Testing instrument validity in sample selection models

Martin Huber and Giovanni Mellace

December 2011 Discussion Paper no. 2011-45
Testing instrument validity in sample selection models\textsuperscript{1}

Martin Huber and Giovanni Mellace

\textsuperscript{1} We have benefited from comments by Alberto Abadie, Joshua Angrist, Guido Imbens, Toru Kitagawa, and seminar participants at Harvard (seminar in econometrics, September 2011). Martin Huber gratefully acknowledges financial support from the Swiss National Science Foundation grant PBSGP1_138770.
Abstract

This paper proposes tests for instrument validity in sample selection models with non-randomly censored outcomes. Such models commonly invoke an exclusion restriction (i.e., the availability of an instrument affecting selection, but not the outcome) and additive separability of the errors in the selection process. These assumptions allow us to both point identify and bound the outcome distribution of the subpopulation of the always selected, whose outcomes are observed irrespective of the instrument value. As the point must lie within its bounds, this yields two testable inequality constraints. We apply our tests to two instruments conventionally exploited for the estimation of female wage equations: non-wife/husband’s income and the number of (young) children. Considering eight empirical applications, our results suggest that the former is not a valid instrument, while the validity of the latter is not refuted on statistical grounds.

Keywords

Sample selection, instrument, test.

JEL Classification

C12, C15, C24, C26.
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 rely on instrumental variables (IV) which are presumably related to selection, but have no direct effect on the outcome of interest. Such IV exclusion restrictions are even more crucial when considering 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 satisfaction of the exclusion restriction and secondly, the monotonicity of selection in the instrument (or equivalently, the additive separability of the errors in the selection equation). However, the validity of these assumptions in applied work has not been assessed on statistical grounds.

Therefore, the first contribution of this paper is the proposition of a novel test for the joint satisfaction of the exclusion restriction and monotonicity. It is closely related to the test suggested in Huber and Mellace (2011), who, however, consider the conceptually different framework of treatment endogeneity. Assuming a binary instrument for the ease of exposition, the test is based on the following intuition. Under IV validity, the outcome distribution of the always selected, who are selected irrespective of the instrument, is point identified. It simply corresponds to the distribution in the selected subpopulation not receiving the instrument. On the other hand, the selected subpopulation receiving the instrument consists of both always selected and compliers, whose selection state reacts on the instrument. In this mixed population, upper and lower bounds on the distribution of the always selected can be derived by using the results of Horowitz and Manski (1995). Clearly, the identified outcome distribution of the always selected in the absence of the instrument must lie within the bounds in the presence of the instrument, otherwise the IV assumptions are violated. This provides us with two inequality constraints, which we verify by using the minimum p-value test with partial recentering proposed by Bennett (2009). While the test is not asymptotically uniformly powerful in the sense that the null may still be violated even if the point identified parameter is within its bounds, a rejection unambiguously points to a violation of IV validity.

It is worth noting that further IV validity tests have been proposed in the context of sample

\[1\text{Such monotonicity restrictions have been prominently discussed in Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996), albeit in the different context of treatment endogeneity.}\]
selection models. Blundell, Gosling, Ichimura, and Meghir (2007) and Kitagawa (2009) suggest methods to verify the IV 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. They do not show asymptotic validity of their inferential bootstrap procedure.

Kitagawa (2009) 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 (2009) 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 will show that one of our constraints is equivalent to the implication of Kitagawa (2009). However, based on the same the model assumptions, we also provide a second testable constraint not considered therein. 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 (2009) requires that some outcomes are observed with certainty.

Our second contribution is an empirical one. We investigate the IV validity of two variables prominently used in female wage regressions to control for sample selection, where selection bias comes from the labor supply decision. The first instrument is non-wife income (such as husband’s income or other sources of family income). Most empirical studies find that non-wife income

\footnote{It has already been noticed by Manski (2003) that the exclusion restriction is violated if the identification region defined by the bounds is empty.}
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 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. We therefore test the IV validity of non-wife income in four data sets that cover various countries and/or socio-economic groups and come from Schafgans (1998), Martins (2001), Chang (2011), and Wooldridge’s online archive (the LABSUP data set on http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html).

The second 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 validity of this instrument, implying that the number of children is not directly related with wages, is not undisputed. There are theoretical arguments that 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 IV exclusion restriction would fail. Therefore, we investigate the IV validity in four data sets previously analyzed by Martins (2001), Mulligan and Rubinstein (2008), Lee (2009), and Chang (2011).

The remainder of this paper is organized as follows. Section 2 characterizes the sample selection model and the testable implications. For ease of exposition, a binary instrument will be assumed in the first place. Section 3 introduces the tests based on Bennett (2009). Section 4 generalizes the discussion to non-binary instruments. The empirical applications to the estimation of female wage equations are presented in Section 5. Section 6 concludes.
2 The selection model and testable implications

This section outlines the sample selection model along with the testable implications. To this end, we introduce some notation. \( Y \) is a (scalar) continuous outcome variable with bounded support, \( X \) denotes a scalar or a vector of covariate(s), and \( U \) represents the unobserved terms affecting the outcome. A general outcome equation can be written as

\[
Y = \varphi(X, U),
\]

where \( \varphi \) is an unknown function. 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. Then, the selection decision may be expressed as

\[
S = I\{\zeta(W, V) > 0\},
\]

where \( I\{\cdot\} \) denotes the indicator function and \( \zeta(\cdot) \) is an unknown function.

