. The number of the additive components is between $\max\{d,r\}$ and $(d+r)$, let us call it s. For the dummies, now x^d and z^d together in v^d, we keep the notation $\beta\in I\!R^p$.

So far known nonparametric estimation procedures for these models are the backfitting algorithm, see Hastie and Tibshirani (1990) and the so-called marginal integration estimator (MIE), see Tjøstheim and Auestad (1994) or Linton and Nielsen (1995). To estimate a particular additive component ψ_k they use a multidimensional preliminar estimator and then integrate out all covariates v_j except v_k.

To get the multidimensional pre-estimate of $\psi(\cdot) := \theta+\sum_{j=1}^{s}\psi_j(\cdot)$, we use the method of Severini and Staniswalis (1994) which considered a model of the form (15). Their approach is based on an iterative application of smoothed local and un-smoothed global likelihood functions. In particular, this method allows for a $\sqrt{N}$-consistent estimation of the parametric component. Afterwards, we apply the integration idea on ψ to obtain estimates for all ψ_j (see Härdle, Huet, Mammen and Sperlich, 1998).

The main reason why we have chosen in our application the MIE is that no asymptotic theory for GAPLM with the backfitting algorithm has been developed so far. Note that the MIE is indeed always estimating the marginal effect of the particular input variable, even when the assumption of additivity is violated. In contrast, the backfitting is looking for an optimal fit in the space of additive models of the regression problem and thus has a different interpretation (cf. Sperlich, Linton, Härdle, 1997).

The Estimation Procedure: Depending on what kind of participation variable ζ we consider, we have a density or a mass point (probability) function $f(\zeta,\mu)$, or $f(Y,\mu)$, where the goal now is to estimate the unknown parameter μ, in the case of our first step with

$\mu_i = F\{\beta^T v_i^d + \theta + \sum_{j=1}^s \psi_j(v_{ij}^c)\}$. Then the (conditional) log-likelihood we consider is

$$\mathcal{L}(\psi, \beta) = \sum_{i=1}^N \ln f(\zeta_i, \mu_i) \ . \tag{16}$$

E.g. for the binary case, as we have it in the first step, this is given by

$$\sum_{i=1}^N \zeta_i \log \mu_i + (1 - \zeta_i) \log(1 - \mu_i) \ . \tag{17}$$

Without loss of generality, we describe now how to estimate the component ψ_1. For a vector $u \in \mathbb{R}^s$ we denote the vector $(u_2, \ldots, u_s)^T$ by $u_{\underline{1}}$, respectively $v_{i\underline{1}}^c = (v_{i2}^c, \ldots, v_{is}^c)^T$. Further, for a kernel function L defined on $\mathbb{R}^{s-1}$ we put $L_g(\cdot) = g^{-(s-1)} L(g^{-1} \cdot)$ and for a kernel function K defined on $\mathbb{R}$ we put $K_h(\cdot) = h^{-1} K(h^{-1} \cdot)$. For L we take the product kernel $L = \prod_{j=2}^s L_j$. The bandwidth g is related to the smoothing in direction of the nuisance covariates, the bandwidth h to the direction of interest (here direction 1). We also make use of the smoothed likelihood defined by

$$\mathcal{L}^S(\psi, \beta) = \int \sum_{i=1}^N K_h(v_1^c - v_{i1}^c) L_g(v_{\underline{1}}^c - v_{i\underline{1}}^c) \ln f \left[\zeta_i, F\left\{v_i^{dT}\beta + \psi^+(v^c)\right\}\right] dv. \tag{18}$$

A good introduction to all these considerations can be found in Staniswalis (1989). Furthermore, following Severini and Staniswalis (1994) we put for $\beta \in B \subset \mathbb{R}^p$:

$$\widehat{\psi}_\beta(v^c) = \arg\min_\psi \sum_{i=1}^N K_{h_1}(v_1^c - v_{i1}^c) L_{h_2}(v_{\underline{1}}^c - v_{i\underline{1}}^c) \ln f \left[\zeta_i, F\{v_i^{dT}\beta + \psi\}\right], \tag{19}$$

$$\widehat{\beta} = \arg\min_{\beta \in B} \mathcal{L}(\widehat{\psi}_\beta, \beta), \tag{20}$$

$$\widehat{\psi} = \widehat{\psi}_{\widehat{\beta}} \ . \tag{21}$$

We remark that $\widehat{\psi}$ is a multivariate kernel estimate of ψ which makes no use of the additive structure but serves as a pre-estimate as mentioned above.

We now apply the marginal integration method. Because of the identifiability conditions, $\psi_1(v_1^c)$ is, up to a constant, equal to $\int w_{\underline{1}}(u)\psi(v_1^c, u)du$, where $w_{\underline{1}}$ is any weight function verifying $\int w_{\underline{1}}(u)du = 1$. In other words, integrating out of the multidimensional function ψ all the nuisance directions $(v_{\underline{1}}^c)$, is giving you, up to a constant, the marginal influence of v_i^c, that is ψ_1. For details see one of the abovementioned papers.

Thus, we put

$$\tilde{\psi}_1(v_1^c) = \frac{\frac{1}{N}\sum_{i=1}^N w_{\underline{1}}(v_{i\underline{1}}^c)\widehat{\psi}(v_1^c, v_{i\underline{1}}^c)}{\frac{1}{N}\sum_{i=1}^N w_{\underline{1}}(v_{i\underline{1}}^c)} \tag{22}$$

and by centering $\tilde{\psi}_1$ to zero, we get an estimate for ψ_1. Here again, introduction of a weight function w_1 may be useful to avoid problems at the boundary. The additive constant θ is simply estimated by

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^N \widehat{\psi}(v_i^c). \tag{23}$$

In the paper of Härdle, Huet, Mammen and Sperlich (1998) under the set of assumptions B, that are included in Appendix I, the consistency for these estimators (including $\hat{\beta}$) is proved and the asymptotic distribution is developed.

3.1.2 Second step: Semiparametric estimation

In this step, our aim is to estimate the structural parameters of the wage equation. To this end, recall that under the previous assumptions then the mean of the wages conditional on the explanatory variables for those individuals who work is

$$E[W|\zeta = 1, Z = z_i, X = x_i] = \gamma_0 + \gamma_1^T z_i + \gamma_2 \lambda \left\{ \beta^T v_i^d + \theta + \sum_{j=1}^{s} \psi_j(v_{ij}^c) \right\}, \qquad i = 1, \cdots, M,$$

where λ stands for the inverse of Mill's ratio. Recall that as it was indicated in the previous section, to make clear that there can be an overlapping of entrances in z and x, in equation (15) we rewrite $\beta^T x^d + \sum_{j=1}^{d} \eta_j(x_j^c)$ as $\beta^T v^d + \sum_{l=1} \psi_l(v_l^c)$. The number of the additive components is between $\max\{d, r\}$ and $(d+r)$, let us call it s. For the dummies, now x^d and z^d together in v^d, we keep the notation $\beta \in I\!R^p$. Then,

$$(24) \qquad w_i = \gamma_0 + \gamma_1^T z_i + \gamma_2 \lambda \left\{ \beta^T v_i^d + \theta + \sum_{j=1}^{s} \psi_j(v_{ij}^c) \right\} + \epsilon_i, \qquad i = 1, \cdots, M,$$

where $\epsilon_i = w_i - E[W|\zeta = 1, Z = z_i, X = x_i]$. In the previous equation, the estimation of the parameter vector $\gamma^T = (\gamma_0 \quad \gamma_1^T \quad \gamma_2)^T$ is infeasible since the parameters of the index, β and $\{\psi_l(\bullet)\}_{j=1}^{s}$ are unknown. Remember that in this case, β stands indeed for the parameter vector β assuming the normalization $\sigma_v = 1$. Heckman (1979) proposed, in a fully parametric setting, to replace them by consistent estimators. Proceeding in the same way then obtain the following regression equation

$$(25) \qquad w_i = \gamma_0 + \gamma_1^T z_i + \gamma_2 \lambda \left\{ \hat{\beta}^T v_i^d + \hat{\theta} + \sum_{j=1}^{s} \hat{\psi}_j(v_{ij}^c) \right\} + \epsilon_i, \qquad i = 1, \cdots, M.$$

The parameter vector $(\gamma_0 \quad \gamma_1^T \quad \gamma_2)^T$ can be estimated by ordinary least squares. However, this estimation procedure does not provide satisfactory rates of convergence for the estimators. This is due to the nonparametric part that is plugged into the inverse of the Mill's ratio. In fact, it can be shown that the o.l.s. estimators of the parameters of the wage equation are consistent and asymptotically normal, but they heritate from the nonparametric estimator of the first step its rate. This rate, of course is optimal for the nonparametric function, but it is suboptimal for parametric estimators. In order to obtain root-n consistent estimators of the structural parameters of the wage equation we propose to use the differencing estimator proposed by Powell (1987) and Ahn and Powell (1993). Then, the parameters γ_1 in the structural wage equation can be estimated as follows,

$$(26) \qquad \hat{\gamma}_1 = \hat{M}_{sz}^{-1} \hat{M}_{sw}$$

where

$$\hat{M}_{sz} = \binom{n}{2}^{-1} \sum_{i=1}^{N} \sum_{j=i+1}^{N-1} \hat{\omega}_{ij} (s_i - s_j)(z_i - z_j)^T, \tag{27}$$

$$\hat{M}_{sw} = \binom{n}{2}^{-1} \sum_{i=1}^{N} \sum_{j=i+1}^{N-1} \hat{\omega}_{ij} (s_i - s_j)(w_i - w_j), \tag{28}$$

and we define the weights for $i, j = 1, \cdots, N$ as

$$\hat{\omega}_{ij} = \frac{1}{h_2} K_2 \left(\frac{\lambda\left\{\hat{\beta}^T v_i^d + \hat{\theta} + \sum_{k=1}^{s} \hat{\psi}_k(v_{ik}^c)\right\} - \lambda\left\{\hat{\beta}^T v_j^d + \hat{\theta} + \sum_{k=1}^{s} \hat{\psi}_k(v_{jk}^c)\right\}}{h_2} \right) \zeta_i \zeta_j$$

with kernel function $K_2(\bullet)$ and bandwidth h_2. Finally, we define the instruments by

$$s_i = s(x_i, z_i), \qquad \text{for some function} \qquad s : \mathbb{R}^{p+d+r} \to \mathbb{R}^r .$$

The estimation of γ_0 and γ_2 has been neglected by Powell (1987), Ahn and Powell (1993), neither a proof for the consistency and rate of $\hat{\gamma}_1$ is provided. The treatment of γ_0 and γ_2 we consider in Section 4 in the context of our application. In the second section of the Appendix we present a detailed proof for the asymptotic result for $\hat{\gamma}_1$. Further, for the sake of simplicity, also the assumptions needed to claim this result have been relegated to the the Appendix, first part. The asymptotics of our estimator are the following:

Theorem 2 *Let us define the index as* $\nu_i = v_i^{dT}\beta + \theta + \sum_{k=1}^{s} \psi(v_{ik}^c)$. *Then, under assumptions (B.1) to (B.12) and if the sequence of bandwidths* h, h_2 *and* g *tends to zero such that* $hg^{d-1}N^{1/2}(\log N)^{-1} \to \infty$, $Nh_2^6 \to \infty$ *and* $Nh_2^8 \to 0$ *then*

$$\sqrt{N}(\hat{\gamma}_1 - \gamma_1) = \Phi_{sz}^{-1} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ f(\nu_i) F(\nu_i) \zeta_i \left[s_i - \frac{E[\zeta s|\nu_i]}{F(\nu_i)} \right] u_{2i} \right.\right. \tag{29}$$
$$\left.\left. + \left\{ A_{1i}\tilde{v}_i^d + A_{2i}\kappa_i(v_i^c) \right\} \frac{\zeta_i - F(\nu_i)}{F(\nu_i)(1 - F(\nu_i))} f(\nu_i) \right\} \right\} + o_p(1)$$

where Φ_{sz} *is defined in assumption (B.12).*

$$A_{1i} = E\left\{ \lambda'(\nu_i) F(\nu_i)^2 f(\nu_i) \left[s_i - \tfrac{E[\zeta s|\nu_i]}{F(\nu_i)} \right] \right.$$
$$\left. \times \left[w_i - \{E[T_i^2|v_i^c]\}^{-1} E\left[T_i^2 v_i^d | v_i^c\right] - E\left[w_i - \{E[T_i^2|v_i^c]\}^{-1} E\left[T_i^2 v_i^d|v_i^c\right] \right] \right]^T \right\},$$

$$A_{2i} = E\left\{ \lambda'(\nu_i) \left\{ E\left[T_i^2|v_i^c\right] \right\}^{-1} F(\nu_i)^2 \left[s_i - \frac{E[\zeta s|\nu_i]}{F(\nu_i)} \right] \right\}$$

$$\tilde{v}_i^d = v_i^d - \left\{ E\left[T_i^2|v_i^c\right] \right\}^{-1} E\left[T_i^2 v_i^d|v_i^c\right] \qquad i = 1, \cdots, N,$$

$$T_i^2 = \frac{f(\nu_i)^2}{F(\nu_i)(1 - F(\nu_i))} \qquad i = 1, \cdots, N$$

$$\kappa_i(v^c) = \frac{K_h(v_1^c - v_{i1}^c) L_g(v_{\underline{1}}^c - v_{i\underline{1}}^c)}{\frac{1}{n}\sum_{j=1}^{n} K_h(v_1^c - v_{i1}^c) L_g(v_{\underline{1}}^c - v_{i\underline{1}}^c)} \qquad i = 1, \cdots, N \tag{30}$$

as N *tends to infinity.*

Finally, we want to mention again, that this second step also could be done non- or semi-parametrically allowing for arbitrary functional forms in equation (2). This would not cause any problems of identification as discussed above but results as in Theorem 2 would be much harder to derive.

