Selection Without Exclusion[*]

Bo E. Honoré[†] Luojia Hu[‡]

November 15, 2019

Abstract

It is well understood that classical sample selection models are not semiparametrically identified without exclusion restrictions. Lee (2009) developed bounds for the parameters in a model that nests the semiparametric sample selection model. These bounds can be wide. In this paper, we investigate bounds that impose the full structure of a sample selection model with errors that are independent of the explanatory variables but have unknown distribution. The additional structure can significantly reduce the identified set for the parameters of interest. Specifically, we construct the identified set for the parameter vector of interest. It is a one-dimensional line segment in the parameter space, and we demonstrate that this line segment can be short in practice. We show that the identified set is sharp when the model is correct and empty when there exist no parameter values that make the sample selection model consistent with the data. We also provide non-sharp bounds under the assumption that the model is correct. These are easier to compute and associated with lower statistical uncertainty than the sharp bounds. Throughout the paper, we illustrate our approach by estimating a standard sample selection model for wages.

Key Word: Sample Selection, Exclusion Restrictions, Bounds, Partial Identification. JEL Code: C10, C14.

[*] This research was supported by the Gregory C. Chow Econometric Research Program at Princeton University and by the National Science Foundation (Grant Numbers SES-1530741 and SES-1824131). The opinions expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System. Leah Plachinski and Rachel Anderson provided excellent research assistance. We thank four anonymous referees, the editor and participants at numerous seminars for helpful comments. The most recent version of this paper can be found at http://www.princeton.edu/~honore/papers/SelectionWithoutExclusion.pdf.

[†] Mailing Address: Department of Economics, Julis Romo Rabinowitz Building, Princeton, NJ 08544. Email: honore@Princeton.edu.

[‡] Mailing Address: Economic Research Department, Federal Reserve Bank of Chicago, 230 S. La Salle Street, Chicago, IL 60604. Email: lhu@frbchi.org.

1 Introduction

This paper considers identification in the classical sample selection model (Heckman (1976))

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

where $y_i = y_i^*$ is observed if $w_i'\gamma + \nu_i \geq 0$. Early applications of the model assumed (ε_i, ν_i) is independent of (x_i, w_i), and distributed according to a bivariate normal distribution where both means are 0 and the variance of ν_i is 1. This allows one to estimate β (and γ) by maximum likelihood or by a two-step procedure. See Heckman (1979). Powell (1987) and others later considered semiparametric estimation of β under the assumption that (ε_i, ν_i) is independent of (x_i, w_i) but without the normality assumption. See, for example, Powell (1994). The key identifying assumption is that x_i must have full rank conditional on $w_i'\gamma$. This is essentially an exclusion restriction that requires that w_i include variables that do not enter in x_i. Ahn and Powell (1993) and Das, Newey, and Vella (2003) make a similar exclusion restriction assumption in more nonparametric settings.[1]

In this paper, we address the question of how much can be learned without an exclusion restriction like the one assumed in the literature discussed above[2]. We consider this important because it is often difficult to find variables that both matter for selection and can be credibly excluded from the main equation. For example, Krueger and Whitmore (2001) assumed normality and wrote, "Identification in these models is based on the assumption of normal errors, as there is no exclusion restriction." Lee (2009) and Krueger and Whitmore (2001) considered set identification in a sample selection model which contains (1) as a special case[3]. Unfortunately, these

[1] Escanciano, Jacho-Chvez, and Lewbel (2016) considered an identified sample selection model in which identification is essentially driven by nonlinearity. We consider our paper a complement to theirs.

[2] We focus on the case where $w_i'\gamma$ is bounded from above, since otherwise, one might use "identification at infinity" arguments to identify β.

[3] Manski (1989) constructed bounds in a model that is neither more general nor more restrictive than our setting. See also Manski (1990). Blundell, Gosling, Ichimura, and Meghir (2007) also constructed bounds in a sample selection model, but in a much more nonparametric setting than the one considered here.

sets are often too large to be informative. For example, Barrow and Rouse (2017) wrote, "Unfortunately, Lee Bounds estimates (Lee, 2009) are quite wide and largely uninformative." This is the motivation for this paper.

We first gain insights by studying the simplest case where the only explanatory variable is binary (Section 3). We demonstrate that in that model, the identified region for the parameter of interest can be quite small, and we provide conditions under which the upper or lower limits of the bound for the parameter coincide with the true parameter value. These results are then generalized to a model with a single potentially non-binary explanatory variable.

We next study the sample selection model with a more general set of explanatory variables in Section 4. We show that in this case, the identified set is one-dimensional. This observation is also implicit in Chamberlain (1986). Combining this insight with the results from Section 3, we then construct the identified set for the parameter vector. We show that if the model is correctly specified, our constructed identified region is sharp, and that it is empty when there exist no parameter values that make the sample selection model consistent with the data.

The population version of the identified set for β can be small enough to be empirically interesting. However, the characterization of the sharp identified set for β relies heavily on the whole distribution of y_i (conditional on selection). This will make estimation of the set based on a sample analog unattractive. We will therefore propose estimators of slightly larger sets.

Throughout the paper, we illustrate our approach by estimating a classical sample selection model for wages. We introduce this application in Section 2 and expand the analysis throughout the paper.

NOTATIONAL NOTE: Throughout this paper, we use f with a subscript letter to denote the density of that variable. If a variable, y, is subject to sample selection, i.e., it is observed with probability less than 1, f_y will integrate to the probability that y_i is observed. For the unobserved error terms, ε_i and ν_i, f_ε and f_ν denote the underlying densities and they each integrate to 1.

2 Empirical Illustration: Wages and Ethnicity

Throughout the paper, we use a simple sample selection model for log-wages to illustrate our approach. The emphasis will be on the effect of ethnicity on wages. Inspired by Mora (2008), we investigate the wage-differential between third-generation Mexican-Americans and other Americans after controlling for sample selection.

Like Mora (2008), we use CPS data on wages from Arizona, California, New Mexico and Texas. Our data spans the years 2003 to 2016, and contains $129,907$ women, of whom $26,698$ are third-generation Mexican-Americans and $103,209$ are non-Hispanic whites. There are $118,418$ men. Of them, $21,402$ are third-generation Mexican-Americans and $97,016$ are non-Hispanic whites. For women, the percentage working is 64% for third-generation Mexican-Americans and 61% for non-Hispanic whites. For men, the shares are 71% and 67%, respectively.

Summary statistics are provided in Table 1. Appendix 2 provides details about the data.

3 Simplest Case: Single Regressor

Consider first the simple case with a scalar binary explanatory variable, x_i:

$$y_i^* = x_i\beta + \varepsilon_i \tag{2}$$

where $y_i = y_i^*$ is observed if $x_i + \nu_i \geq 0$ and (ε_i, ν_i) is independent of x_i. When $x_i + \nu_i < 0$, y_i^* is not observed and y_i is undefined. The coefficient on the sample selection equation is only identified up to scale and its sign is identified. There is therefore no loss of generality by assuming that the coefficient on x_i is[4] 1. Our theorems below consider the more general setting, but since point-mass or limited support of the distribution of ε_i generally help with identification, it is useful to

[4] The exception is when the sample selection is independent of x. However, in that case, estimation of the coefficient on x will not suffer from sample selection bias. Moreover, one can identify whether selection is independent of x.

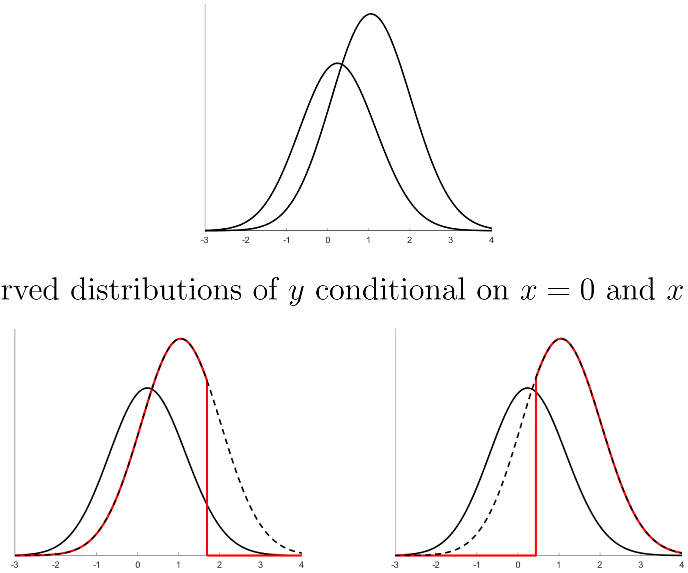
build intuition from the case where ε_i is continuously distributed with full support conditional on ν_i. We will implicitly assume random sampling, and occasionally drop the subscript i to aid readability.

Lee (2009) considers a more general sample selection model in which both the distribution of y_i and the probability of selection depend on selection in a nonparametric manner and the object of interest is the average effect of the "treatment", x_i. His main assumption is a monotonicity assumption that requires that any individual who is selected into the sample when $x_i = 0$ would also have been selected in a counterfactual scenario where $x_i = 1$. As such, he essentially considered the same model, but with (2) replaced by $y_i = h\left(x_i, \varepsilon_i\right)$ for some unknown h. The average treatment effect, $E\left[y_i^* \middle| x_i = 1\right] - E\left[y_i^* \middle| x_i = 0\right]$, is not identified in this case, but Lee constructed the sharp identified set for the parameter $E\left[y_i^* \middle| x_i = 1, s_i\right] - E\left[y_i^* \middle| x_i = 0, s_i\right]$ where s_i is the event that y_i would be observed whether $x_i = 0$ or $x_i = 1$. The bounds are based on the insight that an individual, for whom y_i is observed when $x_i = 0$, would also have had an observed y_i if x_i had been 1. On the other hand, some individuals would have observed y_i only when $x_i = 1$. This follows from his monotonicity assumption, and it implies that one must "trim" some observations from the distribution of y_i conditional on $x_i = 1$ in order to make it comparable to the distribution of y_i conditional on $x_i = 0$. The extreme cases are to trim the top and the bottom of the distribution of y_i conditional on $x_i = 1$.

