Testing exclusion restrictions and additive separability in sample selection models

Martin Huber and Giovanni Mellace

University of St. Gallen, Dept. of Economics

Abstract: Standard sample selection models with non-randomly censored outcomes assume (i) an exclusion restriction (i.e., a variable affecting selection, but not the outcome) and (ii) additive separability of the errors in the selection process. This paper proposes tests for the joint satisfaction of these assumptions by applying the approach of Huber and Mellace (2011) (for testing instrument validity under treatment endogeneity) to the sample selection framework. We show that the exclusion restriction and additive separability imply two testable inequality constraints that come from both point identifying and bounding the outcome distribution of the subpopulation that is always selected/observed. We apply the tests to two variables for which the exclusion restriction is frequently invoked in female wage regressions: non-wife/husband's income and the number of (young) children. Considering eight empirical applications, our results suggest that the identifying assumptions are likely violated for the former variable, but cannot be refuted for the latter on statistical grounds.

Keywords: sample selection, exclusion restriction, additive separability, monotonicity, test.

JEL classification: C12, C15, C24, C26.

An earlier version of this paper was circulated under the title "Testing instrument validity in sample selection models". We have benefited from comments by Alberto Abadie, Joshua Angrist, Guido Imbens, Toru Kitagawa, Alexa Tiemann, seminar participants at Harvard (seminar in econometrics, September 2011), and an anonymous associate editor. Martin Huber gratefully acknowledges financial support from the Swiss National Science Foundation grant PBSGP1_138770. Addresses for correspondence: Martin Huber (martin.huber@unisg.ch), Giovanni Mellace (giovanni.mellace@unisg.ch), SEW, University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland.

1 Introduction

The sample selection problem as discussed in Gronau (1974) and Heckman (1974) arises when the outcome of interest is only observed for a non-randomly selected subpopulation. This is an ubiquitous phenomenon in empirical economics. E.g., when estimating the returns to schooling or training it is an issue when only a selective subgroup is employed, which is a condition for observing earnings. As a second example, several studies in development and educational economics assess the impact of school vouchers or other incentives on test scores in high school and college, see for instance Angrist, Bettinger, and Kremer (2006) and Angrist, Lang, and Oreopoulos (2009). In this context, sample selection bias is an issue because only a selective subgroup of students usually takes the test.

In the presence of sample selection, Heckman (1974, 1976, and 1979) proposed fully parametric maximum likelihood and two step estimators, assuming that the errors in the selection and outcome equations are jointly normally distributed. These assumptions are theoretically sufficient for identification by exploiting the (known) nonlinearity of the selection bias correction term, the so-called inverse Mill's ratio. In practice however, this is often tenuous due to multicollinearity problems. Many empirical applications therefore invoke exclusion restrictions, implying that there exist one or several observed variables that are related to selection, but have no direct effect on the outcome. We will refer to these variables as instruments (for selection), even though they are not be confused with instruments for endogenous regressors in instrumental variable models.

Exclusion restrictions are even more crucial in generalizations of the original sample selection model. E.g., Cosslett (1991), Gallant and Nychka (1987), Powell (1987), and Newey (2009) relax the distributional assumptions but maintain the single index structure in the selection equation and linearity in the outcome equation. In contrast, Ahn and Powell (1993) allow for a nonparametric selection equation, whereas Das, Newey, and Vella (2003) identify a fully nonparametric model with additively separable errors. Finally, Newey (2007) considers a general model with non-separable errors in the outcome equation, which comes at the cost that

interesting parameters such as partial effects are only identified in the selected subpopulation. It is worth noting that all of the mentioned sample selection models, even the most general ones, invoke additive separability of the unobserved term in the selection equation for reasons of identifiability. As shown by Vytlacil (2002), this is equivalent to assuming that the potential selection state of each individual increases or decreases (weakly) monotonically in the value of the instrument. Standard sample selection models, no matter whether parametric, semi-parametric, or non-parametric, therefore rely on two crucial restrictions: Firstly, the exclusion restriction and secondly, the monotonicity of selection in the instrument (or equivalently, the additive separability of the errors in the selection equation).

Therefore, our first contribution is the proposal of tests for the joint satisfaction of the exclusion restriction and monotonicity. To this end, we apply the approach of Huber and Mellace (2011) originally suggested for testing instrument validity under treatment endogeneity, which can be easily adapted to our sample selection framework. Assuming that the variable satisfying the exclusion restriction is binary, the intuition of testing is as follows. Under the exclusion restriction and monotonicity, the outcome distribution of the 'always selected' (those selected irrespective of the instrument) is point identified in the selected subpopulation not receiving the instrument. On the other hand, the selected subpopulation receiving the instrument contains both always selected and 'compliers' (whose selection state reacts to the instrument). Using the results of Horowitz and Manski (1995), upper and lower bounds on the distribution of the always selected can be derived in this mixed population. The exclusion restriction and monotonicity can only be satisfied if the point identified outcome distribution of the always selected in the absence of the instrument lies within the bounds in the presence of the instrument, which yields two testable constraints. While testing is not asymptotically uniformly powerful in the sense that the assumptions may be violated even if the point identified parameter is within its bounds, a rejection asymptotically implies the violation of at least one assumption.

It is worth noting that further tests have been proposed in the context of sample selection models. Blundell, Gosling, Ichimura, and Meghir (2007) and Kitagawa (2010) suggest methods to

verify the exclusion restriction when additive separability of the unobserved term in the selection equation is not assumed. Their framework is therefore more general than the one considered in this paper which is, however, predominant in the empirical literature primarily concerned with the identification of marginal effects. The test of Blundell, Gosling, Ichimura, and Meghir (2007) is based on (i) bounding the outcome distribution of the total population conditional on the instrument and (ii) verifying whether bounds crossing occurs across different values of the instrument.[1] They do not show asymptotic validity of their inferential bootstrap procedure.

Kitagawa (2010) proves that the bounds considered in Blundell, Gosling, Ichimura, and Meghir (2007) are not necessarily sharp. He provides a test relying on sharp bounds which is based on the intuition that under the exclusion restriction, the integral over the envelope of the conditional densities of the observed outcomes given the instrument must not be larger than one. As a second contribution, Kitagawa (2010) also derives a testable implication (without providing a formal test) under additive separability. Not surprisingly, the latter increases asymptotic testing power compared to assuming the exclusion restriction alone. We show in Appendix A.1 that one of our constraints is equivalent to the implication of Kitagawa (2010). Furthermore, we also provide a second testable constraint not considered therein that may increase finite sample testing power. Finally, Crépon (2006) proposes a test for the exclusion restriction at infinity. I.e., testing relies on observations that are selected with probability one, which may not be available in a particular empirical application. In contrast, neither our approach nor the one of Kitagawa (2010) requires some outcomes to be observed with certainty.