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. \( Y \), e.g., see 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 instruments 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 with \( U \). 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 required if we want to give a causal interpretation to the covariates. However, here we are just interested in testing IV validity such the following conditional independence assumption suffices:

**Assumption 1:**
\( Z \) is independent of \((U, V)|X\) (conditional independence of the instrument and unobserved terms).

In addition to the exclusion restriction, virtually all selection models, including the nonparametric framework of Newey (2007), impose additive separability of the unobserved term in the selection equation, which results in the standard single index crossing model

\[
S = I\{\zeta(X, Z) + V > 0\}. \tag{3}
\]

Additive separability is attractive because 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.\(^3\) As discussed in Das, Newey, and Vella (2003), the index restriction is implied by assuming (3) 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, note that Vytlacil (2002) shows that a selection model with additively separable unobservables can be equivalently analyzed by assuming that the potential selection state of each individual increases or decreases weakly monotonically in the value of the instrument.

\(^3\)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.
We propose a test for the joint satisfaction of (i) the exclusion restriction and (ii) monotonicity (and thus, additive separability of the unobserved term). If both assumptions hold, we say that $Z$ satisfies IV validity. Without loss of generality and to keep the exposition simple, we will derive our testable implications for a binary instrument $Z$ in the first place. The results will be extended to multivalued instruments later on. To translate (3) into the monotonicity assumption to be tested, 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 $z$. Note that $S = Z \cdot S(1) + (1 - Z) \cdot S(0)$. For a binary instrument, monotonicity (given $X$) implies the following:

**Assumption 2:**

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

Henceforth, we will only consider positive monotonicity of $S$ in $Z$, as a symmetric argument follows under negative monotonicity. Furthermore, we will omit conditioning $X$ in our discussion for the sake of ease of notation. This is in spite of our awareness that instruments may not be valid unconditionally, but only after controlling for the observed characteristics $X$. In fact, both unconditional and conditional IV validity will be examined in the empirical section. The reader may therefore consider the subsequent discussion to take place within cells defined upon $X$.

Our test exploits the fact that under IV validity, the outcome distribution of a particular subpopulation can be both point identified and bounded and that the point must lie within its bounds. In what follows, we will derive these testable implications. Using the principal stratification framework of Frangakis and Rubin (2002) and a similar notation as in Angrist, Imbens, and Rubin (1996), the population can be divided into four types according to the reaction of the selection state on the instrument. The always selected are those with observed outcomes irrespective of the instrument state, 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.

<table>
<thead>
<tr>
<th>Type ($T$)</th>
<th>$S(1)$</th>
<th>$S(0)$</th>
<th>appellation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>1</td>
<td>1</td>
<td>Always selected</td>
</tr>
<tr>
<td>$c$</td>
<td>1</td>
<td>0</td>
<td>Compliers</td>
</tr>
<tr>
<td>$d$</td>
<td>0</td>
<td>1</td>
<td>Defiers</td>
</tr>
<tr>
<td>$n$</td>
<td>0</td>
<td>0</td>
<td>Never selected</td>
</tr>
</tbody>
</table>

As either $S(1)$ or $S(0)$ but never both are known for any individual, the type of a subject is not directly observed. Without any assumptions we neither identify the proportions of the various types, nor the outcome distributions within each type, which are the ingredients of our test. To see this, note that the observed values of $Z$ and $S$ define four observed subgroups, which all are mixtures of two types. A second complication is that outcomes are only observed for those with $S = 1$. Table 2 summarizes these results.

<table>
<thead>
<tr>
<th>observed values $Z, S$</th>
<th>types</th>
<th>$Y$ observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>${i: Z_i = 1, S_i = 1}$</td>
<td>subject $i$ belongs either to $a$ or to $c$</td>
<td>yes</td>
</tr>
<tr>
<td>${i: Z_i = 1, S_i = 0}$</td>
<td>subject $i$ belongs either to $d$ or to $n$</td>
<td>no</td>
</tr>
<tr>
<td>${i: Z_i = 0, S_i = 1}$</td>
<td>subject $i$ belongs either to $a$ or to $d$</td>
<td>yes</td>
</tr>
<tr>
<td>${i: Z_i = 0, S_i = 0}$</td>
<td>subject $i$ belongs either to $c$ or to $n$</td>
<td>no</td>
</tr>
</tbody>
</table>

However, by Assumption 2, defiers are ruled out. I.e., if monotonicity holds such that defiers do not exist, all observations with $Z = 1, S = 0$ must belong to the never selected while all individuals with $Z = 0, S = 1$ are always selected. For a similar result in the context of selection models, see the discussion in Lee (2009), who, however, considers monotonicity of selection in a binary treatment. The non-existence of defiers allows us to point identify the proportions of the various remaining types $T \in \{a, c, n\}$, denoted by $\pi_T$. To see this, let $P_{s|z} \equiv \Pr(S = s|Z = z)$ the observed selection probability conditional on the instrument. By the independence of $Z$ and $V$ imposed by Assumption 1, the share of any type conditional on the instrument is equal to its unconditional proportion in the entire population. Therefore, the relationship between the observed conditional probabilities and latent type proportions is as shown in Table 3.
Table 3: Observed conditional probabilities and type proportions

<table>
<thead>
<tr>
<th>Observed cond. selection prob.</th>
<th>Type proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_{11} \equiv \Pr(S = 1</td>
<td>Z = 1) )</td>
</tr>
<tr>
<td>( P_{01} \equiv \Pr(S = 0</td>
<td>Z = 1) )</td>
</tr>
<tr>
<td>( P_{10} \equiv \Pr(S = 1</td>
<td>Z = 0) )</td>
</tr>
<tr>
<td>( P_{00} \equiv \Pr(S = 0</td>
<td>Z = 0) )</td>
</tr>
</tbody>
</table>

It is easy to see that \( P_{11} - P_{10} \) identifies the complier’s proportion. This will be important for the identification of the outcome distributions of the always selected, which our test relies upon.