Lee's bounds are illustrated graphically in Figure 1 for a data-generating process with $(\varepsilon_i, \nu_i)'$ distributed according to a bivariate normal distribution, $\beta = 1$, $E\left[\varepsilon_i\right] = 0$, $E\left[\nu_i\right] = \frac{1}{2}$, $\nu\left[\varepsilon_i\right] = 1$, $\nu\left[\nu_i\right] = 1$ and $\operatorname{cov}\left(\varepsilon_i, \nu_i\right) = \frac{1}{2}$. See also Example 1 below. The first panel displays the "densities" of y_i conditional on $x_i = 0$ and $x_i = 1$, respectively. They both integrate to the respective probabilities of selection. The second and third panels display the "density" of y_i conditional on $x_i = 0$ and the "density" of y_i conditional on $x_i = 1$ after being trimmed at the top or at the bottom.

The sample selection model considered here implies the monotonicity assumption in Lee (2009). If y_i is observed when $x_i = 0$, then ν_i must be greater than 0, as a result, y_i will also be observed for the same draw of ν_i when $x_i = 1$. Hence the

Figure 1: Construction of Lee Bounds for Data-Generating Process in Example 1



Observed distributions of y conditional on $x = 0$ and $x = 1$.

Distributions after trimming according to Lee (2009).

sample selection model (2) is the version of Lee's setup in which the treatment effect is constant, and Lee's bounds can be thought of as non-sharp bounds on β.

To illustrate the approach in this paper, it is useful to define a binary variable for whether y_i is observed, $d_i = 1\left\{x_i + \nu_i \geq 0\right\}$. For all $c_1 < c_2$, we then have

$$P\left(c_1 < \varepsilon_i \leq c_2, d_i = 1 \middle| x_i = 0\right) \leq P\left(c_1 < \varepsilon_i \leq c_2, d_i = 1 \middle| x_i = 1\right) \tag{3}$$

or

$$P\left(c_1 < y_i \leq c_2, d_i = 1 \middle| x_i = 0\right) \leq P\left(c_1 < y_i - \beta \leq c_2, d_i = 1 \middle| x_i = 1\right). \tag{4}$$

When the errors are continuously distributed, the restriction (4) can be expressed in terms of the density of the observed y conditional on x_i. Define

$$f_y\left(c \middle| x_i\right) = f_{y^*}\left(c \middle| x_i\right) P\left(d_i = 1 \middle| y_i = c, x_i\right).$$

This is the "density" of the observed y_i, except that it does not integrate to 1 because

y_i is not observed when $d_i = 0$. With this notation, (4) can be expressed as[5]

$$f_y\left(c|\, x_i = 0\right) \leq f_y\left(c + \beta|\, x_i = 1\right) \tag{5}$$

for all values of c.

Equation (5) is illustrated in Figure 2 using the same data-generating process as above. The contour plot in the left panel shows the joint distribution of (ε_i, ν_i) before selection and the solid line in the right panel depicts the corresponding marginal distribution of ε_i. The selection implies that y_i^* is not observed when $\nu_i \leq -x_i$. For $x_i = 0$ and $x_i = 1$, this means that we "lose" the errors below the solid lines in the left panel of Figure 2. The dashed lines in the right panel of Figure 2 show the "density" of the remaining ε's. These densities integrate to the probability that y_i^* is observed conditional on x_i.

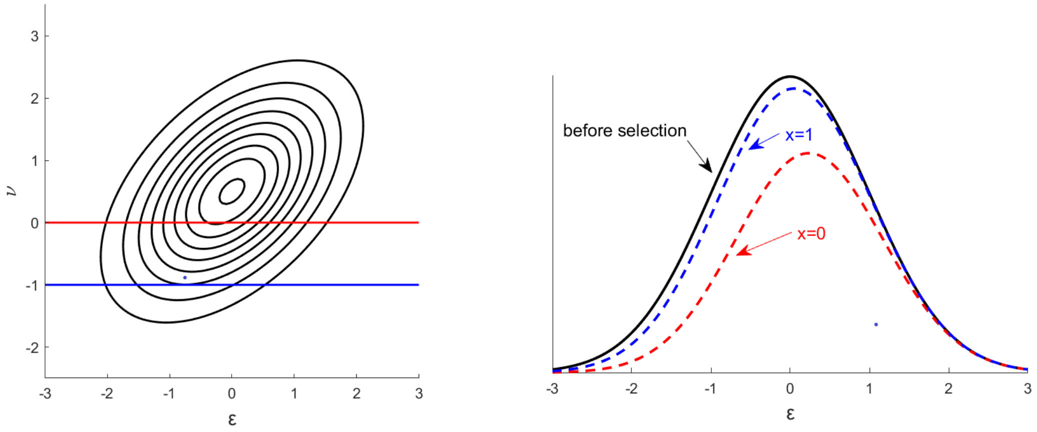
This logic relies on the full structure of the classical sample selection model. This includes independence between the explanatory variable and the error, monotone selection and parameter homogeneity. In Figure 2(b), the distribution of ε before selection needs to be the same whether $x_i = 0$ or $x_i = 1$. This would be violated without independence between ε_i and x_i. Without monotone selection, we could not conclude equation (3). Finally, it is the parameter homogeneity that allows us to translate the statement about the unobserved ε_i in equation (3) into a statement about the observed y_i in equation (4). In contrast, Lee (2009) only assumed monotone selection, which is why our identified region is smaller than his.

The following theorem establishes that the inequalities in equation (4) contain all the available information. As a result they can be used to construct the identified region for β.

Theorem 1 *Let x_i be a scalar, non-degenerate, binary random variable, and let (ε_i, ν_i) be independent of x_i. If $y_i^* = x_i\beta + \varepsilon_i$ and if $y_i = y_i^*$ is observed when*

[5] This is reminiscent of the insight in Kitagawa (2015) who creates a test for instrument validity based on whether one product of a density and a probability lies above a second product of a density and a probability at all points. To map our insight into his, we would have to think of (a) his outcome as our $y - x\beta$, (b) his instrument as our x, and (c) his treatment as our selection dummy.

Figure 2: Distribution of ε Before and After Selection



(a) Joint Distribution of (ε, ν)

(b) Distribution of ε

$d_i = 1\{x_i + \nu_i \geq 0\}$ *equals one, then the identified region for* β *is*

$$\mathbf{B} = \{b \in \mathbb{R} : P(c_1 < y_i \leq c_2, d_i = 1|\, x_i = 0) \\ \leq P(c_1 < y_i - b \leq c_2, d_i = 1|\, x_i = 1) \text{ \textit{for all values of} } c_1, c_2\}$$

provided that $P(d_i = 1|\, x_i = 1) > 0$.

Proof. This is a special case of Theorem 3 below. The proof here is more readable. The discussion above established that the true β belongs to $\mathbf{B}$. We will now argue that for any b in $\mathbf{B}$, there exists a joint (cumulative) distribution[6] G of (ε, ν) such that (G, b) will be consistent with the observed distribution of (y, d) given x. First, define the marginal distribution of ε by

$$\widetilde{F}_{\varepsilon}(a) = \frac{P(y \leq a + b, d = 1|x = 1)}{P(d = 1|x = 1)}.$$

[6] For ease of exposition, we have dropped the subscript i in the proof.

Next, define a conditional distribution function of ν given ε by[7]

$$\widetilde{F}_\nu\left(-1|\varepsilon \leq a\right) = 1 - P\left(d = 1|x = 1\right)$$

and

$$\widetilde{F}_\nu\left(0|\varepsilon \leq a\right) = 1 - \frac{P\left(y \leq a, d = 1|x = 0\right)}{\widetilde{F}_\varepsilon\left(a\right)} = 1 - \frac{P\left(y \leq a, d = 1|x = 0\right)}{P\left(y \leq a + b, d = 1|x = 1\right)} P\left(d = 1|x = 1\right)$$

when $\widetilde{F}_\varepsilon\left(a\right) > 0$, and $\widetilde{F}_\nu\left(0|\varepsilon \leq a\right) = 1$ when $\widetilde{F}_\varepsilon\left(a\right) = 0$.

This construction defines a cumulative distribution function if $\widetilde{F}_{\varepsilon,\nu}\left(a_1, 0\right) + \widetilde{F}\left(a_0, -1\right) - \widetilde{F}\left(a_1, -1\right) - \widetilde{F}\left(a_0, 0\right) \geq 0$ for all $a_0 < a_1$ (Durrett (2019), Theorem 1.1.11). It follows immediately from the expressions above that

$$\begin{aligned}
&\widetilde{F}_{\varepsilon,\nu}\left(a_1, 0\right) + \widetilde{F}\left(a_0, -1\right) - \widetilde{F}\left(a_1, -1\right) - \widetilde{F}\left(a_0, 0\right) \\
&\qquad = P\left(a_0 < y - b \leq a_1, d = 1|x = 1\right) - P\left(a_0 < y \leq a_1, d = 1|x = 0\right)
\end{aligned}$$

when $\widetilde{F}_\varepsilon\left(a_0\right) > 0$; it is $P\left(-\infty < y - b \leq a_1, d = 1|x = 1\right) - P\left(-\infty < y \leq a_1, d = 1|x = 0\right)$ when $\widetilde{F}_\varepsilon\left(a_0\right) = 0$ and $\widetilde{F}_\varepsilon\left(a_1\right) > 0$; finally, it is 0 when $\widetilde{F}_\varepsilon\left(a_0\right) = 0$ and $\widetilde{F}_\varepsilon\left(a_1\right) = 0$. Hence $\widetilde{F}_{\varepsilon,\nu}$ satisfies the conditions for a cumulative distribution function if (and only if) b belongs to $\mathbf{B}$.