Our second contribution is an empirical application of the tests. We consider two supposed instruments frequently used in female wage regressions to control for sample selection, where selection bias comes from the labor supply decision. The first variable is non-wife income (such as husband's income or other sources of family income). Most empirical studies find that non-wife income affects female labor supply negatively, see for instance Mroz (1987) and Zabel (1993). However, the instrument is only valid if it neither exhibits direct effects on the female wage, nor is

[1] It has already been noticed by Manski (2003) that the exclusion restriction is violated if the identification region defined by the bounds is empty.

related to unobserved terms affecting wage. The latter would for instance be violated if unobserved social and economic attributes were related both with the expected wage and the likelihood to find a partner with a particular income level. In fact, Becker (1981) argues that individuals of superior productivity tend to marry one another, which is in line with the empirical finding of Nakosteen, Westerlund, and Zimmer (2004) that spouses tend to be economically similar before marriage, at least in the dimension of earnings. Applying our tests to four data sets coming from Schafgans (1998), Martins (2001), Chang (2011), and Jeffrey Wooldridge's LABSUP data set (available at http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html), the results suggest that the identifying assumptions are likely violated when using non-wife income as instrument.

The second supposed instrument is the number of (young) children in the household. The intuition is that women with young kids are less likely to provide labor due to time constraints related to child rearing. Indeed, the vast majority of empirical studies examining the connection between fertility and female labor supply find a negative correlation. However, the exclusion restriction, implying that the number of children is not directly related to wages, in not undisputed. Theoretical arguments suggest that labor supply, wages, and fertility are endogenous, see for instance the multiple equation family model proposed by Fleisher and Rhodes (1979). E.g., if women with relatively low expected future wages had on average a higher fertility, the exclusion restriction would fail. Considering four data sets previously analyzed by Martins (2001), Mulligan and Rubinstein (2008), Lee (2009), and Chang (2011), test results do however not point to a violation of the identifying assumptions when relying on young children as an instrument.

The remainder of this paper is organized as follows. Section 2 characterizes the sample selection model and its identifying assumptions. Section 3 derives the testable implications of the exclusion restriction and monotonicity/separability. Empirical applications to the estimation of female wage equations are presented in Section 4. Section 5 concludes.

2 The selection model

This section introduces the sample selection model along with the assumptions to be tested. Denote by Y a (scalar) continuous outcome variable with bounded support, by X a scalar or a vector of regressor(s), and by U an unobserved term affecting the outcome. The outcome is modeled as

$$Y = \varphi(X) + U, \tag{1}$$

where $\varphi(\cdot)$ denotes some function (with $\varphi(X) = X\beta$ if the outcome is linear in the regressors). E.g., when assessing the returns to schooling, Y is the wage, X may be education and labor market experience, and U are unobserved factors such as ability and motivation. Researchers and policy makers are typically interested in parameters like the conditional expectation $E[Y|X = x]$ or the marginal or average effect of X. However, in the presence of sample selection, Y is only observed for a non-random subpopulation, e.g., the employed (while X is assumed to be observed for the entire population). To address this problem, let $S \in \{1, 0\}$ be an observed binary selection indicator which is 1 if the outcome of some individual is observed and 0 otherwise. Furthermore, denote by W and V observed and unobserved terms affecting selection, respectively. We assume the following standard selection model with additive separability of the unobserved term:

$$S = I\{\zeta(W) + V \geq 0\}, \tag{2}$$

where $I\{\cdot\}$ denotes the indicator function and $\zeta(\cdot)$ is some function of W.

The sample selection problem arises when the unobserved terms V and U are not independent. In a general model set-up (without imposing tight parametric assumptions) a first requirement for identification is that at least one variable in W satisfies an exclusion restriction w.r.t. Y, see for example the discussion in Newey (2009). I.e., there exists one variable in W that does not belong to X (which affects the outcome). In this case, we may write $W = (X, Z)$,

where Z denotes the instrument(s) not appearing in the outcome equation. Furthermore, the exclusion restriction requires that Z is independent of the unobserved term U in the outcome equation and consequently also of the unobservable V in the selection equation, as the latter is related to U. Therefore, a commonly invoked assumption is joint independence of (X, Z) and (U, V), as for instance in Newey (2007). Including X in the assumption is necessary if we want to give a causal interpretation to the regressors. However, here we are just concerned with testing so that the following conditional version of the assumption (given X) suffices:

Assumption 1:

Z is independent of $(U, V)|X$ (exclusion restriction).

It is interesting to note that virtually all selection models concerned with point identification, including the nonparametric framework of Newey (2007), rely (besides the exclusion restriction) on the additive separability of the unobserved term in the selection model as postulated in equation (2). The reason is that even in general models, identification of causal effects of X typically relies on the index restriction $E(U|S = 1, X, Z) = E(U|S = 1, \Pr(S = 1|X, Z))$, see for instance Newey (2009), which allows using $\Pr(S = 1|X, Z)$ as a control function.[2] As discussed in Das, Newey, and Vella (2003), this index restriction is implied by assuming (2) along with joint independence of (X, Z) and (U, V) (at least conditional on $\Pr(S = 1|X, Z)$) and a strictly monotonic cdf of V. Furthermore, Vytlacil (2002) shows that assuming a selection model with additive separability is equivalent to assuming that the potential selection state of each individual increases or decreases weakly monotonically in the instrument. To translate (2) into the monotonicity assumption, we use the potential outcome notation (see for instance Rubin, 1974) and denote by $S(z)$ the potential selection state if the instrument Z was exogenously set to z: $S(z) = I\{\zeta(X, z) + V \geq 0\}$. For a binary instrument, monotonicity (given X) implies the following:

Assumption 2:

[2] In contrast, Mealli and Pacini (2008) consider identification (for binary treatment variables) when conditioning on a binary instrument directly rather than using $\Pr(S = 1|X, Z)$ as a control function. In this case, point identification is not obtained in general, but requires additional assumptions.