We therefore introduce some further notation. Denote by \( f(y|Z = z, T = t) \) the pdf of \( Y = y \) for a particular type \( T = t \) given \( Z = z \). 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.

\[
f(y|Z = 1, S = 1) = \frac{\pi_a}{\pi_a + \pi_c} \cdot f(y|Z = 1, T = a) + \frac{\pi_c}{\pi_a + \pi_c} \cdot f(y|Z = 1, T = c). \tag{4}
\]

Horowitz and Manski (1995) have shown that whenever it is possible to bound or point identify the mixing probabilities (in our case \( \frac{\pi_a}{\pi_a + \pi_c} \) and \( \frac{\pi_c}{\pi_a + \pi_c} \), respectively), sharp bounds can be obtained for any parameter of the mixture components that respects stochastic dominance. We point identify the mixing probabilities by \( \frac{\pi_a}{\pi_a + \pi_c} = \frac{P_{10}}{P_{11}} \) and \( \frac{\pi_c}{\pi_a + \pi_c} = \frac{P_{11} - P_{10}}{P_{11}} = 1 - \frac{P_{10}}{P_{11}} \). This allows us to bound the outcome distributions of either type in the mixed distribution by applying the findings of Horowitz and Manski (1995). To this end, let \( q \) correspond to the proportion of always selected in the mixed population: \( q = \frac{P_{10}}{P_{11}} \). Furthermore, denote by \( y_q \) the \( q \)th 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. From the results of Horowitz and Manski (1995) follows that the pdf of \( Y \)

\footnotetext{Note that the instrument \( Z \) and the type \( T \) uniquely define the value of the selection indicator \( S \) such that conditioning on the latter is redundant.}
among the always selected in the mixed population is bounded by

$$[0, 1] \cap \left[ \frac{f(y|Z = 1, S = 1) - (1 - q)}{q}, \frac{f(y|Z = 1, S = 1)}{q} \right]$$

for all $y$ in the support of $Y$.

This 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 $(-\infty, y]$ for obtaining the cdf. Then, the probability of $Y \in A$ among the always selected in the mixed population is bounded by

$$[0, 1] \cap \left[ \frac{\Pr(Y \in A|Z = 1, S = 1) - (1 - q)}{q}, \frac{\Pr(Y \in A|Z = 1, S = 1)}{q} \right]$$

for all $A$ in the support of $Y$.

Concerning the mean outcome of the always selected, the results of Horowitz and Manski (1995) imply that its bounds correspond to trimmed averages, see also the related discussion in Lee (2009) (where, however, $S$ is monotonic in a treatment, not in $Z$):

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

I.e., sharp bounds on the mean outcome of the always selected are obtained by averaging $Y$ over the upper and lower share of the distribution which corresponds to the proportion of the always selected in the mixed population. Note, however, that these bounds are (in contrast to those of the probability measures) also valid under the weaker mean independence $E(U,V|Z = 1) = E(U,V|Z = 0)$ rather than full independence. Therefore, if the parameter of interest is the mean outcome $E(Y)$, Assumption 1 might be weakened to mean independence.

In addition to the identification of bounds, our model assumptions also point identify the pdf of the always selected. By Assumption 2 we have that for all $y$ and $A$ in the support of $Y$,
respectively,

\[ f(y|Z = 0, S = 1) = f(y|Z = 0, T = a) \text{ and } \Pr(Y \in A|Z = 0, S = 1) = \Pr(Y \in A|Z = 0, T = a), \]

as defiers are ruled out. Similarly, the mean outcome among the always selected is

\[ E(Y|Z = 0, T = a) = E(Y|Z = 0, S = 1). \]

Furthermore, by the exclusion restriction, \( \Pr(Y \in A|Z = 1, T = a) = \Pr(Y \in A|Z = 0, T = a) \), otherwise the instrument \( Z \) would have a direct effect on \( Y \). Therefore, our model assumptions imply that the point identified \( \Pr(Y \in A|Z = 0, T = a) \) lies within the bounds of \( \Pr(Y \in A|Z = 1, T = a) \) in the mixed population. Hence, it must hold that

\[
\frac{\Pr(Y \in A|Z = 1, S = 1) - (1 - q)}{q} \leq \Pr(Y \in A|Z = 0, S = 1) \leq \frac{\Pr(Y \in A|Z = 1, S = 1)}{q}
\]

for all \( A \) in the support of \( Y \),

or, when considering the mean, that

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

Therefore, if the testable restrictions (5) and (6) are not satisfied, either the exclusion restriction, or monotonicity, or both are violated, which is the base of our tests outlined in the next section.

We conclude this section by linking our results to Kitagawa (2009), who also derives a testable implication based on the same model assumptions. Considering only positive monotonicity, Kitagawa (2009) shows in his Proposition 2.3 that under IV validity,

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

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 must not be smaller than zero, as densities must not be negative.