With this $\left(\widetilde{F}_{\varepsilon,\nu}, b\right)$,

$$\begin{aligned}
\widetilde{P}\left(y \leq c, d = 1|\, x = 1\right) &= \widetilde{F}_\varepsilon\left(c - b\right)\left(1 - \widetilde{F}_\nu\left(-1|\varepsilon \leq c - b\right)\right) \\
&= \frac{P\left(y \leq c, d = 1|x = 1\right)}{P\left(d = 1|x = 1\right)} P\left(d = 1|x = 1\right) = P\left(y \leq c, d = 1|x = 1\right)
\end{aligned}$$

[7] It does not matter what the conditional distribution function of ν given ε is at points other than -1 and 0.

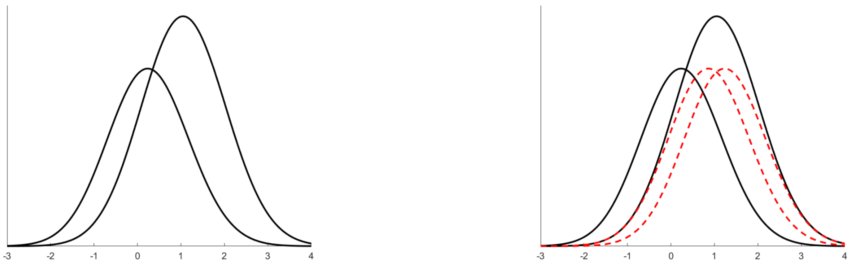
and

$$\begin{aligned}
& \widetilde{P}\left(y \leq c, d=1 \middle| x=0\right) \\
= \; & \widetilde{F}_{\varepsilon}(c)\left(1-\widetilde{F}_{\nu}(0|\varepsilon \leq c)\right) \\
= \; & \frac{P\left(y \leq c+b, d=1|x=1\right)}{P\left(d=1|x=1\right)} \frac{P\left(y \leq c, d=1|x=0\right)}{P\left(y \leq c+b, d=1|x=1\right)} P\left(d=1|x=1\right) \\
= \; & P\left(y \leq c, d=1|x=0\right).
\end{aligned}$$

This proves the theorem. ■

The construction of the identified region in equation (5) is illustrated graphically in Figure 3. The left side of Figure 3 shows the density of the observed y conditional on x multiplied by the conditional probability of selection for x_i equal to 0 and 1. Theorem 1 characterized the identified region for β as the length of the horizontal shifts of one of the curves that will result in one of the curves being above the other. This is illustrated in the right hand side of Figure 3.

Figure 3: Illustration of Bounds Based on Equation (5).



(a) Observed Distributions

(b) Observed Distributions Shifted by b in Identified Region

Example 1 *Let* $(\varepsilon_i, \nu_i)'$ *be distributed according to a bivariate normal distribution with* $\beta = 1$, $E[\varepsilon_i] = 0$, $E[\nu_i] = \frac{1}{2}$, $\nu[\varepsilon_i] = 1$, $\nu[\nu_i] = 1$ *and* $\operatorname{cov}(\varepsilon_i, \nu_i) = \frac{1}{2}$. *With these* $P\left(d_i = 1 \middle| x_i = 0\right) = 0.691$ *and* $P\left(d_i = 1 \middle| x_i = 1\right) = 0.933$. *This is the situation depicted in Figure 3 and the identified region for* β *is* $[0.626, 1.00]$. *In contrast, the Lee bounds are*[8] $[0.389, 1.238]$.

[8] To calculate the bounds, we use equation (5) in Muthén (1990) after correcting a typo in the second line (the next to last subscript-i should be subscript-j).

To estimate the identified region characterized by Theorem 1, one needs to compare the estimated probabilities for all pairs of (c_1, c_2). This is clearly impossible. Moreover, when c_1 and c_2 are close, this is akin to comparing estimated density of the observed y_i conditional on x_i for $x_i = 0$ and $x_i = 1$. This is troublesome because the probabilities will typically both be close to 0 in the tails, and small estimation errors will have a big effect on which one takes on the larger value. This suggests constructing an identified region by exploring (4) for a finite number of pairs of (c_1, c_2). For example, one could calculate the deciles of the observed y conditional on $x_i = 0$ and then use $(c_1, c_2) = (q_{j-1}, q_j)$ for $j = 1, ..., 10$, where q_j is the j-th decile, $q_0 = -\infty$ and $q_{10} = \infty$.

Example 2 *(Example 1 continued) In this setup the crude bounds described above are* $[0.609, 1.025]$.

In Example 1, the upper bound of the identified set equals the true β. This is true in general when the true (unknown) distribution of the errors is a bivariate normal with positive correlation.

Proposition 1 *When the distribution of the errors is bivariate normal with positive correlation, the upper limit of the identified region based on equation (5) is the true parameter value. When the correlation is negative, the lower limit of the identified region is the true value. With no selection, i.e., independence of the errors, the identified region is the true value.*

Proof. See Appendix 1. ■

The proof of Proposition 1 is driven by the tail behavior of the normal distribution. As a result, the proposition can be generalized to

Proposition 2 *Suppose that ε is continuously distributed with a density which has sufficiently thin tails that for $a > 0$, $f(c)/f(c+a) \to \infty$ as $c \to \infty$ and for $a < 0$, $f(c)/f(c+a) \to \infty$ as $c \to -\infty$. Then*

1. *If the distribution of ν given $\varepsilon = c_1$ stochastically dominates the distribution of ν given $\varepsilon = c_2$ whenever $c_1 > c_2$, then the upper limit of the identified region is the true value.*

2. *If the distribution of ν given $\varepsilon = c_1$ stochastically dominates the distribution of ν given $\varepsilon = c_2$ whenever $c_1 < c_2$, then the lower limit of the identified region is the true value.*

3. *If ν and ε are independent, then the identified region is the true value.*

Proof. See Appendix 1. ■

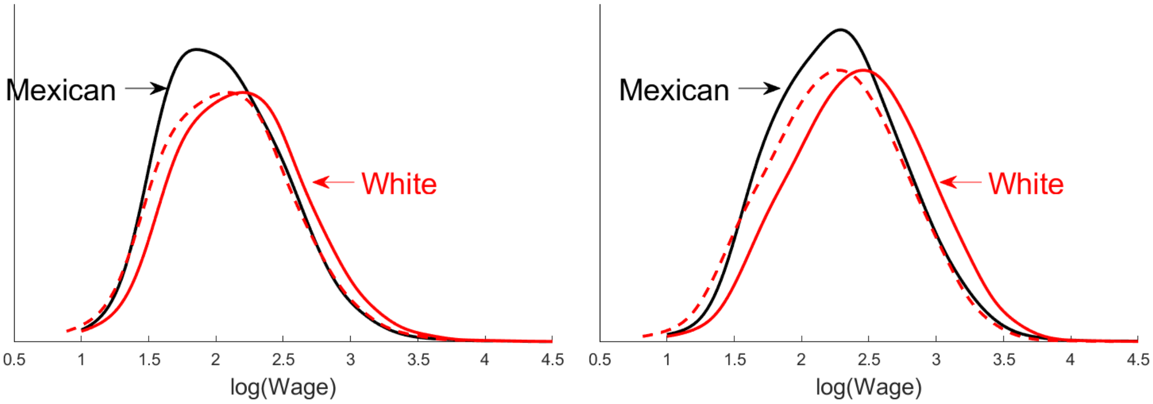
The assumption on the tail behavior of the marginal distribution of ε is slightly stronger than log-concavity and is implied by tail-behavior of the form $\exp(-ax^{\gamma})$ for $a > 0$ and $\gamma > 1$. We interpret the stochastic dominance assumption in 1 as positive selection: larger values of ε are associated with higher probability of selection. Likewise, we interpret the stochastic dominance assumption in 2 as negative selection. The setup in Proposition 2 is different from, but similar in spirit to, the approach in Heckman (1990) and Andrews and Schafgans (1998). Both rely on "identification at infinity," but while Heckman (1990) and Andrews and Schafgans (1998) need an exclusion restriction and rely on extreme values of the selection index, we do not need exclusion restrictions and rely on extreme values of the outcome variable.

3.1 Empirical Illustration Part 1

In this section, we illustrate the insights above graphically. Using the data described in Section 2, we plot the "densities" (the product of the density conditional on selection and the probability of selection) of log-wages for Mexican-Americans and Non-Hispanic white Americans by gender. We restrict the sample to individuals whose highest degree is high school. These are depicted by the solid lines in Figure 4. The areas under the Mexican-American curves are larger than for Non-Hispanic white Americans because the former are more likely to work for pay.

The dashed lines are the curves for the Non-Hispanic whites shifted by -0.11 for women and by -0.18 for men. The shifted curves for whites almost fit under the curves for the Mexican-Americans. This suggests that the assumptions of the classical sample selection model are not too unreasonable in this case. Moreover, it is clear from the figure that shifting the curves by a lot more or a lot less would lead to violations of equation (5). This suggests that the identified regions are relatively small. In this case, the Lee bounds for the log-wage differentials between Mexican-Americans and Non-Hispanic white Americans are $(-0.210, -0.041)$ for women and $(-0.249, -0.074)$ for men while the difference in means of the observed data are -0.123 for women and -0.162 for men.

Figure 4: Shifted and Unshifted Log-Wage Distributions. The dashed lines display the White log-wage densities shifted by -0.11 for women and by -0.18 for men.