$\Pr(S(1) \geq S(0)|X) = 1$ (positive monotonicity) or

$\Pr(S(1) \leq S(0)|X) = 1$ (negative monotonicity).

We will subsequently only consider positive monotonicity of S in Z, as the case of negative monotonicity is symmetric. Furthermore, we will omit conditioning on X in our discussion for the sake of ease of notation. In the empirical application, both unconditional versions of Assumptions 1 and 2 (without controlling for X) as well as conditional ones (given X) will be considered. For the latter case, the reader may regard the following discussion on testing to take place within cells defined upon X.

3 Testing

In this section, we show how the joint satisfaction of Assumptions 1 and 2 can be tested by a straightforward application of the approach of Huber and Mellace (2011) to the sample selection problem. To keep the exposition simple, we derive the testable implications for a binary instrument Z. However, the results could be easily extended to multi-valued discrete instruments with bounded support in an analogous way as in Huber and Mellace (2011). Testing is based on the fact that under Assumptions 1 and 2, the outcome distribution of the subpopulation of always selected can be both point identified and bounded. The point must lie within its bounds, otherwise the assumptions are necessarily violated. To formally discuss this result, we use the principal stratification framework of Frangakis and Rubin (2002) and a similar notation as in Angrist, Imbens, and Rubin (1996) to divide the population into four types according to the reaction of selection to the instrument. The always selected are those with observed outcomes irrespective of the value of the instrument, the compliers are selected under $Z = 1$, but not under $Z = 0$, the defiers are selected under $Z = 0$, but not under $Z = 1$, and the outcomes of the never selected are never observed. Table 1 displays the relationship between the types, denoted by T, and the potential selection states.

The type of a subject is not directly observed, because either $S(1)$ or $S(0)$ but never both

Table 1: Types

type (T)	potential selection states
a (always selected)	S(1)=1, S(0)=1
c (compliers)	S(1)=1, S(0)=0
d (defiers)	S(1)=0, S(0)=1
n (never selected)	S(1)=0, S(0)=0

Table 2: Observed subgroups and types

observed values Z, S	possible types	Y observed
$Z = 1, \quad S = 1$	either a or c	yes
$Z = 1, \quad S = 0$	either d or n	no
$Z = 0, \quad S = 1$	either a or d	yes
$Z = 0, \quad S = 0$	either c or n	no

are known for any individual. To see this, note that the observed values of Z and S define four observed subgroups, which all are mixtures of two types. A further complication is that outcomes are only observed for those with $S = 1$.[3] Table 2 summarizes these results. Assumptions 1 and 2 nevertheless allow point identifying the proportions of the types ($T \in \{a, c, d, n\}$), denoted by π_T. By Assumption 2, defiers do not exist because selection increases weakly monotonically in the instrument such that $\pi_d = 0$. If follows that observations with $Z = 1, S = 0$ (with $Z = 0, S = 1$) are necessarily never selected (always selected).[4] By Assumption 1, Z and V are independent so that the share of any type conditional on the instrument is equal to its unconditional proportion in the entire population. Therefore, the relationship between some observed conditional selection probability given the instrument, $\Pr(S = s|Z = z)$, henceforth denoted by $P_{s|z}$, and the latent type proportions is as shown in Table 3.

Table 3: Observed conditional probabilities and type proportions

observed cond. selection prob.	type proportions
$P_{1\|1} \equiv \Pr(S = 1\|Z = 1)$	$\pi_a + \pi_c$
$P_{0\|1} \equiv \Pr(S = 0\|Z = 1)$	π_n
$P_{1\|0} \equiv \Pr(S = 1\|Z = 0)$	π_a
$P_{0\|0} \equiv \Pr(S = 0\|Z = 0)$	$\pi_c + \pi_n$

[3]This issue does not arise in the endogenous treatment framework of Huber and Mellace (2011), where all outcomes are observed.

[4]For a similar result in the context of selection models see Lee (2009), who in contrast to this paper considers monotonicity of selection in a binary treatment.

It is easy to see that all type proportions are identified, which will be used to bound the outcome distribution of the always selected. To this end, we introduce some further notation. Denote by $f(y|Z=z,T=t)$ the pdf of $Y=y$ for a particular type $T=t$ given Z.[5] Furthermore, let $f(y|Z=z,S=s)$ denote the conditional pdf of $Y=y$ given Z and S. As outlined in Table 2, only two of possibly four conditional pdfs are observed: $f(y|Z=1,S=1)$ and $f(y|Z=0,S=1)$. Equivalent to Imbens and Rubin (1997) in the context of treatment endogeneity, it follows that the former is a mixture of the always selected and the compliers, where the mixing proportions correspond to the relative shares of types in the conditional outcome distribution.

$$f(y|Z=1,S=1)=\frac{\pi_a}{\pi_a+\pi_c}\cdot f(y|Z=1,T=a)+\frac{\pi_c}{\pi_a+\pi_c}\cdot f(y|Z=1,T=c). \tag{3}$$

Table 3 implies that the mixing proportions are identified by $\frac{\pi_a}{\pi_a+\pi_c}=\frac{P_{1|0}}{P_{1|1}}$ and $\frac{\pi_c}{\pi_a+\pi_c}=\frac{P_{1|1}-P_{1|0}}{P_{1|1}}=1-\frac{P_{1|0}}{P_{1|1}}$, which allows bounding the outcome distributions of either type in (3). To this end, let q correspond to the proportion of always selected in the mixed population: $q=\frac{P_{1|0}}{P_{1|1}}$. Furthermore, denote by y_q the qth conditional quantile in the conditional outcome distribution given $Z=1$ and $S=1$, i.e., $y_q=G^{-1}_{Y|Z=1,S=1}(q)$, where G is the cdf. By the results of Horowitz and Manski (1995) we then obtain the following sharp bounds on the pdf of Y among the always selected in the mixed population:

$$[0,1]\cap\left[\frac{f(y|Z=1,S=1)-(1-q)}{q},\frac{f(y|Z=1,S=1)}{q}\right]\text{ for all } y \text{ in the support of } Y. \tag{4}$$

This bounding rule holds for probability measures (also other than the pdf) in general. To extend the discussion to broader definitions of probabilities, let $\Pr(Y\in A|Z=z,S=1)$ denote the conditional probability that the value of Y belongs to some subset A given $S=1$ and $Z=z$. E.g., for some value y in the support of Y, A may be defined as $(-\infty,y]$ to obtain the cdf. Then,