Note that our restriction (5) is equivalent to

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

(8)

for all $A$ in the support of $Y$, because

$$
\frac{\Pr(Y \in A|Z = 1, S = 1)}{q} = \frac{\Pr(Y \in A, S = 1|Z = 1)}{q \cdot \Pr(S = 1|Z = 1)} = \frac{\Pr(Y \in A, S = 1|Z = 1)}{\Pr(Y \in A, S = 1|Z = 1) - (1 - q) - P_{1|1} - P_{1|0}}
$$

$$
\frac{\Pr(Y \in A|Z = 0, S = 1)}{q} = \frac{\Pr(Y \in A|V, D = 1|Z = 1)}{\Pr(Y \in V, D = 1|Z = 1)} = \frac{\Pr(Y \in A|S = 1|Z = 1)}{\Pr(Y \in A|S = 1|Z = 1)} - (1 - q)
$$

by using basic probability theory. (8) in turn implies that for all $A$ in the support of $Y$,

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

(9)

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

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

(10)

I.e., our restriction (10) adds one further testable implication to (7) derived by Kitagawa (2009). If we rearrange the first part in (9) $\Pr(Y \in A, S = 1|Z = 1) - (P_{1|1} - P_{1|0}) \leq \Pr(Y \in A, S = 1|Z = 0)$
to be \( \Pr(Y \in A, S = 1|Z = 1) - \Pr(Y \in A, S = 1|Z = 0) \leq (P_{11} - P_{10}) \), 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 complier, because

\[
\int [f(y, S = 1|Z = 1) - f(y, S = 1|Z = 0)] dy = P_{11} - P_{10}.
\]

(11)

However, it has to be mentioned that asymptotically, (10) does not increase power compared to testing based on (7) alone. The reason is that if under any value of the selected outcome (10) does not hold, there must exist at least one other value under which (7) is violated, otherwise (11) is not satisfied. The same argument carries over to any probability measures defined by non-overlapping subsets \( A \). Therefore, exploiting (10) is at best useful in finite samples, either because its violation occurs in regions where estimation is more precise than in areas where (7) is binding, and/or due to the use of overlapping subsets \( A \) for reasons of efficiency. In the latter case it may theoretically happen that (10) does not hold, while at the same time negative densities of complier outcomes average out with positive ones in the subsets \( A \) considered. For an intuitive example in a related testing problem, we refer to Huber and Mellace (2011).

3 Testing

This section outlines the test procedure based on the method proposed in Bennett (2009). First note that expression (5) provides us with two testable inequality constraints for general probabilities. Under the null hypothesis that the instrument is valid it must hold that

\[
H_0 : \begin{pmatrix}
\frac{\Pr(Y \in A|Z = 1, S = 1) - (1-q)}{q} - \Pr(Y \in A|Z = 0, S = 1) \\
\Pr(Y \in A|Z = 0, S = 1) - \frac{\Pr(Y \in A|Z = 1, S = 1)}{q}
\end{pmatrix} \equiv \begin{pmatrix}
\theta_{p1} \\
\theta_{p2}
\end{pmatrix} \leq \begin{pmatrix}
0 \\
0
\end{pmatrix}.
\]

(12)

As (12) refers to any probability measure, it can be used to construct tests with multiple inequality constraints. E.g., it may be applied to the pdf at various points in the outcome distribution,
which increases asymptotic testing power. Then, the number of constraints obtained is twice the number of probability measures considered. Concerning the mean outcome of the always selected, (6) implies the following two constraints. As already mentioned, the latter only require that the unobserved terms are mean independent of the instrument, which is a necessary, albeit not sufficient condition for full independence.

$$H_0 : \begin{pmatrix} E(Y|Z = 1, S = 1, Y \leq y_q) - E(Y|Z = 0, S = 1) \\ E(Y|Z = 0, S = 1) - E(Y|Z = 1, S = 1, Y \geq y_{1-q}) \end{pmatrix} \equiv \begin{pmatrix} \theta_{m1} \\ \theta_{m2} \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$  \hspace{1cm} (13)

Under the alternative hypothesis that the instrument does not satisfy monotonicity and/or has a direct effect on the outcome, at most one of the two constraints in (12) and (13), respectively, might be binding. This is the case because violations of the first and second constraint are mutually exclusive. Furthermore, note that even if no inequality constraint is violated, IV validity may not be satisfied. I.e., we detect violations only if they are large enough such that the point identified parameter of the always selected lies outside the bounds in the mixed population. Ideally, we would like to test for the equality of the parameters within the always selected across instrument states. However, this is not feasible as it remains unknown which individuals in the mixed population belongs to the group of always selected or compliers. Therefore, without further assumptions, testing based on inequality moment constraints is the best one can get. It is obvious that such tests gain power as the proportion of compliers decreases, implying that the bounds on the probabilities and mean outcomes of the always selected become tighter.

To keep the discussion on testing general, we will denote by \( \theta \) the collection of constraints to be tested. E.g., when considering the mean, \( \theta = (\theta_{m1}, \theta_{m2})' \), when considering probabilities, \( \theta \) may include a large number of constraints depending on the various definitions of \( A \). Furthermore, let \( \hat{\theta} \) denote the estimate of \( \theta \) based on an i.i.d. sample containing \( n \) observations, which can easily be shown to satisfy a standard GMM estimation problem, see Huber and Mellace (2011). As in our companion paper, we will use the novel minimum p-value-type test proposed by Bennett (2009) for joint inequality moment constraints. The test relies on the following, quite general assumptions.
(see Assumption 1 of Bennett, 2009) which are satisfied in a standard GMM framework: (i) i.i.d.
sampling, (ii) bounded second moments, (iv) Lipschitz continuity of the moment functions with
the Lipschitz function having bounded second moments, (v) linear representation of the testing
problem.