(a) Women

(b) Men

Heckman's two-step estimator exploits variation in the conditional mean of the dependent variable. When the only explanatory variable is binary, there will be perfect collinearity between it and the sample selection correction term. The procedure therefore cannot be applied. In contrast, the maximum likelihood estimator for the log-wage differentials between Mexican-Americans and Non-Hispanic white Americans exploits information from the entire distribution of the dependent variable, and this estimator is therefore in principle applicable, although it is likely to be fragile.

For example, when we tried to estimate the model using Stata[9], the routine failed to converge for the men and located a coefficient on Mexican-American to be -0.174 for women and -0.208 for men.

3.2 Single Non-Binary Regressor

Theorem 1 applies to the case where x_i is binary. When x_i is not binary and unbounded from above, identification at infinity arguments like those in Andrews and Schafgans (1998) and Heckman (1990) yield point-identification of β. We therefore focus on the case where x_i is bounded from above.

When x_i is not binary and bounded from above, applying (4) to all pairs of values in the support of x_i yields bounds on the identified region of β. The following theorem establishes that the intersection of these bounds is sharp.

Theorem 2 *Let $(x_i, \varepsilon_i, \nu_i)$ be a random vector such that (ε_i, ν_i) is independent of x_i, and (ε_i, ν_i) has continuous and everywhere positive density. If $y_i = x_i\beta + \varepsilon_i$, $y_i = y_i^*$ is observed when $d_i = 1\{x_i + \nu_i \geq 0\}$ equals one, and the upper bound on the support of x_i is $x_{\max} < \infty$, then the identified region for β is* [10]

$$\mathbf{B} = \{b \in \mathbb{R} : P\left(c_1 < y_i \leq c_2, d_i = 1 | \, x_i = \xi_1\right) \leq P\left(c_1 < y_i - b \leq c_2, d_i = 1 | \, x_i = \xi_2\right)$$
$$\textit{for all values of } c \textit{ and } \xi_1 < \xi_2 \textit{ in the support of } x_i\}.$$

Proof. Follows from Theorem 3 below. ■

As discussed above, the identified set can also be expressed in terms of densities,

$$\mathbf{B} = \left\{b \in \mathbb{R} : f_{y_i|x_i}\left(c + x_i b | \, \xi_1\right) \leq f_{y_i|x_i}\left(c + x_i b | \, \xi_2\right)\right.$$
$$\left.\text{for all values of } c \text{ and } \xi_1 < \xi_2 \text{ in the support of } x_i\right\}$$

provided that these densities are well-defined.

[9] More precisely, the routine `heckman` in Stata Version 14 with all the default options.

[10] Recall that $f_{y_i|x_i}$ does not integrate to 1, since y_i is not always observed.

We finally note that the conclusion of Proposition 2 carries over to general distribution of x_i. Specifically,

Proposition 3 *Let x_i be a random variable and let (ε_i, ν_i) be independent of x_i. Assume that (ε_i, ν_i) has continuous and everywhere positive density. Also assume that $y_i = x_i\beta + \varepsilon_i$, that $y_i = y_i^*$ is observed when $d_i = 1\{x_i + \nu_i \geq 0\}$ equals one, and that the upper bound on the support of x_i is $x_{\max} < \infty$. If the density of ε has sufficiently thin tails that for $a > 0$, $f(c)/f(c+a) \to \infty$ as $c \to \infty$ and for $a < 0$, $f(c)/f(c+a) \to \infty$ as $c \to -\infty$, then*

1. *If the distribution of ν given $\varepsilon = c_1$ stochastically dominates the distribution of ν given $\varepsilon = c_2$ whenever $c_1 > c_2$, then the upper limit of the identified region for β is the true value.*

2. *If the distribution of ν given $\varepsilon = c_1$ stochastically dominates the distribution of ν given $\varepsilon = c_2$ whenever $c_1 < c_2$, then the lower limit of the identified region for β is the true value.*

3. *If ν and ε are independent, then the identified region for β is the true value.*

Proof. See Appendix 1. ■

Needless to say, all of the analysis in this section could be performed conditional on a set of covariates, x_2, in which case the bounds derived here would become bounds on the effect of x conditional on x_2. In the next section, we investigate the alternative approach of explicitly incorporating additional explanatory variables in the standard selection model.

4 More General Sample Selection Model

We now return to the sample selection model with a k–dimensional vector of explanatory variables, x_i,

$$y_i^* = x_i'\beta + \varepsilon = x_{i1}\beta_1 + x_{i2}'\beta_2 + \varepsilon_i \tag{6}$$

where $y_i = y_i^*$ is observed if $x_i'\gamma + \nu_i \geq 0$. When the support of $x_i'\gamma$ is unbounded from above, identification at infinity arguments like those in Andrews and Schafgans (1998) and Heckman (1990) can yield point-identification of β. We therefore focus on the case where $x_i'\gamma$ is bounded from above.

To fix ideas, suppose that β_1 is the parameter of interest.

Conditions under which γ is identified up to scale are well-understood; see for example Powell (1994) and the references therein. In the following, we assume that these conditions hold and that the necessary scale normalization has been imposed by normalizing the first element of γ to be[11] 1. We will then write $\gamma = (1, \gamma_2')'$ to distinguish between the variable of interest, x_{i1}, and the other explanatory variables. As in the previous section, we assume independence between (ε_i, ν_i) and (x_{i1}, x_{i2}), and we define $g(z) = E[\varepsilon_i | \nu_i > -z, x_{i1}, x_{i2}]$. We can then write

$$y_i = x_{i1}\beta_1 + x_{i2}'\beta_2 + g(x_i'\gamma) + u_i \tag{7}$$

with $E[u_i | x_{i1}, x_{i2}] = 0$.

In this section, we will argue that the vector β is identified except for a single scale parameter. In the following subsections we will then show that bounds can be obtained for this parameter. The intuition for why β is identified except for a single scale parameter is very simple. Suppose we knew β_1. We could then define $w_i^* = y_i^* - x_{i1}\beta_1 = x_{i2}'\beta_2 + \varepsilon_i$. The variable x_{i1} would then be excluded from the model for w_i^*. On the other hand, by normalizing the first coefficient in the selection equation to be 1, we have already assumed that x_{i1} matters for selection. Hence we have the necessary exclusion restriction, and the parameter vector, β, is identified except for the one-dimensional component β_1. Here, we give a slightly different argument because it makes the empirical implementation easier.

[11] This implicitly rules out that all coefficients in the selection equation are 0. However, in that case, estimation of β does not suffer from sample selection bias. Moreover, one can identify whether selection is independent of x.

Following, for example Robinson (1988), we start by noting that (7) implies that

$$y_i - E\left[y_i | x_i'\gamma\right] = \left(x_{i1} - E\left[x_{i1} | x_i'\gamma\right]\right)\beta_1 + \left(x_{i2} - E\left[x_{i2} | x_i'\gamma\right]\right)'\beta_2 + u_i. \tag{8}$$

Next note that

$$\begin{aligned}\left(x_{i1} - E\left[x_{i1} | x_i'\gamma\right]\right) + \left(x_{i2} - E\left[x_{i2} | x_i'\gamma\right]\right)'\gamma_2 &= \left(x_{i1} - E\left[x_{i1} | x_i'\gamma\right]\right) + \left(x_{i2}\gamma_2 - E\left[x_{i2} | x_i'\gamma\right]'\gamma_2\right) \\ &= x_i'\gamma - E\left[x_i'\gamma | x_i'\gamma\right] = 0.\end{aligned}$$

In other words, $\left(x_{i1} - E\left[x_{i1} | x_i'\gamma\right]\right) = -\left(x_{i2} - E\left[x_{i2} | x_i'\gamma\right]\right)'\gamma_2$. Equation (8) can then be written as

$$\begin{aligned}y_i - E\left[y_i | x_i'\gamma\right] &= \left(-\left(x_{i2} - E\left[x_{i2} | x_i'\gamma\right]\right)'\gamma_2\right)\beta_1 + \left(x_{i2} - E\left[x_{i2} | x_i'\gamma\right]\right)'\beta_2 + u_i \qquad (9) \\ &= \left(x_{i2} - E\left[x_{i2} | x_i'\gamma\right]\right)'\left(\beta_2 - \gamma_2\beta_1\right) + u_i = \left(x_{i2} - E\left[x_{i2} | x_i'\gamma\right]\right)'\alpha_2 + u_i\end{aligned}$$

where $\alpha_2 = \left(\beta_2 - \gamma_2\beta_1\right)$. We can therefore identify $\alpha_2 = \left(\beta_2 - \gamma_2\beta_1\right)$ subject to a rank condition on $\left(x_{i2} - E\left[x_{i2} | x_i'\gamma\right]\right)$. Since γ_2 is identified, this implies that for a given value of β_1, β_2 is identified. In other words, the identification problem is essentially one-dimensional, and bounds on β_1 will imply bounds of the whole β vector.