3.1.3 Third step: Tobit Local Maximum Likelihood Estimation

Here, we can apply a quite similar procedure as in the first step. We replace all wages with the predicted ones obtained in the second step: $\hat{w}_i = \hat{\gamma}_0 + z_i^T \hat{\gamma}_1$ for $i = 1, \ldots, N$. The goal is to estimate the conditional expectation of Y, that is

$$E[Y|W = \hat{w}, X = x] = F\left\{\frac{\beta^T x^d + \theta + \sum_{j=1}^{d} \eta_j(x_j^c) + \beta_w \hat{w}}{\sigma_1}\right\},$$

where σ_1 is the standard deviation of u_1.

In general, this can again be estimated by the (smoothed) maximum likelihood method presented in Section 3.1.1 following Staniswalis (1989) and Härdle, Huet, Mammen, Sperlich (1999). E.g. if the participation variable Y as presented in equation (1) is the number of working hours, we have a censured variable and thus to estimate a tobit model, compare equation (11) for the parametric analogon. Since to our knowledge this has presented never before explicitly nor implemented, we give the estimation procedure here and a possible numerical algorithm for implementation in Appendix III.

The likelihood function in this model is

$$\begin{aligned}\ln L \;=\; & \frac{-\ln(2\pi)}{2} + \frac{-\ln(\sigma_1^2)}{2} - \frac{\{y - x^{dT}\beta + \alpha + \sum_{j=1}^{d} \eta_j(x_j^c) + \beta_w \hat{w}\}^2}{2\sigma_1^2} \\ & - \ln\left\{1 - F\left(\frac{-(x^{dT}\beta + \alpha + \sum_{j=1}^{d} \eta_j(x_j^c) + \beta_w \hat{w})}{\sigma_1}\right)\right\}\end{aligned}$$

Let η be $\alpha + \sum_{j=1}^{d} \eta_j(x_j^c)$. We estimate β, β_w, σ_1 and η in the following way:

1. $\hat{\eta}_{\beta,\beta_w,\sigma_1} = \arg\max_{\eta} \mathcal{L}^S(\eta, \beta, \beta_w, \sigma_1)$,
2. $(\beta^T, \widehat{\beta_w}, \sigma_1) = \arg\max_{\beta,\beta_w,\sigma_1} \mathcal{L}(\hat{\eta}_{\beta,\beta_w,\sigma_1}, \beta, \sigma_1)$,
3. $\hat{\eta} = \hat{\eta}_{\hat{\beta},\hat{\beta}_w,\hat{\sigma}_1}$,

where

$$\begin{aligned}\mathcal{L}(\eta, \beta, \beta_w, \sigma_1) \;=\; & \sum_{i=1}^{N} \frac{-\ln(2\pi)}{2} + \frac{-\ln(\sigma_1^2)}{2} - \frac{\{Y_i - \beta^T x_i^d - \eta(x_i^c) - \beta_w \hat{w}\}^2}{2\sigma_1^2} \\ & - \ln\left\{1 - F\left(\frac{-\beta^T x_i^d - \eta(x_i^c) - \beta_w \hat{w}}{\sigma_1}\right)\right\},\end{aligned}$$

$$\mathcal{L}^S(\eta(\cdot),\beta,\beta_w,\sigma_1) = \int \sum_{i=1}^{N} \left[\frac{-\ln(2\pi)}{2} + \frac{-\ln(\sigma_1^2)}{2} - \frac{\{Y_i - \beta^T x_i^d - \eta(t) - \beta_w \hat{w}\}^2}{2\sigma_1^2} \right.$$
$$\left. - \ln\left\{1 - F\left(\frac{-\beta^T x_i^d - \eta(t) - \beta_w \hat{w}}{\sigma_1}\right)\right\}\right] K_h(t_1 - t_{i1}) L_g(t_{\underline{1}} - t_{i\underline{1}}) dt.$$

Here, again $t^T = (t_1, t_{\underline{1}}^T)$.

The following result is also shown in the Appendix for $\hat{\beta}$ and $\hat{\beta}_w$.

Theorem 3 *Under the assumptions (B.1) to (B.12) and the rates on h and g considered in Theorem 2, then*

$$\hat{\beta} = \beta + o_p(1)$$
$$\hat{\beta}_w = \beta_w + o_p(1)$$

as N tends to infinity.

3.2 Nonlinearity in wages

Finally, we analyze the case when the structural hours equation presents a nonlinear relationship between hour and log of wages. This case, has been motivated for example by Blundell and Meghir (1986) and it can be taken into account by assuming the following nonparametric additive relationship

$$\begin{aligned} l_1(w_i^*, x_i) &= \theta + \beta^T x^d + \textstyle\sum_{j=1}^{d} \eta_j(x_j^c) + \eta_{d+1}(w_i^*), \\ l_2(z_i) &= \gamma_0 + z_i^T \gamma_1 \end{aligned} \tag{31}$$

The η_j's are unknown functions that need to be estimated jointly with the parameters θ and the vectors β and γ_1. Note that the difference with respect to the linear specification in wages (see equation (14)) is twofold. First, from a theoretical point of view, individual log wages are replaced by 'lifetime' wage rates, $\{w^*\}$. This represents a life-cycle average of single period wage rates (see Killingsworth and Heckman, 1986). For further discussion see also the application in Section 4. Second, the hours of work and *average wages* are related through an unspecified nonlinear relationship, $\eta_{d+1}(\bullet)$, whereas in equation (14)) a linear parametric relationship was assumed. Again, the wage equation is assumed to be linear and parametric for the sake of simplicity.

Although under a nonlinear relationship between hours and log wages the three step method estimation procedure that has been proposed in this paper remains valid it is necessary to make some changes in the estimation stages. For the first step, recall that under the specification that was introduced in (31) and the normality assumption for the error terms, the reduced form model for the hours equation is

$$E(\zeta | Z = z, X = x) = F\left(\frac{\eta(z,x)}{\sigma_v}\right) \tag{32}$$

where $F(\bullet)$ is the c.n.d.f., σ_v a scaling factor, and η is a multivariate unknown function. Here, we could separate now η depending on the intersection between the elements of x and z; we can separate, given that the data generating process is like this, exactly that parts which affect elements of x which are not in z.

A different solution has been chosen in Ahn and Powell (1993). They argue that the wage can be considered in the first step as being endogeneous and be replaced by some instrumental variables. However, they did not discuss the rising problems of identification if this seems inappropriate to the empirical researcher. In that case, we give up the idea of perfectly nested equations and thus instead of straight forward plugging in one equation into the other we only try to approximate the values needed for the bias correction in the second step, see discussion in Section 2. Following their arguments, the reduced form model for the hours equation will be

$$(33) \qquad E(\zeta|Z=z,X=x) \;=\; F\left(\frac{\beta^T x^d+\theta+\sum_{j=1}^{d}\eta_j(x_j^c)+\delta^T z^d+\sum_{j=1}^{r}\varphi_j(z^c)}{\sigma_v}\right)$$

where $F(\bullet)$ is the c.n.d.f., σ_v a scaling factor, and η_j, φ_j are unknown functions from $\mathbb{R}$. At least one element of z is different from the elements of x.

For estimation purposes, the three step estimation method changes depending on whether (32) or (33) are chosen to be the reduced form model. In the second case, the steps and properties developed in Section 3.1 apply. For the first case, a direct application of the local likelihood method proposed by Staniswalis (1989) can be used for the first step: the Pre-estimation for the selection bias correction, i.e.

$$\begin{aligned}\hat{\eta}(x_0,z_0) \;=\; & \arg\max_{\eta}\sum_{i=1}^{N}K\left(\frac{x_0-x_i}{h}\right)K\left(\frac{z_0-z_i}{h}\right)\times \\ & \left[\zeta_i \ln F\left(\frac{\eta(x_0,z_0)}{\sigma_v}\right)+(1-\zeta_i)\ln\left\{1-F\left(\frac{\eta(x_0,z_0)}{\sigma_v}\right)\right\}\right],\end{aligned}$$

where bandwidths h and g not refer to the one chosen before but have to be chosen accordingly this new smoothing problem. For the second step, the semiparametric estimator developed in Section 3.1.2 can be used. The regression equation is now

$$(34) \qquad w_i=\gamma_0+\gamma_1^T z_i+\gamma_2\lambda\{\hat{\eta}(x_i,z_i)\}+v_{2i}, \qquad i=1,\ldots,M$$

and the parameter vector γ_1 can be estimated by the differencing estimator proposed in (26). i.e.

$$(35) \qquad \hat{\gamma}_1=\hat{M}_{sz}^{-1}\hat{M}_{sw}$$

where now

$$(36) \qquad \hat{M}_{sz}=\binom{n}{2}^{-1}\sum_{i=1}^{N}\sum_{j=i+1}^{N-1}\hat{\omega}_{ij}^{*}(s_i-s_j)(z_i-z_j)^T,$$

$$(37) \qquad \hat{M}_{sw}=\binom{n}{2}^{-1}\sum_{i=1}^{N}\sum_{j=i+1}^{N-1}\hat{\omega}_{ij}^{*}(s_i-s_j)(w_i-w_j),$$

and we define the weights for $i, j = 1, \cdots, N$ as

$$\widehat{\omega}_{ij}^{*} = \frac{1}{h_2} K_2 \left(\frac{\lambda\{\hat{\eta}(x_i, z_i)\} - \lambda\{\hat{\eta}(x_j, z_j)\}}{h_2} \right) \zeta_i \zeta_j$$

with kernel function $K_2(\bullet)$ and bandwidth h_2.
In the third step, taking $\hat{w}_i = \hat{\gamma}_1^T z_i$ for $i = 1, \ldots, N$ we estimate

$$E[Y|W = \hat{w}, X = x] = F\left\{ \frac{x^{d^T}\beta + \theta + \sum_{j=1}^{d} \eta_j(x_j^c) + \eta_{d+1}(\hat{w})}{\sigma_1} \right\}$$

using the same method as in Section 3.1 but with likelihood function

$$\begin{aligned} \ln L_i \;=\; & \frac{-\ln(2\pi)}{2} + \frac{-\ln(\sigma_1^2)}{2} - \frac{\{y_i - x_i^{d^T}\beta + \theta + \sum_{j=1}^{d} \eta_j(x_{ij}^c) + \eta_{d+1}(\hat{w}_i^*)\}^2}{2\sigma_1^2} \\ & - \ln\left\{ 1 - F\left(\frac{-(x_i^{d^T}\beta + \theta + \sum_{j=1}^{d} \eta_j(x_{ij}^c) + \eta_{d+1}(\hat{w}_i^*))}{\sigma_1} \right) \right\}. \end{aligned}$$

We conjecture that the results presented in Section 3.1, can be also claimed for the estimators developed in this section. However, to avoid a too long paper we do not present them as formal results.

4 The Application

4.1 Model and Data

The source of the data is the *Encuesta de Población Activa (EPA)* , the Spanish Labor Force Surveys. These surveys have been carried out on a quarterly basis since 1975 and are collected by the National Bureau of Statistics (INE). They cover approximately 60,000 households and contain information about 150,000 individuals that are older than 16 years. It provides information at different levels of disaggregation both at national and regional level. From these surveys, in the second quarter of 1990, the National Bureau of Statistics randomly selected a cross-section of 4,989 individuals and additional information about some variables that were considered relevant for labor market participation analysis were provided. In this paper we consider a subsample of 1010 individuals participating in the labor market, 612 workers and 398 non workers.

The variables included in this data set are defined in Table 1 including some basic statistics. Further, we certainly have the information whether a person has a job (JOB) or not.

