Essays in Labor Econometrics

Von der Wirtschaftswissenschaftlichen Fakultät der
Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften
- Doctor rerum politicarum -

genehmigte Dissertation

von

Diplom-Volkswirt Jörg Schwiebert

geboren am 08.02.1986 in Rotenburg (Wümme)

2014

Referent: Prof. Dr. Patrick Puhani

Korreferent: Prof. Dr. Stephan L. Thomsen

Tag der Promotion: 11.03.2014

Danksagung

Ich bedanke mich bei Patrick Puhani für die Betreuung meiner Promotion und für die Unterstützung bei der Realisierung meiner eigenen Forschungsideen.

Ich bedanke mich bei Olaf Hübler und Melanie Schienle für die fachliche Unterstützung auf dem Gebiet der Ökonometrie. Ich bedanke mich außerdem bei Jeff Wooldridge, Michael Lechner und diversen Seminar- und Konferenzteilnehmern für ihre fachlichen Kommentare.

Ein besonderer Dank geht an meine Kollegen und die studentischen Hilfskräfte für die angenehme Arbeitsatmosphäre.

Ich danke meiner Familie, die mir dies alles ermöglicht hat.

Summary

This dissertation deals with the development of econometric estimators and their application to problems from the field of labor economics. It includes six essays, from whom four are on the sample selection model and the remaining two are on different topics in labor econometrics. Chapters 2 and 4 consider an extension of sample selection models to the case of endogenous covariates. While the focus of Chapter 2 is on issues of interpretation, Chapter 4 centers on consistent estimation of the model parameters under weak assumptions. Both chapters include empirical applications where educational attainment is considered an endogenous variable in wage equations. Thus, both chapters contain estimates of the so-called returns to education. Chapter 3 deals with the estimation of a sample selection model using copulas. Copulas provide a very flexible modeling device and estimation carried out in this way has several advantages over existing estimators proposed in the literature. Chapter 5 considers semiparametric estimation strategies for a sample selection model with a binary dependent variable. The finite sample properties of the proposed estimators are illustrated by means of a Monte Carlo study and an empirical example. Chapter 6 proposes a detailed decomposition method for limited dependent variable models, which has important advantages over approaches already presented in the literature. Chapter 7 considers identification and estimation of endogenous regressor models when the endogenous regressor is discrete. The identification strategy does not require an additional instrumental variable and is, thus, especially valuable if such an instrumental variable is unavailable in empirical applications.

Keywords: sample selection model, endogenous covariates, decomposition

Zusammenfassung

Diese Dissertation befasst sich mit der Entwicklung ökonometrischer Schätzmethoden und deren Anwendung auf Problemstellungen aus dem Bereich der Arbeitsökonomik. Sie enthält sechs Essays, von denen sich vier mit dem Stichprobenselektionsmodell beschäftigen und die übrigen zwei sich mit verschiedenen Themen der Arbeitsökonometrie befassen. Kapitel 2 und 4 betrachten eine Erweiterung des Stichprobenselektionsmodells auf endogene Kovariate. Während sich Kapitel 2 mit Fragen der Interpretation auseinandersetzt, fokussiert sich Kapitel 4 auf die konsistente Schätzung der Modellparameter unter wenig restriktiven Annahmen. Beide Kapitel enthalten empirische Anwendungsbeispiele, in denen die Schulbildung als endogene Variable in Lohngleichungen betrachtet wird. Beide Kapitel enthalten somit Schätzungen der sogenannten Bildungsrendite. Kapitel 3 befasst sich mit der Schätzung eines Stichprobenselektionsmodells mit Hilfe von Copulas. Copulas bieten einen sehr flexiblen Modellierungsansatz, der Vorteile gegenüber bereits bestehenden Ansätzen aufweist. Kapitel 5 betrachtet die semiparametrische Schätzung eines Stichprobenselektionsmodells mit einer binären abhängigen Variablen. Die Eigenschaften der vorgeschlagenen Schätzmethoden wird mit Hilfe einer Monte Carlo Studie sowie eines empirischen Beispiels illustriert. In Kapitel 6 wird eine detaillierte Dekompositionsmethode für Modelle mit begrenzt abhängiger Variable vorgeschlagen, die signifikante Vorteile gegenüber bereits bestehenden Methoden aufweist. Kapitel 7 betrachtet die Identifikation und Schätzung eines Modells mit einem diskreten endogenen Regressor. Die Identifikationsstrategie beansprucht keine zusätzliche Instrumentvariable und ist daher insbesondere nützlich, wenn eine solche Instrumentvariable in empirischen Anwendungen nicht verfügbar ist.

Schlagworte: Stichprobenselektionsmodell, endogene Kovariate, Dekomposition

Contents

List of Tables

Chapter 1

Main Introduction

The field of labor economics is characterized by a considerable amount of empirical studies. The research methods used to conduct such studies are taken from the related field of econometrics. In fact, labor economics on the one hand and econometrics on the other hand are highly interrelated and stimulating each other. Many important researchers are experts in both labor economics and econometrics, most notably Nobel prize laureate James J. Heckman. This dissertation deals with the development and application of econometric methods in the field of labor economics. Thus, this dissertation is concerned with what may be called "labor econometrics".[1]

Heckman won the Nobel prize for his seminal contributions to the analysis of selective samples. The well-known sample selection model is intrinsically tied to his name. The sample selection model plays an important role in this dissertation.

The sample selection model is used when the observed sample is considered a non-random sample from the overall population. The term "observed" should be understood in the sense that all variables are observed. A popular example for an application of the sample selection model is estimating a wage equation for women. A wage equation has the (natural logarithm) of the wage as the dependent variable and a set of

[1] This term has been used by Heckman and MaCurdy (1986), for example.

explanatory variables on the right hand side, also known as covariates. However, the wage is only observed for women who are working, while the remaining women have "missing" wages. Heckman's (1979) crucial point was that performing an ordinary least squares regression on the observed sample of working women only introduces a bias into the estimation results. To overcome this bias, he proposed the well known Heckman correction. This involves augmenting the right hand side of the wage equation with a control function, the inverse Mills ratio term, which controls for the probability of belonging to the observed sample. Heckman (1979) demonstrated that a simple two-step procedure is suited (under certain assumptions) to obtain consistent estimates of the parameters of interest. In the first step, the selection equation which determines the probability of belonging to the observed sample is estimated. These estimates are used to compute the inverse Mills ratio term, which is inserted as an additional covariate into the wage equation (the main equation). The second step is then an ordinary least squares regression of the wage on the explanatory variables and the inverse Mills ratio term.

The sample selection model in its original formulation due to Heckman (1979) relies on some critical assumptions. One assumption is that the covariates are exogenous, i.e., independent of the error terms of the model. However, this assumption may be doubtful in practice. For instance, in a wage equation a typical covariate is educational attainment. However, unobservables such as the ability (e.g., intelligence or social background) of a woman are likely to affect the wage, the probability of belonging to the workforce and educational attainment jointly. In that case, educational attainment cannot be regarded as exogenous. In Chaper 2 of this dissertation, the sample selection model due to Heckman (1979) is extended to the case of endogenous covariates. An appropriate econometric model is developed and applied to a female wage equation example due to Mulligan and Rubinstein (2008). It is shown that a correlation parameter in the extended model has the same interpretation as its counterpart in the selection

model without endogenous covariates, and can be used to study if the observed sample has above-average skills.

Another critical assumption is the bivariate normality assumption in Heckman's (1979) original formulation. Heckman (1979) assumed that the error terms of main and selection equation have a bivariate normal distribution. However, if this assumption is not fulfilled, estimates are generally inconsistent. In Chapter 3, a semiparametric estimation procedure is proposed which relaxes the strict bivariate normality assumption. It is argued that this procedure has several advantages over a competing approach proposed by Gallant and Nychka (1987).

Chapter 4 also studies endogenous covariates in a selection model. In contrast to the model proposed in Chapter 2, the model in Chapter 4 does not rely on parametric distributional assumptions. Chapter 4 focuses on consistent estimation of the parameters of main and selection equation under weak assumptions, while Chapter 2 centers on issues of interpretation. The model is applied to estimating the (married) female returns to education. It is demonstrated that it is important to account for the endogeneity of education, and the empirical results indicate that the returns to education seem to be smaller than those obtained by ordinary least squares or an ordinary Heckman selection model which does not control for the endogeneity of education.

Chapter 5 considers a sample selection model with a binary dependent variable in the main equation. In the ordinary sample selection model, the dependent variable of the outcome equation is continuous. In Chapter 5, semiparametric estimators are proposed which do not rely on strong distributional assumptions. In particular, two different two-step approaches for estimation are presented and discussed, and the performance of the estimators is evaluated by means of a Monte Carlo study and an empirical example.

Chapters 6 and 7 are methodological contributions to two different topics of labor econometrics. In Chapter 6, a detailed decomposition for limited dependent variable models is proposed. While in case of linear models the well-known Blinder-Oaxaca

decomposition can be applied, this is not possible in nonlinear models such as limited dependent variable models. A detailed decomposition is proposed which has significant advantages over two methods already existing in the literature.

Chapter 7 deals with the identification and estimation of endogenous regressor models when the endogenous regressor is discrete. The virtue of this approach is that no additional instrumental variable is needed for identification. It is shown that the discreteness of the endogenous regressor implies a nonlinear relationship between the endogenous regressor and the remaining explanatory variables, which can be exploited for identification. The usefulness of the approach is illustrated by a Monte Carlo study and an empirical application.

Chapter 2

Estimation and Interpretation of a Heckman Selection Model with Endogenous Covariates

This chapter is a revision of the discussion papers No. 483 and 502, Department of Economics and Business Administration, Leibniz University Hannover (Schwiebert, 2011; Schwiebert, 2012a). I thank Olaf Hübler, Patrick Puhani, Bernd Fitzenberger and three anonymous referees for providing valuable comments.

2.1 Introduction

Researchers using the Heckman (1979) selection model often implicitly assume exogenous covariates. In this chapter, we challenge this sometimes questionable assumption and develop a Heckman selection model with endogenous covariates. While the issue of endogeneity in sample selection models is not novel, our approach has two important advantages. First, estimation can be carried out fairly easily; any econometrics software which supports maximum likelihood estimation of the Heckman selection model can be

used to implement our estimator. Second, our approach provides a measure to analyze the composition of the observed sample[1] with respect to unobservables. That is, a measure which indicates whether individuals in the observed sample have higher outcomes on average than people from the unobserved sample (given the covariates). Having a measure to analyze the composition of the observed sample with respect to unobservables is important. For example, our model can be applied to study the composition of the female workforce, as has been done by Mulligan and Rubinstein (2008).

In their 2008 *QJE* paper, Mulligan and Rubinstein (2008) studied the development of the gender wage gap in the U.S. over time. They sought to obtain an explanation why the wage gap *between* genders has narrowed over time, while it has increased *within* gender. Mulligan and Rubistein (2008) hypothesized that these developments can be explained by an increase in the quality of the female workforce. They provided some evidence supporting this hypothesis.

One of the methods used by Mulligan and Rubinstein (2008) to study the quality of the female workforce was the Heckman selection model (other methods involved identification at infinity and some evidence on the IQ of women). The Heckman model consists of two equations, the main equation of interest and the selection equation, where the latter determines whether an observation belongs to the observed sample. In the Mulligan and Rubinstein (2008) study, the main equation of interest is a log wage equation for women, and the selection equation is equivalent to a labor force participation equation. Using this model, Mulligan and Rubinstein (2008) analyzed the selection of women into the full time full year workforce (defined as women who worked at least 35 hours per week and 50 weeks during the year) and found that the selection has become "more positive" over time, indicating a shift in the quality of the female workforce.

The Heckman selection model can be used to analyze such issues because, as it is

[1]By "observed sample" we mean those observations who have nonmissing values in all variables.

well known, the main equation of interest can be augmented by a control function, the inverse Mills ratio term, for the observed sample. The parameter associated with the inverse Mills ratio is the product of a standard deviation parameter (which is necessarily positive) and the correlation coefficient between the error terms of main and selection equation. Since the inverse Mills ratio is always nonnegative, a positive correlation coefficient implies that the individuals in the observed sample have a higher (potential) log wage (if the log wage is the outcome variable) on average than women who are not observed, conditional on the covariates. That is, individuals with identical covariates differ in their (potential) wages depending on whether they belong to the observed sample or not. If the correlation coefficient is positive and thus the individuals in the observed sample get a higher wage than individuals from the non-observed sample, this can be interpreted such that the individuals from the observed sample have a higher quality (or skills) on average. Indeed, the inverse Mills ratio term can be interpreted as a variable which captures differences in unobservables between the observed and the unobserved sample, e.g. the quality or skills of individuals.

In their application of the Heckman selection model, Mulligan and Rubinstein (2008) implicitly assumed that the covariates entering this model were exogenous. This assumption is, however, questionable for a variable like education. It is likely that common unobservable factors like ability drive the (potential) wage, the probability of labor force participation and education jointly. In that case, education cannot be regarded as exogenous.

In this chapter, we develop a Heckman selection model which allows for endogenous covariates. Endogenous covariates are allowed to enter the main equation only, the selection equation only, or both. Thus, our model is sufficiently general to accommodate all cases of endogeneity. It is important to note that we develop an extension of the Heckman selection model with the same basis assumptions. In particular, we extend Heckman's (1979) bivariate normality assumption and assume that the error terms of

main equation, selection equation and the reduced form equations for the endogenous covariates have a multivariate normal distribution.

Our estimators are conceptually similar to estimators for the Tobit model with endogenous covariates as provided by Smith and Blundell (1986) and the probit model with endogenous covariates as provided by Rivers and Vuong (1988); see also Newey (1987). These estimators are implemented in standard econometrics software (such as STATA) and are frequently used by practitioners.

One might argue that our parametric assumptions (i.e., multivariate normality) are too strong. Indeed, following Heckman's (1979) original setup of the model using the bivariate normality assumption, several authors have challenged this assumption and provided semi-nonparametric estimators which are consistent under weaker assumptions; e.g., Gallant and Nychka (1987), Powell (1987), Ahn and Powell (1993), Das et al. (2003) and Newey (2009). However, what these estimators do not provide is an easy-to-interpret measure characterizing the observed sample with respect to unobservables. In the Mulligan and Rubinstein (2008) example, the correlation coefficient is the easy-to-interpret measure which shows if individuals in the observed sample have a higher quality on average than individuals from the non-observed sample.

To analyze issues like the composition of the female workforce, it is thus desirable to have such a measure which describes the quality of the (observed) workforce. We will show that our extension of the Heckman selection model also includes such a parameter with the same interpretation (and which is also a correlation parameter). Hence, our model can be used to study compositional issues such as the composition of the female workforce.

We apply our framework to 1980 U.S. Census data. We specify our model similar to Mulligan and Rubinstein (2008), but we allow the covariate education to be endogenous. In this data set we have information on the quarter of birth of individuals. This information can be used to form instrumental variables for education (Angrist

and Krueger, 1991). Details are given below. We examine if the ordinary Heckman model (without controlling for endogenous covariates) and our extended model (which controls for the endogeneity of education) lead to different results regarding the quality of the female workforce. We thus seek to answer the question whether the conclusions made by Mulligan and Rubinstein (2008) are valid, even if their model might have been misspecified (as they did not control for the endogeneity of education).

Besides providing a measure characterizing the observed sample with respect to unobservables, our parametric modeling approach has some significant advantages. First, as already mentioned, we allow for endogenous covariates in the selection equation. Wooldridge (2010) and Semykina and Wooldridge (2010) have also provided estimators for the Heckman selection model with endogenous covariates, but in their specification covariates are only allowed to be endogenous in the main equation.[2] Second, our approach does not require the existence of a variable (directly) affecting the selection equation but not the main equation. Such an exclusion restriction is generally needed in semi-nonparametric models to identify the parameters of the main equation (also in semi-nonparametric models which allow for endogenous covariates, e.g. Das et al., 2003). In our model, as in the ordinary Heckman selection model, identification is achieved by our functional form assumptions, hence no additional "instrumental" variable is needed to enter the selection equation. Even if one questions our parametric assumptions, our estimation framework may serve as a starting point for an exploratory data analysis, which may be followed by a more appropriate, e.g. semi-nonparametric, estimation strategy afterwards. Our model setup is similar to Chib et al. (2009), but their estimation strategy is Bayesian, whereas ours is not. A great advantage of our estimator is that it is easy to apply. As will be shown below, any econometrics software

[2] In a related approach, Blundell et al. (1998) estimated labor supply elasticities, controlling for endogeneity of the wage and other income, and for selection into the labor force. They augmented the main equation with a control function for the endogeneity of the wage and other income and an inverse Mills ratio term for sample selectivity. However, this approach only works if there is no endogeneity in the selection equation.

which is capable of maximum likelihood estimation of the ordinary Heckman selection model can be used to implement our estimator.

Mroz (1987) suggested to get rid of the endogenous covariate by replacing it with a reduced form equation which depends on exogenous covariates only. Then, the ordinary Heckman model could be applied. Such a strategy is appropriate if one is interested in the parameters of the main equation. However, it is not clear what the correlation coefficient from this model measures, i.e. if it can be used to study compositional issues. Heckman (1978) also considered endogenous covariates in a more general model, but also focused mainly on the coefficients of the explanatory variables (and not on correlation parameters and alike).

The remainder of the chapter is organized as follows. In Section 2.2, we set up the econometric model which allows for the simultaneous presence of sample selectivity and endogeneity. Section 2.3 presents the estimation strategies and shows how the latter can be implemented in standard econometrics software. We also derive the analogue of the correlation coefficient from the ordinary Heckman selection model and show that both have the same interpretation. Moreover, we provide tests which indicate whether endogeneity of covariates and/or sample seletivity are indeed present. In Section 2.4, we apply our model to 1980 U.S. Census data and compare its results with the results obtained from the ordinary Heckman selection model. Section 2.5 concludes the chapter.

2.2 Econometric Model

In this section, we present a rather general framework for incorporating endogenous covariates into the Heckman selection model. The reason is that endogenous covariates may occur in three respects. First, endogenous covariates may only appear in the main but not in the selection equation; second, endogenous covariates may appear only in the selection but not in the main equation; and third, endogenous covariates may appear

in both equations. Thus, we set up a relatively general model to cover all these cases. The model is given by

$$y_i^* = X_{1i}\beta_1 + X_{2i}\beta_2 + C_i\beta_3 + u_i \equiv X_i\beta + u_i \tag{2.1}$$

$$z_i^* = W_{1i}\gamma_1 + W_{2i}\gamma_2 + C_i\gamma_3 + Q_i\gamma_4 + v_i \equiv W_i\gamma + v_i \tag{2.2}$$

$$X_{2i} = [X_{1i}, W_{1i}]\Delta_1 + Z_{1i}\Delta_2 + \varepsilon_{1i} \equiv \tilde{Z}_{1i}\Delta + \varepsilon_{1i} \tag{2.3}$$

$$W_{2i} = [X_{1i}, W_{1i}]\Lambda_1 + Z_{2i}\Lambda_2 + \varepsilon_{2i} \equiv \tilde{Z}_{2i}\Lambda + \varepsilon_{2i} \tag{2.4}$$

$$C_i = [X_{1i}, W_{1i}]\Upsilon_1 + Z_{3i}\Upsilon_2 + \varepsilon_{3i} \equiv \tilde{Z}_{3i}\Upsilon + \varepsilon_{3i} \tag{2.5}$$

$$z_i = 1(z_i^* > 0) \tag{2.6}$$

$$y_i = y_i^* z_i, \tag{2.7}$$

where $i = 1, \ldots, n$ indexes individuals. The first equation is the main equation, where the latent dependent variable y^* is related to a $(1 \times K_1)$-vector of exogenous explanatory variables, X_1, to a $(1 \times K_2)$-vector of endogenous explanatory variables only included in the main equation but not in the selection equation, X_2, and to a $(1 \times P)$-vector of endogenous explanatory variables included in the main and the selection equation, C. The second equation is the selection equation, where the latent variable z^* is related to a $(1 \times L_1)$-vector of exogenous explanatory variables, W_1, to a $(1 \times L_2)$-vector of endogenous explanatory variables, W_2 only included in the selection equation but not in the primary equation, to C and to Q. Q is an exogenous variable (it could also be a vector) which appears only in the selection equation. This is a well-known exclusion restriction serving to identify the parameters of the main equation. In equations (2.3) to (2.5) it is assumed that the endogenous explanatory variables can be explained by a $(1 \times M_1)$-vector, a $(1 \times M_2)$-vector and a $(1 \times M_3)$-vector of instrumental variables, Z_1, Z_2 and Z_3, respectively. Equation (2.6) expresses that only the sign of z^* is observable. Finally, equation (2.7) comprises the selection mechanism, i.e. the latent variable y^* is only observed if the selection indicator z is equal to one. Equations (2.1), (2.2), (2.6),

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

Note that the exclusive presence of Q in the selection equation, i.e. the validity of an exclusion restriction, is not needed to identify the parameters of the main equation, as our functional form assumptions are sufficient for identification. Nevertheless, we include this variable since some researchers may not want to identify parameters by functional form assumptions alone. By contrast, the instrumental variables appearing in the reduced form equations for the endogenous explanatory variables do have to fulfill an exclusion restriction. These variables may not appear in X or W.

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

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

where NID denotes "normally and independently distributed", $J \equiv K_2 + L_2 + P$, and the distribution should be interpreted as conditional on all exogenous variables (the conditioning has been omitted for the ease of notation). The covariance matrix of the error terms consists of four parts. The upper left part is the covariance matrix attributed to the error terms of the main and selection equation, respectively, where

σ_u^2 and σ_v^2 denote the variances of u and v, and ρ denotes the correlation coefficient. If there was no concern about endogeneity, inference would be based solely on this part of the covariance matrix, as it is common in the standard sample selection model. However, the (potential) presence of endogeneity is indicated by the $(J \times 2)$-matrix Ω, which captures the influence of unobserved factors which jointly affect the dependent variables in equation (2.1) and (2.2) and the endogenous explanatory variables. Note that endogeneity is absent if and only if Ω is equal to the null matrix. Finally, the error terms attributed to the endogenous explanatory variables have covariance matrix Σ whose dimension is $(J \times J)$.

Note that it is assumed that the distribution of the endogenous covariates can be reasonably approximated by a normal distribution, which favors continuous regressors and excludes binary regressors. However, even in case of binary regressors our model can be applied for exploratory data analysis.

2.3 Estimation, Interpretation and Testing for Exogeneity

2.3.1 Estimation

First, we lay out a full information maximum likelihood procedure in which all parameters of the model (2.1)-(2.7) are estimated simultaneously. Note that the conditional distribution of $(u_i, v_i)'$ given $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})$ is given by

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \Bigg| \varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i} \sim \text{NID}\left(\Omega'\Sigma^{-1}\begin{bmatrix} \varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i} \end{bmatrix}', B\right) \tag{2.9}$$

where

$$B \equiv \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix} - \Omega'\Sigma^{-1}\Omega. \tag{2.10}$$

Define

$$\Psi \equiv \underset{(2\times J)}{\begin{pmatrix} \underset{(1\times K_2)}{\psi_{11}} & \underset{(1\times L_2)}{\psi_{12}} & \underset{(1\times P)}{\psi_{13}} \\ \underset{(1\times K_2)}{\psi_{21}} & \underset{(1\times L_2)}{\psi_{22}} & \underset{(1\times P)}{\psi_{23}} \end{pmatrix}} \equiv \Omega'\Sigma^{-1} \tag{2.11}$$

$$\Gamma \equiv \begin{pmatrix} \tilde{\sigma}^2 & \tilde{\rho}\tilde{\sigma} \\ \tilde{\rho}\tilde{\sigma} & 1 \end{pmatrix} \equiv \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix} - \Omega'\Sigma^{-1}\Omega, \tag{2.12}$$

where the lower right element of Γ has been set equal to unity due to normalization. Therefore, equation (2.9) can be recast as

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \Bigg| \varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i} \sim \text{NID}\left(\begin{bmatrix} \psi_{11}\varepsilon_{1i}' + \psi_{12}\varepsilon_{2i}' + \psi_{13}\varepsilon_{3i}' \\ \psi_{21}\varepsilon_{1i}' + \psi_{22}\varepsilon_{2i}' + \psi_{23}\varepsilon_{3i}' \end{bmatrix}, \begin{pmatrix} \tilde{\sigma}^2 & \tilde{\rho}\tilde{\sigma} \\ \tilde{\rho}\tilde{\sigma} & 1 \end{pmatrix} \right), \tag{2.13}$$

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

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

$$l(\theta) = \sum_{z_i=0} \log\{\Phi(-W_i\gamma - \psi_{21}\varepsilon_{1i}' - \psi_{22}\varepsilon_{2i}' - \psi_{23}\varepsilon_{3i}')\}$$

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

$$
\begin{aligned}
&+ \sum_{z_i=1} \log\{\tilde{\sigma}^{-1}\phi(\tilde{\sigma}^{-1}(y_i - X_i\beta - \psi_{11}\varepsilon'_{1i} - \psi_{12}\varepsilon'_{2i} - \psi_{13}\varepsilon'_{3i}))\} \\
&+ \sum_{z_i=1} \log\{\Phi((1-\tilde{\rho}^2)^{-1/2}[W_i\gamma + \psi_{21}\varepsilon'_{1i} + \psi_{22}\varepsilon'_{2i} + \psi_{23}\varepsilon'_{3i} \\
&\qquad + \tilde{\rho}\tilde{\sigma}^{-1}(y_i - X_i\beta - \psi_{11}\varepsilon'_{1i} - \psi_{12}\varepsilon'_{2i} - \psi_{13}\varepsilon'_{3i})])\} \\
&- \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n} \begin{bmatrix} \varepsilon_{1i} & \varepsilon_{2i} & \varepsilon_{3i} \end{bmatrix} \Sigma^{-1} \begin{bmatrix} \varepsilon_{1i} & \varepsilon_{2i} & \varepsilon_{3i} \end{bmatrix}', \qquad (2.14)
\end{aligned}
$$

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

$$\varepsilon_{1i} = X_{2i} - \tilde{Z}_{1i}\Delta \qquad (2.15)$$

$$\varepsilon_{2i} = W_{2i} - \tilde{Z}_{2i}\Lambda \qquad (2.16)$$

$$\varepsilon_{3i} = C_i - \tilde{Z}_{3i}\Upsilon, \qquad (2.17)$$

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

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

$$\hat{\theta} = \arg\max_{\theta}\ l(\theta). \qquad (2.18)$$

The FIML estimator actually does not provide estimates of the "structural" variance-covariance parameters, i.e., those parameters in the unconditional distribution of the error terms. Such parameters might be interesting; for example, the variance of the main equation's error term may be used as a measure of inequality of the skill distribution. These structural parameters can be deduced from the FIML estimates by noting that

$$\hat{\Pi} = \hat{\Gamma} + \hat{\Psi}\hat{\Sigma}\hat{\Psi}' \qquad (2.19)$$

$$\hat{\Omega} = \hat{\Sigma}\hat{\Psi}' \tag{2.20}$$

$$\hat{\rho} = \frac{\hat{g}}{\hat{\sigma}_u \hat{\sigma}_v}, \tag{2.21}$$

where $\Pi \equiv \begin{pmatrix} \sigma_u^2 & g \\ g & \sigma_v^2 \end{pmatrix}$ and $g \equiv \rho\sigma_u\sigma_v$. In the Appendix it is shown how standard errors for these structural estimates can be derived by means of the delta method.[4]

The FIML estimator is fully efficient. However, if the number of observations is large and/or the number of covariates is large, estimation may be quite time consuming. As an alternative, one may consider choosing a limited maximum likelihood (LIML) approach. We propose the following procedure:

1) Estimate the reduced form equations (2.3)-(2.5) by OLS and obtain the residuals $\hat{\varepsilon}_1$, $\hat{\varepsilon}_2$ and $\hat{\varepsilon}_3$.

2) Insert these estimated values into the following log-likelihood function

$$\begin{aligned} l(\tilde{\theta}) = &\sum_{z_i=0} \log\{\Phi(-W_i\gamma - \psi_{21}\hat{\varepsilon}'_{1i} - \psi_{22}\hat{\varepsilon}'_{2i} - \psi_{23}\hat{\varepsilon}'_{3i})\} \\ &+ \sum_{z_i=1} \log\{\tilde{\sigma}^{-1}\phi(\tilde{\sigma}^{-1}(y_i - X_i\beta - \psi_{11}\hat{\varepsilon}'_{1i} - \psi_{12}\hat{\varepsilon}'_{2i} - \psi_{13}\hat{\varepsilon}'_{3i}))\} \\ &+ \sum_{z_i=1} \log\{\Phi((1-\tilde{\rho}^2)^{-1/2}[W_i\gamma + \psi_{21}\hat{\varepsilon}'_{1i} + \psi_{22}\hat{\varepsilon}'_{2i} + \psi_{23}\hat{\varepsilon}'_{3i} \\ &\qquad + \tilde{\rho}\tilde{\sigma}^{-1}(y_i - X_i\beta - \psi_{11}\hat{\varepsilon}'_{1i} - \psi_{12}\hat{\varepsilon}'_{2i} - \psi_{13}\hat{\varepsilon}'_{3i})])\}, \end{aligned} \tag{2.22}$$

which is then maximized over $\tilde{\theta} \equiv (\beta', \gamma', \tilde{\rho}, \tilde{\sigma}, \mathrm{vec}(\Psi)')$.

Observe that the log-likelihood function is the same as for the Heckman selection model without endogenous covariates, with the difference that we have the additional covari-

[4] We also provide in the Appendix a small Monte Carlo simulation study which analyzes the finite sample performance of the FIML estimator and compares its estimates to the (biased) estimates based on the ordinary Heckman selection model which does not control for endogeneity. Moreover, we provide an application of our estimator to the well-known Mroz (1987) labor supply data set in order to compare our results with those of Wooldridge (2010), who did the same using his estimator.

ates $\hat{\varepsilon}_1$, $\hat{\varepsilon}_2$ and $\hat{\varepsilon}_3$. Thus, our model can be estimated using any econometrics software which supports maximum likelihood estimation of the Heckman selection model. One must simply add to the set of covariates the estimated residuals $\hat{\varepsilon}_1$, $\hat{\varepsilon}_2$ and $\hat{\varepsilon}_3$.

Of course, using estimated residuals as covariates instead of the true error terms requires an adjustment of the (asymptotic) standard errors. To get appropriate standard errors, one can either

a) use a correction formula which gives that $\sqrt{n}(\hat{\tilde{\theta}} - \tilde{\theta}) \xrightarrow{d} \mathcal{N}(0, C)$, where C is the corrected asymptotic covariance matrix which accounts for the estimation error in $\hat{\varepsilon}_1$, $\hat{\varepsilon}_2$ and $\hat{\varepsilon}_3$. The exact expression for C is provided in the appendix;

b) combine the first order conditions from maximizing the limited information log-likelihood function with the normal equations for estimating the reduced form equations for the endogenous explanatory variables and estimate the parameters jointly in a generalized method of moments (GMM) framework;

c) use the bootstrap.