4.1 Sharp Bounds

With the result of the previous section, we can write

$$\begin{aligned}y_i^* &= x_{i1}\beta_1 + x_{i2}'\beta_2 + \varepsilon_i \\ &= x_{i1}\beta_1 + x_{i2}'\left(\alpha_2 + \gamma_2\beta_1\right) + \varepsilon_i\end{aligned}$$

or

$$y_i^* - x_{i2}'\alpha_2 = \left(x_{i1} + x_{i2}'\gamma_2\right)\beta_1 + \varepsilon_i \tag{10}$$

where γ_2 and α_2 are identified as above and $y_i = y_i^*$ (and hence $y_i - x_{i2}'\alpha_2 = y_i^* - x_{i2}'\alpha_2$) is observed when $d_i = 1\left\{x_{i1} + x_{i2}'\gamma_2 + \nu_i > 0\right\}$. We can then apply Theorem 2 to bound β_1 to the region

$$\begin{aligned}\mathbf{B} = \{b_1 \in \mathbb{R} : & P\left(c_1 < y_i - x_{i2}'\alpha_2 - \left(x_{i1} + x_{i2}'\gamma_2\right) b_1 \leq c_2, d_i = 1 \middle| x_i'\gamma = \xi_1\right) \\ & \leq P\left(c_1 < y_i - x_{i2}'\alpha_2 - \left(x_{i1} + x_{i2}'\gamma_2\right) b_1 \leq c_2, d_i = 1 \middle| x_i'\gamma = \xi_2\right) \\ & \text{for all values of } c_1 < c_2 \text{ and } \xi_1 \leq \xi_2 \text{ in the support of } x_i'\gamma\}.\end{aligned} \tag{11}$$

The identified region for the whole vector, β, is then the one-dimensional line segment

$$\left\{\left(\begin{array}{c} b_1 \\ \alpha_2 + \gamma_2 b_1 \end{array}\right) : b_1 \in \mathbf{B}\right\}.$$

The identified region can also be written in terms of the density of the observed data:

$$\begin{aligned}\{b_1 \in \mathbb{R} : & f_{y|x}\left(c + x_{i2}'\alpha_2 + \left(x_{i1} + x_{i2}'\gamma_2\right) b_1 \middle| x_i'\gamma = \xi_1\right) \\ & \leq f_{y|x}\left(c + x_{i2}'\alpha_2 + \left(x_{i1} + x_{i2}'\gamma_2\right) b_1 \middle| x_i'\gamma = \xi_2\right) \\ & \text{for all values of } c \text{ and all } \xi_1 \leq \xi_2 \text{ in the support of } x_i'\gamma\}\end{aligned} \tag{12}$$

provided that these densities are well-defined.

The bounds implied by $\mathbf{B}$ are sharp by Theorem 3.

Theorem 3 *Suppose that (i) (ε_i, ν_i) is independent of x_i, (ii) $E[\varepsilon_i | \nu_i > a]$ is finite for all a, (iii) there is no proper linear subspace of R^k that contains x_i with probability 1, and (iv)*

$$y_i^* = x_i'\beta + \varepsilon_i$$

is observed if $d_i = 1\{x_i'\gamma + \nu_i \geq 0\}$ equals one for some β and some γ with $\gamma_1 = 1$. If $P(d = 1) > 0$, γ is identified and the support of $x_i'\gamma$ is bounded from above, then $\mathbf{B}$ is the (sharp) identified region for β_1.

Proof. The discussion in the text established that the true β_1 belongs to $\mathbf{B}$ and that $\beta_2 = \alpha_2 + \gamma_2\beta_1$. We next need to argue that for any $b = (b_1, \alpha_2' + \gamma_2' b_1)'$ with b_1 in

$\mathbf{B}$, there exists a joint distribution, $\widetilde{F}$, of (ε, ν) such that the distribution implied by the model combined with $\left(\widetilde{F}, b\right)$ is the same as the observed distribution of (y, d) conditional on $x = \xi$ for all ξ in the support of x.

Let $\overline{x}$ be such that $\overline{x}'\gamma$ is the upper bound of the support of $x'\gamma$.

First, define the marginal distribution of ε by

$$\widetilde{F}_{\varepsilon}(a) = \frac{P\left(y \leq a + \overline{x}'b, d = 1 | x = \overline{x}\right)}{P\left(d = 1 | x = \overline{x}\right)}.$$

The definition of $\mathbf{B}$ guarantees that this gives the same $\widetilde{F}_{\varepsilon}(a)$ for all choices of $\overline{x}$ such that $\overline{x}'\gamma$ is the upper bound of the support of $x'\gamma$. The assumption that $P(d = 1) > 0$ guarantees that the denominator is non-zero. Next, we define the conditional cumulative distribution function of ν given ε over the support of $-x'\gamma$. Let g be a point in that support and let x_g be such that $x_g'\gamma = -g$. We then define

$$\begin{aligned}\widetilde{F}_{\nu}(g | \varepsilon \leq a) = 1 - \frac{P\left(y \leq a + x_g'b, d = 1 | x = x_g\right)}{\widetilde{F}_{\varepsilon}(a)}& \\ = 1 - \frac{P\left(y \leq a + x_g'b, d = 1 | x = x_g\right)}{P\left(y \leq a + \overline{x}'b, d = 1 | x = \overline{x}\right)} P\left(d = 1 | x = \overline{x}\right)&\end{aligned}$$

when $\widetilde{F}_{\varepsilon}(a_0) > 0$, and 1 otherwise. The definition of $\mathbf{B}$ guarantees that this gives the same $\widetilde{F}_{\nu}(g | \varepsilon \leq a)$ for all choices of x_g such that $x_g'\gamma = -g$.

This construction defines a cumulative distribution function if $\widetilde{F}(a_1, g_1) + \widetilde{F}(a_0, g_0) - \widetilde{F}(a_1, g_0) - \widetilde{F}(a_0, g_1) \geq 0$ for all $a_0 < a_1$ and $g_0 < g_1$ (Durrett (2019), Theorem 1.1.11). It follows immediately from the expressions above that

$$\begin{aligned}&\widetilde{F}(a_1, g_1) + \widetilde{F}(a_0, g_0) - \widetilde{F}(a_1, g_0) - \widetilde{F}(a_0, g_1) \\ &= P\left(a_0 < y \leq a_1, d = 1 | x = x_{g_0}\right) - P\left(a_0 < y \leq a_1, d = 1 | x = x_{g_1}\right)\end{aligned}$$

when $\widetilde{F}_{\varepsilon}(a_0) > 0$ and $\widetilde{F}_{\varepsilon}(a_1) > 0$. Since $x_{g_0}'\gamma > x_{g_1}'\gamma$, this is non-negative by the definition of **B**. Also,

$$\begin{aligned}
&\widetilde{F}(a_1, g_1) + \widetilde{F}(a_0, g_0) - \widetilde{F}(a_1, g_0) - \widetilde{F}(a_0, g_1) \\
&= P\left(-\infty < y \le a_1 + x'_{g_0}b, d = 1|x = x_{g_0}\right) - P\left(-\infty < y \le a_1 + x'_{g_1}b, d = 1|x = x_{g_1}\right)
\end{aligned}$$

when $\widetilde{F}_\varepsilon(a_0) = 0$ and $\widetilde{F}_\varepsilon(a_1) > 0$. This is again non-negative by the definition of **B**. Finally, $\widetilde{F}(a_1, g_1) + \widetilde{F}(a_0, g_0) - \widetilde{F}(a_1, g_0) - \widetilde{F}(a_0, g_1) = 0$ when $\widetilde{F}_\varepsilon(a_0) = 0$ and $\widetilde{F}_\varepsilon(a_1) = 0$. Hence $\widetilde{F}$ defines a bivariate cumulative distribution function.

With this $\left(\widetilde{F}, b\right)$, the model yields

$$\begin{aligned}
& \widetilde{P}(y \le c, d = 1|\, x = x_0) \\
= \;& \widetilde{F}_\varepsilon(c - x'_0 b)\left(1 - \widetilde{F}_\nu(-x'_0\gamma | \varepsilon \le c - x'_0 b, x = x_0)\right) \\
= \;& \frac{P(y \le c - x'_0 b + \overline{x}' b, d = 1|x = \overline{x})}{P(d = 1|x = \overline{x})} \frac{P(y \le c - x'_0 b + x'_0 b, d = 1|x = x_0)}{P(y \le c - x'_0 b + \overline{x}' b, d = 1|x = \overline{x})} P(d = 1|x = \overline{x}) \\
= \;& P(y \le c, d = 1|x = x_0).
\end{aligned}$$

This proves the theorem. ■

Theorem 3 states that **B** characterizes the identified region for β_1 when the model is correct. Theorem 4 below establishes that when **B** is not empty, the linear sample selection model cannot be rejected by the data.

Theorem 4 *Suppose that the data-generating process for the observed distribution of (d_i, y_i, x_i) is such that (i) y_i is only observed when $d_i = 1$, (ii) $P(d_i = 1|\, x_i)$ can be written as a non-decreasing, right-continuous function of $x'_i\gamma$ for some $\gamma = (1, \gamma'_2)'$, (iii) the support of $x'_i\gamma$ is bounded from above, and (iv) the density of y_i given x_i is positive everywhere for all x_i in the support of x_i.*

*If, for α_2 defined in (9), **B** in (11) is not empty, then for every β in **B**, there exists a distribution of (ε_i, ν_i) such that the observed distribution is the same as the one generated from a model in which (ε_i, ν_i) is independent of x_i,*

$$y_i^* = x'_i\beta + \varepsilon_i$$

and $y_i = y_i^*$ is observed if $d_i = 1\{x_i'\gamma + \nu_i \geq 0\}$ equals 1.[12]

Proof. Let b_1 be an element of $\mathbf{B}$, $b_2 = \alpha_2 + \gamma_2 b_1$ and $b = (b_1, b_2')'$. As in the proof of Theorem 3, we need to argue that there exists a joint distribution, $\widetilde{F}$, of (ε, ν) such that the distribution implied by the model combined with $\left(\widetilde{F}, b\right)$ is the same as the observed distribution of (y, d) conditional on $x = \xi$ for all ξ in the support of x. This is done exactly as in the proof of Theorem 3, except that there, the model yields that the constructed conditional cumulative distribution function of ν given $\varepsilon \leq a$ is right-continuous. Here it follows from assumption (ii). ■

4.2 Non-sharp Bounds

One could in principle estimate bounds on β_1 based on the density inequalities in (5) above. We do not pursue this approach because the resulting estimates would depend on the tails of nonparametrically estimated densities. In this section, we instead present non-sharp bounds based on moments that can be estimated using sample averages.