In this application we are considering the problem of estimating the conditional expectation of being employed. As discussed in the introduction, we have two problems; for including the wages we have to predict them for non workers, and additionally, estimating the wage

Variable	Description	Whole Sample	Workers
SEXM	dummy, 1 if male	0.680 (0.466)	0.625 (0.484)
AGE1	dummy, age 16 to 19	0.131 (0.338)	0.111 (0.314)
AGE2	dummy, age 20 to 25	0.265 (0.441)	0.256 (0.437)
AGE3	dummy, age 26 to 35	0.278 (0.448)	0.261 (0.439)
AGE4	dummy, older than 45	0.138 (0.345)	0.143 (0.351)
EDUC1	elementary school	0.350 (0.477)	0.339 (0.474)
EDUC2	high school	0.115 (0.320)	0.106 (0.308)
EDUC3	university	0.064 (0.245)	0.039 (0.194)
URATE	unemployment rate of the district	0.171 (0.069)	0.171 (0.071)
SINGLE	dummy, 1 if single	0.689 (0.463)	0.725 (0.446)
NOHH	dummy, 1 if person is not head of household	0.703 (0.456)	0.616 (0.486)
WAGE	earnings per hour	292.735 (313.237)	483.108 (264.402)

Table 1: *Comparative Statistics of the explanatory variables; mean and standard deviation (in brackets).*

equation we are touched by the sample selection problem. Therefore we apply the three step Heckman estimation procedure.

We did the estimation for two competing models, a standard parametric (Model I) and a semiparametric one (Model II), as we described them in Section 3. For Model II we did both, estimation with modelling the influence of log-wages nonlinearly (i)), and modelling it linearly (ii)). Notice that this makes only a difference for the calculations in step 3. Further we investigated the question what happens (for these data) when using w versus W^*, see Section 3.2.

We proceed as follows. We regress in the following steps

1. the variable JOB against AGE1, AGE2, AGE3, AGE4, EDUC1, EDUC2, EDUC3, SEXM, SINGLE, NOHH, SEXM*SINGLE, URATE and a constant (CONST) by

$$\text{(Model I)} \qquad E[\zeta|X=x] = F\left(x^T\beta_I\right) \text{ , respectively by}$$

$$\text{(Model II)} \qquad E[\zeta|X^d=x^d, X^c=x^c] = F\left\{x^{dT}\beta_{II} + \eta(x^c)\right\},$$

with $X = (X^{dT}, X^{cT})^T$ denoting all input variables, X^c =URATE;

2. ln(WAGE) against AGE1, AGE2, AGE3, AGE4, EDUC1, EDUC2, EDUC3, SEXM, SEXM*SINGLE, URATE the inverse of Mill's ratio and a constant (CONST) by

$$E[\ln(\text{WAGE})|Z = z, \Lambda = \lambda] = z^T\gamma_1 + \gamma_2\lambda,$$

again with Z denoting all input variables and Λ, λ the Mills ratio;

3. JOB against AGE1, AGE2, AGE3, AGE4, EDUC1, EDUC2, EDUC3, SEXM, SINGLE, NOHH, URATE, ln(WAGE) and (CONST) by

$$\begin{aligned} &\text{(Model I)} && E[\zeta|X = x] = F\left(x^T\beta_I\right) \text{ , respectively by} \\ &\text{(Model II)} \quad i) && E[\zeta|X^d = x^d, X^c = x^c] = F\left\{x^{dT}\beta_{II} + \alpha + \eta_1(x_1^c) + \eta_2(x_2^c)\right\} \\ &\qquad\qquad ii) && E[\zeta|X^d = x^d, X^c = x^c] = F\left\{x^{dT}\beta_{II} + \alpha + \eta_1(x_1^c) + \beta_w x_2^c\right\}, \end{aligned}$$

$X = (X^{dT}, X^{cT})^T$ denoting all input variables, x_1^c =URATE, x_2^c =ln(WAGE).

We had observed $N = 1010$ people, of which $M = 612$ have a job, for more information see Table 1.

As common, in step 2 is regressed ln(WAGE) versus $\ln(Z)$ since we are interested in the partial increase proportional to wage, i.e. $(\partial\text{WAGE})/(\partial X_j)\cdot(\text{WAGE})^{-1} = (\Delta\ln(\text{WAGE}))\,/\,(\Delta Z_j)$. For comparison reasons this was done first with a simple OLS for both models and afterwards for Model II with our estimation procedure proposed in the preceding sections.

In the latter case we estimated the constant γ_0 apart. A consistent root-N estimator was proposed by Andrews and Schafgans (1998). The estimator they consider is

$$\hat{\gamma}_0 = \frac{\sum_{i=1}^N \left(W_i - Z_i^T\hat{\gamma}_1\right)\zeta_i\kappa\left(Index_i - \delta_N\right)}{\sum_{i=1}^N \zeta_i\kappa\left(Index_i - \delta_N\right)},$$

where $Index$ stands for the index in the selection equation l_1, δ_N for a smoothing parameter with $\delta_N \to \infty$ when $N \to \infty$. The function $\kappa(\cdot)$ is a non-decreasing $[0,1]$-valued function that has three derivatives bounded over $\mathbb{R}$ and for which $\kappa(u) = 0$ for $u \leq 0$ and $\kappa(u) = 1$ for $u \geq B$ for some $0 < B < \infty$. However, they did not present simulations, applications nor discussion how to choose κ, B, δ_N reasonable in practice. An investigation about this topic and robustness of the estimator against their choices as well as against the choice of h_2, the bandwidth of the differencing estimator, would be interesting. As we are also interested in the (partly nonparametric with bandwidth h_e, explained later) estimation of the variance of $\hat{\gamma}_1$, to look at all at the same time would be too much and such a simulation study beyond the scope of this application. Instead, we compare for different bandwidths h_2 the resulting $\hat{\gamma}_0$ with the constant we got out of the simple OLS (for Model II), see Tables 2 and 3. In our application we choose for numerical reasons

$$\kappa(u) = \left[\exp\left\{5.5 - 11.0\frac{u - \delta_N}{B}\right\} + 1\right]^{-1}$$

with $B = \delta_N = \frac{1}{2} Index[0.95N]$, where $Index_{[a]}$ means the a'th order statistic.

As mentioned above we also compare the two cases of (a) taking in step 3 predicted wages for all persons in the sample versus (b) taking the real (observed) wages for workers. This comparison for Model II will strengthen the often found state that it is preferable to take as regressor the predicted log wage instead of the real ones also for the people having a job and thus working with the in average expected log wages $\hat{w}^*$ for all people.

For the nonparametric estimation we applied in all steps the quartic kernel, $K(u) = \frac{15}{16}(1 - u^2)^2 \mathbb{1}\{|u| \leq 1\}$. Since in the first step where we only pre-estimate to get an approximation for the bias correcting factor in the log wage equation, we are interested in keeping the bias small but are not that much worried about slightly bigger variances. Thus we undersmooth a little bit choosing in the first step bandwidth $h = 0.8*$ stdev(URATE), where stdev$(\cdot)$ is the standard deviation of the corresponding input. In step three we used $h_p = (1.5, 1.75)*$ stdev(URATE,ln(WAGE)) when estimating the parametric (linear) part and $h = (1.0, 1.25)*$stdev(URATE,ln(WAGE)), $g = h$ to estimate η_1, η_2. For the explanation of h and g, see Section 3. Not discussed in Section 3, we run the estimation procedure separately to get β, respectively η. Then h_p indicates the bandwidth used for getting β when there does not exist an direction of interest for η. We did all calculations also for bigger as well as smaller bandwidths. Presented are the results giving reasonable smooth estimates, respectively slightly undersmoothed ones for the first step. A nice result and a little bit surprising: qualitatively the estimates almost did not vary with the bandwidth.

We start with estimation of step 1 and 2 for both models. Hereby, we did also for Model II, step 2 a simple OLS. All results are displayed in Table 2. Note that the standard deviation for Model II, step 2 are not corrected for the first step and thus have to be interpreted carefully if at all. For the nonparametric part in Model I see Figure 1, (f1,1). Because step 1 is only for determining the inverse of Mill's ratio, we skip a detailed discussion of its numerical results. We only notice that all coefficients have the expected sign due to economic theory. Aside the signs a comparison of the results for the dummy variables between Model I and II is not possible due to the different normalizations but in any case the influence of URATE seems to be strongly nonlinear.

For the wage equation (step 2) we again have quite similar results for the different Models, except for the inverse of Mill's ratio. Notice that age, education, sex and family status have significant influence with expected signs. They confirm that very young age and low education level have a negative influence on the earnings per hour.

The fact that URATE is perfectly insignificant could indicate that pay policy and wage negotiations are still nationwide in Spain and are not affected by the labor market in the particular district. The Mills ratio in both models is strongly significant, so we indeed deal with a big selection bias in the wage equation. For the semiparametric case (Model II) it is much smaller.

We now look closer to the results of step 2, Model II, giving the estimates of the method

	step 1		step 2	
Variable	Model I	Model II	Model I	Model II
Constant	1.034 (0.213**)	1.003 -	6.305 (0.074**)	6.393 (0.065**)
AGE1	−0.520 (0.183**)	−0.450 (0.112**)	−0.302 (0.097**)	−0.329 (0.097**)
AGE2	−0.267 (0.163*)	−0.259 (0.099**)	0.046 (0.075)	0.044 (0.060)
AGE3	−0.188 (0.146)	−0.192 (0.088**)	0.006 (0.067)	0.010 (0.052)
AGE4	−0.439 (0.161**)	−0.443 (0.096**)	0.071 (0.070)	0.068 (0.055)
EDUC1	0.034 (0.105)	−0.033 (0.066)	−0.027 (0.051)	−0.016 (0.046)
EDUC2	−0.118 (0.142)	−0.178 (0.090**)	0.210 (0.070**)	0.214 (0.063**)
EDUC3	−0.531 (0.182**)	−0.510 (0.115**)	0.645 (0.109**)	0.601 (0.101**)
SEXM	−0.219 (0.181)	−0.247 (0.112**)	0.081 (0.075)	0.060 (0.065)
SINGLE	0.696 (0.199**)	0.690 (0.120**)	–	–
NOHH	−0.895 (0.130**)	−0.910 (0.078**)	–	–
SEXM*SINGLE	−0.102 (0.212)	−0.089 (0.131)	−.1111 (0.066*)	−0.093 (0.059)
U-RATE	−0.504 (0.614)	–	0.073 (0.280)	0.071 (0.215)
Mill's inverse	–	–	−0.480 (0.116**)	−0.299 (0.062**)

Table 2: *Estimation results for step 1 and 2 for the parametric part. Standard deviations are given in brackets. Asterisks indicate significance at 10 (*), respectively 5 (**) percent level.*

proposed in Section 3 including the (correct) standard deviations. Looking at Theorem 2 and a careful check of its proof reveals that we can estimate the variance of $\hat{\gamma}_1$ by

$$\widehat{Var}(\hat{\gamma}_1) = \hat{\sigma}_2^2 \, \Phi_{sz}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ f(\nu_i) F(\nu_i) \zeta_i \left[s_i - \frac{E[\zeta s|\nu_i]}{F(\nu_i)} \right] \right\} \tag{38}$$

$$\left\{ f(\nu_i) F(\nu_i) \zeta_i \left[s_i - \frac{E[\zeta s|\nu_i]}{F(\nu_i)} \right] \right\}^T \Phi_{sz}^{-1} , \tag{39}$$

with $\hat{\sigma}_2^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_{2i}^2$, $\hat{u}_{2i}$ the residuals of the log-wage equation. Unfortunately, we can get only these residuals for the people working. Therefore, we take $\hat{\sigma}_2^2 = \frac{1}{M} \sum_{i=1}^{M} \left(\hat{u}_{2i} - \bar{\hat{u}}_2 \right)^2$ only with the obtained M residuals and $\bar{\hat{u}}_2 = \frac{1}{M} \sum_{i=1}^{M} \hat{u}_{2i}$.

The conditional expectations $E[\zeta s|\nu_i]$, $E[\zeta z|\nu_i]$, see definition of Φ_{sz} in the Appendix I, we estimate using the Nadaraya-Watson estimator with quartic kernel and bandwidth h_e.