In contrast to other tests based on inequality constraints, e.g. Andrews and Jia (2008),
Andrews and Soares (2010), Hansen (2005), and Donald and Hsu (2010), the Bennett (2009)
test invariant to studentization and does not require the choice of any smoothing function as in
Chen and Szroeter (2009). Furthermore, it estimates the distribution of the minimum p-value
\( \min(P_{\hat{\theta}}) \) based on two sequential bootstraps, where the second resamples from the distribution
of the first bootstrap. Therefore, it does not rely on the double (i.e., nested) bootstrap (see
Beran, 1988) as in Godfrey (2005), which may be computationally intensive. Bennett (2009)
considers both full (i.e., standard) recentering of inequality constraints and partial recentering
of only those constraints which are either violated in the sample or not violated but within a
small neighborhood of the boundary of the null hypothesis. He shows that partial recentering
(henceforth minP.p) has weakly superior finite sample properties than full recentering (henceforth
minP.f). The algorithm of both methods can be sketched as follows:

1. Estimate the vector of parameters \( \hat{\theta} \) in the original sample.

2. Draw \( B_1 \) bootstrap samples of size \( n \) from the original sample.

3. In each bootstrap sample, compute the recentered vector \( \tilde{\theta}_b^{f} \equiv \hat{\theta}_b - \hat{\theta} \) for the minP.f test

   and the partially recentered vector \( \tilde{\theta}_b^{p} \equiv \hat{\theta}_b - \max(\hat{\theta}_b, -\delta_n) \) for the minP.p test, where \( \delta_n \) is

   a sequence such that \( \delta_n \to 0 \) and \( \sqrt{n} \cdot \delta_n \to \infty \) as \( n \to \infty \).\(^5\)

\(^5\)In the applications further below, we choose \( \delta_n = \sqrt{\frac{2 \ln(\ln(n))}{n}} \cdot \hat{\sigma}_i, \quad i \in \{1, 2\}, \) where \( \hat{\sigma}_i \) is the estimated (in

the \( B_1 \) first stage bootstrap samples) standard deviation of the \( i \)-th inequality constraint, as suggested by Bennett

(2009). It is, however, not guaranteed that this choice is optimal, see for instance the discussion in Donald and

Hsu (2010).
4. Estimate the vector of p-values for \( \text{minP.f} \), denoted by \( P_{\tilde{\theta}^f} \):

\[
P_{\tilde{\theta}^f} = B_1^{-1} \cdot \sum_{b=1}^{B_1} I\{ \sqrt{n} \cdot \tilde{\theta}^f_b > \sqrt{n} \cdot \hat{\theta} \}.
\]  

(14)

5. Compute the minimum P-values for \( \text{minP.f} \):

\[
\hat{p}_f = \min(P_{\tilde{\theta}^f}).
\]  

(15)

6. Draw \( B_2 \) values from the distributions of \( \tilde{\theta}^f_b \) and \( \tilde{\theta}^p_b \). We denote by \( \tilde{\theta}^f_{b_2} \) and \( \tilde{\theta}^p_{b_2} \) the resampled observations in the second bootstrap.

7. In each bootstrap sample, compute the minimum P-values of \( \text{minP.f} \) and \( \text{minP.p} \), denoted by \( \hat{p}_{f,b_2} \) and \( \hat{p}_{p,b_2} \):

\[
\hat{p}_{f,b_2} = \min(P_{\tilde{\theta}^f_{b_2}}), \quad \hat{p}_{p,b_2} = \min(P_{\tilde{\theta}^p_{b_2}}),
\]  

(16)

where

\[
P_{\tilde{\theta}^f_{b_2}} = B_1^{-1} \cdot \sum_{b=1}^{B_1} I\{ \sqrt{n} \cdot \tilde{\theta}^f_b > \sqrt{n} \cdot \tilde{\theta}^f_{b_2} \}, \quad P_{\tilde{\theta}^p_{b_2}} = B_1^{-1} \cdot \sum_{b=1}^{B_1} I\{ \sqrt{n} \cdot \tilde{\theta}^f_b > \sqrt{n} \cdot \tilde{\theta}^p_{b_2} \}.
\]  

(17)

8. Compute the p-values of the \( \text{minP.f} \) and \( \text{minP.p} \) tests by the share of bootstrapped minimum p-values that are smaller than the respective minimum p-value of the original sample:

\[
\hat{p}_\text{minP.f} = B_2^{-1} \cdot \sum_{b_2=1}^{B_2} I\{ \hat{p}_{f,b_2} \leq \hat{p}_f \}, \quad \hat{p}_\text{minP.p} = B_2^{-1} \cdot \sum_{b_2=1}^{B_2} I\{ \hat{p}_{p,b_2} \leq \hat{p}_f \}.
\]  

(18)

As already mentioned, \( \text{minP.f} \) and \( \text{minP.p} \) only differ in terms of recentering. The former test recenters all four constraints, while the latter recenters only the restrictions that either violate the null or are in the null but close (i.e., within \( \delta_n \)) to equality in the original sample. Partial recentering allows estimating the number of binding constraints from the data and therefore
provides a better approximation of the asymptotic distribution of the test under the null. It dominates minP.f in terms of power while yielding asymptotically exact size, see Bennett (2009).

4 Non-binary instruments

This section generalizes the testable implications to bounded non-binary instruments. We denote by \( z_{\text{min}} \) and \( z_{\text{max}} \) the minimum and maximum in the support of the possibly continuous or multi-valued discrete \( Z \). This requires us to replace Assumption 2 by the following monotonicity assumption:

**Assumption 3:**

\[
\Pr(S(z) \geq S(z')) = 1 \quad \forall \ z, z' \text{ satisfying } z_{\text{min}} \leq z' < z \leq z_{\text{max}} \text{ (positive monotonicity)}.
\]

I.e., \( z \) and \( z' \) are two distinct subsets of the support of \( Z \) such that any element in \( z \) is larger than any element in \( z' \).