[5] Note that the instrument Z and the type T uniquely determine the value of the selection indicator S such that conditioning on the latter is redundant.

the probability of $Y \in A$ among the always selected in the mixed population is bounded by

$$[0,1] \cap \left[\frac{\Pr(Y \in A | Z=1, S=1) - (1-q)}{q}, \frac{\Pr(Y \in A | Z=1, S=1)}{q} \right] \text{ for all } A \text{ in the support of } Y. \quad (5)$$

In addition to the identification of bounds, our model assumptions also point identify the pdf/probability measures of the outcome of the always selected. As defiers do not exist, it holds for all y and A in the support of Y, respectively, that

$$f(y|Z=0,S=1) = f(y|Z=0,T=a) \text{ and } \Pr(Y \in A|Z=0,S=1) = \Pr(Y \in A|Z=0,T=a). \quad (6)$$

Finally, by the exclusion restriction, $\Pr(Y \in A|Z=1,T=a) = \Pr(Y \in A|Z=0,T=a) = \Pr(Y \in A|T=a)$, otherwise the instrument Z would affect Y via U. This implies that $\Pr(Y \in A|Z=0,T=a)$, the point identified probability measure of Y among always selected given $Z=0,S=1$, must lie within the bounds of $\Pr(Y \in A|Z=1,T=a)$ in the mixed population with $Z=1,S=1$. If this is not the case, either the exclusion restriction, or monotonicity, or both assumptions are necessarily violated. Hence, by Assumptions 1 and 2 it must hold that[6]

$$\frac{\Pr(Y \in A|Z=1,S=1) - (1-q)}{q} \leq \Pr(Y \in A|Z=0,S=1) \leq \frac{\Pr(Y \in A|Z=1,S=1)}{q}$$
$$\text{for all } A \text{ in the support of } Y, \quad (7)$$

which yields two testable inequality constraints:

$$H_0 : \begin{pmatrix} \frac{\Pr(Y \in A|Z=1,S=1)-(1-q)}{q} - \Pr(Y \in A|Z=0,S=1) \\ \Pr(Y \in A|Z=0,S=1) - \frac{\Pr(Y \in A|Z=1,S=1)}{q} \end{pmatrix} \equiv \begin{pmatrix} \theta_1^p \\ \theta_2^p \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (8)$$

(8) refers to any probability measure A of the outcome, including its density function. One can therefore construct tests with multiple inequality constraints by using several subsets A. The number of constraints obtained is twice the number of probability measures considered.

[6] In Appendix A.1 we show how this result compares to Kitagawa (2010), who derives a related testable implication based on comparable model assumptions.

In addition to probability measures, we can analogously to Huber and Mellace (2011) also bound the mean outcome of the always selected by making use of trimmed averages in the mixed population, see also the related results in Lee (2009):

$$[E(Y|Z=1,S=1,Y\leq y_q),E(Y|Z=1,S=1,Y\geq y_{1-q})]. \tag{9}$$

I.e., sharp bounds on the mean outcome of the always selected are obtained by averaging Y over the upper and lower shares of the distribution which correspond to the proportion of always selected in the mixed population (given $Z=1$ and $S=1$). Furthermore and in analogy to (6), the mean outcome among the always selected is point identified:

$$E(Y|Z=0,T=a)=E(Y|Z=0,S=1). \tag{10}$$

Under Assumptions 1 and 2, the point identified mean outcome of the always selected conditional on $Z=0,S=1$, which is given in (10), must lie within its bounds in the mixed population with $Z=1,S=1$, as provided in (9). It therefore has to hold that

$$E(Y|Z=1,S=1,Y\leq y_q)\leq E(Y|Z=0,S=1)\leq E(Y|Z=1,S=1,Y\geq y_{1-q}). \tag{11}$$

This again implies two inequality constraints:

$$H_0:\begin{pmatrix} E(Y|Z=1,S=1,Y\leq y_q)-E(Y|Z=0,S=1) \\ E(Y|Z=0,S=1)-E(Y|Z=1,S=1,Y\geq y_{1-q}) \end{pmatrix}\equiv\begin{pmatrix}\theta_1^m\\ \theta_2^m\end{pmatrix}\leq\begin{pmatrix}0\\0\end{pmatrix}. \tag{12}$$

Note that if Assumptions 1 and/or 2 are violated, at most one of the two constraints in (8) and (12), respectively, is binding, because violations of the respective first and second constraint are mutually exclusive. Even if no inequality constraint is binding, the assumptions may not be satisfied if the violations are small enough so that the point identified parameter of the always selected still lies within the bounds in the mixed population. Therefore, testing power increases

as the proportion of compliers decreases, implying that the bounds on the probabilities and mean outcomes of the always selected become tighter.

To test (8) and (12) in the applications, we use the method of Chen and Szroeter (2012) for testing multiple inequality constraints, which has several desirable properties. First, it pre-estimates the constraints that are (close to being) violated to base the test statistic only on these constraints, which increases power. Second, testing is (in the spirit of Horowitz, 1992) based on indicator smoothing of the functions indicating whether the constraints are violated at the origin of the constraints, where the distribution of the test statistic is discontinuous. After smoothing, standard asymptotic theory applies to the test statistic such that bootstrapping is not necessarily required to approximate its distribution as in other tests. Finally, Chen and Szroeter (2012) show that their test has correct asymptotic size in the uniform sense under certain regularity conditions,[7] which is desirable given that the test statistic's asymptotic distribution is discontinuous w.r.t. the number of binding constraints. In the applications, we use the standard normal cdf as smoothing function for the indicators, see the algorithm in Appendix A.2 for more details on the implementation of the test.