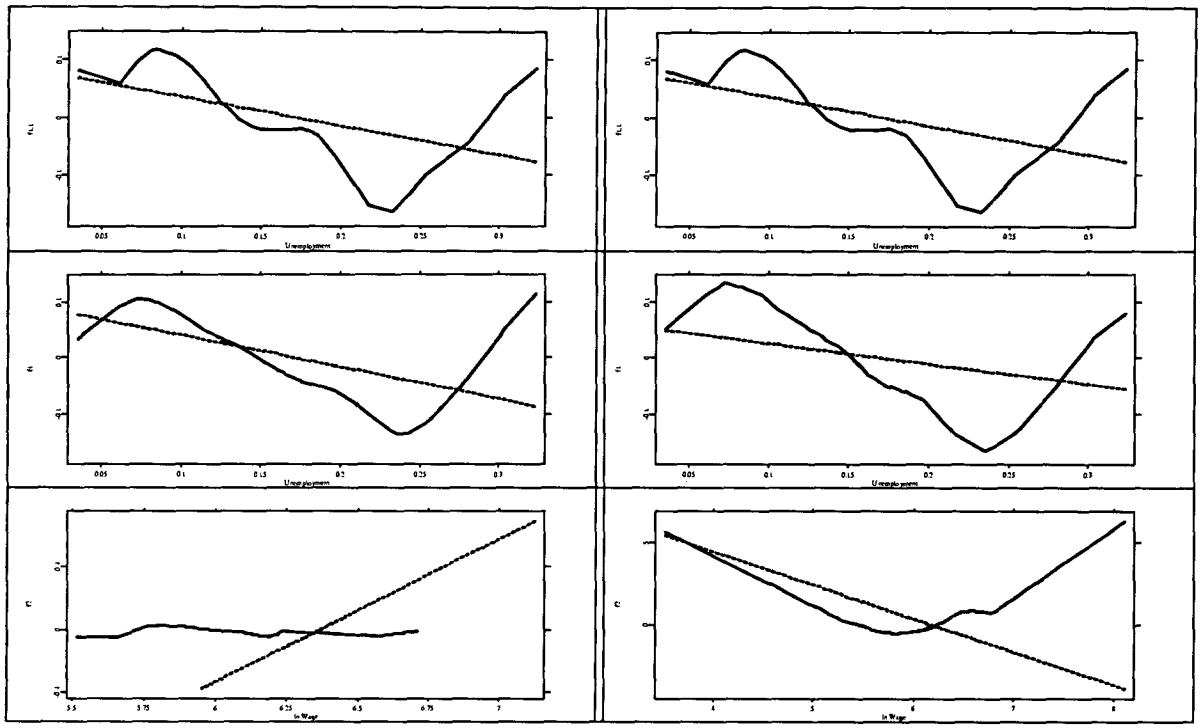

Figure 1: *Nonparametric and parametric estimates for URATE and ln(WAGE). At top (f1,1) for step 1, below (f1,f2) for step 3, left case (a), right case (b).*

Since the robustness of the estimators γ_1, $Var(\gamma_1)$ against the choice of h_2, h_e is of crucial importance as it decides about slope, variance, significance and has direct impact on step 3, we dedicated an own small simulation study on this topic. In this context we also got the different $\hat{\gamma}_0$ to compare with the OLS estimate. All the results can be found in Table 3.

The all over impression is that the results, taking the trade-off between absolute value for the coefficient estimates and their variance into account, are astonishing robust. We also see the closeness to the simple OLS results. Certainly, the standard deviations are bigger now since the one in Table 2 (step 2, Model II) are not corrected for the first step. For all results in Table 3 we have significance at 5% level for AGE1 and EDUC3, whereas EDUC2 is only significant at about 12%.

Unfortunately, in both models the wage regression have a R^2 of only about 21%, i.e. we are not explaining much of the variance of ln(WAGE). This lead us to the comparison of two cases ((a) and (b)) in step 3. In Figure 1 we have plotted the estimates of the influence functions for URATE, and ln(WAGE) for both cases when modelling influence of log-wages nonlinear (case *i*)). The problem in case (b), when we take the observed wages for workers and predicted wages only for the non workers is, that due to the small R^2 we have predicted wages only in a much smaller (but high-level) range than the range is for observed wages. For the case where taking the predicted wages for all, we have the functions on quite different

h_2		$\hat{\gamma}_1$, $\hat{\gamma}_0$	standard deviation			
	h_e		0.5	1.0	1.5	2.0
	$\hat{\gamma}_0$	5.989				
	AGE1	-0.385	0.104	0.103	0.106	0.108
	AGE2	-0.001	0.232	0.224	0.224	0.224
	AGE3	-0.022	0.264	0.252	0.251	0.250
1.0	AGE4	0.051	0.252	0.248	0.248	0.250
	EDUC1	-0.040	0.291	0.280	0.279	0.279
	EDUC2	0.188	0.140	0.133	0.133	0.133
	EDUC3	0.587	0.040	0.039	0.041	0.042
	SEXM	0.007	0.418	0.416	0.417	0.418
	SEXM*SINGLE	-0.066	0.349	0.364	0.364	0.364
	U-RATE	0.054	0.099	0.087	0.085	0.085
	h_e		0.5	1.0	1.5	2.0
	$\hat{\gamma}_0$	5.947				
	AGE1	-0.348	0.105	0.105	0.107	0.110
	AGE2	0.026	0.236	0.227	0.228	0.227
	AGE3	-0.002	0.268	0.256	0.254	0.253
0.75	AGE4	0.054	0.256	0.251	0.252	0.253
	EDUC1	-0.041	0.296	0.284	0.283	0.283
	EDUC2	0.194	0.142	0.135	0.135	0.135
	EDUC3	0.631	0.041	0.040	0.041	0.043
	SEXM	0.028	0.424	0.422	0.423	0.424
	SEXM*SINGLE	-0.080	0.354	0.369	0.369	0.369
	U-RATE	0.073	0.100	0.088	0.086	0.086
	h_e		0.5	1.0	1.5	2.0
	$\hat{\gamma}_0$	5.924				
	AGE1	-0.324	0.106	0.106	0.109	0.111
	AGE2	0.038	0.238	0.229	0.230	0.229
	AGE3	0.007	0.271	0.259	0.257	0.256
0.5	AGE4	0.044	0.259	0.254	0.255	0.256
	EDUC1	-0.044	0.299	0.287	0.286	0.286
	EDUC2	0.203	0.143	0.137	0.136	0.137
	EDUC3	0.676	0.042	0.041	0.042	0.043
	SEXM	0.039	0.429	0.427	0.428	0.429
	SEXM*SINGLE	-0.087	0.358	0.373	0.373	0.373
	U-RATE	0.083	0.101	0.089	0.087	0.087

Table 3: *Estimation results for Model II, in step 2, using the differencing estimtor for different bandwidths.*

scales for $log-\hat{w}^*$. This is due to the fact that maybe the estimator of Andrews and Schafgans (1998) is underestimating the constant. Fortunately, this does not affect the further steps neither the estimation of the standard deviations of $\hat{\gamma}_1$.

Consequently all small wages in that sample belong to workers and vice versa we get a strongly negative estimate for the influence of ln(WAGE) on having a job. Therefore, and to be consistent in the inputs, we rely more on case (a) where we take the predicted wages for all people and thus avoid the problem of having two quite different variations in the same predictor variable. As we will see, wage seems to be absolutely insignificant in the linear model (in this example) and thus a linear modelling (case *ii*)) is not affecting the other results. So we skipped the presentation of the results for that case.

In Table 4 we show the estimation results in step 3, case (a) for the parametric part in all models considered. First, to manifest the difference between a standard probit, as often done in the economic literature, and the probit with corrected standard deviations, we present both for Model I. We can see that the corrected ones are about 5 to 15% bigger than the uncorrected ones.

The coefficient estimates for Model I and II are again different. Certainly, including ln(WAGE), the significance for the other explanatory variables is shrinking compared to step 1. But still age (AGE2, AGE4), SEX, SINGLE and NOHH is highly significant with expected signs. In Model I URATE is only significant at a level of about 21%. But this statement holds only for the linear influence of URATE. Looking at Figure 1, we see a clearly nonlinearity for URATE while for ln(WAGE) insignificance seems to be real and not just caused by a misspecification of its functional form. In Model II now also education is highly significant. In general we find that allowing the influences of the continuous regressors to be nonlinear, increases the significance for the parametric part even compared to the uncorrected standard errors.

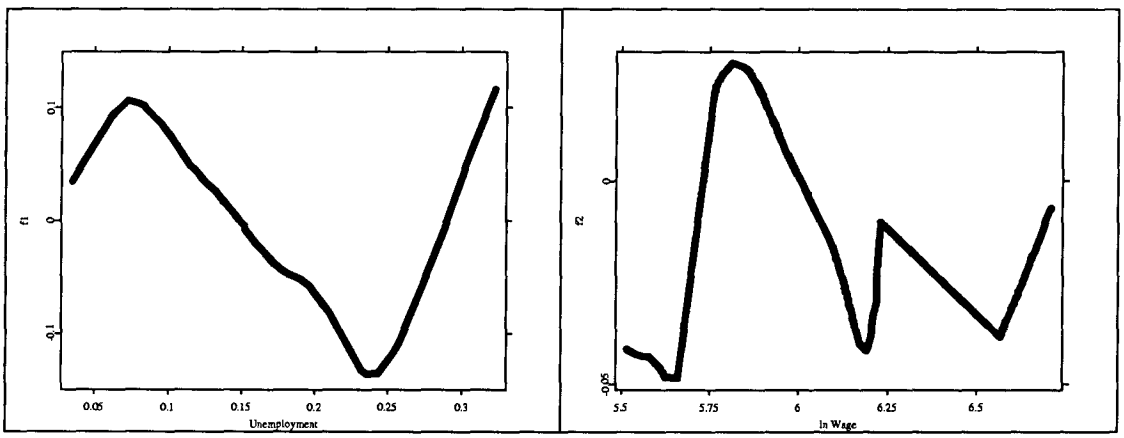

Figure 2: *Nonparametric estimates for URATE and ln(WAGE) for step 3, case (a).*

Variable	Model I	Model I*	Model II
Constant	−4.715 (14.001)	(12.133)	1.044 -
AGE1	−0.244 (0.829)	(0.615)	−0.477 (0.133**)
AGE2	−0.310 (0.201)	(0.181*)	−0.257 (0.098**)
AGE3	−0.193 (0.168)	(0.145)	−0.198 (0.088**)
AGE4	−0.504 (0.213**)	(0.204**)	−0.444 (0.096**)
EDUC1	0.059 (0.134)	(0.118)	−0.035 (0.066)
EDUC2	−0.309 (0.474)	(0.421)	−0.185 (0.097**)
EDUC3	−1.119 (1.351)	(1.228)	−0.660 (0.260**)
SEXM	−0.292 (0.123**)	(0.094**)	−0.308 (0.058**)
SINGLE	0.696 (0.190**)	(0.199**)	0.621 (0.067**)
NOHH	−0.895 (0.130**)	(0.130**)	−0.894 (0.075**)
ln(WAGE)	0.912 (2.209)	(1.907)	–
U-RATE	−0.570 (0.652)	(0.631)	–

Table 4: *Estimation results for step 3, case (a) for the parametric part. Standard deviations are given in brackets. Column Model I* is giving the uncorrected standard deviations. Asterisks indicate significance at 10 (*), respectively 5 (**) percent level. For Model II we give here the uncorrected standard deviations neglecting the first two steps.*

Nevertheless, zooming the graphics, see Figure 2, we get the impression that in Spain there is a small upper-middle class. We have many jobs where people earn a small salary, but for some people with the adequate abilities to earn a lot there is an increasing probability to get such a job.

Looking at the estimated influence of URATE, the functional form is a little bit harder to understand. Let E be the employed people, LF the labor force (all participants), U the unemployed people, u the percentage of U to LF, and P the probability to be employed (that is what we estimate).

Describing the probability of having a job in terms of E, u and LF, we have

$$P = E/LF = E/(E+U) = E/(E+u*LF)$$

and consider

$$\frac{\partial P}{\partial u} = \frac{-E * LF}{(E + u * LF)^2},$$

compare with parametric results in Table 4.

Now we consider this with respect to a possibly change in Labor Force. This consideration makes sense, since there are many situations where at the same time many people give up, e.g. in a recession when nobody is believing in his chance to find a job.

$$\frac{\partial P/\partial u}{\partial LF} = \frac{-E(E + u * LF) + 2E * u * LF}{(E + u * LF)^3} = \frac{-E^2 + E * u * LF}{(E + u * LF)^3}.$$

Here we can see that indeed for certain changes in LF or u this probability P can be increasing or decreasing, compare e.g for especially high unemployment u. In the next data wave the unemployment rate u should be corrected then for this change in Labor Force. Certainly also other reasons could be thought to be responsible for this phenomenon.

5 Conclusions

The purpose of this paper has been to provide a new, practical way of specification and estimation a standard model of labor supply with simultaneous equations of interest. As a byproduct we gave also theoretical results for the so far used fully parametric procedure. The method is based on the ideas developed by Heckman and it can be included in the so called predicted wage methods. We have been able to identify, estimate and give asymptotic theory for all parameters and functions of interest, including the standard deviations.

According to the empirical findings, the relationship between labor supply and some of its explaining variables in the structural equation is nonlinear and this fact not only contradicts the standard assumptions incorporated in previous econometric models of labor supply but also demands for more flexible methods. Our procedure allows for this flexibility and moreover has been shown to be quite robust in its smoothing based parts. With the application we thus could maintain the practical relevance of the proposed methods.

Appendix I: Assumptions

For the asymptotic expansions that we have introduced in Theorem 2, we make the following assumptions.

(B.1) The values $\{w_i, x_i, z_i\}_{i=1}^N$ are realizations from i.i.d random variables, where $W \in I\!R$, $X \in I\!R^{p+d}$ and $Z \in I\!R^r$. Moreover, $\{y_i\}_{i=1}^N$ are realizations from a truncated random variable, Y, and ζ is a binary variable that takes the value 1 as $y > 0$ and 0 otherwise.