2.3.2 The Interpretation of $\tilde{\rho}$

In this subsection we show that $\tilde{\rho}$ has the same interpretation as the correlation coefficient in the ordinary Heckman selection model. To keep the notation easy, we consider the following simple model with one endogenous explanatory variable:

$$y_i^* = X_{1i}\beta_1 + C_i\beta_3 + u_i \equiv X_i\beta + u_i \tag{2.23}$$

$$z_i^* = W_{1i}\gamma_1 + C_i\gamma_3 + v_i \equiv W_i\gamma + v_i \tag{2.24}$$

$$C_i = \tilde{Z}_{3i}\Gamma + \varepsilon_{3i} \tag{2.25}$$

$$z_i = 1(z_i^* > 0) \tag{2.26}$$

$$y_i = y_i^* z_i. \tag{2.27}$$

The notation is the same as in Section 2.2. Moreover, let

$$\begin{bmatrix} u_i \\ v_i \\ \varepsilon_{3i} \end{bmatrix} \sim NID \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} & \omega_u \\ \sigma_{uv} & \sigma_v^2 & \omega_v \\ \omega_u & \omega_v & \sigma_\varepsilon^2 \end{bmatrix} \right) \tag{2.28}$$

For the observable part of the main equation we have that (with a slight abuse of notation[5])

$$E[y^*|z=1, X, W, \tilde{Z}_3] = X\beta + E[u|z=1, \varepsilon_3] \tag{2.29}$$

$$= X_i\beta + E[u|v > -W\gamma, \varepsilon_3], \tag{2.30}$$

where we have suppressed some explanatory variables from the conditional expectations on the RHS because these are not crucial. The term $E[u|v > -W\gamma, \varepsilon_3]$ is a generalization of the inverse Mills ratio term from the ordinary Heckman selection model. In contrast to the ordinary selection model, this terms controls not only for sample selectivity but for the endogeneity of C as well. The term can be written as

$$E[u|v > -W\gamma, \varepsilon_3] = \int u \frac{Pr(v > -W\gamma|u, \varepsilon_3) f_u(u|\varepsilon_3) f_\varepsilon(\varepsilon_3)}{Pr(v > -W\gamma|\varepsilon_3) f_\varepsilon(\varepsilon_3)} du, \tag{2.31}$$

where $f_i(\cdot)$ denotes the probability density function of variable i.

By using laws for conditional normal distributions, we have that v conditional on u and ε_3 has a normal distribution with mean

$$E[v|u, \varepsilon_3] = \begin{pmatrix} \sigma_{uv} & \omega_v \end{pmatrix} \begin{pmatrix} \sigma_u^2 & \omega_u \\ \omega_u & \sigma_\varepsilon^2 \end{pmatrix}^{-1} \begin{pmatrix} u \\ \varepsilon_3 \end{pmatrix} \tag{2.32}$$

$$= (\sigma_u^2\sigma_\varepsilon^2 - \omega_u^2)^{-1}\{(\sigma_{uv}\sigma_\varepsilon^2 - \omega_u\omega_v)u + (\omega_v\sigma_u^2 - \sigma_{uv}\omega_u)\varepsilon_3\} \tag{2.33}$$

[5]We should e.g. write $E[y^*|z=1, X=x, W=w, \tilde{Z}_3=\tilde{z}_3]$ instead of $E[y^*|z=1, X, W, \tilde{Z}_3]$, but we keep the notation simple to avoid long formulas.

and variance

$$Var[v|u,\varepsilon_3] = \sigma_v^2 - \begin{pmatrix}\sigma_{uv} & \omega_v\end{pmatrix}\begin{pmatrix}\sigma_u^2 & \omega_u \\ \omega_u & \sigma_\varepsilon^2\end{pmatrix}^{-1}\begin{pmatrix}\sigma_{uv} & \omega_v\end{pmatrix}' \equiv \lambda^2. \tag{2.34}$$

Furthermore, u conditional on ε_3 has a $N(\omega_u/\sigma_\varepsilon^2\varepsilon_3, \sigma_u^2 - \omega_u^2/\sigma_\varepsilon^2)$ distribution.

By making a change of variables $u = (\sigma_u^2 - \omega_u^2/\sigma_\varepsilon^2)^{1/2}\eta + (\omega_u/\sigma_\varepsilon^2)\varepsilon_3$, we obtain that (after some tedious algebra)

$$E[u|v > -W\gamma, \varepsilon_3] = \int u\frac{Pr(v > -W\gamma|u,\varepsilon_3)f_u(u|\varepsilon_3)f_\varepsilon(\varepsilon_3)}{Pr(v > -W\gamma|\varepsilon_3)f_\varepsilon(\varepsilon_3)}du \tag{2.35}$$

$$= \int((\sigma_u^2 - \omega_u^2/\sigma_\varepsilon^2)^{1/2}\eta + \frac{\omega_u}{\sigma_\varepsilon^2}\varepsilon_3)\frac{\Phi(\frac{W\gamma + a_\eta\eta + a_\varepsilon\varepsilon_3}{\lambda})\phi(\eta)}{Pr(v > -W\gamma|\varepsilon_3)}d\eta \tag{2.36}$$

$$= \frac{\int((\sigma_u^2 - \omega_u^2/\sigma_\varepsilon^2)^{1/2}\eta + \frac{\omega_u}{\sigma_\varepsilon^2}\varepsilon_3)\Phi(a + b\eta)\phi(\eta)d\eta}{Pr(v > -W\gamma|\varepsilon_3)} \tag{2.37}$$

$$= \frac{(\sigma_u^2 - \omega_u^2/\sigma_\varepsilon^2)^{1/2}\frac{b}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right) + \frac{\omega_u}{\sigma_\varepsilon^2}\varepsilon_3\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)}{Pr(v > -W\gamma|\varepsilon_3)} \tag{2.38}$$

$$= \underbrace{\frac{(\sigma_u^2 - \omega_u^2/\sigma_\varepsilon^2)^{1/2}\frac{b}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)}{Pr(v > -W\gamma|\varepsilon_3)}}_{\text{Selection effect}} + \underbrace{\frac{\frac{\omega_u}{\sigma_\varepsilon^2}\varepsilon_3\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)}{Pr(v > -W\gamma|\varepsilon_3)}}_{\text{Endogeneity effect}}, \tag{2.39}$$

where

$$a_\eta \equiv (\sigma_u^2\sigma_\varepsilon^2 - \omega_u^2)^{-1}(\sigma_{uv}\sigma_\varepsilon^2 - \omega_u\omega_v)(\sigma_u^2 - \omega_u^2/\sigma_\varepsilon^2)^{1/2} \tag{2.40}$$

$$a_\varepsilon \equiv (\sigma_u^2\sigma_\varepsilon^2 - \omega_u^2)^{-1}\{(\sigma_{uv}\sigma_\varepsilon^2 - \omega_u\omega_v)\frac{\omega_u}{\sigma_\varepsilon^2} + (\omega_v\sigma_u^2 - \sigma_{uv}\omega_u)\} \tag{2.41}$$

$$a \equiv \frac{W\gamma + a_\varepsilon\varepsilon}{\lambda} \tag{2.42}$$

$$b \equiv \frac{a_\eta}{\lambda}, \tag{2.43}$$

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

As mentioned, equation (2.39) is a generalization of the inverse Mills ratio term

known from the ordinary Heckman selection model. It consists of two parts, a selection effect and an effect due to the endogeneity of covariates. If endogeneity is absent (i.e., $\omega_u = \omega_v = 0$), it can be shown that the endogeneity effect is zero and the selection effect reduces to the inverse Mills ratio term from the ordinary Heckman selection model with exogenous covariates.

The selection effect measures the expected excess outcome of the selected sample holding all explanatory variables constant (including the endogenous covariate). For example, if the outcome is the log wage, then we would talk about positive selection if an individual from the observed sample has a higher log wage on average than an individual from the unobserved sample, *if both individuals have the same values of covariates*. Note that the sign of the selection effect depends entirely on b. b, in turn, depends on $\tilde{\rho}$. We can show this algebraically for the case of one endogenous covariate, but we conjecture that this relationship also holds for the general case. To see the relationship between b and $\tilde{\rho}$, note that

$$\begin{pmatrix} \tilde{\sigma}^2 & \tilde{\rho}\tilde{\sigma} \\ \tilde{\rho}\tilde{\sigma} & 1 \end{pmatrix} = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix} - \begin{pmatrix} \omega_u \\ \omega_v \end{pmatrix} \frac{1}{\sigma_\varepsilon^2} \begin{pmatrix} \omega_u & \omega_v \end{pmatrix}, \tag{2.44}$$

hence $\tilde{\rho}\tilde{\sigma} = \sigma_{uv} - \frac{\omega_u \omega_v}{\sigma_\varepsilon^2}$ and $\tilde{\sigma}^2 = \sigma_u^2 - \frac{\omega_u^2}{\sigma_\varepsilon^2}$. Now observe that the sign of b depends on a_η, which can be rewritten as

$$a_\eta = (\sigma_u^2 - \omega_u^2/\sigma_\varepsilon^2)^{1/2} \tilde{\sigma}^{-2} \tilde{\rho}\tilde{\sigma} = (\sigma_u^2 - \omega_u^2/\sigma_\varepsilon^2)^{1/2} \frac{\tilde{\rho}}{\tilde{\sigma}}. \tag{2.45}$$

Therefore, the sign of the selection effect in a Heckman selection model with endogenous covariates is determined by the correlation coefficient $\tilde{\rho}$. Note that $\tilde{\rho}$ is a conditional correlation coefficient because it has been derived from a conditional distribution (see Section 2.3.1). This distinguishes this correlation coefficient from its counterpart in the ordinary selection model, which is an unconditional correlation coefficient. However, as

we have shown in this subsection, both parameters have the same interpretation as a measure to analyze the composition of the observed sample with respect to unobservables.

2.3.3 Testing for Exogeneity and the Absence of Sample Selection Bias

We now present a simple test which indicates whether endogeneity is indeed a problem in a particular application. The absence of endogeneity means that the matrix Ω is equal to the null matrix. But this implies that Ψ is equal to the null matrix as well. Hence, we can test for the absence of endogeneity by performing a simple test of joint significance of the parameters associated with the additional "covariates" ε_1, ε_2 and ε_3. If we cannot reject the joint hypothesis that these parameters are equal to zero, then this indicates that endogeneity is indeed absent and estimates from an ordinary Heckman selection model would be consistent.

A test of the null hypothesis that Ψ equals the null matrix is a standard task in maximum likelihood estimation. For instance, one can apply a Wald test. Of course, a likelihood ratio test or a Lagrange Multiplier test are also possible. A Wald test based on the FIML estimates can be done using the test statistic

$$W_\Psi = \text{vec}(\hat{\Psi})'(\text{Asy.Cov}[\text{vec}(\hat{\Psi})])^{-1}\text{vec}(\hat{\Psi}) \sim \chi^2(2J), \tag{2.46}$$

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

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, -\mathcal{H}^{-1}), \tag{2.47}$$

where $\mathcal{H} = n^{-1}\mathrm{E}\left(\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\right)$. In practice, $\mathcal{H}$ would be replaced by a consistent estimator.

In a similar manner, it is possible to test for the absence of sample selection bias. In that case, the null hypothesis would be $\tilde{\rho} = 0$.

2.4 Empirical Analysis

In this section, we apply our LIML estimator to the Mulligan and Rubinstein (2008) example. As described in the introduction, Mulligan and Rubinstein (2008) used the Heckman selection model to study the composition (or quality) of the female full time full year (FTFY) workforce. However, Mulligan and Rubinstein (2008) assumed exogeneity of their covariates, which is questionable for a variable like education. Based on their estimation results, they concluded that the female workforce was negatively selected in the late 1970s (1975-1979) and positively selected in the late 1990s (1995-1999). That means, women in the late 1970s who belonged to the FTFY workforce had a lower (potential) expected wage than women (with the same covariates) who did not belong to the FTFY workforce, whereas in the 1990s women who belonged to the FTFY workforce had a higher expected wage than women who did not belong to the FTFY workforce.

We set up a similar model as Mulligan and Rubinstein (2008) did. Our goal is to study if the conclusions regarding the composition of the female FTFY workforce persist if one applies a Heckman selection model which controls for the (potential) endogeneity of education. We use 1980 U.S. Census data[6] , which can be seen as a substitute for the late 1970s in the Mulligan and Rubinstein study. We expect that applying an ordinary Heckman selection model leads to the same conclusions as in Mulligan and Rubinstein (2008), i.e., that the female FTFY workforce was negatively selected. Our goal is thus to check whether this conclusion persists if we apply our proposed Heckman selection

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

model controlling for the (potential) endogeneity of education.

The reason for the choice of the data set (which is different from the data set used by Mulligan and Rubinstein, 2008) is that we need plausible instrumental variables for education. These should be randomly assigned, affect the wage only through the effect on education ("exclusion") and should have a statistically significant relation to education ("first stage"). Instrumental variables satisfying these conditions are hard to find. To resolve this issue, we exploit the idea underlying the Angrist and Krueger (1991) paper. Angrist and Krueger (1991) used the quarter of birth (and various interactions) as an instrumental variable for education. The idea is that children in the United States attend school in the year they turn six, where December the 31st is the cutoff date. Thus, a child who turns six late in the year attends school at the age of five, whereas a child who turns six early in the year attends school at the age of six. Since the legal high school drop out age in the United States is 16 years of age, Angrist and Krueger (1991) argue that children born late in the year attend school at an earlier age and, thus, stay longer in school.

We made sample restrictions that are close to Mulligan and Rubinstein (2008). Our sample consists of white women between 25 and 54 years of age not living in group quarters. We consider selection into the full time full year (FTFY) workforce, i.e., workers who worked at least 35 hours per week and 50 weeks in the last year. Only for these women we calculated an hourly wage given by their annual income divided by (52 times the usual hours of work). The remaining women add to the population who were not selected into the FTFY workforce, which thus comprises women who did not work at all and women who did work but not full time full year. We excluded observations with wages below the 2.5th percentile and above the 97.5th percentile of the wage distribution. Choosing different percentiles did not change the results much. Observations for which incomes have been imputed by a "hot deck" procedure were eliminated as well. We excluded unemployed people as we cannot say whether these (potentially) belong

to the FTFY workforce or to the remaining population. Furthermore, we eliminated self-employed workers.

We consider a simplified version of the Mulligan and Rubinstein (2008) model specification. The most important difference is that Mulligan and Rubinstein (2008) used dummies for different levels of educational attainment, whereas we use only one continuous education variable. The main equation of our model has the natural logarithm of the hourly wage as its dependent variable, so that the estimated coefficients of the explanatory variables can be interpreted as the percentage change in the wage rate due to a one-unit increase in an explanatory variable (in case of continuous variables). Covariates in the main equation include years of education (*educ*), age (*age*), age squared (*age2*), dummies for the census region (*northeast*, *midwest*, *south*; *west* is the baseline) and dummies for the marital status (*widowed*, *divorced*, *separated*, *never married*; *married* is the baseline). The selection equation includes the same variables as the main equation and the number of children younger than five years of age (*nchlt5*). The latter variable is technically not needed for identification, but we used it because Mulligan and Rubinstein (2008) did the same.[7]

Since education is potentially endogenous[8], we followed the LIML approach outlined in Section 2.3.1 and estimated in the first stage a reduced form equation for education. Explanatory variables are the exogenous variables from the main equation and our quarter of birth dummies (where the first quarter is the baseline) as instrumental variables. We did not use the various interactions of quarter of birth with state of residence and time periods as Angrist and Krueger (1991) did, since this might lead us towards a weak instruments problem (Bound et al., 1995).

Table 2.1 provides descriptive statistics of the variables. Wages are measured in 1999

[7]Mulligan and Rubinstein (2008) argued that they did not want to identify the main equation parameters by functional form assumptions alone, hence they selected an instrumental variable for the selection equation.

[8]It might be argued that marital status is endogenous as well. We thus replicated our analysis without dummies for the marital status. However, our results did not change much qualitatively.

U.S. dollars. We have 1,590,646 observations in total, from whom 465,897 (=29.3%) belong to the FTFY workforce.

We begin our empirical analysis with the (maximum likelihood) estimation of an ordinary Heckman selection model, which assumes exogeneity of covariates. This is the approach (implicitly) taken by Mulligan and Rubinstein (2008). Table 2.2 contains the results, which are qualitatively similar to Mulligan and Rubinstein (2008). In particular, the correlation coefficient has an estimated value of -.0265, which points in the same direction as the Mulligan and Rubinstein (2008) estimate of -0.077. Therefore, if we assumed exogenous covariates, we would make the same conclusion as in Mulligan and Rubinstein (2008), i.e., that the female FTFY workforce was negatively selected in the late 1970s/1980.

Now we turn to the estimation of the Heckman selection model controlling for the potential endogeneity of education. First, we estimated the reduced form equation for education, whose results can be found in Table 2.3. From Table 2.3 we can see that the quarter of birth dummies have a significant impact on the education variable, thus fulfilling one basic requirement to be valid instrumental variables. The F statistic testing the joint hypothesis that the coefficients of the instrumental variables are all zero takes a value of 54.44, which is larger than the often-cited value of 10 recommended by Staiger and Stock (1997). This indicates that we are not facing a weak instruments problem. The coefficients on the quarter of birth dummies possess the expected signs, since the coefficient values imply that the educational attainment of people born late in the year is higher.

From these first stage estimates, we obtained the estimated residuals (*eps*) and inserted them as additional covariates into a maximum likelihood estimation procedure of the Heckman selection model (i.e., the LIML approach described in Section 2.3.1). Then we used these estimates as starting values for FIML estimation, which immediately gives the correct standard errors. The FIML estimates are shown in Table 2.4.

The first result to highlight is that the returns to education increase substantially when endogeneity is taken into account (from 6% to 17%). Moreover, the coefficient of education in the selection equation also increases substantially, while the other coefficients remain relatively stable. Hence, when endogeneity of education is taken into account, the impact of education on the (log) wage as well as on the probability of belonging to the FTFY workforce is much larger than suggested by the ordinary Heckman selection model.

An application of a Wald and a likelihood ratio test, as outlined in Section 2.3.3, revealed that the hypothesis of no endogeneity (i.e., the coefficients of *eps* in main and selection equation are jointly equal to zero) was clearly rejected (the p-values of both tests were almost zero).

It is difficult to explain the huge increase in the returns to education when accounting for the endogeneity of education. To exclude the conjecture that this finding is due to a potentially misspecified model (in the sense that the parametric multivariate normality assumption is not valid), we re-estimated the main equation by ordinary least squares (OLS) and two stage least squares (2SLS), using the same instrumental variables as before. We thus ignore the issue of sample selectivity and use well-known estimators which do not rely on strong parametric assumptions. Results are given in Table 2.5. We see that the OLS estimates of the main equation are close to those of the ordinary Heckman selection model. However, the 2SLS estimates also confirm a tremendous increase in the female returns to education. Therefore, we conjecture that the high returns to education found in the Heckman model with endogenous education are not due to the model specification. In his survey article, Card (1999) found that studies using instrumental variable techniques to estimate the returns to education (mostly for men) typically came to the result that instrumental variable estimates were larger than the OLS estimates, sometimes substantially larger. In light of this, our results are not implausible. Card (1999) provided some general economic explanations why the

instrumental variable estimates may be larger than the OLS estimates. In this chapter, however, we do not attempt to provide economic explanations for our results.

What is more important in our analysis is the value of the correlation coefficient $\tilde{\rho}$. As has been shown above, in the case of endogenous covariates $\tilde{\rho}$ can be interpreted analogously to the correlation coefficient ρ from the ordinary selection model, i.e., as a measure of the "quality" of the observed sample. As we can see from Table 2.4, the estimated value of $\tilde{\rho}$ is (almost) identical to the value of the correlation coefficient from the ordinary selection model (see Table 2.2). Hence, despite the fact that the coefficients of education in main and selection equation have so much changed when controlling for the endogeneity of education, the parameter $\tilde{\rho}$ is not very different from its counterpart ρ. Thus, we can confirm the conclusion made by Mulligan and Rubinstein (2008) that the female FTFY workforce was negatively selected in the late 1970s/1980, even after controlling for the endogeneity of education.

2.5 Conclusions

In this chapter, we have developed a Heckman selection model with endogenous covariates. We provided a rather general model which encompasses various scenarios of endogeneity, including endogeneity only in the main equation, only in the selection equation or in both. Although our estimator relies on distributional assumptions which may not be satisfied in particular applications, the estimator nevertheless serves as a starting point for a deeper (semiparametric) analysis. A virtue of our estimator is that it is relatively simple to compute. In fact, any econometrics software which is capable of performing maximum likelihood estimation of the Heckman sample selection model can be used.

The most important advantage of our model is that it provides an easy-to-interpret measure to analyze the composition of the observed sample with respect to unobserv-

ables. As an example, we considered the composition of the female FTFY workforce, as analyzed by Mulligan and Rubinstein (2008). We applied our model to this example and found that the conclusion made by Mulligan and Rubinstein, i.e., that the female FTFY workforce was negatively selected in the late 1970s, is robust to accounting for the endogeneity of education in the Heckman selection model. It would be interesting to see if this is also true for the second time period which was considered by Mulligan and Rubinstein (2008), i.e., the late 1990s. We did not do this because we did not have suitable data.

Our estimation results based on the 1980 U.S. Census data also indicate that accounting for the endogeneity of education leads to a tremendous increase in the estimated female returns to education. Future research may provide economic explanations of this result. The lesson from this finding is that it is important to control for endogeneity of covariates in sample selection models. Although selection models are frequently used in applied econometrics, most authors assume exogeneity of covariates. We hope that this chapter makes the issue of endogenous covariates in selection models more prominent and that it fosters the application of selection models which also control for the endogeneity of covariates.

2.6 Appendix A

In this appendix, we show how the asymptotic covariance matrix of the LIML estimator must be corrected in order to account for the estimation of the regressors ε_1, ε_2 and ε_3. First, let $\alpha \equiv (\text{vec}(\Delta)', \text{vec}(\Lambda)', \text{vec}(\Upsilon)')'$ and $\tilde{l}(\tilde{\theta}, \hat{\alpha}) = \sum_{i=1}^{n} l_i(\tilde{\theta}, \hat{\alpha})$ be the limited information log-likelihood function. Provided there exists an interior solution, we can write the first order condition from maximizing this likelihood function as

$$\sum_{i=1}^{n} \frac{\partial l_i(\hat{\tilde{\theta}}, \hat{\alpha})}{\partial \tilde{\theta}} = 0. \tag{2.48}$$

An asymptotic first order expansion about $\hat{\tilde{\theta}} = \tilde{\theta}$ gives after rearranging and pre-multiplication with $\sqrt{n}$

$$\sqrt{n}(\hat{\tilde{\theta}} - \tilde{\theta}) = \left(-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\tilde{\theta}, \alpha)}{\partial \tilde{\theta}^2} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial l_i(\tilde{\theta}, \hat{\alpha})}{\partial \tilde{\theta}} + o_p(1). \tag{2.49}$$

Expanding the gradient about $\hat{\alpha} = \alpha$ yields

$$\begin{aligned} \sqrt{n}(\hat{\tilde{\theta}} - \tilde{\theta}) = & \left(-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\tilde{\theta}, \alpha)}{\partial \tilde{\theta}^2} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial l_i(\tilde{\theta}, \alpha)}{\partial \tilde{\theta}} \\ & + \left(-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\tilde{\theta}, \alpha)}{\partial \tilde{\theta}^2} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\tilde{\theta}, \alpha)}{\partial \tilde{\theta} \partial \hat{\alpha}} \right) \sqrt{n}(\hat{\alpha} - \alpha) + o_p(1). \end{aligned} \tag{2.50}$$

If

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\tilde{\theta}, \alpha)}{\partial \tilde{\theta}^2} \xrightarrow{p} H \text{ pos. def.} \tag{2.51}$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial l_i(\tilde{\theta}, \alpha)}{\partial \tilde{\theta}} \xrightarrow{d} \mathcal{N}(0, M) \tag{2.52}$$

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\tilde{\theta}, \alpha)}{\partial \tilde{\theta} \partial \hat{\alpha}} \xrightarrow{p} J \tag{2.53}$$

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, V), \tag{2.54}$$

then

$$\sqrt{n}(\hat{\tilde{\theta}} - \tilde{\theta}) \xrightarrow{d} \mathcal{N}(0, C), \tag{2.55}$$

where $C = H^{-1}(M + JVJ')H^{-1}$. This follows because the covariance between $\frac{\partial l_i(\tilde{\theta},\alpha)}{\partial \tilde{\theta}}$ and $(\hat{\alpha} - \alpha)$ is zero, as shown by Smith and Blundell (1986).

Note that implementation of the LIML estimator using an econometrics software yields an asymptotic covariance of $H^{-1}MH^{-1}$, as the software does not know that some regressors have been estimated. Hence, one must add to this expression a correction term of $H^{-1}(JVJ')H^{-1}$ in order to obtain the correct asymptotic covariance.

2.7 Appendix B

In this appendix, we derive formulas for the (asymptotic) variances of the estimates of the structural variance-covariance parameters (based on the FIML estimates). We assume, however, that FIML estimation does not yield estimates of $\tilde{\rho}$, $\tilde{\sigma}$ and Σ, but rather of $\text{atanh}(\tilde{\rho})$, $\ln(\tilde{\sigma})$ and S such that $\Sigma = SS'$. The reason for not directly estimating these parameters is that we have to make sure that $\hat{\tilde{\rho}} \in (-1, 1)$, $\hat{\tilde{\sigma}} > 0$ and $\hat{\Sigma}$ be positive definite. Our reparameterization guarantees that these conditions are fulfilled.

(i) The Asymptotic Distribution of $\hat{\Omega}$

ML estimation yields estimates of[9]

$$s \equiv \begin{bmatrix} s_{11} \\ s_{21} \\ s_{22} \end{bmatrix} = \text{vech}(S) \quad \text{and} \quad \text{vec}(\Psi') = \begin{bmatrix} \psi'_{11} \\ \psi'_{12} \\ \psi'_{21} \\ \psi'_{22} \end{bmatrix}_{(2J \times 1)} . \tag{2.56}$$

Let

$$q \equiv (s', \text{vec}(\Psi')')'. \tag{2.57}$$

Since $\Omega = \Sigma\Psi'$ is a function of q, the asymptotic distribution of $\text{vec}(\hat{\Omega})$ can be obtained by means of the Delta method. If

$$\sqrt{n}(\hat{q} - q) \xrightarrow{d} N(0, M), \tag{2.58}$$

then

$$\sqrt{n}(\text{vec}(\hat{\Omega}) - \text{vec}(\Omega)) \xrightarrow{d} N\left(0, CMC'\right), \tag{2.59}$$

where

$$C = \frac{\partial \text{vec}(\Omega)}{\partial q'} \tag{2.60}$$

$$= \frac{\partial \text{vec}(SS'\Psi')}{\partial q'} \tag{2.61}$$

$$= (\Psi \otimes I_J)\frac{\partial \text{vec}(SS')}{\partial q'} + (I_2 \otimes SS')\frac{\partial \text{vec}(\Psi')}{\partial q'} \tag{2.62}$$

$$= (\Psi \otimes I_J)\begin{bmatrix} \frac{\partial \text{vec}(SS')}{\partial s'} & 0 \end{bmatrix} + (I_2 \otimes \Sigma)\begin{bmatrix} 0 & \frac{\partial \text{vec}(\Psi')}{\partial \text{vec}(\Psi')'} \end{bmatrix} \tag{2.63}$$

[9]Note: The 2-by-2 case has been used here for the sake of illustration. The following analysis does not hinge on this case.

$$= \left[(\Psi \otimes I_J) \tfrac{\partial \text{vec}(SS')}{\partial s'} \quad (I_2 \otimes \Sigma) \right]. \tag{2.64}$$

Furthermore,

$$\frac{\partial \text{vec}(SS')}{\partial s'} = \left\{ (S \otimes I_J) \frac{\partial \text{vec}(S)}{\partial s'} + (I_J \otimes S) \frac{\partial \text{vec}(S')}{\partial s'} \right\} \tag{2.65}$$

$$= \left\{ (S \otimes I_J) \frac{\partial \text{vec}(S)}{\partial s'} + (I_J \otimes S) K_J \frac{\partial \text{vec}(S)}{\partial s'} \right\} \tag{2.66}$$

$$= \{ (S \otimes I_J) L_J' + (I_J \otimes S) K_J L_J' \} \tag{2.67}$$

$$= \{ (S \otimes I_J) + (I_J \otimes S) K_J \} L_J' \tag{2.68}$$

$$= (I_{J^2} + K_J)(S \otimes I_J) L_J', \tag{2.69}$$

with

$$L_J = \sum_{i \geq j} u_{ij} \text{vec}(E_{ij})' \tag{2.70}$$

$$K_J = \sum_{i=1}^{J} \sum_{j=1}^{J} E_{ij} \otimes E_{ij}', \tag{2.71}$$

where u_{ij} denotes a unit vector of size $\frac{1}{2}J(J+1)$ whose $[(j-1)J + i - \frac{1}{2}j(j-1)]$-th element is unity $(1 \leq j \leq i \leq J)$, and E_{ij} is a $(J \times J)$ matrix with one at the (i,j)-th position and zeros elsewhere. Note that L_J and K_J do only depend on J.

Therefore,

$$C = \left[(\Psi \otimes I_J)(I_{J^2} + K_J)(S \otimes I_J) L_J' \quad (I_2 \otimes \Sigma) \right]. \tag{2.72}$$

(ii) The Asymptotic Distribution of $\hat{\Pi} = \begin{pmatrix} \hat{\sigma}_u^2 & \rho \hat{\sigma}_u \hat{\sigma}_v \\ \rho \hat{\sigma}_u \hat{\sigma}_v & \hat{\sigma}_v^2 \end{pmatrix}$

ML estimation yields estimates of

$$s \equiv \text{vech}(S), \quad \text{vec}(\Psi') = \begin{bmatrix} \psi'_{11} \\ \psi'_{12} \\ \psi'_{21} \\ \psi'_{22} \end{bmatrix}, \quad [\ln \tilde{\sigma}], \quad [\text{atanh}(\tilde{\rho})]. \tag{2.73}$$

Let

$$q \equiv (s', \text{vec}(\Psi')', [\ln \tilde{\sigma}], [\text{atanh}(\tilde{\rho})])'. \tag{2.74}$$

Since

$$\Pi = \Gamma + \Psi \Sigma \Psi' \tag{2.75}$$

$$= \begin{bmatrix} (\exp\{[\ln \tilde{\sigma}]\})^2 & \tanh([\text{atanh}(\tilde{\rho})]) \exp\{[\ln \tilde{\sigma}]\} \\ \tanh([\text{atanh}(\tilde{\rho})]) \exp\{[\ln \tilde{\sigma}]\} & 1 \end{bmatrix} + \Psi \Sigma \Psi' \tag{2.76}$$

is a function of q, the asymptotic distribution of $\text{vech}(\hat{\Pi})$ can be obtained by means of the delta method.

If

$$\sqrt{n}(\hat{q} - q) \xrightarrow{d} N(0, M), \tag{2.77}$$

then

$$\sqrt{n}(\text{vech}(\hat{\Pi}) - \text{vech}(\Pi)) \xrightarrow{d} N\left(0, CMC'\right), \tag{2.78}$$

where

$$C = \frac{\partial \text{vech}(\Pi)}{\partial q'} \tag{2.79}$$

$$= L_{2J} \frac{\partial \text{vec}(\Pi)}{\partial q'} \tag{2.80}$$

$$= L_{2J} \left\{ \frac{\partial \text{vec}(\Gamma)}{\partial q'} + \frac{\partial \text{vec}(\Psi \Sigma \Psi')}{\partial q'} \right\}. \tag{2.81}$$

Both components of the RHS have to be investigated in detail.