Equation (11) implies that for $\xi_1 \leq \xi_2$ in the support of $x_{i1} + x_{i2}'\gamma_2$

$$\begin{aligned} &E\left(1\{c_1 < y_i - x_{i2}'\alpha_2 \leq c_2, d_i = 1\} \mid x_i'\gamma = \xi_1\right) \\ &\qquad \leq E\left(1\{c_1 < y_i - x_{i2}'\alpha_2 + (\xi_1 - \xi_2) b_1 \leq c_2, d_i = 1\} \mid x_i'\gamma = \xi_2\right) \end{aligned} \tag{13}$$

for any b_1 in the identified set. Note that c_1 and c_2 in (13) can depend on b_1, ξ_1, and ξ_2.

This implies the moment inequalities

$$\begin{aligned} &E\left[1\{c_1 < y_i - x_{i2}'\alpha_2 - (x_i'\gamma)\beta_1 \leq c_2, d_i = 1\} \mid (x_{i1} + x_{i2}'\gamma_2) \in A_1\right] \\ &\qquad \leq E\left[1\{c_1 < y_i - x_{i2}'\alpha_2 - (x_i'\gamma)\beta_1 \leq c_2, d_i = 1\} \mid (x_{i1} + x_{i2}'\gamma_2) \in A_2\right] \end{aligned} \tag{14}$$

for all sets A_1, A_2 where all elements in A_1 are strictly below the elements in A_2.

[12] Here ν_i is allowed to take the value $-\infty$.

The moment inequalities (14) can be used to estimate non-sharp bounds for β in (6) by considering a finite number of A-sets combined with a finite number of pairs of (c_1, c_2).

The following example suggests that the sizes of these non-sharp identified sets are likely to be small enough to be useful. To do this, we compare the sets to the estimation uncertainty that a researcher would face if she estimated an identified parametric version of the same model.

Example 3 *Consider the data generating process*

- (ν_i, ε_i) *bivariate normal with* $\nu[\nu_i] = 1$, $\nu[\varepsilon_i] = 2$, *cov* $\nu(\nu_i, \varepsilon_i) = 1$, $E[\nu_i] = \frac{1}{2}$ *and* $E[\varepsilon_i] = 0$.
- $x_{ik} = U_{ik} + Z_i$ *for* $k = 1, 2, 3$, *where* $U_{ik} \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$ *and* $Z_i \sim N\left(0, \frac{1}{25}\right)$ *(all independent)*
- $\beta = (1, 1, 1)'$ *and* $\gamma = (0.45, 0.55, 0, 55)$ *(before normalization)*

We calculate the (non-sharp) identified region for β_1 *based on equation (14) with the A's based on quintiles of* $x_{j1} + x_{j2}'\gamma_2$ *and* c_1 *and* c_2 *adjacent deciles of* $y_i - x_{i2}'\alpha_2 - (x_{i1} + x_{i2}'\gamma_2)\, b_1$ *to be* $(0.658, 1.003)$. *When we decreased the number of inequalities by only considering* $A_1 = \left(-\infty, \text{median}\left(x_{j1} + x_{j2}'\gamma_2\right)\right)$ *and* $A_2 = \left(\text{median}\left(x_{j1} + x_{j2}'\gamma_2\right), \infty\right)$ *and* c_1 *and* c_2 *adjacent quintiles of* $y_i - x_{i2}'\alpha_2 - (x_{i1} + x_{i2}'\gamma_2)\, b_1$, *the (non-sharp) identified region for* β_1 *increased to* $(0.529, 1.031)$.

By comparison, the 5^{th} *and* 95^{th} *percentiles of Heckman's two-step estimator for* β_1 *based on 1,000 observations from this design are 0.332 and 1.714.*[13]

4.3 Empirical Illustration Part 2

To investigate the usefulness of the approach from Section 4.2 in empirical settings, we return to the question in Section 2. In this application, the parameter of interest

[13] The bounds based on (14) are calculated using a sample with 100,000,000 observations. The percentiles of Heckman's two-step estimator are calculated in Matlab by Monte Carlo using 100,000 replications.

is the coefficient on being third-generation Mexican-American as opposed to non-Hispanic white. The other explanatory variables are age, age-squared, experience, experience-squared, education dummies (less than high school, some college, college, and advanced degree, with high school as the omitted category), dummies for being a veteran and being married, state dummies, and year dummies.

We first estimate the model under the assumption of joint normality of the errors, using both the maximum likelihood estimator and Heckman's two-step estimator. The estimation results are presented in the first two columns of Tables 2 and 3. To implement the idea in Section 4.2, we define a sample analog of the solutions to the population inequalities in equation (14) as the maximizers of $Q_n(b_1)$ where

$$\begin{aligned} Q_n(b_1) = -\sum_{\ell,k} \max\Big\{ & \widehat{E}\left[1\left\{c_\ell < y_i - x_{i2}'\widehat{\alpha}_2 - (x_i'\widehat{\gamma})\, b_1 \le c_{\ell+1}, d_i = 1\right\} \middle| \, x_i'\widehat{\gamma} \in A_k\right] \\ & - \widehat{E}\left[1\left\{c_\ell < y_i - x_{i2}'\widehat{\alpha}_2 - (x_i'\widehat{\gamma})\, b_1 \le c_{\ell+1}, d_i = 1\right\} \middle| \, x_i'\widehat{\gamma} \in A_{k+1}\right], 0\Big\}^2 . \end{aligned}$$

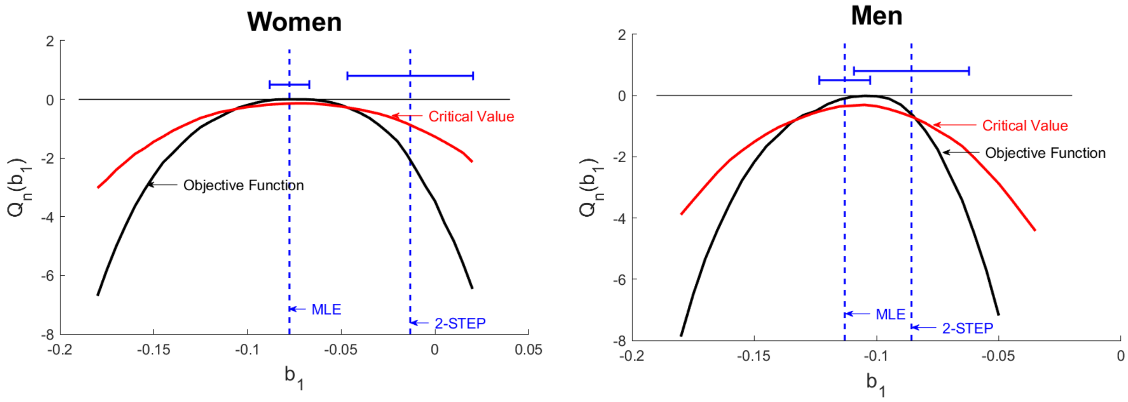
Figure 5 displays the objective function and the 5%-critical value function calculated using sub-sampling (see Canay and Shaikh (2017)) with sub-sample size equal to 15,000 and 1,000 sub-samples. The parameter γ is estimated by logit maximum likelihood[14] and $\alpha_2 = (\beta_2 - \gamma_2\beta_1)$ is estimated from (9), where the conditional expectations are estimated by kernel regressions with standard normal kernel and bandwidth equal to 0.2 times the standard deviation of $x_i'\widehat{\gamma}$ (in the sample where y_i is observed). We choose $c_1 = -\infty$, c_2 to c_9 are the deciles of $\{y_i - x_{i2}'\widehat{\alpha}_2 - (x_i'\widehat{\gamma})\, b_1\}$ in the sample where y_i is observed and $c_{10} = \infty$. The sets A_k corresponds to the intervals between quintiles of $x_i'\widehat{\gamma}$. All parameters, including the bandwidths in the kernel regressions, and c_2 to c_9, are re-calculated in each subsample. The figure also displays the maximum likelihood estimator and Heckman's two-step estimator for β_1 along with their 95% confidence intervals. The estimated bounds[15] on the parameters

[14] Alternatively, one could use a semiparametric estimator such as Han (1987)'s maximum rank correlation estimator in the first step.

[15] As explained in Manski and Tamer (2002), the set of maximizers of Q_n will not (in general) yield a consistent estimator of the identified region of β_1. It is therefore customary to define the estimator as the set of points for which the objective function is within a small distance from its

are presented in the third column of Tables 2 and 3. The 95% confidence interval for the log-wage differentials between Mexican-Americans and Non-Hispanic white Americans case are the points in Figure 5 for which Q_n is larger than the critical value function. They are $(-0.107, -0.046)$ for women and $(-0.134, -0.085)$ for men.

Figure 5: Q_n as a Function of the Coefficient on 3rd Generation Mexican-American



As expected, the implied confidence intervals for the identified sets are longer than the confidence intervals based on the maximum likelihood estimators. On the other hand, they are roughly equivalent to the length of the confidence intervals for the two-step estimator. Our set estimate contains the maximum likelihood estimate for both samples. For men, it is also close to the two-step estimate. For women, the two-step estimate is, however, quite different from our estimated set as well as from the maximum likelihood estimate. This casts doubt on the validity of the normality assumption for women. On the other hand, the moment inequalities implied by the independence assumption (equation (13)) are not rejected by the data in either sample.

maximum. For the bounds in Tables 2 and 3, we have chosen this distance to be 1. Judging from Figure 5, we think that this is a conservative choice.

5 Concluding Remarks

This paper has studied identification in a classical semiparametric sample selection model in which both the selection mechanism and outcome of interest depend linearly on the same explanatory variables, and the errors are independent of the explanatory variables. This model is not semiparametrically point-identified, but the sharp identified set is one-dimensional. Toy examples as well as an empirical application suggest that the identified set can be quite small in practice. In this respect, the practical take-away of this paper is similar to papers in different areas of economics which have demonstrated that the identified regions of non-identified parameters can be small enough to be useful in empirical applications. The papers by Haile and Tamer (2003), Honoré and Lleras-Muney (2006), and Blundell, Gosling, Ichimura, and Meghir (2007) are early examples of this.