(B.2) $X = (X^d, X^c)$, where $X^c \in I\!R^d$ are absolutely continuous random variables and $X^d \in I\!R^p$ are discrete or dummy variables. $\{x_i^d, x_i^c\}_{i=1}^N$ are realizations from (X^d, X^c). X^d and X^c have compact support D_d and D_c. The support D_c is of the form $D_{c,1} \times D_{c,-1}$ with $D_{c,1} \subset I\!R$ and $D_{c,-1} \subset I\!R^{d-1}$. X^c has a twice continuously differentiable density f_c with $\inf_{x^c \in D_c} f_c(x^c) > 0$. Furthermore, the same conditions are assumed for Z.

(B.3) W has finite sixth order moments, and the laplace transform $E \exp t|Y|$ is finite for $t > 0$ small enough.

(B.4) The data satisfy the restrictions (1) and (2) as defined in equation (14).

(B.5) $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim N(0, \Sigma)$ where $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

(B.6) $\{\eta_j(\bullet)\}_{j=1}^d$ are four times continuously differentiable on $I\!R$. The weight functions w, w_{-1} and w_1 are positive and twice continuously differentiable. To avoid problems on the boundary, we assume that for a $\delta > 0$ we have that $w_{-1}(t) = 0$, $w_1(t) = 0$, and $w(t) = 0$ for $t \in D_{c,-1}^- = \{s :$ there exists an $u \notin D_{c,-1}$ with $\|s - u\| \leq \delta\}$, $t \in D_{c,1}^- = \{s :$ there exists an $u \notin D_{c,1}$ with $\|s - u\| \leq \delta\}$ or $t \in S_T^- = \{s :$ there exists an $u \notin D_c$ with $\|s - u\| \leq \delta\}$, respectively. Furthermore, the weight function w_1 is such that $\int_{D_{c,1}} w_1(t_1) m_1(t_1) f_{T_1}(t_1) dt_1 = 0$, where f_{T_1} denotes the density of T_1.

(B.7) The kernel L is a product kernel $L(v) = L_1(v_1) \cdot \ldots \cdot L_{d-1}(v_{d-1})$. The kernels L_j are symmetric probability densities with compact support ($[-1, 1]$, say), $j = 1, \ldots, d-1$. The kernel K is a symmetric probability density with compact support (e.g.$[-1, 1]$),too. Moreover, the kernels K and L are twice continuously differentiable.

(B.8) The kernel function $K_2(\bullet)$ satisfies

i) $K_2(u)$ is twice differentiable, with $K_2'' < k_0$ for some k_0.

ii) $K_2(u) = K_2(-u)$.

iii) $K_2(u) = 0$ if $|u| > l_0$ for some $l_0 > 0$.

iv) $\int u^l K(u) du = 0$ for $l = 1, 2, 3$.

(B.9) $E[\varphi\{\nu\}|v^c = t]$ and $E\left[\varphi\{\nu\} v^d | v^c = t\right]$ are twice continuously differentiable functions for $t \in D_c$. Where

$$\varphi\{\nu\} = \frac{\partial}{\partial \nu}\left[\frac{\zeta - F(\nu)}{F(\nu)(1 - F(\nu))} f(\nu)\right]$$

(B.10) The matrix $E\, T^2 \widetilde{v^d} \widetilde{v^d}^T$ is strictly positive definite. This assumption implies that v^d does not contain an intercept.

(B.11) The conditional expectation functions $E[\nu|\nu = \nu_i]$ $E[\zeta z|\nu = \nu_i]$ and $E[\zeta s|\nu = \nu_i]$ are at least three times continuously differentiable in the index function.

(B.12) The matrix

$$\Phi_{sz} = E\left[f(\nu)F(\nu)^2\left[s - \frac{E[\zeta s|\nu]}{F(\nu)}\right]\left[z - \frac{E[\zeta z|\nu]}{F(\nu)}\right]^T\right]$$

is positive definite.

Appendix II: Proof of Theorems

Proof of Theorem 1

Before to prove the results we introduce the following lemma.

Lemma 1 (Newey and McFadden, 1994) *If z_i is i.i.d., $a(z,\theta)$ is continuous at θ_0 with probability one, and there is a neighborhood N of θ_0 such that $E\left[\sup_{\theta\in N}\|a(z,\theta)\|\right] < \infty$, then for any $\tilde{\theta} \to_p \theta_0$, $n^{-1}\sum_{i=1}^n a\left(z_i,\tilde{\theta}\right) \xrightarrow{P} E\left[a(z,\theta_0)\right]$.*

The proof can be found in Newey and McFadden (1994), p. 2156.

We will show first the convergence in distribution result. In order to do so, first note that

$$\sqrt{N}\left(\hat{\delta} - \delta_0\right) = \frac{1}{\sqrt{N}}\sum \bar{\psi}(x_i, z_i), \tag{40}$$

where $\bar{\psi}(x_i,z_i) = -\left[\frac{1}{N}\sum_{i=1}^N \nabla_\delta m\left(x_i,z_i,\bar{\delta}\right)\right]^{-1} m(x_i,z_i)$ and $\bar{\delta}$ is a mean value. This is because $\hat{\delta}$ makes (7) equal to zero, with probability one, and the mean value theorem. Furthermore, Assumptions (A.1) to (A.6) ensure both that $\hat{\delta}$ is a consistent estimator for δ_0 (see Manski and McFadden, 1994: Theorem 2.5, p. 2131), and $E\left[\sup_{\delta\in N_\delta}\|m(x,z;\delta)\|\right] < \infty$. Therefore, Lemma 1 applies for $\bar{\delta}$ between $\hat{\delta}$ and and δ_0 and then

$$\sqrt{N}\left(\hat{\delta} - \delta_0\right) = \frac{1}{\sqrt{N}}\sum \psi(x_i, z_i) + o_p(1), \tag{41}$$

where now, $\psi(x_i,z_i) = -\left[E\left[\nabla_\delta m(x,z)\right]\right]^{-1} m(x_i,z_i)$. Proceeding in the same way as before for equation (9), it is also possible to show that

$$\begin{aligned}\sqrt{N}(\hat{\gamma} - \gamma_0) = &-\left[\tfrac{1}{N}\textstyle\sum_{i=1}^N \nabla_\gamma g\left(x_i,z_i,\hat{\delta},\bar{\gamma}\right)\right]^{-1} \times \\ &\left\{\tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N g(x_i,z_i) + \left[\tfrac{1}{N}\textstyle\sum_{i=1}^N \nabla_\delta g\left(x_i,z_i,\bar{\delta},\gamma_0\right)\right]\sqrt{N}\left(\hat{\delta} - \delta_0\right)\right\}.\end{aligned} \tag{42}$$

But then, if we substitute (41) into (42) and we apply again Lemma 1 we obtain

$$\sqrt{N}\left(\hat{\gamma}-\gamma_0\right)=-G_\gamma^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{g\left(x_i,z_i\right)+G_\delta\psi\left(x_i,z_i\right)\right\}+o_p(1). \tag{43}$$

The same can be done for equation (12) and then

$$\begin{aligned}&\sqrt{N}\left(\hat{\tau}-\tau_0\right)=\\ &-H_{N\tau}\left(\hat{\delta},\hat{\gamma},\bar{\tau}\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}h\left(x_i,z_i\right)-H_{N\tau}\left(\hat{\delta},\hat{\gamma},\bar{\tau}\right)^{-1}H_{N\delta}\left(\bar{\delta},\gamma_0,\tau_0\right)\sqrt{N}\left(\hat{\delta}-\delta_0\right)\\ &-H_{N\tau}\left(\hat{\delta},\hat{\gamma},\bar{\tau}\right)^{-1}H_{N\gamma}\left(\delta_0,\bar{\gamma},\tau_0\right)\sqrt{N}\left(\hat{\gamma}-\gamma_0\right)\\ &-H_{N\tau}\left(\hat{\delta},\hat{\gamma},\bar{\tau}\right)^{-1}H_{N\delta\gamma}\left(\bar{\delta},\bar{\gamma},\tau_0\right)N\left(\hat{\delta}-\delta_0\right)\left(\hat{\gamma}-\gamma_0\right),\end{aligned} \tag{44}$$

where

$$H_{N\tau}\left(\hat{\delta},\hat{\gamma},\bar{\tau}\right)=\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\nabla_\tau h\left(x_i,z_i,\hat{\delta},\hat{\gamma},\bar{\tau}\right),$$

$$H_{N\delta}\left(\bar{\delta},\gamma_0,\tau_0\right)=\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\nabla_\delta h\left(x_i,z_i,\bar{\delta},\gamma_0,\tau_0\right),$$

$$H_{N\gamma}\left(\delta_0,\bar{\gamma},\tau_0\right)=\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\nabla_\gamma h\left(x_i,z_i,\delta_0,\bar{\gamma},\tau_0\right),\text{ and}$$

$$H_{N\delta\gamma}\left(\bar{\delta},\bar{\gamma},\tau_0\right)=N^{-\frac{3}{2}}\textstyle\sum_{i=1}^{N}\nabla_\delta\nabla_\gamma h\left(x_i,z_i,\bar{\delta},\bar{\gamma},\tau_0\right).$$

Now, substituting back both equations (41) and (43) into (44) and applying Lemma 1 then we obtain

$$\begin{aligned}&\sqrt{N}\left(\hat{\tau}-\tau_0\right)=-H_\tau^{-1}\\ &\times\tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^{N}\left\{h\left(x_i,z_i\right)+H_\delta\psi\left(x_i,z_i\right)+H_\gamma G_\gamma^{-1}\left(g\left(x_i,z_i\right)+G_\delta\psi\left(x_i,z_i\right)\right)\right\}+o_p(1)\end{aligned} \tag{45}$$

and taking into account assumption (A.2), we can apply both the Lindeberg-Levy CLT and the Slutzky theorem (see Serfling, 1980: pp.19 and 28) and the proof is done.

The proof of consistency of the variance covariance matrix $\hat{V}\left(\hat{\tau}\right)$ is immediate by applying Lemma 1 to the different terms. □

Proof of Theorem 2

For the proof of this theorem, we need the following lemmas

Lemma 2 (Powell, Stock and Stoker, 1989) *Consider a second order U-statistic of the form*

$$U_N=\binom{n}{2}^{-1}\sum_{i=1}^{N}\sum_{j=i+1}^{N-1}p_N\left(z_i,z_j\right)$$

where $\{z_i, i = 1, \cdots, N\}$ *is an i.i.d. random sample and* p_N *is a k-dimensional symmetric kernel. Also define*

$$\begin{aligned} r_N(z_i) &= E\left[p_N\left(z_i, z_j\right)|z_i\right], \\ \theta_N &= E\left[r_N\left(z_i\right)\right] = E\left[p_N\left(z_i, z_j\right)\right] \\ \hat{U}_N &= \theta_N + \frac{2}{N}\sum_{i=1}^{N}\left[r_N(z_i) - \theta_N\right]. \end{aligned}$$

If $E\left[\left\|p_N\left(z_i, z_j\right)\right\|^2\right] = o(N)$, *then*

$$U_N = \hat{U}_N + o_p\left(\frac{1}{\sqrt{N}}\right). \tag{46}$$

The proof of this lemma is the proof of lemma 3.1 in Powell, Stock and Stoker (1989), p. 1426.