First,

$$\frac{\partial \text{vec}(\Gamma)}{\partial [\ln \tilde{\sigma}], [\text{atanh}(\tilde{\rho})]} = \begin{bmatrix} 2(\exp\{[\ln \tilde{\sigma}]\})^2 & 0 \\ \tanh([\text{atanh}(\tilde{\rho})]) \exp\{[\ln \tilde{\sigma}]\} & (1 - \tanh^2([\text{atanh}(\tilde{\rho})])) \exp\{[\ln \tilde{\sigma}]\} \\ \tanh([\text{atanh}(\tilde{\rho})]) \exp\{[\ln \tilde{\sigma}]\} & (1 - \tanh^2([\text{atanh}(\tilde{\rho})])) \exp\{[\ln \tilde{\sigma}]\} \\ 0 & 0 \end{bmatrix} \tag{2.82}$$

$$\equiv A \tag{2.83}$$

$$\Rightarrow \frac{\partial \text{vec}(\Gamma)}{\partial q'} = \begin{bmatrix} 0 & A \end{bmatrix}. \tag{2.84}$$

Next,

$$\frac{\partial \text{vec}(\Psi \Sigma \Psi')}{\partial (s', \text{vec}(\Psi')')} = (\Psi \Sigma \otimes I_2) \frac{\partial \text{vec}(\Psi)}{\partial (s', \text{vec}(\Psi')')} + (I_2 \otimes \Psi) \frac{\partial \text{vec}(\Sigma \Psi')}{\partial (s', \text{vec}(\Psi')')} \tag{2.85}$$

$$= (\Psi \Sigma \otimes I_2) K_{2J} \frac{\partial \text{vec}(\Psi')}{\partial (s', \text{vec}(\Psi')')} + (I_2 \otimes \Psi) \frac{\partial \text{vec}(\Sigma \Psi')}{\partial (s', \text{vec}(\Psi')')} \tag{2.86}$$

$$= (\Psi \Sigma \otimes I_2) K_{2J} \begin{bmatrix} 0 & I_{2J} \end{bmatrix} + (I_2 \otimes \Psi) \frac{\partial \text{vec}(\Sigma \Psi')}{\partial (s', \text{vec}(\Psi')')} \tag{2.87}$$

$$\Rightarrow \frac{\partial \text{vec}(\Psi \Sigma \Psi')}{\partial q'} = \begin{bmatrix} \frac{\partial \text{vec}(\Psi \Sigma \Psi')}{\partial (s', \text{vec}(\Psi')')} & 0 \end{bmatrix} \tag{2.88}$$

Hence,

$$C = \frac{\partial \text{vec}(\Psi\Sigma\Psi')}{\partial q'} = \left[\frac{\partial \text{vec}(\Psi\Sigma\Psi')}{\partial (s', \text{vec}(\Psi')')} \quad A \right]. \tag{2.89}$$

(iii) The Asymptotic Distribution of $\hat{\rho}$

Given an estimate of

$$\Pi = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}, \tag{2.90}$$

let

$$g \equiv \rho\sigma_u\sigma_v \Rightarrow \rho = \frac{g}{\sigma_u\sigma_v} = \frac{g}{\sqrt{\sigma_u^2\sigma_v^2}} = g\left(\sigma_u^2\sigma_v^2\right)^{-\frac{1}{2}} \tag{2.91}$$

and

$$q \equiv \left(\sigma_u^2, g, \sigma_v^2\right)' = \text{vech}(\Pi). \tag{2.92}$$

Since ρ is a function of q, the asymptotic distribution of $\hat{\rho}$ can be obtained by means of the delta method.

If

$$\sqrt{n}(\hat{q} - q) \xrightarrow{d} \mathcal{N}(0, G) \tag{2.93}$$

then

$$\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{d} \mathcal{N}(0, FGF') \tag{2.94}$$

with

$$F = \frac{\partial \rho}{\partial q'} = \left[-\frac{1}{2} g \left(\sigma_u^2 \sigma_v^2\right)^{-\frac{3}{2}} \sigma_v^2, \left(\sigma_u^2 \sigma_v^2\right)^{-\frac{1}{2}}, -\frac{1}{2} g \left(\sigma_u^2 \sigma_v^2\right)^{-\frac{3}{2}} \sigma_u^2 \right]. \tag{2.95}$$

(iv) The Asymptotic Distribution of $\hat{\Sigma} = \hat{S}\hat{S}'$

ML estimation yields estimates of[10]

$$s \equiv \begin{bmatrix} s_{11} \\ s_{21} \\ s_{22} \end{bmatrix} = \text{vech}(S). \tag{2.96}$$

The asymptotic distribution is given by

$$\sqrt{n}(\hat{s} - s) \xrightarrow{d} N(0, M). \tag{2.97}$$

Since $\text{vech}(\Sigma) = \text{vech}(SS') = c(s)$ is a function of s, the asymptotic distribution of $\text{vech}(\Sigma)$ can be obtained by using the delta method, which gives

$$\sqrt{n}(\text{vech}(\hat{\Sigma}) - \text{vech}(\Sigma)) \xrightarrow{d} N(0, C(s)MC(s)'), \tag{2.98}$$

[10]Note: The 2-by-2 case has been used here for the sake of illustration. The following analysis does not hinge on this case.

where

$$C(s) = \frac{\partial c(s)}{\partial s'} \tag{2.99}$$
$$= \frac{\partial \text{vech}(SS')}{\partial s'} \tag{2.100}$$
$$= L_J \frac{\partial \text{vec}(SS')}{\partial s'} \tag{2.101}$$
$$= L_J \left\{ (S \otimes I_J) \frac{\partial \text{vec}(S)}{\partial s'} + (I_J \otimes S) \frac{\partial \text{vec}(S')}{\partial s'} \right\} \tag{2.102}$$
$$= L_J \left\{ (S \otimes I_J) \frac{\partial \text{vec}(S)}{\partial s'} + (I_J \otimes S) K_J \frac{\partial \text{vec}(S)}{\partial s'} \right\} \tag{2.103}$$
$$= L_J \left\{ (S \otimes I_J) L_J' + (I_J \otimes S) K_J L_J' \right\} \tag{2.104}$$
$$= L_J \left\{ (S \otimes I_J) + (I_J \otimes S) K_J \right\} L_J' \tag{2.105}$$
$$= L_J (I_{J^2} + K_J)(S \otimes I_J) L_J' \tag{2.106}$$

and

$$L_J = \sum_{i \geq j} u_{ij} \text{vec}(E_{ij})' \tag{2.107}$$
$$K_J = \sum_{i=1}^{J} \sum_{j=1}^{J} E_{ij} \otimes E_{ij}', \tag{2.108}$$

where u_{ij} denotes a unit vector of size $\frac{1}{2}J(J+1)$ whose $[(j-1)J + i - \frac{1}{2}j(j-1)]$-th element is unity ($1 \leq j \leq i \leq J$), and E_{ij} is a $(J \times J)$ matrix with one at the (i,j)-th position and zeros elsewhere. Note that L_J and K_J do only depend on J.

2.8 Appendix C

In this appendix, we use Monte Carlo simulations in order to study the finite sample properties of our FIML estimator and in order to gauge the bias which occurs if one does not account for endogeneity. The results of these simulations are presented in

Table 2.6.

The first column of Table 2.6 contains the specification. We distinguish between four benchmark cases. In the first case, endogeneity is only present in the primary equation. In particular, it is assumed that

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

and

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

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

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

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

and

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

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

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

and

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

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

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

and

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

Throughout, X_{1i}, Z_{1i}, Z_{2i} and Z_{3i}, $i = 1, \ldots, n$, are scalars which have been simulated from a standard normal distribution. For each of the four cases, these random numbers have been drawn once and kept fixed during simulation. In total, each simula-

tion encompasses 1000 repetitions in which parameter estimates have been computed. Table 2.6 presents the mean of these estimates over the repetitions, along with the corresponding standard deviations.

In order to gauge the finite-sample performance of the estimator outlined in Section 2.3.1, Table 2.6 contains simulation results for different sample sizes. For each sample size, Table 2.6 displays the results for the FIML estimator presented in Section 2.3.1 ("IV") and contrasts these results with those obtained when using the ordinary estimator for the sample selection model which does not account for endogeneity ("non-IV"). To save space, only the estimates for the parameters of the primary equation and selection equation are presented.

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

Note that the estimates for the selection equation are subjected to a normalization rule. This is the reason why the performance of the IV estimator seems to be not "perfect". However, as it is well known, in binary choice models only coefficient ratios are identified. Put differently, one should not consider the raw coefficients given in Table 2.6 but rather coefficient ratios. For example, in specification (iii) for $n = 1000$ we can calculate that the mean of the second coefficient divided by the first gives 0.7018, whereas the mean of the third coefficient divided by the first gives 0.2991. Thus, we see that also the parameters of the selection equation are well estimated by the FIML

procedure.

On the contrary, in most cases the non-IV estimator yields severely biased estimates of the parameters of the primary equation among all specifications. For instance, for a sample size of $n = 1000$ the bias ranges from 13 to 248.1 percent. However, the estimates of the selection equation are sometimes relatively close to their true values (specifications (i) and (iii)). This notwithstanding, note especially that the estimates of the parameters of the main equation are severely biased even if endogeneity is only present in the selection equation (specification (ii)). This result, which is due to the nonlinearity of the underlying model, has not gained much attention in the literature yet.

Overall, the results show that the FIML-IV estimator from Section 2.3.1 outperforms the ordinary estimator for the sample selection model, especially with respect to the parameters in the primary equation and in case of large sample sizes. Moreover, the results indicate that the bias in the parameter estimates may be substantial if one does not account for endogeneity.

2.9 Appendix D

In this appendix, we present an application of our FIML estimator to the labor supply data set introduced by Thomas Mroz (1987). Our goal is to compare our results with those of Wooldridge (2010), who also applied his estimator to this data set.

The Mroz data set is quite popular and is often used to illustrate the performance of estimators which account for sample selectivity. The data set consists of 753 married women of whom 428 are working. We not only have information about relevant labor market characteristics of women (such as the wage, educational attainment and experience) but also on private characteristics such as the number of children, the "non-wife income" and the educational attainment of the parents and the husband. The former

variables help identify the selection equation, while the latter variables may serve as instrumental variables for education. These variables are assumed to satisfy an exclusion restriction in the sense that they *directly* affect only the probability of labor market participation and educational attainment, respectively, but not the wage rate.

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

We estimated the following model: The main equation contains the natural logarithm of the hourly wage as its dependent variable; explanatory variables are experience, experience squared and education. The selection equation includes experience, experience squared, non-wife income, age, number of children aged until 6 years of age in the household, number of children aged 6 years or older in the household and education. Since education is treated as endogenous, instrumental variables are needed for estimation. Following Wooldridge (2010), we chose mother's education, father's education and husband's education as instrumental variables for education.[11] Means and standard deviations of these variables are presented in Table 2.7.

[11]For the appropriateness of these instrumental variables, cf. the discussion in Card (1999), pp. 1822-26.

Estimation results are given in Table 2.8. In Table 2.8, estimation results for the ordinary sample selection model ("non-IV") and the sample selection model with endogeneity ("IV") are provided. The first part of this table contains the parameter estimates for the variables of the main equation, as well as estimates of the "reduced form" selection parameter $\tilde{\rho}$ and the endogeneity parameter ψ_{11}. This last parameter indicates whether endogeneity of education is relevant in the primary equation. The second part presents the parameter estimates for the selection equation. Additionally included is the endogeneity parameter ψ_{21}, which indicates whether endogeneity of education is relevant in the selection equation. Finally, the third part includes the parameter estimates of the exogenous variables and instrumental variables with respect to education. In analogy with the instrumental variables terminology, this part has been labeled "first stage".

The results show significance of education in the primary and the selection equation. Moreover, the instrumental variables for education employed in the "first stage" are highly significant. The remaining variables possess the expected signs. However, the estimates of $\tilde{\rho}$, ψ_{11} and ψ_{21} are not significantly different from zero, indicating that there is neither a selection bias nor an endogeneity bias present.[12] These results are in line with those reported by Wooldridge (2010) who draws similar conclusions. However, given that there seems to be neither a sample selection bias nor an endogeneity bias present, this result is not surprising.

[12]In addition, joint significance of ψ_{11} and ψ_{21} is rejected as well (p-value of 0.1907).

2.10 Tables

Table 2.1: Summary statistics

Variable	Mean	Standard. Dev.	Min	Max
lwage	2.4436	0.3828	1.1978	3.2766
educ	12.4301	2.5496	0	17
age	37.8317	8.8395	25	54
northeast	0.2276	0.4193	0	1
midwest	0.2682	0.4430	0	1
west	0.1911	0.3932	0	1
south	0.3131	0.4637	0	1
married	0.7759	0.4170	0	1
widowed	0.0236	0.1516	0	1
divorced	0.0950	0.2932	0	1
separated	0.0238	0.1525	0	1
never_married	0.0818	0.2740	0	1
nchlt5	0.2580	0.5597	0	6
qtr1	0.2487	0.4323	0	1
qtr2	0.2397	0.4269	0	1
qtr3	0.2624	0.4400	0	1
qtr4	0.2492	0.4325	0	1
No. obs.	1,590,646			
No. obs. FTFY	465,897			

Source: 1980 U.S. Census data; own calculations.

Table 2.2: Estimates from Heckman model without endogenous covariates

Variable	Coef.	Std.err.
Main equation		
educ	0.0594	0.0003
age	0.0227	0.0006
age2	-0.0002	0.0000
northeast	-0.0161	0.0017
midwest	-0.0358	0.0016
south	-0.1014	0.0016
widowed	-0.0082	0.0034
divorced	0.0278	0.0023
separated	-0.0191	0.0033
never_married	0.0577	0.0023
constant	1.2373	0.0127
Selection equation		
educ	0.0336	0.0004
age	-0.0658	0.0013
age2	0.0007	0.0000
northeast	0.0215	0.0034
midwest	0.0653	0.0033
south	0.1520	0.0032
widowed	0.2420	0.0070
divorced	0.6962	0.0035
separated	0.3040	0.0069
never_married	0.5783	0.0040
nchlt5	-0.6033	0.0028
constant	0.3563	0.0258
correlation parameter (ρ)	-0.0265	0.0100

Source: 1980 U.S. Census data; own calculations.

Table 2.3: Reduced form estimates for education

Variable	Coef.	Std.err.
qtr2	-0.0022	0.0056
qtr3	0.0410	0.0055
qtr4	0.0555	0.0056
age	-0.0006	0.0023
age2	-0.0006	0.0000
northeast	-0.1118	0.0061
midwest	-0.1630	0.0059
south	-0.4336	0.0057
widowed	-0.6586	0.0132
divorced	-0.0509	0.0068
separated	-0.8154	0.0130
never_married	0.7369	0.0074
constant	13.5899	0.0440
F statistic	54.44	

Source: 1980 U.S. Census data; own calculations.

Table 2.4: Estimates from Heckman model with endogenous covariates

Variable	Coef.	Std.err.
Main equation		
educ	0.1751	0.0240
age	0.0228	0.0007
age2	-0.0002	0.0000
northeast	-0.0031	0.0032
midwest	-0.0169	0.0043
south	-0.0512	0.0106
widowed	0.0680	0.0163
divorced	0.0337	0.0028
separated	0.0753	0.0200
never_married	-0.0275	0.0178
constant	-0.3388	0.3275
eps	-0.1158	0.0240
Selection equation		
educ	0.3753	0.0512
age	-0.0655	0.0015
age2	0.0009	0.0000
northeast	0.0598	0.0070
midwest	0.1210	0.0092
south	0.3001	0.0225
widowed	0.4670	0.0347
divorced	0.7135	0.0049
separated	0.5827	0.0426
never_married	0.3266	0.0380
nchlt5	-0.6032	0.0028
constant	-4.2969	0.6982
eps	-0.3418	0.0512
Reduced form equation for educ		
qtr2	0.0038	0.0047
qtr3	0.0448	0.0049
qtr4	0.0576	0.0051
age	-0.0006	0.0023
age2	-0.0006	0.0000
northeast	-0.1118	0.0061
midwest	-0.1630	0.0059
south	-0.4336	0.0057
widowed	-0.6586	0.0132
divorced	-0.0509	0.0068
separated	-0.8154	0.0130
never_married	0.7370	0.0074
constant	13.5866	0.0440
correlation parameter ($\tilde{\rho}$)	-0.0265	0.0100

Source: 1980 U.S. Census data; own calculations.

Table 2.5: OLS and 2SLS estimates of the main equation

	OLS		2SLS	
Variable	Coef.	Std.err.	Coef.	Std.err.
educ	0.0596	0.0002	0.2292	0.0508
age	0.0224	0.0006	0.0428	0.0062
age2	-0.0002	0.0000	-0.0004	0.0000
northeast	-0.0158	0.0016	0.0176	0.0103
midwest	-0.0353	0.0016	0.0149	0.0152
south	-0.1004	0.0015	-0.0370	0.0191
widowed	-0.0064	0.0033	0.0373	0.0140
divorced	0.0326	0.0014	0.0284	0.0024
separated	-0.0168	0.0032	0.0396	0.0175
never_married	0.0622	0.0016	-0.0813	0.0431
constant	1.2271	0.0121	-1.5016	0.8180

Source: 1980 U.S. Census data; own calculations.

Table 2.6: Monte Carlo results

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

Source: Own calculations.

Table 2.7: Descriptive statistics for the Mroz data

Variable	Mean	Std.dev.
log wage	4.1777	3.3103
exper	10.6308	8.0691
educ	12.2869	2.2802
nwifeinc	20.1290	11.6348
age	42.5379	8.0726
kidslt6	0.2377	0.5240
kidsge6	1.3533	1.3199
motheduc	9.2510	3.3675
fatheduc	8.8088	3.5723
huseduc	12.4914	3.0208
Sample size	753	
No. of obs. with wage>0	428	

Source: Mroz (1987) data; own calculations.

Table 2.8: Estimation of a wage equation for married women based on the Mroz data

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

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

Source: Mroz (1987) data; own calculations.

Chapter 3

Sieve Maximum Likelihood Estimation of a Copula-Based Sample Selection Model

This chapter is a major revision of the discussion paper No. 503, Department of Economics and Business Administration, Leibniz University Hannover (Schwiebert, 2012b). I thank Blaise Melly, Melanie Schienle, Jeffrey M. Wooldridge, Michael Lechner, participants at the 16th IZA summer school and seminar participants from SEW St. Gallen for providing valuable comments.

3.1 Introduction

The sample selection model has become the standard econometric tool when dealing with sample selectivity. The model typically consists of a main equation (of interest) and a selection equation, where the latter determines the probability of being in the observed sample. If sample selectivity is present, ordinary least squares estimation of the main equation is likely to produce inconsistent estimates because the observed sam-

ple is a nonrandom sample from the overall population. Heckman (1979) showed that the sample selection problem can be interpreted as an omitted variable bias problem. He demonstrated that ordinary least squares estimation of the main equation including a selectivity correction term (known as the inverse Mills ratio) leads to consistent estimates of the parameters of interest. Besides estimating the model by ordinary (or weighted) least squares techniques, it is also possible to estimate the model by maximum likelihood.

Gallant and Nychka (1987) have proposed a semi-nonparametric maximum likelihood estimator for estimating the sample selection model. The virtue of their approach is that it is not necessary to assume a parametric (joint) distribution for the error terms of the underlying econometric model. Consequently, consistent estimates of the parameters of interest can be obtained under weak conditions. This is an important advantage over the model proposed by Heckman (1979) who assumed a bivariate normal distribution for the error terms of main and selection equation.

In this chapter, we propose a sieve maximum likelihood estimator for the sample selection model. We make the crucial assumption that the joint distribution of the error terms of main and selection equation can be characterized by a specific copula, but we estimate the marginal distributions semiparametrically by the method of sieves along with the structural parameters of interest. Our estimation concept is thus sieve maximum likelihood estimation (Chen, 2007).

Our modeling and estimation approach has several advantages over the Gallant and Nychka (1987) procedure. First, our approach allows to incorporate prior information on the distribution of error terms into the estimation process. For example, the joint distribution of error terms may be characterized by fat tails, hence a Student t copula may be an appropriate modeling choice (Heckman, 2003). Furthermore, the selection equation may reasonably be estimated by probit or logit, hence the marginal distribution of the selection equation's error term is normal or logistic, respectively.

Since a copula couples two marginal distributions into a joint distribution, such prior information on the joint or marginal distributions can be easily incorporated into our econometric model. This is not possible in the Gallant and Nychka (1987) approach, who estimate the entire joint density function of error terms semi-nonparametrically by a series expansion.

Second, our method is less computationally demanding than the Gallant and Nychka (1987) procedure. In Gallant and Nychka (1987) a two-dimensional density function is approximated semi-nonparametrically by a series expansion (where the number of series term grows with the sample size). The coefficients of the series expansion are then estimated along with the parameters of interest. However, the approximation of a two-dimensional density function requires a considerable number of series terms, which leads to a computationally demanding estimation process. Our approach, on the other hand, requires only the approximation of the one-dimensional marginal distributions, which is far easier than approximating a (bivariate) joint distribution.[1]

Third, Gallant and Nychka (1987) have proved the consistency of their estimator, but no (asymptotic) distribution results have been provided. Yet, such distribution results are necessary for hypotheses testing and obtaining confidence intervals. Of course, one could obtain estimates under the assumption that the number of series terms is fixed rather than increasing with the sample size; in that case, distribution results would follow from standard (parametric) maximum likelihood theory. However, this procedure is in general not justified due to the semiparametric nature of the estimation problem. Concerning our proposed method, conditions under which a sieve maximum likelihood estimator is consistent and asymptotically normally distributed have been provided by Chen et al. (2006) and Chen (2007). As will be shown below, under suitable assumption these conditions are fulfilled in case of our estimator, hence we are able to provide distribution results.

[1] The same argument has been used by Chen et al. (2006).

Fourth, our approach offers an easy way to test for the validity of parametric assumptions. Incorporating correct parametric prior information into an econometric model is desirable since this typically leads to efficiency gains. However, prior information may not be correct, hence it is important to test for the validity of such assumptions. Our copula framework provides an easy way to do so because one can separately test for the validity of the assumed copula and for the validity of the assumed marginal distributions. Details are given in Section 3.4.

Besides Gallant and Nychka (1987), several other authors have developed semi-non-parametric estimators for the sample selection model which do not rely on strong parametric assumptions. Examples include Powell (1987), Ahn and Powell (1993), Das et al. (2003) and Newey (2009). These authors propose least-squares based estimation procedures to consistently estimate the structural parameters of the main equation. These estimation procedures are typically two-step. In a first step, the selection equation is estimated by some semi-nonparametric technique. As in case of the model with normally distributed error terms, one augments the main equation with a selectivity correction term (a generalization of the inverse Mills ratio term). Then one either gets rid of the selectivity correction term by differencing out (Powell, 1987; Ahn and Powell, 1993), or approximates the term by, e.g., a series expansion (Das et al., 2003; Newey, 2009). In a second step, estimation of the main equation is carried out by some variant of ordinary or weighted least squares.

Our and the Gallant and Nychka (1987) approach differ from these least-squares based techniques in three important ways. First, our and the Gallant and Nychka (1987) approach are one-step. This facilitates the computation of standard errors and confidence intervals (in case of our estimator) because one does not have to adjust for the uncertainty associated with the first-step estimation. Second, no exclusion restriction is needed, i.e., the selection equation need not contain a variable (with a nonzero coefficient) which may not appear in the main equation. Such an exclusion is

generally needed in least-squares based approaches, yet it is sometimes difficult to justify economically why a variable should appear in the selection equation but not in the main equation. Third, our and the Gallant and Nychka (1987) approach are not based on least-squares but maximum likelihood estimation. This *requires* a specification of the joint distribution of error terms of main and selection equation. Considered conversely, a *virtue* of our and the Gallant and Nychka (1987) approach is that they also provide information on the joint distribution of the error terms.

Information on the joint distribution of error terms is useful for a couple of reasons. First, sample selectivity is a problem only if the error terms are dependent. Distributional information helps to identify these dependencies, and thus reveals how the sample selection mechanism works. Second, from the joint distribution one can derive the marginal distributions of error terms. For instance, if the main equation is a wage equation, an object of interest might be if wage densities are fat-tailed (Heckman and Sedlacek, 1990). Third, the joint distribution is interesting because treatment parameters depend on the tail behavior of error terms (Heckman et al., 2003).

A drawback of our proposed approach might be that it is necessary to specify a parametric copula for the joint distribution of error terms in advance. However, if one has prior information (e.g., from economic theory or empirical regularities) on the features of the joint distribution (such as fat tails), then the copula framework provides a very flexible environment to include such prior information into the econometric model. Chen et al. (2006) also estimate a copula model with unknown marginal distributions and note that "this class of semiparametric multivariate distributions is able to jointly model any type of dependence with any types of marginal behaviors and has proven useful in diverse fields" (Chen et al., 2006, p. 1228). Hence, our approach exhibits the same flexibility as the semiparametric approach of Gallant and Nychka (1987), but may be preferred due to the reasons given above.

The remainder of the chapter is organized as follows. In Section 3.2 we provide the

model and our proposed estimation strategy. In Section 3.3 we derive the asymptotic properties of our proposed estimator. Section 3.4 contains remarks and extensions concerning different aspects of estimation, testing, and model specification. Finally, Section 3.5 concludes the chapter.

3.2 Model Setup and Estimation

We consider an ordinary sample selection model given by

$$y_i^* = x_i'\beta + \varepsilon_i \tag{3.1}$$

$$d_i^* = w_i'\gamma + u_i \tag{3.2}$$

$$d_i = 1(d_i^* > 0) \tag{3.3}$$

$$y_i = \begin{cases} y_i^* \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases}, \tag{3.4}$$

where $i = 1, \ldots, n$ indexes individuals. The first equation is the main equation, where y^* is a latent outcome variable, x is a vector of (exogenous) explanatory variables with corresponding parameter vector β and ε denotes the error term. The second equation is the selection equation, where d^* is the latent dependent variable, w is a vector of (exogenous) explanatory variables with corresponding parameter vector γ and u denotes the error term. The last two equations comprise the selection mechanism. The latent variable y^* can only be observed if $d^* > 0$, or, equivalently, if the selection indicator d is equal to one.

We make the following assumptions:

Assumption 1: $\{(x_i, w_i, \varepsilon_i, u_i)\}_{i=1}^n$ *are i.i.d. from some underlying distribution.*

Assumption 2: *The joint distribution function of* ε *and* u *is given by* $H_{\varepsilon,u}(a,b) = C(F_\varepsilon(a), F_u(b); \tau)$*, where* $C : [0,1]^2 \to [0,1]$ *is a known copula with dependence param-*

eter τ, and F_ε and F_u denote the marginal distribution functions of ε and u. Furthermore, the marginal density functions f_ε and f_u are absolutely continuous with respect to Lebesgue measure.

ASSUMPTION 3: *(x, w) and (ε, u) are independent.*

ASSUMPTION 4: *(i) x and w do not contain a constant term. (ii) The first element of γ is equal to one in absolute value.*

Assumptions 1 and 2 imply that our model can be estimated by maximum likelihood. Assumptions 3 and 4 are basic conditions for identification. Assumption 4 (i) is needed because constant terms are not identified, as they cannot be separated from the constants in the unknown functions f_ε and f_u. Assumption 4 (ii) imposes a scale normalization on the parameters of the selection equation, since these are only identified up to scale.

Note that we do not require that an exclusion restriction holds. That is, there need not be a variable only appearing in the selection equation (with a nonzero coefficient) but not in the main equation. Gallant and Nychka (1987, p. 383) derived an identification condition on their model which explicitly does not require the existence of an exclusion restriction. Since the Gallant and Nychka (1987) and our model are to some extent similar, the same applies to our model. For a more elaborate discussion of the identification conditions, we refer the reader to the Gallant and Nychka (1987) paper. In practice, our approach may work better than Gallant and Nychka (1987) since we put an initial parametric restriction on the joint distribution of error terms (i.e., the copula). Since without an exclusion restriction parameters will be identified by functional forms, putting some restriction on the joint distribution of error terms may lead to better results in practice (i.e., the likelihood function may not have too many local maxima).

The joint probability density function (p.d.f.) of ε and u is given by

$$h_{\varepsilon,u}(a,b) = c(F_\varepsilon(a), F_u(b); \tau) f_\varepsilon(a) f_u(b), \tag{3.5}$$

where $c(\cdot,\cdot;\tau)$ denotes the p.d.f. associated with $C(\cdot,\cdot;\tau)$. The log-likelihood function then follows as

$$\begin{aligned}
&\ln L(\beta, \gamma, \tau, f_\varepsilon, f_u; Z) \\
&= \sum_{i=1}^{n} \left\{ (1-d_i) \ln \int_{-\infty}^{\infty} \int_{-\infty}^{-w_i'\gamma} h_{\varepsilon,u}(\varepsilon, u) du d\varepsilon + d_i \ln \int_{-w_i'\gamma}^{\infty} h_{\varepsilon,u}(y_i - x_i'\beta, u) du \right\} \\
&= \sum_{i=1}^{n} \left\{ (1-d_i) \ln F_u(-w_i'\gamma) + d_i \ln \left(f_\varepsilon(y_i - x_i'\beta) - \frac{\partial H_{\varepsilon,u}(\varepsilon, -w_i'\gamma)}{\partial \varepsilon}\bigg|_{\varepsilon = y_i - x_i'\beta} \right) \right\},
\end{aligned} \tag{3.6}$$

where $Z = \{z_i\}_{i=1}^n$ and $z_i = (y_i, x_i, d_i, w_i)$ denotes the observed data. Note that the log-likelihood function is not only maximized over the structural parameters β, γ and τ but over the unknown functions f_ε and f_u as well. Furthermore, note that it suffices that the log-likelihood function depends on f_ε and f_u and not additionally on F_ε and F_u, because we have that $F_\varepsilon(x) = \int_{-\infty}^{x} f_\varepsilon(v) dv$ and $F_u(x) = \int_{-\infty}^{x} f_u(v) dv$. Our interest focuses on estimation of the structural parameters $\theta = (\beta', \gamma', \tau)'$, while the unknown functions f_ε and f_u are considered as nuisance parameters. Remember that the first element of γ is equal to one in absolute value due to identification, hence it need not be estimated. This restriction will be suppressed in the following in order to ease the notation.

Since f_ε and f_u are of infinite dimension, estimation requires that we approximate these functions. We follow Chen et al. (2006) and Chen (2007) and approximate these densities by the method of sieves. That means, we approximate an unknown function (the densities) by a (e.g., linear) combination of known basis functions (such as polynomials or splines) and unknown sieve coefficients. The unknown sieve coefficients are

then estimated along with the structural parameters β, γ and τ. Since we approximate density functions, we have to restrict the approximating functions to satisfy two fundamental properties of densities, i.e., that they are not negative and that they integrate to one. The former property can be satisfied if we approximate not the density function by the method of sieves but the square root of the density function instead. This is the approach taken in Chen et al. (2006), who propose the following sieve space:

$$\mathcal{F}_{n,\eta} = \left\{ f_{n,\eta}(x) = \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta} A_{k,\eta}(x) \right]^2, \int f_{n,\eta,}(x)dx = 1 \right\}, K_{n,\eta} \to \infty, \frac{K_{n,\eta}}{n} \to 0, \tag{3.7}$$

where $f_{n,\eta}$ is an approximation to $f_\eta, \eta \in \{\varepsilon, u\}$, based on $K_{n,\eta}$ sieve coefficients, $\{A_{k,\eta}(\cdot) : k \geq 0\}$ denote known basis functions and $\{a_{k,\eta}(\cdot) : k \geq 0\}$ are unknown sieve coefficients which must be estimated. Note that $K_{n,\eta}$ depends on the sample size n but grows at a slower rate. For the basis functions Chen et al. (2006) suggest to use Hermite polynomials or splines; for details, see Chen et al. (2006). To ensure that the approximation of the density function integrates to one in applications, one can set

$$f_{n,\eta}(x) = \frac{\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta} A_{k,\eta}(x) \right]^2}{\int \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta} A_{k,\eta}(v) \right]^2 dv}. \tag{3.8}$$

Let

$$h_{n,\varepsilon,u}(a,b) = c(F_{n,\varepsilon}(a), F_{n,u}(b); \tau) f_{n,\varepsilon}(a) f_{n,u}(b), \tag{3.9}$$

where $F_{n,\eta}(x) = \int_{-\infty}^{x} f_{n,\eta}(v)dv, \eta \in \{\varepsilon, u\}$. Then, our proposed sieve maximum likelihood estimator $\hat{\theta}_n$ of θ is obtained by maximizing

$$\ln L(\beta, \gamma, \tau, a_{0,\varepsilon}, \ldots, a_{K_{n,\varepsilon},\varepsilon}, a_{0,u}, \ldots, a_{K_{n,u},u}; Z)$$

$$= \sum_{i=1}^{n} \left\{ (1-d_i) \ln \int_{-\infty}^{\infty} \int_{-\infty}^{-w_i'\gamma} h_{n,\varepsilon,u}(\varepsilon, u) du d\varepsilon + d_i \ln \int_{-w_i'\gamma}^{\infty} h_{n,\varepsilon,u}(y_i - x_i'\beta, u) du \right\}$$
$$= \sum_{i=1}^{n} \left\{ (1-d_i) \ln F_{n,u}(-w_i'\gamma) + d_i \ln \left(f_{n,\varepsilon}(y_i - x_i'\beta) - \frac{\partial H_{n,\varepsilon,u}(\varepsilon, -w_i'\gamma)}{\partial \varepsilon} \bigg|_{\varepsilon = y_i - x_i'\beta} \right) \right\} \tag{3.10}$$

over θ and the unknown sieve coefficients $(a_{0,\varepsilon}, \ldots, a_{K_{n,\varepsilon},\varepsilon}, a_{0,u}, \ldots, a_{K_{n,u},u})$.