The numerical calculations presented in this paper illustrate that the bounds obtained under the semiparametric model considered here are much tighter than those obtained in Lee (2009)'s nonparametric setting. We leave it for future research to investigate intermediate assumptions that are weaker than those imposed here, but strong enough to generate identified sets that are small enough to be empirically informative.

References

Ahn, H., and J. L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58(1-2), 3–29.

Andrews, D. W. K., and M. M. A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *The Review of Economic Studies*, 65(3), 497–517.

Barrow, L., and C. E. Rouse (2017): "Financial Incentives and Educational

Investment: The Impact of Performance-Based Scholarships on Student Time Use," *Education Finance and Policy.*

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(2), 323–363.

Canay, I. A., and A. M. Shaikh (2017): "Practical and Theoretical Advances in Inference for Partially Identified Models," in *Advances in Economics and Econometrics: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, vol. 2, pp. 271–306. Cambridge University Press.

Chamberlain, G. (1986): "Asymptotic Efficiency in Semi-Parametric Models with Censoring," *Journal of Econometrics*, 32(2), 189–218.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *The Review of Economic Studies*, 70(1), 33–58.

Durrett, R. (2019): *Probability: Theory and Examples*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edn.

Escanciano, J. C., D. Jacho-Chvez, and A. Lewbel (2016): "Identification and estimation of semiparametric two-step models," *Quantitative Economics*, 7(2), 561–589.

Haile, P., and E. Tamer (2003): "Inference with an Incomplete Model of English Auctions," *Journal of Political Economy*, 111(1), 1–51.

Han, A. (1987): "Nonparametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35, 303–316.

Heckman, J. J. (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(4), 475–92.

——— (1979): “Sample Selection Bias as a Specification Error,” *Econometrica*, 47(1), 153–61.

——— (1990): “Varieties of Selection Bias,” *The American Economic Review*, 80(2), 313–318.

Honoré, B. E., and A. Lleras-Muney (2006): “Bounds in Competing Risks Models and the War on Cancer,” *Econometrica*, 74(6), 1675–1698.

Kitagawa, T. (2015): “A Test for Instrument Validity,” *Econometrica*, 83(5), 2043–2063.

Krueger, A. B., and D. M. Whitmore (2001): “The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project Star,” *The Economic Journal*, 111(468), 1–28.

Lee, D. S. (2009): “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects,” *The Review of Economic Studies*, 76(3), 1071–1102.

Manski, C. F. (1989): “Anatomy of the Selection Problem,” *The Journal of Human Resources*, 24(3), 343–360.

——— (1990): “Nonparametric Bounds on Treatment Effects,” *The American Economic Review*, 80(2), 319–323.

Manski, C. F., and E. Tamer (2002): “Inference on Regressions with Interval Data on a Regressor or Outcome,” *Econometrica*, 70(2), 519–546.

Mora, R. (2008): “A nonparametric decomposition of the Mexican American average wage gap,” *Journal of Applied Econometrics*, 23(4), 463–485.

Muthén, B. (1990): “Moments of the censored and truncated bivariate normal distribution,” *British Journal of Mathematical and Statistical Psychology*, 43(1), 131–143.

Powell, J. L. (1987): "Semiparametric Estimation of Bivariate Latent Models," Working Paper no. 8704, Social Systems Research Institute, University of Wisconsin–Madison.

——— (1994): "Estimation of Semiparametric Models," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, no. 4 in Handbooks in Economics,, pp. 2443–2521. Elsevier, North-Holland, Amsterdam, London and New York.

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56(4), 931–954.

Appendix 1: Proofs of Propositions

Proof of Proposition 1. This proposition is a special case[16] of Proposition 2. Here we provide a more readable proof that explicitly uses properties of the normal distribution. Recall that if

$$\left.\begin{pmatrix} \nu \\ y \end{pmatrix}\right| x \sim N\left(\begin{pmatrix} \mu \\ x'\beta \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}\right),$$

then

$$\nu|\, y, x \sim N\left(\mu + \frac{\rho}{\sigma}\left(y - x'\beta\right), 1 - \rho^2\right).$$

Hence

$$\begin{aligned} f\left(y|\, \nu > -x, x\right) &= \frac{f\left(y|\, x\right) P\left(\nu > -x|\, y, x\right)}{P\left(\nu > -x|\, x\right)} \\ &= \frac{1}{\sigma}\varphi\left(\frac{y - x'\beta}{\sigma}\right)\Phi\left(\frac{x + \mu + \frac{\rho}{\sigma}\left(y - x'\beta\right)}{\sqrt{1-\rho^2}}\right)\Big/\ \Phi\left(x + \mu\right) \end{aligned}$$

and therefore

$$f_y\left(\cdot|\, x\right) = \frac{1}{\sigma}\varphi\left(\frac{y - x'\beta}{\sigma}\right)\Phi\left(\frac{x + \mu + \frac{\rho}{\sigma}\left(y - x'\beta\right)}{\sqrt{1-\rho^2}}\right).$$

Now consider a b in the identified region, $\mathbf{B}$. For that b, the inequality $f_y\left(c|\, x = 0\right) \leq f_y\left(c + b|\, x = 1\right)$ holds for all values of c. This can be written as

$$\frac{f_y\left(c|\, x = 0\right)}{f_y\left(c + b|\, x = 1\right)} \leq 1.$$

[16] We have kept this because the concreteness of the calculation helped us understand the results.

Under normality, the inequality becomes

$$\begin{aligned}\frac{f_y\left(c|\,x=0\right)}{f_y\left(c+b|\,x=1\right)} &= \frac{\frac{1}{\sigma}\varphi\left(\frac{c}{\sigma}\right)\Phi\left(\frac{\mu+\frac{\rho}{\sigma}c}{\sqrt{1-\rho^2}}\right)}{\frac{1}{\sigma}\varphi\left(\frac{c+b-\beta}{\sigma}\right)\Phi\left(\frac{1+\mu+\frac{\rho}{\sigma}(c+b-\beta)}{\sqrt{1-\rho^2}}\right)} \\ &= \exp\left(\left(b-\beta\right)c+\left(b-\beta\right)^2\big/\,2\sigma^2\right)\frac{\Phi\left(\frac{\mu+\frac{\rho}{\sigma}c}{\sqrt{1-\rho^2}}\right)}{\Phi\left(\frac{1+\mu+\frac{\rho}{\sigma}(c+b-\beta)}{\sqrt{1-\rho^2}}\right)} \leq 1.\end{aligned}$$

Now assume that $\rho > 0$ and consider the limit as $c \to \infty$. If $b > \beta$, the first term in the product increases to ∞, while the second term converges to 1. This contradicts the inequality, and we conclude that $b \leq \beta$. Hence β is the upper endpoint of $\mathbf{B}$.

When $\rho < 0$, we consider the limit as $c \to -\infty$ and conclude that $b \geq \beta$. Hence β is the lower endpoint of $\mathbf{B}$.

Finally, when $\rho = 0$, the inequality becomes

$$\left(b-\beta\right)c \leq -\log\left(\frac{\Phi\left(\mu\right)}{\Phi\left(1+\mu\right)}\right) - \left(b-\beta\right)^2\big/\,2$$

for all values of c. This can only be true if $b = \beta$, and β is point-identified.

This completes the proof. ■

Proof of Proposition 2.

Assumptions:

1. The distribution of ν given $\varepsilon = c_1$ stochastically dominates the distribution of ν given $\varepsilon = c_2$ if $c_1 > c_2$.

2. The density of ε has sufficiently thin tails that for $a > 0$, $f\left(c\right)/f\left(c+a\right) \to \infty$ as $c \to \infty$ and for $a < 0$, $f\left(c\right)/f\left(c+a\right) \to \infty$ as $c \to -\infty$.

Recall that

$$\begin{aligned} f_y\left(c|\,x\right) &= f_{y^*}\left(c\right)P\left(\nu > -x|\,y^* = c\right) \\ &= f_\varepsilon\left(c - x\beta\right)P\left(\nu > -x|\,\varepsilon = c - x\beta\right). \end{aligned}$$

Now consider a b in the identified region, $\mathbf{B}$. For that b, the inequality $f_y\left(c|\,x=0\right) \leq f_y\left(c+b|\,x=1\right)$ holds for all values of c. In other words

$$\frac{f_y\left(c|\,x=0\right)}{f_y\left(c+b|\,x=1\right)} = \frac{f_\varepsilon\left(c\right)P\left(\nu>0|\,\varepsilon=c\right)}{f_\varepsilon\left(c+b-\beta\right)P\left(\nu>-1|\,\varepsilon=c+b-\beta\right)} \leq 1.$$

Suppose that $b>\beta$. Then

$$\frac{P\left(\nu>0|\,\varepsilon=c\right)}{P\left(\nu>-1|\,\varepsilon=c+b-\beta\right)} > \frac{P\left(\nu>0|\,\varepsilon=c\right)}{1}.$$

The right hand side is increasing in c by the stochastic dominance assumption. Hence it is bounded from below by some positive constant, k. Therefore

$$\frac{f_y\left(c|\,x=0\right)}{f_y\left(c+b|\,x=1\right)} > k\frac{f_\varepsilon\left(c\right)}{f_\varepsilon\left(c+b-\beta\right)},$$

where the ratio on the right hand side increases to ∞ as c goes to ∞. This contradicts the inequality, and we conclude that no b in $\mathbf{B}$ can be greater than the true β. Hence β is the upper endpoint of $\mathbf{B}$.