Lemma 3 *Under assumptions (B.1) to (B.12) and if the sequence of bandwidths* h, h_2 *and* g *tends to zero such that* $Nhg^{2(s-1)}\left(\log N\right)^{-2} \to \infty$, $Nh_2^6 \to \infty$ *and* $Nh_2^8 \to 0$ *then*

$$\hat{M}_{sz} = 2\Phi_{sz} + o_p(1)$$

where Φ_{sz} *is defined in assumption (B.12), and*

$$\sqrt{N}\left(M_{sw} - \gamma_1 M_{sz}\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} f\left(\nu_i\right) F\left(\nu_i\right) \zeta_i \left[s_i - \frac{E[\zeta s|\nu_i]}{F\left(\nu_i\right)}\right] u_{2i} + o_p(1)$$

where u_{2i} *is the error term of the structural wage equation,*

$$\nu_i = \beta^T v_i^d + \theta + \sum_{k=1}^{s} \psi_k(v_{ik}^c),$$

$$M_{sw} = \left(\begin{array}{c} n \\ 2 \end{array}\right)^{-1} \sum_{i=1}^{N}\sum_{j=i+1}^{N-1} \omega_{ij}\left(s_i - s_j\right)\left(w_i - w_j\right),$$

$$M_{sz} = \left(\begin{array}{c} n \\ 2 \end{array}\right)^{-1} \sum_{i=1}^{N}\sum_{j=i+1}^{N-1} \omega_{ij}\left(s_i - s_j\right)\left(z_i - z_j\right)^T, \tag{47}$$

and

$$\omega_{ij} = \frac{1}{h_2} K_2\left(\frac{\lambda\left\{\beta^T v_i^d + \theta + \sum_{k=1}^{s}\psi_k(v_{ik}^c)\right\} - \lambda\left\{\beta^T v_j^d + \theta + \sum_{k=1}^{s}\psi_k(v_{jk}^c)\right\}}{h_2}\right)\zeta_i\zeta_j$$

$f(\bullet)$ *and* $F(\bullet)$ *are respectively the gaussian density and the gaussian distribution function.*

Proof of Lemma 3

In order to prove the fist statement of the lemma, we first show that

$$\hat{M}_{sz} - M_{sz} = o_p(1).$$

To do so, from equations (27) and (47) we have that

$$\hat{M}_{sz} - M_{sz} = \binom{n}{2}^{-1} \sum_{i=1}^{N} \sum_{j=i+1}^{N-1} \{\hat{\omega}_{ij} - \omega_{ij}\} (s_i - s_j) (z_i - z_j)^T \leq$$

$$\left[\binom{n}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|s_i - s_j\| \, \|z_i - z_j\| \right] \times \max_{ij} \{\hat{\omega}_{ij} - \omega_{ij}\} . \tag{48}$$

Now, with the same arguments as in Lemma B.1 from Ahn and Powell (1992), p. 23, we can obtain the following bound

$$\max_{ij} \{\hat{\omega}_{ij} - \omega_{ij}\} \leq 2k_0 h_2^{-2} \left\{ \sup_{v} \|\lambda \left(\hat{\beta}^T v_i^d + \hat{\theta} + \sum_{k=1}^{s} \hat{\psi}_k(v_{ik}^c) \right) - \lambda \left(\beta^T v_i^d + \theta + \sum_{k=1}^{s} \psi_k(v_{ik}^c) \right) \| \right\} .$$

Using a third order Taylor expansion around the true values, assumption (B.5) and the uniform convergence properties of the first step estimators (Härdle, Huet, Mammen and Sperlich, 1999; p. 32) then

$$\max_{ij} \{\hat{\omega}_{ij} - \omega_{ij}\} = O_p \left(h_2^{-2} N^{-\frac{2}{4+s}} \right) . \tag{49}$$

If $s \leq 3$, then equation (49) is $o_p(1)$ because we have assumed that $Nh_2^6 \to \infty$. This constraint in s can be weakened by assumption of higher order smoothness in $\psi_1, \cdots, \psi_s$ and by use of higher order kernels.

Finally, applying Lemma 2 and the strong law of large numbers for U-statistics (see Serfling, 1980; p. 190) the first term of the r.h.s. of equation (48) is $O_p(1)$, and since $Nh_2^6 \to \infty$ the second term is $o_p(1)$, and the first result is proved.

The proof of the first part of the lemma will be closed by proving that

$$M_{sz} = 2\Phi_{sz} + o_p(1).$$

The proof of the previous statement relies on the arguments of Lemma 5.1 and Theorem 5.1 of Powell (1987). Taking expectations conditional on the index functions and applying assumptions (B.5), (B.8) and (B.11) then it can be shown that

$$E \left[\left\| \omega_{ij} (s_i - s_j)(z_i - z_j)^T \right\|^2 \right] = O \left(h_2^{-1} \right) + O \left(h_2^4 \right) .$$

This expression is $o(N)$ because $Nh_2^6 \to \infty$. Proceeding as in the previous expression then

$$E \left[\omega_{ij} (s_i - s_j)(z_i - z_j)^T \right] = 2E \left[f(\nu) F(\nu)^2 \left[s - \frac{E[\zeta s|\nu]}{F(\nu)} \right] \left[z - \frac{E[\zeta z|\nu]}{F(\nu)} \right]^T \right] + O \left(h_2^4 \right)$$

where $\nu = \beta^T v^d + \theta + \sum_{k=1}^{s} \psi_k(v_k^c)$. Lemma 2 and the strong law of large numbers for U-statistics (see Serfling, 1980; p. 190) applies and the proof is done.
Next we show the second part of the lemma, i.e.

$$\sqrt{N}\left(M_{sw} - \gamma_1 M_{sz}\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} f\left(\nu_i\right) F\left(\nu_i\right) \zeta_i \left[s_i - \frac{E[\zeta s|\nu_i]}{F\left(\nu_i\right)}\right] u_{2i} + o_p(1)$$

Taking into account that $w_i = z_i^T \gamma_1 + \gamma_2 \lambda\left(\beta^T v_i^d + \theta + \sum_{k=1}^{s} \psi_k(v_{ik}^c)\right) + u_{2i}$, then

$$M_{sw} = \gamma_1 M_{sz} + \binom{n}{2}^{-1} \sum_{i=1}^{N} \sum_{j=i+1}^{N-1} \omega_{ij}\left(s_i - s_j\right)\left(\gamma_2\left(\lambda(\nu_i) - \lambda(\nu_j)\right) + u_{2i} - u_{2j}\right)$$

To apply lemma 2 we have to show that

$$E\left[\left\|\omega_{ij}\left(s_i - s_j\right)\left(\gamma_2\left(\lambda(\nu_i) - \lambda(\nu_j)\right) + u_{2i} - u_{2j}\right)\right\|^2\right] = o(N). \tag{50}$$

We claim,

$$\gamma_2^2 E\left[\omega_{ij}^2\left(\lambda(\nu_i) - \lambda(\nu_j)\right)^2 \left(s_i - s_j\right)^T \left(s_i - s_j\right)\right] = O(h_2), \tag{51}$$

$$E\left[\omega_{ij}^2\left(u_{2i} - u_{2j}\right)^2 \left(s_i - s_j\right)^T \left(s_i - s_j\right)\right] = O(h_2^{-1}), \tag{52}$$

$$2\gamma_2 E\left[\omega_{ij}^2\left(\lambda(\nu_i) - \lambda(\nu_j)\right)\left(u_{2i} - u_{2j}\right)\left(s_i - s_j\right)^T \left(s_i - s_j\right)\right] = 0. \tag{53}$$

Taking expectations conditional on the index functions and applying assumptions (B.1), (B.2), (B.5) and (B.11) then (51) and (52) hold. Moreover, equation (53) holds because $E\left(u_2|\nu = \nu_i\right) = 0$. Therefore, under the conditions previously stated on the bandwidths the result in equation (50) has been shown. Then Lemma 2 and the strong law of large numbers for U-statistics (see Serfling, 1980; p. 190) applies and the proof of the desired result is done by noticing that under the conditions previously stated in the lemma

$$\begin{aligned} \theta_N &= E\left[\omega_{ij}\left(s_i - s_j\right)\left(\gamma_2\left(\lambda(\nu_i) - \lambda(\nu_j)\right) + u_{2i} - u_{2j}\right)\right] = 0 \\ r_N\left(z_i\right) &= f\left(\nu_i\right) F\left(\nu_i\right) \zeta_i \left[s_i - \frac{E[\zeta s|\nu_i]}{F\left(\nu_i\right)}\right] u_{2i} + O\left(h_2^4\right). \end{aligned}$$

This closes the proof of lemma 3.

□

Lemma 4 *Under assumptions (B.1) to (B.12) and if the sequence of bandwidths h, h_2 and g tends to zero such that $h = g = o(N^{-1/8})$, $hg^{s-1}N^{1/2}(\log N)^{-1} \to \infty$, $Nh_2^6 \to \infty$ and $Nh_2^8 \to 0$ then*

$$\sqrt{N}\left(\hat{M}_{sw} - M_{sw}\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{A_{1i}\tilde{v}_i^d + A_{2i}\kappa_i\left(v_i^c\right)\right\} \frac{\zeta_i - F\left(\nu_i\right)}{F\left(\nu_i\right)\left(1 - F\left(\nu_i\right)\right)} f\left(\nu_i\right) + o_p(1)$$

where where Φ_{sz} is defined in assumption (B.12).

$$A_{1i} = E\left\{\lambda'(\nu_i) F(\nu_i)^2 f(\nu_i) \left[s_i - \tfrac{E[\zeta s|\nu_i]}{F(\nu_i)}\right] \times \left[w_i - \{E[T_i^2|v_i^c]\}^{-1} E\left[T_i^2 v_i^d|v_i^c\right] - E\left[w_i - \{E[T_i^2|v_i^c]\}^{-1} E\left[T_i^2 v_i^d|v_i^c\right]\right]\right]^T\right\},$$

$$A_{2i} = E\left\{\lambda'(\nu_i)\left\{E\left[T_i^2|v_i^c\right]\right\}^{-1} F(\nu_i)^2 \left[s_i - \frac{E[\zeta s|\nu_i]}{F(\nu_i)}\right]\right\}$$

$$\tilde{v}_i^d = v_i^d - \left\{E\left[T_i^2|v_i^c\right]\right\}^{-1} E\left[T_i^2 v_i^d|v_i^c\right] \qquad i = 1, \cdots, N,$$

$$T_i^2 = \frac{f(\nu_i)^2}{F(\nu_i)(1 - F(\nu_i))} \qquad i = 1, \cdots, N$$

$$\kappa_i(v^c) = \frac{K_h(v_1^c - v_{i1}^c) L_g(v_{\underline{1}}^c - v_{i\underline{1}}^c)}{\frac{1}{n}\sum_{j=1}^n K_h(v_1^c - v_{i1}^c) L_g(v_{\underline{1}}^c - v_{i\underline{1}}^c)} \qquad i = 1, \cdots, N \tag{54}$$

as N tends to infinity.

Proof of Lemma 4

In order to show the lemma we first prove that

$$\left\|\hat{S}_{zw} - S_{zw} - \binom{n}{2}^{-1} \sum_{i=1}^N \sum_{j=i+1}^{N-1} \Xi_{ij} \left[(\lambda(\hat{\nu}_i) - \lambda(\nu_i)) - (\lambda(\hat{\nu}_j) - \lambda(\nu_j))\right]\right\| = o_p\left(N^{-1/2}\right). \tag{55}$$

where

$$\Xi_{ij} = \left(\frac{1}{h_2}\right)^2 K_2'\left(\frac{\lambda(\nu_i) - \lambda(\nu_j)}{h_2}\right) \zeta_i \zeta_j (s_i - s_j)(\gamma_2(\lambda(\nu_i) - \lambda(\nu_j)) + u_{2i} - u_{2j}) \tag{56}$$

The l.h.s. of equation (55) is bounded by

$$\left\|\hat{M}_{sz} - M_{sz}\right\| + \left\{\max_{ij}\left|\hat{\omega}_{ij} - \omega_{ij} - \left(\tfrac{1}{h_2}\right)^2 K_2'\left(\tfrac{\lambda(\hat{\nu}_i) - \lambda(\nu_i)}{h_2}\right) \zeta_i \zeta_j \left[(\lambda(\hat{\nu}_i) - \lambda(\nu_i)) - (\lambda(\hat{\nu}_j) - \lambda(\nu_j))\right]\right|\right\} \times \binom{n}{2} \|s_i - s_j\| \|\gamma_2(\lambda(\nu_i) - \lambda(\nu_j)) + u_{2i} - u_{2j}\|.$$

By Lemma 3,

$$\left\|\hat{M}_{sz} - M_{sz}\right\| = o_p(1). \tag{57}$$

Furthermore, using Lemma 2,

$$\binom{n}{2}^{-1}\sum_{i=1}^{N}\sum_{j=i+1}^{N-1}\|s_i-s_j\|\,\|\gamma_2(\lambda(\nu_i)-\lambda(\nu_j))+u_{2i}-u_{2j}\|=O_p(1). \tag{58}$$

Finally, (55) is shown by following the same arguments in Lemma B.2 from Ahn and Powell (1992), p. 24. More concretely, note that under the rates defined in the lemma for h_2, and using a third order Taylor expansion around the true values, assumption (B.5) and the uniform convergence properties of the first step estimators, then

$$\begin{aligned}&\max_{ij}\left|\hat{\omega}_{ij}-\omega_{ij}\right.\\&\left.-\left(\tfrac{1}{h_2}\right)^2K_2'\left(\tfrac{\lambda(\hat{\nu}_i)-\lambda(\nu_i)}{h_2}\right)\zeta_i\zeta_j\left[(\lambda(\hat{\nu}_i)-\lambda(\nu_i))-(\lambda(\hat{\nu}_j)-\lambda(\nu_j))\right]\right|=O_p\left(h_2^{-3}N^{-\frac{4}{4+s}}\right).\end{aligned} \tag{59}$$

If $s\leq 4$, then equation (60) is $o_p(1)$ because we have assumed that $Nh_2^6\to\infty$. This constraint in s can be weakened again as in the proof of lemma 3, by assumption of higher order smoothness in $\psi_1,\cdots,\psi_s$ and by use of higher order kernels. To finish the proof of the lemma recall first that

$$\begin{aligned}&\binom{n}{2}^{-1}\sum_{i=1}^{N}\sum_{j=i+1}^{N-1}\Xi_{ij}\left[(\lambda(\hat{\nu}_i)-\lambda(\nu_i))-(\lambda(\hat{\nu}_j)-\lambda(\nu_j))\right]\\&=2\binom{n}{2}^{-1}\sum_{i=1}^{N}\sum_{j=i+1}^{N-1}\Xi_{ij}\left(\lambda(\hat{\nu}_i)-\lambda(\nu_i)\right)\end{aligned} \tag{60}$$