As an example, we consider the well-known Gaussian copula. In that case, the joint cumulative distribution function of ε and u is given by

$$H_{\varepsilon,u}(a,b) = \Phi_2(\Phi^{-1}(F_\varepsilon(a)), \Phi^{-1}(F_u(b)); \tau), \tag{3.11}$$

where $\Phi_2(\cdot,\cdot,\tau)$ is the c.d.f. of the bivariate standard normal distribution with correlation coefficient τ, i.e.,

$$\Phi_2(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} \frac{1}{2\pi\sqrt{1-\tau^2}} \exp\left(-\frac{1}{2(1-\tau^2)} (x^2 + y^2 - 2\tau xy) \right) dy dx, \tag{3.12}$$

and $\Phi^{-1}(\cdot)$ is the inverse of the c.d.f. of the univariate standard normal distribution. This implies that the joint p.d.f. of ε and u is given by

$$h_{\varepsilon,u}(a,b) = \left| \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix} \right|^{-1/2} \exp\left(-\frac{1}{2} \begin{pmatrix} \Phi^{-1}(F_\varepsilon(a)) \\ \Phi^{-1}(F_u(b) \end{pmatrix}' \left(\begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}^{-1} - I_2 \right) \begin{pmatrix} \Phi^{-1}(F_\varepsilon(a)) \\ \Phi^{-1}(F_u(b) \end{pmatrix} \right)$$
$$\times f_\varepsilon(a) f_u(b), \tag{3.13}$$

where I_2 is the 2-by-2 identity matrix. Lee (1983) was the first who applied the Gaussian copula to sample selection models. He showed that the log-likelihood function is given by

$$\ln L = \sum_{i=1}^{n} \{ (1-d_i) \ln(1 - F_u(w_i'\gamma))$$

$$+d_i \ln f_\varepsilon(y_i - x_i'\beta) + d_i \ln \Phi\left(\frac{\Phi^{-1}(F_u(w_i'\gamma)) + \tau\Phi^{-1}(F_\varepsilon(y_i - x_i'\beta))}{\sqrt{1-\tau^2}}\right)\Bigg\}. \quad (3.14)$$

Besides the Gaussian copula, there exist many other copulas which can be used to model dependencies among the error terms. Popular examples are copulas from the Farlie-Gumbel-Morgenstern (FGM) family and the Archimedean class of copulas. The Archimedean class encompasses some well-known copulas such as the Clayton copula, the Frank copula and the Gumbel copula. We refer the reader to Smith (2003) for a description of these copulas. Smith (2003) also provides the likelihood functions for sample selection models based on these copulas.

3.3 Asymptotic Properties

In this section, we derive consistency and asymptotic normality of our proposed sieve maximum likelihood estimator using the results in Chen et al. (2006) and Chen (2007). First, let $\mathcal{A} = \Theta \times \mathcal{F}_\varepsilon \times \mathcal{F}_u$ denote the parameter space. As in the last section, the sieve MLE is defined as

$$\hat{\alpha}_n = \arg\max_{\alpha \in \mathcal{A}_n} \ln L(\alpha; Z) = \sum_{i=1}^{n} \ln l(\alpha, z_i), \quad (3.15)$$

where $\alpha = (\theta', f_\varepsilon, f_u)'$ and $\hat{\alpha}_n = (\hat{\theta}_n', \hat{f}_{n,\varepsilon}, \hat{f}_{n,u})' \in \Theta \times \mathcal{F}_{n,\varepsilon} \times \mathcal{F}_{n,u} = \mathcal{A}_n$. The true value of the parameter vector is denoted as $\alpha_0 = (\theta_0', f_{0,\varepsilon}, f_{0,u})' \in \mathcal{A}$.

Our first goal is to derive consistency of our proposed estimator. Suppose that $d(\cdot,\cdot)$ is a (pseudo) metric on $\mathcal{A}$. We make the following assumptions (in addition to Assumptions 1-4), which are taken from Conditions 3.1', 3.2', 3.3', 3.4 and 3.5 in Chen (2007):

Assumption 5: *(i)* $E[\ln L(\alpha, Z)]$ *is continuous at* $\alpha_0 \in \mathcal{A}$, $E[\ln L(\alpha_0, Z)] > -\infty$
(ii) for all $\epsilon > 0$, $E[\ln L(\alpha_0, Z)] > \sup_{\{\alpha \in \mathcal{A}:\, d(\alpha,\alpha_0) \geq \epsilon\}} E[\ln L(\alpha, Z)]$.

ASSUMPTION 6: *$\mathcal{A}_k \subseteq \mathcal{A}_{k+1} \subseteq \mathcal{A}$ for all $k \geq 1$; and for any $\alpha \in \mathcal{A}$ there exists a sequence $\pi_k \alpha_0 \in \mathcal{A}_k$ such that $d(\alpha_0, \pi_k \alpha_0) \to 0$ as $k \to \infty$.*

ASSUMPTION 7: *For each $k \geq 1$,*

(i) $\ln L(\alpha, Z)$ is a measurable function of the data Z for all $\alpha \in \mathcal{A}_k$; and

(ii) for any data Z, $\ln L(\alpha, Z)$ is upper semicontinuous on $\mathcal{A}_k$ under the metric $d(\cdot, \cdot)$.

ASSUMPTION 8: *The sieve spaces, $\mathcal{A}_k$, are compact under $d(\cdot, \cdot)$.*

ASSUMPTION 9: *For all $k \geq 1$, $\text{plim}_{n\to\infty} \sup_{\alpha \in \mathcal{A}_k} |\ln L(\alpha) - E[\ln L(\alpha)]| = 0$.*

Assumption 5 is an identification condition which implies that the true parameter vector α_0 uniquely maximizes the expected value of the log-likelihood function. Assumptions 6 and 8 contain assumptions on the sieve spaces. In particular, it is assumed that asymptotically the difference between an (unknown) function and its sieve approximation tends to zero. Assumption 7 is a continuity condition, while Assumption 9 assumes uniform convergence of the sample log-likelihood to its population counterpart over the sieves.

We establish the following consistency theorem:

THEOREM 1: *Suppose that Assumptions 1-9 hold. Then $d(\hat{\alpha}_n, \alpha_0) = o_p(1)$.*

PROOF: *See Chen (2007), pp. 5589-5591.* ■

In order to establish asymptotic normality, we show that Conditions 4.1-4.5 in Chen (2007) are fulfilled. We derive asymptotic normality only for the structural parameters of interest contained in θ. Our exposition closely follows Chen (2007, ch. 4).

Let

$$\frac{\partial l(\alpha_0, z)}{\partial \alpha'}[\alpha - \alpha_0] = \lim_{\omega \to 0} \frac{l(\alpha_0 + \omega[\alpha - \alpha_0], z) - l(\alpha_0, z)}{\omega} \tag{3.16}$$

be the directional derivative of $l(\alpha_0, z)$ in the direction $[\alpha - \alpha_0]$ and suppose that it is well defined for almost all z. Let V be the completion of the space spanned by $\mathcal{A} - \alpha_0$

. As in Chen et al. (2006), we define the Fisher norm on this space as

$$||v||^2 = E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha'}[v]\right]^2, \tag{3.17}$$

which induces the Fisher inner product

$$\langle v, \tilde{v}\rangle = E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha'}[v]\frac{\partial l(\alpha_0, z)}{\partial \alpha'}[\tilde{v}]\right]. \tag{3.18}$$

Let $f(\theta_0) = \lambda'\theta_0$, where λ is an arbitrary unit vector with the same dimension as θ. It follows from the Riesz representation theorem that there exists $v^* \in V$ such that, for any $\alpha - \alpha_0 \in V$,

$$\lambda'(\theta - \theta_0) = \langle \alpha - \alpha_0, v^*\rangle \tag{3.19}$$

with $||v^*|| < \infty$.

To proceed further, it is necessary to compute the Riesz representer v^*. Define

$$D_{w_j}(z) = \frac{\partial l(\alpha_0, z)}{\partial \theta_j} - \frac{\partial l(\alpha_0, z)}{\partial f'}[w_j], \quad j = 1, \ldots, \dim(\theta), \tag{3.20}$$

where $f = (f_\varepsilon, f_u)'$. Then, the Riesz representer $v^* = ((v_\theta^*)', (v_f^*)')'$ is given by

$$v_f^* = -(w^*)'v_\theta^* \tag{3.21}$$

$$v_\theta^* = (E[D_{w^*}(z)D_{w^*}(z)'])^{-1}\lambda \tag{3.22}$$

$$w_j^* = \arg\inf_{w_j} E[(D_{w_j}(z))^2], \tag{3.23}$$

where $w = (w_1, \ldots, w_{\dim(\theta)})'$ and $D_w(z) = (D_{w_1}(z), \ldots, D_{w_{\dim(\theta)}(z)})'$.

We make the following assumptions:

ASSUMPTION 10: $\theta_0 \in \text{int}(\Theta), \Theta$ *a compact subset of* $\mathbb{R}^{\dim(\theta)}$.

ASSUMPTION 11: *The log-likelihood function* $\ln L(\alpha, z)$ *is twice continuously pathwise differentiable with respect to* $\alpha \in \mathcal{A}$ *and* $||\alpha - \alpha_0|| = o(1)$*, and the derivatives are uniformly bounded with respect to* $\alpha \in \mathcal{A}$ *and* z*.*

ASSUMPTION 12: $E[D_{w^*}(z)D_{w^*}(z)']$ *is positive definite.*

ASSUMPTION 13: *There is* $\pi_n v^* \in \mathcal{A}_n$ *such that* $||\pi_n v^* - v^*|| = O(K^{-\psi}) = o(n^{-1/2})$.

Assumptions 10-12 are standard. Assumption 13 places a smoothness condition on the Riesz representer v^*, which is similar to Assumption 3 of Newey (1997). We establish the following theorem:

THEOREM 2: *Suppose that Assumptions 1-4 and 10-13 hold, and that* $||\hat{\alpha}_n - \alpha_0|| = o_p(1)$*. Then* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_*(\theta_0)^{-1})$*, where* $I_*(\theta_0) = E[D_{w^*}(z)D_{w^*}(z)']$.

PROOF: *See the Appendix.* ■

Furthermore, $\hat{\theta}_n$ is semiparametrically efficient (see Chen, 2006).

In order to calculate standard errors or confidence intervals for $\hat{\theta}_n$, one needs an estimate of the asymptotic covariance matrix $I_*(\theta_0)^{-1}$. Such an estimate can be obtained in the following way (see Chen, 2007, p. 5616). Let

$$\hat{w}_j^* = \arg \min_{w_j \in (\mathcal{F}_{n,\varepsilon} \times \mathcal{F}_{n,u})} \frac{1}{n} \sum_{i=1}^{n} [(\hat{D}_{w_j}(z_i))^2], \tag{3.24}$$

with

$$\hat{D}_{w_j}(z) = \frac{\partial l(\hat{\alpha}_0, z)}{\partial \theta_j} - \frac{\partial l(\hat{\alpha}_0, z)}{\partial f'}[w_j], \quad j = 1, \ldots, \dim(\theta). \tag{3.25}$$

Define $\hat{D}_w(z) = (\hat{D}_{w_1}, \ldots, \hat{D}_{w_{\dim(\theta)}})'$. Then an estimate $\hat{I}_*(\hat{\theta}_n)^{-1}$ of $I_*(\theta_0)^{-1}$ is given by

$$\hat{I}_*(\hat{\theta}_n)^{-1} = \left(\frac{1}{n} \sum_{i=1}^{n} [\hat{D}_{\hat{w}^*}(z_i) \hat{D}_{\hat{w}^*}(z_i)'] \right)^{-1}. \tag{3.26}$$

The following theorem establishes the consistency of $\hat{I}_*(\hat{\theta}_n)^{-1}$ for $I_*(\theta_0)^{-1}$:

THEOREM 3: *Suppose that Assumptions 1-4 and 10-13 hold, and that* $||\hat{\alpha}_n - \alpha_0|| = o_p(1)$.

Then $\hat{I}_*(\hat{\theta}_n)^{-1} = I_*(\theta_0)^{-1} + o_p(1)$.

PROOF: *Follows from Lemma 2 in Ackerberg et al. (2012), pp. 493-494.* ■

In fact, Ackerberg et al. (2012) showed that there is a simpler way to obtain $\hat{I}_*(\hat{\theta}_n)^{-1}$. Suppose there is a fictitious practitioner who uses the same sieve space (7) to approximate the unknown densities f_ε and f_u, but she treats the number of sieve terms, $K_{n,\eta}, \eta \in \{\varepsilon, u\}$, as fixed rather than as increasing with the sample size. Consequently, she has a finite dimensional parameter vector and maximum likelihood estimation and inference can be carried out as usual. However, since the number of sieve terms is considered fixed, the maximum likelihood estimator will not be consistent for the parameters of interest (i.e., θ) and will not have the correct limiting distribution proposed in Theorem 2.

To fix ideas, let $\tilde{\alpha} = (\theta', \kappa')'$ denote the parameter vector to be estimated by our fictitious practitioner, where $\kappa = (a_{0,\varepsilon}, \ldots, a_{K_{n,\varepsilon},\varepsilon}, a_{0,u}, \ldots, a_{K_{n,u},u})'$ contains the sieve coefficients. Note that the practitioner faces the same problem as in our sieve estimation approach, but the difference is that the practitioner treats $K_\eta, \eta \in \{\varepsilon, u\}$ as fixed. The information matrix of the practitioner is given by

$$\tilde{I}(\tilde{\alpha}_0) = \begin{bmatrix} E\left[\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \theta}\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \theta'}\right] & E\left[\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \theta}\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \kappa'}\right] \\ E\left[\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \kappa}\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \theta'}\right] & E\left[\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \kappa}\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \kappa'}\right] \end{bmatrix}, \tag{3.27}$$

which can be consistently estimated by

$$\hat{\tilde{I}}(\hat{\tilde{\alpha}}_n) = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \theta}\frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \theta'} & \frac{1}{n}\sum_{i=1}^n \frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \theta}\frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \kappa'} \\ \frac{1}{n}\sum_{i=1}^n \frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \kappa}\frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \theta'} & \frac{1}{n}\sum_{i=1}^n \frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \kappa}\frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \kappa'} \end{bmatrix}, \tag{3.28}$$

where $\hat{\theta}_n$ and $\hat{\kappa}_n$ denote the practitioner's estimates of θ_0 and κ_0. An estimate of the asymptotic covariance of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is then given by the upper left block of the inverse of $\hat{\tilde{I}}(\hat{\tilde{\alpha}}_n)$. Ackerberg et al. (2012) derived the following result: Despite the fact

that the likelihood function is misspecified (since $K_{n,\varepsilon}$ and $K_{n,u}$ are treated as fixed), the practitioner's estimate of the asymptotic covariance of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is numerically equivalent to $\hat{I}_*(\hat{\theta}_n)^{-1}$.

The practical implication of this result is simple but powerful. A researcher who wants to carry out sieve maximum likelihood estimation just has to maximize the log-likelihood function over θ and the unknown sieve coefficients, and the (asymptotically) correct standard errors for $\hat{\theta}_n$ can be easily obtained from the inverse of the information matrix, *provided that the information matrix is based on the outer product of gradients* (and not on the Hessian matrix). Hence, any statistical software package which is capable of dealing with user-supplied likelihood functions can be used for sieve maximum likelihood estimation and inference, as long as the researcher is allowed to specify how the information matrix shall be calculated.

3.4 Remarks and Extensions

3.4.1 Closed Form Likelihood Function

The log-likelihood function in (3.10) does not exhibit a closed form expression due to the presence of integral terms. Integrals arise because of the presence of the distribution functions $F_{n,\varepsilon}$ and $F_{n,u}$, which are related to $f_{n,\varepsilon}$ and $f_{n,u}$ via $F_{n,\varepsilon}(x) = \int_{-\infty}^{x} f_{n,\varepsilon(v)} dv$ and $F_{n,u}(x) = \int_{-\infty}^{x} f_{n,u}(v) dv$. Moreover, the copula function may contain integrals as well; the Student t copula would be an example where this is the case (Demarta and McNeil, 2005). Calculating the integrals within an optimization routine is of course possible, but may be computationally demanding if the sample size and/or the number of parameters increases. Put differently, it may take a quite long time until the optimization routine finds the maximum likelihood estimates. In this subsection we describe a method how the integrals in $F_{n,\varepsilon}$ and $F_{n,u}$ can be replaced by closed form expressions, which may facilitate maximum likelihood estimation in practice. The integrals

appearing through the copula function are not considered here.[2] Fortunately, many well known copulas (such as the Gaussian copula, Archimedean copulas) indeed have closed form expressions. Sample selection models based on these copulas are analyzed in Smith (2003).

Our method to obtain closed form expressions for $F_{n,\varepsilon}$ and $F_{n,u}$ essentially relies on an expansion of the unknown densities f_ε and f_u by Hermite polynomials. More specifically, we propose to approximate the unknown density functions by

$$f_{n,\eta}(x) = \frac{\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta}(x/\sigma_\eta)^k\right]^2 \phi(x/\sigma_\eta)/\sigma_\eta}{\int_{-\infty}^{\infty}\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta}(v/\sigma_\eta)^k\right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv}, \quad \eta \in \{\varepsilon, u\}, \tag{3.29}$$

where $\sigma_\eta > 0$ is a scale parameter which must be estimated, and $\phi(\cdot)$ is the standard normal probability density function.

An important advantage of using Hermite polynomials as basis functions in these expansions is that $F_{n,\varepsilon}$ and $F_{n,u}$ have closed form expressions.[3] Note that

$$F_{n,\eta}(x) = \frac{\int_{-\infty}^{x}\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta}(v/\sigma_\eta)^k\right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv}{\int_{-\infty}^{\infty}\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta}(v/\sigma_\eta)^k\right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv}, \quad \eta \in \{\varepsilon, u\}. \tag{3.30}$$

To see that $F_{n,\varepsilon}$ and $F_{n,u}$ have closed forms, consider the denominator of (3.30) first. To ease notation, we suppress η in the following formulas. The denominator can be

[2] To deal with such integrals, Maximum Simulated Likelihood techniques may be employed.

[3] Using Hermite polynomials to approximate the square root of a density has been suggested by e.g. Gallant and Nychka (1987). Of course, there may exist other basis functions which imply closed form distribution functions. However, such other basis functions will not be considered here. To be precise, $F_{n,\varepsilon}$ and $F_{n,u}$ contain the standard normal c.d.f., which could also be regarded as an integral which needs to be approximated. However, the standard normal c.d.f. is included in most statistical software as a standard function and is computed immediately. Hence, using Hermite polynomials to approximate the unknown densities may be interpreted as a transformation of a complicated integral into a term involving integrals (the standard normal c.d.f.) which can be computed very quickly.

simplified by making a change of variables $z = x/\sigma$, which yields

$$\int_{-\infty}^{\infty} \left[\sum_{k=0}^{K} a_k (x/\sigma)^k\right]^2 \phi(x/\sigma)/\sigma dx \tag{3.31}$$

$$= \int_{-\infty}^{\infty} [Z' a a' Z] \phi(z) dz \tag{3.32}$$

$$= \text{tr}\left[aa' \int_{-\infty}^{\infty} ZZ' \phi(z) dz\right], \tag{3.33}$$

where $a = (a_0, \ldots, a_{K_n})'$ and $Z = (z^0, z^1, z^2, \ldots, z^{K_n})'$. The integral term represents moments of the standard normal distribution. For example, if $K_n = 2$, we have that $Z = (1, z, z^2)'$ and

$$\int_{-\infty}^{\infty} ZZ' \phi(z) dz = \int_{-\infty}^{\infty} \begin{bmatrix} 1 & z & z^2 \\ z & z^2 & z^3 \\ z^2 & z^3 & z^4 \end{bmatrix} \phi(z) dz = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{bmatrix} \tag{3.34}$$

with

$$\int_{-\infty}^{\infty} z^k dz = 0, \quad k = 1, 3, 5, \ldots \tag{3.35}$$

$$\int_{-\infty}^{\infty} z^k dz = (k-1)(k-3) \cdot \ldots \cdot 1, \quad k = 0, 2, 4, \ldots. \tag{3.36}$$

Hence, the denominator does not involve integrals any more and thus has a closed form.

Next, consider the numerator of (3.30). We have that

$$\int_{-\infty}^{x} \left[\sum_{k=0}^{K_n} a_k (v/\sigma)^k\right]^2 \phi(x/\sigma)/\sigma dv \tag{3.37}$$

$$= \int_{-\infty}^{x/\sigma} \left[\sum_{k=0}^{K_n} a_k (z)^k\right]^2 \phi(z) dz \tag{3.38}$$

$$= \text{tr}\left[aa' \int_{-\infty}^{x/\sigma} ZZ' \phi(z) dz\right] \tag{3.39}$$

and

$$\int_{-\infty}^{x/\sigma} ZZ'\phi(z)dz = \begin{bmatrix} b'_{0:K_n} \\ b'_{1:(K_n+1)} \\ \vdots \\ b'_{K_n:(2K_n)} \end{bmatrix}_{([K_n+1]\times[K_n+1])}, \tag{3.40}$$

where $b'_{i:j} = (b_i, \ldots, b_j)$ with

$$b_0 = \Phi(x/\sigma) \tag{3.41}$$

$$b_1 = -\phi(x/\sigma) \tag{3.42}$$

$$b_k = -\phi(x/\sigma)(x/\sigma)^{k-1} + (k-1)b_{k-2}, \quad k = 2, \ldots, 2K_n, \tag{3.43}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function. Hence, by these transformations the integrals in the numerator of (3.30) vanish as well. Therefore, $F_{n,\varepsilon}$ and $F_{n,u}$ have closed form expressions.

3.4.2 Initial Values for Maximum Likelihood Estimation

The likelihood function (3.10) usually contains a lot of parameters to be estimated, since the sieve coefficients must be estimated as well. As in case of integral terms, this may be associated with further computational complexity. However, having good initial values for the maximum likelihood estimation routine may reduce this computational burden. Such initial values can be easily obtained for the parameters β and γ if consistent estimates are available. For instance, the parameters of the selection equation, γ, may be estimated by a suitable semiparametric estimator for binary choice models. The Klein and Spady (1993) semiparametric estimation procedure can be used in this case. The parameters of the main equation, β, can be estimated by the approaches proposed by Powell (1987) or Newey (2009).

3.4.3 Testing for the Validity of Parametric Assumptions

As described in the introduction, a great advantage of our estimation approach is that it is easy to test for the validity of parametric assumptions. Testing for the validity of parametric assumptions is important since incorporating (correct) parametric information into a model typically results in efficiency gains.

It is easy to test for the validity of parametric assumptions in our copula framework because one can separately test for the validity of a certain joint distribution (represented by the copula) and for the validity of certain marginal distributions. Suppose we want to test if a certain copula is valid to describe the joint distribution of error terms. We could then estimate the model by the Gallant and Nychka (1987) procedure which does not make any (parametric) assumptions on the joint distribution. Then we would estimate the model by our approach, including the assumed copula whose validity we seek to test. Since the Gallant and Nychka (1987) and our approach are based on maximum likelihood estimation, one can test whether the parametric copula assumption is justified by applying the Vuong (1989) test for nonnested models.[4]

In a similar manner, one can test for the validity of certain parametric marginal distributions. Given a valid parametric copula, we would estimate the model by our approach with unspecified marginal distributions, and then with one or both marginal distributions parametrically specified. Again the Vuong (1989) test may help decide if the parametric assumptions on the marginal distribution(s) are correct. In case of the selection equation only one may also apply the Horowitz and Härdle (1994) testing procedure to test if a certain parametric marginal distribution is valid for the selection equation's error term.[5]

[4]The validity of the Vuong test crucially depends on whether both models can be considered nonnested for a given n. As in the case of the asymptotic variance, it might be conjectured that the Vuong test is valid when treating the semiparametric estimation problem as if it were parametric. Future research may resolve this issue.

[5]In the context of sample selection models, the Horowitz and Härdle testing procedure has been applied by e.g. Martins (2001) and Genius and Strazzera (2008).

3.4.4 Binary Dependent Variable

This subsection focuses on an extension of our semiparametric copula model to the case of a binary dependent variable. Sample selection models with a binary dependent variable have been used by van de Ven and van Praag (1981), Boyes et al. (1989), Greene (1992) and Mohanty (2002). These authors, however, assumed a bivariate normal distribution for the error terms of main and selection equation, as Heckman (1979) did. Thus, the following exposition generalizes these models by allowing for distributions apart from the bivariate normal.

The model is now given by

$$y_i^* = x_i'\beta + \varepsilon_i \tag{3.44}$$

$$d_i^* = w_i'\gamma + u_i \tag{3.45}$$

$$d_i = 1(d_i^* > 0) \tag{3.46}$$

$$y_i = \begin{cases} 1(y_i^* > 0) \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases} . \tag{3.47}$$

The difference between this model and the benchmark model from Section 3.2 is that the dependent variable associated with the main equation, y_1, now assumes only the values one or zero.

Under the same assumptions as in Section 3.2 (except that a scale normalization must be put on β), the log-likelihood function for this model is given by

$$\begin{aligned} \ln L(\beta, \gamma, f_\varepsilon, f_u; Z) &= \sum_{i=1}^{n} \Bigg\{ (1-d_i) \ln \int_{-\infty}^{\infty} \int_{-\infty}^{-w_i'\gamma} h_{\varepsilon,u}(\varepsilon, u) du d\varepsilon \\ &+ d_i(1-y_i) \ln \int_{-\infty}^{-x_i'\beta} \int_{-w_i'\gamma}^{\infty} h_{\varepsilon,u}(\varepsilon, u) du d\varepsilon + d_i y_i \ln \int_{-x_i'\beta}^{\infty} \int_{-w_i'\gamma}^{\infty} h_{\varepsilon,u}(\varepsilon, u) du d\varepsilon \Bigg\} \\ &= \sum_{i=1}^{n} \{(1-d_i) \ln F_u(-w_i'\gamma) + d_i(1-y_i) \ln[F_\varepsilon(-x_i'\beta) - H_{\varepsilon,u}(-w_i'\gamma, -x_i'\beta)] \end{aligned}$$

$$+ d_i y_i \ln H_{\varepsilon,u}(-w_i'\gamma, -x_i'\beta)\} \,. \tag{3.48}$$

Estimation and inference can be carried out as described above for the benchmark model. In fact, there is no conceptual difference between the model considered in this section and the benchmark model, apart from the binary nature of the dependent variable.

3.4.5 Endogenous Covariates

In this subsection we show how our semiparametric copula model can be extended to the case of endogenous covariates. Taking the potential endogeneity of covariates into account is important since parameter estimates will be inconsistent otherwise. To provide an illustration, we consider the classical example for which sample selection models have been used. Suppose a researcher wants to estimate a wage equation for females, and that her interest centers on the female returns to education. If she fitted a wage regression to the observed sample of working females only, she would obtain inconsistent estimates due to sample selectivity. So she would instead fit a sample selection model to the observed data. But is sample selectivity the only source of endogeneity in this example? Indeed, there may be sociological or intelligence-related factors (which we will summarize by the term "ability") which affect not only the wage (main equation) and the probability of labor force participation (selection equation), but education as well. If the researcher does not take the potential endogeneity of education into account, she will obtain an inconsistent estimate of the female returns to education.

To conceptualize these ideas, we consider the following extension of the model from Section 3.2:

$$y_{1i}^* = \tilde{x}_i'\tilde{\beta} + \delta_1 y_{2i} + \tilde{\varepsilon}_i \tag{3.49}$$

$$d_i^* = \tilde{x}_i'\tilde{\gamma} + \delta_2 y_{2i} + \delta_3 \tilde{w}_i + \tilde{u}_i \tag{3.50}$$

$$d_i = 1(d_i^* > 0) \tag{3.51}$$

$$y_{1i} = \begin{cases} y_{1i}^* \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases} \tag{3.52}$$

$$y_{2i} = \tilde{x}_i'\alpha_1 + q_i'\alpha_2 + v_i, \tag{3.53}$$

where y_2 is the endogenous covariate. The fifth equation is a reduced form equation for y_2 which includes an *instrumental variable* q which is not contained in $\tilde{x}$ or $\tilde{w}$ (exclusion restriction). Furthermore, v is an error term which is assumed to be independent of $\tilde{x}$, $\tilde{w}$ and q, but correlated with $\tilde{\varepsilon}$ and $\tilde{u}$. For instance, v, $\tilde{\varepsilon}$ and $\tilde{u}$ may be affected by a common variable like ability in the aforementioned example.

To estimate this model, we can insert the reduced form equation for y_2 into the main and selection equation, which gives the following reduced form model:

$$y_{1i}^* = x_i'\beta + \varepsilon_i \tag{3.54}$$

$$d_i^* = w_i'\gamma + u_i \tag{3.55}$$

$$d_i = 1(d_i^* > 0) \tag{3.56}$$

$$y_{1i} = \begin{cases} y_{1i}^* \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases} \tag{3.57}$$

$$y_{2i} = x_i'\alpha + v_i, \tag{3.58}$$

where $x = (\tilde{x}', q')', w = (x', \tilde{w}')', \varepsilon = \delta_1 v + \tilde{\varepsilon}, u = \delta_2 v + \tilde{u}, \beta = ((\tilde{\beta} + \delta_1\alpha_1)', \delta_1\alpha_2')', \gamma = ((\tilde{\gamma} + \delta_2\alpha_1)', \delta_2\alpha_2', \delta_3')'$ and $\alpha = (\alpha_1', \alpha_2')'$. Note that this model is conceptually similar to the model from Section 3.2. If only the reduced form parameters β and γ were of interest, the model could be estimated as in Section 3.2. However, one usually seeks

to estimate the structural parameters $(\tilde{\beta}', \delta_1, \tilde{\gamma}', \delta_2, \delta_3, \alpha)'$. We propose the following estimation strategy: Obtain the first order conditions associated with the likelihood function resulting from the reduced form equations (3.54)-(3.57). The likelihood function is the same as in Section 3.2, because the reduced form equations contain exogenous covariates only. Then estimate the structural parameters by using the first order conditions *and* the normal equations associated with the reduced form equation for y_2 in a Generalized Method of Moments or minimum distance framework. This procedure will give consistent estimates of the structural parameters. Asymptotic normality results can be derived as well, but may be different from those in Section 3.2 (depending on the estimation procedure). However, the results in Ackerberg et al. (2012) can still be applied: The estimation problem may be treated as if it were parametric, and parameter estimates and estimates of standard errors and confidence intervals may be obtained in the usual parametric way. As demonstrated by Ackerberg et al. (2012) for quite general classes of estimators, the standard error estimates are numerically equivalent to those which would be obtained under the correct presumption that the estimation problem was semiparametric.

One word of caution remains, though. The joint distribution implied by the copula and the (unknown) marginal distributions is now the joint distribution of the composite error terms ε and u. This has to be taken into account when interpreting the joint distribution of the error terms associated with the reduced form model.

3.5 Conclusions

In this chapter we proposed a sieve maximum likelihood estimation approach for a copula-based sample selection model. We also provided the asymptotic properties of our proposed estimator and showed that its asymptotic covariance matrix can be easily obtained using statistics software which is capable of dealing with user-supplied

likelihood functions. To facilitate estimation, we showed how closed form likelihood functions can be obtained and how appropriate initial values for maximum likelihood estimation may be chosen. We demonstrated that parametric assumptions on the joint or marginal distributions of error terms can be easily tested for in our framework. We also extended our basis model to the cases of a binary dependent variable and endogeneity of covariates.