4 Applications

In this section, we present eight applications to test Assumptions 1 and 2 when considering two variables prominently used as instruments in sample selection models concerned with the estimation of female wage equations. The first variable is non-wife or husband's income in Martins (2001), Schafgans (1998), Chang (2011), and the instructional data set "LABSUP" of Wooldrigde (http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html). The second one is the number of (young) children in the household available in Martins (2001), Mulligan and Rubinstein (2008), Chang (2011), and Lee (2009). In our analysis we discretize both instruments. In three out of four cases, the first instrument is equal to one if non-wife or husband's income is larger than the

[7] As discussed in Chen and Szroeter (2012), a sufficient condition for correct asymptotic size in the uniform sense is that the first four moments exist for each of the i.i.d. data points used to estimate the constraints.

median value in the sample and zero otherwise. The second instrument indicates whether the number of children is larger than zero. By discretization, we sacrifice asymptotic testing power. On the other hand, this will help to ensure that the number of observations is not too small when investigating whether Assumptions 1 and 2 hold conditionally in subsamples defined upon observables X. We test the constraints on the mean outcome of the always selected postulated in (12), as well as on the probability measures, see (8). In the latter case, we use four subsets A which are defined by an equidistant grid over the support of the observed outcomes. This provides us with eight testable inequality constraints.[8]

The results are presented in Tables 4 and 5. Under the satisfaction of Assumptions 1 and 2, the third column provides the complier or defier share (depending on whether the parameter is positive or negative), which is relevant because testing power decreases in its absolute value. The fourth column reports the standardized maximum of the mean constraints in (12). I.e., the maximum of (θ_1^m, θ_2^m) is divided by the standard deviation of the observed outcomes. A negative value or zero implies that no constraint is violated, while the converse is true for positive differences. This standardized distance (denoted by st.dist) therefore indicates how severely a constraint is violated (however, without saying anything about precision). Columns 5 and 6 contain the p-values of the Chen and Szroeter (2012) tests of the mean- and probability-based constraints (12) and (8), denoted by p-val(mean) and p-val(prob), respectively.

We first consider the data of Schafgans (1998) which come from the Second Malaysian Family Life Survey (MFLS-2) conducted between August 1988 and January 1989 in Peninsular Malaysia. The author investigates the wage gap between ethnic Chinese and Malays for both males and females. Here, we only use the female subsample with non-missing information on unearned income, which consists of 2,770 observations. The selection variable S equals 1 if an individual is a wage worker and 0 otherwise. The category of non-wage workers therefore also includes self-employed and individuals engaged in home production. The outcome variable Y is the log of

[8]Which number and definition of the subsets A is optimal for testing is an unsolved issue. We therefore also considered more or less subsets, but the results did not differ in an important way and are for this reason not reported here.

Table 4: Applications - non-wife income

Study	n	% comp.	st.dist	p-val(mean)	p-val(prob)
Schafgans (1998)- female sample	2,770	0.001	0.658	0.000	0.001
Malay (M)	1,477	0.053	0.532	0.001	0.078
M, 11 or 12 yrs of schooling	296	0.032	0.571	0.007	0.131
M, 11/12y.s., age 25-35, pot.exp.10-20yrs	56	0.078	0.480	0.202	0.361
Martins (2001) - full sample	2,338	0.033	0.340	0.000	0.000
yrs of schooling < 12	1,999	-0.013	0.143	0.126	0.002
yrs of sch. < 12, pot. exp. 20-30 yrs	732	-0.002	0.306	0.009	0.000
yrs of sch. < 12, pot. exp. 20/21 yrs	165	-0.002	0.359	0.099	0.014
Chang (2011) - 1985 sample	1,627	-0.049	0.247	0.000	0.000
white (W)	1,282	-0.067	0.185	0.004	0.000
W, 12 yrs of schooling	609	-0.058	0.278	0.004	0.040
W, 12 yrs of s., age 30-35, recent job	89	-0.116	0.450	0.037	0.088
Wooldridge - full sample of mothers	31,857	0.034	0.134	0.000	0.046
hispanic (H)	18,897	0.021	0.064	0.082	0.476
H, < 10 yrs of schooling	7,085	-0.003	0.192	0.014	0.716
H, 12 yrs of schooling, age 25	341	-0.067	0.245	0.176	0.981

the hourly wage rate, which is only observed conditional on $S = 1$. In contrast, the regressors X are observed for the entire sample and comprise potential experience, education, and a dummy for living in an urban area. The instrument Z to be tested is unearned income, i.e., income not coming from paid work. We dichotomize this variable such that it is equal to one whenever unearned income is larger than zero, which is the case for 1,430 observations.

The tests reject Assumptions 1 and/or 2 in the entire sample on the 0.1% level of significance. Also the rather large standardized distance of 0.658 indicates that the point estimate of the mean outcome of the always selected is well outside its bounds. When considering the subsample with Malay ethnicity, the mean- and probability-based tests reject the assumptions at the 0.1% and 10% levels, respectively. The violation of the mean constraints remains highly significant when conditioning on both Malay ethnicity and 11 or 12 years of schooling. Finally, we additionally restrict the sample to include only individuals in the age bracket of 25-35 years with 10-20 years of potential experience. Only then both tests are not significant at any conventional level any more, which, however, may be due to the small sample size of just 56 observations and the related low finite sample power. As a further worrisome matter in addition to the test results, note that compliance remains always positive even when conditioning on observed characteristics, while in the remaining three applications, conditional 'compliance' is negative (which is more in line with

the empirical literature). This points to the violation of the monotonicity assumption, because the former is consistent with the nonexistence of defiers and the latter with the nonexistence of compliers.

Our second application is based on Portuguese female labor market data from Martins (2001), who compared parametric and semiparametric estimators of sample selection models. The sample stems from the 1991 wave of the Portuguese Employment Survey and consists of 2,339 married women aged below 60 years whose husbands earned labor income. The outcome variable is log hourly wage, which is only observed for those 1,400 women who participate in the labor market (such that $S = 1$). The regressors X include years of education and potential experience. The instrumental variable Z is the log of husband's wage, which we use to create a binary variable indicating whether husband's wage is higher than the median (11.085). Applying our tests to the full sample (i.e., testing Assumptions 1 and 2 unconditionally) clearly rejects the identifying assumptions. When restricting the data to individuals with a low level of education (less than 12 years), the mean-based test is only borderline significant, while the p-value of the probability-based test is again close to zero. Conditioning on particular brackets of potential education on top of the previous restrictions yields p-values below 10% for both tests. This suggests that Assumptions 1 and 2 are violated given the regressors available in Martins (2001).