Now consider the case where the distribution of ν given $\varepsilon=c_1$ stochastically dominates the distribution of ν given $\varepsilon=c_2$ if $c_1<c_2$. Suppose that $b<\beta$. Then again

$$\frac{P\left(\nu>0|\,\varepsilon=c\right)}{P\left(\nu>-1|\,\varepsilon=c+b-\beta\right)} > P\left(\nu>0|\,\varepsilon=c\right) > P\left(\nu>0|\,\varepsilon=0\right)$$

for all $c<0$. Therefore

$$\frac{f_y\left(c|\,x=0\right)}{f_y\left(c+b|\,x=1\right)} > k\frac{f_\varepsilon\left(c\right)}{f_\varepsilon\left(c+b-\beta\right)}$$

for $c<0$. Taking the limit as $c\to-\infty$ brings the right hand side above 1, and we conclude that a b for which $b<\beta$ cannot belong to the set $\mathbf{B}$. Hence β is the lower endpoint of $\mathbf{B}$.

Finally, when ε and ν are independent, the inequality defining $\mathbf{B}$ is

$$\frac{f_\varepsilon(c)\,P(\nu > 0|\,\varepsilon = c)}{f_\varepsilon(c+b-\beta)\,P(\nu > -1|\,\varepsilon = c+b-\beta)} = \frac{f_\varepsilon(c)\,P(\nu > 0)}{f_\varepsilon(c+b-\beta)\,P(\nu > -1)} \leq 1.$$

Taking the limit as $c \to -\infty$ generates a contradiction when $b < \beta$ and taking the limit as $c \to \infty$ generates a contradiction when $b > \beta$. Therefore $\mathbf{B} = \{\beta\}$.

This completes the proof. ■

Proof of Proposition 3. First consider case 1 (with positive selection). The identified set can be written as

$$\begin{aligned}
\mathbf{B} &= \big\{b \in \mathbb{R} : f_{y|x}(c+xb|\,\xi_1) \leq f_{y|x}(c+xb|\,\xi_2) \\
&\qquad\qquad \text{for all values of } c \text{ and } \xi_1 < \xi_2 \text{ in the support of } x\big\} \\
&= \bigcap_{\xi_1<\xi_2} \big\{b \in \mathbb{R} : f_{y|x}(c+\xi_1 b|\,\xi_1) \leq f_{y|x}(c+\xi_2 b|\,\xi_2) \text{ for all values of } c\big\} \\
&= \bigcap_{\xi_1<\xi_2} \big\{b \in \mathbb{R} : f_{y-\xi_1 b|x}(c|\,\xi_1) \leq f_{y-\xi_1 b|x}(c+(\xi_2-\xi_1)\,b|\,\xi_2) \text{ for all values of } c\big\}.
\end{aligned}$$

Now consider one of the sets on the right hand side above. By Proposition 2, the upper limit of $(\xi_2 - \xi_1)\,b$ will be $(\xi_2 - \xi_1)\,\beta$. This implies that the upper limit of all the sets in the intersection above is the true β. Hence, the upper limit on the intersection is β.

The proofs of cases 2 and 3 are similar. ■

Appendix 2: Data Details

This analysis utilizes the Merged Outgoing Rotation Groups (MORG) files of the Current Population Survey (CPS), which were prepared by the National Bureau of Economic Research (NBER). Following Mora, we restrict our sample to non-Hispanic whites and Mexican-Americans between the ages of 25 and 62 (inclusive) who live in Arizona, California, New Mexico, or Texas. We further limit our analysis to those who have at least one parent born in the United States (i.e., third-generation Americans).

We also drop the top 1.67% of earners in each year's income distribution from our analysis, and we multiply top-coded earnings by 1.33. Finally, in our wage samples, we exclude self-employed workers, as well as individuals who report that they are working but do not report either hours worked or earnings.

The variables are:

- Log hourly wage: Calculated by taking the natural log of an individual's weekly earnings divided by his usual hours works, adjusted for inflation.

- Veteran status: Indicator variable that equals one if an individual ever reported serving in the U.S. military, and zero otherwise.

- Married: Indicator variable that equals one if an individual reports that she or he is either (1) a married civilian with spouse present, (2) a married Armed Forces member with spouse present, or (3) married with spouse absent or separated, and zero otherwise.

- Experience: For individuals who have completed at least seventh grade, their labor market experience is defined as their age (in single years) minus their education-years minus 6. Individuals whose educational attainment is less than seventh grade are assigned an experience level equal to their age minus thirteen.

- Education-years: Following Mora, we assign education-years based on the level of education attainment reported in the data as follows:

    - Less than 1st grade = 0 years of education
    - 1st – 4th grade = 2.5 years of education
    - 5th or 6th grade = 5.5 years of education
    - 7th or 8th grade = 7.5 years of education
    - 9th = 9 years of education
    - 10th = 10 years of education
    - 11th = 11 years of education

- 12th grade (no diploma) = 12 years of education
- High school graduate, diploma, or GED = 12 years of education
- Some college but no degree = 13.5 years of education
- Associate degree – occupational/vocational =14 years of education
- Associate's degree – academic program = 14 years of education
- Bachelor's degree (i.e. BA, AB, BS) = 16.5 years of education
- Master's degree (i.e. MA, MS, MEng, MSW, MBA) = 18 years of education
- Professional school degree (i.e. MD, DDS, DVM, LLB, JD) = 18 years of education
- Doctorate degree (i.e. PhD, EdD) = 20 years of education

Table 1: Summary Statistics

	Mexican Women		White Women		Mexican Men		White Men	
	mean	sd	mean	sd	mean	sd	mean	sd
California	0.38	0.49	0.46	0.50	0.37	0.48	0.47	0.50
Arizona	0.08	0.26	0.11	0.31	0.08	0.28	0.11	0.31
Texas	0.47	0.50	0.34	0.47	0.46	0.50	0.34	0.47
Real Wage	2.16	0.57	2.42	0.63	2.34	0.58	2.61	0.61
Working	0.64	0.48	0.61	0.49	0.71	0.45	0.67	0.47
Age	40.66	10.76	44.34	10.81	40.56	10.74	43.99	10.92
Experience	21.63	11.19	23.88	11.15	21.65	10.97	23.63	11.08
Less than HS	0.16	0.37	0.04	0.20	0.16	0.37	0.05	0.21
Some College	0.33	0.47	0.34	0.48	0.31	0.46	0.33	0.47
College	0.12	0.32	0.27	0.44	0.11	0.31	0.26	0.44
Advanced Degree	0.05	0.21	0.12	0.33	0.04	0.18	0.12	0.32
Married	0.53	0.50	0.62	0.49	0.55	0.50	0.61	0.49
Veteran	0.01	0.10	0.02	0.13	0.12	0.32	0.16	0.37
No. Observations	26,698		103,209		21,402		97,016	

Table 2: Estimated Wage Regression (Women)

	MLE	Two-Step	Estimated Bounds
Mexican–American	-0.078	-0.013	$[-0.086, -0.080]$
	(0.005)	(0.017)	
Age	0.113	0.213	$[0.096, 0.106]$
	(0.007)	(0.026)	
Age–squared	-0.047	-0.118	$[-0.000, -0.000]$
	(0.005)	(0.018)	
Experience	-0.070	-0.127	$[-0.067, -0.062]$
	(0.006)	(0.016)	
Experience–squared	0.006	0.036	$[-0.000, -0.000]$
	(0.004)	(0.009)	
Less than HS	-0.177	-0.372	$[-0.193, -0.178]$
	(0.015)	(0.050)	
Some College	0.033	0.017	$[0.026, 0.028]$
	(0.011)	(0.014)	
College	0.155	0.084	$[0.136, 0.142]$
	(0.025)	(0.036)	
Advanced Degree	0.199	0.113	$[0.167, 0.174]$
	(0.034)	(0.047)	
Veteran	0.030	0.037	$[0.029, 0.030]$
	(0.016)	(0.020)	
Married	0.033	-0.079	$[0.042, 0.052]$
	(0.005)	(0.028)	
California	0.204	0.178	$[0.206, 0.208]$
	(0.007)	(0.011)	
Arizona	0.098	0.103	$[0.097, 0.097]$
	(0.009)	(0.012)	
Texas	0.031	0.064	$[0.025, 0.028]$
	(0.008)	(0.013)	
Year Dummies	yes	yes	yes
No. Observations	127,738	127,738	127,738

Standard errors for point identified parameters in parentheses.

Table 3: Estimated Wage Regression (Men)

	MLE	Two-Step	Estimated Bounds
Mexican–American	-0.113	-0.084	$[-0.109, -0.097]$
	(0.005)	(0.012)	
Age	0.079	0.112	$[0.077, 0.091]$
	(0.006)	(0.014)	
Age–squared	-0.047	-0.072	$[-0.001, -0.000]$
	(0.004)	(0.010)	
Experience	-0.025	-0.045	$[-0.032, -0.023]$
	(0.005)	(0.009)	
Experience–squared	-0.014	-0.007	$[-0.000, -0.000]$
	(0.004)	(0.005)	
Less than HS	-0.170	-0.222	$[-0.193, -0.174]$
	(0.012)	(0.023)	
Some College	0.051	0.043	$[0.048, 0.050]$
	(0.009)	(0.011)	
College	0.235	0.205	$[0.222, 0.232]$
	(0.023)	(0.026)	
Advanced Degree	0.257	0.194	$[0.231, 0.254]$
	(0.031)	(0.041)	
Veteran	-0.001	0.015	$[-0.001, 0.005]$
	(0.006)	(0.008)	
Married	0.136	0.185	$[0.133, 0.154]$
	(0.005)	(0.019)	
California	0.151	0.140	$[0.147, 0.151]$
	(0.007)	(0.009)	
Arizona	0.042	0.052	$[0.042, 0.045]$
	(0.009)	(0.010)	
Texas	0.015	0.045	$[0.013, 0.026]$
	(0.008)	(0.014)	
Year Dummies	yes	yes	yes
No. Observations	118, 250	118, 250	118, 250

Standard errors for point identified parameters in parentheses.