The semi-nonparametric maximum likelihood estimation approach of Gallant and Nychka (1987) has not often been used in applied econometrics. One reason may be that no distribution theory is available, which is necessary to compute standard errors and confidence intervals. Another reason may be that the approximation of a two-dimensional density function is rather complex, hence the whole estimation problem is complex as well. The approach derived in this chapter reduces the complexity since only one-dimensional densities have to be approximated. Furthermore, standard errors and confidence intervals can be easily obtained in practice by treating the estimation problem *as if* it were parametric. We thus hope that our exposition fosters the application of semi-nonparametric maximum likelihood estimators to sample selection models, especially if the distribution of the error terms of main and selection equation is of interest.

3.6 Appendix

Proof of Theorem 2:

We prove Theorem 2 by verifying that the Conditions 4.1-4.5 in Chen (2007) are fulfilled. For convenience, we restate these conditions here. In the following, $\mu_n(g(z)) = \frac{1}{n}\sum_{i=1}^{n}(g(z_i) - E[g(z_i)])$ denotes the empirical process indexed by the function g.

Condition 4.1:

(i) There is $\omega > 0$ *such that* $|f(\theta) - f(\theta_0) - \frac{\partial f(\theta_0}{\theta}[\theta - \theta_0]| = O(||\theta - \theta_0||^{\omega})$ *uniformly in* $\theta \in \Theta$ *with* $||\theta - \theta_0|| = o(1)$.

(ii) $||\frac{\partial f(\theta_0)}{\partial \theta}|| < \infty$.

(iii) There is $\pi_n v^* \in \mathcal{A}_n$ such that $||\pi_n v^* - v^*|| \times ||\hat{\alpha}_n - \alpha_0|| = o_p(n^{-1/2})$.

Condition 4.2':

$$\sup_{\{\bar{\alpha}\in\mathcal{A}_n:||\bar{\alpha}-\alpha_0||<\delta_n\}} \mu_n\left(\frac{\partial l(\bar{\alpha}, z)}{\partial \alpha}[\pi_n v^*] - \frac{\partial l(\alpha_0, z)}{\partial \alpha}[\pi_n v^*]\right) = o_p(n^{-1/2}). \tag{3.59}$$

Condition 4.3':

$$E\left[\frac{\partial l(\hat{\alpha}_n, z)}{\partial \alpha}[\pi_n v^*]\right] = \langle \hat{\alpha}_n - \alpha_0, \pi_n v^* \rangle + o(n^{-1/2}). \tag{3.60}$$

Condition 4.4:

(i) $\mu_n(\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*]) = o_p(n^{-1/2})$.

(ii) $E[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^*]]$.

Condition 4.5: $n^{1/2}\mu_n(\frac{\partial l(\alpha_0,z)}{\partial \alpha}[v^*]) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{v^*})$, with $\sigma^2_{v^*} > 0$.

Condition 4.1 (i) is fulfilled with $\omega = \infty$. Condition 4.1 (ii) is fulfilled by Assumption 12. Condition 4.1 (iii) is satisfied by Assumption 13 and the consistency of $\hat{\alpha}_n$.

Condition 4.2 can be verified using Theorem 3 from Chen et al. (2003). Theorem 3 requires continuity conditions on $m(z, \alpha) = \frac{\partial l(\alpha,z)}{\partial \alpha}[\pi_n v^*] - E[\frac{\partial l(\alpha,z)}{\partial \alpha}[\pi_n v^*]]$, which are satisfied in our case because of Assumption 11.

Condition 4.3 is trivially satisfied because we have used the Fisher norm (Chen 2007, p. 5617). Condition 4.4 (i) is fulfilled because we have i.i.d. observations and

$$E\left[\mu_n(\frac{\partial l(\alpha_0, z)}{\partial \alpha}[\pi_n v^* - v^*])\right]^2 = n^{-1} E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha}[\pi_n v^* - v^*]\right]^2 \tag{3.61}$$

$$= n^{-1}||\pi_n v^* - v^*||^2 = o(n^{-1}). \tag{3.62}$$

Hence, by the Markov inequality we have that $\mu_n(\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*]) = o_p(n^{-1/2})$. Condition 4.4 (ii) is satisfied since

$$\begin{aligned} E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha}[\pi_n v^*]\right] &= E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha}[\pi_n v^*]\right] - E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha}[v^*]\right] + E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha}[v^*]\right] \\ &= E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha}[\pi_n v^* - v^*]\right], \end{aligned} \tag{3.63}$$

and by Jensen's inequality,

$$\begin{aligned} \left(E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha}[\pi_n v^* - v^*]\right]\right)^2 &\leq E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha}[\pi_n v^* - v^*]\right]^2 \\ &= ||\pi_n v^* - v^*||^2 = O(K^{-2\psi}) = o(n^{-1}) \end{aligned} \tag{3.64}$$

by Assumption 13, hence $E\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*]\right] = o(n^{-1/2})$. Condition 4.5 is fulfilled because we have i.i.d. observations and

$$\sigma^2_{v^*} = Var\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha}[v^*]\right] \tag{3.65}$$

$$= Var\left[\frac{\partial l(\alpha_0, z)}{\partial \theta} - \frac{\partial l(\alpha_0, z)}{\partial f'}[w](v^*_\theta)\right] \tag{3.66}$$

$$= (v_\theta^*)E[D_{w^*}(z)D_{w^*}(z)'](v_\theta^*)' \tag{3.67}$$

$$= \lambda'(E[D_{w^*}(z)D_{w^*}(z)'])^{-1}E[D_{w^*}(z)D_{w^*}(z)'](E[D_{w^*}(z)D_{w^*}(z)'])^{-1}\lambda \tag{3.68}$$

$$= \lambda'(E[D_{w^*}(z)D_{w^*}(z)'])^{-1}\lambda > 0 \tag{3.69}$$

by Assumption 12. By Theorem 4.3 in Chen (2007) it follows that $n^{1/2}(f(\hat{\theta}_n) - f(\theta_0)) \xrightarrow{d} \mathcal{N}(0, \sigma_{v^*}^2)$, hence

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, (E[D_{w^*}(z)D_{w^*}(z)'])^{-1}) = \mathcal{N}(0, I_*(\theta_0)^{-1}). \tag{3.70}$$

■

Chapter 4

One-Step Sieve Estimation of a Sample Selection Model with Endogeneity - with an Application to Estimating the Female Returns to Education

This chapter is a major revision of the discussion paper No. 504, Department of Economics and Business Administration, Leibniz University Hannover (Schwiebert, 2012c). I thank Melanie Schienle and Jeffrey M. Wooldridge for providing valuable comments.

4.1 Introduction

The sample selection model is used when the observed data is considered a nonrandom sample from the underlying population. It is well-known that not accounting for sample selectivity may result in inconsistent estimates of the parameters of interest. Heckman

(1979) demonstrated that the sample selection problem can be interpreted as an omitted variable bias problem, and suggested procedures to overcome the selection bias. Heckman assumed that the error terms of the main equation of interest and the selection equation have a bivariate normal distribution. With this assumption, the model can be estimated by ordinary least squares including a control function (the inverse Mills ratio term) or by maximum likelihood. However, since the bivariate normality assumption may be quite restrictive, several authors have proposed semiparametric estimation procedures for sample selection models which give consistent estimates under far weaker conditions (e.g., Gallant and Nychka, 1987; Powell, 1987; Ahn and Powell, 1993; Das et al., 2003; Newey, 2009).

In this chapter, we consider another semiparametric estimation procedure for a sample selection model based on the method of sieves (Chen, 2007). As it is common in the literature, we augment the main equation of interest with an unknown control function term which accounts for the sample selectivity. This term can be considered a generalization of the inverse Mills ratio term known from the Heckman (1979) model. The main equation then consists of finite dimensional structural parameters and the unknown control function which is an infinite dimensional nuisance parameter. The selection equation, on the other hand, is also associated with finite dimensional structural parameters and the cumulative distribution function (c.d.f.) of the selection equation's error term, where the latter is an infinite dimensional nuisance parameter. Our strategy is to estimate the main equation including the control function and the selection equation simultaneously (i.e., one-step) by the method of sieves in a Generalized Method of Moments framework. To do this, we approximate the unknown control function of the main equation and the c.d.f. of the selection equation's error term by simpler basis functions. The coefficients associated with these basis functions are then estimated jointly with the structural parameters of interest.

Our model setup and estimation procedure differ in some aspects from those pro-

posed in the literature. First, our estimation procedure is one-step. Two-step methods seem to dominate the literature on sample selection models because they are computationally less demanding than one-step procedures. However, to conduct inference the researcher has to adjust the estimator's covariance matrix from the second step for the uncertainty due to the first-step estimation. This typically involves the computation of the derivative of the optimized objective function with respect to the first-step parameters. Researchers sometimes avoid the effort associated with such adjustment and use bootstrap procedures instead. This, however, may be computationally demanding depending on the number of bootstrap replications, the sample size and the numbers of parameters to be estimated. A one-step procedure, on the other hand, does not require adjustments of the estimator's covariance matrix, but provides a valid estimate immediately. Furthermore, a one-step procedure is typically more efficient than a two-step method since the correlation between the error terms of the estimating equations can be exploited, which results in efficiency gains.

Second, our econometric model allows covariates to be endogenous. Endogeneity of covariates in sample selection models has been analyzed by Das et al. (2003), Dustmann and Rochina-Barrachina (2007), Wooldridge (2010) and Semykina and Wooldridge (2010). These authors, however, do only consider endogeneity of covariates in the main equation.[1] We extend this setting by allowing the covariates of the selection equation to be endogenous as well. Such a setting is quite realistic since many covariates enter both the main and the selection equation.

We consider an important empirical application of our estimator. Our goal is to obtain an estimate of the female returns to education. Our sample includes married women only, as these may decide whether to be a homemaker or to participate in the labor market. Our main equation is a wage equation, and the coefficient associated

[1] To be precise, Das et al. (2003) mention the possibility to extend their model to the case of endogenous covariates in the selection equation, but do not provide asymptotic distribution theory for this extension.

with the education variable in this wage equation is known as the returns to education. Obtaining a consistent estimate of the coefficient of education is difficult, though. First, the wage is only observed for women who participate in the labor market. Since only women participate in the labor market whose (potential or actual) wage exceeds their reservation wage, the sample of working women is nonrandom. Ordinary least squares estimation of the wage equation is thus inappropriate due to sample selectivity. On the other hand, employing a sample selection model to account for the selectivity may be not sufficient. A variable like education is affected by unobserved latent factors like ability. Ability is also likely to affect the wage as well as the probability of labor market participation. Since ability is unobserved it is captured by the error terms of our econometric model. As a consequence, the education variable will be correlated with these error terms and must thus be considered endogenous. Hence, an appropriate econometric model on which estimation of the female returns to education is based should account for sample selectivity and endogeneity jointly. Our econometric model accounts for both issues and is thus suited to estimate the female returns to education.

We apply our estimation procedure to females from the 1980 U.S. Census. This data set provides us with information on the quarter of birth of individuals. As demonstrated by Angrist and Krueger (1991), the quarter of birth of an individual is correlated with its educational attainment. Since the quarter of birth may be considered to be randomly assigned to individuals, this variable is suited as an instrumental variable for education.

Our empirical results show that it is indeed important to account for selectivity and endogeneity of education jointly. In particular, we find that the returns to education are *smaller* than those obtained under estimation strategies which do not account for the joint presence of selectivity and endogeneity of education.

The remainder of this chapter is organized as follows. In Section 4.2 we present the econometric model and propose our estimation procedure. Section 4.3 deals with the asymptotic properties of our estimator. Section 4.4 contains the empirical application

of the estimator to the female returns to education. Section 4.5 concludes the chapter.

4.2 Model Setup and Estimation

To facilitate the exposition, we analyze a model where a single endogenous variable enters the main and the selection equation. A generalization to several endogenous variables is straightforward. We consider the following model:

$$y_{1i}^* = \tilde{x}_i'\tilde{\beta} + \delta_1 y_{2i} + \tilde{\varepsilon}_i \tag{4.1}$$

$$d_i^* = \tilde{x}_i'\tilde{\gamma} + \delta_2 y_{2i} + \delta_3 \tilde{w}_i - \tilde{u}_i \tag{4.2}$$

$$d_i = 1(d_i^* > 0) \tag{4.3}$$

$$y_{1i} = \begin{cases} y_{1i}^* & \text{if } d_i = 1 \\ \text{“missing”} & \text{otherwise} \end{cases} \tag{4.4}$$

$$y_{2i} = \tilde{x}_i'\alpha_1 + q_i'\alpha_2 + v_i, \tag{4.5}$$

where $i = 1, \ldots, n$ indexes individuals. The first equation is the main equation (of interest), where y_1^* is the latent dependent variable, $\tilde{x}$ is a vector of exogenous explanatory variables, y_2 is an endogenous explanatory variable and $\tilde{\varepsilon}$ is an error term. The second equation is the selection equation, where d^* is the latent dependent variable, $\tilde{w}$ is a vector of exogenous explanatory variables appearing only in the selection equation, and $\tilde{u}$ is the error term. The third equation expresses that only the sign of d^* is observable. The fourth equation comprises the sample selection mechanism: y_1^* is only observable if the selection indicator d is equal to one. The fifth equation is the reduced form equation for the endogenous explanatory variable y_2, where q is a vector of exogenous explanatory (instrumental) variables and v is an error term. Note that this model contains the sample selection model without endogenous covariates as a special case.

Absorbing equation (4.5) into equations (4.1) and (4.2) gives the following reduced

form model:

$$y_{1i}^* = x_i'\beta + \varepsilon_i \tag{4.6}$$

$$d_i^* = w_i'\gamma - u_i \tag{4.7}$$

$$d_i = 1(d_i^* > 0) \tag{4.8}$$

$$y_{1i} = \begin{cases} y_{1i}^* & \text{if } d_i = 1 \\ \text{“missing”} & \text{otherwise} \end{cases} \tag{4.9}$$

$$y_{2i} = x_i'\alpha + v_i, \tag{4.10}$$

where $x = (\tilde{x}', q')', w = (x', \tilde{w}')', \beta = ((\tilde{\beta} + \delta_1\alpha_1)', \delta_1\alpha_2')', \gamma = ((\tilde{\gamma} + \delta_2\alpha_1)', \delta_2\alpha_2', \delta_3')'$, $\alpha = (\alpha_1', \alpha_2')'$, and ε and u denote appropriate error terms.

We make the following assumptions:

Assumption 1: *$\{(\tilde{x}_i, \tilde{w}_i, q_i, \tilde{\varepsilon}_i, \tilde{u}_i, v_i)\}_{i=1}^n$ are i.i.d. from some underlying distribution.*

Assumption 2: (i) *$(\tilde{x}, \tilde{w}, q)$ and $(\tilde{u}, v)$ are independent. (ii) The first element of δ_3 is equal to one or minus one. (iii) q contains at least one variable (except from a constant term) which is not included in $\tilde{w}$. (iv) $\tilde{x}$ and $\tilde{w}$ do not contain a constant term. (v)* $E[\varepsilon|d = 1, x = a, w = b] = E[\varepsilon|w'\gamma = b'\gamma] = g(b'\gamma)$.

Assumption 1 is a standard i.i.d. assumption. Assumptions 2 contains identification conditions. Assumption 2 (i) is a standard condition to identify the reduced form parameters of the selection equation. Assumption 2 (ii) and (iv) also identify the reduced form parameters of the selection equation by imposing location and scale restrictions; cf. Klein and Spady (1993). Assumption 2 (iii) imposes an exclusion restriction associated with the selection equation, which is a standard condition in semi-nonparametric estimation of sample selection models. Assumption 2 (iv) and (v) identify the reduced form parameters of the main equation. The function $g(\cdot)$ is the control function which accounts for the selectivity effect.

For the observations with $d_i = 1$, we may rewrite the main equation as

$$y_{1i} = x_i'\beta + g(w_i'\gamma) + r_i, \tag{4.11}$$

where r is a mean-zero error term. Then, we have the following three conditional moment restrictions which form the basis of our minimum distance estimation approach:

$$E[d_i(y_{1i} - x_i'\beta - g(w_i'\gamma))|w_i] = 0 \tag{4.12}$$

$$E[d_i - H(w_i'\gamma)|w_i] = 0 \tag{4.13}$$

$$E[y_{2i} - x_i'\alpha|w_i] = 0, \tag{4.14}$$

where $H(\cdot)$ is the c.d.f. of u. Note that $Pr(d = 1|w = b) = Pr(d = 1|w'\gamma = b'\gamma) = Pr(u < b'\gamma) = H(b'\gamma)$.

If g and H were known, estimation would be straightforward in a (parametric) minimum distance (or GMM) setting. However, since these functions are nuisance parameters which are unknown, a semiparametric estimation procedure is needed. We propose estimation based on the method of sieves. That is, we approximate the unknown functions g and H by simpler basis functions and estimate the parametric part of the model jointly with the coefficients associated with these basis functions. For example, g may be approximated by a linear expansion:

$$g_n(t) = \sum_{k=0}^{K_n} a_k A_k(t), K_n \to \infty, \frac{K_n}{n} \to 0, \tag{4.15}$$

where g_n is an approximation to g based on K_n sieve coefficients, $\{A_k(\cdot) : k \geq 0\}$ denote known basis functions and $\{a_k : k \geq 0\}$ are unknown sieve coefficients which must be estimated. Note that K_n increases with the sample size but at a slower rate. Other classes of approximating functions are analyzed in Chen (2007). The idea of sieve estimation is that in the limit, as n approaches infinity, the approximating function

becomes equal to the actual function. When approximating the c.d.f. H, however, we have to make sure that the approximating function only takes values between zero and one (because it is a c.d.f.). Instead of approximating H in a similar manner as g, it may be more convenient to write $H(t) = \Phi(h(t))$, where $\Phi(\cdot)$ is a known c.d.f., and approximate h instead of H. h may indeed be approximated in a similar manner as g, while $\Phi(\cdot)$ ensures that the approximating c.d.f. always takes values between zero and one.

Let $\theta = (\tilde{\beta}', \delta_1, \tilde{\gamma}', \delta_2, \tilde{\delta}_3', \alpha_1', \alpha_2')'$ denote the finite dimensional parameter vector of interest, where $\tilde{\delta}_3$ contains all elements of δ_3 except for the first (which has been set equal to one due to identification). Furthermore, define

$$\rho(\theta, g, h, z_i) = \begin{bmatrix} d_i(y_{1i} - x_i'\beta - g(w_i'\gamma)) \\ d_i - \Phi(h(w_i'\gamma)) \\ y_{2i} - x_i'\alpha \end{bmatrix}, \tag{4.16}$$

where $z_i = (y_{1i}, y_{2i}, d_i, x_i, w_i)$ denotes the data. We propose to obtain an estimate $\hat{\theta}_n$ of θ by minimizing the criterion function

$$\frac{1}{n}\sum_{i=1}^{n} \rho(\theta, g_n, h_n, z_i)'\Sigma(w_i)^{-1}\rho(\theta, g_n, h_n, z_i) \tag{4.17}$$

over θ and the unknown sieve coefficients associated with g_n and h_n, where $\Sigma(\cdot)$ denotes a positive definite weighting matrix.

4.3 Asymptotic Properties

In this section, we derive consistency and asymptotic normality of our proposed sieve minimum distance estimator using the results in Chen (2007). First, let $\mathcal{A} = \Theta \times \mathcal{G} \times \mathcal{H}$ denote the parameter space. $\mathcal{G}$ and $\mathcal{H}$ denote the function spaces in which the true

functions g and h are included. On the other hand, $\mathcal{G}_n$ and $\mathcal{H}_n$ denote the sieve spaces, i.e., the classes of functions used to approximate g and h. For instance, if we consider linear sieves as in Section 4.2, we would have

$$\mathcal{F}_n = \left\{ f_n(x) = \sum_{k=0}^{K_n} a_{k,n} A_{k,n}(x) \right\}, K_n \to \infty, \frac{K_n}{n} \to 0, (\mathcal{F}, f) \in \{(\mathcal{G}, g), (\mathcal{H}, h)\}. \quad (4.18)$$

As in the last section, the sieve minimum distance estimator is defined as

$$\hat{\alpha}_n = \arg\min_{\alpha \in \mathcal{A}_n} Q(\alpha; Z) = \frac{1}{n} \sum_{i=1}^{n} \rho(\alpha, z_i)' \Sigma(w_i)^{-1} \rho(\alpha, z_i), \quad (4.19)$$

where $Q(\cdot)$ denotes the criterion function, $Z = \{z_i\}_{i=1}^n$, $\alpha = (\theta', g, h)'$ and $\hat{\alpha}_n = (\hat{\theta}_n', \hat{g}_n, \hat{h}_n)' \in \Theta \times \mathcal{G}_n \times \mathcal{H}_n = \mathcal{A}_n$. The true value of the parameter vector is denoted as $\alpha_0 = (\theta_0', g_0, h_0)' \in \mathcal{A}$.

Our first goal is to derive consistency of our proposed estimator. Suppose that $d(\cdot, \cdot)$ is a (pseudo) metric on $\mathcal{A}$. We make the following assumptions (in addition to Assumptions 1-2), which are taken from Conditions 3.1', 3.2', 3.3', 3.4 and 3.5 in Chen (2007):

ASSUMPTION 3: *(i)* $E[Q(\alpha, Z)]$ *is continuous at* $\alpha_0 \in \mathcal{A}$, $E[Q(\alpha_0, Z)] < \infty$
(ii) for all $\epsilon > 0, E[Q(\alpha_0, Z)] < \inf_{\{\alpha \in \mathcal{A}:\, d(\alpha, \alpha_0) \geq \epsilon\}} E[Q(\alpha, Z)]$.

ASSUMPTION 4: $\mathcal{A}_k \subseteq \mathcal{A}_{k+1} \subseteq \mathcal{A}$ *for all* $k \geq 1$*; and for any* $\alpha \in \mathcal{A}$ *there exists a sequence* $\pi_k \alpha_0 \in \mathcal{A}_k$ *such that* $d(\alpha_0, \pi_k \alpha_0) \to 0$ *as* $k \to \infty$.

ASSUMPTION 5: *For each* $k \geq 1$,
(i) $Q(\alpha, Z)$ *is a measurable function of the data* Z *for all* $\alpha \in \mathcal{A}_k$*; and*
(ii) for any data Z, $Q(\alpha, Z)$ *is upper semicontinuous on* $\mathcal{A}_k$ *under the metric* $d(\cdot, \cdot)$.

ASSUMPTION 6: *The sieve spaces,* $\mathcal{A}_k$*, are compact under* $d(\cdot, \cdot)$.

ASSUMPTION 7: *For all* $k \geq 1, \operatorname{plim}_{n \to \infty} \sup_{\alpha \in \mathcal{A}_k} |Q(\alpha) - E[Q(\alpha)]| = 0$.

Assumption 3 is an identification condition which implies that the criterion function is uniquely minimized at the true parameter vector α_0. Assumptions 4 and 6 contain

assumptions on the sieve spaces. In particular, it is assumed that asymptotically the difference between an (unknown) function and its sieve approximation tends to zero. Assumption 5 is a continuity condition, while Assumption 7 assumes uniform convergence of the sample criterion function to its population counterpart over the sieves.

We establish the following consistency theorem:

THEOREM 1: *Suppose that Assumptions 1-7 hold. Then* $d(\hat{\alpha}_n, \alpha_0) = o_p(1)$.

PROOF: *See Chen (2007), pp. 5589-5591.* ■

In order to establish asymptotic normality, we show that Assumptions 4.1 and 4.2 in Chen (2007) are fulfilled. We derive asymptotic normality only for the structural parameters of interest contained in θ. Our exposition closely follows Chen (2007, ch. 4).

Let

$$\frac{\partial \rho(\alpha_0, z)}{\partial \alpha'}[\alpha - \alpha_0] = \lim_{\omega \to 0} \frac{\rho(\alpha_0 + \omega[\alpha - \alpha_0], z) - \rho(\alpha_0, z)}{\omega} \tag{4.20}$$

be the directional derivative of $\rho(\alpha_0, z)$ in the direction $[\alpha - \alpha_0]$ and suppose that it is well defined for almost all z. Let V be the completion of the space spanned by $\mathcal{A} - \alpha_0$. As in Chen (2007), we define the norm on this space as

$$||v||^2 = E\left[\frac{\partial \rho(\alpha_0, z)'}{\partial \alpha'}[v]\Sigma(w)^{-1}\frac{\partial \rho(\alpha_0, z)}{\partial \alpha'}[v]\right], \tag{4.21}$$

which induces the inner product

$$\langle v, \tilde{v} \rangle = E\left[\frac{\partial \rho(\alpha_0, z)'}{\partial \alpha'}[v]\Sigma(w)^{-1}\frac{\partial \rho(\alpha_0, z)}{\partial \alpha'}[\tilde{v}]\right]. \tag{4.22}$$

Let $f(\theta_0) = \lambda'\theta_0$, where λ is an arbitrary unit vector with the same dimension as θ. It follows from the Riesz representation theorem that there exists $v^* \in V$ such that, for

any $\alpha - \alpha_0 \in V$,

$$\lambda'(\theta - \theta_0) = \langle \alpha - \alpha_0, v^* \rangle \tag{4.23}$$

with $||v^*|| < \infty$.

To proceed further, it is necessary to compute the Riesz representer v^*. Define

$$D_{r_j}(w) = \frac{\partial \rho(\alpha_0, z)}{\partial \theta_j} - \frac{\partial \rho(\alpha_0, z)}{\partial f'}[r_j], \quad j = 1, \ldots, \dim(\theta), \tag{4.24}$$

where $f = (g, h)'$. Then, the Riesz representer $v^* = ((v_\theta^*)', (v_f^*)')'$ is given by

$$v_f^* = -r^* v_\theta^* \tag{4.25}$$

$$v_\theta^* = (E[D_{r^*}(w)'\Sigma(w)^{-1}D_{r^*}(w)])^{-1}\lambda \tag{4.26}$$

$$r_j^* = \arg\inf_{r_j} E[D_{r_j}(w)'\Sigma(w)^{-1}D_{r_j}(w)], \tag{4.27}$$

where $r = (r_1, \ldots, r_{\dim(\theta)})$, and $D_r(w) = (D_{r_1}(w), \ldots, D_{r_{\dim(\theta)}}(w))$ is a $(\dim(\rho) \times \dim(\theta))$-matrix.

We make the following assumptions:

ASSUMPTION 8: *$\theta \in \text{int}(\Theta), \Theta$ a compact subset of $\mathbb{R}^{\dim(\theta)}$.*

ASSUMPTION 9: *$\rho(\alpha, z)$ is twice continuously pathwise differentiable with respect to $\alpha \in \mathcal{A}$ and $||\alpha - \alpha_0|| = o(1)$, and the derivatives are uniformly bounded with respect to $\alpha \in \mathcal{A}$ and z.*

ASSUMPTION 10: *$E[D_{r^*}(w)'\Sigma(w)^{-1}D_{r^*}(w)]$ is positive definite.*

ASSUMPTION 11: *$\Sigma(w)$ and $\Sigma_0(w) = Var[\rho(\alpha, z)|w]$ are positive definite and bounded uniform over w.*

ASSUMPTION 12: *There is $\pi_n v^* \in \mathcal{A}_n$ such that $||\pi_n v^* - v^*|| = O(K^{-\psi}) = o(n^{-1/2})$.*

Assumptions 8-11 are standard. Assumption 12 places a smoothness condition on the Riesz representer v^*, which is similar to Assumption 3 of Newey (1997). We establish

the following theorem:

THEOREM 2: *Suppose that Assumptions 1-2 and 8-12 hold, and that* $||\hat{\alpha}_n - \alpha_0|| = o_p(1)$. *Then* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_1^{-1} V_2 V_1^{-1})$, *where*

$$V_1 = E[D_{r^*}(w)' \Sigma(w)^{-1} D_{r^*}(w)] \tag{4.28}$$

$$V_2 = E[D_{r^*}(w)' \Sigma(w)^{-1} \Sigma_0(w) \Sigma(w)^{-1} D_{r^*}(w)]. \tag{4.29}$$

PROOF: *Chen (2007) proves that under Assumptions 4.1 and 4.2 in that paper, the sieve minimum distance estimator has the asymptotic distribution stated in Theorem 2. Hence, we have to check whether these assumptions are satisfied in our case. Assumptions 4.1 (i), (ii) and 4.2 (i) are identical to our Assumptions 8, 10 and 11. Assumption 4.1 (iii) is implied by the consistency of* $\hat{\theta}_n$ *and Assumption 12. Assumptions 4.2 (ii) and (iii) are implied by Assumption 9, the definition of the norm (4.21), and our conditional moment restrictions (4.12)-(4.14).* ■

In order to calculate standard errors or confidence intervals for $\hat{\theta}_n$, one needs estimates of the the matrices V_1 and V_2. Such estimates can be obtained in the following way (see Remark 4.2 in Chen, 2007). Let

$$\hat{r}_j^* = \arg \min_{r_j \in (\mathcal{G}_n \times \mathcal{H}_n)} \frac{1}{n} \sum_{i=1}^{n} [(\hat{D}_{r_j})(w)' \Sigma(w_i)^{-1} \hat{D}_{r_j}(w)], \tag{4.30}$$

with

$$\hat{D}_{r_j}(w) = \frac{\partial \rho(\hat{\alpha}_0, z)}{\partial \theta_j} - \frac{\partial \rho(\hat{\alpha}_0, z)}{\partial f'}[r_j], \quad j = 1, \ldots, \dim(\theta). \tag{4.31}$$

Define $\hat{D}_r(w) = (\hat{D}_{r_1}(w), \ldots, \hat{D}_{r_{\dim(\theta)}}(w))'$. Then, consistent estimators $\hat{V}_1$ and $\hat{V}_2$ of V_1 and V_2 are given by

$$\hat{V}_1 = \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{\hat{r}^*}(w_i)' \Sigma(w_i)^{-1} \hat{D}_{\hat{r}^*}(w_i) \tag{4.32}$$

$$\hat{V}_2 = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_{\hat{r}^*}(w_i)'\Sigma(w_i)^{-1}\hat{\Sigma}_0(w_i)\Sigma(w_i)^{-1}\hat{D}_{\hat{r}^*}(w_i), \tag{4.33}$$

where $\hat{\Sigma}_0(w)$ is a consistent estimator of $\Sigma_0(w) = Var[\rho(\alpha, z)|w]$.

However, Ackerberg et al. (2012) show that there is actually a simpler way to obtain an estimate of the asymptotic covariance matrix $V_1^{-1}V_2V_1^{-1}$. Suppose there is a fictitious practitioner who uses the same sieve spaces to approximate the unknown functions g and h, but she treats the number of sieve terms, K, as fixed rather than increasing with the sample size. Hence, the practitioner faces a parametric estimation problem, and inference can be done as usual. Let $\tilde{\alpha} = (\theta', \kappa')'$ denote the parameter vector to be estimated by our fictitious practitioner, where κ contains the sieve coefficients. The parametric practitioner may calculate the following consistent estimator (consistent from the practitioner's perspective) of the asymptotic covariance matrix of $\hat{\tilde{\alpha}}_n$:

$$\widehat{aVar(\hat{\tilde{\alpha}}_n)} = \hat{\tilde{V}}_1^{-1}\hat{\tilde{V}}_2\hat{\tilde{V}}_1^{-1}, \tag{4.34}$$

where

$$\hat{\tilde{V}}_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial\rho(\hat{\tilde{\alpha}}_n, z_i)'}{\partial\tilde{\alpha}'}\Sigma(w_i)^{-1}\frac{\partial\rho(\hat{\tilde{\alpha}}_n, z_i)}{\partial\tilde{\alpha}'} \tag{4.35}$$

$$\hat{\tilde{V}}_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial\rho(\hat{\tilde{\alpha}}_n, z_i)'}{\partial\tilde{\alpha}'}\Sigma(w_i)^{-1}\hat{\Sigma}_0(w_i)\Sigma(w_i)^{-1}\frac{\partial\rho(\hat{\tilde{\alpha}}_n, z_i)}{\partial\tilde{\alpha}'}. \tag{4.36}$$

Note that $\hat{\tilde{V}}_1$ and $\hat{\tilde{V}}_2$ have larger dimensions than $\hat{V}_1$ and $\hat{V}_2$ due to the sieve coefficients. The practitioner's estimate of the asymptotic covariance matrix of $\hat{\theta}_n$ is then given by the upper left block of $\widehat{aVar(\hat{\tilde{\alpha}}_n)}$.