Note that the complier share may well be small or even zero for some pairs \( z, z' \), as monotonicity implies that each individual switches its selection state at most once under the null as a reaction to different values of the instrument. While small or zero complier shares point to a weak instrument problem and appear undesirable for estimation, the converse is true for testing, as a complier share of zero maximizes asymptotic power. A further dimension relevant to testing power is the number of subsets considered. I.e., it is useful to look at all possible pairs of neighboring \( z \) and \( z' \) for which the inequality constraints must hold under instrument validity. In large samples small subsets therefore appear preferable, firstly to minimize the complier share and secondly to maximize the number of neighboring pairs of \( z \) and \( z' \). However, in small samples a trade-off between finite sample power and asymptotic power may well occur.

\[6\]Note that for any fixed \( z \), neighboring \( z \) and \( z' \) give weakly lower complier shares than non-neighboring pairs and thus, entail a higher asymptotic power.
To generalize the testable implications to non-binary instruments, define \( \tilde{Z} \) to be

\[
\tilde{Z} = \begin{cases} 
1 & \text{if } Z \in z \\
0 & \text{if } Z \in z' 
\end{cases}.
\] (19)

Under Assumptions 1 and 3, our previous results also hold when replacing \( Z \) by \( \tilde{Z} \) (and ignoring all other observations). This implies that for any \( \tilde{Z} \), we obtain the following inequality constraints:

\[
H_0 : \left( \frac{\Pr(Y \in A|\tilde{Z} = 1, S = 1) - \Pr(Y \in A|\tilde{Z} = 0, S = 1)}{q} - \frac{\Pr(Y \in A|\tilde{Z} = 1, S = 1)}{1 - q} \right) - \frac{\Pr(Y \in A|\tilde{Z} = 0, S = 1)}{q},
\] (20)

for any \( A \), and

\[
H_0 : \left( E(Y|\tilde{Z} = 1, S = 1, Y \leq y_q) - E(Y|\tilde{Z} = 1 = 0, S = 1) \right) - \left( E(Y|\tilde{Z} = 1 = 0, S = 1) - E(Y|\tilde{Z} = 1, S = 1, Y \geq y_{1-q}) \right),
\] (21)

Let \( n_{\tilde{Z}} \) be the number of possible choices of \( \tilde{Z} \) with neighboring subsets. Testing IV validity amounts to applying the test procedures outlined in Section [3], where the number of inequality constraints is now \( 2 \cdot n_{\tilde{Z}} \) instead of 2. To give an example, consider the case that \( Z \) may take the values 0, 1, or 2. The number of possible definitions of \( \tilde{Z} \) with neighboring \( z, z' \) is 4:

\[
z' = 0 \quad z = 1,
\]

\[
z' = 1 \quad z = 2,
\]

\[
z' = 0 \quad z = \{1, 2\},
\]

\[
z' = \{0, 1\} \quad z = 2.
\]

This implies that we have \( 2 \times 4 = 8 \) testable inequality constraints based on neighboring pairs.

Notice that also considering the non-neighboring pair \( z' = 0, z = 2 \) does neither increase finite nor asymptotic power: A test base on the non-neighboring pair is weakly dominated by using \( z' = \{0, 1\}, z = 2 \) and \( z' = 0, z = \{1, 2\} \) in terms of the sample size (which influences finite sample power) and entails a weakly higher complier share than any other neighboring pair.
final remark, note that $n_2$ becomes infinite when the instrument is continuous. In practice, the researcher will have to define a finite number of subsets that depends on the richness of the data in the application considered and will, thus, again face a trade-off between asymptotic and finite sample power.

5 Applications

In this section, we present eight applications to test the IV validity of two variables prominently used in sample selection models concerned with the estimation of female wage equations. The first instrument is non-wife or husband’s income, which we verify in Martins (2001), Schafgans (1998), Chang (2011), and the instructional data set “LABSUP” of Wooldridge (http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html). The second one is the number of (young) children in the household, which we consider 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 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 conditional IV validity in subsamples defined upon observables $X$. We test the constraints on the mean outcome of the always selected postulated in (13), as well as on the probability measures, see (12). In the latter case, we use four subsets $A$ which are defined by an equidistant grid on the support of the observed outcomes. This provides us with 8 testable inequality constraints.

The results are presented in Tables 4 and 5. Under the null the third column provides, depending on whether the parameter is positive or negative, the complier or defier share, respectively, which is relevant because testing power increases as the absolute value decreases.

\footnote{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 such that they are not reported here. However, they are available from the authors upon request.}
The fourth column reports the maximum of the standardized differences of the constraints on the mean in (13). I.e., the maximum of \((\theta_1^{m}, \theta_2^{m})\) is divided by the standard deviation of the observed outcomes. A negative value or zero implies that no constraint is binding, while the converse is true for positive differences. I.e., the standardized difference gives us an idea of how severely a constraint is violated (however, without saying anything about precision). Columns 5 to 8 contain the p-values with partial (p) and full (f) recentering for the mean- and probability-based tests, respectively.