Assumption (B.5) allows us to make a Taylor expansion around the true values of the index in the inverse of the Mill's ratio, and therefore (60) can be writen as

$$\begin{aligned}&2\binom{n}{2}^{-1}\textstyle\sum_{i=1}^{N}\sum_{j=i+1}^{N-1}\Xi_{ij}\lambda'(v_i^{dT}\beta+\theta+\sum_{k=1}^{s}\hat{\psi}_k(v_{ik}^c))\\&\times\left[v_i^{dT}\left(\hat{\beta}-\beta\right)+\hat{\theta}-\theta+\textstyle\sum_{k=1}^{s}\left(\hat{\psi}_k(v_{ik}^c)-\psi_k(v_{ik}^c)\right)\right]+O_p\left(\tfrac{1}{N}\right)+O_p\left(\tfrac{1}{Nh_2}\right).\end{aligned} \tag{61}$$

Under assumptions (B.1) to (B.8) we have available the following asymptotic expansions for $\hat{\beta}$ and $\sum_{k=1}^{s}\hat{\psi}_k(v_{ik}^c)$ (see Härdle, Huet, Mammen and Sperlich, 1999; p. 34)

$$\begin{aligned}\hat{\beta}=\beta+\{E(T^2\tilde{v}^d\tilde{v}^{dT})\}^{-1}\frac{1}{N}\sum_{i=1}^{N}\tilde{v}_i^d\frac{\zeta_i-F(\nu_i)}{F(\nu_i)(1-F(\nu_i))}f(\nu_i)\\+O_p\left(\left(h^2+(\log N)^{1/2}(Ngh^{s-1})^{-1/2}\right)^2\right),\end{aligned} \tag{62}$$

and

$$\sup_{v^c\in D_c^*}\left|\Delta(v^c)\right|=O_p\left(\left(h^2+(\log N)^{1/2}(Ngh^{s-1})^{-1/2}\right)^2\right), \tag{63}$$

with

$$\begin{aligned}\Delta(v^c)\;=\;&\hat{\psi}(v^c)-\left\{\overline{\psi}(v^c)\right.\\&+\{E(T^2|v^c)\}^{-1}E(T^2v^{dT}|v^c)\{E(T^2\tilde{v}^d\tilde{v}^{dT})\}^{-1}\\&\left.\times\frac{1}{N}\sum_{i=1}^{N}\tilde{v}_i^d\frac{\zeta_i-F(\nu_i)}{F(\nu_i)(1-F(\nu_i))}f(\nu_i)\right\},\end{aligned} \tag{64}$$

$$\overline{\psi}(v^c) = \psi^+(v^c) + \{E(T^2|v^c)\}^{-1} \frac{1}{N} \sum_{i=1}^{N} \kappa_i(v^c) \frac{\zeta_i - F(\nu_i)}{F(\nu_i)(1-F(\nu_i))} f(\nu_i),$$

$$D_c^* = \{v^c \in D_c : v^c + \eta \in D_c \text{ for all } \eta \text{ with } |\eta_1| \leq g \text{ and } |\eta_j| \leq h\,(j = 2, \ldots, s)\},$$

$$\tilde{v}_i^d = v_i^d - \{E[T_i^2|v_i^c]\}^{-1} E[T_i^2 v_i^d|v_i^c],$$

$$T_i^2 = \frac{f(\nu_i)^2}{F(\nu_i)(1-F(\nu_i))}.$$

If we substitute expressions (62) and (64) into (61) and apply the corresponding rates then the proof is closed.

□

Proof of Theorem 2

The proof of this theorem follows inmediately from Lemmas 3 and 4. To show this, let

$$\hat{\gamma}_1 = \hat{M}_{sz}^{-1} \hat{M}_{sw}. \tag{65}$$

Then, taking into account that $\hat{M}_{sz} = M_{sz} + o_p(1)$, and applying Lemma 4 we obtain

$$\hat{\gamma}_1 = M_{sz}^{-1} \left\{ M_{sw} + \frac{1}{N} \sum_{i=1}^{N} \left\{ A_{1i} \tilde{v}_i^d + A_{2i} \kappa_i(v_i^c) \right\} \frac{\zeta_i - F(\nu_i)}{F(\nu_i)(1-F(\nu_i))} f(\nu_i) \right\} + o_p(1) \tag{66}$$

Apply Lemma 3 and the proof is done.

□

Proof of Theorem 3

The proof of this result follows directly form the proof of Theorem 1 in Härdle, Huet, Mammen and Sperlich (1999).
Let $k(\beta) = E[\ln f\left(\zeta; F\{X^{dT}\beta + \beta_w Z^T \gamma_1 + \eta^+_{\beta,\gamma_1}(X^c)\}\right)]$. We will show that

$$\sup_{\beta \in B} \left| \frac{1}{n} \mathcal{L}(\widehat{\eta}^+_{\beta,\hat{\gamma}_1}, \hat{\gamma}_1, \beta) - k(\beta) \right| \to 0 \qquad \text{(in probability)}. \tag{67}$$

This proves the result we want to show because

$$k''(\beta_0) = E\left[\lambda''\{X^{dT}\beta + \beta_w Z^T \gamma_1 + \eta^+_{\beta,\gamma_1}(X^c)\} \left\{ X + \gamma_{10}^T Z + \frac{\partial \eta^+_{\beta,\gamma_1}}{\partial \beta}(\beta_0, \gamma_{10}, X^c) \right\} \right.$$
$$\left. \times \left\{ X + \gamma_{10}^T Z + \frac{\partial \eta^+_{\beta,\gamma_1}}{\partial \beta}(\beta_0, \gamma_{10}, X^c) \right\}^T \right]$$

is strictly negative definite and $k(\beta_0) = \sup_{\beta \in H} k(\beta)$.

It remains to prove (67). This follows from the following two properties:

$$\sup_{\beta \in B} \left| \frac{1}{n} \mathcal{L}(\eta^+_{\beta,\hat{\gamma}_1}, \hat{\gamma}_1, \beta) - k(\beta) \right| \to 0 \qquad \text{(in probability)}, \tag{68}$$

$$\sup_{\beta \in B} \left| \frac{1}{n} \mathcal{L}(\widehat{\eta}^+_{\beta,\hat{\gamma}_1}, \hat{\gamma}_1, \beta) - \frac{1}{n} \mathcal{L}(\eta^+_{\beta,\hat{\gamma}_1}, \hat{\gamma}_1, \beta) \right| \to 0 \qquad \text{(in probability)}. \tag{69}$$

Claim (68) holds because $\mathcal{L}(\eta^+_{\beta,\hat{\gamma}_1}, \hat{\gamma}_1, \beta)/n$ converges to $k(\beta)$ by the law of large numbers and because $\{\mathcal{L}(\eta^+_{\beta,\hat{\gamma}_1}, \hat{\gamma}_1, \beta), \beta \in B\}$ is tight.
For the proof of tightness note first that

$$\begin{aligned} \left| \frac{1}{n} \mathcal{L}(\eta^+_{\beta_1,\hat{\gamma}_1}, \hat{\gamma}_1, \beta_1) - \mathcal{L}(\eta^+_{\beta_2,\hat{\gamma}_1}, \hat{\gamma}_1, \beta_2) \right| &\leq T_{n,1} \|\beta_1 - \beta_2\| + T_{n,2} \sup_{x^c} |\eta^+_{\beta_1,\hat{\gamma}_1}(x^c) - \eta^+_{\beta_2,\hat{\gamma}_1}(x^c)| \\ &\leq T_{n,1} \|\beta_1 - \beta_2\| + T_{n,2} \sup_{x^c,\beta} \left\| \frac{\partial}{\partial \beta} \eta^+_{\beta,\hat{\gamma}_1}(x^c) \right\| \|\beta_1 - \beta_2\|, \end{aligned}$$

where

$$\begin{aligned} T_{n,1} &= \sup_{\beta,\eta^+} \frac{1}{n} \sum_{i=1}^n \lambda'(X_i^{dT}\beta + \beta_w Z_i^T \hat{\gamma}_1 + \eta^+_{\beta,\hat{\gamma}_1}(X^c)) \|X_i^c\| \|Z_i^T \hat{\gamma}_1\|, \\ T_{n,2} &= \sup_{\beta,\eta^+} \frac{1}{n} \sum_{i=1}^n \lambda'(X_i^{dT}\beta + \beta_w Z_i^T \hat{\gamma}_1 + \eta^+_{\beta,\gamma_1}(X_i^c)). \end{aligned}$$

It is easy to see that, under our conditions, and the consistency of $\hat{\gamma}_1$ (see Theorem 2), $T_{n,1}$ and $T_{n,2}$ are bounded in probability. Following the arguments of Härdle, Huet, Mammen and Sperlich (1999), p.33, it can be seen that $\frac{\partial}{\partial \beta} \eta^+_{\beta,\hat{\gamma}_1}(x^c)$ is uniformly bounded in β and x^c. This shows (68). Claim (69) follows from

$$\begin{aligned} &\sup_{\beta \in B} \left| \frac{1}{n} \mathcal{L}(\widehat{\eta}^+_{\beta,\hat{\gamma}_1}, \hat{\gamma}_1, \beta) - \frac{1}{n} \mathcal{L}(\eta^+_{\beta,\hat{\gamma}_1}, \hat{\gamma}_1, \beta) \right| \\ \leq \sup_{\beta,\eta^+} &\left| \lambda' \left(X_i^{dT}\beta + \beta_w Z_i^T \hat{\gamma}_1 + \eta^+_{\beta,\gamma_1}(X_i^c) \right) \right| \sup_{x^c,\beta} |\widehat{\eta}^+_{\beta,\hat{\gamma}_1}(x^c) - \eta^+_{\beta,\hat{\gamma}_1}(x^c)|. \end{aligned}$$

But then, the result claimed in the theorem has been shown, because

$$\sup_{\beta,\eta^+} \left| \lambda' \left(X^{dT}\beta + \beta_w Z^T \gamma_1 + \eta^+_{\beta,\gamma_1}(X^c) \right) \right| = o_p(1)$$

and

$$\sup_{x^c,\beta} |\widehat{\eta}^+_{\beta,\hat{\gamma}_1}(x^c) - \eta^+_{\beta,\hat{\gamma}_1}(x^c)| = o_p(1).$$

The last two terms are inmediate from the consistency of $\hat{\gamma}_1$ and the proof of Theorem 3.1 from Härdle, Huet, Mammen and Sperlich (1999), p.32.

□

Appendix III: Newton-Raphson algorithm for local likelihood

To facilitate the notation we just write σ for σ_1, and set $K_h = K_{h_1} \times L_{h_2}$. In case of modeling the influence of wage linear, we write nevertheless just β^T for (β^T, β_w) and adjust the vector of regressors accordingly. We start with calculating $\partial \mathcal{L}^S(\eta_j, \beta, \sigma)/\partial \eta_j$ and $\partial^2 \mathcal{L}^S(\eta_j, \beta, \sigma)/\partial \eta_j^2$, where η_j is the function $\eta(\cdot)$ at point t_j. For ease of notation we set $u_{ij} = \{-\beta^T v_i^d - \eta_j\}\sigma^{-1}$ and get

$$\frac{\partial \mathcal{L}^S(\eta_j, \beta, \sigma)}{\partial \eta_j} = \frac{1}{\sigma} \sum_{i=1}^{n} A_n(u_{ij}) K_h(t_j - t_i), \tag{70}$$

$$\frac{\partial^2 \mathcal{L}^S(\eta_j, \beta, \sigma)}{\partial \eta_j^2} = \frac{-1}{\sigma^2} \sum_{i=1}^{n} \left[1 + \frac{u_{ij} f(u_{ij})}{1 - F(u_{ij})} + \frac{f^2(u_{ij})}{\{1 - F(u_{ij})\}^2} \right] K_h(t_j - t_i). \tag{71}$$

where

$$A_n(u_{ij}) = \frac{Y_i}{\sigma} + u_{ij} - \frac{f(u_{ij})}{1 - F(u_{ij})}$$

Moreover, we have used $\partial f(-u/s)/\partial u = -us^{-2} f(-u/s)$.