Ackerberg et al. (2012) provide the following important result: *Despite the fact that the parametric model is misspecified, the parametric estimate of the asymptotic covariance matrix of $\hat{\theta}_n$ is numerically equivalent to the semiparametric estimate $\hat{V}_1^{-1}\hat{V}_2\hat{V}_1^{-1}$.*

Therefore, a researcher who wants to carry out sieve minimum distance estimation can treat the semiparametric estimation problem *as if* it were parametric and estimate the asymptotic covariance of the estimator in the usual parametric way. In particular, any econometrics software which is capable of performing (nonlinear) minimum distance estimation can be used to obtain estimates and valid standard errors of these estimates. Thus, the practical implementation of our proposed sieve minimum distance estimator is fairly simple.

Finally, we discuss the efficiency of our estimator. In parametric minimum distance estimation, it is well known that the optimal weighting matrix is $\Sigma(w) = \Sigma_0(w) = Var[\rho(\alpha, z)|w]$. The same result holds for sieve minimum distance estimation. Ai and Chen (1999) suggest the following procedure to obtain an efficient estimator:

1. Obtain a consistent (but inefficient) estimator $\hat{\hat{\alpha}}_n$ by using the identity matrix as the weighting matrix.

2. Use these estimates to compute a consistent estimator $\hat{\Sigma}_0(w)$ of $\Sigma_0(w) = Var[\rho(\alpha, z)|w]$.

3. Use $\hat{\Sigma}_0(w)$ as the weighting matrix to obtain the final estimator $\hat{\alpha}_n$.

Then, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_1^{-1})$, and $\hat{\theta}_n$ is semiparametrically efficient (see Chen, 2007, p. 5621).

4.4 Empirical Application

In this section, we apply our proposed estimation procedure to 1980 U.S. Census data[2] to obtain an estimate of the female returns to education, where we consider married women only. Hence, our goal is to estimate the average percentage wage increase of married women if educational attainment is raised by one year of schooling. The reason why we consider married women only is that these women may indeed be able to select

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

among being a homemaker or participating in the labor market, as the husband's income may be sufficiently high to make a living in either case.

The "selected women" in our analysis comprise those women who worked full time full year (FTFY) in the previous year, i.e. who worked at least 36 hours per week and at least 50 weeks in the previous year. The reason for this sample restriction is that it is difficult to distinguish whether women who did not work the full year belong to the workforce or not. Moreover, the weekly worked hours of women who worked part time only may be contaminated by measurement error. To avoid dealing with such difficulties, we define the workforce to be the women who worked full time full year in the previous year. Hence, the selection decision of married women amounts to working full time full year or not working full time full year.

Our sample consists of white non-Hispanic women between 25 and 54 years of age not living in group quarters. The hourly wage of women belonging to the workforce as defined in the last paragraph is calculated as the annual salary income divided by (52 times the usual hours of work). We restricted our sample to FTFY working women above the 5th and below the 95th percentile of the overall wage distribution (including non-married women), since we are interested in the results for women located in the the main part of the wage distribution (as results in the tails may be different).[3] We also eliminated self-employed workers and observations for which incomes have been imputed by a "hot deck" procedure. Furthermore, we excluded unemployed women as we cannot say whether these (potentially) belong to the FTFY women or to the remaining population.

Our basic model is given by

$$lwage_i = \tilde{\beta}_1 age_i + \tilde{\beta}_2 age2_i + \delta_1 educ_i + \tilde{\varepsilon}_i \quad \text{if } d_i = 1 \tag{4.37}$$

$$d_i = 1(\tilde{\gamma}_1 age_i + \tilde{\gamma}_2 age2_i + \delta_2 educ_i - nchlt5_i - \tilde{u}_i > 0) \tag{4.38}$$

[3]Since our main equation is linear in the coefficients, we expect that linearity is more likely to hold in the main part of the overall wage distribution.

$$educ_i = \alpha_0 + \alpha_1 age_i + \alpha_2 age2_i + \alpha_3 qtr1_i + \alpha_4 qtr2_i + \alpha_5 qtr3_i + v_i, \tag{4.39}$$

where $lwage$ is the natural logarithm of the hourly wage, age is the years of age, $age2$ is age squared, $educ$ is the years of education, $nchlt5$ is the number of children less than five years of age, and $qtr1, qtr2, qtr3$ are quarter of birth dummies. The remaining notation is the same as in Section 4.2. Note that in light of our discussion on the sample restriction above the selection indicator d is equal to one if a woman belongs to the FTFY workforce and zero otherwise.

In order to identify the parameters of our model, we made the following decisions (recall Assumption 2). First, the number of children less than five years of age ($nchlt5$) is supposed to directly affect the labor market participation decision[4], but not (directly) the wage and educational attainment. This is the exclusion restriction associated with the selection equation. Note that the coefficient of $nchlt5$ has been set equal to -1, which is in accordance with Assumption 2. We set the coefficient to -1 because it seems plausible that the number of children has a negative impact on the probability of labor market participation. Parametric estimation of the selection equation using probit and logit models confirmed that $nchlt5$ has a strong negative impact on the probability of labor market participation, hence our choice seems to be justified.

Our second identification decision concerns the choice of the instrumental variables for education. These have to fulfill two requirements. First, they have to be independent from the error terms of our econometric model (exogeneity of instruments). This implies that only education is directly affected by the instrumental variables, but neither the wage nor the probability of labor market participation. This is the exclusion restriction associated with the reduced form equation for education. The second requirement for instrumental variables is that they be (highly) correlated with the variable to be instrumented, which is education in our case. Instrumental variables which fulfill these

[4]Labor market participation refers only to FTFY working women, as discussed above.

requirements (especially the first one) are hard to find. We exploit the idea used by Angrist and Krueger (1991) to resolve this issue. In their well-known study, Angrist and Krueger (1991) used the quarter of birth (and various interactions) as an instrumental variable for education. The idea is that children in the United States attend school in the year they turn six, where December 31st is the cutoff date. Thus, a child who turns six late in the year attends school at the age of five, whereas a child who turns six early in the year attends school at the age of six. Since the legal high school drop out age in the United States is 16 years of age, Angrist and Krueger (1991) argue that children born late in the year attend school at an earlier age and, thus, stay longer in school. Hence, the quarter of birth has an impact on education and, moreover, it can be considered to be randomly assigned. Therefore, both requirements for (valid) instrumental variables are fulfilled, at least in theory.

Table 4.1 contains some summary statistics for the variables appearing in our model formulation. Note that the selection indicator d has a mean of about 0.31, meaning that 31 percent of the women are working full time full year. In total, we have 840,173 observations. Note further that the quarter of birth dummies have rather similar means, which indicates that the instrumental variable quarter of birth is indeed randomly assigned.

Before we proceed, we provide some evidence that the second requirement for valid instrumental variables, i.e. that the instrumental variables have an impact on education, is fulfilled. In Table 4.2 we listed the means of education by quarter of birth. As can be seen, the mean of education is largest in the last two quarters of the year, which is in accordance with the Angrist and Krueger (1991) idea. Furthermore, we estimated the reduced form equation for education in advance. Table 4.3 contains the results. The estimated coefficients of $qtr1$ to $qtr3$ support the descriptive evidence that women born late in the year have higher education on average (note that the fourth quarter, i.e. $qtr4$ serves as the base category). Important is the value of the F statistic in the

last row of Table 4.3. The F statistic is associated with a test of the joint hypothesis that the coefficient of each instrumental variable is equal to zero. A large value of the F statistic indicates that the instrumental variables have a strong impact on education. In our case the F statistic is 36.23, which is not too low but may be considered relatively small, given the sample size of 840,173 observations. We will make some remarks on the strength of the instruments when we present our estimation results below.

We now turn to our estimation procedure. As described in Section 4.2, we used a series expansion to approximate the unknown functions g and h. As basis functions we chose polynomials, so that we approximated g and h by polynomial expansions. For the c.d.f. $\Phi(\cdot)$ we selected the standard normal distribution function. However, before we estimated the system of equations in a minimum distance framework, we determined in advance the number of sieve terms (i.e., K) by estimating reduced form versions of the main equation and the selection equation (i.e., both including the reduced form expression for education) separately and considering which K seemed appropriate. Concerning h we found that $K = 1$ is appropriate. We obtained this conclusion by trying different K's and found that estimates of the structural parameters were rather stable for low K, but became unstable thereafter due to the multicollinearity caused by the increased sieve terms. We then estimated the main equation, using polynomials of the estimated index $w'\hat{\gamma}_n$ (from the selection equation) to approximate the (unknown) control function. Due to similar reasoning as in case of the selection equation, we selected $K = 3$ as the appropriate number of sieve terms for g.

We estimated four different models. Model I is the basic model given by eq. (4.37)-(4.39). In this model, both sample selectivity and endogeneity of education are being accounted for. Model II treats education as exogenous, hence only the first two equations (4.37)-(4.38) are estimated. Note that we can estimate Model II in the same way as Model I, since a sample selection model without endogeneity is a special case of our proposed sample selection model with endogeneity. We selected the number of sieve

terms in an analogous manner as in case of Model I and found that $K = 3$ is appropriate for approximating both g and h. Model III assumes that there is no sample selectivity, but endogeneity of education. We estimate this model by instrumental variable techniques. Model IV comprises the main equation only, hence it assumes that neither sample selectivity nor endogeneity of education is present. This model is estimated by ordinary least squares (OLS).

Estimation results are presented in Table 4.4. The estimated standard errors are heteroskedasticity-robust in case of all models. In particular, the estimated standard errors for Models I and II have been obtained according to the formula in Theorem 2. Moreover, we utilized the results by Ackerberg et al. (2012) and obtained the standard errors from a nonlinear minimum distance estimation routine, proceeding *as if* the estimation problem was parametric. As described in Section 4.3, these standard error estimates are numerically equivalent to those which would have been obtained if we treated the estimation problem as semiparametric (what is indeed the case!).

From Table 4.4 we see that the estimated coefficients of the main equation are rather similar across the four models. In particular, the estimate of the returns to education is approximately 5 percent, so that a one year increase in education is associated with a wage increase of approximately 5 percent on average. Since the estimates of Model I are rather similar to those of Models II (which controls for sample selectivity only) and IV (which does not account for sample selectivity and/or endogeneity), one may raise the question whether it is important at all to account for sample selectivity and/or endogeneity in case of our data. The instrumental variable results in Model III reveal that the coefficient of education may be larger than the OLS estimate, although the standard error of this coefficient is relatively large. Hence, endogeneity seems to be important to some extent. The question is why Model I nevertheless yields an estimate which is similar to the OLS estimate, although endogeneity seems to be present. The answer lies in the selection equation. Note that the coefficient of education in the

selection equation is very different in Models I and II. In Model I, the coefficient is seven times larger. Thus, endogeneity seems to be a very important issue in the selection equation.

The estimates of the selection equation in Models I and II and the instrumental variable estimates of the main equation in Model III indicate that endogeneity is indeed present in main and selection equation. However, the effect of endogeneity in the main equation seems to be offset by the corresponding effect in the selection equation, as both equations are interrelated through the control function. This may explain why the OLS (Model IV) estimates are not very different from those in Model I, which accounts for both endogeneity and sample selectivity.

In sum, we have evidence that Model I which accounts for sample selectivity and endogeneity jointly is the appropriate model to estimate the female returns to education. Regarding the fact that the estimated coefficient of education is similar to those obtained under models which only account for selectivity (Model II) or neither selectivity nor endogeneity (Model IV), we note that the standard error of the estimate in Model I is rather large. This may indicate that the effect of the instrumental variables on the education variable is not sufficiently strong to get precise estimates. Since the confidence interval around the coefficient of education is quite large, the "true" coefficient may be far larger or far smaller than 0.05. One could argue that better instruments would be needed to identify whether OLS under- or overstates the "true" female returns to education.

However, we can also get more precise estimates by choosing the weighting matrix of our minimum distance approach optimally, as described in Section 4.3. For the calculations up to now we simply used the identity matrix as the weighting matrix, which is not optimal of course as the conditional moment restrictions are correlated. To exploit this correlation pattern, we followed Ai and Chen (1999) and used the three-step procedure described in Section 4.3 to obtain efficient estimates. That is, we first

obtained an estimate of the model parameters using the identity matrix as the weighting matrix. We then calculated the residuals

$$\hat{\eta}_{1i} = d_i(y_{1i} - x_i'\hat{\beta}_n - \hat{g}_n(w_i'\hat{\gamma}_n)) \tag{4.40}$$

$$\hat{\eta}_{2i} = d_i - \Phi(\hat{h}_n(w_i'\hat{\gamma}_n)) \tag{4.41}$$

$$\hat{\eta}_{3i} = y_{2i} - x_i'\hat{\alpha}_n \tag{4.42}$$

for each individual $i = 1, \ldots, n$, where we used the notation from Section 4.2. Since we have discrete covariates in w, a nonparametric estimator of $\Sigma_0(w)$ is given by

$$\hat{\Sigma}_0(w) = \frac{\frac{1}{n}\sum_{i=1}^n \hat{\eta}_i\hat{\eta}_i' \, 1(w_i = w)}{\frac{1}{n}\sum_{i=1}^n 1(w_i = w)}, \tag{4.43}$$

where $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3)'$. Put differently, we computed the cell means of $\hat{\eta}\hat{\eta}'$ for each combination of w. We then used $\hat{\Sigma}_0(w)$ as the (optimal) weighting matrix to obtain efficient estimates.

These efficient estimates are presented in Table 4.5. Overall, the estimates are similar to those in Table 4.4. However, the returns to education are estimated with more precision, and they are *smaller* than those in Table 4.4 (3.3 percent instead of 5 percent). Therefore, the efficient estimates indicate that Models II-IV *overstate* the returns to education. That is, when both selectivity and endogeneity issues are jointly accounted for, the estimate of the returns to education is smaller than obtained under models which control for selectivity only (Model II), endogeneity only (Model III) or neither (Model IV).

4.5 Conclusions

In this chapter, we proposed a one-step sieve estimation strategy for a sample selection model with endogenous covariates. We showed that our estimator is consistent

and asymptotically normally distributed. With regard to the conjecture that inference based on a semiparametric model may be complicated, we demonstrated that this is not the case for our estimator. Actually one can treat the estimation problem as if it were parametric and obtain standard errors and confidence intervals in the usual parametric way. As demonstrated by Ackerberg et al. (2012), these (estimated) standard errors and confidence intervals are numerically equal to those obtained under the (correct) presumption that the estimation problem was semiparametric. Put differently, with respect to coefficient estimates and estimated standard errors it does not matter whether we treat the estimation problem as parametric or semiparametric. Hence, our estimation strategy can be easily implemented by practitioners, who might favor parametric models.

We presented an application of our strategy to the (married) female returns to education. Our empirical results clearly demonstrate that both accounting for sample selectivity as well as for endogeneity of education is important, as the returns to education are smaller than obtained under models which do not account for the joint presence of selectivity and endogeneity.

Many researcher using selection models assume (implicitly) that covariates are exogenous. However, such an assumption may not be appropriate, as in our empirical example on the female returns to education. We hope that future research puts more emphasis on endogeneity issues in sample selection models, since it is likely that such issues are important in empirical work.

4.6 Tables

Table 4.1: Summary statistics

Variable	Mean	Standard deviation	Min	Max
lwage	2.4264	0.3312	1.7084	3.1137
d	0.3091	0.3091	0	1
age	38.3789	8.8685	25	54
educ	12.1587	2.3881	0	17
nchlt5	0.3369	0.6312	0	6
qtr1	0.2477	0.4317	0	1
qtr2	0.2411	0.4278	0	1
qtr3	0.2628	0.4402	0	1
qtr4	0.2484	0.4321	0	1
Number obs.:	840,173			

Source: 1980 U.S. Census data; own calculations.

Table 4.2: Mean of education by quarter of birth

	educ
qtr1 (Jan-Feb-March)	12.1338
qtr2 (April-May-June)	12.1169
qtr3 (July-Aug-Sept)	12.1846
qtr4 (Oct-Nov-Dec)	12.1965

Source: 1980 U.S. Census data; own calculations.

Table 4.3: Reduced form estimates for education

	Coeff.	(Std.err)
age	-0.0053	(0.0030)
age2	-0.0006	(0.0000)
qtr1	-0.0516	(0.0073)
qtr2	-0.0625	(0.0073)
qtr3	-0.0088	(0.0072)
const	13.2851	(0.0572)
F statistic	36.23	

Source: 1980 U.S. Census data; own calculations.

Table 4.4: Estimation results

	Model I	Model II	Model III	Model IV
Main equation (dep. var.: lwage)				
age	0.0123	0.0164	0.0145	0.0104
	(0.0009)	(0.0008)	(0.0067)	(0.0007)
age2	-0.0001	-0.0002	-0.0001	-0.0001
	(0.0000)	(0.0000)	(0.0001)	(0.0000)
educ	0.0500	0.0500	0.0734	0.0520
	(0.0240)	(0.0004)	(0.0341)	(0.0003)
const			1.1579	1.5271
			(0.5889)	(0.0142)
Selection equation				
age	-0.1291	-0.0995		
	(0.0022)	(0.0021)		
age2	0.0016	0.0010		
	(0.0001)	(0.0000)		
educ	0.4601	0.0615		
	(0.0717)	(0.0009)		
nchlt5	-1	-1		
	(-)	(-)		
Reduced form equation for educ				
age	-0.0031			
	(0.0030)			
age2	-0.0006			
	(0.0000)			
qtr1	-0.0555			
	(0.0070)			
qtr2	-0.0627			
	(0.0072)			
qtr3	-0.0141			
	(0.0066)			
const	13.2415			
	(0.0562)			

Note: Heteroskedasticity-robust standard errors in parentheses.
Source: 1980 U.S. Census data; own calculations.

Table 4.5: Efficient estimates

	Coeff.	(Std.err)
Main equation (dep. var.: lwage)		
age	0.0118	(0.0005)
age2	-0.0001	(0.0000)
educ	0.0329	(0.0138)
Selection equation		
age	-0.1246	(0.0024)
age2	0.0015	(0.0001)
educ	0.5268	(0.0737)
nchlt5	-1	(-)
Reduced form equation for educ		
age	0.0101	(0.0029)
age2	-0.0007	(0.0000)
qtr1	-0.0538	(0.0064)
qtr2	-0.0639	(0.0068)
qtr3	-0.0152	(0.0056)
const	12.9174	(0.0549)

Source: 1980 U.S. Census data; own calculations.

Chapter 5

Semiparametric Estimation of a Binary Choice Model with Sample Selection

This chapter is a revision of the discussion paper No. 505, Department of Economics and Business Administration, Leibniz University Hannover (Schwiebert, 2012d). I thank Melanie Schienle for providing valuable comments.

5.1 Introduction

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

In this chapter, we consider semiparametric estimation of a binary choice model

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

For these reasons, several authors have analyzed semi-nonparametric methods to estimate the *ordinary* sample selection model which assumes a continuous dependent variable; examples include Gallant and Nychka (1987), Powell (1987), Ahn and Powell (1993), Das et al. (2003) and Newey (2009). Only Klein et al. (2011) provided a semiparametric maximum likelihood estimator for a sample selection model with a binary dependent variable. However, their estimator is one-step, and thus may be computationally demanding if the sample size and/or the number of covariates is large.

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

Our second estimation strategy is based on augmenting the main equation with a "control function" term which accounts for sample selectivity. This term is simply a generalization of the inverse Mills ratio term in the ordinary sample selection model.

We will show how combining "similar" observations makes it possible to get rid of the unknown control function, so that the resulting model can be estimated by known techniques. In particular, we employ the maximum score estimator due to Manski (1975) and the smoothed maximum score estimator due to Horowitz (1992). This approach is conceptually similar to Powell (1987).

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

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

However, the van de Ven and van Praag (1981) model is parametric since it relies on a joint normality assumption on the error terms in the (latent) main equation and the selection equation. As raised above, parametric estimation leads to inconsistent

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

parameter estimates if the parametric assumptions are not fulfilled.

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

The remainder of this chapter is organized as follows. In Section 5.2 we set up the econometric model. In Section 5.3 we review parametric estimation of the model, and in Section 5.4 we propose our semiparametric estimation strategies. In Section 5.5, we conduct a small Monte Carlo study to compare the performance of the parametric and semiparametric estimators in small samples. Section 5.6 contains an empirical example where we apply our estimators to real data. In Section 5.7, we extend our model to the case where explanatory variables are allowed to be endogenous. Finally, Section 5.8 concludes the chapter.

5.2 The Model

The model we consider is given by

$$y_i^* = x_i'\beta + \varepsilon_i \tag{5.1}$$

$$d_i^* = w_i'\gamma + u_i \tag{5.2}$$

$$d_i = 1(d_i^* > 0) \tag{5.3}$$

$$y_i = \begin{cases} 1(y_i^* > 0) \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases}, \tag{5.4}$$

where $i = 1, \ldots, N$ indexes individuals. The first equation is the main equation of interest, where y^* is the latent dependent variable, x is a vector of exogenous explanatory variables and ε is an error term. The second equation is the selection equation, where d^* is the latent dependent variable, w is a vector of exogenous explanatory variables and u is an error term. The third equation expresses that only the sign of d^* is observable. By equation (5.4), the same is true for y^*, but only if d is equal to one. Otherwise, y^* cannot be observed ("missing"). This model differs from the ordinary sample selection model by the fact that the dependent variable of the outcome equation is binary, taking only the values one or zero.

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

ASSUMPTION 1: *w contains at least one variable (with a nonzero coefficient) which is not included in x.*

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

Our next assumption is on the sampling process:

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

We further require that there is no "multicollinearity":

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

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

Having made these basic assumptions, we proceed to consider parametric and semi-

parametric estimation of our model.

5.3 Parametric Estimation

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

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

The log-likelihood function for this model is given by

$$\begin{aligned}\log L(\beta, \gamma) = &\sum_{i=1}^{N} \log(1 - \Phi(w_i'\gamma))1(d_i = 0) + \sum_{i=1}^{N} \log(\Phi_2(x_i'\beta, w_i'\gamma; \rho))1(d_i = 1, y_i = 1) \\ &+ \sum_{i=1}^{N} \log(\Phi_2(-x_i'\beta, w_i'\gamma; -\rho))1(d_i = 1, y_i = 0), \end{aligned} \tag{5.5}$$

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

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

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

We will denote the (parametric) maximum likelihood estimator of β by $\hat{\beta}_H$, where the "H" is a shortcut for "Heckprob", named after the STATA command for estimat-

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

ing a probit model with sample selection. Note that Assumption 1 is technically not needed for identification, since identification is already ensured by the parametric assumptions. However, in practice researchers might not want to identify β by functional form assumptions alone.

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

5.4 Semiparametric Estimation

In order to estimate the model set up in Section 5.2 semiparametrically, we first have to make an identifying assumption. Assumption 1 from Section 5.2 is a necessary assumption to identify the model parameters but it is not sufficient.[3] Here we give two identifying assumptions which give rise to different estimation strategies.

ASSUMPTION 4: *Either*

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

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

holds with probability one.

Assumption 4 (a) allows to estimate the model parameters by semiparametric maximum likelihood. In particular, we propose to estimate β by Rothe's (2009) extension of the Klein and Spady (1993) semiparametric estimation procedure for binary choice

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

models. Note that the log-likelihood function of our observed sample is given by

$$\log L(\beta|\gamma = \hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} y_i \log(G(x_i'\beta, w_i'\hat{\gamma})) + (1 - y_i)\log(1 - G(x_i'\beta, w_i'\hat{\gamma})), \tag{5.6}$$

where $n < N$ is the number of observations for which y is observable. Note that we used a preliminary estimate of γ in the log likelihood function. In principle, we could estimate the parameters of main and selection equation simultaneously which would be efficient. However, two-stage estimators are often preferred due to a reduction of dimensionality and computational issues regarding the stability of numerical optimization routines. Consequently, we assume that the parameters in γ can be consistently estimated by some first-stage estimation procedure:

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

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

$$\hat{G}(x_i'\beta, w_i'\hat{\gamma}) = \frac{\frac{1}{n}\sum_{j\neq i}^{n} y_j \frac{1}{h_x h_w} K(x_i'\beta/h_x)K(w_i'\hat{\gamma}/h_w)}{\frac{1}{n}\sum_{j\neq i}^{n} \frac{1}{h_x h_w} K(x_i'\beta/h_x)K(w_i'\hat{\gamma}/h_w)}, \tag{5.7}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a univariate kernel density function (e.g., the standard normal probability density function) and h_x and h_w are bandwidth parameters satisfying $h_x \to 0$ and $h_w \to 0$ as $n \to \infty$. Then, estimation can be performed as usual with $G(\cdot)$ in (5.6) replaced by (5.7), i.e.,

$$\hat{\beta}_{KS} = \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{G}(x_i'\beta, w_i'\hat{\gamma})) + (1 - y_i)\log(1 - \hat{G}(x_i'\beta, w_i'\hat{\gamma})). \tag{5.8}$$

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

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

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

ASSUMPTION 6:

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

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

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

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

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

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

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

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

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

e) The matrix

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

is positive definite.

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

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

We can now establish the following theorem:

THEOREM 1: *Under Assumptions 1-3, 4 (a), 5 and 6, we have that* $\hat{\tilde{\beta}}_{KS} - \tilde{\beta} = o_p(1)$

.

PROOF: *Our estimation approach is conceptually the same as in Rothe (2009). The difference is that Rothe proposes a control function approach to control for endogeneity of covariates instead of sample selectivity. In his derivations, a reduced form error term (resulting from the reduced form equation of the endogenous explanatory variable) plays the same role as* $w'\gamma$ *does for our estimator. We can thus follow the arguments in Rothe (2009), who derives consistency (and asymptotic normality) of his estimator (by checking whether the conditions in Chen, Linton and van Keilegom, 2003, are fulfilled).*

■

Instead of deriving the asymptotic distribution to conduct inference, we follow Rothe's (2009) arguments and propose to employ the bootstrap for inference. The reason is that the asymptotic distribution of this estimator depends on unknown derivatives which would have to be computed in order to calculate the asymptotic variance. Hence, using the bootstrap is a simpler way to obtain standard errors in practice.

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

$$y_i^* = x_i'\beta + g(w_i'\gamma) + v_i, \quad i = 1, \ldots, n, \tag{5.9}$$

where $v_i \equiv \varepsilon_i - median[\varepsilon_i | w_i'\gamma]$. Since, by construction, v has a conditional median of zero, we could apply Manski's (1975) maximum score estimator to obtain parameter estimates. Again, this is not feasible as the function $g(\cdot)$ is unknown. However, suppose we have two individuals with the same value of $w'\gamma$. In that case, we can subtract equation (5.9) for individual i from the equation for individual j, i.e.,

$$y_i^* - y_j^* = (x_i - x_j)'\beta + g(w_i'\gamma) - g(w_j'\gamma) + v_i - v_j \tag{5.10}$$

$$= (x_i - x_j)'\beta + v_i - v_j. \tag{5.11}$$

The differencing in equations (5.10) and (5.11) resembles the underlying idea of Manski's (1987) conditional maximum score approach for binary panel data. In the panel data approach, an individual specific "fixed effect" is removed by differencing over time for a given individual, while in our case we have a cross sectional data set and use differencing to remove an unknown function.

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

[4]This strategy has also been used by Ahn and Powell (1993). In their case, the control function

Note that despite of the model transformation in equations (5.10) and (5.11) due to differencing we are still able to identify the parameters in β. We simply combine only observations for which $y_i \neq y_j$. Then, we have the following correspondence:

$$y_i^* - y_j^* \begin{cases} > 0 \text{ if } y_i = 1 \wedge y_j = 0 \\ \\ < 0 \text{ if } y_i = 0 \wedge y_j = 1 \end{cases}, \tag{5.12}$$

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

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

$$\hat{\beta}_{MS} = \arg\max_{\beta} -\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |\tilde{y}_{ij} - 1(\tilde{x}_{ij}'\beta > 0)| \frac{1}{h} K(\tilde{w}_{ij}'\hat{\gamma}/h) 1(y_i \neq y_j), \tag{5.13}$$

where $\tilde{y}_{ij} = 1(y_i^* - y_j^* > 0)$, $\tilde{x}_{ij} = x_i - x_j$, $\tilde{w}_{ij} = w_i - w_j$, $K : \mathbb{R} \to \mathbb{R}$ is a univariate kernel density function which is bounded, absolutely integrable and symmetric about zero, and h is a bandwidth parameter which converges to zero when the sample size approaches infinity. Note that the minimization problem in equation (5.13) uses only observations for which $y_i \neq y_j$, and, for the same reasons as given above, preliminary

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

estimates of γ.

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

However, since the objective function in (5.13) is not differentiable it may be difficult to obtain parameter estimates. Horowitz (1992) proposes a smoothed maximum score estimator which features a smooth objective function. Using that estimator, our estimation problem may be written as

$$\hat{\beta}_{SMS} = \arg\max_{\beta} \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (2\tilde{y}_{ij} - 1) \frac{1}{h_x} \Phi(\tilde{x}'_{ij}\beta / h_x) \frac{1}{h_w} K(\tilde{w}'_{ij}\hat{\gamma} / h_w) 1(y_i \neq y_j), \tag{5.14}$$

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

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

In order to establish consistency of $\hat{\tilde{\beta}}_{MS}$ and $\hat{\tilde{\beta}}_{SMS}$ we need some further assumptions which lead to consistency of the maximum score and smoothed maximum score estimators in general, i.e. without sample selectivity. We take these assumptions from Horowitz (1992) and summarize them in Assumption 7:

ASSUMPTION 7:

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

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

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

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

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

We establish the following theorem:

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

PROOF: *First, let*

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

and

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

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

(see Billingsley, 1995, Theorem 16.4.) to handle $\hat{\gamma}$, we obtain that

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

and

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

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

We do not provide asymptotic distribution theory for these estimators since in case of the maximum score estimator the form of the asymptotic distribution is very complicated and not suitable for practical inference; as an alternative, Manski and Thompson (1986) examined the performance of the bootstrap and found encouraging results. In case of the smoothed maximum score estimator Horowitz (1992) derived the asymptotic distribution and reported a relatively weak finite sample performance of the asymptotic

theory, hence he also proposes to use the bootstrap.

We follow these lines of reasoning and propose to use the bootstrap for obtaining standard errors, too; for instance, the standard errors in our empirical example in Section 5.6 have been obtained in that way.

5.5 Monte Carlo Evidence

In this section, we provide some (limited) Monte Carlo evidence on the finite sample performance of our proposed estimators. We not only consider the behavior of the semiparametric estimators from Section 5.3, but also the behavior of the parametric "Heckprob" estimator from Section 5.2. Our simulated model is given by

$$y_i^* = \beta_1 q_i + \beta_2 x_i + \varepsilon_i \tag{5.15}$$

$$d_i^* = x_i + w_i + u_i \tag{5.16}$$

$$\varepsilon_i = u_i + \nu_i \tag{5.17}$$

$$d_i = 1(d_i^* > 0) \tag{5.18}$$

$$y_i = \begin{cases} 1(y_i^* > 0) \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases}, \tag{5.19}$$

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

For u and ν, we consider the following distributions:

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

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

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

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

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

Note that our proposed estimators each require a normalization. We implemented such a normalization by setting β_1 equal to its true value of one. That means, the only parameter to be estimated in the main equation is β_2.

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

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

Klein and Spady estimator, the bandwidths are given ad hoc rather than being part of the optimization problem. We did this because computation of the maximum score and smoothed maximum score estimator is relatively difficult due to the presence of local optima. Instead, we specified the bandwidths in advance so that there is only one parameter, i.e. β_2, over which optimization is performed. To find the optimal value of $\hat{\beta}_2$, we performed a grid search over the interval $[-1, 3]$ with a step length of 0.005.