Table 4: Applications - non-wife income

<table>
<thead>
<tr>
<th>Study</th>
<th>n</th>
<th>% comp.</th>
<th>st.dist</th>
<th>mean constraints</th>
<th>prob. constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p-val(p)</td>
<td>p-val(f)</td>
</tr>
<tr>
<td>Schafgans (1998)- female sample</td>
<td>2770</td>
<td>0.001</td>
<td>0.658</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Malay (M)</td>
<td>1477</td>
<td>0.053</td>
<td>0.532</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>M, 11 or 12 yrs of schooling</td>
<td>296</td>
<td>0.032</td>
<td>0.571</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>M, 11/12yrs., age 25-35, pot.exp.10-20yrs</td>
<td>56</td>
<td>0.078</td>
<td>0.480</td>
<td>0.050</td>
<td>0.106</td>
</tr>
<tr>
<td>Martins (2001) - full sample</td>
<td>2338</td>
<td>0.033</td>
<td>0.340</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>yrs of schooling &lt; 12</td>
<td>1999</td>
<td>-0.013</td>
<td>0.143</td>
<td>0.020</td>
<td>0.041</td>
</tr>
<tr>
<td>yrs of sch. &lt; 12, pot. exp. 20-30 yrs</td>
<td>732</td>
<td>-0.002</td>
<td>0.306</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>yrs of sch. &lt; 12, pot. exp. 20/21 yrs</td>
<td>165</td>
<td>-0.002</td>
<td>0.359</td>
<td>0.017</td>
<td>0.043</td>
</tr>
<tr>
<td>Chang (2011) - 1985 sample</td>
<td>1627</td>
<td>-0.049</td>
<td>0.247</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>white (W)</td>
<td>1282</td>
<td>-0.067</td>
<td>0.185</td>
<td>0.011</td>
<td>0.015</td>
</tr>
<tr>
<td>W, 12 yrs of schooling</td>
<td>609</td>
<td>-0.058</td>
<td>0.278</td>
<td>0.004</td>
<td>0.009</td>
</tr>
<tr>
<td>W, 12 yrs of s., age 30-35, recent job</td>
<td>89</td>
<td>-0.116</td>
<td>0.450</td>
<td>0.027</td>
<td>0.056</td>
</tr>
<tr>
<td>Wooldridge - full sample of mothers</td>
<td>31857</td>
<td>0.034</td>
<td>0.134</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>hispanic (H)</td>
<td>18897</td>
<td>0.021</td>
<td>0.064</td>
<td>0.059</td>
<td>0.110</td>
</tr>
<tr>
<td>H, &lt; 10 yrs of schooling</td>
<td>7085</td>
<td>-0.003</td>
<td>0.192</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>H, 12 yrs of schooling, age 25</td>
<td>341</td>
<td>-0.067</td>
<td>0.245</td>
<td>0.061</td>
<td>0.124</td>
</tr>
</tbody>
</table>

Note: Tests are based on 1999 bootstrap draws.

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 encounters 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 the hourly wage rate, which is only observed conditional on \(S = 1\). In contrast, the covariates \(X\) 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 suggest that IV validity is violated in the entire sample, as all p-values are smaller than 1%. Also the rather large standardized difference of 0.658 indicates that the point estimate of the mean outcome of the always selected is well outside its bounds. When conditioning on Malay ethnicity all tests remain significant at the 1% level. Considering the subsample with Malay ethnicity and a level of schooling equivalent to high school graduation, three out of four tests reject the null at the 5% level. The only exception is the probability-based test with full recentering (which is, however, less attractive than that with partial recentering as already mentioned in Section 3). Finally, we additionally restrict the sample to include only individuals in the age bracket 25-35 with 10-20 years of potential experience. The partially recentered mean test still rejects IV validity at the 5% level despite using just 56 observations, while the fully recentered method is just border line significant. However, the sample appears to be too small for the probability-based tests to work properly and give meaningful results. Both yield a p-value of 100%. 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 the Portuguese female labor market data of 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 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 covariates $X$ 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 for unconditional IV validity) shows that the null is rejected at any conventional level of significance. Restricting the data to individuals with a low level of education (less than 12 years) still entails rejections at the 5% or an even lower level of significance. Conditioning on particular brackets of potential education on top of the previous restrictions does not change the picture: Husband’s income appears to be an invalid instrument given the covariates 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 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. Furthermore, the data contain information on education, age, race, and the recent employment history as conditioning set \( X \). All test statistics are significant at the 1% level when applied to the entire sample and to the white subsample. The p-values of any test remain below 5% when also conditioning on (i) white ethnicity, (ii) white ethnicity and years of schooling and (iii) white ethnicity, years of schooling, the age bracket 30 to 35, and recent employment. Our results therefore suggest that IV validity is unlikely to hold in this data set.

The fourth application we consider is the instructional “LABSUP” data set on married mothers in the US provided by J. Wooldridge on his website with 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 1000 USD, which is only observed for the 18,789
individuals providing positive labor supply \((S = 1)\). The covariates \(X\) include ethnicity, years of schooling, and age. The instrument \(Z\) is non-wife income per year in 1000, the median of which is 29.399. Applied to the entire data, the tests again suggest hat IV validity does not hold. In the subsample of Hispanics, the probability-based tests reject the null at the 5% level while the partially recentered mean test does so at the 10% level. When also conditioning on poor education, only the fully recentered probability-based test yields a p-value larger than 5%, but the test statistic is still significant at the 10% level. Finally, we restrict the sample to 25 years old Hispanics with high school education. Three out of four tests reject the null at the 10% and one at the 5% level. Again, the results cast serious doubts on the validity of the instrument.