Next, we calculate $\partial \mathcal{L}(\eta_{\beta,\sigma}, \beta, \sigma)/\partial \beta$ and $\partial^2 \mathcal{L}(\eta_{\beta,\sigma}, \beta, \sigma)/\partial \beta^2$ and denote $\eta_i = \eta_{\beta,\sigma}(t_i)$, $\tilde{s}_{ii} = s_i + \partial \eta_i / \partial \beta$:

$$\frac{\partial \mathcal{L}(\eta_{\beta,\sigma}, \beta, \sigma)}{\partial \beta} = \frac{1}{\sigma} \sum_{i=1}^{n} \left[\frac{Y_i}{\sigma} + u_{ii} - \frac{f(u_{ii})}{1 - F(u_{ii})} \right] \tilde{s}_{ii}. \tag{72}$$

For the Hessian matrix we neglect the dependency of $\tilde{s}_{ii}$ on β and get

$$\frac{\partial \mathcal{L}^2(\eta_{\beta,\sigma}, \beta, \sigma)}{\partial \beta^2} = \frac{-1}{\sigma^2} \sum_{i=1}^{n} \left[1 + \frac{f(u_{ii}) u_{ii}}{1 - F(u_{ii})} - \frac{f^2(u_{ii})}{\{1 - F(u_{ii})\}^2} \right] \tilde{s}_{ii} \tilde{s}_{ii}^T.$$

Here we used $\partial f(u_{ii})/\partial \beta = f(u_{ii}) u_{ii} \tilde{s}_{ii}$.

A little bit more complicated is to get $\partial \mathcal{L}(\eta_{\beta,\sigma}, \beta, \sigma)/\partial \sigma$ and $\partial^2 \mathcal{L}(\eta_{\beta,\sigma}, \beta, \sigma)/\partial \sigma^2$; set $\eta'_\sigma = \partial \eta_{\beta,\sigma}/\partial \sigma$:

$$\frac{\partial \mathcal{L}(\eta_{\beta,\sigma}, \beta, \sigma)}{\partial \sigma} = \frac{-1}{\sigma} \sum_{i=1}^{n} B_n(u_{ii}), \tag{73}$$

where

$$B_n(u_{ii}) = 1 - \left(\frac{Y_i}{\sigma} + u_{ii} \right)^2 - \left\{ \frac{Y_i}{\sigma} + u_{ii} \right\} \eta'_\sigma + \frac{f(u_{ii})(u_{ii} + \eta'_\sigma)}{1 - F(u_{ii})}.$$

For the Hessian matrix we again neglect the dependency of η'_σ on σ and so get with $B_n(u_i i)$ from (73) and get

$$\frac{\partial^2 \mathcal{L}(\eta_{\beta,\sigma}, \beta, \sigma)}{\partial \sigma^2} = \sum_{i=1}^{n} \sigma^{-2} B_n(u_{ii}) + \sigma^{-1} \Bigg[2\sigma^{-1} \{ \frac{Y_i}{\sigma} + u_{ii} \} \{ -\eta' - (\frac{Y_i}{\sigma} + u_{ii}) \} +$$
$$\frac{\eta'_\sigma}{\sigma} \Big\{ -\eta'_\sigma - (\frac{Y_i}{\sigma} + u_{ii}) \Big\} - \frac{f'(u_{ii})(u_{ii} + \eta'_\sigma)}{1 - F(u_{ii})} - \frac{f(u_{ii}) u'_{ii}}{1 - F(u_{ii})} - \frac{f^2(u_{ii}) u'_{ii} (u_{ii} + \eta'_\sigma)}{\{1 - F(u_{ii})\}^2} \Bigg],$$

with

$$u'_{ii} = \frac{\partial u_{ii}}{\partial \sigma} = -\sigma^{-1}\left(u_{ii} + \eta'_\sigma\right)$$

and

$$f'(u_{ii}) = \frac{\partial f(u_{ii})}{\partial \sigma} = \frac{f(u_{ii})}{\sigma}\big(-1 + \eta'_\sigma u_{ii} + u_{ii}^2\big).$$

The question is how to get $\eta'_\beta = \partial\eta/\partial\beta$ and $\eta'_\sigma = \partial\eta/\partial\sigma$. For the likelihood maximizing η_j expression (70) is equal to zero. First we derive it with respect to β:

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}^S(\eta_j,\beta,\sigma)}{\partial\eta_j\partial\beta} &= -\sigma^{-2}\sum_{i=1}^n\left[1 + \frac{f(u_{ij})u_{ij}}{1-F(u_{ij})} + \left\{\frac{f(u_{ij})}{1-F(u_{ij})}\right\}^2\right]K_h(t_j - t_i)\tilde{s}_{ii} = 0 \\
\Longleftrightarrow \eta'_\beta &= \frac{-\sum_{i=1}^n\left[1 + \frac{f(u_{ij})u_{ij}}{1-F(u_{ij})} + \left\{\frac{f(u_{ij})}{1-F(u_{ij})}\right\}^2\right]K_h(t_j - t_i)s_i}{\sum_{i=1}^n\left[1 + \frac{f(u_{ij})u_{ij}}{1-F(u_{ij})} + \left\{\frac{f(u_{ij})}{1-F(u_{ij})}\right\}^2\right]K_h(t_j - t_i)}.
\end{aligned}$$

Second we derive this expression with respect to σ:

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}^S(\eta_j,\beta,\sigma)}{\partial\eta_j\partial\sigma} &= -\sigma^{-2}\sum_{i=1}^n\Bigg[A_n\left(u_{ij}\right) + \eta'_\sigma + \frac{Y_i}{\sigma} + u_{ij} - \frac{f(u_{ij})}{1-F(u_{ij})} + \eta'_\sigma\frac{f(u_{ij})u_{ij}}{1-F(u_{ij})} \\
&\quad + \frac{f(u_{ij})u_{ij}^2}{1-F(u_{ij})} - \frac{f^2(u_{ij})u_{ij}}{\{1-F(u_{ij})\}^2} - \eta'_\sigma\frac{f^2(u_{ij})}{\{1-F(u_{ij})\}^2}\Bigg]K_h(t_j - t_i), \\
\Longleftrightarrow \eta'_\sigma &= \frac{-\sum_{i=1}^n\left[A_n\left(u_{ij}\right) + \frac{Y_i}{\sigma} + u_{ij} - \frac{f(u_{ij})}{1-F(u_{ij})} + \frac{f(u_{ij})u_{ij}^2}{1-F(u_{ij})} - \frac{f^2(u_{ij})u_{ij}}{\{1-F(u_{ij})\}^2}\right]K_h(t_j - t_i)}{\sum_{i=1}^n\left[1 + \frac{f(u_{ij})u_{ij}}{1-F(u_{ij})} - \frac{f^2(u_{ij})}{\{1-F(u_{ij})\}^2}\right]K_h(t_j - t_i)}.
\end{aligned}$$

Finally we need the mixed derivatives $\partial\mathcal{L}(\eta_{\beta,\sigma},\beta,\sigma)/\partial\beta\partial\sigma$ and $\partial\mathcal{L}(\eta_{\beta,\sigma},\beta,\sigma)/\partial\sigma\partial\beta$.

$$\begin{aligned}
&\frac{\partial\mathcal{L}(\eta_{\beta,\sigma},\beta,\sigma)}{\partial\beta\partial\sigma} = \\
&-\sigma^{-2}\sum_{i=1}^n\left[\eta'_\sigma + 2u_{ij} + \frac{\sigma f'_\sigma(u_{ij})}{1-F(u_{ij})} - \frac{f^2(u_{ij})(\eta'_\sigma + u_{ij})}{\{1-F(u_{ij})\}^2} - \frac{f(u_{ij})}{1-F(u_{ij})}\right]\tilde{s}_{ii}
\end{aligned}$$

with $\eta'_\sigma = \partial\eta_{\beta,\sigma}/\partial\sigma$, $\tilde{s}_{ii}$ as above and

$$f'_\sigma(u_{ij}) = \frac{\partial f(u_{ij})}{\partial\sigma} = \frac{f(u_{ij})}{\sigma}\Big(-1 + \eta'_\sigma u_{ij} + u_{ij}^2\Big),$$

$$\begin{aligned}
\frac{\partial\mathcal{L}(\eta_{\beta,\sigma},\beta,\sigma)}{\partial\sigma\partial\beta} &= -\sigma^{-2}\sum_{i=1}^n\Big[2u_{ij} + \eta'_\sigma + \\
&\quad \frac{u_{ij}f(u_{ij})(u_{ij} + \eta'_\sigma)}{1-F(u_{ij})} - \frac{f(u_{ij})}{1-F(u_{ij})} - \left\{\frac{f(u_{ij})}{1-F(u_{ij})}\right\}^2(u_{ij} + \eta'_\sigma)\Big]\tilde{s}_{ii}.
\end{aligned}$$

Here, we have neglected the dependency of η'_σ on β and the dependency of η'_β on σ.

The Hessian matrix for $\mathcal{L}^S$ is simply given by (71).

The Hessian matrix for $\mathcal{L}$ is given by

$$H_{\mathcal{L}} = \begin{pmatrix} \frac{\partial^2\mathcal{L}}{\partial\beta^2} & \frac{\partial^2\mathcal{L}}{\partial\beta\partial\sigma} \\ \frac{\partial^2\mathcal{L}}{\partial\sigma\partial\beta} & \frac{\partial^2\mathcal{L}}{\partial\sigma^2} \end{pmatrix}.$$

References

ANDREWS, D.W. AND N.A. SCHAFGANS (1998) Semiparametric Estimation of the Intercept of a Sample Selection Model. *Review of Economic Studies,* **65**: 497-517.

BOSKIN, M.J. (1973) The economics of Labor Supply. in G. Cain and H. Watts (eds), *Income, Maintenance and Labor Supply.* New-York.

BLUNDELL, R. AND C. MEGHIR (1986) Selection Criteria for a Microeconometric Model of Labour Supply. *Journal of Applied Econometrics,* **1**: 55-80.

FERNANDEZ, A. I. AND J. M. RODRIGUEZ-POO (1997) Estimation and Specification Testing in Female Labor Participation Models: Parametric and Semiparametric Methods. *Econometric Reviews,* **16 (2)**: 229-248.

HALL, R.E. (1973) Wages, Income, and Hours of Work in the U.S. Labour Force. in G. Cain and H. Watts (eds), *Income, Maintenance and Labor Supply.* New-York.

HASTIE, T. J. AND R. J. TIBSHIRANI. (1990) Generalized Additive Models. *Chapman and Hall: London.*

HÄRDLE, H., S. HUET, E. MAMMEN, AND S. SPERLICH (1998) Semiparametric additive indices for binary response and generalized additive models. *Discussion Paper 95, SFP 373, Humboldt-Universität zu Berlin, Germany.*

HECKMAN, J.J. (1979) Sample Selection Bias as a Specification Error. *Econometrica,* **47**: 153-161.

HECKMAN, J.J. (1993) What has been learned about labor supply in the past twenty years? *American Economic Review, Papers and Proceedings,* **83**: 116-121.

KILLINGSWORTH, M AND J. J. HECKMAN (1986) Female labor supply: a survey. *Handbook of Labor Economics,* **vol 1:** chapter 2.

LEE, L. G.S. MADDALA AND R. P. TROST. (1980) Asymptotic covariance matrices of two stage probit and two stage tobit methods for simultaneous equations models with selectivity. *Econometrica,* **48**: 491-503.

LINTON, O. AND J. P. NIELSEN. (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika,* **82**: 93-101.

MANSKI, C.F. (1993) The selection problem in Econometrics and Statistics. in G.S. Maddala, C.R. Rao and H.D. Vinod (eds.), *Handbook of Statistics,* **13**: 73-84.

NEWEY, W.K. AND D. MCFADDEN(1994) Large sample estimation and hypothesis testing. in R.F. Engle and D. McFadden (eds.), *Handbook of Econometrics,* **4**: 2113-2241.

MROZ, T.A. (1987) The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions. *Econometrica,* **55 (4)**: 765-799.

OLSEN, R.J. (1978) Note on the uniqueness of the maximum likelihood estimator for the Tobit model. *Econometrica,* **46**: 1211-1215.

ROSEN (1976) Taxes in a Labor Supply Model with Joint Wage-Hours determination. *Econometrica,* **44**: 485-507.

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

SEVERINI, T.A. AND J. G. STANISWALIS. (1994) Quasi-Likelihood estimation in semi-parametric models. *J. Amer. Statist. Assoc.,* **89**: 501-511.

SPERLICH, S., O.B. LINTON AND W. HÄRDLE. (1999) Integration and Backfitting methods in additive models: Finite sample properties and comparison. *Forthcoming in Test.*

STANISWALIS, J.G.. (1989) The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association,* **84**: 276-283.

STONE, C.J. (1986) The dimensionality reduction principle for generalized additive models. *Ann. Statist.,* **14**: 590-606.

TJØSTHEIM, D. AND B.H. AUESTAD. (1994) Nonparametric identification of nonlinear time series: projections. *J. American Statistical Association,* **89**: 1398-1409.

VELLA F. (1998) Estimating Models with Sample Selection Bias: A Survey. *Journal of Human Resources,* **33 (1)**: 127-169.

VIJVERBERG, W.P. (1991) Selectivity and Distributional Assumptions in Static Labor Supply Models. *Southern Economic Journal,* **57**: 822-840.

WALES, T.J. AND A.D. WOODLAND (1980) Sample Selectivity and the Estimation of labor Supply Functions. *International Economic Review,* **21**: 437-468.