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

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

Table 5.1 contains the results for distribution (i) and a known selection index. We see from Table 5.1 that, in terms of RMSE, the estimators perform better as the sample size increases (as expected). However, we also see that the mean of the estimates differs slightly from the true value of one even for the relatively large sample size of $N = 1000$. The reason is that the estimates exhibit a lot of variation, as indicated by the standard deviations. Among the three estimators, the maximum score and the smoothed maximum score estimator have lower RMSE's then the Klein and Spady

estimator due to lower standard deviations, which means that these estimators seem to be slightly more precise. We will investigate if this property also holds true for the remaining distributions and in case that the selection index is estimated rather than known in advance.

In Table 5.2, we reconsider distribution (i) but now the selection index is estimated. For obtaining these estimates, we used the same type of estimator in the first stage as in the second stage. That means, for the Klein and Spady estimator we used a Klein and Spady estimator in the first stage, for the maximum score estimator we used a maximum score estimator in the first stage and for the smoothed maximum score estimator we used a smoothed maximum score estimator in the first stage. The idea is that in practice it would seem a bit uncommon to use one semiparametric estimator in the first stage and to use a different semiparametric estimator in the second stage, at least in principle. In the empirical example in Section 5.6 we will, however, provide a practical reason why using different estimators in first and second stage might be sensible.

Note that Table 5.2 also contains results for the parametric "Heckprob" model from Section 5.2. Since distribution (i) implies a normal distribution of the error terms in main and selection equations, one might expect that the "Heckprob" model should perform quite well. Table 5.2 confirms this conjecture. We see that the estimators perform relatively similar with respect to the standard deviation. The differing means are again a result of the relatively great deal of variation of the estimators. When comparing these results with those from Table 5.1 we see that there is not much difference in standard deviations. Hence we may conclude that using the same type of estimator for first and second stage does not lead to stark distortions between the estimators.

In Table 5.3 we consider the mixed normal distribution (ii). We can see that the "Heckprob" estimator performs surprisingly well, having the least bias and the least RMSE among all estimators and for all sample sizes. The standard deviations of the

estimators are generally lower when compared to distribution (i), which is due to the fact that the error term variance is smaller for distribution (ii). Among the semiparametric estimators, the maximum score estimator has the least bias but the largest RMSE.

Table 5.4 contains results for distribution (iii) where we have conditional heteroskedasticity in the selection equation but not in the main equation. In this case, all three semiparametric estimators are consistent, whereas the "Heckprob" estimator is not. However, from Table 5.4 we see that the "Heckprob" estimator performs very well. All estimators exhibit a great deal of variation, which again explains the slight biases of these estimators.

Finally, we consider distribution (iv), where we have conditional heteroskedasticity in the main equation but not in the selection equation. In this case, only the Klein and Spady estimator is consistent. From Table 5.5 we see that not only the Klein and Spady estimator but also the remaining semiparametric estimators perform relatively well. The "Heckprob" estimator, however, exhibits a larger bias than one might have expected. Nevertheless, the "Heckprob" estimator has the smallest RMSE among all estimators.

From these results, we can draw two major conclusions. First, in all considered designs the estimators exhibit a lot of variation (as indicated by the standard deviations). Moreover, we also experienced considerable variation between the estimators. Hence, the first major conclusion is that one needs substantial sample sizes to obtain precise estimates. Second, the parametric "Heckprob" estimator performs relatively well even in situations where it should be biased. Of course, these results may be an artifact of our simulation designs and need not hold in general. However, especially in small sample sizes the parametric estimator may be a sensible alternative due to its favorable RMSE properties. At least one could test the parametric estimator against a semiparametric alternative (at least in a heuristic way, e.g. by considering whether the confidence intervals overlap). When considering the standard deviations of the

semiparametric estimators over the simulations, it seems relatively likely that results based on the parametric estimator would not be rejected empirically. For large sample sizes, however, a semiparametric estimator should be preferred as it relies on considerably fewer assumptions than the parametric estimator. Put differently, the larger the sample size the more obvious it should be when the parametric assumptions are not fulfilled.

5.6 Empirical Example

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

We emphasize that our example is mainly of illustrative purpose. In particular, our empirical specification may be considered to contain not all relevant variables. Our specification is mainly practically motivated, as a large number of explanatory variables makes semiparametric estimation of the model computationally challenging (especially since standard errors are obtained by bootstrapping).

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

may require the capability of getting along with electronic equipment (e.g., personal computers).

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

Our data is taken from the German Socio-Economic Panel (GSOEP) for the year 2002. Our sample consists of 989 married women aged 25 to 35 with German nationality. From these women, 565 are working (57.1 %). Summary statistics of the variables are given in Table 5.6.

We specify our estimators as in the last section. That means, in case of the Klein and Spady estimator we selected the standard normal p.d.f. as the kernel function and let the optimal bandwidth be obtained simultaneously with the parameters of interest; in case of the maximum score estimator, we chose the standard normal p.d.f. as the kernel function and selected a bandwidth according to the rule $h = n^{-1/6.5}$; for the smoothed maximum score estimator we chose the standard normal c.d.f. for $\Phi(\cdot)$ and the standard normal p.d.f. for $K(\cdot)$. We set $h_x = h_w = n^{-1/6.5}$.

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

education equal to one.

Table 5.7 contains the Klein and Spady estimates of the selection equation parameters. As expected, the number of children has a negative impact on the probability of working. For a woman's age we get a U-shaped pattern which is plausible for the sample under consideration, since women start working when they are young, then leave the labor market to raise their children and return thereafter. Standard errors of these estimates have been obtained by performing 100 bootstrap replications.

Table 5.8 contains the second-stage results for the Klein and Spady estimator (KS), the maximum score estimator (MS) and the smoothed maximum score (SMS) estimator. The coefficient of education has been set equal to one due to normalization. We also provide estimates using the "Heckprob" estimator. Standard errors are again based on 100 bootstrap replications. As can be seen from Table 5.8, the coefficient of the number of children is positive over all estimators. However, only in case of the "Heckprob" and smoothed maximum score estimator the coefficient is significantly different from zero (as suggested by the bootstrap standard errors). We get the same picture as in the Monte Carlo simulations from the last section: The semiparametric estimates exhibit a lot of variation and relatively large standard errors. However, the semiparametric estimates also indicate that the effect of the number of children on the latent dependent variable may be larger than suggested by the estimate of the "Heckprob" model. Although it is unlikely that the parametric "Heckprob" model would be rejected by the data when compared to one of these semiparametric alternatives, the semiparametric estimates at least hint that the parametric estimates may be biased, i.e. that the effect of the number of children is larger than the parametric estimate indicates.[5]

Finally, we conducted a small robustness check. While in case of the Klein and Spady estimator the bandwidth is selected optimally by being part of the optimization problem,

[5]It would probably be more interesting to study the effects of the explanatory variables on the dependent variable instead of the latent dependent variable. However, since this chapter is concerned with the estimation of the index parameters in β, we did not consider such marginal effects.

the bandwidths for the maximum score and smoothed maximum score estimator have been selected ad hoc. We, thus, provide some robustness analysis by varying these bandwidths. From Table 5.9 we see that variations of the bandwidths alter the estimates for the maximum score and smoothed maximum score estimator to some extent, but the differences are relatively small. We conclude that estimation results are not very sensitive with respect to bandwidth choice.

5.7 Endogenous Covariates

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

$$x_i^e = z_i'\alpha + \eta_i, \tag{5.20}$$

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

ASSUMPTION 4': *Either*

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

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

holds with probability one.

We can once again implement the estimators proposed above. In case of Assumption

4' (a), we choose a modified Klein and Spady estimator such that

$$\hat{\beta}^e_{KS} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} y_i \log(\hat{G}(x_i'\beta, w_i'\hat{\gamma}, \hat{\eta}_i)) + (1 - y_i) \log(1 - \hat{G}(x_i'\beta, w_i'\hat{\gamma}, \hat{\eta}_i)), \quad (5.21)$$

where

$$\hat{G}(x_i'\beta, w_i'\hat{\gamma}, \hat{\eta}_i) = \frac{\frac{1}{n}\sum_{j\neq i}^{n} y_j \frac{1}{h_x h_w h_\eta} K(x_i'\beta/h_x) K(w_i'\hat{\gamma}/h_w) K(\hat{\eta}_i/h_\eta)}{\frac{1}{n}\sum_{j\neq i}^{n} \frac{1}{h_x h_w h_\eta} K(x_i'\beta/h_x) K(w_i'\hat{\gamma}/h_w) K(\hat{\eta}_i/h_\eta)}. \quad (5.22)$$

Note that the only difference between equation (5.22) and equation (5.7) above is that we have to take the (estimated) reduced form error term of our endogenous variable into account, so that we need an additional kernel function. It is obvious that augmenting the function $G(\cdot)$ with more kernel functions requires large sample sizes to produce reliable estimation results. This problem is even more severe when we have several endogenous explanatory variables. In that case, estimation results might be contaminated by the curse of dimensionality.

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

$$\hat{\beta}^e_{MS} = \arg\min_{\beta} -\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |\tilde{y}_{ij} - 1(\tilde{x}_{ij}'\beta > 0)| \frac{1}{h} K(\tilde{w}_{ij}'\hat{\gamma}/h) \times \frac{1}{h_\eta} K(\tilde{\hat{\eta}}_{ij}/h_\eta) 1(y_i \neq y_j), \quad (5.23)$$

while in the second case

$$\hat{\beta}^e_{SMS} = \arg\max_{\beta} \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (2\tilde{y}_{ij} - 1) \frac{1}{h_x} \Phi(\tilde{x}_{ij}'\beta/h_x) \frac{1}{h_w} K(\tilde{w}_{ij}'\hat{\gamma}/h_w) \times \frac{1}{h_\eta} K(\tilde{\hat{\eta}}_{ij}/h_\eta) 1(y_i \neq y_j), \quad (5.24)$$

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

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

5.8 Conclusion

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

The conclusions from these findings are that (a) one should use the semiparametric estimators in case of large sample sizes and (b) in small samples, the parametric estimator may be preferred if it is successfully tested against a semiparametric alternative. The reason for preferring parametric estimates is that coefficient estimates, especially

in small samples, are estimated with higher precision. However, in large samples it may become obvious that the parametric model is misspecified, hence a semiparametric estimation procedure should be chosen.

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

We also outlined an extension of our semiparametric estimators to the case of endogenous covariates. Endogenous covariates may be a concern in many empirical applications, and not accounting for this endogeneity will lead to inconsistent parameter estimates in general. Extending our estimators to handle endogenous covariates is quite straightforward. However, given the variability of the semiparametric estimators shown in Section 5.4 (which do not control for endogenous covariates), we conjecture that this problem may be even more severe if our estimation procedures also have to account for endogeneity of covariates. This indicates that one needs even larger sample sizes to obtain reliable estimates.

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

5.9 Tables

Table 5.1: Design I - normal + known index

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

Source: Own calculations.

Table 5.2: Design I - normal + unknown index

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

Source: Own calculations.

Table 5.3: Design II - mixed normal

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

Source: Own calculations.

Table 5.4: Design III - heteroskedasticity in selection equation

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

Source: Own calculations.

Table 5.5: Design IV - heteroskedasticity in main equation

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

Source: Own calculations.

Table 5.6: Summary statistics

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

Source: GSOEP data (2002 wave); own calculations.

Table 5.7: Estimates of selection equation parameters

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

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

Source: GSOEP data (2002 wave); own calculations.

Table 5.8: Estimation results

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

Note: Standard errors in parentheses. Standard errors are based on 100 bootstrap replications.
Source: GSOEP data (2002 wave); own calculations.

Table 5.9: Varying the bandwidth

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

Source: GSOEP data (2002 wave); own calculations.

Chapter 6

A Detailed Decomposition for Limited Dependent Variable Models

This chapter is a revision of the discussion paper No. 506, Department of Economics and Business Administration, Leibniz University Hannover (Schwiebert, 2012e). I thank Olaf Hübler, Patrick Puhani and Melanie Schienle for providing valuable comments.

6.1 Introduction

Decomposition methods in economics have been a nascent field of research over the last years. Recently, in the fourth volume of the *Handbook of Labor Economics* a full chapter has been devoted to this topic (Fortin et al., 2011).

In this chapter, we consider a detailed decomposition method for limited dependent variable models, such as probit, logit and tobit models. In contrast to models which are linear in parameters and explanatory variables, a detailed decomposition in limited dependent variable models is not straightforward and comes along with some difficulties, as shown below. Approaches already presented in the literature to tackle these difficulties are not satisfactory as they do not lead to a unique decomposition or do

not take into account the nonlinearity of the model. On the contrary, we propose a decomposition approach which leads to a unique decomposition and accounts for the nonlinearity of the model in a rather intuitive manner.

Our decomposition approach is in the spirit of the famous Oaxaca-Blinder decomposition method. The Oaxaca (1973) and Blinder (1973) decomposition is a well-known and often applied technique to decompose the mean differential in some outcome variable between two groups into a part which is due to differences in observable characteristics (*explained differential*) and another part which is due to differences in unobservable characteristics (*unexplained differential*). A typical example is an analysis of the mean wage differential between, e.g., men and women or white and black people. Under some conditions, the unexplained differential can be attributed to discriminatory behavior of firms, households or other economic institutions; hence the Oaxaca-Blinder decomposition has often been applied to analyze the impact of discrimination.

The Oaxaca-Blinder decomposition in its original version can be applied to econometric models which are linear in parameters and explanatory variables. An extension to limited dependent variable models has been suggested by Bauer and Sinning (2008), for instance. However, Bauer and Sinning only provide a decomposition into the total explained and unexplained differential. We proceed further and consider a *detailed* decomposition of the explained differential, which means that we seek to decompose the explained differential into the contribution of each explanatory variable. In case of wage differentials, a detailed decompositions allows the researcher to make statements like "10 percent of the mean wage differential between men and women can be explained by differences in educational attainment, 20 percent by differences in working experience", and so on.

In this chapter, we focus our attention on the *explained* differential only since the unexplained differential is hard to interpret. In the linear Oaxaca-Blinder decomposition, the unexplained differential is given by differences in coefficients multiplied by a

vector of characteristics of a particular group (e.g., men or women). In nonlinear models such as limited variable models, however, differences in coefficients could also be the result of a misspecified model. Moreover, nonlinear models typically involve nuisance parameters (such as a variance parameter); a detailed decomposition of the unexplained differential would also have to attribute differences in nuisance parameters to the effects of specific factors. A detailed decomposition is then hard to justify economically. A further critique which applies to linear and nonlinear decompositions has been pointed out by Jones (1983). As Jones has shown, a detailed decomposition of the unexplained differential is not unique if there are dummy variables among the list of explanatory variables. The detailed decomposition then depends on the reference category chosen for the dummy variable, hence the decomposition is not unique.

On the contrary, a detailed decomposition of the *explained* differential assumes an identical model structure for the analyzed groups. That means, we relate the mean differential in the outcome variable only to differences in explanatory variables, but holding constant the model structure. In case of the Oaxaca-Blinder decomposition that means we consider differences in (mean) explanatory variables, evaluated at a constant coefficient vector of *one* particular group.

A detailed decomposition in linear models is rather straightforward, since the mean differential in the outcome variable can directly be attributed to the mean differential in the explanatory variables. This, however, is not true for limited dependent variable models. Fairlie (1999, 2005) and Yun (2004) have proposed approaches to obtaining detailed decompositions in such models. As will be shown below, Fairlie's decomposition is path-dependent, which means that the decomposition relies on the ordering of explanatory variables. Since different orderings imply different decomposition results, Fairlie's approach has the drawback that it does not lead to a unique decomposition.

Yun (2004) seeks to tackle the difficulties associated with the nonlinear model structure by two linearizations, thus bringing the model back to the linear case where mean

differences in the outcome variable can directly be related to mean differences in the explanatory variables. However, such a procedure has the drawback that it ignores the nonlinear model structure. For instance, if the outcome differential is located in the tails of the distribution or in case of large differences in the explanatory variables (see Fortin et al., 2011, p. 52), such a linearization is likely to be inadequate.[1]

Our approach is based on a linearization using marginal effects, hence we explicitly account for the nonlinearity of the model in a way which is familiar from the general analysis of limited dependent variable models. Fortin et al. (2011) have already mentioned such a possibility (without providing details, though), but have also remarked that the contribution of each variable derived in such a decomposition would not add up to the total differential. This remark is only partly true. By applying the mean value theorem, we will show that there is exactly one marginal effect which not only leads to a detailed decomposition that adds up to the total differential, but which also leads to a unique decomposition and which has a very appealing interpretation.

The remainder of the chapter is structured as follows. In Section 6.2, we set up the econometric framework. In Section 6.3 we derive the detailed decomposition theoretically, whereas Section 6.4 shows how to estimate the detailed decomposition. In Section 6.5, we compare our decomposition method to the approaches of Fairlie (2005) and Yun (2004). Finally, Section 6.6 concludes the chapter.

6.2 Econometric Framework

We consider the following latent representation of a limited dependent variable model:

$$y_i^* = x_i'\beta + \varepsilon_i, \tag{6.1}$$

[1] This is the same argument why one should at all use a limited dependent variable model.

where $i = 1, \ldots, n$ indexes individuals, y^* is the latent dependent variable, x is a vector of explanatory variables associated with a coefficient vector $\beta \in \mathbb{R}^{K+1}$ and ε is a zero-mean error term. We assume that x contains a constant term in its first component and K "real" explanatory variables. We denote the observable dependent variable by y which is functionally related to y^*. For instance, in a binary choice model we would have that $y = 1(y^* > 0)$, where $1(\cdot)$ denotes the indicator function. Furthermore, we let d be a group indicator, taking a value of one if an individual belongs to a certain group and zero otherwise.

We make the following assumptions:

ASSUMPTION 1: $(y_i, x_i, d_i), i = 1, \ldots, n$, *are i.i.d. observations.*

ASSUMPTION 2: $E[y_i|x_i, d_i] = G(x_i'\beta, \psi), \quad \forall i = 1, \ldots, n$ *almost surely, where* $G : \mathbb{R} \times \Psi \to \mathbb{R}$ *is a known (link) function which (a) is differentiable, (b) depends on* x *only through* $x'\beta$*;* $\psi \in \Psi$ *denotes a vector of nuisance parameters.*

We need these assumptions for deriving our proposed detailed decomposition in the next section. Note that Assumption 2 covers some well-known limited dependent variable models such as probit, logit and tobit. For these three models, we have the following link functions:

- Probit: $G(x_i'\beta, \sigma) = \Phi(x_i'\beta/\sigma)$;
- Logit: $G(x_i'\beta, \sigma) = \Lambda(x_i'\beta/\sigma) = \frac{\exp\{x_i'\beta/\sigma\}}{1+\exp\{x_i'\beta/\sigma\}}$;
- Tobit with truncation from the left at zero: $G(x_i'\beta, \sigma) = \Phi(x_i'\beta/\sigma)\left(x_i'\beta + \sigma\frac{\phi(x_i'\beta/\sigma)}{\Phi(x_i'\beta/\sigma)}\right)$,

where $\sigma = \sqrt{E[\varepsilon_i^2|x_i]}, \quad \forall i = 1, \ldots, n$; $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cumulative distribution function and density function, respectively.

Note that Assumption 2 imposes the same conditional expectation $G(x_i'\beta, \psi)$ for all individuals, i.e. irrespective of whether d is equal to one or zero. Since we are only concerned with a detailed decomposition of the explained differential, we can ignore issues such as group-dependent parameters or other group-dependent model structures.

When reviewing the Oaxaca-Blinder decomposition for the linear model in the context of discrimination, Oaxaca and Ransom (1994) suggest that the explained differential should be evaluated not at the coefficient vector of one particular group, but at the coefficient vector in the absence of discrimination. We generalize this point of view and consider the framework in equation (6.1) and Assumption 1 and 2 to represent a model structure in the absence of discrimination.

This point will also be important for the economic interpretation of our proposed detailed decomposition, which will be derived in the next section.

6.3 Derivation of the Detailed Decomposition

In this section we derive (and define) the detailed decomposition. Consider two individuals i and j, where i belongs to the group with $d = 1$ and j to the group with $d = 0$, respectively. We begin with a formal notation of the total explained differential, which we define as

$$\Delta = E[y_i|d_i = 1] - E[y_j|d_j = 0]. \tag{6.2}$$

This definition has also been proposed by Fortin et al. (2011, p. 52). The explained differential is thus given by the expected difference in the outcomes of each group. By the law of iterated expectations and Assumption 2, it follows that

$$E[y_i|d_i = 1] - E[y_j|d_j = 0] = E[G(x_i'\beta, \psi)|d_i = 1] - E[G(x_j'\beta, \psi)|d_j = 0]. \tag{6.3}$$

Since observations are i.i.d. (due to Assumption 1), we can write

$$E[G(x_i'\beta, \psi)|d_i = 1] - E[G(x_j'\beta, \psi)|d_j = 0] \tag{6.4}$$

$$= E[G(x_i'\beta, \psi)|d_i = 1, d_j = 0] - E[G(x_j'\beta, \psi)|d_i = 1, d_j = 0] \tag{6.5}$$

$$= E[G(x_i'\beta, \psi) - G(x_j'\beta, \psi)|d_i = 1, d_j = 0]. \tag{6.6}$$

In order to obtain a detailed decomposition of the explained differential, we linearize the term in the expectations operator by applying the mean value theorem. This yields

$$E[G(x_i'\beta, \psi) - G(x_j'\beta, \psi)|d_i = 1, d_j = 0] \tag{6.7}$$

$$= E[g((x_{ij}^*)'\beta, \psi)(x_i - x_j)'\beta|d_i = 1, d_j = 0], \tag{6.8}$$

where $g(u, \psi) = \partial G(u, \psi)/\partial u$ and $(x_{ij}^*)'\beta$ is a scalar lying on the line segment joining $x_i'\beta$ and $x_j'\beta$. Note that x_{ij}^* can also be represented as

$$x_{ij}^* = \lambda x_i + (1 - \lambda)x_j \tag{6.9}$$

for some $\lambda \in (0, 1)$.

Due to the linearization, we define the contribution of each variable to the explained differential as follows:

DEFINITION 1: ***Detailed Decomposition***. *The contribution of a variable x_k to the explained differential is given by $c_k = E[g((x_{ij}^*)'\beta, \psi)\beta_k(x_{i,k} - x_{j,k})|d_i = 1, d_j = 0]$, $\forall k = 1, \ldots, K$.*

Note that the mean value theorem guarantees that the contributions of the variables add up to the total explained differential. Furthermore, note that Definition 1 implies that the contribution of each variable is given by the difference in explanatory variables multiplied with the marginal effect of this variable. Hence, as suggested by Fortin et al. (2011) our decomposition approach evaluates differences in variables between two groups at the marginal effects of these variable, thus taking into account the nonlinearity of the underlying model.

But we have a specific marginal effect. In general, marginal effects could be evaluated at any value of the explanatory variables. However, it makes an intuitive sense that we choose the marginal effect evaluated at x_{ij}^* in order to define the detailed decomposition. To see this, note that, by equation (6.9), the marginal effect in Definition 1 is based on a convex combination of the explanatory variables of two individuals which belong to different groups. Suppose for the moment that one group consists of males and the other one of females. The convex combination may be interpreted so as to represent a *synthetic individual*, so that the marginal effect in Definition 1 is the marginal effect of a synthetic individual which is a combination of the male and female individual. Now suppose that our synthetic individual is initially endowed like the female individual (j). Then, after receiving the difference $x_i - x_j$, the marginal effect implies a change of the synthetic individual from the female (j) to the male individual (i) in terms of the value of the link function G. Given that our model represents a situation in the absence of discrimination, this marginal effect can thus be interpreted as the *marginal effect in the absence of discrimination*. This is a generalization of the suggestion by Oaxaca and Ransom (1994) that the explained differential in linear models should be evaluated at a coefficient vector which would be prevalent in the absence of discrimination.

6.4 Estimation of the Detailed Decomposition

The detailed decomposition proposed in Definition 1 is, of course, a theoretical one and represents a population concept (due to the expectations operator). In this section we show how the detailed decomposition can be estimated. Furthermore, we prove consistency and asymptotic normality of our proposed estimator. The latter allows to obtain standard errors for the decomposition results.

Let $\mathcal{D} = \{i : d_i = 1\}$ denote the set of individuals belonging to the group with $d = 1$ and $m = \sum_{i=1}^{n} 1(d_i = 1)$ be the corresponding number of group members. We propose

to estimate $c_k = E[g((x_{ij}^*)'\beta, \psi)\beta_k(x_{i,k} - x_{j,k})|d_i = 1, d_j = 0]$ by

$$\hat{c}_k = \frac{1}{m(n-m)} \sum_{i \in \mathcal{D}} \sum_{j \notin \mathcal{D}} g((x_{ij}^*)'\hat{\beta}, \hat{\psi})\hat{\beta}_k(x_{i,k} - x_{j,k}). \tag{6.10}$$

Thus, we take all possible pairs between members of both groups, compute the detailed decomposition as in Definition 1 for each pair and then average over these pair-specific decompositions to obtain an approximation to the theoretical expectation in Definition 1. Note that estimates $\hat{\theta} = (\hat{\beta}', \hat{\psi}')'$ of $\theta = (\beta', \psi')'$ enter this expression. For the link functions listed in Section 6.2, estimates could be obtained by using the probit, logit or tobit model; estimation routines for these models are contained in any standard statistical software package. Furthermore, note that the estimator $\hat{c}_k$ contains x_{ij}^* which follows from the mean value theorem. However, it is not necessary to calculate x_{ij}^* explicitly. It suffices to calculate $g((x_{ij}^*)'\hat{\beta}, \hat{\psi})$ by

$$g((x_{ij}^*)'\hat{\beta}, \hat{\psi}) = \frac{G(x_i'\hat{\beta}, \hat{\psi}) - G(x_j'\hat{\beta}, \hat{\psi})}{(x_i - x_j)'\hat{\beta}}. \tag{6.11}$$

Hence, in practice it is not complicated to calculate $\hat{c}_k$ for each explanatory variable.

Instead of estimating θ and $(c_1, \ldots, c_K)$ separately, we propose to estimate these parameters simultaneously in a generalized method of moments (GMM) framework. Let $\alpha = (\theta', c_1, \ldots, c_K)'$ denote the parameter vector to be estimated whose true value is denoted by α_0, and $\mathcal{A}$ is the parameter space. Furthermore, suppose that $\hat{\theta}$ is obtained from solving

$$\frac{1}{n} \sum_{i=1}^{n} \tau(y_i, x_i; \hat{\theta}) = 0. \tag{6.12}$$

This equation may come from the first order condition of a maximum likelihood or minimum distance estimation approach or from the empirical counterpart of a population

moment restriction. Define $\delta_n = m/n$ and observe that

$$\frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=1}^{n-m} \{\delta_n \tau(y_i, x_i; \hat{\theta}) + (1-\delta_n)\tau(y_j, x_j; \hat{\theta})\} \tag{6.13}$$

$$= \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=1}^{n-m} \delta_n \tau(y_i, x_i; \hat{\theta}) + \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=1}^{n-m} (1-\delta_n)\tau(y_j, x_j; \hat{\theta}) \tag{6.14}$$

$$= \delta_n^{-1} \frac{1}{n} \sum_{i=1}^{m} \delta_n \tau(y_i, x_i; \hat{\theta}) + (1-\delta_n)^{-1} \sum_{j=1}^{n-m} (1-\delta_n)\tau(y_j, x_j; \hat{\theta}) \tag{6.15}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \tau(y_i, x_i; \hat{\theta}). \tag{6.16}$$

Let

$$h(y_i, y_j, x_i, x_j; \alpha) = \begin{bmatrix} \delta_n \tau(y_i, x_i; \theta) + (1-\delta_n)\tau(y_j, x_j; \theta) \\ c_1 - g((x_{ij}^*)'\beta, \psi)\beta_1(x_{i1} - x_{j1}) \\ \vdots \\ c_K - g((x_{ij}^*)'\beta, \psi)\beta_K(x_{iK} - x_{jK}) \end{bmatrix}. \tag{6.17}$$

Then, the GMM estimator $\hat{\alpha}$ of α_0 is defined by

$$\frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=1}^{n-m} h(z_i, z_j, \hat{\alpha}) = 0, \tag{6.18}$$

where $z_t = (y_t, x_t),\ t \in \{i, j\}$.

We make a set of assumptions which are summarized in Assumption 3:

Assumption 3:

(a) $\alpha_0 \in int[\mathcal{A}]$, *where* $\mathcal{A}$ *is a compact set.*

(b) $E[\tau(y, x; \theta)] = 0$ *only at* $\theta = \theta_0$.

(c) $\tau(y, x; \theta)$ *is continuously differentiable with respect to* θ.

(d) $G(x'\beta, \psi)$ *is twice continuously differentiable with respect to its arguments.*

(e) $E[h(z_i, z_j; \alpha_0)h(z_i, z_j; \alpha_0)']$ *exists and is positive definite.*

(f) $E[\frac{\partial h(z_i, z_j; \alpha_0)}{\partial \alpha'}]$ *exists and is positive definite.*

These assumptions are technical and needed to prove the following theorem, which provides consistency and asymptotic normality results for the estimator $\hat{\alpha}$ of α_0:

THEOREM 1: *Suppose that Assumptions 1-3 hold and that* $\delta_n \to \delta$. *Then,*

(a) $\hat{\alpha} \xrightarrow{p} \alpha_0$.

(b) $\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{d} \mathcal{N}(0, M^{-1}CM^{-1})$,

where M *and* C *are defined in the Appendix.*

Proof. See the Appendix. ■

Statistical software packages like STATA supply a user-friendly GMM estimation procedure. The researcher has to specify the moment equations $h(z_i, z_j; \alpha)$, and the software will produce the estimates. Hence, it is not complicated to implement our estimation procedure in practice. To obtain standard errors, a bootstrap procedure may be applied, since obtaining the sample analogue of the asymptotic covariance matrix from part b) of Theorem 1 may be difficult in practice.

6.5 Comparison to Existing Decomposition Methods

In this section, we compare our proposed decomposition method to competing approaches proposed by Fairlie (1999, 2005) and Yun (2004). Both authors derive their detailed decompositions in a finite-sample-context and do not provide population considerations (as expressed by the expectations operator on Defintion 1 above). We briefly discuss their methods and show that our proposed detailed decomposition overcomes the main drawbacks of these approaches.

We begin with Fairlie (1999, 2005) who proposes what is called a *sequential decomposition*. Fairlie analyzes a detailed decomposition for binary choice models, i.e., probit and logit models. For simplicity, we consider the case of only two explanatory variables ($K = 2$). For notational ease, we index individuals with $d = 1$ by w and individuals with $d = 0$ by b. Fairlie's procedure works as follows:

1. Reduce the size of the larger group (by randomly selecting individuals) so that both groups have the same size l.

2. Rank observations by their predicted probability that y is equal to one (i.e., by $G(x'\beta, \psi)$) *within* each group.

3. Match observations from both groups which have the same rank.

4. Let $x_{v,i,j}$ denote the value of variable x_j for an individual i from group $v \in \{w, b\}$. Fairlie's sequential decomposition would then be given by

$$\Delta \approx \frac{1}{l}\sum_{i=1}^{l}\{G(\beta_0 + x_{w,i,1}\beta_1 + x_{w,i,2}\beta_2, \psi) - G(\beta_0 + x_{b,i,1}\beta_1 + x_{b,i,2}\beta_2, \psi)\} \tag{6.19}$$

$$= \underbrace{\frac{1}{l}\sum_{i=1}^{l}\{G(\beta_0 + x_{w,i,1}\beta_1 + x_{w,i,2}\beta_2, \psi) - G(\beta_0 + x_{b,i,1}\beta_1 + x_{w,i,2}\beta_2, \psi)\}}_{\hat{c}_1} \tag{6.20}$$

$$+ \underbrace{\frac{1}{l}\sum_{i=1}^{l}\{G(\beta_0 + x_{b,i,1}\beta_1 + x_{w,i,2}\beta_2, \psi) - G(\beta_0 + x_{b,i,1}\beta_1 + x_{b,i,2}\beta_2, \psi)\}}_{\hat{c}_2},$$

(6.21)

where the index i runs over the matched observations from both groups.