Chang (2011) proposes a simulation estimator for two-tiered dynamic panel tobit models which is applied to a 9-year panel data set from the Panel Study of Income Dynamics (PSID). The sample contains observations for 1,627 married women between 1984 and 1992 who are aged between 19 and 60 years in 1985. Here, we focus on the 1985 wave. The selection indicator S is equal to one if the woman provided a positive supply of hours worked in that year. Wife's income in 1985 serves as outcome variable Y. The instrument Z is husband's income 1985, the median of which is 25,228 USD. Furthermore, the data contain information on education, age, race, and the recent employment history as conditioning set X. The p-values of the tests are close to zero when considering the entire sample or the subsample of whites. Either test statistic also remains significant at the 10% or a lower level when controlling for (i) white ethnicity and

years of schooling and (ii) white ethnicity, years of schooling, the age bracket 30-35 years, and recent employment. The results therefore suggest that our assumptions are unlikely to hold.

The fourth application we consider is the instructional "LABSUP" data set on married mothers in the US provided by J. Wooldridge on his website of textbook data sets (http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html). For a sample of Blacks and Hispanics, it contains similar information as the U.S. Census Public Use Micro Samples used by Angrist and Evans (1998) to estimate the effect of fertility on female labor supply. With 31,857 observations, it is substantially larger than the samples investigated so far. The outcome variable is mother's labor income per year in 1,000 USD, which is only observed for the 18,789 individuals providing positive labor supply ($S = 1$). The regressors X include ethnicity, years of schooling, and age. The instrument Z is non-wife income per year in 1,000, the median of which is 29.399. Applied to the entire data, the tests again suggest that our assumptions do not hold. When considering the subsamples of (i) Hispanics and (ii) Hispanics with low education, the probability-based test statistic becomes insignificant, but the mean-based test still rejects the identifying assumptions at the 10 and 5% levels, respectively. Finally, we restrict the sample to 25 years old Hispanics with high school education. Only then, the p-values of both tests exceed the 10% level, albeit the standardized distance of 0.245 still points to a violation of Assumptions 1 and/or 2.

We now turn to our second supposed instrument, (young) children in the household. We first reconsider the data of Martins (2001), who uses kids under 3 as instrument, too. This time, however, testing does not point to a violation of the exclusion restriction and/or monotonicity. When considering the entire sample and the same subsamples as before, the p-values are quite large. Also the standardized distance remains negative throughout, implying that none of the constraints are binding.

Mulligan and Rubinstein (2008) investigate two repeated cross sections (1975-1979) and (1995-1999) of the US Current Population Survey (CPS) to determine the selection of females into the full-time work force over time using Heckman two-step estimation. Individuals are classified

Table 5: Applications - young kids

Study	n	% comp.	st.dist	p-val(mean)	p-val(prob)
Martins (2001) - full sample	2,338	0.058	-0.097	0.994	0.982
yrs of schooling < 12	1,999	0.037	-0.051	0.959	0.559
yrs of s. < 12, pot. exp. 20-30 yrs	732	-0.113	-0.247	1.000	0.994
yrs of s. < 12, pot. exp. 20/21 yrs	165	-0.199	-0.382	1.000	0.984
MR (2008) - 1995-1999 married females	53,966	-0.151	-0.351	1.000	0.735
high school graduate (hsg)	18,383	-0.130	-0.289	1.000	0.999
hsg, pot. exp. 20-30 yrs	7,889	-0.107	-0.159	0.997	0.910
hsg, pot. exp. 20 yrs, south	239	-0.017	-0.021	0.895	0.679
Lee (2009) - female sample	4,044	-0.046	-0.024	0.815	0.985
married, black (B)	2,021	-0.050	0.002	0.722	0.988
mar., B, yrs of s. < 12	1,475	-0.058	-0.163	1.000	1.000
mar., B, yrs of s. < 12, no recent job	699	-0.030	-0.294	0.996	0.938
Chang (2011) - 1985 sample	1,627	-0.110	0.016	0.463	1.000
white (W)	1,282	-0.134	0.043	0.371	1.000
W, 12 yrs of schooling	609	-0.126	0.040	0.480	1.000
W, 12 yrs of s., age 30-35, recent job	89	-0.034	0.219	0.462	0.927

as working ($S = 1$) if they work 35+ hours per week and at least 50 weeks during the year. Self-employed and persons in the military, agriculture, or private household sectors as well as individuals with inconsistent reports on earnings or with allocated earnings are excluded from the sample with observed wages, see Mulligan and Rubinstein (2008) for further details. The outcome variable is log hourly wage, which is computed based on total annual earnings deflated by the US Consumer Price Index. Here, we focus on the subsample of married white females aged between 25 and 54 years in the second repeated cross section, in total 53,966 observations. The instrument Z is the incidence of children aged 0-6 in the household. The regressors X include education, potential work experience, the marital status, and regional dummies. Assumptions 1 and 2 are neither rejected in the entire sample, nor in subsamples with (i) high school graduation, (ii) high school graduation and potential experience of 20-30 years, and (iii) high school graduation, 20 years of potential experience and living in the Southern states.

Next, we consider data from a labor market policy experiment which was conducted in the U.S. in the mid-1990s in order to assess the publicly funded Job Corps program. This program targeted young individuals (aged 16-24 years) that had a legal residence in the U.S. and came from a low-income household, see Schochet, Burghardt, and Glazerman (2001) for further details. It provided participants with approximately 1,100 hours of vocational training and education

as well as with housing, board, and health services over an average duration of roughly eight months. Here, we use the female subsample of the experimental data as also analyzed by Lee (2009), which includes 4,044 observations. The selection indicator S states if someone is working one year after program start, which is the case for 1,454 individuals. The outcome Y is the hourly wage. The baseline survey prior to the program contains (among other factors) information on the marital status, ethnicity, education, and recent labor market history, which serve as regressors X. The instrument Z is the incidence of children in the household. We do not find violations of Assumptions 1 and 2, neither in the entire sample, nor in subsamples defined upon marital status, ethnicity, low education, and recent unemployment.

In our last application, we return to Chang (2011), who uses the number young children as instrument, too. Our binary Z indicates whether at least one child under the age of six is present in the household. Relying on the same sample restrictions as before, all p-values are considerably larger than 10%. However, the standardized distances are positive throughout, indicating that one constraint is binding (albeit insignificantly so).

We conclude that the tests do not provide evidence against Assumptions 1 and 2 when using young children as instrument. In contrast, our results evoke serious concerns about non-wife/husband's income or similar variables. This is bad news for the literature on female wage equations, because in general, at least one element in Z needs to be continuous when using flexible (semi- or non-parametric) sample selection models. With this respect, non-wife income appeared to be much more suitable than the number of children, which typically consists of very few mass points. However, our tests suggest that non-wife income should not be used as instrument, at least conditional on the regressors commonly found in applications.