We now turn to our second instrumental variable to be investigated, (young) children in the household. We first reconsider the data of Martins (2001), who also uses kids under 3 as instrument. This time, however, none of our tests indicate that the exclusion restriction and/or monotonicity are violated. When examining IV validity in the entire sample and in 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.

![Table 5: Applications - young kids](image)

<table>
<thead>
<tr>
<th>Study</th>
<th>n</th>
<th>% comp.</th>
<th>st.dist</th>
<th>mean constraints</th>
<th>prob. constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p-val(p)</td>
<td>p-val(f)</td>
</tr>
<tr>
<td>Martins (2001) - full sample</td>
<td>2338</td>
<td>0.058</td>
<td>-0.097</td>
<td>0.467</td>
<td>0.984</td>
</tr>
<tr>
<td>yrs of schooling &lt; 12</td>
<td>1999</td>
<td>0.037</td>
<td>-0.051</td>
<td>0.572</td>
<td>0.957</td>
</tr>
<tr>
<td>yrs of s. &lt; 12, pot. exp. 20-30 yrs</td>
<td>732</td>
<td>-0.113</td>
<td>-0.247</td>
<td>0.742</td>
<td>0.999</td>
</tr>
<tr>
<td>yrs of s. &lt; 12, pot. exp. 20/21 yrs</td>
<td>165</td>
<td>-0.199</td>
<td>-0.382</td>
<td>0.881</td>
<td>0.997</td>
</tr>
<tr>
<td>MR (2008) - 1995-1999 married females</td>
<td>53966</td>
<td>-0.151</td>
<td>-0.351</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>high school graduate (hs)</td>
<td>18383</td>
<td>-0.130</td>
<td>-0.289</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>hsg, pot. exp. 20-30 yrs</td>
<td>7889</td>
<td>-0.107</td>
<td>-0.159</td>
<td>0.531</td>
<td>1.000</td>
</tr>
<tr>
<td>hsg, pot. exp. 20 yrs, south</td>
<td>239</td>
<td>-0.017</td>
<td>-0.021</td>
<td>0.466</td>
<td>0.559</td>
</tr>
<tr>
<td>Lee (2009) - female sample</td>
<td>4044</td>
<td>-0.046</td>
<td>-0.024</td>
<td>0.520</td>
<td>0.929</td>
</tr>
<tr>
<td>married, black (B)</td>
<td>2021</td>
<td>-0.050</td>
<td>0.002</td>
<td>0.477</td>
<td>0.793</td>
</tr>
<tr>
<td>mar., B, yrs of s. &lt; 12</td>
<td>1475</td>
<td>-0.058</td>
<td>-0.163</td>
<td>0.532</td>
<td>0.999</td>
</tr>
<tr>
<td>mar., B, yrs of s. &lt; 12, no recent job</td>
<td>699</td>
<td>-0.030</td>
<td>-0.294</td>
<td>0.666</td>
<td>0.987</td>
</tr>
<tr>
<td>Chang (2011) - 1985 sample</td>
<td>1627</td>
<td>-0.110</td>
<td>0.016</td>
<td>0.456</td>
<td>0.740</td>
</tr>
<tr>
<td>white (W)</td>
<td>1282</td>
<td>-0.134</td>
<td>0.043</td>
<td>0.243</td>
<td>0.456</td>
</tr>
<tr>
<td>W, 12 yrs of schooling</td>
<td>609</td>
<td>-0.126</td>
<td>0.040</td>
<td>0.387</td>
<td>0.657</td>
</tr>
<tr>
<td>W, 12 yrs of s., age 30-35, recent job</td>
<td>89</td>
<td>-0.034</td>
<td>0.219</td>
<td>0.166</td>
<td>0.320</td>
</tr>
</tbody>
</table>

Note: Tests are based on 1999 bootstrap draws.

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 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 between the ages of 25 and 54 in the second repeated cross section, in total 53,966 observations. The instrument \(Z\) to be tested is the incidence of children aged 0-6 in the household. The covariates \(X\) include education, potential work experience, the marital status, and regional dummies. As before, instrument validity is not rejected by any test in the entire sample. This result does not change when restricting the data to observations with (i) high school graduation, (ii) high school graduation and potential experience of 20 to 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 targets young individuals (aged 16-24 years) that have a legal residence in the U.S. and come from a low-income household, see Schochet, Burghardt, and Glazerman (2001) for further details. It provides 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 8 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 covariates \(X\). The instrument \(Z\) is the incidence of children in the household. We do not find violations of
the IV assumptions, 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. Our binary $Z$ indicates whether at least one child under 6 is present in the household. Relying on the same sample restrictions as before, all p-values are considerably larger than 10%. However, the standardized differences are positive throughout, indicating that one constraint is binding (albeit insignificantly so).

We conclude that the tests do not provide evidence against the IV validity of young children. In contrast, our results evoke serious concerns over the appropriateness of 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 it is an invalid instrument, at least conditional on the covariates commonly used in applications.

6 Conclusion

This paper has proposed a new test for instrumental validity in sample selection models. The latter commonly invoke two restrictions: (i) the instrument does not directly affect the outcome 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 distribution of the always selected, i.e., the subpopulation that is selected irrespective of the instrument. As the point must lie within the bounds, this provides us with two testable inequality constraints. Using tests based on a sequential bootstrap method proposed by Bennett (2009), we have considered eight empirical applications to verify the validity of two instruments that are often used in the estimation of female wages under sample selection: non-wife income and the number of young
children. Our results provide evidence that non-wife income is not a valid instrument, at least conditional on the covariates commonly available in such studies. However, the tests do not refute the IV validity of the number of children.
References


27