Hence, the contribution of a variable is given by the average change of the link function G if the variable of interest is changed while holding all other variables constant. Note that the decomposition ensures that the sum of the contributions of the explanatory variables is equal to the individual gap in the values of the link function

between the matched individuals. A disadvantage is, however, that the contributions of each variable depend on the ordering of variables. If the order of variables during the decomposition is interchanged, different decomposition results will be obtained, so that the decomposition is not unique. This problem, which is known as *path-dependency* (cf. Fortin et al., 2011, p. 27), is a drawback of any sequential decomposition. On the contrary, our approach derived in Section 6.3 is not a sequential one, implying that our decomposition results are unique in the sense that they do not depend on the ordering of variables.

However, our decomposition results are not only unique in this sense. As mentioned before, Fairlie's methodology is based on a matching procedure for individuals from both groups. However, the matching procedure is arbitrary and lacks a theoretical foundation. Our approach, on the other hand, is theoretically founded and uses all between-group-pairs (recall equation (6.10)) of individuals, thus avoiding an arbitrary matching procedure.

The decomposition approach proposed by Yun (2004)[2] is based on two linearizations to bring the model back to the linear case, where a detailed decomposition is straightforward. His decomposition is given by

$$\Delta \approx \frac{1}{m}\sum_{i\in\mathcal{D}} G(x_i'\beta,\psi) - \frac{1}{n-m}\sum_{i\notin\mathcal{D}} G(x_i'\beta,\psi) \tag{6.22}$$

$$= G(\bar{x}_w'\beta,\psi) - G(\bar{x}_b'\beta,\psi) + R_M \tag{6.23}$$

$$= (\bar{x}_w - \bar{x}_b)'\beta g(\bar{x}_w'\beta,\psi) + R_M + R_T, \tag{6.24}$$

where $\bar{x}_w = \frac{1}{m}\sum_{i\in\mathcal{D}} x_i$ and $\bar{x}_b = \frac{1}{n-m}\sum_{i\notin\mathcal{D}} x_i$; R_M and R_T denote appropriate remain-

[2]For the explained differential, Even and Macpherson (1990, 1993) used the same decomposition methodology as derived by Yun (2004) in order to explain the decline of unionism in the United States. However, they just stated the decomposition method without providing a formal derivation.

der terms. The contribution of a variable x_k is given by

$$\hat{c}_k = \frac{(\bar{x}_{w,k} - \bar{x}_{b,k})\beta_k g(\bar{x}'_w\beta, \psi)}{(\bar{x}_w - \bar{x}_b)'\beta g(\bar{x}'_w\beta, \psi)} \left\{ \frac{1}{m} \sum_{i \in \mathcal{D}} G(x'_i\,\beta, \psi) - \frac{1}{n-m} \sum_{i \notin \mathcal{D}} G(x'_i\,\beta, \psi) \right\}. \quad (6.25)$$

Note that Yun develops weights based on equation (6.24) which are given by

$$\frac{(\bar{x}_{w,k} - \bar{x}_{b,k})\beta_k g(\bar{x}'_w\beta, \psi)}{(\bar{x}_w - \bar{x}_b)'\beta g(\bar{x}'_w\beta, \psi)}. \quad (6.26)$$

These weights are then multiplied with the observed differential $\frac{1}{m}\sum_{i \in \mathcal{D}} G(x'_i\,\beta, \psi) - \frac{1}{n-m}\sum_{i \notin \mathcal{D}} G(x'_i\,\beta, \psi)$ in order to yield the contribution of a variable x_k.

However, note that the weights in equation (6.26) reduce to

$$\frac{(\bar{x}_{w,k} - \bar{x}_{b,k})\beta_k}{(\bar{x}_w - \bar{x}_b)'\beta}, \quad (6.27)$$

so that we have the same weights as in decompositions for linear models since the nonlinear component $g(\bar{x}'_w\beta, \psi)$ cancels out. Put differently, the Yun procedure ignores the nonlinear model structure. As mentioned in the introduction, this may be problematic if the outcome differential is located in the tails of the distribution or in case of large differences in the explanatory variables (see Fortin et al., 2011, p. 52).

We illustrate this point by means of a small numerical example. Our model is given by

$$y_i^* = -6 + x_1 + x_2 + \varepsilon_i \quad (6.28)$$

$$y_i = 1(y_i^* > 0) \quad (6.29)$$

$$\varepsilon_i \sim \mathcal{N}(0, 1). \quad (6.30)$$

Hence, we consider a probit model with a link function given by $G(x'_i\beta) = \Phi(x'_i\beta)$, where $\Phi(\cdot)$ is again the standard normal cumulative distribution function. We set $n = 2,000$

with 1,000 individuals belonging to each group. Moreover, for the group with $d = 1$ we specified that $x_1 \sim \mathcal{N}(6, 1)$ and $x_2 \sim \mathcal{N}(3, 1)$. For the group with $d = 0$ we took $x_1 \sim \mathcal{N}(2, 1)$ and $x_2 \sim \mathcal{N}(2, 1)$. Hence, the group with $d = 1$ has larger mean values for both explanatory variables, in particular with respect to the variable x_1.

We simulated this model with 1,000 replications and performed our proposed decomposition and Yun's method. We then averaged over the 1,000 replications in order to obtain results. Over these replications, the (averaged) mean of our dependent variable y is 0.959 for the group with $d = 1$ and 0.124 for the group with $d = 0$, so that the averaged differential in the outcome variable is given by 0.835. The averaged decomposition results are given in the following table:

	$\hat{c}_1$	$\hat{c}_2$
Our decomposition	.6912	.1432
Yun's decomposition	.6675	.1669

Hence, we see that our proposed decomposition and Yun's method yield different results. In this example, the differences are not large. Nevertheless, they indicate that the Yun method may be too rough since it does not properly account for the nonlinearity of the underlying model.

6.6 Conclusion

In this chapter we derived a detailed decomposition (of the explained differential) for limited dependent variable models. We first defined the detailed decomposition theoretically and then showed how the theoretical decomposition can be consistently estimated using sample data. We also provided (asymptotic) distribution results for obtaining standard errors for the decomposition results and demonstrated that our estimation procedure can be easily implemented in practice. Unlike existing approaches discussed in the literature to perform detailed decompositions in nonlinear econometric models,

our method leads to a unique decomposition and accounts for the nonlinearity of the model in a rather intuitive way (i.e., by using marginal effects to evaluate differences in explanatory variables). Moreover, in light of the suggestion by Oaxaca and Ransom (1994) that the explained differential should be evaluated at a parameter vector which would be prevalent in the absence of discrimination, our decomposition approach provides a natural extension of this idea to nonlinear models.

A detailed decomposition of the explained differential in a limited dependent variable model is important because it allows to relate differences in non-continuous outcome variables to differences in characteristics. For instance, one can analyze which characteristics contribute most to the differential in, say, labor force participation rates between men and women. Another field of research where our method can be applied is an analysis of the erosion of union membership over time, where the erosion can be attributed to changes in the characteristics of the workforce (see Fitzenberger et al., 2011). Hence, our method cannot only be applied to group differences at a given point in time, but it can also be used to analyze changes over time (where two points in time serve as "groups"). However, such applications are left for future research.

6.7 Appendix

Proof of Theorem 1.

A first order expansion of

$$\frac{1}{m(n-m)}\sum_{i=1}^{m}\sum_{j=1}^{n-m}h(z_i,z_j;\hat{\alpha})=0 \tag{6.31}$$

about $\hat{\alpha}=\alpha_0$ yields (by Assumption 3 (a), (c) and (d))

$$\frac{1}{m(n-m)}\sum_{i=1}^{m}\sum_{j=1}^{n-m}h(z_i,z_j;\alpha_0)+\frac{1}{m(n-m)}\sum_{i=1}^{m}\sum_{j=1}^{n-m}\frac{\partial h(z_i,z_j;\alpha_0)}{\partial\alpha'}(\hat{\alpha}-\alpha_0)+o_p(n^{-1/2})=0. \tag{6.32}$$

Hence,

$$\hat{\alpha}-\alpha_0=M_n^{-1}U_n+o_p(n^{-1/2}), \tag{6.33}$$

where

$$M_n=\frac{1}{m(n-m)}\sum_{i=1}^{m}\sum_{j=1}^{n-m}\frac{\partial h(z_i,z_j;\alpha_0)}{\partial\alpha'}, \tag{6.34}$$

$$U_n=\frac{1}{m(n-m)}\sum_{i=1}^{m}\sum_{j=1}^{n-m}h(z_i,z_j;\alpha_0). \tag{6.35}$$

By Assumption 3 (e) and (f), and a law of large numbers for 2-sample U-statistics (e.g. Serfling, 1980, p. 190-191), we can establish that $M_n \xrightarrow{p} M$, where

$$M=E\left[\frac{\partial h(z_i,z_j;\alpha_0)}{\partial\alpha'}\right]. \tag{6.36}$$

and $U_n \xrightarrow{p} 0$ (since $E[h(z_i, z_j; \alpha_0)|d_i = 1, d_j = 0] = 0$). Hence, $\hat{\alpha} \xrightarrow{p} \alpha_0$, and part (a) of Theorem 1 is established.

To prove part (b), we apply a central limit theorem for 2-sample U-statistics to $\sqrt{n}U_n$; see van der Vaart (1998, p. 166) for the univariate case and Yu et al. (2011, p. 461) for the extension to multivariate U-statistics. This yields $\sqrt{n}U_n \xrightarrow{d} \mathcal{N}(0, C)$, where

$$\begin{aligned} C =& \delta^{-1} \left\{ E[h(z_i, z_j; \alpha_0) h(z_i, \tilde{z}_j; \alpha_0)' | d_i = 1, d_j = 0, \tilde{d}_j = 0] \right\} \\ &+ (1 - \delta)^{-1} \left\{ E[h(z_i, z_j; \alpha_0) h(\tilde{z}_i, z_j; \alpha_0)' | d_i = 1, \tilde{d}_i = 1, d_j = 0] \right\} \end{aligned} \tag{6.37}$$

We thus obtain that $\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{d} \mathcal{N}(0, M^{-1} C M^{-1})$, which establishes part (b) of Theorem 1. ■

Chapter 7

Identification and Estimation of Endogenous Regressor Models When the Endogenous Regressor is Discrete

I thank Patrick Puhani for providing valuable comments.

7.1 Introduction

In this chapter, we consider identification and estimation of a linear regression model with an endogenous regressor, where we assume that the endogenous regressor is discrete. The virtue of our approach is that we do not need an additional instrumental variable for identification. Instead, identification is fully achieved through the nonlinear relationship between the discrete endogenous regressor and the remaining (exogenous) variables included into the model.

Instrumental variable techniques have been used a number of times in applied econo-

metrics in order to consistently estimate the parameters of interest in models with endogenous regressors. However, the requirements for suitable instrumental variables are quite high. In particular, the instrumental variable should be highly correlated with the endogenous regressor and it must not be correlated with the error term of the equation of interest. Finding instrumental variables which fulfill these conditions is a hard task, and sometimes virtually impossible.

Klein and Vella (2010) and Lewbel (2012) have used heteroskedasticity to identify models with endogenous regressors. By putting restrictions on the higher moments of the error terms, these authors established estimators which do not require additional instrumental variables to achieve identification. The approach presented in this chapter also uses nonlinearities for identification. However, unlike Klein and Vella (2010) and Lewbel (2012), no heteroskedasticity is needed to achieve identification. Indeed, if heteroskedasticity is absent, these approaches fail to estimate the parameters of interest, whereas ours still works.

We proceed as follows. In the next section, our identification strategy is presented along with the underlying assumptions. In Section 7.3, we provide a small Monte Carlo study to analyze the properties of our estimator. Section 7.4 contains an empirical application to the returns to education. Finally, Section 7.5 concludes the chapter.

7.2 Identification and Estimation

We consider the following linear regression model:

$$y_{1i} = x_i'\beta + \gamma y_{2i} + \varepsilon_i, \tag{7.1}$$

where $i = 1, \ldots, n$, y_1 is the dependent variable, x is a vector of exogenous variables, y_2 is the *discrete* endogenous regressor and ε is the error term. By the exogeneity of x we mean that ε is mean independent of x, i.e., $E[\varepsilon|x] = 0$. Note that the mean independence

assumption implies that ε is uncorrelated with any function of x, which will become crucial for the identification of γ. Furthermore, note that mean independence is a stronger assumption than uncorrelatedness of x and ε, which is usually imposed for identification of instrumental variables estimates.

The discreteness of y_2 entails that the relationship between y_2 and the exogenous variables in x is intrinsically nonlinear. To see this, suppose that y_2 is generated from some underlying latent model:

$$y_{2i}^* = x_i'\delta + u_i \tag{7.2}$$

$$y_{2i} = h(y_{2i}^*), \tag{7.3}$$

where y_2^* is a (continuous) latent variable which is related linearly to the explanatory variables in x, $h(\cdot)$ is a function which transforms the latent variable into the discrete variable and u is the error term. For instance, if y_2 is binary, then $h(y_2^*) = 1(y_2^* > 0)$, where $1(\cdot)$ is the indicator function. Moreover, we can write

$$y_{2i} = E[y_{2_i}|x_i] + v_i, \tag{7.4}$$

where $v_i \equiv y_{2_i} - E[y_{2_i}|x_i]$. Since y_2 is discrete, the conditional expectation $E[y_{2_i}|x_i]$ will typically be nonlinear. For instance, if we assume that y_2 is binary and $u \sim \mathcal{N}(0,1)$, we have that

$$y_{2i} = \Phi(x_i'\delta) + v_i, \tag{7.5}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

In a typical instrumental variables or two stage least squares estimation procedure one would obtain the linear projection of y_2 on x and an instrumental variable and use the projected values as an instrument for y_2. In that case, identification of γ would

be achieved solely through the instrumental variable which guarantees that the linear projection and the variables in x do not exhibit perfect multicollinearity.

In our identification approach we use a nonlinear projection. In the probit example above, we would just estimate a probit model for y_2 using x as the explanatory variables. The (estimated) predicted values are a valid instrumental variable for y_2 since they are correlated with y_2 but uncorrelated with ε (provided the mean independence assumption holds). Moreover, the nonlinear projection guarantees that our instrumental variable is not perfectly correlated with the variables in x, which implies that γ is identified. Hence, identification is achieved through the nonlinearity of the nonlinear projection of y_2 on x.[1]

The reason why our approach works is that the nonlinear relationship between y_2 and x is intrinsically present through the discreteness of y_2. Our approach could also be interpreted as an ordinary instrumental variable approach where nonlinear transformations of x are used as instrumental variables. Of course, using a correct nonlinear form (instead of using various transformations of x as instrumental variables) results in efficiency gains.

Finally, we make some cautionary remarks. Our proposed procedure only works well in finite samples if the nonlinearity between y_2 and x is substantial. Put differently, if the relationship between y_2 and x is close to being linear, γ will not be properly identified (in the sense that standard errors will become unbounded). Furthermore, x may not include too many nonlinear terms, because in that case the nonlinear projection of y_2 on x and the variables in x may be perfectly correlated as well. In any case, a sufficiently large sample size is needed to get reliable results. The Monte Carlo results in the next section illustrate this point.

[1]Even more nonlinearities can be exploited for identification when conditional heteroskedasticity is present in equation (7.2). The proposed instrumental variable would then be a nonlinear transformation of the exogenous variables due to the discreteness of the endogenous variable and due to the heteroskedasticity. This may be useful, since, as shown below, the higher the degree of nonlinearity the better performs our proposed estimator.

7.3 Monte Carlo Evidence

The goal of our small Monte Carlo study is to illustrate how our proposed estimator performs depending on (a) the degree of nonlinearity of the nonlinear projection of y_2 on x and (b) the sample size. Concerning the former, we expect that the higher the degree of nonlinearity the better our model will be identified, and thus the better our estimator should perform. Regarding the latter, we expect that a fairly large sample size is needed to obtain precise estimates of γ.

We consider the following model:

$$y_{1i} = 1 + x_i + y_{2i} + \varepsilon_i \tag{7.6}$$

$$y_{2i}^* = \alpha + x_i + u_i \tag{7.7}$$

$$y_{2i} = 1(y_{2i}^* > 0) \tag{7.8}$$

$$\varepsilon_i = u_i + v_i, \tag{7.9}$$

where $x \sim \mathcal{N}(1,1)$, $u \sim \mathcal{N}(0,1)$ and $v \sim \mathcal{N}(0,1)$. The object of interest is the parameter γ associated with the endogenous explanatory variable y_2, whose true value is set equal to one. The parameter α is varied in order to obtain different degrees of nonlinearity. To see this, note that the model corresponds to the probit example from the last section. For different values of α we obtain different means of y_2. As it is well known from the properties of the standard normal cumulative distribution function, the link between y_2 and the exogenous variables exhibits the highest degree of nonlinearity if the mean of y_2 is either very small or very large. We choose $\alpha \in \{-3, -1, 1\}$, which corresponds to means of y_2 given by approximately 0.08, 0.5 and 0.92, respectively. Furthermore, we consider sample sizes of $n \in \{250, 500, 1000, 2000, 5000, 10000\}$.

As noted in the last section, we used the following procedure to obtain estimates of γ. We estimated a probit model for y_2 using x and a constant as explanatory variables.

The predicted values $\hat{y}_2$ were then used as instrumental variables for y_2 in the framework of a two stage least squares regression.

Our Monte Carlo results are based upon 1,000 replications. For each sample size and for each value of α, we computed the mean of the estimates of γ, the standard deviation as well es the root mean squared error (RMSE). Moreover, we computed the mean of the F statistics from the "first stage" of an instrumental variable estimation approach. In our case, the F statistic is the test statistic associated with the hypothesis $H_0 : \psi_2 = 0$ in a regression

$$y_{2_i} = \psi_0 + \psi_1 x_i + \psi_2 \hat{y}_{2i} + \eta_i. \tag{7.10}$$

The Monte Carlo results are presented in Table 7.1. We see that a sample size of $n = 1,000$ is sufficient to obtain estimates whose mean is close to the true value of one, irrespective of the value of α. Apart from a sample size of $n = 250$, we also see that the root mean squared error of our estimator is smaller for $\alpha = -3$ or $\alpha = 1$ compared to $\alpha = -1$, which provides evidence for our hypothesis that the higher the nonlinearity the better the identification of γ, and thus the better the estimator performance. Regarding our second hypothesis, we see that the sample size has to be quite large in order to obtain precise estimates. This is of great practical importance, since if ordinary least squares (OLS) and IV estimates are quite close one needs large sample sizes to decide whether OLS and IV estimates are actually different from each other.

Considering the averaged F statistics, we see that these are larger for higher degrees of nonlinearity (i.e., $\alpha = -3$ or $\alpha = 1$). However, there is no clear link between the F statistic and the estimator performance as measured by the RMSE. For instance, when $\alpha = -1$ an F statistic of about 225 is associated with a RMSE of about 0.23. On the other hand, when $\alpha = -3$ or $\alpha = 1$, an F statistic of about 225 is associated with a RMSE of about 0.41, which is nearly twice as much! However, for a given degree of

nonlinearity we still have (as it should be) that the larger the F statistic the better the estimator performance.

7.4 Empirical Application

In this section, we consider an application of our estimation strategy to the measurement of the returns to college education. We use data from the 2000 U.S. Census (Ruggles et al., 2010). This provides us with a very large sample size, which is needed to obtain precise results with regard to the results of the last section. Our sample consists of white men not living in group quarters and between 25 and 54 years of age, and who are not self-employed. Moreover, we consider only full time full year (FTFY) workers. We are interested in results for the main part of the wage distribution, hence we keep only observations whose wage is located between the 5th and 95th percentile of the wage distribution.[2]

We estimate the following model: Our dependent variable is the natural logarithm of the hourly wage. Explanatory variables are age (*age*), age squared (*age2*), dummy variables for the Census region (*northeast, midwest, south*), a dummy variable for the marital status (*married*), and a dummy variable for college education (*college*). The *college* indicator is equal to one if the years of education are larger than twelve. Summary statistics for all variables are given in Table 7.2.

Since college education is potentially endogenous, we apply the estimation methodology proposed in Section 7.2 to obtain a consistent estimate of the coefficient of college education, which may also be called the "returns to college education". For comparison, we also present the OLS estimates in the first panel of Table 7.3.

College education is a binary variable, hence we could estimate the relationship between college education and the exogenous variables by means of a probit model,

[2]Since our main equation is linear in the coefficients, we expect that linearity is more likely to hold in the main part of the wage distribution (as opposed to the tails).

as outlined in the last two sections. However, this "first stage" relationship may be characterized by even more nonlinearities (such as conditional heteroskedasticity). To obtain a nonlinear projection of college education on the exogenous variables, we thus use a very flexible modeling device: We simply create cells for each combination of our exogenous variables, and then compute the cell means with respect to college education. This is equivalent to fitting a saturated model for the regression of college education on the exogenous variables. Hence, we model the "first stage" of our estimation procedure fully nonparametrically. The cell means are then used as an instrumental variable for education.

A two stage least squares (2SLS) procedure yields the results given in the second panel of Table 7.3. We see that the coefficient of education is significantly larger when estimated by 2SLS. This suggests that OLS underestimates the returns to education. Since the value of the F statistic is very large, we may have some confidence that the 2SLS procedure indeed gives a correct (and precise) estimate of the returns to college education.

7.5 Conclusions

In this chapter, we have provided an identification strategy for an endogenous regressor model in case that the endogenous regressor is discrete. Our identification and estimation strategy may prove useful in situations where "classical" instrumental variables are hard to find. Moreover, our method may serve as a robustness check to compare estimates obtained under our strategy with estimates which have been obtained using "classical" instrumental variables.

While in settings with "classical" instrumental variables the exogeneity assumption of the instrument may be doubtful, our strategy crucially hinges on the degree of nonlinearity associated with the discreteness of the endogenous regressor. If the nonlinearity

is substantial and the sample size is sufficiently large, our approach may provide a valuable estimation strategy if a "classical" instrumental variable is unavailable.

7.6 Tables

Table 7.1: Monte Carlo results, 1,000 replications

$\alpha = -3$				
n	Mean	Std.dev.	RMSE	Mean F
250	0.6709	2.7117	2.7316	31.4890
500	0.9449	0.9778	0.9793	57.9315
1,000	0.9825	0.6020	0.6022	113.5824
2,000	0.9943	0.4148	0.4148	223.8416
5,000	1.0005	0.2594	0.2594	554.4310
10,000	0.9961	0.1847	0.1847	1104.5113
$\alpha = -1$				
n	Mean	Std.dev.	RMSE	Mean F
250	0.8256	2.0697	2.0771	6.4913
500	0.9096	1.1765	1.1800	11.8646
1,000	0.9741	0.7817	0.7821	23.5757
2,000	0.9954	0.5325	0.5325	45.7948
5,000	0.9946	0.3299	0.3299	113.2809
10,000	1.0003	0.2257	0.2257	225.8149
$\alpha = 1$				
n	Mean	Std.dev.	RMSE	Mean F
250	0.6628	5.7067	5.7166	31.9808
500	0.9012	0.9434	0.9486	59.9713
1,000	0.9416	0.5997	0.6025	113.1646
2,000	0.9720	0.4155	0.4164	224.1928
5,000	0.9880	0.2570	0.2572	555.3603
10,000	0.9952	0.1838	0.1839	1104.2445

Source: 2000 U.S. Census data; own calculations.

Table 7.2: Summary Statistics

Variable	Mean	Std. Dev.
log hourly wage	2.87	0.42
age	39.42	8.10
northeast	0.19	0.39
midwest	0.28	0.45
south	0.34	0.47
married	0.72	0.45
college	0.59	0.49

Source: 2000 U.S. Census data; own calculations.

Table 7.3: Estimation results

	OLS		2SLS	
Variable	Coefficient	(Std. Error)	Coefficient	(Std. Error)
college	0.2549	(0.0008)	0.4470	(0.0128)
age	0.0582	(0.0005)	0.0630	(0.0006)
age2	-0.0006	(0.0000)	-0.0007	(0.0000)
northeast	0.0239	(0.0013)	0.0382	(0.0017)
midwest	-0.0368	(0.0012)	-0.0203	(0.0017)
south	-0.0776	(0.0011)	-0.0617	(0.0017)
married	0.1134	(0.0009)	0.1111	(0.0010)
constant	1.3348	(0.0095)	1.1195	(0.0178)
n	961,224			
F statistic			3,861.45	

Note: Estimation results are based on Census weights and robust standard errors.

Source: 2000 U.S. Census data; own calculations.

References

Ackerberg, D., Chen, X., and Hahn, J. (2012). A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics*, 94: 481-498.

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

Ai, C. and Chen, X. (1999). Efficient sieve minimum distance estimation of semiparametric conditional moment models. Manuscript. London School of Economics.

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

Angrist, J.D. and Krueger, A.B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106: 979-1014.

Bauer, T. and Sinning, M. (2008). An extension of the Blinder-Oaxaca decomposition to nonlinear models. *AStA Advances in Statistical Analysis*, 92: 197-206.

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

Blinder, A.S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8: 436-455.

Blundell, R., Duncan, A., and Meghir, C. (1998). Estimating Labor Supply Responses Using Tax Reforms. *Econometrica*, 66: 827-861.

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

Bound, J.B., Jaeger, D.A., and Baker, R.M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, 90: 443-450.

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

Card, D. (1999). The causal effect of education on earnings. In Ashenfelter, O. and Card, D. (eds.) *Handbook of Labor Economics*, volume 3, chapter 30, pp. 1801-1863. Elsevier, Amsterdam.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In Heckman, J. and Leamer, E. (eds.) *Handbook of Econometrics*, volume 6, chapter 76, pp. 5549-5632. Elsevier, Amsterdam.

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

Chen, X., Fan, Y., and Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association*, 101: 1228-1240.

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

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

Davidson, R. and MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, New York, NY.

Demarta, S. and McNeil, A.J. (2005). The t copula and related copulas. *International Statistical Review*, 73: 111-129.

Dustmann, C. and Rochina-Barrachina, M.E. (2007). Selection correction in panel data models: An application to the estimation of females' wage equations. *Econometrics Journal*, 10: 263-293.

Even, W.E. and Macpherson, D.A. (1990). Plant size and the decline of unionism. *Economics Letters*, 32: 393-398.

Even, W.E. and Macpherson, D.A. (1993). The decline of private-sector unionism and the gender wage gap. *Journal of Human Resources*, 28: 279-296.

Fairlie, R.W. (1999). The absence of the african-american owned business: An analysis of the dynamics of self-employment. *Journal of Labor Economics*, 17: 80-108.

Fairlie, R.W. (2005). An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. *Journal of Economic and Social Measurement*, 30: 305-316.

Fitzenberger, B., Kohn, K., and Wang, Q. (2011). The erosion of union membership in Germany: determinants, densities, decompositions. *Journal of Population Economics*, 24: 141-165.

Fortin, N., Lemieux, T., and Firpo, S. (2011). Decomposition Methods in Economics. In: Ashenfelter, O. and Card, D. (eds.) *Handbook of Labor Economics*, volume 4A, chapter 1, pp. 1-102. Elsevier, Amsterdam.

Gallant, A.R. and Nychka, D.W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55: 363-390.

Genius, M. and Strazzera, E. (2008). Applying the copula approach to sample selection modelling. *Applied Economics*, 40: 1443-1455.

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

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

Heckman, J.J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, 46: 931-959.

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

Heckman, J.J. and MaCurdy, T.E. (1986). Labor Econometrics. In: Griliches, Z. and Intriligator, M.D. (eds.) *Handbook of Econometrics*, volume 3, chapter 32, pp. 1917-1977. Elsevier, Amsterdam.

Heckman, J.J. and Sedlacek, G.L. (1990). Self-selection and the distribution of hourly wages. *Journal of Labor Economics*, 8: S329-S363.

Heckman, J.J., Tobias, J.L., and Vytlacil, E. (2003). Simple estimators for treatment parameters in a latent-variable framework. *Review of Economics and Statistics*, 85: 748-755.

Horowitz, J.L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60: 505-531.

Horowitz, J.L. and Härdle, W. (1994). Testing a parametric model against a semi-parametric alternative. *Econometric Theory*, 10: 821-848.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

Jones, F.L. (1983). On decomposing the wage gap: a critical comment on Blinder's method. *Journal of Human Resources*, 18: 126-130.

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

Klein, R.W. and Vella, F. (2010). Estimating a class of triangular simultaneous equations models without exclusion restrictions. *Journal of Econometrics*, 154: 154-164.

Klein, R.W., Shen, C., and Vella, F. (2011). Semiparametric Selection Models with Binary Outcomes. IZA Discussion Paper No. 6008.

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

Lee, L.F. (1983). Generalized econometric models with selectivity. *Econometrica*, 51: 507-12.

Lewbel, A. (2012). Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models. *Journal of Business and Economics Statistics*, 30: 67-80.

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

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

Manski, C.F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, 55: 357-362.

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

Martins, M.F.O. (2001). Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal. *Journal of Applied Econometrics*, 16: 23-39.

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

Mohanty, M.S. (2002). A bivariate probit approach to the determination of employment: a study of teem employment differentials in Los Angeles county. *Applied Economics*, 34: 143-156.

Mroz, T.A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55: 765-799.

Mulligan, C.B. and Rubinstein, Y. (2008). Selection, investment, and women's relative wages over time. *The Quarterly Journal of Economics*, 123: 1061-1110.

Nelsen, R.B. (2006). *An Introduction to Copulas*. Springer, New York, NY.

Newey, W.K. (1987). Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics*, 36: 231-250.

Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79: 147-168.

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

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

Oaxaca, R.L. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14: 693-709.

Oaxaca, R.L. and Ransom, M.R. (1994). On discrimination and the decomposition of wage differentials. *Journal of Econometrics*, 61: 5-21.

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

Prieger, J.E. (2002). A flexible parametric selection model for non-normal data with application to health care usage. *Journal of Applied Econometrics*, 17: 367-392.

Rivers, D. and Vuong, Q.H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39: 347-366.

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

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

Schwiebert, J. (2011). A Full Information Maximum Likelihood Approach to Estimating the Sample Selection Model with Endogenous Covariates. Discussion Paper No. 483, Department of Economics and Business Administration, Leibniz University Hannover.

Schwiebert, J. (2012a). Revisiting the Composition of the Female Workforce - A Heckman Selection Model with Endogeneity. Discussion Paper No. 502, Department of Economics and Business Administration, Leibniz University Hannover.

Schwiebert, J. (2012b). Analyzing the Composition of the Female Workforce - A Semiparametric Copula Approach. Discussion Paper No. 503, Department of Economics and Business Administration, Leibniz University Hannover.

Schwiebert, J. (2012c). Semiparametric Estimation of a Sample Selection Model in the Presence of Endogeneity. Discussion Paper No. 504, Department of Economics and Business Administration, Leibniz University Hannover.

Schwiebert, J. (2012d). Semiparametric Estimation of a Binary Choice Model with Sample Selection. Discussion Paper No. 505, Department of Economics and Business Administration, Leibniz University Hannover.

Schwiebert, J. (2012e). A Detailed Decomposition for Limited Dependent Variable Models. Discussion Paper No. 506, Department of Economics and Business Administration, Leibniz University Hannover.

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

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

Smith, R.J. and Blundell, R.W. (1986). An exogeneity test for a simultaneous equation tobit model with an application to labor supply. *Econometrica*, 54: 679-685.

Smith, M.D. (2003). Modelling sample selection using Archimedean copulas. *Econometrics Journal*, 6: 99-123.

Staiger, D. and Stock, J.H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65: 557-586.

Trivedi, P.K. and Zimmer, D.M. (2007). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, 1: 1-111.

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

Van der Vaart, A.W. (1998): *Asymptotic Statistics*. Cambridge University Press, New York, NY.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *Journal of Human Resources*, 33: 127-169.

Vuong, Q.H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57: 307-333.

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

Yu, Q., Tang, W., Kowalski, J., and Tu, X.M. (2011). Multivariate U-statistics: a tutorial with applications. *Wiley Interdisciplinary Review: Computational Statistics*, 3: 457-471.

Yun, M.-S. (2004). Decomposing differences in the first moment. *Economics Letters*, 82: 275-280.