5 Conclusion

This paper has proposed tests for the joint satisfaction of two identifying assumptions in sample selection models by applying the approach of Huber and Mellace (2011) (who consider the

conceptually different framework of treatment endogeneity). Sample selection models commonly invoke two restrictions: (i) at least one variable is correlated with selection but does not directly affect the outcome (exclusion restriction) and (ii) the unobserved term in the selection equation is additively separable. We have shown that these assumptions allow us to both bound and point identify the outcome distribution of the always selected. As the point must lie within the bounds, this provides us with two testable inequality constraints. Furthermore, we have tested the identifying assumptions in eight empirical applications of female wage regressions where the exclusion restriction was invoked for two different variables: non-wife income and the number of young children. Our results suggest that the assumptions are likely violated for non-wife income, at least conditional on the regressors commonly available in such studies. In contrast, the tests do not reject the identifying assumptions for the number of young children.

A Appendix

A.1 Link to Kitagawa (2009)

The subsequent discussion links the testable implications of Section 3 to Kitagawa (2010), who derives a testable implication based on comparable model assumptions. Considering only positive monotonicity, Kitagawa (2010) shows in his Proposition 2.3 that under Assumptions 1 and 2,

$$f(y, S = 1|Z = 0) \leq f(y, S = 1|Z = 1) \text{ for all } y \text{ in the support of } Y. \tag{A.1}$$

I.e., the joint density of Y and $S = 1$ given $Z = 1$ must nest the joint density of Y and $S = 1$ given $Z = 0$ for any value of Y. Rearranging terms such that $f(y, S = 1|Z = 1) - f(y, S = 1|Z = 0) \geq 0$ gives the intuitive interpretation that the pdf of the compliers' outcome cannot be smaller than zero, as densities must not be negative.

Note that (7) in Section 3 is equivalent to

$$\frac{\Pr(Y \in A, S = 1|Z = 1)}{P_{1|0}} - \frac{P_{1|1} - P_{1|0}}{P_{1|0}} \quad \leq \quad \frac{\Pr(Y \in A, S = 1|Z = 0)}{P_{1|0}} \leq \frac{\Pr(Y \in A, S = 1|Z = 1)}{P_{1|0}} \tag{A.2}$$

for all A in the support of Y, because

$$\begin{aligned}
\frac{\Pr(Y \in A|Z = 1, S = 1) - (1 - q)}{q} &= \frac{\Pr(Y \in A, S = 1|Z = 1)}{q \cdot \Pr(S = 1|Z = 1)} - \frac{(1 - q)}{q} \\
&= \frac{\Pr(Y \in A, S = 1|Z = 1)}{P_{1|0}} - \frac{P_{1|1} - P_{1|0}}{P_{1|0}}, \\
\frac{\Pr(Y \in A|Z = 1, S = 1)}{q} &= \frac{\Pr(Y \in A, S = 1|Z = 1)}{q \cdot \Pr(S = 1|Z = 1)} \\
&= \frac{\Pr(Y \in V, D = 1|Z = 1)}{P_{1|0}}, \\
\Pr(Y \in A|Z = 0, S = 1) &= \frac{\Pr(Y \in A, S = 1|Z = 0)}{P_{1|0}},
\end{aligned}$$

by using basic probability theory. (A.2) in turn implies that for all A in the support of Y,

$$\Pr(Y \in A, S = 1|Z = 1) - (P_{1|1} - P_{1|0}) \quad \leq \quad \Pr(Y \in A, S = 1|Z = 0) \leq \Pr(Y \in A, S = 1|Z = 1), \tag{A.3}$$

and when applied to the pdf, that for all y in the support of Y

$$f(y, S = 1|Z = 1) - (P_{1|1} - P_{1|0}) \leq f(y, S = 1|Z = 0) \leq f(y, S = 1|Z = 1). \tag{A.4}$$

I.e., (A.4) yields one additional testable implication compared to (A.1). If we rearrange the first part in (A.3) $\Pr(Y \in A, S = 1|Z = 1) - (P_{1|1} - P_{1|0}) \leq \Pr(Y \in A, S = 1|Z = 0)$ to be $\Pr(Y \in A, S = 1|Z = 1) - \Pr(Y \in A, S = 1|Z = 0) \leq (P_{1|1} - P_{1|0})$, our additional implication gets an intuitive interpretation: The joint probability of being a complier and having a particular value of the outcome (and any sum of joint probabilities defined by non-overlapping subsets A) must not be larger than the unconditional probability of being a complier, because

$$\int [f(y, S = 1|Z = 1) - f(y, S = 1|Z = 0)]dy = P_{1|1} - P_{1|0}. \tag{A.5}$$

It is worth noting that if testing is based on subsets A that are non-overlapping and jointly cover the entire support of Y, then our additional testable implication in (A.4) is are already taken into account by (A.1) and thus redundant. The prevalence of some $\Pr(Y \in A, S = 1|Z = 1) - \Pr(Y \in A, S = 1|Z = 0) > (P_{1|1} - P_{1|0})$ then necessarily implies the existence of at least one distinct A' for which $\Pr(Y \in A', S = 1|Z = 1) - \Pr(Y \in A', S = 1|Z = 0) < 0$ so that (A.1) is violated, too. Therefore, power gains from the additional testable implication might possibly only be realized when using subsets A that overlap (so that violations may be averaged out) and/or do not cover the entire support of Y, see also the discussion in Huber and Mellace (2011).

A.2 Chen and Szroeter's test algorithm

This section provides the algorithm of the Chen and Szroeter (2012) test when testing the constraints on the mean outcome given in (12), but testing the probability constraints in (8) is analogous. Let $\hat{\theta}$ denote the sample analog of $\theta = (\theta_1^m, \theta_2^m)'$. The algorithm can be sketched as follows:

1. Estimate the vector of parameters $\hat{\theta}$ and the asymptotic variance $\hat{J}$ of $\sqrt{n} \cdot (\hat{\theta} - \theta)$.
2. Let $\hat{\eta}_i = 1/\sqrt{\hat{J}_i},\ i = 1, 2$, where $\hat{J}_i$ is the i-th element of the main diagonal of $\hat{J}$, and compute the smoothing function $\hat{\Psi}_i(\delta_n^{-1} \cdot \hat{\eta}_i \cdot \hat{\theta}_i) = \Phi(\delta_n^{-1} \cdot \hat{\eta}_i \cdot \theta_i)$, where Φ is the standard normal cdf and the tuning parameter δ_n is a sequence satisfying $\delta_n \to 0$ and $\sqrt{n} \cdot \delta_n \to \infty$ as $n \to \infty$. In the applications, we choose $\delta_n = \sqrt{\frac{2 \cdot \ln(\ln(n))}{n}} \cdot \hat{\sigma}_{\theta_i}$, where $\hat{\sigma}_{\theta_i}$ is the estimated standard deviation of the i-th inequality constraint.
3. Compute the approximation term $\hat{\Lambda}_i = \phi(\delta_n^{-1} \cdot \hat{\eta}_i \cdot \hat{\theta}_i) \cdot \frac{1}{\delta_n \cdot \sqrt{n}}, \quad i = 1, 2$, with ϕ being the standard normal pdf.
4. Define the vectors $\hat{\Psi} = \left(\hat{\Psi}_1(\delta_n^{-1} \cdot \hat{\eta}_1 \cdot \hat{\theta}_1), \hat{\Psi}_2(\delta_n^{-1} \cdot \hat{\eta}_2 \cdot \hat{\theta}_2)\right)^T$, $\hat{\Lambda} = \left(\hat{\Lambda}_1, \hat{\Lambda}_2\right)^T$, $\iota_2 = (1, 1)^T$, $\hat{\Delta} = diag(\hat{J}_1, \hat{J}_2)$.
5. Let $\hat{Q}_1 = \sqrt(n) \cdot \hat{\Psi}^T \hat{\Delta} \hat{\theta} - \iota_2^T \hat{\Lambda}$ and $\hat{Q}_2 = \sqrt{\hat{\Psi}^T \hat{\Delta} \hat{J} \hat{\Delta} \hat{\Psi}}$.
6. Compute the p-value as $\hat{p} = \begin{cases} 1 - \Phi\left(\frac{\hat{Q}_1}{\hat{Q}_2}\right) & \text{if } \hat{Q}_2 > 0 \\ 1 & \text{if } \hat{Q}_2 = 0. \end{cases}$

References

Ahn, H., and J. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

Angrist, J., E. Bettinger, and M. Kremer (2006): "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *American Economic Review*, 96, 847–862.

Angrist, J., and W. Evans (1998): "Children and their parents labor supply: Evidence from exogeneous variation in family size," *American Economic Review*, 88, 450–477.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects using Instrumental Variables," *Journal of American Statistical Association*, 91, 444–472 (with discussion).

Angrist, J., D. Lang, and P. Oreopoulos (2009): "Incentives and Services for College Achievement: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, 1, 136–163.

Becker, G. (1981): *A Treatise on the Family*. Harvard University Press, Cambridge, Mass.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds," *Econometrica*, 75, 323–363.

Chang, S.-K. (2011): "Simulation estimation of two-tiered dynamic panel Tobit models with an application to the labor supply of married women," *Journal of Applied Econometrics*, 26, 854–871.

Chen, L.-Y., and J. Szroeter (2012): "Testing Multiple Inequality Hypotheses: A Smoothed Indicator Approach," *CeMMAP working paper 16/12*.

Cosslett, S. (1991): "Distribution-Free Estimator of a Regression Model with Sample Selectivity," in *Nonparametric and semiparametric methods in econometrics and statistics*, ed. by W. Barnett, J. Powell, and G. Tauchen, pp. 175–198. Cambridge University Press, Camdridge, UK.

Crépon, B. (2006): "Testing Exclusion Restrictions at Infinity in the Semiparametric Selection Model," *IZA Discussion Paper no. 2035*.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

Fleisher, B. M., and J. Rhodes, George F. (1979): "Fertility, Women's Wage Rates, and Labor Supply," *The American Economic Review*, 69, 14–24.

Frangakis, C. E., and D. B. Rubin (2002): "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

Gallant, A., and D. Nychka (1987): "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica*, 55, 363–390.

Gronau, R. (1974): "Wage comparisons-a selectivity bias," *Journal of Political Economy*, 82, 1119–1143.

Heckman, J. J. (1974): "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 42, 679–694.

——— (1976): "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement*, 5, 475–492.

——— (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153–161.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

Horowitz, J. L., and C. F. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.

Huber, M., and G. Mellace (2011): "Testing instrument validity for LATE identification based on inequality moment constraints," *University of St Gallen, Dept. of Economics Discussion Paper no. 2011-43*.

Imbens, G. W., and D. Rubin (1997): "Estimating outcome distributions for compliers in instrumental variables models," *Review of Economic Studies*, 64, 555–574.

Kitagawa, T. (2010): "Testing for instrument independence in the selection model," *unpublished manuscript, UCL*.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Manski, C. F. (2003): *Partial Identification of Probability Distributions*. New York: Springer Verlag.

Martins, M. (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.

Mealli, F., and B. Pacini (2008): "Exploiting instrumental variables in causal inference with nonignorable outcome nonresponse using principal stratification," *mimeo*.

Mroz, T. (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 Y. Rubinstein (2008): "Selection, Investment, and Women's Relative Wages Over Time," *Quarterly Journal of Economics*, 123, 1061–1110.

Nakosteen, R. A., O. Westerlund, and M. A. Zimmer (2004): "Marital Matching and Earnings: Evidence from the Unmarried Population in Sweden," *The Journal of Human Resources*, 39, 1033–1044.

Newey, W. K. (2007): "Nonparametric continuous/discrete choice models," *International Economic Review*, 48, 1429–1439.

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

Powell, J. L. (1987): "Semiparametric Estimation of Bivariate Latent Variable Models," *unpublished manuscript*, University of Wisconsin-Madison.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Schafgans, M. M. A. (1998): "Ethnic wage differences in Malaysia: parametric and semiparametric estimation of the ChineseMalay wage gap," *Journal of Applied Econometrics*, 13, 481–504.

Schochet, P. Z., J. Burghardt, and S. Glazerman (2001): "National Job Corps Study: The Impacts of Job Corps on Participants Employment and Related Outcomes," *Report (Washington, DC: Mathematica Policy Research, Inc.)*.

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

Zabel, J. E. (1993): "The Relationship between Hours of Work and Labor Force Participation in Four Models of Labor Supply Behavior," *Journal of Labor Economics*, 11, 387–416.